Perspectives on Adversarial Machine Learning in Intelligent Information Systems

Zhuoran Liu

Institute for Computing and Information Sciences

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Perspectives on Adversarial Machine Learning in Intelligent Information Systems

Zhuoran Liu

This work was partially funded by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). Part of this work was carried out on the Dutch national einfrastructure with the support of SURF Cooperative.

Perspectives on Adversarial Machine Learning in Intelligent Information Systems

Zhuoran Liu

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Zhuoran Liu
Cover: Lexi Liu, Proefschrift AIO
Printing: DPN Rikken/Pumbo

ISBN: 9789465150727
DOI: 10.54195/9789465150727
Free download at: https://doi.org/10.54195/9789465150727

RADBOUD
UNIVERSITY
PRESS

Perspectives on Adversarial Machine Learning in Intelligent Information Systems

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 17 april 2025
om 12.30 uur precies

door

Zhuoran Liu

geboren op 30 januari 1990
te Qinghai, China

Promotor:

Prof. dr. Martha Larson

Manuscriptcommissie:

Prof. dr. Marco Loog (voorzitter)
Prof. dr. Marcel Worring (Universiteit van Amsterdam)
Dr. Jennifer Williams (University of Southampton, Verenigd Koninkrijk)

Perspectives on Adversarial Machine Learning in Intelligent Information Systems

Dissertation

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Thursday, April 17, 2025
at 12.30 pm

by

Zhuoran Liu

born on January 30, 1990
in Qinghai, China

Supervisor:

Prof. dr. Martha Larson

Manuscript Committee:

Prof. dr. Marco Loog (Chair)
Prof. dr. Marcel Worring (University of Amsterdam)
Dr. Jennifer Williams (University of Southampton, United Kingdom)

Contents

Summary

Information systems provide users with tailored information services by collecting, processing, and managing data from diverse sources. Two important, representative, modern intelligent information systems are retrieval systems, which take a user query as input to provide query-relevant information, and recommender systems, which take user-item interaction logs as input to provide related personalized recommendations. The current advancement of intelligent information systems is substantially driven by machine learning, especially deep learning, where models are built on large-scale data sets.

Accumulated externally sourced data is an essential part of the data used in building intelligent systems, where user-contributed media content and interactions are taken to represent users' preferences and to model patterns. The current state-of-the-art commercial intelligent systems are primarily advanced by leveraging externally sourced content with the help of machine learning. However, the data-driven nature of these systems gives rise to issues and opportunities for both owners and users of intelligent information systems. On the one hand, system owners cannot guarantee the integrity and quality of externally sourced data, so publicly accessible systems may encounter unknown data-caused threats. Such threats, when exploited by malicious parties, could severely harm the end users and owners of information systems. On the other hand, users may be unaware of and disagree with how intelligent systems are using and benefiting from their data, especially when they are not certain about their willingness to consent. Misuse of externally sourced data could also potentially harm the users and, later, the owners in the long term.

Adversarial machine learning studies the threats to machine learning models, where most attacks focus on data modifications by exploiting the models' data dependency. This thesis focuses on implications that arise for intelligent information systems because these systems make use of data that is drawn from the outside and can be modified by a user standing outside of the system using adversarial machine learning. The issues arise for all possible different types of input and they fall into two categories. First, from the system owner's perspective, adversarial machine learning can give rise to security issues. Second, from the user's perspective, it can represent opportunities for users to improve their privacy or protect their data. Each chapter in this thesis takes one of the two perspectives on adversarial machine learning in intelligent information systems.

In Chapter 2, from the system owner's perspective, we explore the influence of externally sourced adversarial items in the background collections of recommender systems. We investigate the threats to representative recommender systems that

use images to address the cold start, including systems with existing countermeasures. In particular, we look into the practical vulnerabilities of visually-aware recommender systems by conducting adversarial item promotion in different threat scenarios where adversaries have gradually less knowledge of the system. We demonstrate that adversarial images targeting the recommendation ranking mechanism may open recommender systems to potential adversarial threats.

In Chapter 3, from the user's perspective, we show that users can protect their data by poisoning, but special attention needs to be paid to stronger adversaries. In particular, we revise the methodologies of availability poisoning for data misuse protection and find that poisoning samples are surrogate-dependent. According to this finding, we introduce a series of compression-based mitigation methods and demonstrate their effectiveness against different types of poisoning methods. In addition, we conduct an in-depth analysis of poisons' dependency on different training stages of surrogate models and provide an analysis of possible adaptive poisoning methods against compression-based mitigation methods. We show that availability poisoning is fragile but still promising in mitigating the misuse of externally sourced data for training.

In Chapter 4, from the user's perspective, we examine and mitigate the privacy risks of externally sourced profiles against bag-based attribute profiling. We provide experiments showing that deep bag-based profile-level classifiers pose a strong privacy threat. Especially, bag-based classifiers that use early or intermediate fusion are potentially more dangerous than approaches that use late fusion, i.e., predict at the item level before aggregating to reach a final prediction. We introduce three pivoting additions to resist bag-based profiling, which we study under different threat scenarios. We show that it is possible for users to resist bag-based attribute profiling by adversarially adding pivoting additions to existing profiles.

In Chapter 5, from the user's perspective, we investigate the influence of externally sourced adversarial image queries on content-based image information retrieval systems. We propose an unsupervised method to generate adversarial image queries that misdirect content-based image retrieval models. We demonstrate the influences of adversarial queries against local, global, and neural feature-based image retrieval systems. We show that the similarity between images in an intelligent information retrieval system can be adjusted in a guided manner to change the results that match a given image query, which benefits users' privacy. Adversarial queries benefit the privacy of users who want to share the image content with others but wish to withhold the semantics.

Based on our findings, we suggest that system owners revisit the necessity of leveraging externally sourced data, and we suggest that users pay attention to potential privacy risks caused by private data exploitation and take the initiative. We emphasize the uncertainty of data collection outside information systems and recommend future research directions to combat privacy and security threats.

Samenvatting

Informatiesystemen bieden gebruikers op maat gemaakte informatiediensten door gegevens uit verschillende bronnen te verzamelen, verwerken en beheren. Twee belangrijke, representatieve, moderne intelligente informatiesystemen zijn retrievalsystemen, die een gebruikersquery als invoer nemen om relevante informatie te bieden, en aanbevelingssystemen, die interactielogboeken van gebruikers en items als invoer nemen om gerelateerde gepersonaliseerde aanbevelingen te doen. De huidige ontwikkeling van intelligente informatiesystemen wordt grotendeels gedreven door machine learning, in het bijzonder deep learning, waarbij modellen worden gebouwd op grootschalige datasets.

Opgebouwde extern verzamelde gegevens vormen een essentieel onderdeel van de data die wordt gebruikt bij het ontwikkelen van intelligente systemen, waarbij door gebruikers aangeleverde media-inhoud en interacties worden beschouwd als representatief voor gebruikersvoorkeuren en worden ingezet om patronen te modelleren. De huidige geavanceerde commerciële intelligente systemen zijn voornamelijk tot stand gekomen door gebruik te maken van externe inhoud in combinatie met machine learning. Het datagestuurde karakter van deze systemen brengt echter zowel problemen als kansen met zich mee voor eigenaren en gebruikers van intelligente informatiesystemen. Enerzijds kunnen systeemeigenaren de integriteit en kwaliteit van extern verkregen gegevens niet garanderen, waardoor openbaar toegankelijke systemen geconfronteerd kunnen worden met onbekende, door data veroorzaakte bedreigingen. Dergelijke bedreigingen, indien uitgebuit door kwaadwillende partijen, kunnen ernstige schade toebrengen aan zowel eindgebruikers als eigenaren van informatiesystemen. Anderzijds zijn gebruikers zich mogelijk niet bewust van, of niet akkoord met, de wijze waarop intelligente systemen hun gegevens gebruiken en hier voordeel uit halen, vooral wanneer ze onzeker zijn over hun bereidheid om hierin bij te dragen.

Adversarial machine learning onderzoekt bedreigingen voor machine learning modellen, waarbij de meeste aanvallen zich richten op het manipuleren van gegevens door gebruik te maken van de afhankelijkheid van modellen van die gegevens. Dit proefschrift richt zich op de implicaties voor intelligente informatiesystemen die gegevens van buitenaf gebruiken en die kunnen worden gewijzigd door een externe gebruiker via adversarial machine learning. Deze problematiek doet zich voor bij alle mogelijke soorten invoergegevens en kan worden onderverdeeld in twee categorieën. Ten eerste kan adversarial machine learning, vanuit het perspectief van de systeemeigenaar, leiden tot beveiligingsproblemen. Ten tweede kan het vanuit gebruikersperspectief juist mogelijkheden bieden om de privacy van gebruikers te verbeteren of hun gegevens beter te beschermen. Elk hoofdstuk in dit proefschrift

behandelt een van deze twee perspectieven op adversarial machine learning binnen intelligente informatiesystemen.

In hoofdstuk 2 onderzoeken we, vanuit het perspectief van de systeemeigenaar, de invloed van extern aangeleverde adversarial items binnen de achtergrondcollecties van aanbevelingssystemen. We analyseren bedreigingen voor representatieve aanbevelingssystemen die afbeeldingen gebruiken om het cold-start-probleem aan te pakken, inclusief systemen met bestaande tegenmaatregelen. In het bijzonder bestuderen we de praktische kwetsbaarheden van visueel-georiënteerde aanbevelingssystemen door middel van adversarial item-promotie binnen verschillende dreigingsscenario's, waarin aanvallers geleidelijk steeds minder kennis van het systeem hebben. We tonen aan dat adversarial afbeeldingen, gericht op het beïnvloeden van het aanbevelings-rangschikkingsmechanisme, aanbevelingssystemen kunnen blootstellen aan potentiële bedreigingen.

In hoofdstuk 3 laten we vanuit gebruikersperspectief zien dat gebruikers hun gegevens kunnen beschermen door middel van poisoning, waarbij bijzondere aandacht nodig is voor sterkere tegenstanders. We herzien hierbij met name de methodologieën van availability poisoning voor bescherming tegen gegevensmisbruik, en tonen aan dat poison-voorbeelden afhankelijk zijn van surrogaatmodellen. Op basis van deze bevinding introduceren we een reeks compressiegebaseerde mitigatiemethoden en demonstreren we de effectiviteit daarvan tegenover verschillende poisoning-aanvallen. Daarnaast voeren we een diepgaande analyse uit naar de afhankelijkheid van poison-gegevens ten opzichte van verschillende trainingsfasen van surrogaatmodellen en geven we een analyse van mogelijke adaptieve poisoning-methoden die zijn gericht tegen compressiegebaseerde mitigatiestrategieën. We tonen aan dat availability poisoning kwetsbaar is, maar desondanks veelbelovend blijft als methode om misbruik van extern verkregen gegevens voor trainingsdoeleinden tegen te gaan.

In hoofdstuk 4 bestuderen en verminderen we, vanuit gebruikersperspectief, de privacyrisico's van extern verkregen gebruikersprofielen bij zogeheten bag-based attribute profiling. We tonen door middel van experimenten aan dat diepe, op "bags" gebaseerde profielclassificatiemodellen aanzienlijke privacybedreigingen vormen. Vooral classificatiemethoden die gebruikmaken van vroege of tussentijdse fusie ("early/intermediate fusion") zijn potentieel gevaarlijker dan methoden die gebruikmaken van late fusie ("late fusion"), dat wil zeggen methoden die eerst op itemniveau voorspellingen doen en deze daarna aggregeren tot een finale voorspelling. We introduceren drie pivot-gebaseerde toevoegingen om weerstand te bieden tegen bag-based profiling, die we vervolgens analyseren in verschillende dreigingsscenario's. We laten zien dat gebruikers zich kunnen beschermen tegen bag-based attribute profiling door op adversariële wijze pivot-toevoegingen aan bestaande profielen te doen.

In hoofdstuk 5 onderzoeken we, vanuit gebruikersperspectief, de invloed van extern verkregen adversarial beeldqueries op inhoud gebaseerde beeldinformatiesystemen. We stellen een niet-gesuperviseerde methode voor om adversarial beeldqueries te

genereren die modellen voor content-based beeldretrieval misleiden. We demonstreren hierbij de invloed van adversarial queries op beeldinformatiesystemen in verschillende scenario's. Daarbij laten we de kwetsbaarheden zien van systemen die gebaseerd zijn op lokale inhoudskenmerken, wanneer deze blootgesteld worden aan adversarial beeldqueries gericht op lokale kenmerken, zonder toegang tot het retrievalmodel zelf. We tonen aan dat adversarial queries de privacy bevorderen van gebruikers die beeldinhoud met anderen willen delen, maar daarbij toch de semantische betekenis willen verbergen.

Op basis van onze bevindingen adviseren wij systeemeigenaren om kritisch te heroverwegen of het gebruik van extern verkregen gegevens noodzakelijk is. Daarnaast adviseren wij gebruikers zich bewust te zijn van potentiële privacyrisico's die voortkomen uit misbruik van persoonlijke gegevens, en hierbij proactief maatregelen te treffen. We benadrukken de onzekerheid van gegevensverzameling buiten informatiesystemen en bevelen toekomstige onderzoeksrichtingen aan om privacy- en veiligheidsbedreigingen te bestrijden.

Chapter 1

Introduction

Intelligent information systems provide tailored services to users to fulfill their information needs. Current state-of-the-art information systems, including retrieval and recommender systems, are driven to a substantial degree by machine learning models built on large-scale data sets. These large-scale data sets mainly consist of externally sourced data including media content and behavioral data.

Adversarial machine learning studies the threats to machine learning models. We focus on *data modification* methods guided by adversarial machine learning. In the thesis, we take two perspectives: the *system owner's perspective* and the *user's perspective*. From the system owner's perspective, adversarial data modification is a threat that can compromise the availability and integrity of intelligent systems. Systems that overly rely on externally sourced data are especially vulnerable. From the user's perspective, adversarial data modification is a promising method to combat potential privacy threats in order to decrease harm. The risks include unauthorized data usage, data usage without users' consent, and profiling users' private attributes, for purposes that users do not agree with.

This thesis brings together intelligent information systems and adversarial machine learning, to gain understanding into the implications of adversarial machine learning from the system owner's perspective and the user's perspective. In this introduction, we motivate the importance of the topic, present the necessary background, and provide and overview of the rest of the thesis.

1.1 Intelligent Information Systems

Intelligent information systems collect, process, and manage data to provide tailored information services. Information retrieval and recommender systems are widely deployed information systems that filter externally sourced data to fulfill users' information needs. Information retrieval systems collect and organize task-related content items to create a *background collection*. In a search session, users issue *queries*, and then similarity-based ranking scores between the queries and the background collection items are calculated, based on which relevant items are

returned to users. Recommender systems process *user-item interactions* to build personalization models that provide filtered recommendations to users. In a recommendation session, the recommender system takes users' *behavioral data* as input to return personalized recommended items.

Users use smart devices to create multimedia content and share it through the web where information systems are deployed as fundamental elements. On the system side, the scale of *externally sourced media content* is rapidly growing, and more data needs to be managed in the information system. On the user side, the exploding amount of accumulated multimedia data makes it harder for end-users to discover the contents they need. For these reasons, retrieval and recommender systems have become important in providing end-users with the content they are interested in. To this end, intelligent information systems need some key components to process the externally sourced data, including *semantic extraction* and *personalized filtering* [144]. Current state-of-the-art intelligent information systems advance due to good management of large-scale externally sourced data and exploitation of this data for machine learning.

Learning-based intelligent information systems include different modules to process media data. Previously, media processing modules of information systems are knowledge-based feature extractors, e.g., SIFT [115] and GIST [129], where learning models are trained on engineered features. Deep neural networks depend heavily on large-scale data sets, so an intelligent information system's critical driving force is shifting to data. Externally sourced data, which is usually collected from outside sources, is consistently growing in importance as the most widely available data resource for system development. In the next sections, we explain how externally sourced data helps, but also the harms that it can introduce.

1.2 Externally Sourced Data Helps

Several types of externally sourced data, such as externally sourced media content and behavioral data, can enter the information system in different stages [166]. When users interact with information systems, condensed information needs, user-system interactions, user profiles, users' uploaded content, and other externally sourced information are stored, processed, and analyzed to model users' intent and preferences and to learn common data patterns.

Externally sourced data is critical because it represents users' preferences. On the one hand, users share their data with the service providers, including profiles, search queries, uploaded content, interaction history, etc, when interacting with information systems. Based on the collection of personalized user-related data, user-specific tailored services can be provided [176; 35; 150], which can improve user satisfaction, engagement, and retention.

On the other hand, service providers can train large data-driven models on publicly available externally sourced data to improve the general system performance [148;

147; 164]. Better systems can provide more accurate information service that fulfills users' information needs.

In general, users are willing to interact with intelligent information systems in return for high-quality services, and the systems keep updating themselves based on externally sourced data to model the users' interests better. Such a mechanism, boosted by learning models, is seen as important to keep improving service providers' profits and users' engagement. Externally sourced data has become an indispensable part of the development of learning-based information systems.

At the same time, systems are becoming greedy for externally sourced data because system performance is highly correlated with the scale of collected data. For example, fine-grained personalized information systems, such as recommender systems, need more private session data to provide accurate, personalized services [158]. Chatbot applications need more web-scraped corpora to train the large language models [2].

1.3 Externally Sourced Data Harms

From the system owner's perspective, exploiting externally sourced data can be harmful, compromising the availability and integrity of data-driven information systems. For example, a poisoning attack can be conducted to manipulate a small proportion of commonly used data sets if the integrity check is missing in the future download [19]. Malicious parties can manipulate part of the data set to mislead the information system. The security of information systems is important when exploiting externally sourced data from the system owner's perspective.

From the user's perspective, the learning algorithm's data-driven and black-box nature also makes it difficult for users, the people who provide the data, to gain insight into intelligent information systems' decision-making process. Users may disagree with their assigned *collective tag* in the system. For example, in a recommender system, users can conjecture that they receive recommendations based on similar users but may disagree with the group that defines the similarity. Users may also not agree with the information systems to profile other users using models built on their generated content. A prime example is the Facebook Cambridge Analytica data scandal [79] where the externally sourced data was used for the purposes that the data generators are unaware of and possibly harmed.

Malicious machine learning further worsens the situation by scraping and exploiting data on the web. Users may disagree with the design *purposes* of malicious systems that exploit their data. For example, Flickr users who share images on social media may not agree with the potential that their data could be used in military applications [55]. Meanwhile, users may not be aware of the *sensitive redundant information* in the media they share. For example, the personality trait of a user can be predicted based on a set of externally sourced images. This lack of transparency between the users' intents and the information about users exploited

by the system leads to a trust illusion for users that only non-sensitive, non-harmful information is shared with the information systems for purposes they agree with.

Conventional methods attempt to address the problem from the service provider side. For example, bias-aware algorithms are deployed in practical information systems [58]. ICT law is advancing to protect users [95]. Adversarial machine learning is promising to empower users to protect themselves. It holds promise to complement and support both legal and technical approaches to protecting privacy. In this thesis, we study the potential adversarial machine learning methods that modify externally sourced data entering the information systems. From the system owner's perspective we look at the threats posed by adversarial machine learning. From the user's perspective, we look at the potential of adversarial machine learning to push back against the potential harms caused to users by information systems exploiting externally sourced data.

1.4 Adversarial Machine Learning

Adversarial machine learning investigates the vulnerabilities of machine learning algorithms to malicious inputs that are designed to deceive or manipulate these algorithms. Adversarial machine learning research typically focuses on studying the security and robustness of machine learning algorithms. In this thesis, adversarial machine learning provides the opportunity to explore the influence of externally sourced data on information systems, focusing on investigating systems' security and improving users' privacy.

Several heuristic methods explored the influence of externally sourced data on information systems [64], e.g., the bandwagon methods. With the development of learning-based information systems, adversarial machine learning sheds light on investigating the impacts of externally sourced data. Current adversarial machine learning methods that modify data can be categorized as evasion and poisoning. Evasion methods take place in the inference stage of the machine learning model. A specifically modified input can confidently deceive the trained model into making a wrong prediction, while the modifications to the input are mostly non-suspicious to humans. Given the classification model information, an adversarial example generated by an iterative evasion method PGD [118] can be formulated as,

$$\boldsymbol{x}^{t+1} = \Pi\left(\boldsymbol{x}^{t} + \alpha \cdot \operatorname{sign}(\nabla_{\boldsymbol{x}} \mathcal{L}(F(\boldsymbol{x}^{t}; \boldsymbol{\theta}), y))\right) \tag{1.1}$$

where $\boldsymbol{\theta}$ represents the parameters of the classifier F. $\mathcal{L}$ is the model loss, t is the number of iterations, y is the classification target, and $\boldsymbol{x}^t$ is the adversarial example at step t. α represents the step size, and Π is the projection function. Adversarial data can be generated with respect to a surrogate model loss that can be formulated to represent users' or attackers' intents.

Poisoning methods take place in the training stage of a machine learning model, where the training data of a learning model is modified to achieve pre-defined goals. Specifically, backdoor or availability poisoning manipulates training data by adding triggers or noise to undermine the model's performance.

Given a classification model, availability poisoning samples can be generated by solving the following objective,

$$\max_{\boldsymbol{\delta}} \ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\Big[\mathcal{L}\left(F(\boldsymbol{x};\boldsymbol{\theta}'(\boldsymbol{\delta})),y\right)\Big] \tag{1.2}$$

$$\text{s.t. } \boldsymbol{\theta}'(\boldsymbol{\delta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{S}} \mathcal{L}(F(\boldsymbol{x}_i+\boldsymbol{\delta}_i;\boldsymbol{\theta}),y_i), \|\boldsymbol{\delta}\|_p \leq \epsilon, \tag{1.3}$$

where $\boldsymbol{\theta}'(\boldsymbol{\delta})$ represents the parameters of the poisoned classifier F. $\boldsymbol{\delta}$ denotes the additive perturbations whose L_p norm is restricted by ϵ.[1] $\mathcal{L}$ is the model loss, which takes as input a pair of model output $F(\boldsymbol{x}_i;\boldsymbol{\theta})$ and the corresponding label y_i. $\mathcal{S}$ represents the training set, and $\mathcal{D}$ is the test set. Availability poisoning modifies training data to influence the performance of learning models on real-world input data, which can be used by users who do not authorize the usage of their data for training.

Externally sourced data can be adversarially modified following the guidance from the pre-defined objective. Such modification could be made imperceptible to humans but effective against learning models that are working as the core of an information system [215]. Transfer learning-based techniques can gradually decrease the modifications' dependencies on a specific learning model to increase its general utility [216].

We show that the adversarial machine learning techniques, even at an early stage of development as a protection tool, is a promising solution to narrow the gap of control between information system owners and users. From the information system owner's perspective, adversarial machine learning can be used to attack the system, but it can also be used to help to distill the specific data systems need, strengthening the system's robustness and explainability. From the user's perspective, adversarial machine learning helps data processors strategically modify their data to eliminate what they don't want to expose to improve privacy. Leveraging adversarial machine learning, this thesis focuses on externally sourced data and explores the influence of adversarial data modification in different stages of information systems.

[1] Note that Eq. 1.3 is relevant for Ch. 3 and occurs again as Eq. 3.2. Note that there is a small difference in the two, namely, in Eq. 1.3 the specification of the δ constraint and also the parameterization of theta is explicit and in Eq. 3.2 it is left implicit.

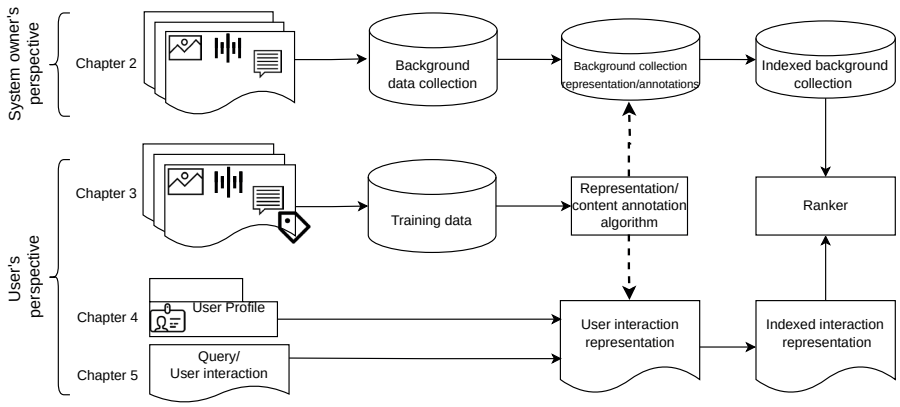


Figure 1.1: Working diagram intelligent information system indicating the stages where externally sourced data can enter the system. From the system owner's perspective, system security, we discuss the influence of adversarial background data (Chapter 2). From the user's perspective, we discuss privacy protection by training data modification (Chapter 3), profile modification (Chapter 4), and query modification (Chapter 5).

Externally sourced data is both a threat to the system owner and an opportunity for users. We examine the security and privacy threats caused by dependence on externally sourced data at each of the four stages of the information system's pipeline at which data can enter the system. Figure 2.1 presents an overall working diagram of an information system illustrating the data processing stages where externally sourced data can enter the system. Different data entry stages discussed in this thesis are marked with corresponding chapter numbers. Externally sourced data are linked to different system modules where the data is processed. In this thesis, we examine information systems from the system owner's perspective and user's perspective using adversarial machine learning and explore the influence of externally sourced data in each stage when these data enter information systems.

Publicly-available externally sourced data can be used directly as background collections. For example, uploaded images on social media platforms and item descriptions in the recommender system can be retrieved or recommended as background items. **Chapter 2** takes the system owner's perspective to look at the externally sourced background collection in information systems. We show that externally sourced background collection data can be exploited to compromise the integrity of recommender systems by promoting specific target items. **Chapter 3** takes the user's perspective to look at externally sourced data that are used as training data to build representation models, where we investigate the potential of availability poisoning for data privacy protection. In addition, externally sourced data represent users' characteristics, where specific data can be used for profiling and personalization. For example, externally sourced (media) profiles and user interactions (e.g., queries and clicks) can be used to predict user attributes. **Chapter 4** takes the user's perspective to look into the user profile that includes media content,

and **Chapter 5** takes the user's perspective to focus on externally sourced image queries, where we explore the adversarial examples to protect the item semantics and *private attributes* that are predictable from user-soured data.

1.5 Thesis Scope

This thesis focuses on issues that arise for intelligent information systems because these systems make use of data that is drawn externally and can be modified by a user standing outside of the system using adversarial machine learning. Issues arise for all possible different types of input and fall into two categories. First, they can represent security issues for the system's owner, and second, they can represent opportunities for users to improve their privacy or protect their data. In this section, we introduce the thesis scope in more detail, chapter by chapter, following the overview in Figure 2.1.

1.5.1 On Background Collection Data

Leveraging the fact that the systems heavily depend on the data, adversaries can modify their data to achieve pre-defined goals. In particular, adversaries can maliciously modify the background collection data to shift the system outputs.

In **Chapter 2**, from the system owner's perspective, we focus on the background collection data in a multimedia recommender system. We look at malicious merchants who want to promote their goods on the e-commerce platform, assuming three different kinds of malicious merchants based on their knowledge levels. By gradually decreasing merchants' knowledge levels, we investigate insider, expert, and black-box semantic attacks. All attacks are evaluated on recommender systems with three representative re-rankers: a neural feature-based similarity model pre-trained on ImageNet, a Collaborative-Filtering model leveraging visual features, and an end-to-end learning-based neural model. We demonstrate that using images to address cold start opens recommender systems to potential threats with clear, practical implications.

1.5.2 On Training Data

A huge amount of externally sourced data is publicly available on the web. Users can adversarially modify their data to decrease the data utility for unauthorized model training.

In **Chapter 3**, from the user's perspective, we explore the influence of the training data modification on the image annotation model. We first revisit availability poisoning (i.e., unlearnable examples) techniques, where labeled training data can be made unexploitable in supervised learning. By introducing compression-based pre-processing techniques, we show the possibilities of recovering the effectiveness of unlearnable examples in model training. We also provide an adaptive study on our compression-based methods and point out that availability poisoning is a promising method to mitigate misuse of externally sourced data for training purposes.

1.5.3 On Interaction Data

Externally sourced media content is archived and analyzed to provide and improve information services. In information systems, user profile information can be used for recognition purposes, where the predicted profile semantics can be further used as auxiliary information to increase the system performance. Users can sanitize their data before interaction to decrease the chance of being profiled. Media content can be modified according to user preferences without influencing the data's original social and sharing utility.

In **Chapter 4**, from the user's perspective, we explore how to pivot existing profiles against bag-based attribute profiling models. We point out the privacy threat posed by deep bag-based multiple-instance learning classifiers. Such classifiers can be used to infer privacy-sensitive attributes from sets of images posted online. We propose an adversarial machine learning approach to produce "pivoting additions" that can help users resist these classifiers by posting additional images, without requiring them to delete images. In particular, untouched adversarially selected natural images can resist profiling even in cases where the user has nearly no information about the privacy attack.

In **Chapter 5**, from the user's perspective, we further explore the influence of adversarial image query on different types of content-based image retrieval systems. As shown in Figure 2.1, search queries represent part of the information need from users. Our work on adversarial queries makes contributions to both understanding the working mechanism of modifications against content-based image retrieval systems and hiding user intent against malicious retrieval systems. Specifically, we propose a new unsupervised feature space retrieval attack, perturbations for image retrieval error (PIRE), for neural feature-based image retrieval model and evaluate it against three representative content-based image retrieval systems. Given an image query, PIRE generates adversarial perturbations by maximizing the Euclidean distance between the perturbed query image and original query image in feature space. We demonstrate that adversarial queries generated by PIRE are effective against neural, local, and global feature-based image retrieval systems.

1.6 List of Publications and Contributions

The author has published the following work during the Ph.D study. The remaining chapters in this thesis are based on the publications, as indicated.

This thesis consists of four chapters based on collaborative works. In each chapter, I have contributed to the formulation and conception of the work, the implementation of software, carrying out experiments, the interpretation of research data, and the writing.

Publications in This Thesis

Zhuoran Liu and Martha Larson. Adversarial Item Promotion: Vulnerabilities at the Core of Top-N Recommenders that Use Images to Address Cold Start. The Web Conference (WWW), 2021. –[**Chapter 2**]

Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Image Shortcut Squeezing: Countering Perturbative Availability Poisons with Compression. International Conference on Machine Learning (ICML), 2023. –[**Chapter 3**]

Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Resisting Bag-based Attribute Profiling by Adding Items to Existing Media Profiles. Under review. Preliminary version was published at Conference on User Modeling, Adaptation and Personalization (UMAP), 2021. –[**Chapter 4**]

Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Who's Afraid of Adversarial Queries?: The Impact of Image Modifications on Content-based Image Retrieval. International Conference on Multimedia Retrieval (ICMR). 2019. –[**Chapter 5**]

Other Publications during the Ph.D Study

Mingliang Liang, **Zhouran Liu**, and Martha Larson. Mutant Texts: A Technique for Uncovering Unexpected Inconsistencies in Large-Scale Vision-Language Models. International Conference on Multimedia Modeling (MMM), 2024.

Loes van Bemmel, **Zhuoran Liu**, Nik Vaessen, and Martha Larson. Beyond Neural-on-Neural Approaches to Speaker Gender Protection. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

Mingliang Liang, **Zhuoran Liu**, and Martha Larson. Textual Concept Expansion with Commonsense Knowledge to Improve Dual-Stream Image-Text Matching. International Conference on Multimedia Modeling (MMM), 2023.

Rui Wen, Zhengyu Zhao, **Zhuoran Liu**, Michael Backes, Tianhao Wang, and Yang Zhang. Is Adversarial Training Really a Silver Bullet for Mitigating Data Poisoning? International Conference on Learning Representations (ICLR), 2023.

Zhengyu Zhao, **Zhuoran Liu**, and Martha Larson. Adversarial Image Color Transformations in Explicit Color Filter Space. IEEE Transactions on Information Forensics and Security, 2023.

Dirren van Vlijmen, Alex Kolmus, **Zhuoran Liu**, Zhengyu Zhao, and Martha Larson. Generative Poisoning Using Random Discriminators. ECCV Workshop on Responsible Computer Vision, 2022.

Carlos Javier Hernández-Castro, **Zhuoran Liu**, Alex Serban, Ilias Tsingenopoulos, and Wouter Joosen. Adversarial machine learning. Book chapter of Security and Artificial Intelligence. Springer, 2022.

Zhengyu Zhao, **Zhuoran Liu**, and Martha Larson. On Success and Simplicity: A Second Look at Transferable Targeted Attacks. Annual Conference on Neural Information Processing Systems (NeurIPS), 2021

Zhuoran Liu, Niels Samwel, Léo Weissbart, Zhengyu Zhao, Dirk Lauret, Lejla Batina, and Martha Larson. Screen Gleaning: A Screen Reading TEMPEST Attack on Mobile Devices Exploiting an Electromagnetic Side Channel. Network and Distributed System Security Symposium (NDSS), 2021.

Zhuoran Liu, Zhengyu Zhao, Martha Larson, and Laurent Amsaleg. Exploring Quality Camouflage for Social Images. Working Notes Proceedings of the MediaEval Workshop, 2020.

Zhengyu Zhao, **Zhuoran Liu**, and Martha Larson. Adversarial color enhancement: Generating unrestricted adversarial images by optimizing a color filter. British Machine Vision Conference (BMVC), 2020.

Zhengyu Zhao, **Zhuoran Liu**, and Martha Larson. Towards Large yet Imperceptible Adversarial Image Perturbations with Perceptual Color Distance. Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

Zhengyu Zhao, **Zhuoran Liu**, and Martha Larson, Ahmet Iscen, Naoko Nitta. Reproducible Experiments on Adaptive Discriminative Region Discovery for Scene Recognition. ACM International Conference on Multimedia (MM), 2019.

Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Pixel Privacy 2019: Protecting Sensitive Scene Information in Images. Working Notes Proceedings of the MediaEval Workshop, 2019.

Sam Sweere, **Zhuoran Liu**, and Martha Larson. Maintaining Perceptual Faithfulness of Adversarial Image Examples by Leveraging Color Variance. Working Notes Proceedings of the MediaEval Workshop, 2019.

Zhuoran Liu, and Zhengyu Zhao. Adversarial Photo Frame: Concealing Sensitive Scene Information of Social Images in a User-Acceptable Manner. Working Notes Proceedings of the MediaEval Workshop, 2019.

Zhuoran Liu and Zhengyu Zhao. First Steps in Pixel Privacy: Exploring Deep Learning-based Image Enhancement against Large-Scale Image Inference. Working Notes Proceedings of the MediaEval Workshop, 2018.

Martha Larson, **Zhuoran Liu**, Simon Brugman, and Zhengyu Zhao. Pixel Privacy. Increasing Image Appeal while Blocking Automatic Inference of Sensitive Scene Information. Working Notes Proceedings of the MediaEval Workshop, 2018.

Chapter 2

Security Threats by Adversarial Background Collection

E-commerce platforms provide their customers with ranked lists of recommended items matching the customers' preferences. Merchants on e-commerce platforms would like their items to appear as high as possible in the top-N of these ranked lists. In this work, we demonstrate how unscrupulous merchants can create item images that artificially promote their products, improving their rankings. Recommender systems that use images to address the cold start problem are vulnerable to this security risk. We describe a new type of attack, Adversarial Item Promotion *(AIP), that strikes directly at the core of Top-N recommenders: the ranking mechanism itself. Existing work on adversarial images in recommender systems investigates the implications of conventional attacks, which target deep learning classifiers. In contrast, our AIP attacks are embedding attacks that seek to push features representations in a way that fools the ranker (not a classifier) and directly leads to item promotion. We introduce three AIP attacks* insider attack*,* expert attack*, and* semantic attack*, which are defined with respect to three successively more realistic attack models. Our experiments evaluate the danger of these attacks when mounted against three representative visually-aware recommender algorithms in a framework that uses images to address cold start. We also evaluate potential defenses, including adversarial training and find that common, currently-existing, techniques do not eliminate the danger of AIP attacks. In sum, we show that using images to address cold start opens recommender systems to potential threats with clear practical implications.*

This Chapter is published as Zhuoran Liu and Martha Larson. Adversarial item promotion: Vulnerabilities at the core of top-n recommenders that use images to address cold start. The Web Conference (WWW), 2021

2.1 Introduction

Visually-aware recommender systems [119; 71; 91] incorporate image information into their ranking mechanism. This information helps to address the challenge of cold start since it compensates for insufficient interactions associated with new users or items. In this work, we show how the use of image content for cold start opens visually-aware recommenders to vulnerability.

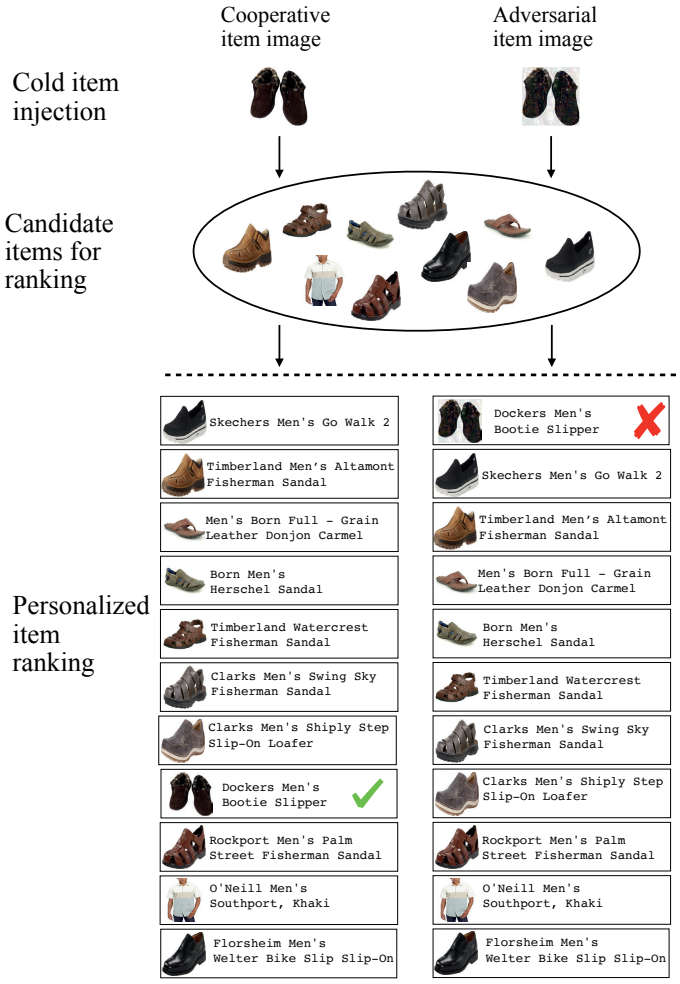


Figure 2.1: The cooperative cold item image and its corresponding adversarial cold item image (top) are each injected into the candidate set, which is generated by a first-stage ranker (here, BPR). In the personalized ranked list generated by the visually-aware second-stage ranker (here, DVBPR) the advesarial item (right) lands much higher than the cooperative item (left). (Diagram shows a real example. The adversarial item image is generated by our INSA attack with $\epsilon = 32$ and epoch $= 5$ explained in detail in Section 2.4.1).

The vulnerability is due to adversarial examples, which are samples deliberately designed to cause a machine learning system to make mistakes. The computer vision community has developed an in-depth understanding of how adversarial images can be used to attack classifiers, starting with [9; 184]. Classifier-targeted adversarial images can have an impact on recommender systems that leverage image content, as has been demonstrated by TAaMR [37]. However, until now, recommender system researchers have not considered how images can be modified to create adversarial items that attack visually-aware Top-N recommender systems by directly targeting the ranker, rather than a classifier.

We expose the vulnerability of visually-aware recommender systems to adversarial items by presenting a series of attacks and by experimentally assessing the threat that they pose. We also examine possible defenses. Adversarial training has been proposed in order to improve the general performance of multimedia recommender systems. The dominant approach is currently AMR [185]. Our experiments show that AMR is not sufficient to defend against our adversarial attacks. Further, other common defenses, such as image compression also fall short. In sum, the vulnerabilities of visually-aware recommender systems that we investigate here are serious and require further attention of the research community.

Our work is part of the long tradition of research devoted to the security and robustness of recommender system algorithms [132; 100; 121; 16; 105; 48; 30; 47; 186]. Most work, however, focuses on vulnerabilities related to user profiles. Early work looked at *shilling* [100], which uses fake users. Shilling was later generalized to *profile injection attacks* [121] or *poisoning attack* [105]. In our work, in contrast, we are looking at attackers who are able to manipulate items directly, and, specifically, to choose item images. In other words, instead of looking at profile-related attacks we are looking at an *item representation attacks*. Concretely, the risk of such attacks presents itself in the case of e-commerce platforms that sell the items of individual merchants, e.g., e-commerce or customer to customer (C2C) marketplaces. The merchants create their own item description, including images. We show that if such merchants act unscrupulously they can artificially promote their items and compromise the security of the recommender system.

Figure 2.1 illustrates the mechanics of the attack that we consider in this work, called an *Adversarial Item Promotion* (AIP) attack. On the left, we see personalized item ranking for a user when the recommender system is *not* under attack (i.e., the cold start item is a "cooperative item"). On the right, we see the ranking when the recommender system is under attack by an unscrupulous merchant, who has used a manipulated image in an item representation (i.e., the cold start item is an "adversarial item"). This setup reflects the way that recommender system platform would add a certain number of cold items to the personalized ranked lists of users in order to allow the items to start accumulating interactions. We choose a two-stage recommender, since they are used in industry [34; 195]. With the two stage recommender, we ensure that the adversarial cold item is competing against selected candidate items that are already very relevant to the user.

We provide a short walk-through of Figure 2.1, which illustrates the attack at the level of a single cold item and a single user. A set of "candidate items for ranking" for that user has been selected using a conventional personalized Top-N recommender. Then the cold item is injected into that set. Finally, a visually-aware personalized Top-N recommender is used to rank the candidate item set before it is presented to the user. We see that in the case of the cooperative image (left), the cold item lands somewhere in the ranked list, but probably not at the top. In contrast, in the case of the adversarial image (right), the cold item lands at the top of the personalized item ranking.

The overall impact of the attack depends on the accumulated effect of the attack over all users, and not just a single instance of the attack shown in Figure 2.1. It is important to understand that the final rank position of the cooperative vs. adversarial item will be different for each instance of the attack. However, in general, if the adversarial item of an unscrupulous merchant lands consistently farther towards the top of users' personalized recommendation lists than it deserves to, then, at large scale, the merchant will accrue considerable benefit.

We choose to focus on the cold-start problem because of its importance for recommender systems. However, there is also another reason. A straightforward, practical approach to blocking adversarial image promotion attacks on non-cold-start items is to prevent merchants from being able to change images once their items have started to accumulate interactions. Cold start is the most important moment of opportunity for a merchant to introduce an adversarial image into a representation. Every item starts in some way cold, and the issue particularly extreme in C2C marketplaces selling many unique items.

With this work, we make the following contributions:

- We propose three Adversarial Item Promotion (AIP) attacks on the ranking mechanism of visually-aware recommender systems in a cold item scenario and experimentally assess their impact. The attacks correspond to three different levels of knowledge and we test them using two real-world data sets.

- We show that there is no easy defense against AIP attacks. The currently dominant adversarial training, as well as conventional defenses such as compression, are not sufficient to eliminate the vulnerability.

- We release an implementation of our attacks and defenses that allows for testing and extension.

This work follows the standard procedure for security research. First, we specify a framework including the types of attacks expected (attack models), the systems to be attacked, and a means of measuring the impact of the attacks. Then, we propose attacks for each attack model and evaluate their success. The systems that we attack are representative of visually-aware recommender systems, i.e., a visual

feature-based similarity model (AlexRank) based on AlexNet [97], a Collaborative-Filtering (CF) model leveraging visual features (VBPR [71]), and state-of-the-art learning-based neural model (DVPBR [91]). Finally, we turn to the analysis of possible defenses and close with a conclusion and outlook.

2.2 Related Work

2.2.1 Robustness of Recommender System

In this section, we review previous work on recommender robustness. Note that this work focused on user profiles, not image content. O'Mahony *et al.* [132] introduce the definition of recommender system robustness and present several attacks to identify characteristics that influence robustness. Lam and Riedl [100] explore shilling attacks in recommender systems by evaluating recommendation performance under different scenarios. In particular, they find that new or obscure items are more especially susceptible to attack, and they suggest that obtaining ratings from a trusted source for these items could make them less vulnerable. Mobasher *et al.* [121] propose a formal framework to characterize different recommender system attacks, and they also propose an approach to detect attack profiles. In [121] and [16], evaluation metrics, e.g., hit rate for item and prediction shift, for the robustness of recommender systems are discussed. Recently, instead of model-agnostic profile injection attacks, poisoning attacks that leverage exact recommendation model information have been proposed. Li *et al.* [105] propose poisoning attacks on factorization-based CF algorithms that approximate the gradient based on first-order KKT conditions. Christakopoulou and Banerjee [30] propose a generative approach to generate fake user profiles to mount profile injection attacks. Fang *et al.* propose poisoning attacks to graph-based recommender systems [48]. They also propose to generate fake user-item interactions based on influence function [47]. Tang *et al.* [186] propose effective transfer-based poisoning attacks against recommender systems, but they mention that their approach is less effective on cold items. Our "item representation attack" is distinct from a "profile injection attack" or "poisoning attack", but both kinds of attacks have similar impacts, namely, pushing items that have been targeted for promotion.

2.2.2 Visually-aware Recommender System

Visually-aware recommender systems incorporate visual information into their recommendation ranking mechanism. Originally, visually-aware recommenders relied on image content retrieval to make preference predictions. Given a query image, Kalantidis *et al.* [90] gather segmentation parts and retrieve visually similar items within each of the predicted classes. Later, semantic information of images is also incorporated to improve retrieval performance. Jagadeesh *et al.* [82] collect a large-scale dataset, Fashion-136K, with detailed annotations and propose several retrieval-based approaches to recommend a matching item based on the query image.

Beyond image retrieval-based recommendation approaches, user-item interactions

are leveraged in visually-aware recommenders. IBR [119] models human notions of similarity by considering alternative or complementary items. Algorithms also incorporate the visual signal into CF models so as to exploit user feedback and visual features simultaneously, e.g., VBPR [71] and Fashion DNA [13]. Recently, with the advances in computational resources, learning-based neural frameworks have been proposed and achieve state of the art performance on fashion recommendation (DVBPR [91]) and reciprocal recommendations (ImRec [126]). In our work, to comprehensively evaluate AIP attacks in different recommenders, we select three representatives: an image-retrieval-based similarity model, a CF model leveraging visual features and a learning-based neural model.

2.2.3 Adversarial Machine Learning

Adversarial examples are data samples that are deliberately designed in order to mislead machine learning algorithms [9; 184]. A limited amount of work, as mentioned above, has addressed adversarial images for recommender systems. The work most closely related to our own [37] looks only at classification-based issues. Di Noia *et al.* [37] propose Targeted Adversarial Attack against Multimedia Recommender Systems (TAaMR), and they use two classification-based adversarial attacks, namely Fast Gradient Sign Method (FGSM) [184] and Projected Gradient Descent (PGD) [98], to evaluate two visually-aware recommender systems. In contrast to [37], we show that the problem of adversarial examples in recommender system goes beyond the problem of classifier-targeted adversarial examples.

Adversarial training is a promising techniques to tackle adversarial examples [118]. As stated in Section 2.1, research on adversarial training in visually-aware recommnder systems has, until this point, focused on improving general performance. Specifically, AMR [185] aims to improve recommendation with adversarial training (cf. Section 2.7.1 for details) and considers the robustness of recommender systems perturbations in system-internal representations. In contrast, our goal is to investigate security vulnerability originating from an external adversary who attacks item images. We show that simple adversarial training (i.e., AMR) is not a guarantee for robustness against AIP attacks (cf. Section 2.7.1).

2.3 Background and Framework

This section introduces the background and framework in which the attack models are developed and evaluated. Figure 2.2 gives the overview of the setup. As introduced in Section 2.1, we use a two-stage approach. The first-stage recommender generates a personalized set of candidate items. For this purpose, we choose Bayesian Personalized Ranking (BPR) [158], a representative CF model that is trained on the user-item interaction data. We use the visually-aware second-stage recommender to make a comparison between the cold start of a cooperative item and an adversarial item.

In this section, we first present our three attack models (Section 2.3.1), which are the basis for three specific AIP attacks, INSA, EXPA, and SEMA, explained in

Section 2.4. Then, we present the three representative visually-aware recommenders that we attack (Section 2.3.2). Finally, we explain the dimensions along which we evaluate the impact of the attacks (Section 2.3.3).

Table 2.1: The three AIP attack models characterized by the knowledge to which the attacker has access for each.

Attack model	General Knowledge	Visual feature extraction model	Embeddings
High knowledge cf. INSA (Section 2.4.1)			×
Medium knowledge cf. EXPA (Section 2.4.2)	×	×	
Low knowledge cf. SEMA (Section 2.4.3)	×		

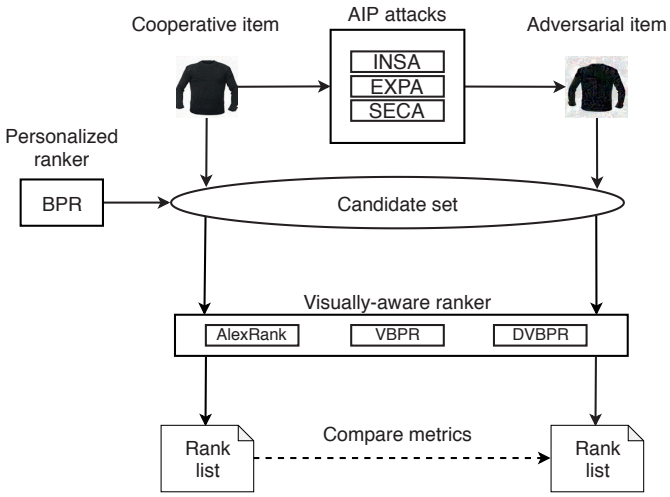


Figure 2.2: Setup for attack and attack evaluation.

2.3.1 Attack Models

We define three attack model following the three dimensions relevant for trustworthy recommender systems [121]. The *Intent* dimension captures the objective of the attacker. All our models use 'push' intent, i.e., attackers are merchants who want their items promoted to higher ranks in users' personalized recommendation lists. The *Knowledge* dimension captures how much information the attacker has about the system being attacked. Defining knowledge levels is common in adversarial machine learning research [9; 20]. Our three attack models correspond to three different levels of knowledge: high, medium, and low. The *Scale* dimension captures the scope of the interference. In all our attack models, we assume an attack with minimum scale, e.g., only a single item is attacked at any given moment. Such a small-scale attack is least likely to be noticed. It is clear that the harm caused by the attack will increase with scale.

Table 2.1 summarizes our attack models in terms of the level of knowledge involved. For our high-knowledge attack model we assume that the attacker is an insider at the recommender system platform and has access to the user embeddings of the trained recommender model. This scenario is not particularly realistic, but it is important because it demonstrates an upper bound for the potential damage that can be inflicted by an AIP attack.

The medium- and low-knowledge attack models are more realistic and assume that the attacker has general knowledge of the market in which the recommender system operates. In particular, the attacker must be able to identify (by observing sales trends or advertising) at least one item that sells well on the platform. We refer to this item as a *hook* item. The attack is strongest when the hook item image is an image used by the recommender, but it could also be an image of the item acquired elsewhere. As will be explained in Section 2.4, the adversarial item will use the hook item to pull itself up in the ranked list. In the medium-knowledge attack model, in addition to general knowledge, the attacker must have access to the pre-trained visual feature extractor used by the visually-aware recommender systems. Recommender systems leveraging a pre-trained visual feature extractor are prevalent in both academic research and industry, e.g., [119; 71; 103; 59; 145; 161]. These models are often released as publicly available resources in transfer learning.

2.3.2 Visually-aware Recommender Systems

In this section, we introduce the three representative visually-aware recommender systems that we will attack in our experiments: the visual feature-based similarity model (AlexRank), the CF model leveraging visual features (VBPR) and the learning-based neural approach (DVBPR). We chose these recommenders because they represent the three types of commonly used visually-aware recommender systems.

2.3.2.1 AlexRank

Image-content-retrieval-based recommendation is a nearest neighbor approach that ranks items by visual feature similarity of product images. Such methods are commonly used as baseline approaches in visually-aware recommender system research [91]. Here we use the output of the second fully-connected layer of AlexNet [97] as the visual feature of item images. Given an image of item i, the average Euclidean distance between the visual feature of item i and all items that user u has interacted with is calculated, so smaller distance means higher preference prediction. Equation 2.1 show the calculation of similarity predictor:

$$p_{u,i} = \sum_{j \in I_u^+} \frac{-\|\Phi_f(\mathbf{X}_i) - \Phi_f(\mathbf{X}_j)\|^2}{|\mathcal{I}_u^+|}, \tag{2.1}$$

where $\mathcal{I}_u^+$ is the set of items that user u has interacted with, and $\mathbf{X}_i, \mathbf{X}_j$ represent images of item i and j. Φ_f is the pre-trained model for image feature extraction. The final ranking of item i for user u is solely determined by preference score $p_{u,i}$.

2.3.2.2 VBPR

Extended from BPR [158], VBPR [71] incorporates visual features into the CF model. By leveraging image features of pre-trained CNN models, VBPR improves the recommendation performance of BPR. The preference prediction of VBPR is described in Equation 2.2:

$$p_{u,i} = \alpha + \beta_u + \beta_i + \gamma_u^T \gamma_i + \theta_u^T (\boldsymbol{E} \Phi_f(\mathbf{X}_i)), \tag{2.2}$$

where θ_u is the user content embedding, and $\boldsymbol{E}$ is the parameter for the visual feature. $\Phi_f(\mathbf{X}_i)$ represents the visual feature of item $\mathbf{X}_i$ from pre-trained model Φ_f, and α, β_u and β_i are user, item biases and global offset term. γ_u and γ_i represent the latent interaction-based embeddings for user and item. For model learning, VBPR adopts the pairwise ranking optimization framework from BPR. The training triples set $\mathcal{D}_s$ is described in Equation 2.3:

$$\mathcal{D}_s = \{(u, i, j) | u \in \mathcal{U} \wedge i \in \mathcal{I}_u^+ \wedge j \in \mathcal{I} / \mathcal{I}_u^+\} \tag{2.3}$$

where $\mathcal{U}$ and $\mathcal{I}$ represent the user and item sets, i represents the interacted item, and j represents the non-interacted item. A bootstrap sampling of training triples is used for model training. The optimization objective of VBPR is described in Equation 2.4:

$$\underset{\Theta}{\operatorname{argmin}} \sum_{(u,i,j) \in \mathcal{D}_s} - \ln \sigma(p_{u,i} - p_{u,j}) + \lambda_\Theta ||\Theta||^2 \tag{2.4}$$

where Θ represents all model parameters and λ_Θ is the weight of regularization term.

2.3.2.3 DVBPR

DVBPR [91] is a concise end-to-end model whose visual feature extractor is trained directly in a pair-wise manner. DVBPR achieves the state-of-the-art performance on several data sets for visually-aware recommendations [91]. The preference prediction of DVBPR is described in Equation 2.5:

$$p_{u,i} = \theta_u^T \Phi_e(\mathbf{X}_i), \tag{2.5}$$

where θ_u is the user content embedding, and $\Phi_e(\mathbf{X}_i)$ is the item content embedding where the CNN model Φ_e is trained directly in a pair-wise manner.

2.3.3 Attack Evaluation Dimensions

We evaluate attacks according to the aspects of *integrity* and *availability* distinguished in machine learning security [6]. Our main concern is the ability of the attack to compromise the integrity of the recommender system, which is related to the success of the 'push' intent of our attack models (cf. Section 2.3.1). We measure the ability of our attacks to raise the rank of cold-start items such that adversarial cold start items land higher than the corresponding cooperative cold start items in users' personalized recommendation lists. Specifically, we report the change in rank of cold-start items with prediction shift and change in hit rate (cf. Section 2.5.2).

In addition, we measure availability, which is related to whether the recommender system remains useful to other merchants and to customers while under attack. For this purpose, we use 'ordinary' test items. (We adopt the test items defined by the data sets.) We calculate change in hit rate when a ordinary test item is in a candidate set with a cooperative cold item and when the same ordinary test item is in a candidate set with an adversarial cold item.

We also consider the extent to which the attack is noticeable to the human eye. As attacks increase in strength, perturbations become visible in images and adversarial images can be identified by experts who know what they were looking for. However, e-commerce platforms are so large and the turnover of items so fast that it is impossible to manually vet all of the images representing items, cf. the known difficulty of filtering item collections for banned or unsafe products [8]. For these reason, strong attacks (i.e., larger image perturbations) are quite realistic. By focusing our experiments on strong attacks, we can evaluate the extent of the vulnerability of the system. We also carry out additional experiments that demonstrate the effect of an attack increasing in strength from weak to strong. In this way, we shed light on what might happen if user clicks are affected by an impression of low quality due to the presence of perturbations in images. Our additional experiments show that the success of the attack is not dependent on a highly noticeable change to the image appearance.

2.4 Adversarial Item Promotion Attacks

In this section, we introduce three adversarial item promotion (AIP) attacks corresponding to the three attack models previously introduced (cf. Section 2.3.1).

2.4.1 Insider Attack (INSA)

The high-knowledge attack model assumes the attacker has insider access to the user embeddings of the trained model (see Section 2.3.1). Some visually-aware recommenders (e.g., AlexRank) only use visual embedding (feature) to build nearest neighbor-based recommender, and other recommenders (e.g., DVBPR) model the visual content embedding together with user content embedding using dot product in a CF manner (cf. Section 2.3.2). For instance, in DVBPR, the inner product of user embedding θ_u and item embedding $\Phi_e(\mathbf{X}_i)$ represents the preference of user u on item i.

We propose an insider attack (INSA) in which the attacker can modify the embeddings of item images in order to increase the predicted preference score that solely determines the recommendation ranking. Specifically, INSA changes item embeddings by adding perturbations $\boldsymbol{\delta}$ on the item images. The perturbations are optimized iteratively such that the strength of preference for the item is maximized over all user profiles. In this work, the magnitude of $\boldsymbol{\delta}$ is restricted by L_∞ norm, which represents the maximum value of $\boldsymbol{\delta}$ and is commonly used in computer vision research to measure imperceptibility [184; 20]. Formally, given a product image $\mathbf{X}_i$ of item i, we optimize perturbations $\boldsymbol{\delta}$ to increase the preference $p_{u,i}$ of all users

on item i. The optimization objectives for different recommenders are specified in Equation 2.6:

$$\begin{aligned}
\text{AlexRank: } & \underset{\boldsymbol{\delta}}{\operatorname{argmax}} \sum_{u \in \mathcal{U}} \sum_{j \in I_u^+} \frac{-\|\Phi_f(\mathbf{X}_i + \boldsymbol{\delta}) - \Phi_f(\mathbf{X}_j)\|^2}{|\mathcal{I}_u^+|} \\
\text{VBPR: } & \underset{\boldsymbol{\delta}}{\operatorname{argmax}} \sum_{u \in \mathcal{U}} \theta_u^T (\boldsymbol{E}\Phi_f(\mathbf{X}_i + \boldsymbol{\delta})) \\
\text{DVBPR: } & \underset{\boldsymbol{\delta}}{\operatorname{argmax}} \sum_{u \in \mathcal{U}} \exp\left(\theta_u^T \Phi_e(\mathbf{X}_i + \boldsymbol{\delta})\right)
\end{aligned} \tag{2.6}$$

$\Phi : \mathbf{X}_i \rightarrow \theta_i$ is the feature extraction or embedding model where θ_i represents the content embedding for item i. θ_u represents the user content embedding. The optimization can be implemented by mini-batch gradient descent, and it stops when certain conditions are met, e.g., it reaches certain number of iterations.

2.4.2 Expert Attack (EXPA)

The medium-knowledge attack model assumes that the attacker can select a hook (i.e., popular) item. It also assumes that the attacker has access to the visual feature extraction model (see Section 2.3.1) and has the expertise needed to use it in a transfer learning pipeline. We propose an expert attack (EXPA) in which the attacker uses the hook item to mark the region of item space to which the adversarial item should be moved. Specifically, the EXPA attack generates perturbations added to the cooperative item in order to create the adversarial item by decreasing the representation distance to the hook item.

Formally, generating an adversarial item image by EXPA is described in Equation 2.7:

$$\underset{\boldsymbol{\delta}}{\operatorname{argmin}} \quad \|\Phi(\mathbf{X}_i + \boldsymbol{\delta}) - \Phi(\mathbf{X}_{\mathbf{hook}})\|_2, \tag{2.7}$$

where Φ is the feature extraction or embedding model. The EXPA attack leverages the same mechanism as the targeted visual feature attack proposed by [160]. The novelty of EXPA is its use of a hook image that moves the adversarial image in image space in a way that makes it rise in personalized recommendation lists. Note that the hook image itself is not necessarily present in candidate set, which is selected by BPR, and thereby also not necessarily in the recommendation lists.

Algorithm 2.1 describes the process to generate adversarial product images with INSA and EXPA. $\mathbf{X}_i$ is the original image of cold item, and $\mathbf{X}_{\mathbf{hook}}$ is the hook item. Φ is the neural network that extracts embeddings or features from the image content. Our aim is to find perturbations $\boldsymbol{\delta}$ that could increase the personalized preference predictions by optimization through all user content embeddings (INSA) or targeting a hook item (EXPA). The magnitude of perturbations can be adjusted by $\boldsymbol{\delta}$. To make sure that the output images are valid with respect to standard image

Algorithm 2.1 Adversarial Item Promotion Attack

Input:
$\mathbf{X}$: cold item image, $\mathbf{X_{hook}}$: hook item image
$\boldsymbol{\delta}$: adversarial perturbations, ϵ: L_∞ norm bound
Φ: neural network, θ: user content embedding
K: number of iterations, $\mathcal{A}$: attack to mount (INSA or EXPA)
Output:
$\mathbf{X}^{'}$: adversarial product image

Initialize $\boldsymbol{x}'_0 \leftarrow \mathbf{X}$, ▷ $\boldsymbol{x}'_k$ represents adversarial image in iteration k
$\boldsymbol{\delta} \leftarrow \mathbf{0}$
for $k \leftarrow 1$ to K **do**
  if $\mathcal{A}$ is INSA **then**
    AlexRank: $\boldsymbol{\delta} \leftarrow \underset{\boldsymbol{\delta}}{\operatorname{argmax}} \sum_{u \in \mathcal{U}} \sum_{j \in I_u^+} \frac{-\|\Phi_f(\boldsymbol{x}'_{k-1}+\boldsymbol{\delta}) - \Phi_f(\mathbf{X}_j)\|^2}{|\mathcal{I}_u^+|}$

    VBPR: $\boldsymbol{\delta} \leftarrow \underset{\boldsymbol{\delta}}{\operatorname{argmax}} \sum_{u \in \mathcal{U}} \theta_u^T (\boldsymbol{E}\Phi_f(\boldsymbol{x}'_{k-1} + \boldsymbol{\delta}))$

    DVBPR: $\boldsymbol{\delta} \leftarrow \underset{\boldsymbol{\delta}}{\operatorname{argmax}} \sum_{u \in \mathcal{U}} \exp\left(\theta_u^T \Phi_e(\boldsymbol{x}'_{k-1} + \boldsymbol{\delta})\right)$ ▷ Eq.(2.6)
  else if $\mathcal{A}$ is EXPA **then**
    $\boldsymbol{\delta} \leftarrow \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \quad \|\Phi(\boldsymbol{x}'_{k-1} + \boldsymbol{\delta}) - \Phi(\mathbf{X_{hook}})\|_2$ ▷ Eq.(2.7)
  else
    break
  end if
  $\boldsymbol{\delta} \leftarrow \text{clip}(\boldsymbol{\delta}, -\epsilon, \epsilon)$ ▷ Make sure that the magnitude of perturbations are in pre-defined L_∞ norm range
  $\boldsymbol{x}'_k \leftarrow \boldsymbol{x}'_{k-1} + \boldsymbol{\delta}$
  $\boldsymbol{x}'_k \leftarrow \text{clip}(\boldsymbol{x}'_k + \boldsymbol{\delta}, 0, 1)$ ▷ Ensure perturbed image stays in valid image range
end for
$\boldsymbol{x}'_k \leftarrow \text{quantize}(\boldsymbol{x}'_k)$ ▷ Ensure $\boldsymbol{x}'_k$ is valid in the 8-bit image format
return $\mathbf{X}^{'} \leftarrow \boldsymbol{x}'_k$ is the adversarial item image

encoding format, a clip function restricts adversarial item image in range [0, 1], and a quantization function ensures that the output image can be saved in the 8-bit format. The resulting adversarial image $\mathbf{X}^{'}$ is the summation of the original image and the clipped perturbations.

2.4.3 Semantic Attack (SEMA)

The low-knowledge attack model assumes nothing beyond general knowledge needed to choose hook items (see Section 2.3.1). We propose a semantic attack (SEMA) uses the semantic content of the image, i.e., what is shown in the image, in order to achieve the promotion of items. The attack differs considerably from INSA and EXPA, which add perturbations to existing images without changing what the images depict.

Figure 2.3 c-SEMA illustrates the semantic attack that we will test here, which we

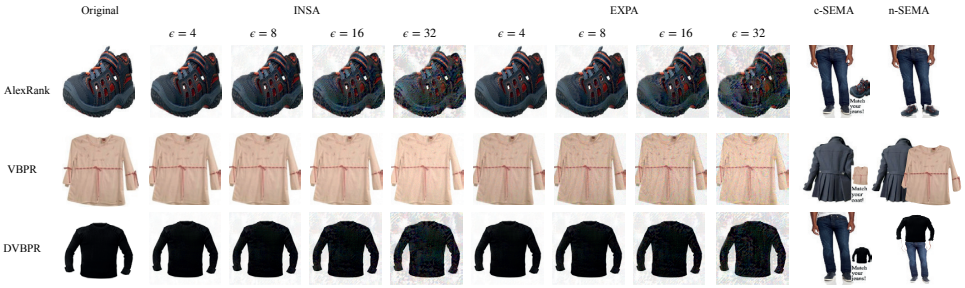


Figure 2.3: Examples of adversarial item images by different approaches. INSA attack is based on Equation 2.6. EXPA, c-SEMA and n-SEMA select a popular item, e.g., Levi's 501 jeans, as the hook item. c-SEMA applies simple co-depiction approach, and n-SEMA incorporates the target item in a more natural way. More details about the influence of ϵ on recommendation performance can be found in Section 2.6.2

call compositing semantic attack (c-SEMA). With c-SEMA, the attacker creates an adversarial image by editing the original image into the hook item image as an inset. Here, the c-SEMA attack is promoting a pair of shoes and the hook item is the jeans. A text (here, "match your jeans/coat") can be included to contribute to the impression that the composite image is a fair-play attempt to raise the interest of potential customers.

Figure 2.3 n-SEMA shows another type of semantic attack, which we call natural semantic attack (n-SEMA). Here, the integration of the hook item is natural. n-SEMA images can be created in a photo studio or a professional photo editor. We do not test them here, since creation is time consuming and we are using a cold test set of 1000 item images. However, an unscrupulous merchant would have the incentive to invest the time to create n-SEMA images. One highly successful adversarial item image could already lead to increased buyers and increased profit.

The semantic attack is particularly interesting for two reasons. SEMA achieves the change in image embeddings needed to push an adversarial image close to a hook image in image space by manipulating the depicted content of the image. First, this means that there are no limits on the quality of a SEMA adversarial image. Contrast c-SEMA and n-SEMA with the INSA and EXPA photos in Figure 2.3. The item is visible in the image, and consumers who decide to purchase the product will not find that they have been misled. However, not all of the images are crisp, and stronger attacks introduce artifacts affecting the perceived image quality. Second, the impact of a SEMA image attack is not dependent on the algorithm used by the recommender systems. In fact, SEMA images can effectively attack any recommender system using visual features, and not just systems using neural embedding as studied here.

Table 2.2: Statistics of the data sets

	#Users	#Items	#Interactions
Amazon Men	34244	110636	254870
Tradesy.com	33864	326393	655409

2.5 Experimental Setup

In this section, we first introduce data sets used for our experiments (Section 4.4.1) and introduce the details of evaluation setup including metrics (Section 2.5.2). Then, we describe implementation of experiments (Section 2.5.3).

2.5.1 Data

2.5.1.1 Data statistics

We perform our experiments on two data sets: Men's Clothing in Amazon.com and Tradesy.com, which are publicly available and widely used in visually-aware recommender system research. The statistics of the two data sets are described in Table 2.2. The Men's Clothing category is an important subset of the Amazon.com data set, where the effectiveness of visual features has been validated in previous work [119; 71; 91]. Tradesy.com is a C2C second-hand clothing trading website where users can buy and sell fashion items. The nature of the Tradesy.com inventory makes visually-aware cold item recommendation crucial, because of its "one-off" characteristics. For both datasets, one descriptive image is available for each item, and we follow the protocol of [91] and treat users' review histories as implicit feedback. For each user, one item is selected among all interacted items as the test item, so we have the same number of test items as the number of users.

2.5.1.2 Cold test item election

To validate the effectiveness of the attacks in cold start scenario, in each of Amazon Men and Tradesy.com data sets, we randomly select 1000 cold test items that no user has interacted with and leave them out as the cold test set. These cold items are excluded from the training process. Later, they are injected as cold-start items into the candidate item set before feeding the set into the visually-aware ranker.

2.5.2 Evaluation Metrics

The change in rank of cold-start items is measured by the prediction shift and the change in hit rate ($HR@N$), following the evaluation metrics for top-N recommender system robustness [121] and [16]. Equation 2.8 defines the average prediction shift Δ_{p_i} for item i and also the mean average prediction shift for a set of test items $\Delta_{\mathbf{set}}$. Our results will report $|\Delta_{\mathbf{set}}|$ and the direction separately for clarity. $p'_{u,i}$ is the post-attack predictor score and $p_{u,i}$ is the original predictor score for item i.

$\mathcal{I}_{\mathbf{test}}$ represents the set of test items.

$$\Delta_{p_i} = \sum_{u \in \mathcal{U}} \frac{(p'_{u,i} - p_{u,i})}{|\mathcal{U}|} \quad \Delta_{\mathbf{set}} = \sum_{i \in \mathcal{I}_{\mathbf{test}}} \frac{\Delta_{p_i}}{|\mathcal{I}_{\mathbf{test}}|}, \tag{2.8}$$

Equation 2.9 defines the average hit rate $HR_i@N$ for item i in terms of $H_{u,i}@N$ for item i for user u. The mean average hit rate $HR@N$ for test items averages $HR_i@N$ over the test set. $\Delta_{HR@N}$ is the change in mean average hit rate, where $HR_i^{'}@N$ is the post-attack hit rate for item i. Our results will report $|\Delta_{HR@N}|$ and the direction separately for clarity.

$$\begin{aligned} HR_i@N &= \sum_{u \in \mathcal{U}} \frac{H_{u,i}@N}{|\mathcal{U}|} & HR@N &= \sum_{i \in \mathcal{I}_{\mathbf{test}}} \frac{HR_i@N}{|\mathcal{I}_{\mathbf{test}}|} \\ \Delta_{HR@N} &= \sum_{i \in \mathcal{I}_{\mathbf{test}}} \frac{HR_i^{'}@N - HR_i@N}{|\mathcal{I}_{\mathbf{test}}|} \end{aligned} \tag{2.9}$$

It is important to note that low metric values can still result in large impact due to the large number of users involved. For example, in Amazon Men, an increase of 0.01 on $HR@5$ means that adversarial cold items are pushed into the top-5 list of about 340 users.

2.5.3 Implementation

In the first stage, we use BPR to generate a candidate set of top 1000 items that are selected by personalized preference ranking. Note that we use a set, which means that the original rank order is not taken into account by the visually-aware ranker. Then we inject the ordinary test item in the top 1000 candidate set and get a set of 1001 items for each user. To compare before and after the attack, we inject one cooperative cold item or its corresponding adversarial cold item in the candidate set. So, for each cold-start item, we have two sets of 1002 items, which each include one test item and also include either one cooperative or one adversarial cold item. We use our three visual ranking models, AlexRank, VBPR, and DVBPR (see Section 2.3.2), to rank the 1002 items and evaluate with respect to both integrity and availability (see Section 2.3.3).

2.5.3.1 Model training

We implement BPR, AlexRank, VBPR, and DVBPR in PyTorch [137]. For the first stage model BPR, we set the number of factors to 64. Stochastic Gradient Desecent (SGD) is used to optimize BPR with learning rate 0.01 for Amazon Men and 0.5 for Tradesy.com, where the weight decay for L_2 penalty is set to 0.001 on both data sets. The feature dimension of AlexRank is 2048, and the embedding length of both VBPR and DVBPR is 100. A grid search of learning rate in $\{0.1, 0.01, 0.001, 0.005\}$ and weight decay in $\{0.001, 0.0001, 0.0001\}$ is conducted for both VBPR and DVBPR to select hyperparameters, and we select the model with best validation performance.

2.5.3.2 AIP attacks

If not specifically mentioned, the maximal size of perturbations ϵ for AIP attacks is set to 32. In INSA, the number of epochs is set to 10 to control the attacking time, and our implementation takes about 2 hours to generate 1000 adversarial item images on a single NVIDIA RTX 2080Ti GPU. We use the Adam optimizer with the learning rate of 0.001 for DVBPR and 0.0001 for both VBPR and AlexRank. In EXPA, the hook items are "Levi's Men's 501 Original Fit Jean" in Amazon Men and a gray coat in Tradesy.com. These two products are most commonly interacted items in training data of these two data sets. Recall, however, that hooks can be chosen without direct access to interaction statistics. We use a Adam optimizer with a learning rate of 0.01 in EXPA, and the number of iterations is set as 5000. In c-SEMA, we resize the hook item image and paste it on the right side of the cooperative item image as shown in Figure 2.3. To make the combination more natural, we also add a text description. More implementation details can be found in our released code.

2.6 Experimental Results

In this section, we carry out experimental analysis of our attacks using two real-world data sets (Section 2.6.1) and also investigate the influence of hyperparameter selections (Section 2.6.2). Finally, we analyze and discuss classification-based attack (Section 2.6.3).

2.6.1 Attack Evaluation

Table 2.3: Absolute mean average prediction shifts of adversarial cold items $|\Delta_{\text{set}}|$ on Amazon Men (AM) and Tradesy.com (TC), where ↑ represents positive prediction shift (score increased) and ↓ represents negative prediction shift (score decreased). Positive shift means a successful attack.

	AlexRank			VBPR			DVBPR		
	INSA	EXPA	c-SEMA	INSA	EXPA	c-SEMA	INSA	EXPA	c-SEMA
AM	↑16.13	↑15.94	↑11.1	↑3.27	↑3.16	↑0.88	↑13.54	↑4.80	↑4.82
TC	↑26.89	↓0.67	↓3.44	↓0.79	↑1.60	↑1.45	↑3.64	↑1.89	↑1.19

We mount AIP attacks and assess their effects with respect to our metrics (cf. Section 2.5.1.2). Table 2.3 shows the absolute mean average prediction shift $|\Delta_{\text{set}}|$ for adversarial cold items vs. cooperative items. The upwards arrow represents a positive prediction shift, meaning that the attack has successfully promoted the item. We see that for nearly all combinations of AIP attack and visually-aware recommender system the attack is successful. The high-knowledge attack, INSA, achieves a larger shift than EXPA and c-SEMA. c-SEMA is surprisingly successful, given the very minimal amount of knowledge that it requires. Note that it is only meaningful to compare the size of the prediction shift for the same recommender system and the same data set.

Table 2.4: Average *HR*@5 of cooperative (adversarial) cold item and ordinary test item in AlexRank, VBPR, DVBPR and adversarial training-based AMR. Three attacks, INSA, EXPA, and c-SEMA, are evaluated on Amazon Men and Tradesy.com data sets. ** indicates that the increase/decrease over ordinary test/cooperative cold test are statistically significant ($p < 0.01$).

(a) AlexRank, VBPR, and DVBPR.

			Amazon Men			Tradesy.com		
Attack method	**Dimension**	**Test set**	AlexRank	VBPR	DVBPR	AlexRank	VBPR	DVBPR
–	–	Cooperative cold test	0.0011	0.0017	0.0010	0.0003	0.0022	0.0011
	–	Ordinary test	0.0187	0.0191	0.0195	0.0338	0.0241	0.0234
INSA	Integrity	Adversarial cold test	0.8296**	0.2678**	0.7211**	0.9399**	0.0783**	0.2711**
	Availability	Ordinary test	0.0140**	0.0190**	0.0170**	0.0284**	0.0224**	0.0228**
EXPA	Integrity	Adversarial cold test	0.0036**	0.0150**	0.0646**	0.0002**	0.0048**	0.0384**
	Availability	Ordinary test	0.0188**	0.0198**	0.0192**	0.0338	0.0228**	0.0232**
c-SEMA	Integrity	Adversarial cold test	0.0017**	0.0027**	0.0181**	0.0000**	0.0048**	0.0005**
	Availability	Ordinary test	0.0187**	0.0198**	0.0194**	0.0338	0.0227**	0.0234

(b) AMR.

Amazon Men	Tradesy.com
AMR	
0.0013	0.0013
0.0289	0.0381
0.7476**	0.2858**
0.0238**	0.0362**
0.0185**	0.0039**
0.0289**	0.0381**
0.0007**	0.0047**
0.0289**	0.0381**

Table 2.5: $|\Delta_{HR@5}|$ before and after AIP attacks: Cold items (attack is successful if *HR*@5 rises).

(a) AlexRank, VBPR, and DVBPR.

	Amazon Men			Tradesy.com		
	AlexRank	VBPR	DVBPR	AlexRank	VBPR	DVBPR
INSA adversarial cold vs. cooperative cold	$\uparrow$ 0.8285	$\uparrow$ 0.2661	$\uparrow$ 0.7201	$\uparrow$ 0.9396	$\uparrow$ 0.0761	$\uparrow$ 0.2700
EXPA adversarial cold vs. cooperative cold	$\uparrow$ 0.0025	$\uparrow$ 0.0133	$\uparrow$ 0.0636	$\downarrow$ 0.0001	$\uparrow$ 0.0026	$\uparrow$ 0.0373
c-SEMA adversarial cold vs. cooperative cold	$\uparrow$ 0.0006	$\uparrow$ 0.0010	$\uparrow$ 0.0171	$\downarrow$ 0.0003	$\uparrow$ 0.0026	$\downarrow$ 0.0006

(b) AMR.

Amazon Men	Tradesy.com
AMR	AMR
$\uparrow$ 0.7463	$\uparrow$ 0.2845
$\uparrow$ 0.0172	$\uparrow$ 0.0026
$\downarrow$ 0.0006	$\uparrow$ 0.0034

A negative mean average prediction shift does not necessarily mean that an attack is unsuccessful since it is the rank position and not the preference prediction score that translates into benefit for the attacker. We go on to examine hit rate related metrics.

Table 2.4(a) presents results in terms of the mean average hit rate *HR*@5. The first two rows report the original situation: the hit rate for the cooperative (i.e., not adversarial) version of the cold items and the hit rate for the ordinary test items in the case that no adversarial items have been added to the candidate set. The rest of the table reports on attacks. Cases marked with ** indicate a statistically significant difference between the original situation and the case of the attack (item-level paired sample t-test $p < 0.01$).

First, we consider "Integrity", namely, the success of the attacks in pushing items. It can be observed in Table 2.4 that in nearly all cases the hit rate for the adversarial version of the cold items exceeds that of the cooperative version of the cold items, meaning that all AIP attacks are generally effective. For INSA, the impact of the attack is dramatic. The cooperative cold item makes it to one of the top-5 position in the lists of only 35 users (averaged over the three recommender system), but after the attack, the adversarial cold item makes it into the top-5 position of over 20,000 users. Since INSA uses the most knowledge, it is not surprising that it is the most effective attack. However, even with much less knowledge, both EXPA and c-SEMA pose serious threats. For example, for Amazon Men, c-SEMA pushes cold items into the top-5 list of 582 users in the case of DVBPR. We also calculated *HR*@10 and *HR*@20 for all conditions. These are not reported here since the trends were overwhelming the same as for *HR*@5.

Next, we turn to discuss "Availability", namely, the extent to which promotion occurs at the expense of other items, which at scale can impede the functioning of the entire recommender system. In Table 2.4, we see that INSA has a strong impact on availability than EXPA and c-SEMA.

Notice that the different performance of EXPA and c-SEMA on Amazon Men and Tradesy.com is not solely attributable to the adversarial attack itself, since the selection of the hook item also has an impact. For Tradesy.com, the *HR*@5 for the selected hook item (which is a gray coat) is originally rather low, so after EXPA or c-SEMA, the rank cannot increase dramatically. In general, we observe that AIP attacks are more damaging to integrity than to availability. However, in real-world situations it would be important to study cases involving the simultaneous presence of multiple adversarial items.

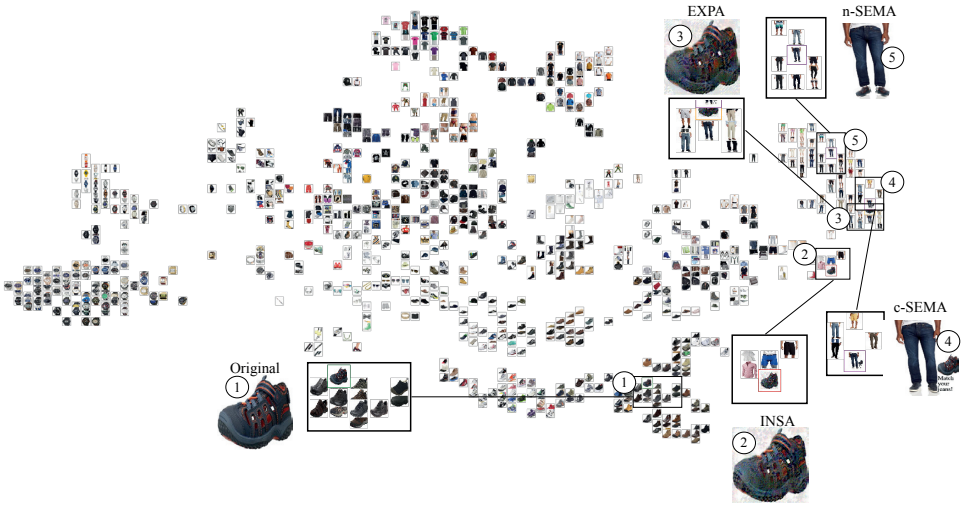


Figure 2.4: 2-D visualization of item embedding by pre-trained DVBPR model (#factors = 100) for randomly selected 1000 items in Amazon Men data set, cooperative item ① embedding and its corresponding adversarial item embedding. Attacks include generated images by INSA ②, EXPA ③, c- SEMA ④ and n- SEMA ⑤.

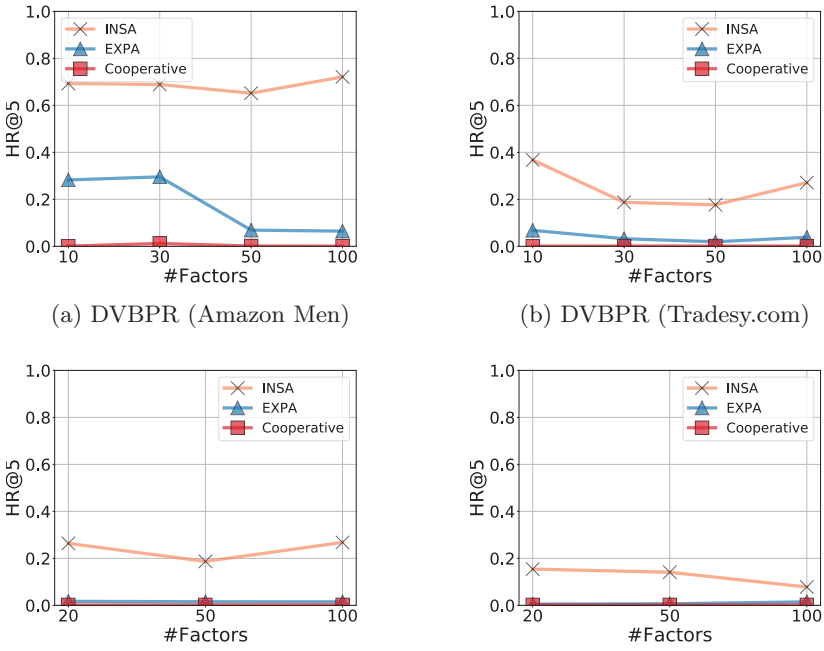
In order to directly illustrate the magnitude of the impact of AIP attacks on integrity, we report the change in mean average hit rate $|\Delta_{HR@5}|$ in Table 2.5(a). Again, the expected large effect of INSA can be observed. Also, again, the c-SEMA attack is surprisingly effective, given the minimal knowledge it involves. Remember that the recommender system platforms we are concerned about are enormous, and even a small boost in average rank of the magnitude of that afforded by c-SEMA could translate in to a substantial increase in interactions and profit.

Figure 2.4 shows a 2-D visualization (using t-SNE [192]) of the image space defined by the DVBPR item embeddings. It allows us to directly observe the influences of different AIP attacks. The position of the original cold image (i.e., a cooperative image) is shown by ①. We can see that it is positioned next to items with which it is visually similar. The attacks move this image to the other positions. The hook item is a pair of jeans. We can see that EXPA, c-SEMA, and n-SEMA push the cold item to a cluster related to the hook item. Note that it is difficult to reason about the position of INSA, since it is optimized with respect to all user embeddings. Here the n-SEMA image is manually generated using photo-editing software, in particular, after cropping and rotation, the shoes to be promoted are edited into the hook image. However, recall that in the real world an image could easily be taken of a model wearing both items.

2.6.2 Influence of Hyperparameters

Choice of the hyperparameters can influence the impact of the attack. Here, we take a look at the two most important hyperparameters, embedding length and attack strength (ϵ).

To study the impact of embedding length, we gradually reduce the embedding length and measure *HR*@5. Specifically, we conduct experiments for different numbers of factors (for VBPR in $\{20, 50, 100\}$ and for DVBPR in $\{10, 30, 50, 100\}$) with same adversarial budget (i.e., same iterations and learning rates). Results are presented in Figure 2.5 for VBPR and DVBPR. We discovered that the embedding length is quite important, with evidence pointing towards systems using shorter embedding length being more vulnerable to AIP attacks. This finding is valuable since without this knowledge a visually-aware recommender systems might use short embeddings to save storage.



(a) DVBPR (Amazon Men)

(b) DVBPR (Tradesy.com)

(c) VBPR (Amazon Men)

(d) VBPR (Tradesy.com)

Figure 2.5: *HR*@5 of cooperative cold items and adversarial cold items by INSA/EXPA with different number of factors in VBPR and DVBPR on Amazon Men's Clothing and Tradesy.com data set.

To study the impact of attack strength we vary the magnitude of ϵ and measure *HR*@5. We carry out experiments with ϵ in $\{4, 8, 26, 32\}$ for INSA and EXPA on subset of 100 cold item images. Results are presented in Table 2.6 for all three visually-aware recommender systems. We found that increasing ϵ from 4 to 32 leads to improved adversarial effects, but large perturbation size is not necessary

in most cases for a successful attack. Comparably, DVBPR is more sensitive to the magnitude of ϵ than the other two approaches. Recall that EXPA has sensitivities to the choice of the hook item.

Figure 2.3 provides examples that correspond to different levels of ϵ. This figure confirms that it is not necessary for the adversarial modifications to be highly noticeable in an image in order for an attack to be effective.

Table 2.6: Average *HR*@5 of a subset of 100 adversarial cold test items by INSA/EXPA with different magnitude of ϵ in DVBPR and VBPR for Amazon Men's Clothing and Tradesy.com data set. (cf. Figure 2.3 for adversarial item images with different ϵ)

		Amazon Men					Tradesy.com				
	Attack	Cooperative cold	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$	$\epsilon = 32$	Cooperative cold	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$	$\epsilon = 32$
AlexRank	INSA	0.0023	0.7856	0.8180	0.8258	0.8297	0.0054	0.9280	0.9351	0.9381	0.9400
	EXPA		0.0036	0.0035	0.0036	0.0036		0.0020	0.0000	0.0000	0.0000
VBPR	INSA	0.0018	0.2932	0.3159	0.3498	0.3599	0.0025	0.0734	0.0755	0.0815	0.0842
	EXPA		0.0008	0.0075	0.0135	0.0167		0.0005	0.0009	0.0009	0.0009
DVBPR	INSA	0.0007	0.0049	0.0391	0.3706	0.7039	0.0018	0.0015	0.0194	0.1116	0.2731
	EXPA		0.0021	0.0088	0.0313	0.0554		0.0025	0.0076	0.0291	0.0397

2.6.3 Classifier-targeted Adversarial Images

Table 2.7: $|\Delta_{HR@5}|$ before and after TAaMR attack: Cold items (attack is successful if *HR*@5 rises) and ordinary test items (successful attacks cause a drop in *HR*@5) .

		Amazon Men			Tradesy.com		
		AlexRank	VBPR	DVBPR	AlexRank	VBPR	DVBPR
Integrity dimension	FGSM adversarial cold vs. cooperative cold	$\downarrow 0.0010$	$\uparrow 0.0014$	$\uparrow 0.0070$	$\downarrow 0.0002$	$\uparrow 0.0183$	$\downarrow 0.0008$
	PGD20 adversarial cold vs. cooperative cold	$\uparrow 0.0100$	$\downarrow 0.0011$	$\downarrow 0.0005$	$\uparrow 0.0019$	$\downarrow 0.0014$	$\downarrow 0.0004$
Availability dimension	FGSM ordinary test vs. ordinary test	$\uparrow << 0.0001$	$\downarrow << 0.0001$	$\uparrow 0.0016$	$\downarrow << 0.0001$	$\downarrow << 0.0001$	$\uparrow << 0.0001$
	PGD20 ordinary test vs. ordinary test	$\downarrow << 0.0001$	$\uparrow << 0.0001$	$\uparrow 0.0016$	$\uparrow << 0.0001$	$\downarrow << 0.0001$	$\uparrow << 0.0001$

In Section 2.1, we pointed out that previous work on the vulnerability of recommender systems that use images has focused on classifier-targeted adversarial examples, which are already well studied in the computer vision literature. In contrast, our EXPA and INSA approaches target the ranking mechanism, and attack the user content embedding. In this section, we confirm that our attack poses a greater threat than the classifier-targeted attack TAaMR [37]. The TAaMR attack works by calculating the most popular class of items using the data set on which the recommender system was trained. As such, it can be considered an insider attack, which, like our INSA, requires knowledge that would only be available inside the company running the recommender system platform. TAaMR also requires access to a visual feature extraction model, like our EXPA. It uses this model as a source of class definitions and to create images that are adversarial with respect to the most popular class. For our experiments we create class-targeted adversarial examples using two approaches from the computer vision literature, FGSM [63] and PGD [98], on pre-trained AlexNet [97]. We use the training data to calculate the most popular item classes. There are 'jersey, T-shirt, tee shirt' (9468 interactions)

for Amazon Men and 'suit, suit of clothes' (19841 interactions) for Tradesy.com. The inventory of classes is taken from ImageNet [36], as in [37]. Table 2.7 shows the change in mean average *HR*@5 for FGSM and PGD20 attack. Generally, PGD20 has larger impacts on the AlexRank, and FGSM is more effective on VBPR. Both attacks have little impact on DVBPR, because the architecture of AlexNet is fairly distinct from the CNN-F architecture [23] in DVBPR. Although it uses information comparable to that used by INSA and EXPA, the adversarial impact is on par with the c-SEMA attack (cf. Table 2.5). This experiment shows the importance of AIP attacks, which directly attacks the user content embeddings and thereby the ranker.

2.7 Defense

2.7.1 Adversarial Training

AMR (Adversarial Multimedia Recommendation) [185] uses the preference prediction function from VBPR (cf. Equation 2.2) and adds on-the-fly adversarial information to the training process. Recall it was proposed to improve recommendation performance, but here we will study its potential for defending against AIP attacks. The optimization of AMR is implemented by mini-batch gradient descent. In each step, given a subset $\mathcal{D}_a$ of $\mathcal{D}_s$ (cf. Equation 2.3), AMR first perturbs the Θ to increase the loss:

$$\Theta' = \underset{\Theta}{\operatorname{argmax}} \sum_{(u,i,j)\in\mathcal{D}_a} -\ln \sigma(p_{u,i} - p_{u,j}) \tag{2.10}$$

In AMR, adversarial perturbations with respect to parameters are calculated by model gradients and added to current parameters in each step. Then it feeds forward the visual features, calculates combined loss, and back-propagates to update the parameters of the model. Specifically, $p'_{u,i}$ and $p'_{u,j}$ are calculated by the model with perturbed parameters Θ' (cf. Equation 2.2). Then, model parameters are updated by back-propagation based on the combined normal and adversarial loss:

$$\underset{\Theta}{\operatorname{argmin}} \sum_{(u,i,j)\in\mathcal{D}_a} -\ln \sigma(p_{u,i} - p_{u,j}) - \lambda_{\text{adv}} \ln \sigma(p'_{u,i} - p'_{u,j}) + \lambda_\Theta ||\Theta||^2 \tag{2.11}$$

where λ_{adv} is the weight hyperparameter for adversarial loss. To train AMR, we adopt the same hyperparameters from VBPR and set $\lambda_{\text{adv}} = 1$.

Table 2.4(b) presents the *HR*@5 of AMR under AIP attacks and Table 2.5(b) presents the mean average hit rate changes. Although AMR increases the general performance by including adversarial information into training process, the *HR*@5 jumps noticeably when AIP attacks are applied, which means AMR is vulnerable to AIP attacks. Our finding here is consistent with recent research in the machine learning community [18; 191], which shows that achieving adversarial robustness is non-trivial.

Table 2.8: Visualization of the level at which a defense is successful at lowering the *HR*@5 of an adversarial cold item equal or less the average *HR*@5 of a cooperative cold item. For JPEG compression, levels are specified as compression percents and for bit depth reduction levels are specified as number of bits with which the image is encoded. (▲: Amazon.com; ★: Tradesy.com)

		JPEG compression					Bit depth reduction					
		90	70	50	30	10	7	6	5	4	3	2
AlexRank	INSA											
	EXPA										▲	
VBPR	INSA											
	EXPA	▲★									▲★	
DVBPR	INSA						★					
	EXPA					▲						▲★

2.7.2 Defense by Image Compression

In the computer vision literature, simple defenses have been shown to be effective against adversarial images that cause neural classifiers to misclassify [66; 33; 44; 207]. Here, we evaluated two common defenses: JPEG compression and bit depth reduction in order to test whether they are effective against AIP attacks. These are known to be able to erase the effect of image perturbations. We carried out an evaluation by applying progressively stronger versions of the defense to the 100-item subset of our larger test set that was previously selected. We do not evaluate SEMA, since the semantic attack does not involve perturbations and if these defenses would destroy the effectiveness of SEMA they would destroy the usefulness of all images to the recommender system.

In Table 2.8, we visualize the level of strength of defense that must be applied to the adversarial image in order for its rank to be lowered to the average *HR*@5 of a cooperative image. If the defenses presented effective protection against adversarial item images then we would expect the ▲ and ★ to appear consistently to the far left in the boxes. This is clearly not the case. We see in Table 2.8 that INSA is more difficult to defeat than EXPA, which is expected because it leverages insider knowledge. However, EXPA is clearly not easy to beat across the board. It is important to note that this test is a strong one. If these defenses would be applied in practice, they would need to be applied to all images and not just adversarial images. Image content becomes indistinguishable as compression increases, and an image 10% the size of the original image or encoded with only 2-3 bits can be expected to contain little to no item information.

2.8 Conclusion and Outlook

This work has investigated the vulnerability at the core of Top-N recommenders that use images to address cold start. We have shown that Adversarial Item Promotion (AIP) attacks allow unscrupulous merchants to artificially raise the rank of their products when a visually-aware recommender system is used for candidate ranking. Our investigation has led us to conclude that AIP attacks are a potential threat with clear practical implications. Compared with existing profile injection attacks [132;

100; 121; 16] and poisoning attacks [105; 48; 30; 47; 185] that promote items by injecting fake profiles, AIP only needs to modify the descriptive image of the item. Effective AIP attacks are easy to mount, as demonstrated by the minimal scale attack that we have studied here (cf. attack model in Section 2.3.1). In short, our work reveals that the promise of hybrid recommender systems to provide a higher degree of robustness [121] is not an absolute, and that we must proceed with caution when using images to address cold start.

Future work should dive more deeply into connection between adversarial items and user experience with the recommender system. One aspect is the relevance of adversarial items to users. Like any cold start item, users click an adversarial cold start item because it piques their interest. As an adversarial item accumulates more clicks, and enters more users' personalized lists, the main issue may be not be relevance, but rather fair competition with other potentially relevant items.

Another aspect related to user experience is the impact of image quality. If users have sensitivities that cause them to avoid products with images affected by perturbations, then attackers would need to back off to weaker attacks that make perturbations unnoticeable. In this case, defenses such as adversarial training could be more effective. More work is needed to understand approaches such as SEMA, which do not involve trading off image quality and attack strength. Alternatively, approaches that make adversarial images effective yet non-suspicious, such as [215; 89], can also be studied.

Future work must develop effective defenses against AIP attacks. An approach that easily comes to mind is the use a gatekeeper classifier to flag adversarial images at the moment that merchants upload them. It is clear that for SEMA such a classifier would be difficult to build, since SEMA attacks are created in a natural manner and are indistinguishable from cooperative images. For INSA and EXPA, a gateway filter could be built if the exact specifications of the adversarial attack, including the parameter settings, were known. However, we need to be aware that in the worst case scenario where the information of the gatekeeper is available (i.e., white-box scenario), variants on INSA and EXPA can still bypass such a classifier by constructing new loss functions [20].

We have shown in our work (cf. Section 2.7.1) that simply incorporating on-the-fly adversarial information into model training cannot guarantee a robust recommender. In addition, adversarial training requires strict hypothesis about the attack strength (ϵ) [118], and it also needs vast computational resources in practice [205]. Therefore, building a robust visually-aware recommender system is non-trivial and needs more research attention.

Future work must look at the impact of multipliers. If a single item has multiple descriptive images, attacks are more likely to go unnoticed, in particular semantic attacks that require no perturbations. Further, multiple merchants (or fake merchant profiles) could collaborate in a collusion attack. Finally, we note that al-

though, here, we have focused on e-commerce, entertainment recommender systems are vulnerable: an adversarial signal could be embedded into a thumbnail or the content itself. In sum, AIP attacks constitute an important, practical risk of using images in recommender systems and serious challenges remain to be addressed.

Chapter 3

Privacy Improvement by Availability Poisons

Perturbative availability poisons (PAPs) add small changes to images to prevent their use for model training. Current research adopts the belief that practical and effective approaches to countering PAPs do not exist. In this work, we argue that it is time to abandon this belief. We present extensive experiments showing that 12 state-of-the-art PAP methods are vulnerable to Image Shortcut Squeezing (ISS), which is based on simple compression. For example, on average, ISS restores the CIFAR-10 model accuracy to 81.73%*, surpassing the previous best preprocessing-based countermeasures by* 37.97% *absolute. ISS also (slightly) outperforms adversarial training and has higher generalizability to unseen perturbation norms and also higher efficiency. Our investigation reveals that the property of PAP perturbations depends on the type of surrogate model used for poison generation, and it explains why a specific ISS compression yields the best performance for a specific type of PAP perturbation. We further test stronger, adaptive poisoning, and show it falls short of being an ideal defense against ISS. Overall, our results demonstrate the importance of considering various (simple) countermeasures to ensure the meaningfulness of analysis carried out during the development of PAP methods.*

This Chapter is published as Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Image shortcut squeezing: Countering perturbative availability poisons with compression. International Conference on Machine Learning (ICML), 2023.

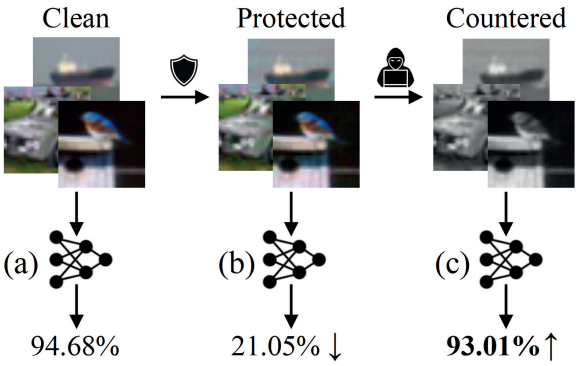


Figure 3.1: An illustration of our Image Shortcut Squeezing (ISS) for countering perturbative availability poisons (PAPs). The model accuracy is reduced by PAPs but is then restored by our ISS. Results are reported for EM [76] poisons on CIFAR-10.

3.1 Introduction

The ever-growing amount of data that is easily available online has driven the tremendous advances of deep neural networks (DNNs) [167; 102; 70; 14]. However, online data may be proprietary or contain private information, raising concerns about unauthorized use. Perturbative availability poisons (PAPs) are recognized as a promising approach to data protection and recently a large number of PAP methods have been proposed that add perturbations to images which block training by acting as shortcuts [175; 76; 52; 51]. As illustrated by Figure 4.3 (a)→(b), the high test accuracy of a DNN model is substantially reduced by PAPs.

Existing research has shown that PAPs can be compromised to a limited extent by preprocessing-based-countermeasures, such as data augmentations [76; 52] and pre-filtering [52; 26]. However, a widely adopted belief is that no approaches exist that are capable of effectively countering PAPs. Adversarial training (AT) has been proven to be a strong countermeasure [187; 200]. However, it is not considered to be a practical one, since it requires a large amount of computation and also gives rise to a non-negligible trade-off in test accuracy of the clean (non-poisoned) model [118; 213]. Further, AT trained with a specific L_p norm is hard to generalize to other norms [190; 99].

In this work, we challenge the belief that it is impossible to counter PAP methods both easily and effectively by demonstrating that they are vulnerable to simple compression. First, we categorize 12 PAP methods into three categories with respect to the surrogate models they use during poison generation: slightly-trained [49; 76; 211; 57; 193], fully-trained [175; 187; 52; 26], and surrogate-free [203; 210; 163]. Then, we analyze perturbations/shortcuts that are learned with these methods and demonstrate that they are strongly dependent on features that are learned in different training stages of the model. Specifically, we find that the methods using a slightly-trained surrogate model prefer *low-frequency* shortcuts, while those using a fully-trained model prefer *high-frequency* shortcuts.

Building on this new understanding, we propose Image Shortcut Squeezing (ISS), a simple, compression-based approach to countering PAPs. As illustrated by Figure 4.3 (b)→(c), the low test accuracy of the poisoned DNN model is restored by our ISS to be close to the original accuracy. In particular, grayscale compression is used to eliminate low-frequency shortcuts, and JPEG compression is used to eliminate high-frequency shortcuts. We also show that our understanding of high vs. low frequency can also help eliminate surrogate-free PAPs [203; 210; 163]. Our ISS substantially outperforms previously studied data augmentation and pre-filtering countermeasures. ISS also achieves comparable results to adversarial training and has three main advantages: 1) generalizability to multiple L_p norms, 2) efficiency, and 3) low trade-off in clean model accuracy (see Section 3.4.2 for details).

We further test the performance of ISS against potentially stronger PAP methods that are aware of ISS and can be adapted to it. We show that they are not ideal against our ISS. Overall, we hope our study can inspire more meaningful analyses of PAP methods and encourage future research to evaluate various (simple) countermeasures when developing new PAP methods.

In sum, we make the following main contributions:

- We identify the strong dependency of the perturbation frequency patterns on the nature of the surrogate model. Based on this new insight, we show that 12 existing perturbative availability poison (PAP) methods are indeed very vulnerable to simple image compression.

- We propose Image Shortcut Squeezing (ISS), a simple yet effective approach to countering PAPs. ISS applies image compression operations, such as JPEG and grayscale, to poisoned images for restoring the model accuracy.

- We demonstrate that ISS outperforms existing data augmentation and pre-filtering countermeasures by a large margin and is comparable to adversarial training but is more generalizable to multiple L_p norms and more efficient.

- We explore stronger, adaptive poisons against our ISS and provide interesting insights into understanding PAPs, e.g., about the model learning preference of different perturbations.

3.2 Related Work

3.2.1 Perturbative Availability Poison (PAP)

Perturbative availability poison (PAP) has been extensively studied. TensorClog (TC) [175] optimizes the poisons by exploiting the parameters of a pre-trained surrogate to cause the gradient to vanish. Deep Confuse (DC) [49] collects the training trajectories of a surrogate classifier for learning a poison generator, which is computationally intensive. Error-Minimizing (EM) poisons [76] minimizes the

classification errors of images on a surrogate classifier with respect to their original labels in order to make them "unlearnable examples". The surrogate is also alternatively updated to mimic the model training dynamics during poison generation. Hypocritical (HYPO) [187] follows a similar idea to EM but uses a pre-trained surrogate rather than the above bi-level optimization. Targeted Adversarial Poisoning (TAP) [52] also exploits a pre-trained model but minimizes classification errors of images with respect to incorrect target labels rather than original labels.

Robust Error-Minimizing (REM) [57] improves the poisoning effects against adversarial training (with a relatively small norm) by replacing the normally-trained surrogate in EM with an adversarially-trained model. Similar approaches [198; 200] on poisoning against adversarial training are also proposed. The usability of poisoning is also validated in scenarios requiring transferability [157] or involving unsupervised learning [68; 214].

There are also studies focusing on revising the surrogate, e.g., Self-Ensemble Protection [26], which aggregates multiple training model checkpoints, and NTGA [211], which adopts the generalized neural tangent kernel to model the surrogate as Gaussian Processes [80]. ShortcutGen (SG) [193] learns a poison generator based on a randomly initialized fixed surrogate and shows its efficiency compared to the earlier generative method, Deep Confuse.

Different from all the above methods, recent studies also explore surrogate-free PAPs [45; 210; 163]. Intuitively, simple patterns, such as random noise [76] and semantics (e.g., MNIST-like digits) [45], can be used as learning shortcuts when they form different distributions for different classes. Very recent studies also synthesize more complex, linear separable patterns to boost the poisoning performance based on sampling from a high dimensional Gaussian distribution [210] and further refining it by introducing the autoregressive process [163]. One Pixel Shortcut (OPS) specifically explores the model vulnerability to sparse poisons and shows that perturbing only one pixel is sufficient to generate strong poisons [203].

In the domain of facial recognition, PAP methods, e.g., Fawkes [173] and LowKey [28], have also been studied. However, their protection algorithms closely resemble the PAPs as discussed above. Specifically, Fawkes adopts a feature-layer loss similar to SEP and a robust surrogate model similar to REM, to boost transferability. LowKey adopts ensemble surrogate models similar to SEP and a pre-processing step similar to TAP, to boost transferability and imperceptibility.

In this work, we evaluate our ISS against 12 representative PAP methods as presented above. In particular, we consider poisons constrained by different L_p norms. Because of their technical similarity to two of the 12 approaches, we do not consider Fawkes and LowKey in our evaluation.

3.2.2 PAP Countermeasures

As mentioned in Section 3.1, existing research has mainly relied on adversarial training (AT) for countering PAPs [187; 200]. However, AT is not practical due to the requirement of large computations and the non-negligible trade-off in test accuracy of the clean model [118; 213]. In addition, image preprocessing, e.g., data augmentations [76; 52] and pre-filtering [52; 26], also show substantial effects but not comparable to AT. In the domain of face recognition, countermeasures are also discussed but either require stronger assumptions or lack a concrete algorithm [149]. See more discussions in Appendix 3.6.4.

In this work, we compare our ISS against existing countermeasures and particularly highlight its generalizability to unknown norms [190; 99] and simplicity.

3.2.3 Adversarial Perturbations and Countermeasures

Simple image compressions, such as JPEG, bit depth reduction, and smoothing, are effective for countering adversarial perturbations based on the assumption that they are inherently high-frequency noise [44; 32; 207]. Other image transformations commonly used for data augmentations, e.g., resizing, rotating, and shifting, are also shown to be effective [204; 188; 42]. However, such image pre-processing operations may be bypassed when the attacker is aware of them and then adapted to them [18]. Differently, adversarial training (AT) is effective against adaptive attacks and is considered to be the most powerful defense so far [191].

Besides (training-time) data poisons, adversarial perturbations can also be used for data protection, but at inference time. Related research has explored person-related recognition [127; 128; 165; 152] and social media mining [101; 106; 114]. An overview of inference-time data protection in images is provided by [130].

Our ISS is based on compression. We specifically evaluate its compression effects in Section 3.4.6.

3.3 Analysis of Perturbative Availability Poisons

3.3.1 Problem Formulation

We formulate the problem of countering perturbative availability poisons (PAPs) in the context of image classification. There are two parties involved, the data *protector* and *exploiter*. The data protector poisons their own images to prevent them from being used by the exploiter for training a well-generalizable classifier. Specifically, here the poisoning is achieved by adding imperceptible perturbations. The data exploiter is aware that their collected images may contain poisons and so apply countermeasures to ensure their trained classifier is still well-generalizable. The success of the countermeasure is measured by the accuracy of the classifier on clean test images, and the higher, the more successful.

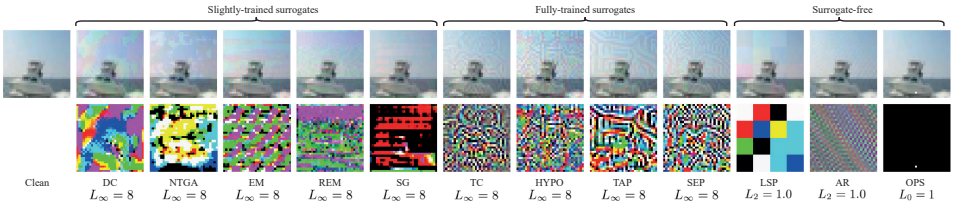


Figure 3.2: Poisoned CIFAR-10 images with corresponding perturbations. Perturbations are re-scaled to $[0, 1]$ for visualization.

Formally stated, the protector aims to make a classifier F generalize poorly on the clean image distribution $\mathcal{D}$, from which the clean training set $\mathcal{S}$ is sampled:

$$\max_{\boldsymbol{\delta}} \ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\Big[\mathcal{L}\left(F(\boldsymbol{x};\boldsymbol{\theta}'(\boldsymbol{\delta})),y\right)\Big] \tag{3.1}$$

$$\text{s.t. } \boldsymbol{\theta}'(\boldsymbol{\delta}) = \underset{\boldsymbol{\theta}(\boldsymbol{\delta})}{\operatorname{argmin}} \sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{S}} \mathcal{L}(F(\boldsymbol{x}_i+\boldsymbol{\delta}_i;\boldsymbol{\theta}(\boldsymbol{\delta}),y_i), \tag{3.2}$$

where $\boldsymbol{\theta}(\boldsymbol{\delta})$ represents the parameters of the poisoned classifier, F, where $\boldsymbol{\delta}$ denotes the additive perturbations with ϵ as the L_p bound. $\mathcal{L}(\cdot;\cdot)$ is the cross-entropy loss, which takes as input a pair of model output $F(\boldsymbol{x}_i;\boldsymbol{\theta})$ and the corresponding label y_i.

The exploiter aims to counter the poisons by applying a countermeasure C to restore the model accuracy even when it is trained on poisoned data $\mathcal{P}$:

$$\min_{\boldsymbol{\theta}} \sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{P}} \mathcal{L}(F(C(\boldsymbol{x}_i+\boldsymbol{\delta}_i);\boldsymbol{\theta}),y_i). \tag{3.3}$$

3.3.2 Categorization of Existing PAP Methods

We carried out an extensive survey of existing PAP methods, which allowed us to identify three categories of them regarding the type of their used surrogate classifiers. These three categories are: Generating poisons 1) with a slightly-trained surrogate, 2) with a fully-trained surrogate, and 3) in a surrogate-free manner. Table 4.2 provides an overview of this categorization. In the first category, the surrogate is at its early training stage. Existing methods in this category either fixes [211; 193] or alternatively updates [49; 76; 57] the surrogate during optimizing the poisons. In the second category, the surrogate has been fully trained. Existing methods in this category fix the surrogate [175; 187; 52; 26] but in principle, it may also be possible that the model is alternatively updated. In the third category, no surrogate is used but the poisons are synthesized by sampling from Gaussian distributions [210; 163] or optimized with a perceptual loss [203].

Table 3.1: Categorization of existing PAP methods.

PAP Methods	Surrogate Model
DC [49] NTGA [211] EM [76] REM [57] SG [193]	Slightly-Trained
TC [175] HYPO [187] TAP [52] SEP [26]	Fully-Trained
LSP [210] AR [163] OPS [203]	Surrogate-Free

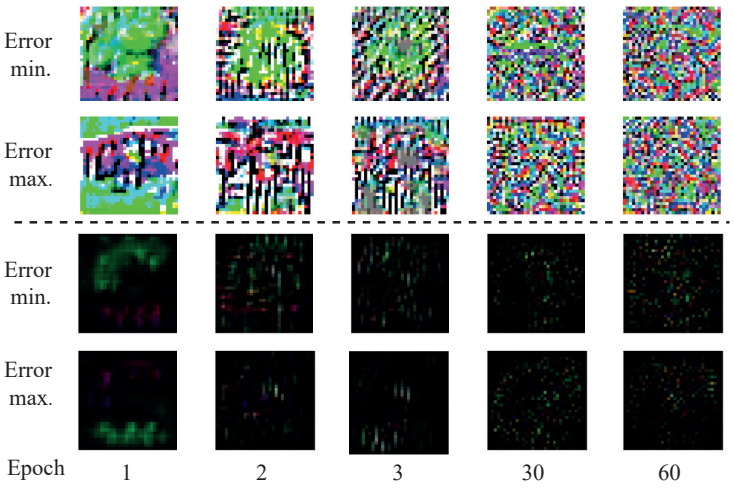


Figure 3.3: Perturbation visualizations for poisons generated using surrogate at its various training epochs. Perturbations with $L_\infty = 8$ (top) and $L_2 = 1$ (bottom) are shown. Both the error minimizing and maximizing losses are considered. Perturbations at later epochs exhibit higher frequency.

3.3.3 Frequency-based Interpretation of Perturbations

Poisoned CIFAR-10 images and their corresponding perturbations for the 12 methods are visualized in Figure 3.2. As can be seen, the four methods that adopt a fully-trained surrogate tend to generate perturbations in patterns having a high spatial frequency. This is consistent with the common finding in the adversarial example literature that adversarial perturbations are normally high-frequency [65]. In contrast, the five methods that adopt a slight-trained surrogate exhibit spatially low-frequency patterns but large differences across color channels.

We hypothesize that the above phenomenon can be explained by the frequency principle [151; 208; 117], that is, deep neural networks often fit target functions from low to high frequencies during training. Accordingly, the poisons optimized against a slightly-trained model capture low-frequency patterns while those optimized against a fully-trained model capture high-frequency patterns [151; 208; 117]. In order to validate this hypothesis, we further try optimizing poisons using either the error-minimizing or error-maximizing loss against the surrogate at its various training epochs. We visualize the resulting poisoned images and their corresponding perturbations in Figure 3.3. As can be been, the spatial frequency of the perturbations gets increasingly higher as the surrogate goes to a later training epoch.

Different from those surrogate-based methods, the three surrogate-free methods have full control of the perturbation patterns they aim to synthesize. However, we notice that they still follow our frequency-based interpretation of perturbation patterns. Specifically, the perturbations of LSP [210] are uniformly upsampled from a Gaussian distribution and so exhibit patch-based low-frequency patterns. On the other hand, the perturbations of AR [163] are generally based on sliding convolutions over the image and so exhibit texture-based high-frequency patterns. OPS [203] perturbations only contain one pixel and so can be treated as an extreme case of high-frequency patterns.

3.3.4 Our Image Shortcut Squeezing

Based on the above new frequency-based interpretation, we propose Image Shortcut Squeezing (ISS), a simple, image compression-based countermeasure against PAPs. We rely on different compression operations suitable for eliminating different types of perturbations. Overall, a specific compression operation is applied to the $C(\cdot)$ in Eq. 3.3.

For perturbations with low frequency but large differences across color channels, we use grayscale transformation to suppress such color differences. We expect grayscale transformation to not sacrifice too much the test accuracy of a clean model because color information is known to contribute little to the DNNs' performance in differentiating objects [206]. For perturbations with high frequency, we follow existing research on eliminating adversarial perturbations to use common image compression operations, such as JPEG and bit depth reduction (BDR) [44; 32; 207]. We expect grayscale transformation to not sacrifice too much the test accuracy of a clean model because DNNs are known to be resilient to small amounts of image compression, e.g., JPEG with a higher quality factor than 10 [41].

3.4 Experiments

In this section, we evaluate our Image Shortcut Squeezing (ISS) and other countermeasures against 12 representative PAP methods. We focus our experiments on the basic setting in which the surrogate (if it is used) and target models are the same and the whole training set is poisoned. We also explore more challenging poisoning

scenarios with unseen target models or partial poisoning (poisoning a randomly selected proportion or a specific class).

3.4.1 Experimental Settings

Datasets and models. We consider three datasets: CIFAR-10 [96], CIFAR-100 [96], and a 100-class subset of ImageNet [36]. If not mentioned specifically, on CIFAR-10 and CIFAR-100, we use 50000 images for training and 10000 images for testing. For the ImageNet subset, we select 20% images from the first 100 classes of the official ImageNet training set for training and all corresponding images in the official validation set for testing. If not mentioned specifically, ResNet-18 (RN-18) [70] is used as the surrogate model and target model. To study transferability, we consider target models with diverse architectures: ResNet-34 [70], VGG-19 [178], DenseNet-121 [75], MobileNet-V2 [162], and ViT [43].

Training and poisoning settings. We train the CIFAR-10/100 models for 60 epochs and the ImageNet models for 100 epochs. We use SGD with a momentum of 0.9, a learning rate of 0.025, and cosine weight decay. We adopt the torchvision transforms module for implementing Grayscale, JPEG, and bit depth reduction (BDR) in our Image Shortcut Squeezing (ISS). We consider 12 representative existing poisoning methods as listed in Table 4.2 under various L_p norm bounds. A brief description of 12 methods can be found in Appendix 3.6.1. Specifically, we follow existing work and use $L_\infty = 8$, $L_2 = 1.0$, and $L_0 = 1$.

3.4.2 Evaluation in the Common Scenario

We first evaluate our ISS against 12 representative poisoning methods in the common scenario where the surrogate and target models are the same and the whole training dataset is poisoned. Experimental results on CIFAR-10 shown in Table 3.2 demonstrate that ISS can substantially restore the clean test accuracy of poisoned models in all cases. Consistent with our new insight in Section 3.3.3, grayscale yields the best performance in countering methods that rely on low-frequency perturbations with large color differences (see more results by other color compression methods on EM in Appendix 3.6.3). In contrast, JPEG and BDR are the best against methods that rely on high-frequency perturbations. Additional results for other hyperparameters of JPEG and BDR in Table 3.13 of Appendix 3.6.2 show that milder settings yield worse results. In addition, we can also apply ISS without determining the specific poisons by directly using Gray+JPEG. The results demonstrate that this combination is globally effective against all 12 PAP methods, with only a small decrease in clean model accuracy.

Table 3.2: Clean test accuracy (%) of models trained on CIFAR-10 poisons and with our Image Shortcut Squeezing (Gray and JPEG) vs. other countermeasures. Note that TC is known to not work well under small norms, e.g., our $L_\infty = 8$ [52]. Hyperparameters for different countermeasures can be found in Appendix 3.6.2.

Norm	Poisons/Countermeasures	w/o	Cutout	CutMix	Mixup	Gaussian	Mean	Median	BDR	Gray	JPEG	Gray+JPEG	AT
	Clean (no poison)	94.68	95.10	95.50	95.01	94.17	45.32	85.94	88.65	92.41	85.38	83.79	84.99
$L_\infty = 8$	DC [49]	16.30	15.14	17.99	19.39	17.21	19.57	15.82	61.10	**93.07**	81.84	83.09	78.00
	NTGA [211]	42.46	42.07	27.16	43.03	42.84	37.49	42.91	62.50	**74.32**	69.49	69.86	70.05
	EM [76]	21.05	20.63	26.19	32.83	12.41	20.60	21.70	36.46	**93.01**	81.50	83.06	84.80
	REM [57]	25.44	26.54	29.02	34.48	27.44	25.35	31.57	40.77	**92.84**	82.28	83.00	82.99
	SG [193]	33.05	24.12	29.46	39.66	31.92	46.87	49.53	70.14	**86.42**	79.49	79.21	76.38
	TC [175]	88.70	86.70	88.43	88.19	82.58	72.25	84.27	84.85	79.75	85.29	82.43	84.55
	HYPO [187]	71.54	70.60	67.54	72.54	72.46	40.27	65.53	83.50	61.86	**85.45**	82.94	84.91
	TAP [52]	8.17	10.04	10.73	19.14	9.26	21.82	32.75	45.99	9.11	**83.87**	81.94	83.31
	SEP [26]	3.85	4.47	9.41	15.59	3.96	14.43	35.65	47.43	3.57	**84.37**	82.18	84.12
$L_2 = 1.0$	LSP [210]	19.07	19.87	20.89	26.99	19.25	28.85	29.85	66.19	82.47	**83.01**	79.05	84.59
	AR [163]	13.28	12.07	12.39	13.25	15.45	45.15	70.96	31.54	34.04	**85.15**	82.81	83.17
$L_0 = 1$	OPS [203]	36.55	67.94	76.40	45.06	19.29	23.50	**85.16**	53.76	42.44	82.53	79.10	14.41

Table 3.3: Additional results on CIFAR-10 with larger perturbation norms: $L_2 = 2.0$ for LSP and $L_\infty = 16$ for the rest.

Poisons	w/o	Cutout	CutMix	Mixup	Gray	JPEG	AT
Clean	94.68	95.10	95.50	95.01	92.41	88.65	84.99
EM	16.33	14.0	13.41	20.22	60.85	**63.44**	61.58
REM	24.89	25.0	22.85	29.51	42.85	**76.59**	80.14
HYPO	58.3	54.22	48.26	57.27	45.38	**85.07**	84.90
TAP	10.98	10.96	9.46	17.97	6.94	**84.19**	83.35
SEP	3.84	8.90	15.79	9.27	5.70	**84.35**	84.07
LSP	19.07	19.87	20.89	26.99	**82.47**	83.01	84.59

Table 3.4: Additional results on CIFAR-100.

Poisons	w/o	Cutout	CutMix	Mixup	Gray	JPEG
Clean	77.44	76.72	80.50	78.56	71.79	57.79
EM	7.25	6.70	7.03	10.68	**67.46**	56.01
REM	9.37	12.46	10.40	15.05	**57.27**	55.77
TC	57.52	60.56	59.19	59.77	47.93	58.94
TAP	9.00	10.30	8.73	19.16	8.84	**83.77**
SEP	3.21	3.21	3.98	7.49	2.10	**58.18**
LSP	3.06	4.43	6.12	5.61	44.62	**53.49**
AR	3.01	2.85	3.49	2.19	24.99	**57.87**
OPS	23.78	**57.98**	56.03	22.71	32.62	54.92

Table 3.5: Additional results on ImageNet subset. Following their original papers, NTGA and DC are tested with only two classes.

Poisons	w/o	Cutout	CutMix	Mixup	Gray	JPEG
Clean	62.04	61.14	65.100	64.32	58.24	58.20
EM	31.52	30.42	42.98	21.44	49.78	**49.88**
REM	11.12	11.62	12.50	17.62	**44.70**	18.16
TAP	24.64	23.00	18.72	28.62	24.30	**44.74**
LSP	26.32	27.64	17.22	2.5	**31.42**	30.78
NTGA	70.79	63.42	70.53	68.42	**83.42**	76.58
DC	65.00	-	-	-	**85.00**	74.00

3.4.3 Evaluation in Challenging Scenarios

Partial poisoning. In practical scenarios, it is common that only a proportion of the training data can be poisoned. Therefore, we follow existing work [52; 77] to test such partial poisoning settings. We poison a certain proportion of the training data and mix it with the rest clean data for training the target model.

Specifically, we test two partial poisoning settings: first, randomly selecting a certain proportion of the images, and, second, selecting a specific class. In the first setting, as shown in Table 3.6, the poisons are effective only when a very large proportion of the training data is poisoned. For example, on average, even when 80% of data are poisoned, the model accuracy is only reduced by about 10 %. In the second setting, we choose to poison all training samples from class `automobile` on CIFAR-10. Table 3.7 demonstrates that almost all poisoning methods are very effective in the full poisoning setting. In both settings, our ISS is effective against all PAP methods.

Table 3.6: Clean test accuracy (%) of CIFAR-10 target models under different poisoning proportions. TC is tested with $L_\infty = 26$.

Poisons	ISS	0.1	0.2	0.4	0.6	0.8	0.9
DC	w/o	94.29	94.26	93.20	91.66	87.19	80.14
	Gray	92.73	92.57	92.37	91.51	90.49	89.50
	JPEG	84.89	85.26	84.43	83.61	83.02	82.69
EM	w/o	94.37	93.63	92.62	91.07	86.63	79.57
	Gray	92.60	92.62	92.52	92.23	90.96	89.69
	JPEG	84.61	84.79	84.96	84.86	84.93	84.40
REM	w/o	94.39	94.56	94.37	94.43	94.19	81.39
	Gray	92.63	92.81	92.78	92.82	92.73	86.62
	JPEG	84.64	85.53	84.82	85.37	85.38	82.44
SG	w/o	94.47	94.40	93.46	91.21	87.75	83.40
	Gray	92.81	92.65	91.90	90.65	88.44	85.26
	JPEG	84.94	84.61	84.11	82.66	80.76	79.38
TC	w/o	93.81	94.09	93.70	93.59	93.02	91.47
	Gray	91.98	92.38	92.03	91.96	91.03	87.71
	JPEG	85.24	85.01	85.23	85.28	85.23	84.37
HYPO	w/o	93.94	94.43	93.34	92.56	90.64	89.35
	Gray	92.59	92.39	91.37	90.06	88.03	86.37
	JPEG	85.61	85.18	85.39	85.21	85.25	85.10
TAP	w/o	94.09	93.94	92.75	91.27	88.42	85.98
	Gray	92.62	91.94	90.73	89.26	85.93	83.18
	JPEG	85.24	84.42	84.86	84.98	84.51	84.36
SEP	w/o	94.12	93.45	92.76	91.22	87.82	85.01
	Gray	92.57	92.04	91.09	89.25	86.31	82.95
	JPEG	85.27	85.27	85.25	84.71	84.07	84.80
LSP	w/o	94.69	94.42	92.81	91.38	88.07	82.26
	Gray	93.12	92.56	92.67	92.20	90.78	89.65
	JPEG	85.01	84.58	84.88	83.49	83.27	81.67
AR	w/o	94.66	94.38	93.82	91.80	88.42	82.36
	Gray	92.85	92.69	92.53	91.24	89.88	85.35
	JPEG	85.37	84.75	85.35	85.35	85.07	87.27
OPS	w/o	94.47	94.11	92.61	91.49	87.19	82.65
	Gray	92.65	92.27	91.36	89.34	85.24	81.37
	JPEG	84.75	84.88	84.55	83.98	82.87	81.33

Table 3.7: Partial poisoning for class `automobile` on CIFAR-10. TC is tested with $L_\infty = 26$.

Poisons	w/o	Gray	JPEG	BDR
DC	1.60	69.00	88.30	52.20
NTGA	51.70	94.20	90.40	75.30
EM	0.10	48.60	94.30	9.60
REM	0.80	34.40	90.40	2.50
SG	27.75	88.39	78.59	70.05
TC	0.50	0.20	92.50	37.20
HYPO	4.00	3.00	94.90	56.80
TAP	0.00	0.10	93.90	38.10
SEP	0.00	0.00	94.70	15.50
LSP	67.30	86.90	95.10	83.20
AR	97.70	97.60	94.60	95.10
OPS	28.90	28.50	93.60	72.10

Transferability to unseen models. In realistic scenarios, the protector may not know the details of the target model. In this case, the transferability of the poisons is desirable. Table 3.8 demonstrates that all PAP methods achieve high transferability to diverse model architectures and our ISS is effective against all of them. It is also worth noting that there is no clear correlation between the transferability and the similarity between the surrogate and target models. For example, transferring from ResNet-18 to ViT is not always harder than to other CNN models.

3.4.4 Adaptive Poisons to ISS

In the adversarial example literature, image compression operations can be bypassed when the attacker is adapted to them [177; 18]. Similarly, we evaluate strong adaptive poisons against our ISS using two PAP methods, EM (L_∞) and LSP (L_2). We assume that the protector can be adapted to grayscale and/or JPEG in our ISS. Specifically, for EM, we add a differentiable JPEG compression module [177] and/or a differentiable grayscale module into its bi-level poison optimization process. For LSP, we increase the patch size to 16×16 to decrease high-frequency features so that JPEG will be less effective, and we make sure the pixel values are the same for three channels to bypass grayscale.

Table 3.9 demonstrates that for EM, the adaptive grayscale poisons are effective against grayscale, but adaptive JPEG-10 noises are not effective against JPEG. As hinted by [177], using an ensemble of JPEG with different quality factors might be necessary for better adaptive poisoning. We also implement BPDA [3] with the same JPEG quality factor (i.e., JPEG-10) and find that our ISS still ensures a very high model accuracy, i.e., 83.70 %. For LSP, we observe that even though adaptive LSP is more effective against the combination of JPEG and grayscale than the other two individual compressions, it is insufficient to serve as a good adaptive protector. On the other hand, adaptive LSP also fails against the model without

ISS, indicating that the additional operations (grayscale and larger patches) largely constrain its poisoning effects.

Table 3.8: Clean test accuracy (%) of CIFAR-10 target models in the transfer setting. Note that AR, LSP, and OPS are surrogate-free. Four CNN models (ResNet-34, VGG-19, DenseNet-121, and MobileNet-V2) and one ViT are considered as the target model. TensorClog (TC) is tested with $L_\infty = 26$.

Poisons	ISS	R34	V19	D121	M2	ViT
DC	w/o	18.06	16.59	16.05	17.81	24.09
	Gray	83.13	80.32	83.93	78.78	44.83
	JPEG	82.64	80.34	83.38	80.30	53.35
NTGA	w/o	40.19	47.13	16.67	40.75	31.82
	Gray	71.84	76.89	64.07	62.28	58.25
	JPEG	67.00	72.17	73.76	70.18	53.00
EM	w/o	29.96	34.70	30.61	30.10	18.84
	Gray	86.97	87.03	84.84	82.81	63.28
	JPEG	84.21	82.46	84.86	82.20	56.33
REM	w/o	25.88	29.04	28.31	24.08	32.22
	Gray	75.20	77.99	70.53	66.21	63.00
	JPEG	82.35	80.70	81.74	80.01	56.13
SG	w/o	29.64	48.5	28.88	30.75	18.11
	Gray	86.53	87.12	86.07	81.34	42.22
	JPEG	79.57	77.78	79.77	75.87	56.27
TC	w/o	87.71	85.47	78.04	78.51	69.86
	Gray	78.48	75.14	66.72	62.39	61.86
	JPEG	84.56	82.66	83.95	82.60	55.51
HYPO	w/o	80.64	81.59	81.48	78.27	67.49
	Gray	75.25	76.65	74.29	69.81	53.02
	JPEG	85.55	83.39	85.03	83.95	55.17
TAP	w/o	7.89	8.59	8.64	10.02	41.32
	Gray	9.38	11.51	8.77	8.29	42.49
	JPEG	84.42	81.95	84.28	82.24	57.35
SEP	w/o	3.11	6.70	4.41	5.29	25.56
	Gray	4.00	5.40	4.20	4.70	22.23
	JPEG	84.64	83.38	84.55	83.25	56.94
LSP	w/o	15.98	17.39	19.79	17.32	26.65
	Gray	71.10	82.11	73.06	70.61	53.36
	JPEG	79.57	78.72	79.66	76.79	60.41
AR	w/o	21.31	19.78	13.54	16.08	22.91
	Gray	70.54	76.92	67.35	62.01	53.22
	JPEG	85.62	83.95	85.46	83.50	54.88
OPS	w/o	37.06	36.3	40.03	27.35	30.25
	Gray	44.29	42.21	38.32	38.71	21.77
	JPEG	82.84	80.70	82.83	80.42	62.93

3.4.5 Adaptive Poisons to ISS

In the adversarial example literature, image compression operations can be bypassed when the attacker is adapted to them [177; 18]. Similarly, we evaluate strong adaptive poisons against our ISS using two PAP methods, EM (L_∞) and LSP (L_2). We assume that the protector can be adapted to grayscale and/or JPEG in our ISS. Specifically, for EM, we add a differentiable JPEG compression module [177] and/or a differentiable grayscale module into its bi-level poison optimization process. For LSP, we increase the patch size to 16×16 to decrease high-frequency features so that JPEG will be less effective, and we make sure the pixel values are the same for three channels to bypass grayscale.

Table 3.9 demonstrates that for EM, the adaptive grayscale poisons are effective against grayscale, but adaptive JPEG-10 noises are not effective against JPEG. As hinted by [177], using an ensemble of JPEG with different quality factors might be necessary for better adaptive poisoning. We also implement BPDA [3] with the same JPEG quality factor (i.e., JPEG-10) and find that our ISS still ensures a very high model accuracy, i.e., 83.70 %. For LSP, we observe that even though adaptive LSP is more effective against the combination of JPEG and grayscale than the other two individual compressions, it is insufficient to serve as a good adaptive protector. On the other hand, adaptive LSP also fails against the model without ISS, indicating that the additional operations (grayscale and larger patches) largely constrain its poisoning effects.

Given that the protector may have full knowledge of our ISS, we believe that better-designed adaptive poisons can bypass our ISS in the future.

Table 3.9: Clean test accuracy (%) of four different target models under EM poisoning and its adaptive variants on CIFAR-10. Results are reported for $L_\infty = 8$ and Table 3.14 in Appendix reports results of EM for $L_\infty = 16$, which follow the same pattern.

Poisons	w/o	Gray	JPEG	G&J	Ave.
EM	21.05	93.01	81.50	83.06	69.66
EM-Gray	17.81	**16.60**	76.71	74.16	**46.32**
EM-JPEG	**17.11**	89.18	83.11	82.85	68.06
EM-G&J	48.93	46.29	**69.48**	**66.26**	57.74
LSP	19.07	82.47	83.01	79.05	65.90
LSP-G&J	93.01	90.34	84.38	**82.13**	87.47

3.4.6 Further Analyses

Working Mechanism of ISS. Here we illustrate the working mechanism of our ISS by ensuring that the poisons are the exact factor that is used by the poisoned model for prediction. To this end, we follow [52] to use poisoned images to both train and test the model. In this case, if the test accuracy (on poisoned images) is high, it demonstrates that the perturbations in the poisoned images are actually learned

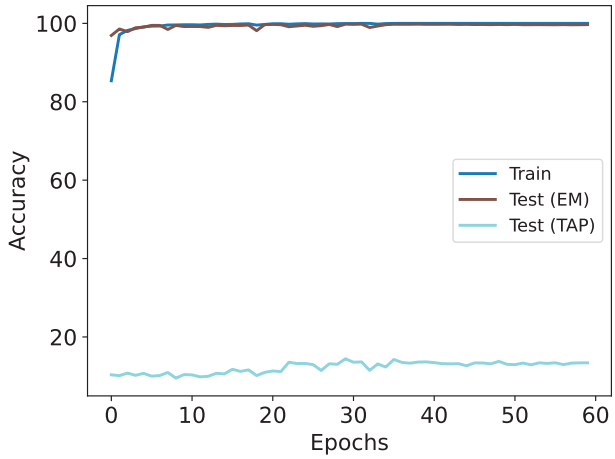


Figure 3.4: Relative model preference of different poisons.

by the model. In addition, we also train and test the model on poisoned images but differently, the testing (poisoned) images are pre-processed using our ISS. In this case, if the test accuracy (on poisoned images) decreases, it demonstrates that ISS can suppress the perturbations at inference time. The results in Table 3.10 validate our hypotheses.

Table 3.10: Test accuracy (%) on clean, poisoned, and ISS-preprocessed poisoned test sets of models that are trained on different poisons.

Test/ Poisons	DC	NTGA	EM	REM	SG	TC	HYPO	TAP	SEP	LSP	AR	OPS
Clean	17.96	-	16.77	26.04	37.50	87.86	73.06	11.63	4.91	15.29	16.37	17.50
Poisoned	97.20	97.85	99.85	99.97	96.72	93.79	99.98	100.0	99.99	100.0	99.94	99.83
Poisoned+ISS	11.17	24.86	11.89	20.68	25.39	24.16	16.68	11.33	10.13	10.06	13.83	12.61

Relative Model Preference of different poisons. We explore the relative model preference of low-frequency vs. high-frequency poisons. This scenario is practically interesting because the same online data might be poisoned by different methods. Inspired by the experiments on the model preference of MNIST vs. CIFAR data in [171], we simply add up the EM and TAP perturbations for each image. The perturbation norm is doubled accordingly. For example, for perturbations with $L_\infty = 8$, the composite perturbations range from -16 to 16. We train a model (using the original image labels) on the composite perturbations of EM and TAP and test it on either EM or TAP perturbations.

As shown in Figure 3.4, the model converges fast and reaches a high test accuracy on EM but not on the TAP. It indicates that TAP perturbations are less preferred than EM perturbations by the model during training.

ISS for a combination of different types of poisons. We create poisons

by combining the two well-known low-frequency and high-frequency methods, i.e., EM and TAP. Specifically, we take the average of the perturbations of these two methods. As shown in Table 3.11, our ISS is still effective against this combination.

Table 3.11: Clean test accuracy (%) of models trained on CIFAR-10 poisons that is a combination of low-frequency poison EM and high-frequency TAP.

Poisons/ISS	w/o	Gray	JPEG
EM	21.05	93.01	81.50
TAP	8.17	9.11	83.87
EM+TAP	36.07	18.93	84.62

ISS for both training and testing. Our ISS only applies to the training data for removing the poisons. However, in this case, it may cause a possible distribution shift between the training and test data. Here we explore such a shift by comparing ISS with another variant that applies compression to both the training and test data. Table 3.12 demonstrates that in most cases, these two versions of ISS do not lead to substantial differences.

Table 3.12: Clean test accuracy (%) for ISS (Gray and JPEG), which applies compression only to training data or to both training and test data (denoted with suffix-TT).

Poisons	Gray-TT	Gray	JPEG-TT	JPEG
Clean	**92.62**	92.41	79.56	**85.38**
DC	83.79	**93.07**	79.41	**81.84**
NTGA	65.42	**74.32**	62.84	**69.49**
EM	90.75	**93.01**	78.96	**81.50**
REM	73.38	**92.84**	79.39	**82.28**
SG	**88.26**	86.42	72.96	**79.49**
TC	**76.41**	75.88	79.42	**83.69**
HYPO	**75.20**	61.86	79.63	**85.60**
TAP	**9.53**	9.11	78.65	**83.87**
SEP	2.93	**3.57**	79.28	**84.37**
LSP	**76.23**	75.77	68.73	**78.69**
AR	68.95	**69.37**	79.26	**85.38**
OPS	**46.53**	42.44	76.87	**82.53**

3.5 Conclusion and Outlook

In this work, we challenge the common belief that there are no practical and effective countermeasures to perturbative availability poisons (PAPs). Specifically, we show that 12 state-of-the-art PAP methods can be substantially countered by Image Shortcut Squeezing (ISS), which is based on simple compression. ISS outperforms other previously studied countermeasures, such as data augmentations and adversarial training. Our in-depth investigation leads to a new insight that the property of PAP perturbations depends on the type of surrogate model used during

poison generation. We also show the ineffectiveness of adaptive poisons to ISS. We hope that further studies could consider various (simple) countermeasures during the development of new PAP methods.

For future work, on the countermeasure side, we would further improve ISS on the trade-off between its effectiveness and the decrease of clean model accuracy by exploring other (simple) accuracy-preserving operations. In addition, to achieve a countermeasure that is more effective against unknown poisons, it would be promising to explore more advanced combination strategies of operations or conduct automatic attack identification and then apply attack-specific operations. On the protection side, we encourage future work to develop effective (adaptive) protection methods against our ISS and other potential countermeasures.

3.6 Appendix

3.6.1 Brief Descriptions of Implemented PAP Methods

- **Deep Confuse (DC)** [49]: Perturbations are generated from a U-Net [159] on CIFAR-10 and encoder-decoder model on two-class ImageNet. The generators are trained on the output of a pseudo-updated classifier, where the classification model is first trained on clean data and then trained on adversarial data to update the generator. We use the implementation from the official GitHub repository.

- **Neural Tangent Generalization Attacks (NTGA)** [211] (target model-agnostic): NTGA uses Neural Tangent Kernels to approximate target networks and then leverages the approximation to generate perturbations. We use the poisons provided in the official GitHub repository.

- **Error-Minimizing perturbations (EM)** [77]: Bi-level optimizing error-minimizing perturbations after certain steps of training on perturbed samples that are from the last iteration. The surrogate model is trained on-the-fly with perturbed training samples. We use the implementation from the official GitHub repository.

- **Robust Error-Minimizing perturbations (REM)** [57]: Same as EM, but the model is adversarially trained and the perturbations generation is equipped with expectation over transformation technique (EOT) [4]. We use the implementation from the official GitHub repository.

- **Shortcut generator (SG)** [193]: Perturbations are generated from a ResNet-like encoder-decoder model from [125]. Different from another generative poi-

`https://github.com/kingfengji/DeepConfuse`
`https://github.com/lionelmessi6410/ntga`
`https://github.com/HanxunH/Unlearnable-Examples/`
`https://github.com/fshp971/robust-unlearnable-examples/`

soning Deep Confuse, the discriminator model is randomly initialized without training. We use the CIFAR-10 poisons (version 'SG') provided by the authors by private communication.

- **TensorClog (TC)** [175]: A second-order derivative with respect to training data is calculated to iteratively optimize the perturbations to minimize the gradients of model loss with respect to the weights of model layers. We use the implementation from the official GitHub repository for poisons ($L_\infty = 26$) on CIFAR-10. We also use the implementation from `https://github.com/lhfowl/adversarial_poisons` for poisons ($L_\infty = 8, 16$) on CIFAR-10.

- **Hypocritical perturbations (HYPO)** [187]: Similar to EM, but the error-minimizing perturbations are generated on a pre-trained surrogate model which is trained on clean data. We use the implementation from the official GitHub repository.

- **Targeted Adversarial Poisoning (TAP)** [52]: Targeted adversarial examples by PGD [118] and Spatial Transformer Networks (STN) module [81]. The poisoning target labels are different from the original labels, but the target labels are the same for poisoning images whose clean versions are from the same class. We use the implementation from the official GitHub repository.

- **Self-Ensemble Protection (SEP)** [26] SEP ensembles intermediate checkpoints when training on the clean training set to create perturbations. SEP is currently the state-of-the-art protection on CIFAR-10. We use the implementation from the official GitHub repository.

- **Linear separable Synthetic Perturbations (LSP)** [210]: Linearly separable Gaussian samples are listed by order and then up-scaled to the size of the image. Perturbations that are sampled from the same Gaussian are added to the same class. We use the implementation from the official GitHub repository.

- **AutoRegressive poisoning (AR)** [163] Autoregressive process generates perturbations that CNN favors during training. We use the CIFAR-10 poisons provided by the authors in the official GitHub repository.

- **One Pixel Shortcut (OPS)** [203]: OPS generates one pixel shortcut by searching the pixel that creates the most significant mean pixel value change for all images from one class. The perturbations are *dataset-dependent*.

`https://github.com/JC-S/TensorClog_Public`
`https://github.com/TLMichael/Delusive-Adversary`
`https://github.com/lhfowl/adversarial_poisons`
`https://github.com/Sizhe-Chen/SEP`
`https://github.com/dayu11/Availability-Attacks-Create-Shortcuts/`
`https://github.com/psandovalsegura/autoregressive-poisoning`
`https://github.com/cychomatica/One-Pixel-Shotcut`

Table 3.13: JPEG with different quality factors and BDR with different bit depth.

Poisons	w/o	JPEG Compression					Bit depth reduction				
		10	30	50	70	90	2	3	4	5	6
Clean (no poison)	94.68	85.38	89.49	90.80	91.85	93.06	88.65	92.22	93.45	94.46	94.55
DC [49]	16.30	81.84	79.35	69.69	58.53	34.79	61.10	27.03	17.34	16.42	15.11
NTGA [211]	42.46	69.49	66.83	64.28	60.19	53.24	62.58	53.48	47.30	44.39	43.29
EM [76]	21.05	81.50	70.48	54.22	42.23	21.98	36.46	24.99	22.57	21.54	20.60
REM [57]	25.44	82.28	77.73	71.19	63.39	37.89	40.77	28.81	28.39	25.38	26.49
SG [193]	33.05	79.49	77.15	74.49	73.03	70.76	69.32	58.03	47.33	31.67	31.56
HYPO [187]	71.54	85.45	89.14	90.16	88.10	70.66	83.17	80.33	76.91	73.22	72.05
TAP [52]	8.17	83.87	84.82	77.98	57.45	11.97	45.99	18.29	14.16	8.590	7.38
SEP [26]	3.85	84.37	87.57	82.25	59.09	8.06	43.48	10.01	7.89	4.99	3.66
LSP [210]	15.09	78.69	42.11	33.99	29.19	26.66	48.27	29.56	25.14	16.88	14.27
AR [163]	13.28	85.15	89.17	86.11	80.01	54.41	31.54	12.64	11.66	9.96	12.99
OPS [203]	36.55	82.53	79.01	68.58	59.81	53.02	53.76	48.46	46.79	38.44	42.27

Table 3.14: Clean test accuracy (%) of target models under EM poisoning and its adaptive variants on CIFAR-10. Results are reported for $L_\infty = 16$.

Poisons	w/o	Gray	JPEG	G&J	Ave.
EM	19.32	80.60	84.32	82.12	66.59
EM-Gray	10.01	12.14	50.14	52.07	31.09
EM-JPEG	21.63	64.83	68.21	81.22	58.97
EM-G&J	19.71	22.68	28.94	30.51	25.46

3.6.2 Hyperparameters for Different Countermeasures

If not explicitly mentioned, we use JPEG with a quality factor of 10 and bit depth reduction (BDR) with 2 bits. For grayscale compression, we use the torchvision implementation where the weighted sum of three channels are first calculated and then copied to all three channels. For adversarial training (AT), PGD-10 is used with a step size of $\frac{2}{255}$, where the model is trained on CIFAR-10 for 100 epochs. We use a kernel size of 3 for both median, mean, and Gaussian smoothing (with a standard deviation of 0.1).

3.6.3 Color Channel Difference Mitigation Methods on EM

We show that grayscale compression is a special case where the weighted sum of different channels are used. Table 3.16 demonstrates that other approaches that reduce color channel differences can also be applied to counter poisons.

3.6.4 PAP Countermeasures in Facial Recognition

In the domain of facial recognition, [149] propose two countermeasures against two PAP methods, Fawkes [173] and LowKey [28]. Their first countermeasure is based on robust training via data augmentation and assumes that an additional clean pre-trained model is available to the data exploiter. In contrast, our work explores robust training, via adversarial training, but does not assume the exploiter has access to additional (clean) data and model. Table 3.15 demonstrates that models

`https://pytorch.org/vision/stable/generated/torchvision.transforms.Grayscale`

trained by their robust training on one type of poison would not generalize to others, limiting the effectiveness of the robust data augmentation against PAPs.

Their second countermeasure is more conceptual, which is to "wait for better facial recognition systems to be developed in the future." This method clearly depends on the potential progress of future models and obviously cannot act as an effective solution at this moment. In contrast, our ISS requires no change to the existing model but only applies pre-processing operations.

Table 3.15: Clean test accuracy (%) of CIFAR-10 when train and test on different poisons.

Train / Test	DC	NTGA	EM	REM	SG	TC	HYPO	TAP	SEP	LSP	AR	OPS
DC	**97.20**	18.76	17.24	19.39	19.01	18.05	18.39	17.86	18.02	14.09	17.43	17.79
NTGA	20.74	**97.85**	28.06	33.24	37.98	33.10	38.71	30.54	32.69	36.07	40.35	38.41
EM	14.49	15.65	**99.85**	20.31	19.57	15.79	15.08	14.52	14.06	12.58	16.99	16.22
REM	21.34	22.44	23.31	**99.97**	24.13	24.42	24.94	22.48	23.69	25.89	25.93	26.41
SG	31.99	36.63	31.53	26.89	**96.71**	33.41	35.87	29.55	30.49	39.59	39.43	36.43
TC	64.27	67.41	50.04	61.93	75.41	**93.79**	61.42	61.57	72.03	76.81	73.81	87.28
HYPO	63.58	67.51	71.55	67.50	71.63	21.59	**99.98**	1.59	0.56	70.60	73.69	72.31
TAP	9.15	10.68	10.81	10.61	11.28	18.96	9.10	**100.00**	9.16	10.61	12.48	11.65
SEP	4.50	5.28	4.65	5.11	4.86	5.53	6.45	4.58	**99.99**	5.06	4.99	5.50
LSP	16.35	18.33	25.47	22.07	17.79	16.95	21.08	20.03	19.36	**100.00**	16.35	15.31
AR	14.06	16.15	10.42	15.60	20.53	17.64	16.15	14.28	14.55	16.82	**99.94**	12.85
OPS	13.11	18.26	16.78	13.53	16.69	12.29	14.04	15.98	12.39	16.79	17.45	**99.83**

Table 3.16: Clean test accuracy (%) of ResNet-18 trained on EMs that are pre-processed by another three channel-wise color suppression methods. C-mean calculates the mean value and copies the mean to three channels. R-3/G-3/B-3 copies the values from the red/green/blue channel to three channels.

Methods	C-Mean	R-3	G-3	B-3	Gray
Acc.	91.83	86.60	86.73	87.91	93.01

Chapter 4

Privacy Improvement by Pivoted Profiles

Bag-based classification is a supervised machine learning method that makes a prediction based on a bag of items. Unfortunately, it can be misused as an attribute profiling attack, where the attacker's objective is to infer a privacy-sensitive attribute of a target user from that user's shared social media profile, i.e., a bag of images or other media. Despite this threat, existing studies on profiling attacks are limited to the item-level perspective, i.e., attack and defense of a single item. In this work, we move obfuscation defenses against attribute profiling beyond the existing single-item research to study the multi-item, bag-based case, which is more practically relevant because it considers the full attack surface. Defense against bag-based profiling is difficult, because, in general, content shared on social media can never be completely deleted. For this reason, we study defenses that involve extensions, referred to as pivoting additions*, to existing profiles, which aim to change (i.e., pivot) the output of the bag-based classifier without removing items contained in the original profile. We propose three different pivoting additions: Adversarial Noise (AdvN), Adversarially Perturbed Items (AdvPI), and Natural Items (NatI). We experimentally demonstrate the ability of these pivoting additions to compromise the performance of three deep bag-based classifiers, representing late-, intermediate- and early-fusion approaches. Overall, our work provides an introduction to the risk of bag-based profiling and a systematic study of defenses.*

This Chapter is under review as Resisting Bag-based Attribute Profiling by Adding Items to Existing Media Profiles. IEEE Transactions on Information Forensics and Security. Preliminary version was published at Conference on User Modeling, Adaptation and Personalization (UMAP), 2021.

4.1 Introduction

Machine learning classifiers can infer potentially privacy-sensitive information from images that users post online. Particularly concerning is work that infers attributes that would not be readily evident to a person casually inspecting the images, such as personality [170] and depression [155]. Recent research has addressed this risk using adversarial techniques. Building on the initial discovery of adversarial images [9; 63], these techniques take an privacy-by-obfuscation approach [128; 101; 173]. Specifically, they create versions of items that have been modified so that machine learning classifiers are fooled, i.e., the accuracy with which a classifier can infer a privacy-sensitive attribute is drastically reduced to near zero or to the point of being no better than a random decision. Adversarial examples have been studied in white-box [21], gray-box [194], and black-box [135] threat models, which differ in the amount of information that targets users, who are seeking to protect themselves, can access about the classifier being used to infer their privacy-sensitive attributes.

Given the importance of online privacy, it is surprising that until now all research on adversarial images for privacy protection has taken an *item-level perspective*. In other words, until now, research has focused on studying attacks and defenses of individual images posted by users. This focus has been present since early work, represented by e.g., [67; 202; 127; 29; 128; 101; 173; 172], in which individual images are modified and the ability of the change to block inference is measured, without considering multiple images posted by a single user.

In this work, we introduce the *profile-level perspective* and investigate how privacy-sensitive attributes can be inferred from a user's profile containing multiple media items and how adversarial techniques make it possible to resist this inference. We are specifically interested in attributes, like personality and emotion, that are not easily observable by a person manually inspecting the profile. We are also specifically interested in cases in which the evidence for the attribute is assumed to be spread out over multiple items in the profile and not concentrated in one or two. For example, we are not interested in attacks that seek to determine whether a user has ever met with a particular person by identifying a clear image of that user with that person in the user's profile.

Our goal of moving research beyond the item-level perspective to the profile-level perspective is important for two reasons. First, it is not particularly realistic to assume that the attack surface is limited to a single media item. Instead profilers, who aim to infer privacy-sensitive information of target users, can acquire multiple items, or an entire profile, creating a set that potentially contains more information than a single item alone. Profilers working from the inside of an online platform can acquire such a set by accumulating images or other media, including items a user has deleted. Profilers with no inside access to an online platform can do it by scraping posted content. User linking between platforms [139] and authorship attribution [183] can potentially expand the number of items available via scraping. Our concern is that an attacker can exploit information in the set of items using *bag-based attribute profiling* carried out with multiple instance learning (MIL). Bag-

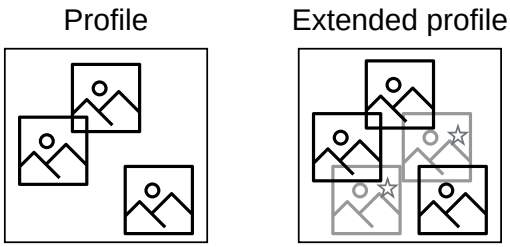


Figure 4.1: *Pivoting additions* are additional items added to an existing profile, which we consider an unordered set of items, or "bag". In this work, we demonstrate that such additions can resist bag-based attribute profiling carried out by a deep bag-based classifier, and deletion of existing items is not necessarily required.

based attribute profiling takes an unordered set of items (i.e., a bag of items) as input and returns a prediction of the value of a profile-level privacy-sensitive attribute.

Second, the item-level perspective assumes that the user protects all of their images or other media *before* uploading them. The dominant view in the literature, represented e.g., by [149], is that users who have already uploaded unprotected items are irreversibly exposed. We agree that it is necessary to assume that deleting items that have already been posted is not a reliable way to protect privacy, since these items may be retained indefinitely on servers or already have been scraped. However, in this work, we argue that it is time to drop the assumption that the existence of already-uploaded images is incompatible with obfuscation approaches to privacy protection. Instead, we demonstrate that the profile-based perspective opens up the possibility that users wishing to resist attribute profiling extend their currently existing profiles by posting additional items, as illustrated in Fig. 4.1. We refer to the added images as *pivoting additions* because they are extensions to the profile that are intending to shift or "pivot" the prediction of the profiling classifier, with the aim of changing a correct prediction into an incorrect one.

This work makes three main contributions:

- We introduce the profile-level perspective for studying the use of adversarial examples in online privacy, including a definition of the threat models in the white-box, gray-box, and near-black-box scenarios.

- We demonstrate that deep bag-based classifiers using early or intermediate fusion are potentially more dangerous, than approaches that use late fusion, i.e., predict at the item level before aggregating to reach a final prediction.

- We introduce three pivoting additions to resist bag-based profiling, which we study under different threat scenarios. Our experiments and analysis demonstrate that it is possible for users to resist bag-based attribute profiling by adding items to existing profiles.

The three pivoting additions that we propose gradually increase the level of stealth, i.e., how obvious it is that the user has pivoted the profile. The first approach, Adversarial Noise (AdvN) addition, adds items consisting solely of noise, which allows us to demonstrate the upper bound of the protection that can be achieved if stealth is not an issue. The second approach, Adversarially Perturbed Item (AdvPI), adds adversarial perturbed items, which is an extension of current single item-level perturbation approaches from adversarial machine learning [101; 128] to newer MIL classifiers. To our knowledge, AdvN and AdvPI represent the first time that optimization of the bag-level MIL model loss has been used to create adversarial profiles. The AdvPI pivot is less easily noticeable than AdvN, yet technically speaking, could still be distinguished. The third approach, Natural Item (NatI), adds adversarial items which are not perturbed but rather selected. This approach follows the trend of naturalness in the item-level adversarial image literature [196; 217]. Since the pivoting addition consists of entirely untouched images, only the posting user knows that a pivot has taken place. Further, natural image additions can be expected to remain more effective than modification-based additions if the attacker applies a typical approach to negate the effects of pertubation, such as compression [207].

This work builds on a previous short work [113] in which we introduced the problem of bag-based attribute profiling and presented a framework for defining the problem along with a single white box defense against late-fusion bag-based classification. Compared to our previous work, this work studies the problem systematically, investigating three different defenses in white-box, gray-box, and near-black-box scenarios against late-, intermediate- and early-fusion bag-based classification.

4.2 Related Work

In this section, we cover the relevant literature most closely related to our work. In Section 4.2.1, we discuss attribute profiling and the obfuscation defenses that have been proposed to resist it. As mentioned, the previous work has focused on the item-level perspective. Moving from item level to the profile level requires considering attacks by recently-developed deep bag-based Multiple Instance Learning approaches, which attack multiple media items at once. These are covered in Section 4.2.2.

4.2.1 Obfuscation Defenses against Attribute Profiling

Attribute profiling attacks use machine-learning classifiers to infer sensitive attributes of a user based on that user's content, especially publicly available social media posts. Many attributes can be considered privacy-sensitive, including personality traits, sexual orientation, geo-location, political views, and occupations [10; 29]. When users post content (i.e., text, images, video) on social media, they may be unaware that their private attributes can be inferred from their posts, which do not seem to be privacy sensitive [169]. Note that a profiling attack is different from a model inversion attack. In model inversion, the adversary has access to a machine learning model and is targeting users in the data that was used to

train that model [53; 54; 120]. In a profiling attack, the target users data is not assumed to be part of the training data.

To carry out a profiling attack, a profiler first trains a model on a data set of user profiles that have been annotated with privacy-sensitive attributes. Often, both the training data and ground truth attribute labels can be scraped from online sources. Then, using the trained model, the profiler can make predictions on target users for whom values of sensitive attributes are unknown. Numerous authors have demonstrated the threat of attribute profiling on the basis of social media data [69; 61; 86; 60; 62; 94; 22; 219]. However, these authors study profiles that consist of interactions and not of media items, that are of interest to us here.

In this work, we focus on obfuscation defenses. As such, our work contributes to research on obfuscation approaches to privacy [15]. We follow in the footsteps of authors who have applied adversarial machine learning to create image modifications that protect users against profiling attacks [88; 127; 128; 101; 173; 172; 201; 154; 202; 67; 29; 111; 199; 85; 180]. These have included work that has studied protection against social media analysis [154; 101], face recognition [127; 173; 201], person recognition [127], and object detection [111], where the main focus is to protect single item [88; 101] or interaction data used to train recommender systems [199; 85; 180]. We, instead, focus on profile-level privacy to protect against attribute inference attacks on profiles consisting of multiple media items.

Early work also explored different scenarios defined in terms of threat models specificying the amount of information that a target user can access about the attack that they are defending against. Studies start with a white box scenario, in which the target user has full information about the profiler, in order to demonstrate the potential of adversarial machine learning to offer protection [111]. Then, gray box [154] and black box [173] scenarios are studied to demonstrate the level of protection that can be provided in more realistic settings. Following previous work, we test our proposed pivoting in threat model scenarios that gradually reduce the assumptions on the knowledge that the target user has about the attacker.

4.2.2 Deep Bag-based Multiple Instance Learning

In the general case, multiple instance learning (MIL) designates approaches to a class of problems in which items occur in sets and the ground truth is labeled with a set-level label. Bag-based MIL is an approach that addresses a subset of the overall MIL problems, namely the case in which items occur in unordered sets in the ground truth, called bags, *and* the predicted output must also be at bag level. In this work, the attribute inference attacks we studied are carried out with bag-based MIL because we take user profiles to be unordered sets of items and because the attacker targets a single user-level attribute. Bag-based MIL can be characterized by the level at which the information from the individual items is combined to make the bag-level decision. Here, we discuss late fusion, and then early and intermediate fusion.

Late fusion first classifies the individual items in the profile, and then combines the item-level decisions into a bag-level decision. Common late fusion strategies include *voting* and *pooling* [17]. For voting, each item is labeled by a classifier, and the labels are combined with majority vote, while for pooling item-level scores are aggregated first and only then converted to a label for the bag. In this work, we choose Majority Vote (MV) [170; 113] as a representative deep learning late fusion method. The item-level predictions are first made by a CNN classifier and then the predictions are combined by way of majority vote. In practical scenarios, it is sometimes impossible to annotate all items in the training bags [27]. In this case, a bag-level annotation can be projected to the item level for training [169], which is the approach that we adopt here when training the MV classifier.

Early fusion combines representations at the item level, including raw inputs or unimodal features, into a unified representation [181]. Previous early fusion methods are mainly based on average or max pooling of representations [50; 17]. Recently, learning-based aggregation methods have become prevalent due to the flexibility of the learning-based aggregation module [50]. In this work, we chose a state-of-the-art method with average pooling, Deep Sets [212] as representative of early fusion. Deep Sets incorporates the aggregation functions into the learning.

Intermediate fusion. Like early fusion, intermediate fusion involves combining representations before the final classification model [11]. However, the representations are at a higher level of abstraction than with early fusion. Compared to early fusion, intermediate fusion provides an extra layer of flexibility. Only a limited number of intermediate fusion approaches exist for deep bag-based MIL. The dominant use of self-attention in deep bag-based MIL is Set Transformer (SetT) [104], which models the inter-relationships between item embeddings. We choose SetT here as representative of intermediate fusion. Interpretability in intermediate fusion approaches to deep bag-based MIL has been addressed by Attention-based deep Multiple Instance Learning (AttMIL) [78]. This approach trains an auxiliary network that provides interpretable weights that are used to calculate weighted averages in the aggregation layer. In this work, we carry out experiments with AttMIL because its interpretible weights allow us to gain insight in to the way that attention is distributed before and after the pivoting additions.

4.3 Threat Model

In this section we present the threat model, i.e., the characterization of the profiler (attacker) and the target user (protector) that we adopt in this work. We provide a description and motivation for each of the three scenarios that we study, white box, gray box, and near-black box.

Specification of profiler. The objective of the profiler is to acquire the values of a specific private attribute y for each of a set of target users $\mathbf{U}_{target} = \{u_1, u_2, ..., u_R\}$. For each target user, the profiler has obtained possession (i.e., by theft or scraping) of a user profile u, which we consider to be a bag of N items $\{x_1, x_2, ..., x_N\}$.

The profiler has collected additional resources (i.e., again by theft or scraping) in order to train a bag-based classifier. The resources consist of a set of user profiles $\mathbf{U}_{train} = \{u_1, u_2, ..., u_M\}$ again each containing N items $\{x_1, x_2, ..., x_N\}$. For each user in $\mathbf{U}_{train}$, the profiler also has the corresponding privacy-sensitive attribute $\{y_1, y_2, ..., y_M\}$. Using this data, the profiler trains a deep bag-based classifier, $f_\theta : u \rightarrow y$, where θ represents parameters of the model f. As previously mentioned, for our late-fusion approach, which requires instance-level labels, we assume the profiler projects the profile-level attribute label y to the level of the individual item x, by assigning every x in the profile u the same label y.

The profiler attacks by using f_θ to infer the value of the privacy sensitive attribute for each target user u in $\mathbf{U}_{target}$. Note that technically it is possible for the profiler to leverage the unlabeled target users to improve f_θ, using a unsupervised approach. However, we leave this possibility to be studied in future work. In our experiments, we also consider *proactive profilers*, who are aware of the pivoting additions in order to study the effectiveness of profiler countermeasures (Section 4.7.5).

Specification of target user. The objective of the target user is to prevent the profiler from inferring the value of a specific privacy-sensitive attribute. The target user has some level of knowledge of the attack, making the threat model either white-box, gray-box or near-black-box, as described in “Threat model scenarios" just below.

The defense of the target user is a pivot-based obfuscation approach. We assume that the target users were unaware of the risks when they began sharing on social media, but subsequently realized the danger of attribute profiling after they had already posted a number of items. Users have no guarantee that items deleted from their profiles are no longer available to the profiler, since their data may already have been scraped. Also, they may be reluctant to delete since it might draw attention to specific items. For this reason, users protect their profiles by a *pivoting addition*, i.e., by adding new items, without modifying or deleting existing items. We consider the case that there is only a single pivot, i.e., the user does not alternate between the original posting behavior and posting images from the pivoting addition.

Threat model scenarios. We study three specific scenarios within the overall threat model: white box, gray box, and near-black box. These scenarios represent three versions of the threat model that are progressively more realistic in terms of what target users know about the attack that they are trying to defend themselves against. For each scenario, we assume that the user has knowledge of which attribute the profiler is trying to acquire.

White box: The target users have full access to the attack classifier that the profiler used for attribute profiling. Under these white box assumptions, defense is the easiest and attack is the hardest. We include this case in our experiments because it represents an upper bound on the success of a pivoting addition.

Gray box: The target users have access to some (i.e., a subset) of the data (i.e., the user profiles) that the profiler used to train the attack classifier. They also have knowledge of the architecture of the classifier and train their own classifier. This scenario can be considered to correspond to the case in which the target user makes an educated guess as to the profiles that the profiler has collected from training or does not have sufficient resources to obtain an exact version of the profiler's training data.

Near-black box: The target users do not have knowledge of the architecture of the attack classifier and do not have access to any of the profiles that the attacker used to train the classifier. However, the target users can download publicly-available pre-trained feature extractors that have been trained on other data. Because there is a limited number of large-scale data sets (e.g., ImageNet) suitable for training such feature extractors, it is plausible that the feature extractors used by the target user and the profiler are trained on the same data. In our work, we make this assumption. Because the protector and the attacker use a feature extractor trained on the same data, we refer to this scenario as near-black box rather than black box.

4.4 Bag-based Profiling

In this section, we introduce the data sets used in our experiments (Section 4.4.1) and describe how we implement the deep bag-based classifiers that we use to study attribute profiling attacks (Section 4.4.2). Note that these classifiers have not previously been used in the literature of attribute profiling attacks. For this reason, we first demonstrate their effectiveness (Section 4.4.3) before moving on to studying the use of pivoting additions for defense.

4.4.1 Data Sets

In our work, we make use of two data sets consisting of image profiles, Personality Profile and Emotion Profile, which are used for our main experiments. We also make use of one data set consisting of speech profiles, Speech Gender Profile, which is used for an additional experiment that demonstrates the applicability of pivoting extensions beyond visual data. Our profile data sets are composed using items from existing data sets, which is common practice of multiple instance learning (MIL) research [78; 104]. This practice is particularly important when studying privacy-sensitive attributes since there are a limited number of publicly available data sets annotated with privacy sensitive attributes that have been collected based on ethical standards. Also due to data availability limitations, the profiles that we use in our experiments are relatively short (10-20 items). Short profiles are not unrealistic since a large number of users do post a limited number of images online or profilers might also have access to only a limited number of items. Also, in the case of the speech content we investigate gender, which in many contexts is not a particularly privacy sensitive attribute. This decision is made because gender information is available and gender is often used as a surrogate for more sensitive attributes in privacy research, as in e.g., [180]. Examples items from the profile data sets are

shown in Figure 4.2 and the data set properties are summarized in Table 4.1. In this section, we describe each data set in turn.

Personality Profile. The Personality Profile data set contains profiles annotated with five personalities that were composed of images from the PsychoFlickr data set [170]. The PsychoFlickr data set contains images from 300 Flickr users who agreed that their images could be used for research and were asked to fill the self-assessment version of the BFI-10. Our Personality Profile data set consists of five subsets, one for each personality trait in the PsychoFlickr data set, namely each of the Big Five OCEAN personality traits: `Openness` (O), `Conscientiousness` (C), `Extraversion` (E), `Agreeableness` (A), and `Neuroticism` (N) [153]. First, for each personality trait we isolate the users in the data who self-reported the strongest and weakest association with that trait (i.e., the first quartile and the fourth quartile). This leaves us with approximately 75 positive and 75 negative users for this personality traits, each with 200 images. The user numbers are approximate due to ties in user scores at the quartile boundaries. Then, we break each the images for each user into 10 smaller, disjoint, user profiles each containing 20 images. The result is one data subset for each profile trait that contains about 750 profiles positive for that trait and about 750 profiles negative for that trait.

Emotion Profile. The Emotion Profile data set contains image profiles annotated with five emotion labels that were composed from two data sets, Emotion6 [138] and the data set of [209]. The Emotion6 [138] data set contains images annotated with six emotion classes and is annotated via human intelligence task on Amazon Mechanical Turk. The data set of [209] consist of continuous images from eight emotion classes collected from Flickr and Instagram by keywords matching. After collection, the emotion labels are verified via human intelligence task on Amazon Mechanical Turk.

Our Emotion Profile data set consists of five subsets, one for each emotion, namely: `anger`, `disgust`, `fear`, `joy`, and `sadness`. Specially, the subsets were created as follows. Because there are six emotion categories from [138] and eight from [209] it was necessary to merge and discard categories. Both sets have `anger`, `disgust`, `fear`, and `sadness`. Then, we conflate `joy` from [138] and `amusement` from[209] to `joy`. We discarded `surprise` from [138] and `excitement`, `contentment` and `awe` from [209].

The subsets were composed by first selecting 1380 images for each emotion, where 1380 is the number of images in the emotion class with the least amount of images, and we aim to make the number of images from all emotion classes balanced. Then, we break images from each emotion into 138 smaller, disjoint user profiles, each containing 10 images. The final Emotion Profile subset for each emotion contains 138 profiles from the target emotion and 138 profiles randomly sampled from other emotions.

As usual with emotion data sets, images are associated with emotion classes and

Table 4.1: Overview of data sets used in our experiments. The Personality Profile data and the Emotion Profile data each consist of five data subsets, one for each profile class.

	# PROFILES	# ITEMS PER PROFILE	ITEM TYPE	PRIVATE ATTRIBUTE
PERSONALITY PROFILE (O)	1560	20	IMAGE	PERSONALITY TRAIT
PERSONALITY PROFILE (C)	1570	20	IMAGE	PERSONALITY TRAIT
PERSONALITY PROFILE (E)	1540	20	IMAGE	PERSONALITY TRAIT
PERSONALITY PROFILE (A)	1570	20	IMAGE	PERSONALITY TRAIT
PERSONALITY PROFILE (N)	1570	20	IMAGE	PERSONALITY TRAIT
EMOTION PROFILE (ANGER)	276	10	IMAGE	EMOTION
EMOTION PROFILE (DISGUST)	276	10	IMAGE	EMOTION
EMOTION PROFILE (FEAR)	276	10	IMAGE	EMOTION
EMOTION PROFILE (JOY)	276	10	IMAGE	EMOTION
EMOTION PROFILE (SADNESS)	276	10	IMAGE	EMOTION
SPEECH GENDER PROFILE	80	20	SPOKEN AUDIO	BINARY GENDER

not users. In contrast to the Personality Profile data, in the Emotion Profile data the images in the profiles that we define were not originally all posted by the same user.

Speech Gender Profile. The Speech gender profile data set used for our study contains short speech recordings and the privacy sensitive attribute is gender. Our *Speech Profile* data set is constructed from the test development and sets of the widely-used LibriSpeech [133] data set. For each speaker in the test and development set, 2 seconds of speech consisting of 20 recordings are included in their profiles as items. The Speech Gender Profile data set includes 80 distinct speakers. We use this data set to provide a demonstration that our pivoting additions are able to protect profiles containing other media beyond speech in Section 4.7.2, which supplements our main experiments on the image data.

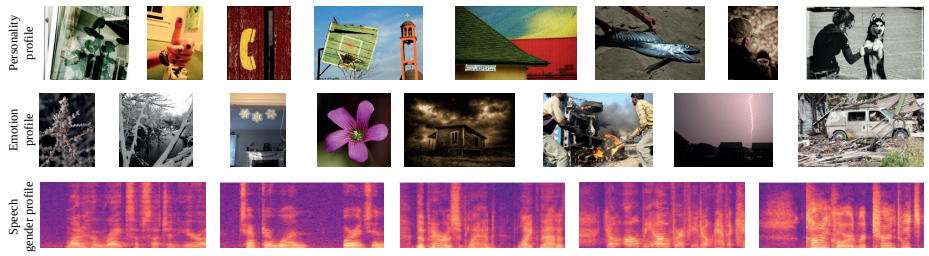


Figure 4.2: Media item examples from Personality Profile, Emotion Profile, and Speech Gender Profile data sets. Personality Profile and Emotion Profile consist of images, while Speech Gender Profile consists of spoken audio clips.

Table 4.2: Deep-bag based classifiers used for attribute profiling attack. The classifiers are chosen to be representative of the different fusion stages. AttMIL is not directly comparable since it does not leverage ILSVRC2012 data [36].

Approach	Method	Fusion stage	Additional Resources
MV	majority vote	Late	Pre-trained model
Deep Sets [212]	Average pooling	Early	Pre-trained features
SetT [104]	self-attention	Intermediate	Pre-trained features
AttMIL [78]	Weighted average pooling	intermediate (Interpretable)	none

4.4.2 Implementation of Bag-based Profiling Models

We carry out experiments using four representative deep bag-based classifiers for profiling, which are summarized in Table 4.2. The models were chosen so that they cover the different categories of approach (late, early, and intermediate fusion). Our main focus is on MV, Deep Sets, and SetT, for which we leverage pre-training on the ILSVRC20212 [36]. We also study AttMIL in order gain insight into how the attention of the model is shifted to the pivoting addition, i.e., to the images that are added to extend the profile and resist profiling attack. AttMIL is separated by a dashed line in Table 4.2 because the different in training resources means that its performance is not directly comparable to the others. We first study the effectiveness of the approaches in carrying out an attribute profiling attack (Section 4.4.3). For each classification task, we randomly select 80% profiles of each data set for training, 10% for validation, and 10% for testing. Note that we use accuracy as the evaluation metric since binary classes are well-balanced in all profile data sets. Then, we test our pivoting additions as defense against Deep Sets, SetT and AttMIL (Section 4.6), but not MV since our experiments show it is not an effective attack on image profiles.

Majority Vote (MV) is a late-fusion method also studied in [113]. For the image experiments, we fine-tune ResNet-50 [70] on the Personality Profile or the Emotion Profile data set, and for the speech experiment, we train a vanilla X-Vector model [182] on the Speech Gender Profile data set. Recall that bag-level ground truth is projected to item level for this purpose. When testing the profiling attack, the deep bag-based classifier makes predictions on one profile by majority vote over the predictions for all items in this profile.

Deep Sets [212] is an early-fusion method that includes a fully connected encoder and decoder to aggregate features of the items in a bag. Deep Sets aggregates the individual transformed representations into a single set representation using a permutation-invariant function that is either mean or max pooling similar to traditional methods. The feature transformation network is trained to aid the aggregation. Formally, Deep Sets can be represented in the following:

$$f(\{x_1, \ldots, x_n\}) = \rho(\text{mean}(\{\phi(x_1), \ldots, \phi(x_n)\})). \tag{4.1}$$

where ϕ represents the transformation network that outputs item representation before aggregation and ρ represents the bag-level aggregation. The aggregation function, here, mean, is considered in the training process, which helps Deep Sets outperform two-stage approaches in which feature extraction and aggregation are done in separate stages.

In our implementation, as the input of the encoder, image features are extracted from a pre-trained VGG-16 [178] feature extractor pre-trained on ILSVRC2012 [36], and speech features are extracted from a pre-trained speech transformer WaveLM [25].

SetT [104] is an intermediate fusion method that incorporates several modules for self-attention. SetT treats each item as an individual element and utilizes self-attention across all items in a bag. The SetT approach facilitates the modeling of the relationships between different items, highlighting each element's significance by considering its relation to the other elements within the set. Formally, SetT is represented as,

$$f(\{x_1, \ldots, x_n\}) = \rho(\text{attention}(\{\phi(x_1), \ldots, \phi(x_n)\})) \tag{4.2}$$

where attention represents self-attention layer and ϕ represents fully-connected layers that transforms input before attention, and ρ represents the bag-level aggregation Self-Attention is calculated on extracted features and then fed to fully connected layers to calculate logits. Like Deep Sets, in our implementation, image features are extracted from a pre-trained VGG-16 [178] feature extractor and speech features are extracted from a pre-trained transformer WaveLM [25].

AttMIL [78] is an interpretable intermediate fusion method that includes two neural networks, where one model extracts features from items and the other assigns aggregation weights to all items in the bag. Formally, AttMIL can be formulated as,

$$f(\{x_1, \ldots, x_n\}) = \rho(\sum(\{a_1\phi(x_1), \ldots, a_n\phi(x_n)\})) \tag{4.3}$$

$$a_i = \frac{\exp\{\mathbf{w}^\top \tanh\left(\mathbf{V}\phi(x_i)^\top\right)\}}{\sum_{j=1}^{N} \exp\{\mathbf{w}^\top \tanh\left(\mathbf{V}\phi(x_j)^\top\right)\}}, \tag{4.4}$$

where $\mathbf{w}$ and $\mathbf{V}$ are trainable parameters, and a_i are weights calculated for the ith

Table 4.3: Bag-based classification accuracy on our Personality Profiles data set. (20 images per profile.)

Methods	O	C	E	A	N	Avg.
MV	57.69	67.52	55.19	58.60	57.96	59.39
Deep Sets	84.62	67.52	75.97	61.78	76.43	73.26
SetT	69.87	77.07	66.88	61.78	68.79	68.89
AttMIL	58.33	67.52	51.30	69.43	71.97	63.71

Table 4.4: Bag-based classification accuracy on our Emotion Profiles data set. (10 images per profile.)

Methods	Anger	Disgust	Fear	Joy	Sadness	Avg.
MV	53.57	60.71	50.00	85.71	50.00	59.28
Deep Sets	100.00	100.00	78.57	100.00	100.00	95.71
SetT	96.43	100.00	78.57	100.00	96.43	94.29
AttMIL	57.14	89.29	64.29	82.14	60.71	70.71

item.

The auxiliary attention module, determined by $\mathbf{w}$ and $\mathbf{V}$, provides normalized weights for all items for one input bag, whereby a larger weight indicates more contribution of an item. In contrast to SetT, AttMIL must be trained in an end-to-end fashion. Because AttMIL does not leverage a pre-trained model, the amount of training data is critical. In this work, we carry out experiments with AttMIL because its interpretible weights allow us to gain insight into the way that attention is distributed before and after the pivoting additions.

Implementation details. We train all deep bag-based classifiers on one NVIDIA RTX 3090 GPU. For MV, we train ResNet-50 [70] for 30 epochs on Personality Profile and Emotion Profile and X-Vector for 20 epochs on Speech Gender Profile. For Deep Sets, both the encoder and decoder have four fully connected layers, and for SetT, we employ a multi-head transformer. We use the VGG-16 [178] pre-trained on ILSVRC2012 as the feature extractor for both Deep Sets and SetT. In optimization, SGD is implemented with a learning rate of 0.1, a momentum of 0.9, and cosine weight decay.

4.4.3 Performance of Bag-based Profiling

Table 4.3 and Table 4.4 show the classification accuracy for the Personality Profile data and the Emotion Profile data. We observe that early (Deep Sets) and intermediate fusion (SetT) outperform late fusion (i.e., majority vote), although all three approaches leverage pre-trained features. Further, Deep Sets substantially outperforms SetT. Our interpretable intermediate fusion approach (AttMIL) outperforms late fusion (majority vote).

Overall, these experiments demonstrate that majority vote is outperformed by the

newer deep bag-based classifiers. For this reason, in the rest of the work, we set majority vote aside and study the other classifiers. Also, we analyze AttMIL separately from Deep Sets an SetT because it differs in the amount of training data that it uses, i.e., it does not use pre-trained features, and its performance is not directly comparable. Were more training data available, AttMIL could possibly be improved. However, we take the performance level achieved by AttMIL to be adequate for our purpose of using AttMIL to study the affect of the pivoting addition on the distribution of attention (Section 4.7.1).

4.5 Pivoting Additions

In this section, we introduce three types of pivoting additions, Adversarial Noise (AdvN), Adversarially Perturbed Items (AdvPI), and Natural Items (NatI), and describe how we test them under the white-box, gray-box, and near-black-box scenarios, which were defined in Section 4.3. An illustration of the three pivoting additions is provided in Figure 4.3. The pivoting additions change the inference of the classifier. In white, gray, and near-black-box scenarios, our main experiments are with pivoting extensions that double the profile length, i.e., we add 20 pivoting items in profiles on the Personality Profile and 10 pivoting items in the Emotion Profile data sets. We provide an analysis on the impact of varying the number of pivoting items in Section 4.7.4.

4.5.1 Adversarial Machine Learning-based Pivoting Additions

Adversarial Noise (AdvN) addition leverages optimization methods to generate items consisting of adversarial noise to add to the existing profile. Noise additions are generated by maximizing profile-level model loss leveraging adversarial machine learning techniques [87]. As mentioned, to our knowledge, we are the first to apply these techniques to bag-based MIL classifiers.

White-box and gray-box scenarios: Target users generate adversarial noise items by increasing the surrogate model loss with respect to the pivoting addition in an iterative manner, considering the existing items in the user profile. Specifically, the loss is calculated on the profile consisting of original and randomly generated images, and the randomly generated items are updated to generate adversarial examples. Existing items in the profile are not touched. Formally, AdvN can be formulated as,

$$u_*^{t+1} = \Pi(u_*^t + \alpha \ \text{sign}(\nabla_{u_*} L(\theta, u_*, y))) \tag{4.5}$$

where L is the profile-level model loss, u_* is the pivoting addition, Π is the projection function, and t is the number of iterations. Also, α is the step size, and θ represents the model parameters.

We use Projected Gradient Descent (PGD) [118] to create pivoting items because

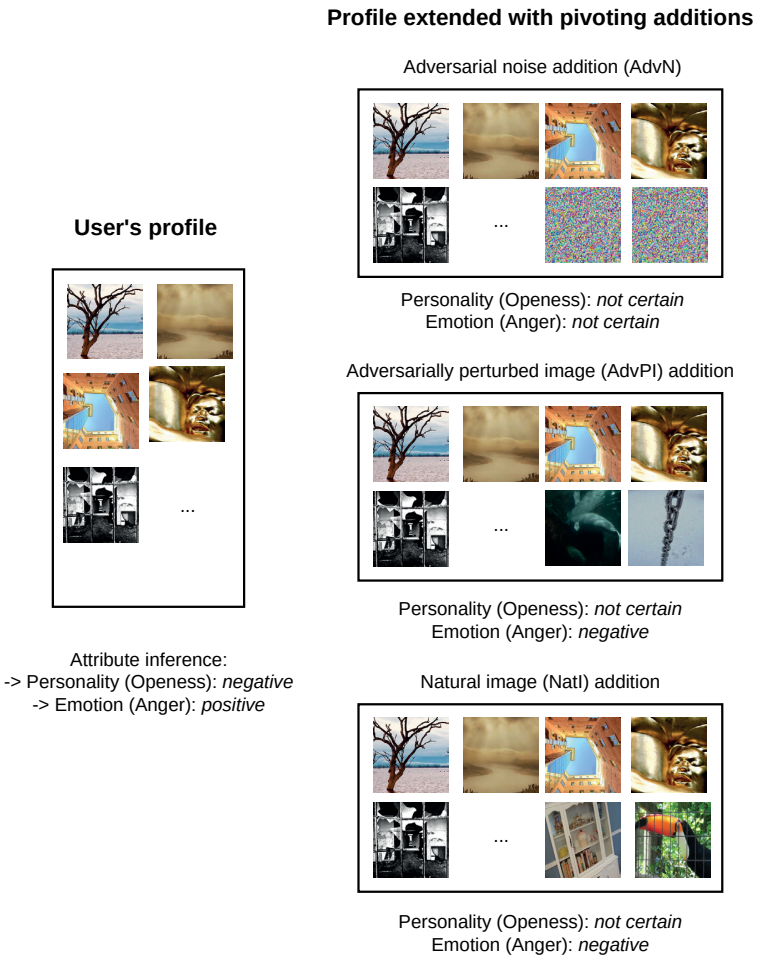


Figure 4.3: Illustration of the three proposed pivoting additions. The original user profile contains several images and is classified as negative for Openness personality and positive for Anger emotion. The predictions are changed AdvN and AdvPI use adversarial techniques to *create* or *modify* the images in the pivoting addition and NatI exploits the model loss to *select* the additional images from a set of unmodified images.

PGD is an iterative approach to generate adversarial noise, which has been shown to be effective in different applications [191]. Several hyper-parameters need to be specified in PGD, including the step size, the number of iterations, and L_p norm restrictions on perturbations. Previous research shows that a sufficient number of iterations is important when testing PGD [216], so we use PGD with 100 iterations in all experiments. In the white-box scenario, users have access to the attack classifier of the attacker, while in the gray-box scenario, users have access to only 50% of the training data and the architecture of the attack classifier.

Near-black-box scenario: Target users use a pre-trained feature extractor as a surrogate feature extractor. Features are extracted for all profile items, and then pivoting additions are generated by minimizing the cosine similarity between the existing profile and the items in the addition. Our approach is inspired by feature-level adversarial perturbations that modify the sample-wise similarity [160; 112]. Specifically, all pivoting addition items are optimized together to minimize the cosine similarity between the profile and the pivoting additions. Again, existing items in the profile are left untouched. The optimization can be formulated as,

$$u_* = \underset{u_*}{\operatorname{argmin}} \ \text{CosineSimilarity}(z_\theta(u), z_\theta(u_*)) \tag{4.6}$$

where z_θ is the feature extraction model, u represents original items, and u_* represents pivoting additions.

In our experiment, target users use ResNet-50 [70] pre-trained on ILSVRC2012 as the surrogate feature extractor to generate pivoting images and a Wave2Vec [70] pre-trained on 960 hours of LibriSpeech data to generate pivoting audios. Note that the profiling classifiers, Deep Sets and Set Transformer, also use publicly available feature extractors that are trained on ILSVRC2012 data.

Adversarially Perturbed Image (AdvPI) addition follows the same optimization as AdvN. The only difference is that in AdvPI, u_* in Equation 4.5 and 4.6 is not noise but rather ordinary items that are selected randomly (i.e., images or speech clips). As can been seen in Figure 4.3, AdvPI samples are less perceptible than AdvN to human observers because the additional items have recognizable content.

4.5.2 Natural Image Additions

Natural Image (NatI) additions do not create or modify the content of pivoting addition items. Instead, NatI selects items from a collection of natural (i.e., ordinary and unmodified) items that the user might potentially wish to post. In our experiments, we use the validation set of ILSVRC2012 [36] as the background collection set to pick natural images for the image pivoting additions. We use the LibriSpeech "test-other" data set [133] as the background collection set to pick natural audio clips for speech pivoting additions. Recall, that the three pivoting additions progress in stealth from AdvN to AdvPI to NatI, with NatI being the most stealth, i.e., the pivoting addition is not visible to the human eye and not detectable by a classifier trained to identify image modifications.

White-box and gray-box scenarios: Target users make use of the same surrogate model as used with AdvN and AdvPI for the white-box and gray-box scenarios. For NatI, items from the background collection are first fed to the surrogate model and items with the highest or lowest predicted confidence are selected to create positive and negative sets of candidate items. Note that although the model is a

Table 4.5: Average profiling accuracy (%) on the image profile data sets after pivoting under the **white-box** scenario. The difference from the original average accuracy (absolute %) is shown in parentheses.

Data set	Pivots	Deep Sets	SetT	AttMIL
Personality Profile	AdvN	0.00 (-73.26)	0.00 (-68.89)	7.81 (-55.90)
	AdvPI	0.00 (-73.26)	0.00 (-68.89)	6.14 (-58.57)
	NatI	48.02 (-25.24)	31.44 (-37.45)	20.61 (-44.10)
Emotion Profile	AdvN	0.00 (-95.71)	0.00 (-94.29)	25.00 (-45.71)
	AdvPI	0.00 (-95.71)	0.00 (-94.29)	19.29 (-51.42)
	NatI	42.86 (52.85)	34.29 (60.00)	31.86 (-38.85)

bag-based classifier, it is capable of classifying individual items (since they are a bag of one). Given a profile and its label, the NatI pivoting addition is created by randomly selecting items from the candidate set with the opposite label.

Near-black-box scenario: Target users make use of the same pre-trained feature extractor as they use with AdvN and AdvPI in the black-box scenario, i.e., ResNet-50 [70] for images and a pre-trained Wave2Vec [70] model for speech. The feature extractor calculates features for all items in the target profile and all items in the background collection, based on which natural items are selected. Specifically, given one target profile, the cosine similarity between each natural item in the background collection and the target profile set is calculated. Natural items that have the lowest cosine similarity scores are selected as pivoting additions.

4.6 Main Experimental Results

This section contains the main experimental results and analysis of our pivoting additions. We report on experiments with image profiles, i.e., Personality Profiles and Emotion Profiles. Section 4.6.1 covers the white- and gray-box scenarios and Section 4.6.2 covers the most realistic near-black-box scenario.

4.6.1 White-box and Gray-box Pivoting Additions

Table 4.5 and Table 4.6 present white-box and gray-box results for the Personality Profile data set and the Emotion Profile data set. Looking first at Deep Sets and SetT, AdvN and AdvPI are both effective because they are able to drop the accuracy of the classifier substantially, in most cases to near random performance (i.e., 50) or below. In the white-box scenario, AdvN and AdvPI did not decrease the accuracy of AttMIL to 0, because gradients vanish for profiles with very high prediction confidence. In theory, the white-box can be improved by adopting adjusted gradients [3], but we use the same PGD implementations for all bag-based deep classifiers for fair comparisons. In the gray-box scenario, as expected, AdvN, which adds optimized noise, is more effective than AdvPI, which adds perturbed images. In the item-level adversarial example literature, it is typical that perturbative approaches are effective at fooling a classifier [101]. However, we observe that for profile-level AttMIL, a priori, it is not obvious that adding adversarial pertur-

Table 4.6: Average profiling accuracy (%) on the image profile data sets after pivoting under the **gray-box** scenario. The difference from the original average accuracy (absolute %) is shown in parentheses.

DATA SET	PIVOTS	DEEP SETS	SETT	ATTMIL
PERSONALITY PROFILE	ADVN	12.21 (-61.05)	31.17 (-37.72)	41.56 (-22.15)
	ADVPI	51.95 (-21.31)	58.70 (-10.19)	63.45 (-0.26)
	NATI	48.40(24.86)	35.16(34.73)	21.10 (-42.61)
EMOTION PROFILE	ADVN	7.86 (-87.85)	8.57 (-85.72)	27.14 (-43.57)
	ADVPI	33.57 (-62.14)	40.71 (-53.58)	32.86 (-37.85)
	NATI	46.43 (49.28)	46.43 (49.28)	48.57 (-22.14)

Table 4.7: Average profiling accuracy (%) on the image profile data sets after pivoting under the **near-black-box** scenario. The difference from the original average accuracy (absolute %) is shown in parentheses.

DATA SET	PIVOTS	DEEP SETS	SETT	ATTMIL
PERSONALITY PROFILE	ADVN	55.33 (-17.93)	58.67 (-10.22)	36.80 (-26.91)
	ADVPI	60.33 (-12.93)	67.33 (-1.56)	39.40 (-24.31)
	NATI	54.00 (-19.26)	53.80 (-15.09)	43.20 (-20.51)
EMOTION PROFILE	ADVN	72.14 (-23.57)	83.57 (-10.72)	42.20 (-28.51)
	ADVPI	65.71 (-30.00)	74.00 (-20.29)	49.29 (-21.42)
	NATI	62.14 (-33.57)	74.29 (-20.00)	39.28 (-31.43)

bations will help, since we do not replace the original profile images, but only add images. Recall, we also restrict our addition to a doubling of the original profile length. NatI is effective, given its stealth. We see all three pivoting additions resist AttMIL attacks.

Next, we compare the white-box and gray-box scenarios. Recall that the architecture of the gray-box surrogate attack model is the same as the white-box surrogate model, except that it is trained on 50% of the original data (randomly selected). Given that the protector has radically less of the attacker's data and profiling classifier, gray-box protection is effective. Note that it is to be expected that the defense has a certain level of sensitivity to the exact nature of the surrogate, e.g., training data sets. For NatI, we see protection is generally less effective in the gray-box case, but still able to drop the performance level to around 50.

4.6.2 Near-Black-box Pivoting Addition

Table 4.7 contains the near-black-box results for the Personality Profile Emotion Profile data sets. These results demonstrate that our pivoting additions are effective when users have no exact knowledge of the deep bag-based classifier used for attack. All methods are effective and can shift the predictions of the bag-based classifier substantially, and often below random. Interestingly, NatI outperforms AdvN and AdvPI in decreasing the profiling accuracy. For AdvN and AdvPI, pivoting additions are transferable, which accords with the research on the transferability of adversarial examples [218]. Also, note that the NatI pivoting extensions are

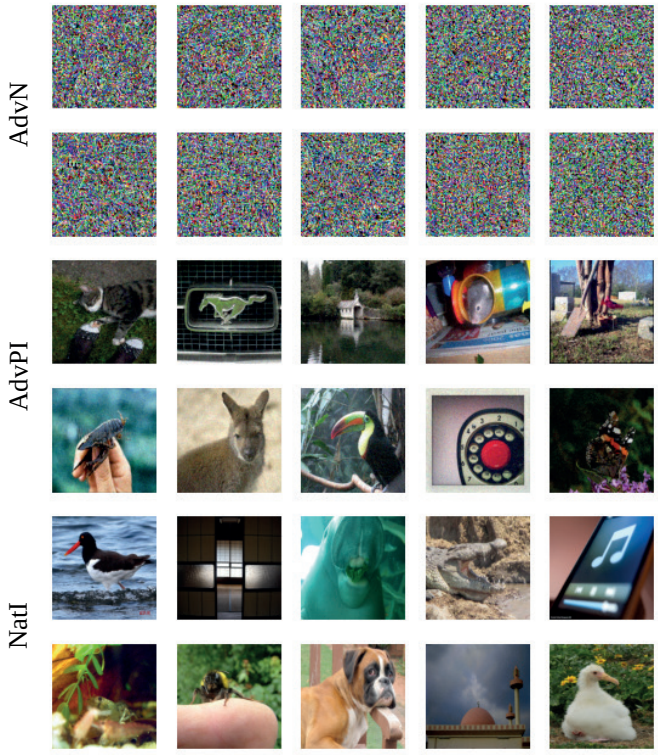


Figure 4.4: Examples of pivoting additions on Emotion Profile data set. The original emotion prediction on the profile is "anger" by a binary bag-based classifier. Pivoting additions are added to decrease the prediction confidence.

the same for all attacks, and for Speech Profile data set, NatI selects audios from same-gender speakers. Note that NatI additions are effective across different models making NatI the most generic of the three pivoting additions.

In order to provide a concrete impression of the pivoting additions that our approaches generate, we provide examples in Figure 4.4 and Figure 4.5. These examples illustrate that stealth increases from AdvN to AdvPI to NatI, i.e., from pure noise to natural images. Figure 4.4 shows some Emotion Profile images that pivot the prediction of Deep Sets away from the correct class `anger`. Here we have the impression that NatI is selecting wildlife and pet images. However, in general it is difficult to give an interpretation about the kinds of images that will pivot the classifier. This point is important because it means that visual inspection does not lead to the impression that a profile has been pivoted. Figure 4.5 shows some Personality Profile images that pivot the prediction away from `Openness (O)` in the near-black-box scenario. Here there is no clear pattern. In fact, the presence of wildlife and pets might simply reflect the images available in the background collection.

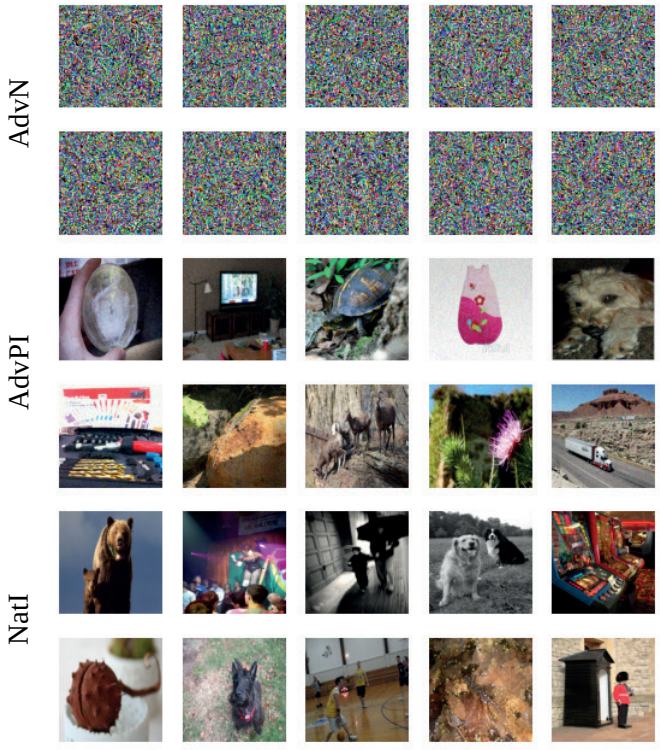


Figure 4.5: Examples of pivoting additions on Personality Profile data set. The original personality trait prediction is "openness" by a binary bag-based classifier. Pivoting additions are added to decrease the prediction confidence.

Table 4.8: Profiling accuracy (%) on Speech Gender Profile data set after pivoting under the near-black box scenario. The test set of the Speech Gender Profile consists of 8 profiles. The difference from the original accuracy (absolute %) is shown in parentheses.

PIVOTS	DEEP SETS	SETT	ATTMIL
ADVN	50.00 (-50.00)	50.00 (-50.00)	37.50 (-62.50)
ADVPI	62.50 (-37.50)	50.00 (-50.00)	37.50 (-62.50)
NATI	37.50 (-62.50)	37.50 (-62.50)	62.50 (-37.50)

4.7 Additional Results and Analysis

This section dives deeper into the analysis of adversarial additions. First, in Section 4.7.1, we take a closer look at the impact of the adversarial additions by analyzing AttMIL, our interpretable intermediate fusion approach. Then, in Section 4.7.2, we provide a demonstration that our pivoting additions can be used on other media beyond images, namely, speech data. Finally, in Sections 4.7.3 and 4.7.4 return to

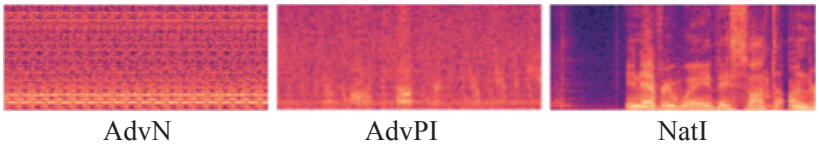


Figure 4.6: Examples of pivoting additions on Speech Gender Profile data set. Pivoting additions are added to the profile to decrease the confidence of original gender predictions.

Table 4.9: Average accumulated attention weights (%) of near-black-box pivoting addition items on Emotion Profile data. Accumulated attention is calculated as the averaged summation of assigned weights for all pivoting items across all profiles. The difference from the original average accuracy (absolute %) is shown in parentheses.

Pivots	Profiling accuracy(%)	Accumulated attention (%)
Random Noise	61.20 (-9.51)	27.88
AdvN	42.20 (-28.51)	58.78
AdvPI	49.29 (-21.42)	58.65
NatI	39.28 (-31.43)	43.01

images to study the robustness of pivoting additions in two practical edge cases and Section 4.7.5 examines proactive profilers, who have knowledge of the defense.

4.7.1 Attention Analysis on AttMIL

In this section, we use AttMIL, which assigns aggregation weights to all items in the bag (See Section 4.4.2), in order to gain insight into the impact of the pivoting addition on the attack classifier. We carry out experiments on the Emotion Profile data set and also focus on the near-black-box scenario, since it is the most realistic. Table 4.9 demonstrates that pivoting additions alter the distribution of attention of items, with the items in profile extension receiving proportionally more attention than the items in the original profile. Interestingly, NatI, which provides the strongest protection demonstrates the least pull of attention towards the images in the extension. This property of NatI could be advantageous because it makes it more difficult for the attacker to examine the attention weights to conclude that a pivot has taken place.

4.7.2 Applicability on Speech-based Gender Classification

In this section, we turn to the speech data. Results on the Speech Gender Profile data set are shown in Table 4.8. Spectrograms of pivoting speech items are shown in Figure 4.6. For NatI, we observe that audio from a speaker of the same gender can also be used to pivot the prediction of deep bag-based gender classifier, indicating that the pivoting addition has the potential to be made non-suspicious to human listeners.

Figure 4.7: Examples "strong items" in the Personality Profile data set for the *Openess* class.

Table 4.10: Profiling accuracy (%) for user profiles with strong items (i.e., items with high prediction confidence) on Personality Profile and Emotion Profile data sets. The difference from the original average accuracy (absolute %) is shown in parentheses.

DATA SET	PIVOTS	DEEP SETS	SETT	ATTMIL
PERSONALITY PROFILE	ADVN	83.20 (-16.80)	76.40 (-23.60)	92.8 0(-4.00)
	ADVPI	85.60 (-13.40)	75.20 (-24.80)	96.80 (-0.00)
	NATI	71.60 (-28.40)	66.80 (-33.20)	67.60 (-32.40)
EMOTION PROFILE	ADVN	96.00 (-4.00)	98.00 (-2.00)	68.00 (-10.00)
	ADVPI	100.00 (-0.00)	100.00 (-0.00)	68.00 (-10.00)
	NATI	80.00 (-20.00)	90.00 (-10.00)	68.00 (-10.00)

4.7.3 Profiles with "Strong" Items

In the profiling attack, deep bag-based classifiers predict attributes of a profile by considering all items in the profile. The influence of pivoting additions on the deep bag-based classifier is not uniform across all different profiles. We conduct a case study where users' profiles contain "strong items" predicted as positive by a classifier with high prediction confidence. When a profile contains "strong items", we call such profiles "strong profiles" and hypothesize that the pivoting additions will be less effective. We measure the confidence by cross-entropy loss to select strong profiles. In particular, we calculate the profile-level loss for all profiles and only implement the near-black-box pivoting additions on "strong" profiles that have lower losses. For both Personality Profile and Emotion Profile data sets, profiles with the top 30% high confidence items are defined as "strong" profiles. Table 4.10 demonstrates that "strong profiles" are less sensitive to pivoting additions, but are still influenced to a degree. We see that NatI is more effective than AdvN and AdvPI for "strong profiles".

Based on the most accurate model, Deep Sets, on the Personality Profile data set, we show in Figure 4.7 examples of images from "strong profiles" for Openness prediction. The "strongest profile" includes more outdoor images with an open environment.

4.7.4 Number of Addition Samples in NatI

We further look at the influence of different numbers of additions in NatI on the deep bag-based classifier. Each profile has 10 images in the Emotion Profile data set. Figure 4.8 demonstrates that the effectiveness of NatI is correlated with the number of added images. In particular, black box NatI is more effective than the other

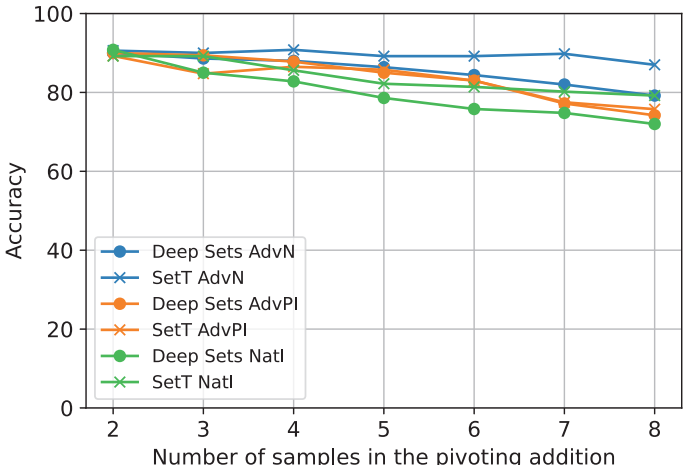


Figure 4.8: Averaged profiling accuracy (%) on five Emotion Profile data sets when different numbers of pivoting items are added to profiles. The effectiveness of pivoting additions is correlated with the number of added items.

two approaches for both models. We can also observe that SetT is less sensitive to pivoting additions than Deep Sets. The effectiveness of the deep bag-based classifier is negatively correlated with the number of pivoting items.

4.7.5 Proactive Profilers

In this section, we extend the threat model specification to include the assumption that the profiler is proactive, i.e., has knowledge of the pivoting additions.

Back-off as a countermeasure Profilers who suspect their target users have pivoted can back off, using less recent items in the profiles for the attribute inference attack. We carry out a test of the backoff strategy on the Emotion Profile data set. Figure 4.9 illustrates the performance as the attacker backs off to using less and less, older and older data within the target user's profile. When the profile has been pivoted (orange line), the accuracy increases. However, when the profile has not been pivoted (blue line), it decreases. These results suggest that in order to use back-off to counteract a pivoting addition, it is necessary to be able to actually predict whether and when a user has pivoted, which would be particularly challenging in the NatI case where the pivoting addition is composed of natural images. Assuming all profiles are pivoted does not increase the overall success of the attack. Also, back-off may not be applicable in a realistic scenario in which case users use multiple pivots.

Training on data augmented with pivoted profiles Profilers that suspect that profiles have been pivoted can also use adversarial training, i.e, train their bag-based clas-

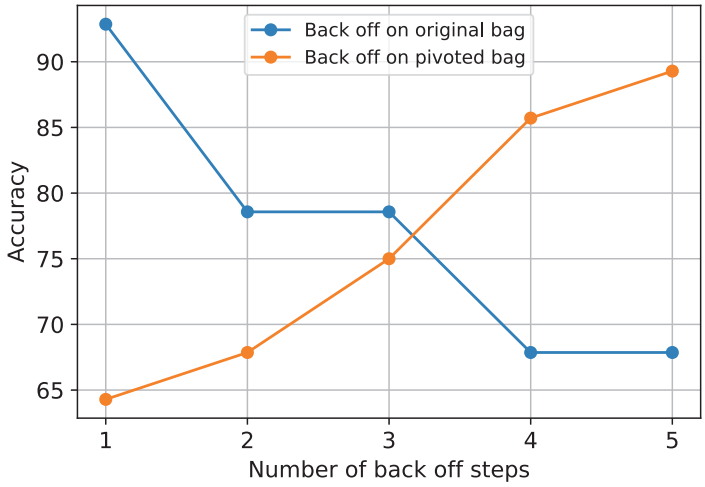


Figure 4.9: Averaged profiling accuracy (%) on five Emotion Profile data sets when different steps of backoff are implemented as mitigation to pivoting addition. Back off means that the profiler drops recent training items across all profiles. The blue curve represents the accuracy on normal profiles, and the orange curve represents the accuracy on pivoted profiles. Note that in this case, we assume original profiles are mixed with pivoted profiles, and the accuracies represent the influence of backoff on these two different groups of profiles.

Table 4.11: Accuracy (%) of profiling models trained on pivoted profiles to predict the emotion "anger" on Emotion Profile data set. The profiling model is a binary deep bag-based classifier that predicts the level of "anger".

	DEEP SETS	SETT
ORIGINAL CLASSIER (NON-PIVOTED)	84.62	69.87
ORIGINAL CLASSIER (NATI)	52.00	51.00
ADVERSARIALLY-TRAINED CLASSIFIER (NON-PIVOTED)	71.15	67.23
ADVERSARIALLY-TRAINED CLASSIFIER (NATI)	61.00	54.00

sifiers on privoted profiles. Table 4.11 presents experimental results that show that adversarial training can increase the robustness of the deep bag-based classifier to pivots, but compromises the performance on non-pivoted profiles.

4.8 Conclusion and Outlook

In this work, we have studied the use of deep bag-based classifiers for attribute profiling attacks and proposed obfuscating profile extensions that are capable of offering resistance. Because, to our knowledge, we are first to study profiling attacks

from the profile-level perspective, we have focused on systematic coverage of core cases, leaving exploration of the entire scope of the threat and possible defenses to future work. The key dimensions of attribute profiling attacks were outlined in our initial short work [113], in the form of a framework for defining the problem, and can help to guide future research directions. In this section, we discuss a selection of aspects we see as important to obtaining a broad understanding of bag-based attribute profiling, from both the attacker and the defender perspectives.

4.8.1 Patterns within Profiles

We expect a wide range of differences in the way that the evidence for a specific privacy sensitive attribute is distributed over user profiles. For this reason, we expect that the most effective bag-based classifier will different depending on the nature of the profile. In Section 4.1, we stated that in this work we assume that evidence for the privacy-sensitive attribute is spread over multiple images in the profile, rather being focused in a single image. Future work can investigate cases in which the evidence is distributed across a profile in a variety of ways. With respect to image profiles, both the proportion of profile images containing evidence and the clarity (or detectability) of such evidence are important dimensions of variation to consider.

Future work should also investigate the case in which the temporal order of items within the profile contains information relevant for profiling. For example, temporal ordering would be important if the attacker is trying to infer whether a user has suffered a serious illness or a traumatic life event. In this case, relevant MIL classifiers are no longer bag-based, but are rather classifiers that take the ordering of instances into account.

4.8.2 Knowledge of Patterns within Profiles

Future work should expand the threat model to include knowledge of the patterns within the target profiles. The late-fusion approach did not perform well given the focus of our work on profiles with diffuse evidence. However, it may be appropriate if the profiler knows that evidence is focused in one or more clear images, for example, if the attacker is trying to acquire knowledge of whether a person has ever visited a particular location.

Similarly, the knowledge of the target user is also important. If target users are aware of the importance of patterns within profiles, they might take action to obfuscate these patterns. For example, instead of a single pivot, they might switch between pivoted or non-pivoted examples to throw the profiler off track.

When developing new pivoting additions, it is important to keep in mind which characteristics of the profile that the user wishes to maintain and how radical of a change they will tolerate. For example, in the NatI approach studied here, the user creates the pivot by selecting images they would have posted anyway, leading to a minimal disruption in their sharing activity. To obfuscate privacy-sensitive

information that is localized within the profile, like location, a user could choose a pivot that contains images of them at locations that they did not visit, which introduces deniability. However, they might not be willing to change their sharing activity in this way since it could be perceived on social media as duplicitous, undermining social relationships.

4.8.3 More Sophisticated Attacks and Defenses

Moving forward, we expect that profilers will use increasingly sophisticate attacks. Previously we have mentioned that profilers can leverage unannotated profiles to develop semi-supervised deep bag-based classifiers. Target users could also develop data poisoning approaches to make their data less useful for profiling attacks. In sum, this work has laid the groundwork for a rich variety of follow-up work. Continued investigates will provide additional insight on how to protect users in real world cases in which profile-level attacks and not item-level attacks are the threat.

Chapter 5

Privacy Improvement by Adversarial Queries

An adversarial query is an image that has been modified to disrupt content-based image retrieval (CBIR), while appearing nearly untouched to the human eye. This work presents an analysis of adversarial queries for CBIR based on neural, local, and global features. We introduce an innovative neural image perturbation approach, called Perturbations for Image Retrieval Error (PIRE), that is capable of blocking neural-feature-based CBIR. PIRE differs significantly from existing approaches that create images adversarial with respect to CNN classifiers because it is unsupervised, i.e., it needs no labeled data from the data set to which it is applied. Our experimental analysis demonstrates the surprising effectiveness of PIRE in blocking CBIR, and also covers aspects of PIRE that must be taken into account in practical settings, including saving images, image quality and leaking adversarial queries into the background collection. Our experiments also compare PIRE (a neural approach) with existing keypoint removal and injection approaches (which modify local features). Finally, we discuss the challenges that face multimedia researchers in the future study of adversarial queries.

This Chapter is published as Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Who's afraid of adversarial queries? The impact of image modifications on content-based image retrieval. International Conference on Multimedia Retrieval (ICMR). 2019.

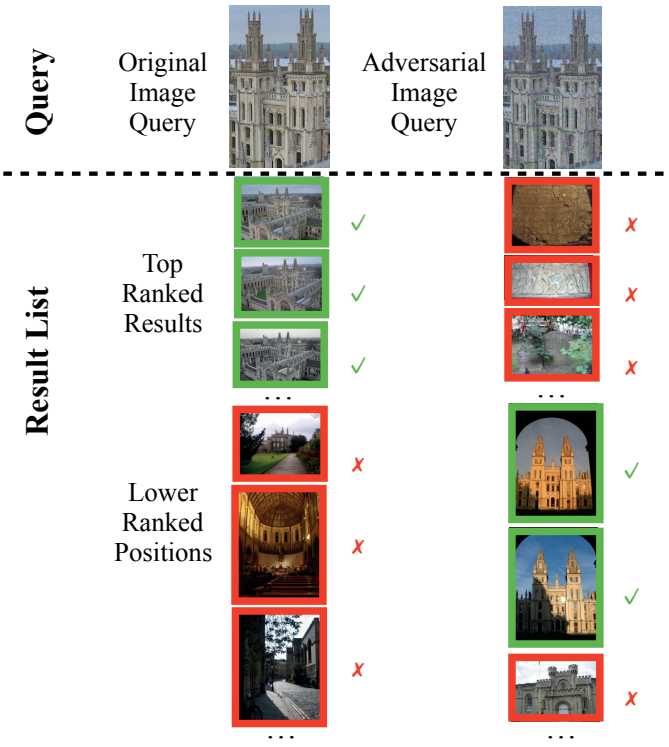


Figure 5.1: A successful query image (top left) and the corresponding adversarial query (top right). The two are visually nearly identical to the human eye. A CBIR system ranks relevant results high for the original image query and low for the adversarial query.

5.1 Introduction

Recently, researchers working on deep learning for image classification have started to study adversarial images intensively and to develop techniques to create them [184; 63; 123; 21; 122; 124]. Their work defines an *adversarial example* to be an image that a human can easily interpret, but that a CNN-based classifier assigns to an unexpected class. Typically, adversarial examples are created by taking an image that is correctly classified by a classifier, and perturbing the pixels. The perturbations are small, such that humans can look at the modified image and judge it to be nearly untouched. The perturbations are also carefully chosen, such that the modified image is no longer classified correctly, but rather is moved over the decision boundary of the classifier and is classified incorrectly. Generally, an image set labeled with the target classes is used to train the perturbations.

In this work, we extend the idea of adversarial examples from image classification to content-based image retrieval (CBIR). We define an *adversarial query* as an image that a human can easily interpret, but that causes a CBIR system unexpected difficulties. The principle is illustrated by the example in Figure 5.1. The adversarial

query resembles the original image as closely as possible.

The fundamental difference between creating adversarial examples in the case of image classification and creating adversarial queries in the case of CBIR is the information available for guiding the image modifications. In contrast to classification systems, which assume a set of discrete classes, CBIR systems are designed to handle arbitrary queries and unconstrained background collections. Specifically, in the deep learning image classification scenario, adversarial modifications are informed by class boundaries. Decision boundary information is lacking in the CBIR scenario. As such, in the CBIR scenario, there is no obvious direction, or directions, in which to move an image in pixel space in order to create a query that is adversarial with respect to the CBIR system. In order to address this challenge, we propose an approach called Perturbations for Image Retrieval Error (PIRE).

PIRE is able to generate perturbations without needing guiding information (i.e., PIRE requires no class labels, or relevance judgments from the data set to which it is applied). PIRE perturbs images such that they can still be interpreted to the human eye, but that they no longer can be used as successful queries for CBIR.

In sum, this work makes the following contributions: (1) We explain why it is important to study adversarial queries. (2) We present PIRE, our neural perturbation approach to creating adversarial queries, and experimentally demonstrate its impact on different CBIR systems (i.e., systems using neural, local, and global features). (3) We discuss and analyze practical aspects of adversarial queries. The work is organized as follows: after introducing the importance of adversarial queries in Section 5.2, we present the relevant related work in Section 5.3. Section 5.4 describes the framework in which we carry out our experiments. Then, Sections 5.5 and 5.6 present our experiments and analyses. Finally, Section 5.7 pulls everything together, and provides an outlook on future work.

5.2 Why study Adversarial Queries?

The study of adversarial image examples is motivated by specific *threat models*. Informally defined, a threat model expresses what we should be worried about, i.e., the dangers that a specific system or technology must be able to ward off. Adversarial image queries play a role in widely different threat models, which are described in this section. Section 5.2.1 looks at adversarial queries as being dangerous. From this perspective, the practical application of our research is understanding attacks on CBIR systems in order to defend against them. Section 5.2.2 looks at adversarial queries as being protective. From this perspective, the practical application of our research is preventing, or at least disincentivizing, harmful use of CBIR.

5.2.1 Threat of image modification technology

The assumption behind many widely-adopted threat models is that modified images are a source of danger. Here, we discuss three familiar examples of such threat

models. First, researchers working on image classification are generally worried about scenarios in which modified images cause misclassification. This threat model applies, for example, to scenarios in which computer vision technology is used by self-driving cars [46]. Adversarial image queries are relevant to such threat models since memory-based image classifiers are generally based on CBIR systems. Second, researchers working on keypoint removal and injection are generally worried about scenarios in which modified images cause the identification of duplicate or near-duplicate images to be blocked. Such work, e.g., [40], [39] and [38], is carried out in the use scenario of preventing copyright violation and image forgery via copy-move. Third, researchers working on image forensics also care about the post-processing operations, such as resampling [143], double JPEG compression [140] and denoising [93], since they are of interest in a forensic examination of an image and can affect forensic methods in various ways [93]. If we consider these threat models, then our reason for studying adversarial queries is to understand how modified images can harm image matching systems.

5.2.2 Threat of image retrieval technology

The assumption behind another more recently emerging class of threat models is that the multimedia retrieval system itself is a source of danger. These retrieval-specific threat models are commonly adopted by researchers working on multimedia privacy. The specific threat is a privacy violation, specifically, harm that people suffer caused by malicious actors who misuse an existing retrieval system (for example an online image search engine) or who build their own retrieval system to search in a collection of misappropriated images. This danger was first articulated by [56], who described the threat of 'cybercasing': criminals using online search engines to mine the Web for users whose online sharing behavior reveals that they own valuable items, and when they are away on vacation. The concern has been recently grown stronger because of high profile data breaches, e.g., [7], which have made clear that sharing images in 'private' mode is not a perfect solution for protection. Unscrupulous actors can implement their own CBIR system if they can get their hands on enough data. The interesting and surprising aspect of retrieval-specific threat models is that giving people access to image modification technology actually would help them to protect themselves against those seeking to misuse their images. Instead of a danger, image modification is a form of protection. If we consider retrieval-specific threat models, then our reason for studying adversarial queries is to understand the conditions under which the matching ability of CBIR systems can be blocked.

A recent investigation concerned with the threat of cybercasing [29] examines the potential of image enhancements to block the inference of the geo-location of the photos that users take and post online. Our work differs from [29] in that we focus specifically on CBIR and we consider image queries that are explicitly designed to be adversarial. However, we adopt the same threat model, and we focus our investigation on a CBIR problem that is related to location because it involves images of buildings in cities.

5.3 Related Work

We first cover work on neural adversarial examples for image classification and then work on blocking local-feature-based CBIR.

5.3.1 Adversarial examples and classification

Research on adversarial examples in the deep learning community was launched by [184], who demonstrated the possibility of constructing images adversarial with respect to a convolutional neural network image classifier (which we will refer to as the 'CNN-model'). As mentioned in the introduction, the basic mechanism used to create adversarial examples is to perturb pixels to construct a misclassified image while at the same time minimizing the distance between the original image (input image) and adversarial image. Work on adversarial examples started with 'whitebox' approaches, which have full knowledge of the CNN-model that they are attempting to delude. The Fast Gradient Sign Method (FGSM) [63] makes use of the gradient of the model with respect to the input image. It increases the model's loss on the input image given the correct class label by perturbing it in the ascending direction of the gradient. DeepFool [123] extends FGSM with more precise control over the size of the perturbations. For both FGSM and DeepFool, the perturbations are specific to the input image, and the correct class label of that image is known. Comparatively, PIRE only operates on neural features without accessing any ground truth (e.g., relevance judgements) of the CBIR system.

Subsequently, researchers have worked to extend 'whitebox' methods so that they require less information about the input images and less information about the CNN-model. Universal Adversarial Perturbations (UAP) [122] took a first step in this direction. UAP produces perturbations that do not require prior knowledge of the input images, however it does need a labeled training set. UAP adversarial examples have been shown to have an adversarial effect on CNN-models other than the one originally used to generate the perturbations. The 'universal' in UAP means that the perturbations are generated to be effective for a majority of images, although in practice they fail for a subset of images. Another whitebox method that is universal in this respect is Fast Feature Fool (FFF) [124], which generates adversarial images by calculating the maximal spurious activations in each convolutional layer while constraining the size of perturbations. FFF, like UAP, produces perturbations without knowledge of the images to be modified. However, whereas UAP requires training data, FFF can make use of the CNN-model with no additional training needed.

Currently, 'blackbox' techniques, which can create images adversarial to an arbitrary CNN-model remain elusive. Attempts at 'blackbox' solutions leverage existing 'whitebox' solutions. An ensemble method has been proposed, which creates examples that are adversarial with respect to a number of known CNN-models, and then tests them against a blackbox model [110]. Also, reconstruction methods have been proposed, namely [136] and [134], which probe the blackbox model with test examples, and then train substitute models that mimic the real model.

In our work, we focus on the case where we have access to the trained CNN-model used by the CBIR system at the moment at which we create our adversarial queries. However, we point out that our approach is not a completely 'whitebox' approach. Labeled training data is used to pre-train and fine-tune the CNN-model, but PIRE is ultimately applied to images from a third semantically related, but yet completely unseen, data set. For this reason, we refer to our approach as unsupervised, and not requiring class labels.

5.3.2 Keypoint Removal and Injection (KR&I)

In order to provide a complete picture of the behavior of adversarial queries, we consider not only neural features, but also local features. We focus on SIFT-based methods because they are representative of local-feature-based CBIR and also due to the rich literature on SIFT KR&I. The first work to consider influences of KR&I in SIFT-based CBIR systems was [40], [39] and [38]. Here, blocking CBIR means blocking the retrieval of exact duplicate or near duplicate images. In contrast, we are interested in blocking the retrieval of images containing the same subject matter as the query, without a specific focus on matching duplicates or near duplicates.

Other KR&I work is not directly connected to CBIR, but focuses on image forensics and multimedia security. With security issues in mind, [74] proposed to modify SIFT features while simultaneously keeping image quality. Later the authors proposed an optimization-based approach [116]. Combining multiple techniques, Classification-based Attack (CLBA) [31] proposed to use an iterative procedure to apply different methods on keypoints in different classes. [109] proposed SIFT keypoint removal and injection methods which remove keypoints with minimized distortion on the processed image. Recently, [108] proposed Removal via Directed Graph Construction (RDG) method to remove SIFT keypoints for colour images while maintaining high visual quality.

5.4 Experimental Framework

Figure 5.2 depicts the framework in which we carry out our experimental analysis. Our experiments test different combinations of image modification approach and CBIR system. The top of the figure shows the query modification step, which either uses PIRE or KR&I. The bottom of the figure shows the image retrieval step, which uses a CBIR system based on either neural, local, or global features. In this section, we describe the design choices that we use for implementing the framework, before introducing PIRE in detail.

5.4.1 Content-based Image Retrieval Systems

A CBIR system accepts an image as a query and returns a list of relevant images as a result. The images are drawn from a larger collection, which we refer to as the background collection. In a basic CBIR system, such as the one we adopt here, ranking occurs by comparing the vector representing the query image with the vectors representing each of the images in the background collection. The results list

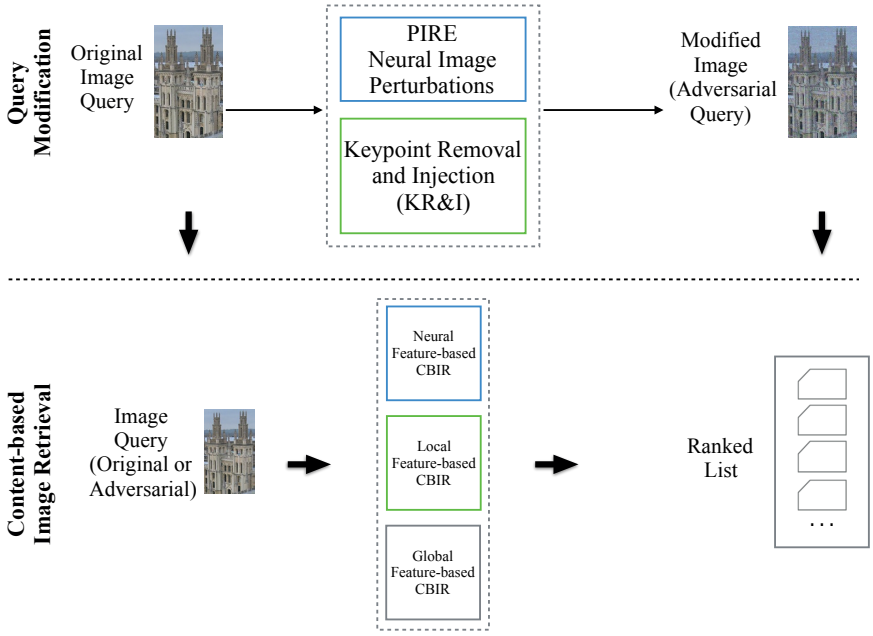


Figure 5.2: Our experimental analysis tests combinations of query modifications (Top) and CBIR systems (bottom). Blue boxes are neural-representation approaches and green boxes are local-feature approaches.

consists of the images from the background collection ranked in order of closeness to the query. CBIR systems are differentiated by the features that they use to create the image feature vectors. As previously mentioned, in our experiments we use neural, local, and global features. We describe each in turn.

Neural representations are compact representations that are extracted from an image using a pre-trained, and possibly then also fine-tuned, CNN-model. For our experiments, we need the currently best available neural representations, and for this reason we adopt GeM [146]. GeM is a fully convolutional CNN-model with a Generalized-Mean pooling layer. Using GeM as a feature extractor achieves the current state of the art on the data sets that we will use for our experiments, Oxford5k [141] and Paris6k [142], which are described in more detail below. We chose to use the structure of ResNet-101. GeM discards the fully-connected layer and replaces the average pooling layer of ResNet-101 with a Generalized-Mean pooling layer. The model is pre-trained on ImageNet and fine-tuned using a data set that consists of 120k Flickr images provided in [146] following a structure-from-motion (SfM) pipeline. The fine-tuning data set is a subset the data set of [168], which contains 7.4 million images from Flickr with keywords of landmarks, cities and countries across the world. The subset excludes Oxford and Paris.

GeM has been shown to outperform previous state-of-the-art approaches, which we mention here briefly for completeness. First, [174] used off-the-shelf neural networks. To improve retrieval performance, Neural Codes [5] used a fine-tuned CNN-model for neural feature extraction. Finally, [189] proposed regional maximum activation of convolutions (R-MAC) to improve image retrieval by adding an additional pooling layer to CNN-model.

The representations that are used by local-feature-based CBIR systems are generally Bag-of-Visual-Word (BoVW) models, dating back to [179]. Codebooks containing a certain number of visual words are trained on extracted SIFT features. We adopt a classic BoVW model with Hamming Embedding (HE) [83], which provides binary signatures that refine the visual-word-based matching. Following [83], we extract SIFT feature of images and train codebooks of size 20,000. Binary signatures of length 64 are used in the HE setting, and the threshold is set to 24. Note that the basic BoVW system that we adopt performs competitively with approaches such as VLAD in [84]. For this reason, we are confident that it meets the needs of the experiments we perform here. We save more detailed investigation of techniques such as geometric matching and query expansion for future work.

The representations used by a global-feature-based CBIR system capture information about overall image texture and image color, rather than information about specific keypoints. Color histograms and Edge histograms (MPEG-7 descriptors) are commonly used for extracting global features. For our experiments we adopt two widely-used global feature representations: Color and Edge Directivity Descriptor (CEDD) [24], which combines image color and texture information, and GIST [129], which extracts a holistic image representation reflecting the shape of a scene.

We perform experiments with two types of image queries: whole image queries (designated WI) and bounding box queries (designated BB). The BB queries use only the content of a bounding box that focuses on the main subject matter of the image. This bounding box is pre-defined (it is included with the queries in the data sets). We use BB queries in order to make our work comparable to other papers who test on the same data sets.

To evaluate, we compare the quality of the results list returned using the original image as a query with the results list returned by adversarial query (i.e., the modified image). We adopt mean average precision (mAP), a standard information retrieval evaluation metric, to measure results list quality. An image modification approach is successful if we observe a decrease in mAP when we move from the original query to the modified query. Finally, to evaluate visual quality, we use structural similarity (SSIM) [197], which assesses the degradation of structural information to be presumed related to the human-perceived quality.

5.4.2 Data

We perform our experiments on two data sets: Oxford5k and Paris6k, which are publicly available and widely used in CBIR research. Because so much work has

been done on these data sets, what constitutes state-of-the-art performance is well understood, and we can be certain that when we test the blocking effects of adversarial queries on CBIR, we are testing a strong CBIR system. The Oxford5k data set consists of 5063 images and includes 55 standard queries representing different views/parts of 11 Oxford buildings. Paris6k data set consists of 6412 images and also includes 55 standard queries from 11 different Paris landmarks. Both data sets include distractor images, which are not related to any of the queries in data set.

5.5 Neural-Feature-based CBIR

In this section, we propose a simple yet effective algorithm, Perturbations for Image Retrieval Error (PIRE), which blocks neural-feature-based CBIR by perturbing pixels of the image query.

5.5.1 Adversarial Queries with PIRE

The basic innovation of PIRE is to modify the original image by pushing its feature representation away from the original position in feature space. Specifically, PIRE maximizes the distance between the feature representation of the original image and that of the modified image, while at the same time limiting the overall size of the permutation. Recall that PIRE is designed with the assumption that the CNN-model (GeM [146] in this work) is available, and it aims to modify the input image with perturbations that are barely perceptible to the human eye.

PIRE is presented in Algorithm 5.1. $\boldsymbol{x}$ represents the image query, and $\boldsymbol{v}$ represents the perturbation vector. We start with a random perturbation feature vector that has the same size of the image and update it by optimizing the following objective function:

$$\begin{aligned} &\text{maximize} && \|f(\boldsymbol{x}) - f(\boldsymbol{x}+\boldsymbol{v})\|_2^2 \\ &\text{subject to} && \|\boldsymbol{v}\|_\infty \leq \epsilon \end{aligned} \tag{5.1}$$

This optimization process will stop when the iterative conditions are met. We create the final perturbation using a multiplicative factor, here, set to 10, to guarantee that the perturbations are retained when the image is saved in an 8-bit format. We return to address this factor in more detail in Section 5.5.2.1.

In each iteration, the perturbation vector is updated using the Adam optimization algorithm [92]. In our experiments we look at the impact of T, the number of rounds iterated. When the iterative conditions are met, perturbation vector $\boldsymbol{v_i}$ is the calculated perturbation vector.

In order to test PIRE, we apply it to all the query images of our data sets to create adversarial images, which are then saved. Table 5.1 reports results for the original queries and for adversarial queries created with PIRE (T = 500). It can be seen

Algorithm 5.1 Perturbations for Image Retrieval Error (PIRE)

Input:
$\mathbf{x}$: Image query; f:Neural feature function; T: Iteration limit; ϵ Perturbation vector range;
Output:
$\mathbf{x}^{'}$: Adversarial image query $\mathbf{x} + 10 * \mathbf{v}_i$

$w, h = \text{size}(\boldsymbol{x}_1)$;
$i = 1$;
Generate a random matrix $\boldsymbol{v}_{0_{(w \times h)}}$;
while $i < T$ **do**
    Calculate the distance between original image $\boldsymbol{x}$ and perturbed image $\boldsymbol{x} + \boldsymbol{v}_{i-1}$;
    $\boldsymbol{v_i} = \text{argmax}_{\boldsymbol{v}} \|(f(\boldsymbol{x}) - f(\boldsymbol{x} + \boldsymbol{v}_{i-1}))\|_2^2$;
    Project $\boldsymbol{v_i}$ into a L_∞ norm sphere;
    $\boldsymbol{v_i} = \text{clip}(\boldsymbol{v}_i, -\epsilon, \epsilon)$;
    $i = i + 1$;
end while
Return perturbed image query;
return $\mathbf{x} + 10 * \mathbf{v}_i$;

Table 5.1: Performance (mAP) of neural-feature-based CBIR (GeM [146]) on Oxford5k and Paris6k data sets before and after original PIRE ($10 * v$) modification with T=500 iterations.

	Oxford5k (BB)	Paris6k (BB)
Original	78.39	87.27
PIRE (T = 500)	5.51	9.34

that the mAP drops dramatically, indicating that PIRE is highly successful. Note that here we report bounding box (BB) queries only, and we are not yet concerned with the visual appearance of queries. As we will in Section 5.5.2, the choice of T allows us to control the trade off between PIRE's adversarial effect and its visual impact.

5.5.2 Adversarial Queries in Practice

5.5.2.1 Saving queries

In order for PIRE to be used in practice, it is necessary that adversarial queries remain adversarial when they are saved. When saving an image in JPEG format (uint8), float values that do not fit into 8 bits are approximated. This means that the perturbations that PIRE adds to an image should not be so small that they disappear when the image is saved. In [21], a method based on greed search was proposed to avoid the rounding effects discussed above. Saving images is obviously important, and so in Algorithm 5.1, on the last line, we use a multiplicative factor

Table 5.2: Performance (mAP) of neural-feature-based CBIR (GeM [146]) on Oxford5k and Paris6k data sets before and after query modification by PIRE ($p(\boldsymbol{v})$) (T=200 and T=500).

	Oxford5k (BB / WI)	Paris6k (BB / WI)
Original	78.39/74.42	87.27/87.26
PIRE (T = 200)	22.98/18.00	34.49/26.53
PIRE (T = 500)	3.93/2.31	10.53/7.18

10 in order to make sure that our perturbations survive rounding.

However, this approach of blindly making perturbations large is not elegant, since large perturbations lead to artifacts that are visually obvious. To tackle this issue, we propose a refinement to PIRE. The refinement adds a function $p(\boldsymbol{v_i})$ that magnifies the original perturbation vector to be just large enough not to be rounded away when the image is saved. These controlled perturbations improve the visual appearance of the adversarial queries. As the final step, the refined PIRE algorithm returns $\boldsymbol{x}_q + p(\boldsymbol{v}_i)$ as the modified image query. Exploratory experiments allowed us to observe that refined PIRE is able to substantially reduce the amount of perturbation needed to achieve an adversarial effect, and, as will be discussed below, also improves the visual appearance. We point out that the example in Figure 5.1 is an actual query from the Oxford5k data set tested with respect to our neural-feature-based CBIR system. The original query image achieves an AP of 93.95 and for the adversarial query (created with refined PIRE T=500) the AP plunges to 3.77. The PIRE results in the rest of the work are for refined PIRE.

5.5.2.2 Viewing queries

Next, we focus on the experience of users viewing adversarial queries. In order to get further insight into the visual impact of PIRE perturbations, we experimented with different levels of perturbation. Specifically, we prepared adversarial image queries with refined PIRE for two different representative values of the threshold T, which controls the number of iterations used to calculate the perturbations. Table 5.2 shows that adversarial image query generated with fewer rounds (T=200) still strongly decreases the performance of neural-feature-based CBIR.

In Figure 5.3, we present example queries to illustrate the contrast between PIRE using different values of the threshold on the number of iterations (T=200 and T=500). In order to quantify the relative difference in impact on the visual appearance, we report SSIM values in Table 5.3. Although more iterations lower the SSIM, the quality is still acceptable at both levels. In addition, we compared the SSIM value of the original PIRE ($10 * v$) and the refined PIRE ($p(\boldsymbol{v})$). SSIM on BB queries from Oxford5k (T = 200) went from 0.757 (PIRE) to 0.801 (refined PIRE). For BB queries on Paris6k (T = 200) results went from 0.690 (PIRE) to 0.771 (refined PIRE).

Figure 5.3: Examples of original image queries vs. adversarial queries generated using PIRE with different number of iterations (T=200 and T=500). Fewer iterations lead to less visible perturbations. (Best viewed on screen with magnification.)

Table 5.3: Average image quality (SSIM values) of adversarial queries from Oxford5k and Paris6K data sets generated by PIRE. (The SSIM value of the original query equals 1; BB= Bounding Box and WI=Whole Image.)

	Oxford5k (BB / WI)	Paris6k (BB / WI)
PIRE (T=200)	**0.801/0.754**	**0.793/0.771**
PIRE (T=500)	0.738/0.687	0.727/0.716

5.5.2.3 Protecting queries

Next, we demonstrate that PIRE has potential to cause a drop in the mAP of a CBIR system when the neural network used for indexing is unknown. We used a new neural network architecture, VGG-GeM, as the basis for generating adversarial queries with PIRE. We tested these queries against our original neural-feature-based CBIR system, which uses ResNet-GeM. We observed a mAP drop from 74.42 to 2.91 (Oxford5k), and from 87.26 to 9.39 (Paris6k) (T=500; cf. Table 5.2). We note that when VGG-GeM is used for both PIRE and retrieval, the effect is comparable to when ResNet-GeM is used for both PIRE and retrieval. We do not investigate VGG-GeM in more detail here, since the SSIM is ca. 0.15 lower for PIRE queries created with VGG-GeM than for PIRE queries created with ResNet-GeM. Our conclusion here is that PIRE has the potential to lower mAP without access to information on the architecture used for indexing. This conclusion is consistent with a further set of exploratory CBIR experiments we carried out with Google Images (`https://images.google.com`). The details of this system are unknown to us, but we assume that advanced neural representations are used, and that the background collection (index) is very large. We found the existence of a unexpectedly high number of cases in which the results returned by Google Images are impacted by PIRE. Future work on the investigation of nature of this impact promises to yield further interesting insight.

5.5.2.4 Editing Queries

Simple image transformations, such as resizing and cropping, may destroy the specific structure of adversarial perturbations. This effect was pointed out by [204]

Table 5.4: Performance (mAP) of neural-feature-based CBIR (GeM [146]) on the Oxford5k data set with bounding box queries and different resizing/cropping settings. Original image queries are compared with PIRE adversarial queries.

Resizing	50%	80%	100%	150%	200%
Original	65.09	74.31	78.39	71.30	62.03
PIRE (T = 500)	55.91	41.41	3.93	16.08	12.06
Cropping	100%	90%	80%	60%	40%
Original	78.39	76.01	76.01	69.54	46.89
PIRE (T = 500)	3.93	24.13	27.26	25.81	10.35

for image classification. Here, we use the Oxford5k data set to test the robustness of the perturbations generated by our PIRE against resizing and cropping of the adversarial query. These transformations obviously will also affect the performance of the original queries, so we report results for those as well.

For image resizing, we implement upscaling and downscaling operations, resulting in resized image queries with 200%, 150%, 80% and 50% of the original size. From the results, which are reported in Table 5.4, it can be observed that upscaling has only a small influence on PIRE (i.e., PIRE mAP remains lower than original mAP), while downscaling has larger influence (i.e., PIRE mAP and original mAP are closer). We suspect that the effect of adversarial queries lies in the subtle perturbation of the original pixels, downscaling changes most of the perturbed pixels. On the other hand, upscaling only interpolates new pixels between the perturbed pixels and for this reason does not impact the structure of perturbation as strongly.

For image cropping, we apply four different settings, i.e., 40%, 60%, 80% and 90% of the original size. From the results, which are also reported in Table 5.4, we can observe that image cropping has more impact than resizing on the original performance of CBIR, which we attribute to the loss of image content. However, PIRE remains effective, and causes substantial performance drops in different settings of image cropping.

5.5.2.5 Leaking queries

If PIRE is used in practice, it can be expected that some images that have been perturbed with PIRE find their way (i.e., “leak") into the background collection. We use one query from each of our data sets (christ-church-4 for Oxford5k and triomphe-3 for Paris6k) to explore what happens when not only queries, but also background images are perturbed with PIRE. For each query, we replace all its original relevant images (i.e., the ones labeled *good* or *ok*) in the background collection with adversarial versions using PIRE (T=200). We test two cases: one in which the image queries are perturbed with exactly the same setting of PIRE (T=200), and one in which they are perturbed with a different setting (T=500).

Table 5.5: Impact of PIRE on the performance (AP) of neural-feature-based CBIR (GeM [146]) for two specific queries before and after replacing the relevant images for these queries in the background collection.

Background	Query	christ-church-4	triomphe-3
Original	Original	93.88	89.52
Original	PIRE (T=200)	2.83	7.39
Replaced	PIRE (T=200)	34.52	48.85
Replaced	PIRE (T=500)	22.43	36.42

The results, reported in Table 5.5, demonstrate that adversarial queries can still maintain an adversarial effect when the relevant background images have been perturbed. If the relevant background images are perturbed with a different T than the adversarial query, the adversarial effect is stronger (mAP is lower) than when they are perturbed with the same T. These results suggest that adversarial queries leaking into the background collection might diminish, but will not negate, the adversarial effect over time. An approach to maintaining the strength of the adversarial effect would be to promote the use of diverse perturbation settings.

5.6 CBIR beyond Neural Features

In order to understand the larger implications of adversarial queries, we now turn to look at local and global image features.

5.6.1 Local-feature-based CBIR

Here, we test PIRE against the SIFT-based CBIR system introduced in Section 5.4.1, and compare it to existing KR&I-modifications that have been developed to block retrieval with local features.

5.6.1.1 KR&I-modification

We use the central methods from previous work to remove and inject SIFT keypoints. From [39], we test Removal with Minimum local Distortion (RMD) and Local Smoothing (LS), as well as the Forge new keypoints with Minimum local Distortion (FMD) method, which is representative of keypoint injection. In addition, we also test Removal via Directed Graph Construction (RDG) [108], a SIFT keypoint removal method that explicitly addresses visual quality.

We carry out experiments with SIFT-based CBIR on the original queries, on queries modified with five different KR&I methods, and on adversarial queries created with PIRE. Results are presented in Table 5.6. Only two KR&I methods (RMD + LS and LS + FMD) achieve substantial success in lowering the mAP compared to the mAP of the original queries, and some increase it (by unintentionally streamlining

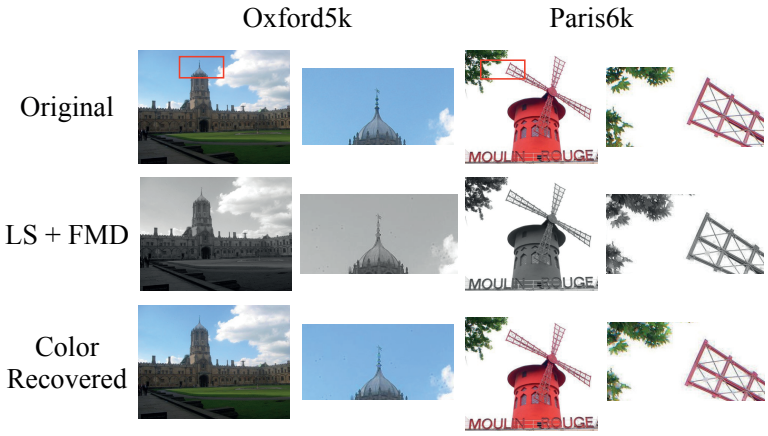


Figure 5.4: Examples of SIFT KR&I: original image queries with specific-region enlargements (top row), gray-scale images modified with LS + FMD (middle row), and the color recovered version of the modified images (bottom row).

Table 5.6: Performance (mAP) of SIFT-based CBIR on Oxford5k and Paris6k data sets: original queries and after modification with KR&I and PIRE. (BB=Bounding Box and WI=Whole Image.)

	Oxford5k (BB / WI)	Paris6k (BB / WI)
Original	52.57/51.59	45.46/44.63
RDG	53.00/51.08	44.45/44.44
RMD	53.81/53.02	44.47/45.23
RMD + LS	42.54/46.90	32.75/33.60
FMD	54.20/51.77	42.38/44.58
LS + FMD	41.23/**42.21**	**29.86**/**32.64**
PIRE (T = 500)	**40.90**/44.05	39.23/40.73

the visual word vocabulary). Interestingly, PIRE (T=500), although it is designed to be adversarial with respect to neural-feature-based CBIR, shows a blocking effect with respect to SIFT-based CBIR. For the Oxford5k data set, this effect is on par with the best of KR&I methods. We point out that KR&I methods maintain a better image quality than PIRE, as can be seen from Table 5.7, which reports SSIM for RMD + LS and LS + FMD.

5.6.1.2 SIFT color recovery

Since, essentially, SIFT features are extracted from single-channel images, in general, KR&I-modifications can only be applied to gray-scale images. However, because we are interested in visual appearance, we would like to compare color versions of KR&I-modified images. To this end, we propose a naive method to recover color after KR&I modification. Color recovery also allows us to make a fair comparison

Table 5.7: Average image quality (SSIM values) of adversarial queries from Oxford5k and Paris6K data sets generated by KR&I methods. (The SSIM value of the original query equals 1; BB= Bounding Box and WI=Whole Image.)

	Oxford5k (BB / WI)	Paris6k (BB / WI)
LS + FMD	0.917/0.938	0.953/0.971
RMD + LS	0.915/0.940	0.952/0.972

between the impact of PIRE and KR&I modification on global-feature-based CBIR in Section 5.6.2.

Given an original color image $\boldsymbol{I}_{rgb}$ with three channels $\boldsymbol{I}_r$,$\boldsymbol{I}_g$ and $\boldsymbol{I}_b$, its gray-scale version $\boldsymbol{I}_{gray}$ can be calculated by the widely-used formula $\boldsymbol{I}_{gray} = 0.30 * \boldsymbol{I}_r + 0.59 * \boldsymbol{I}_g + 0.11 * \boldsymbol{I}_b$. A successful recovery method should guarantee that the restored color image $\hat{\boldsymbol{I}}_{rgb}$ can be transformed back to the modified gray-scale image $\boldsymbol{I}_{mod}$ without the loss of modification effects, i.e., $\boldsymbol{I}_{mod} = 0.30 * \hat{\boldsymbol{I}}_r + 0.59 * \hat{\boldsymbol{I}}_g + 0.11 * \hat{\boldsymbol{I}}_b$. In order to recover the color information, we multiply the pixel at each location (i,j) by the same ratio α for the three channels of the original image $\boldsymbol{I}_{rgb}$. The process can be formalized as

$$\boldsymbol{\alpha}(i,j) = \boldsymbol{I}_{\text{mod}}(i,j) / \boldsymbol{I}_{\text{gray}}(i,j)$$
$$\{\hat{\boldsymbol{I}}_c(i,j) | c \in \{r,g,b\}\} = \boldsymbol{\alpha}(i,j) * \{\boldsymbol{I}_c(i,j) | c \in \{r,g,b\}\}$$

Figure 5.4 provides a impression of the image quality after modification with LS + FMD, using two example queries. For each example, details in the red square are enlarged and shown alongside the whole image query. Our simple color recovery method appears to achieve its aim well. We can observe that artifacts are present in the gray-scale modified images. These are echoed in the color-recovered images.

5.6.2 Global-Feature-based CBIR

Finally, we turn to investigating global-feature-based CBIR, using the CEDD and GIST systems described in Section 5.4.1. CEDD is a low computational-cost feature that incorporates color and texture information in a histogram, while GIST features can represent perceptual dimensions (naturalness, openness, roughness, expansion and ruggedness) of a semantic scene by encoding coarsely localized information in the energy spectrum of an image [108].

The results in Table 5.8 reveal that image modifications operating on local features (LS + FMD) do not block global-feature-based CBIR. However, our PIRE adversarial queries have a quite strong blocking effect on CEDD-based CBIR and a quite noticeable blocking effect on GIST-based CBIR. These results are interesting since PIRE was not trained to block global-feature-based retrieval. In order to

Table 5.8: Performance (mAP) of global-feature-based CBIR (Whole image queries): original queries, PIRE (T=500), KR&I modification, and Gaussian noise baseline.

	Oxford5k (CEDD / GIST)	Paris6k (CEDD / GIST)
Original	10.77/19.29	9.61/18.30
LS+FMD	10.67/18.58	9.92/17.86
PIRE (T=500)	**2.54**/**14.71**	**5.06**/12.11
Gaussian Noise	10.68/**14.71**	8.27/**9.93**

understand their implications, we must know whether PIRE is acting specifically to disrupt pixel patterns that are important for global-feature-based CBIR, or it is merely acting as a sophisticated method of introducing noise throughout the image, which then has a blocking effect because it makes the overall image quality worse. To this end, we carry out a baseline experiment using queries modified with Gaussian noise. Specifically, we generate noise for each image query such that the result is a SSIM value similar to the one caused by PIRE. (On average, for Oxford5k, SSIM equals 0.687 for PIRE, and 0.652 for Gaussian noise; for Paris6k, SSIM equals 0.716 for PIRE, and 0.708 for Gaussian noise.)

As shown in Table 5.8, Gaussian noise degrades performance in the case of GIST-based CBIR. We assume that the reason is that GIST extraction is based on spectral information, with which high-frequency noise interferes, and that PIRE is having a similar effect.

More interesting is how PIRE degrades the performance of CEDD-based CBIR. Our explanation for these results is that CEDD captures texture information, and that the structural perturbations generated by PIRE interfere with texture more effectively than the random changes of Gaussian noise. For completeness, we confirm that this effect does not account for the ability of PIRE to block neural-feature-based CBIR. In Table 5.2, we saw that PIRE (T = 500) drops the mAP of a neural-based CBIR system from 74.42 to 2.31 for the Oxford5k data set (Whole Image queries). Here, the effect of Gaussian noise contrasts with the effect of PIRE. If Gaussian Noise instead of PIRE is used to modify the query, the drop is from 74.42 to only 71.45. Behavior on the Paris6k data set and with Bounding Box queries is comparable.

5.7 Conclusion and Outlook

This work has made the case for studying adversarial queries in content-based image retrieval. We have proposed a new algorithm called PIRE, which is a neural perturbation approach for creating adversarial queries. In contrast to previous work on adversarial examples, PIRE does not require supervision (i.e., no labels from the data set to which it is applied) and is for this reason suited for image retrieval scenarios. Our experimental analysis of PIRE and of other, more traditional, ap-

proaches for blocking image matching with keypoint injection and removal (KR&I) has provided valuable insight into adversarial queries. We summarize these insights in terms of their implications for different groups of researchers.

Researchers in deep learning: Our work opens interesting topics in CBIR for researchers in deep learning. First, improvements on PIRE can be explicitly designed to generate queries that are adversarial with respect to a CBIR system for which little or no information is available, i.e., that uses arbitrary neural representations. Next, we point out again that the data set used to fine-tune our CNN-model is semantically related to the data set to which PIRE is applied. Specifically, the fine-tuning data depicts buildings, but not specifically in our cities. In the future, the impact of this semantic relationship both on CBIR performance and on the ability of PIRE to block CBIR performance should be better understood. We also point to [107], work on universal perturbations for image retrieval that came to our attention while preparing the camera-ready version of this work. Future work should further develop the ideas of [107], such as universal perturbations (PIRE is image specific) and pseudo-supervision.

Researchers interested in local and global features: Our experimental analysis suggests that neural perturbations have potential to block local-feature-based CBIR and global-feature-based CBIR, opening interesting paths for future work. Our results with a neural-feature-based CBIR system show that adversarial queries created with neural perturbations lose their blocking ability after certain edits. KR&I approaches may not have these weakness.

Multimedia privacy researchers: Not everyone who is able to deploy CBIR on a large collection of users' images can be expected to have the users' best interests in mind, and actors with ill intent are an inevitable risk. Our results suggest that adversarial queries are a promising topic of study for multimedia privacy researchers. Note that modest reductions in CBIR performance may already be enough to deincentivize malicious actors from abusing CBIR systems. However, much research still lies ahead. In order to implement privacy protection, it is necessary to apply modifications not only to query images, but also to images in the background collection. Our experimental results suggest that more work should be devoted to understanding the rate at which image modifications need to change dynamically in order to block CBIR over time. Finally, in order for users to adopt image modifications to protect their privacy, it is necessary to pay close attention to the visual acceptability of the modified images. Future work must focus both on minimizing the visual impact of perturbations, as well as understanding how to make visual changes that are acceptable to the user, in cases in which it is necessary to make visible changes.

In closing, we return to question in the work's title: 'Who's afraid of adversarial queries?' Depending on the threat model that a researcher adopts, adversarial queries might be considered part of the problem or part of the solution, and at first consideration might be more or less scary. However, our overall answer is that no

one should be afraid of adversarial queries, since they are important to understand and open up interesting new research questions.

Chapter 6

Conclusion and Outlook

6.1 Wrap-up

This thesis has focused on challenges and opportunities brought by adversarial machine learning to intelligent information systems from both the system owner's perspective and the user's perspective. We have studied cases in which the system outputs are influenced by externally sourced data from outside of the system. Adversarial data modifications on externally sourced data can compromise system security, but they also hold the potential to improve user privacy.

In **Chapter 2**, from the system owner's perspective, we demonstrated that when externally sourced background collection of a recommender system is adversarially modified, the output recommendations can be manipulated for malicious purposes. In the two-stage recommender constituted by a collaborative filtering algorithm with a visual re-ranker, we systematically showed that merchants can promote certain items by using adversarial machine learning under white-box, gray-box, and black-box settings. We also showed that existing countermeasures are ineffective in mitigating adversarial item promotion. Adversarial item promotion constitutes a practical threat to real-world recommender systems that use images to address cold start.

In **Chapter 3**, from the system user's perspective, we looked at visual feature representation models in intelligent information systems. Visual representation models pre-trained on large-scale datasets can be used in information systems in an off-the-shelf manner as the core of the image content processing module. Using the adversarial data poisoning method, we looked at how externally sourced data can influence the representation performance of the pre-trained model. In particular, we conducted a systematic study to demonstrate that compression methods can mitigate the current perturbative adversarial poisoning methods that compromise representation models' availability. From the users' perspective, we showed that availability poisoning is a promising method to mitigate misuse of user-originated data for training.

In **Chapter 4**, from the system user's perspective, we studied the use of deep bag-based classifiers for attribute profiling attacks and proposed obfuscating profile extensions that are capable of offering resistance. We formulated a threat model for bag-based user profiling from the perspectives of different groups of users. Assuming that users are protecting specific attributes, we showed that pivoting additions can resist private attribute inference by extending existing profiles. Questioning the assumption that the existence of already uploaded items is incompatible with obfuscation approaches to privacy protection, pivoting additions provided users with a promising solution to improve privacy on social media. Especially, adversarially selected non-touched natural items can effectively improve users' privacy while simultaneously being non-suspicious to human observers.

In **Chapter 5**, from the system user's perspective, we showed that adversarial queries could solely misdirect the performance of the image retrieval systems. We discussed cases where adversarial queries could be treated differently, covering the common concerns when applying adversarial queries. In addition, we showed that adversarial queries created on neural-feature models can transfer to image retrieval systems built on global and local features. Adversarial queries can affect the performance of content-based image retrieval systems by mismatching the query and background images. Adversarial queries can improve privacy against content-based image retrieval systems, where users' intents are maintained but extracted semantics are mismatched.

6.2 Outlook

In this section, we share our reflections regarding the security concerns and privacy opportunities associated with intelligent information systems and make suggestions for future research and practice.

6.2.1 Important Next Steps

Adversarial training. In this thesis, we have discussed adversarial examples from the user and the system point of view. To gain deeper understanding in both cases, it is important to carry out research on minimizing the impact of adversarial approaches. Although we study adversarial training techniques in Chapter 2 from the system owner's perspective and Chapter 3 from the user's perspective, this work needs to be continued moving forward due to the emerging gap between the state-of-the-art attacks and current countermeasures.

From the system owner's perspective, one promising defense to explore against information system attacks is adversarial training [118], which has been extensively studied in mitigating evasion attacks in computer vision. Adversarially-trained models are designated to be more robust to adversarially perturbed inputs than regular models. Adversarial training on information system models has been shown to be promising in improving regular retrieval [185] or recommendation [72] performance but has not been thoroughly discussed with regards to its ability to improve the robustness of information systems.

Future work could investigate dedicated adversarial training methods for information systems, especially focusing on solving the model performance and robustness trade-off. For example, for recommender systems that leverage media content, adversarially-trained multimedia feature extractors can substitute a conventionally-trained feature extractor, which could potentially benefit the robustness of recommenders against adversarially-perturbed or corrupted media content.

Future work could also explore detection strategies to filter out malicious content before it enters the systems. However, other perspectives on data filtering are also interesting. Adversarially-trained models have also been shown to provide more robust representations than regular models, which could be exploited by future work to build information systems that are less dependent on privacy-sensitive aspects of user data. In particular, features that are both spurious and privacy-sensitive can be spotted and removed from the externally-sourced data, which benefit both the system owners for system security and users for privacy.

Practical adversaries. Privacy and security threats can be realized differently when adversaries' capabilities are different. In this thesis, we have studied strong white-box adversaries and weaker, but more realistic, gray-box adversaries. Specifically, in Chapter 2, from the system owner's perspective, we showed that adversarial merchants can exploit the information about popular items for adversarial item promotion, and in Chapter 4, from the user's perspective, we showed that users can generate adversarial queries that transfer to unknown retrieval systems. In short, this thesis has made a step towards practical adversaries that could occur in real-world situations. However, our work has remained to a great extent focused on the machine learning algorithm itself and the full range of possibilities for adversaries in the real world stretches far beyond what was have considered here.

Real-world adversaries can comprehensively exploit available information to further boost the adversarial effectiveness. Future work can explore model extraction to facilitate the adversarial tasks, where the adversary queries the target system to build a system surrogate. For example, regarding both adversarial item promotion in Chapter 2 and adversarial queries in Chapter 4, target information systems can return results based on the exact queries/interactions.

From the system owner's perspective, real-world adversaries can exploit all available resources to substantially harm the information system security. For example, strong adversaries could reverse engineer the ranking mechanism of the information systems by carefully selected queries [131]. When the system is deployed on the edge, physical probing on information systems is also feasible by exploiting different physical side channels [73].

From the user's perspective, users can exploit queries/interactions and outputs pairs to build a surrogate model to improve their privacy. For example, users can log their interactions with the system. Together with the service-provided tags [156], users could analyze and pivot their behaviors to change the unintended personalization

tags from the service provider.

Natural adversarial items. Natural item-based obfuscations have the potential for privacy improvement with the advantage of being robust against mitigation and being non-suspicious. In this thesis, we took first steps to explore the influence of natural adversarial items on intelligent information systems. In Chapter 2, from the system owner's perspective, we explored semantic attacks that blend a part of a popular item image with the image of a regular item in by image editing. In Chapter 5, from the user's perspective, we added natural non-touched items to user profiles to obfuscate bag-based attribute profiling. These examples demonstrate the potential of natural adversarial items to improve privacy protection, and complement other obfuscation-based protections.

From the system owner's perspective, figuring out the factors that differentiate natural adversarial items from regular items could defend against adversarial attack with natural items. However, the system owners should respect user's wishes to make privacy-sensitive information less readily available to be exploited by information systems.

From the user's perspective, future research could systematically analyze failure cases of information systems, in order to determine the characteristics of natural adversarial items given different types of information systems. It would also be important to study the patterns of naturally occurring groups of the natural adversarial items. Based on the findings, stronger privacy improvements that are non-suspicious and robust can be developed.

6.2.2 Moving Further With the System View

This thesis has studied adversarial machine learning with a system view. In other words, our threat models, attacks, and countermeasures take the entire system into account, and not just the machine learning algorithms in individual system modules. Future work on adversarial machine learning should move forward studying this system view in order to remain as close as possible to real-world issues and opportunities. We close the thesis by discussing areas of future work in which the system view can help further develop adversarial machine learning approaches to protect user privacy.

Privacy improvement by collaboration between users. Current adversarial machine learning-based privacy improvement research mainly assumes that the protector is acting alone. In Chapters 4 and 5, we discussed cases where one query is adversarially modified or one user profile is adversarially extended, which provides a basic understanding of the formulated problem. Privacy improvements by collaboration between users are still to be explored.

Future research can investigate the potential of interconnections between users for privacy improvement. In the social profiling case, multiple users can modify their

profiles to bypass the profiling algorithm or even influence the future training of profiling algorithms. In addition, when more users show a similar changing pattern, it will be challenging for the system to differentiate normal patterns from pivoting patterns. Consequently, mitigation against profile pivoting will be non-trivial since similar patterns occur with multiple users, making it hard to localize the outliers.

Privacy improvement by considering multiple modalities. For privacy improvements, this thesis mainly focused on a single modality, i.e., image. Adversarial modifications on single modality show the potential for privacy improvement but resemble less the practical threat scenarios where samples belonging to different modalities can be used together for profiling. For example, the social profiling of user profiles in Chapter 5 could exploit images, text, and interactions to improve its effectiveness.

Data modification-based privacy improvements could consider leveraging all different modalities to mitigate profiling. Profile images can be used together with text to mitigate the profiling. At the same time, interconnections between different modalities could be considered to further improve the effectiveness of privacy improvement. For example, the level of coherence between images and surrounding textual descriptions could be an important indicator for personalization, based on which users could deliberately create image-caption pairs that obfuscate.

Toward dynamic adversarial modifications This thesis has discussed the influence of adversarially-modified data entering different processing stages of information systems. Initial discussions on mitigation techniques have also been provided. For example, we have shown in Chapter 3 that adaptive poisoning samples can bypass the compression-based defenses provided by the system owner. We also show in Chapter 5 that adversarially pivoted profiles can be slightly counteracted when the system owner implements some countermeasures.

Such static one-step mitigation is useful with clear practical implications. However, regarding adaptive adversaries, we should study scenarios simulating the potential dynamics between two parties. For instance, the availability poisons in Chapter 3 can be discussed with respect to more detailed threat models to show the potential under different adversaries' knowledge-level assumptions, providing deeper insight into real-world implications.

6.2.3 Final Word

Summarizing our outlook, we believe that externally sourced data should be managed so that users are informed and empowered to impact all parts of the system that are directly or indirectly based on their provided data. The first step is to provide users with details about how information systems depend on their data [1]. Making the data exploitation transparent can help users understand their willingness to contribute to the development of the system. The second step is to allow users to remove their data and the corresponding influences of their data on the information system, where machine unlearning [12] technique is a promising direction

to explore for information systems. We also believe that the dynamic and active interactions between users and information system owners could eventually benefit both parties, improving user's privacy and system's security and utility.

Bibliography

[1] (2024). What personal data and information can an individual access on request? https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/dealing-citizens/what-personal-data-and-information-can-individual-access-request_en, Online; accessed 2024-10-16.

[2] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

[3] Athalye, A., Carlini, N., and Wagner, D. (2018a). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*.

[4] Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2018b). Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*.

[5] Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V. (2014). Neural codes for image retrieval. In *European Conference on Computer Vision (ECCV)*.

[6] Barreno, M., Nelson, B., Joseph, A., and Tygar, D. (2010). The security of machine learning. *Machine Learning*, 81(2):121–148.

[7] Barrett, R. (2018). Facebook exposed 6.8 million users' photos to cap off a terrible 2018. https://www.wired.com/story/facebook-photo-api-bug-millions-users-exposed/, Online; accessed 2024-10-16.

[8] Berzon, A., Shifflett, S., and Scheck, J. (2019). Amazon has ceded control of its site. the result: Thousands of banned, unsafe or mislabeled products. https://www.wsj.com/articles/amazon-has-ceded-control-of-its-site-the-result-thousands-of-banned-unsafe-or-mislabeled-products-11566564990. Accessed: 2024-10-16.

[9] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. (2013). Evasion attacks against machine learning at test time. In *European Conference on Machine Learning and Data Mining (ECML)*.

[10] Bogen, M., Rieke, A., and Ahmed, S. (2020). Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

[11] Boulahia, S., Amamra, A., Madi, M., and Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition.

Machine Vision and Applications, 32(6):121.

[12] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. (2021). Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (S&P)*.

[13] Bracher, C., Heinz, S., and Vollgraf, R. (2016). Fashion DNA: Merging content and sales data for recommendation and article mapping. In *KDD Fashion Workshop*.

[14] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[15] Brunton, F. and Nissenbaum, H. (2015). *Obfuscation: A user's guide for privacy and protest*. Mit Press.

[16] Burke, R., O'Mahony, M., and Hurley, N. (2015). Robust collaborative recommendation. In *Recommender systems handbook*, pages 961–995. Springer.

[17] Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

[18] Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. (2019). On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*.

[19] Carlini, N., Jagielski, M., Choquette-Choo, C., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., and Tramèr, F. (2024). Poisoning web-scale training datasets is practical. In *IEEE Symposium on Security and Privacy (S&P)*.

[20] Carlini, N. and Wagner, D. (2017a). Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*.

[21] Carlini, N. and Wagner, D. (2017b). Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*.

[22] Chaabane, A., Acs, G., Kaafar, M., et al. (2012). You are what you like! information leakage through users' interests. In *Network and Distributed System Security Symposium (NDSS)*.

[23] Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference (BMVC)*.

[24] Chatzichristofis, S. and Boutalis, Y. (2008). CEDD: Color and edge directivity ddescriptor: A compact descriptor for image indexing and retrieval. In *International Con-*

ference on Computer Vision Systems.

[25] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

[26] Chen, S., Yuan, G., Cheng, X., Gong, Y., Qin, M., Wang, Y., and Huang, X. (2023). Self-ensemble protection: Training checkpoints are good data protectors. In *International Conference on Learning Representations (ICLR)*.

[27] Cheplygina, V., Tax, D., and Loog, M. (2015). On classification with bags, groups and sets. *Pattern Recognition Letters*, 59:11–17.

[28] Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J., Taylor, G., and Goldstein, T. (2021). LowKey: Leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations (ICLR)*.

[29] Choi, J., Larson, M., Li, X., Li, K., Friedland, G., and Hanjalic, A. (2017). The geo-privacy bonus of popular photo enhancements. In *ACM International Conference on Multimedia Retrieval (ICMR)*.

[30] Christakopoulou, K. and Banerjee, A. (2019). Adversarial attacks on an oblivious recommender. In *ACM Conference on Recommender Systems (RecSys)*.

[31] Costanzo, A., Amerini, I., Caldelli, R., and Barni, M. (2014). Forensic analysis of SIFT keypoint removal and injection. *IEEE Transactions on Information Forensics and Security*, 9(9):1450–1464.

[32] Das, N., Shanbhogue, M., Chen, S.-T., Hohman, F., Chen, L., Kounavis, M., and Chau, D. (2017). Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*.

[33] Das, N., Shanbhogue, M., Chen, S.-T., Hohman, F., Li, S., Chen, L., Kounavis, M. E., and Chau, D. H. (2018). SHIELD: Fast, practical defense and vaccination for deep learning using jpeg compression. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

[34] Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al. (2010). The youtube video recommendation system. In *ACM Conference on Recommender Systems (RecSys)*.

[35] Deldjoo, Y., Schedl, M., Cremonesi, P., and Pasi, G. (2020). Recommender systems leveraging multimedia content. *ACM Computing Surveys*, 53(5):1–38.

[36] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR).

[37] Di Noia, T., Malitesta, D., and Antonio, F. (2020). TAaMR: Targeted adversarial attack against multimedia recommender systems. In *International Workshop on Dependable and Secure Machine Learning.*

[38] Do, T.-T., Kijak, E., Amsaleg, L., and Furon, T. (2012). Enlarging Hacker's Toolbox: Deluding image recognition by attacking keypoint orientations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[39] Do, T.-T., Kijak, E., Furon, T., and Amsaleg, L. (2010a). Deluding image recognition in sift-based cbir systems. In *ACM Workshop on Multimedia in Forensics, Security and Intelligence.*

[40] Do, T.-T., Kijak, E., Furon, T., and Amsaleg, L. (2010b). Understanding the security and robustness of sift. In *ACM International Conference on Multimedia (MM).*

[41] Dodge, S. and Karam, L. (2016). Understanding how image quality affects deep neural networks. In *International Conference on Quality of Multimedia Experience.*

[42] Dong, Y., Pang, T., Su, H., and Zhu, J. (2019). Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[43] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR).*

[44] Dziugaite, G., Ghahramani, Z., and Roy, D. (2016). A study of the effect of JPG compression on adversarial images. *arXiv preprint arXiv:1608.00853.*

[45] Evtimov, I., Covert, I., Kusupati, A., and Kohno, T. (2021). Disrupting model training with adversarial shortcuts. In *ICML Workshop AML.*

[46] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[47] Fang, M., Gong, N., and Liu, J. (2020). Influence function based data poisoning attacks to top-n recommender systems. In *The Web Conference (WWW).*

[48] Fang, M., Yang, G., Gong, N., and Liu, J. (2018). Poisoning attacks to graph-based recommender systems. In *Annual Computer Security Applications Conference (ACSAC).*

[49] Feng, J., Cai, Q.-Z., and Zhou, Z.-H. (2019). Learning to confuse: generating training time adversarial data with auto-encoder. In *Advances in Neural Information Processing*

Systems (NeurIPS).

[50] Feng, J. and Zhou, Z.-H. (2017). Deep MIML network. In *AAAI Conference On Artificial Intelligence (AAAI)*.

[51] Fowl, L., Chiang, P.-y., Goldblum, M., Geiping, J., Bansal, A., Czaja, W., and Goldstein, T. (2021a). Preventing unauthorized use of proprietary data: Poisoning for secure dataset release. In *Security and Safety in Machine Learning Systems Workshop*.

[52] Fowl, L., Goldblum, M., Chiang, P.-y., Geiping, J., Czaja, W., and Goldstein, T. (2021b). Adversarial examples make strong poisons. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[53] Fredrikson, M., Jha, S., and Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[54] Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., and Ristenpart, T. (2014). Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Security Symposium*.

[55] Frenkel, S. (2024). Israel deploys expansive facial recognition program in gaza. *New York Times*.

[56] Friedland, G. and Sommer, R. (2010). Cybercasing the joint: On the privacy implications of geo-tagging. In *USENIX Conference on Hot Topics in Security*.

[57] Fu, S., He, F., Liu, Y., Shen, L., and Tao, D. (2021). Robust unlearnable examples: Protecting data privacy against adversarial learning. In *International Conference on Learning Representations (ICLR)*.

[58] Fu, Z., Xian, Y., Gao, R., Zhao, J., Huang, Q., Ge, Y., Xu, S., Geng, S., Shah, C., Zhang, Y., et al. (2020). Fairness-aware explainable recommendation over knowledge graphs. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

[59] Gomes, J. (2017). Boosting recommender systems with deep learning. In *ACM Conference on Recommender Systems (RecSys)*.

[60] Gong, N. and Liu, B. (2016). You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *USENIX Security Symposium*.

[61] Gong, N. and Liu, B. (2018). Attribute inference attacks in online social networks. *ACM Transactions on Privacy and Security*, 21(1):1–30.

[62] Gong, N., Talwalkar, A., Mackey, L., Huang, L., Shin, E. C. R., Stefanov, E., Shi, E., and Song, D. (2014). Joint link prediction and attribute inference using a social-attribute

network. *ACM Transactions on Intelligent Systems and Technology*, 5(2):1–20.

[63] Goodfellow, I., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

[64] Gunes, I., Kaleli, C., Bilge, A., and Polat, H. (2014). Shilling attacks against recommender systems: a comprehensive survey. *Artificial Intelligence Review*, 42:767–799.

[65] Guo, C., Frank, J., and Weinberger, K. (2019). Low frequency adversarial perturbation. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

[66] Guo, C., Rana, M., Cisse, M., and Van Der Maaten, L. (2018). Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*.

[67] Harvey, A. (2012). CV Dazzle: Camouflage from computer vision.

[68] He, H., Zha, K., and Katabi, D. (2023). Indiscriminate poisoning attacks on unsupervised contrastive learning. In *International Conference on Learning Representations (ICLR)*.

[69] He, J., Chu, W., and Liu, Z. (2006). Inferring privacy information from social networks. In *International Conference on Intelligence and Security Informatics*.

[70] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[71] He, R. and McAuley, J. (2016). VBPR: Visual bayesian personalized ranking from implicit feedback. In *AAAI Conference On Artificial Intelligence (AAAI)*.

[72] He, X., He, Z., Du, X., and Chua, T.-S. (2018). Adversarial personalized ranking for recommendation. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

[73] Horváth, P., Lauret, D., Liu, Z., and Batina, L. (2024). Sok: Neural network extraction through physical side channels. In *USENIX Security Symposium*.

[74] Hsu, C.-Y., Lu, C.-S., and Pei, S.-C. (2009). Secure and robust SIFT. In *ACM International Conference on Multimedia (MM)*.

[75] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. (2017). Densely connected convolutional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[76] Huang, H., Ma, X., Erfani, S., Bailey, J., and Wang, Y. (2021). Unlearnable examples: Making personal data unexploitable. In *International Conference on Learning*

Representations (ICLR).

[77] Huang, W., Geiping, J., Fowl, L., Taylor, G., and Goldstein, T. (2020). MetaPoison: Practical general-purpose clean-label data poisoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[78] Ilse, M., Tomczak, J., and Welling, M. (2018). Attention-based deep multiple instance learning. In *International Conference on Machine Learning (ICML)*.

[79] Isaak, J. and Hanna, M. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59.

[80] Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[81] Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[82] Jagadeesh, V., Piramuthu, R., Bhardwaj, A., Di, W., and Sundaresan, N. (2014). Large scale visual recommendations from street fashion images. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

[83] Jégou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *European Conference on Computer Vision (ECCV)*.

[84] Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[85] Jia, J. and Gong, N. (2018). AttriGuard: A practical defense against attribute inference attacks via adversarial machine learning. In *USENIX Security Symposium*.

[86] Jia, J., Wang, B., Zhang, L., and Gong, N. (2017). Attriinfer: Inferring user attributes in online social networks using markov random fields. In *The Web Conference (WWW)*.

[87] Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. (2020). Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI Conference On Artificial Intelligence (AAAI)*.

[88] Joon Oh, S., Fritz, M., and Schiele, B. (2017). Adversarial image perturbation for privacy protection–a game theory perspective. In *IEEE International Conference on Computer Vision (ICCV)*.

[89] Joshi, A., Mukherjee, A., Sarkar, S., and Hegde, C. (2019). Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *IEEE International Conference on Computer Vision (ICCV)*.

[90] Kalantidis, Y., Kennedy, L., and Li, L.-J. (2013). Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos. In *ACM International Conference on Multimedia Retrieval (ICMR)*.

[91] Kang, W.-C., Fang, C., Wang, Z., and McAuley, J. (2017). Visually-aware fashion recommendation and design with generative image models. In *International Conference on Data Mining (ICDM)*.

[92] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

[93] Kirchner, M. and Fridrich, J. (2010). On detection of median filtering in digital images. *Media Forensics and Security II*, 7541:754110.

[94] Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

[95] Krebs, L. M., Alvarado Rodriguez, O. L., Dewitte, P., Ausloos, J., Geerts, D., Naudts, L., and Verbert, K. (2019). Tell me what you know: Gdpr implications on designing transparency and accountability for news recommender systems. In *Extended abstracts of the CHI Conference on Human Factors in Computing Systems*.

[96] Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

[97] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[98] Kurakin, A., Goodfellow, I., and Bengio, S. (2017). Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR)*.

[99] Laidlaw, C., Singla, S., and Feizi, S. (2021). Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations (ICLR)*.

[100] Lam, S. and Riedl, J. (2004). Shilling recommender systems for fun and profit. In *The Web Conference (WWW)*.

[101] Larson, M., Liu, Z., Brugman, S., and Zhao, Z. (2018). Pixel privacy. increasing image appeal while blocking automatic inference of sensitive scene information. In *Working Notes Proceedings of the MediaEval Workshop*.

[102] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[103] Lee, J. and Abu-El-Haija, S. (2017). Large-scale content-only video recommendation.

In *ICCV Workshops*.

[104] Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. (2019). Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning (ICML)*.

[105] Li, B., Wang, Y., Singh, A., and Vorobeychik, Y. (2016a). Data poisoning attacks on factorization-based collaborative filtering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[106] Li, C. Y., Shamsabadi, A. S., Sanchez-Matilla, R., Mazzon, R., and Cavallaro, A. (2019a). Scene privacy protection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[107] Li, J., Ji, R., Liu, H., Hong, X., Gao, Y., and Tian, Q. (2019b). Universal perturbation attack against image retrieval. *arXiv preprint arXiv:1812.00552*.

[108] Li, Y., Zhou, J., and Cheng, A. (2017). SIFT keypoint removal via directed graph construction for color images. *IEEE Transactions on Information Forensics and Security*, 12(12):2971–2985.

[109] Li, Y., Zhou, J., Cheng, A., Liu, X., and Tang, Y. Y. (2016b). Sift keypoint removal and injection via convex relaxation. *IEEE Transactions on Information Forensics and Security*, 11(8):1722–1735.

[110] Liu, Y., Chen, X., Liu, C., and Song, D. (2017a). Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations (ICLR)*.

[111] Liu, Y., Zhang, W., Yu, N., et al. (2017b). Protecting privacy in shared photos via adversarial examples based stealth. *Security and Communication Networks*, 2017.

[112] Liu, Z., Zhao, Z., and Larson, M. (2019). Who's afraid of adversarial queries? the impact of image modifications on content-based image retrieval. In *ACM International Conference on Multimedia Retrieval (ICMR)*.

[113] Liu, Z., Zhao, Z., and Larson, M. (2021). Pivoting image-based profiles toward privacy: Inhibiting malicious profiling with adversarial additions. In *ACM Conference on User Modeling, Adaptation and Personalization (UMAP)*.

[114] Liu, Z., Zhao, Z., Larson, M., and Amsaleg, L. (2020). Exploring quality camouflage for social images. In *Working Notes Proceedings of the MediaEval Workshop*.

[115] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

[116] Lu, C.-S. and Hsu, C.-Y. (2012). Constraint-optimized keypoint inhibition/insertion attack: Security threat to scale-space image feature extraction. In *ACM International*

Conference on Multimedia (MM).

[117] Luo, T., Ma, Z., Xu, Z.-Q. J., and Zhang, Y. (2021). Theory of the frequency principle for general deep neural networks. *CSIAM Transactions on Applied Mathematics.*

[118] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR).*

[119] McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR).*

[120] Mehnaz, S., Dibbo, S. V., De Viti, R., Kabir, E., Brandenburg, B., Mangard, S., Li, N., Bertino, E., Backes, M., De Cristofaro, E., et al. (2022). Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In *USENIX Security Symposium.*

[121] Mobasher, B., Burke, R., Bhaumik, R., and Williams, C. (2007). Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology*, 7(4):23.

[122] Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. (2017). Universal adversarial perturbations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[123] Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[124] Mopuri, K., Garg, U., and Babu, R. (2017). Fast Feature Fool: A data independent approach to universal adversarial perturbations. In *British Machine Vision Conference (BMVC).*

[125] Naseer, M., Khan, S., Khan, M., Shahbaz Khan, F., and Porikli, F. (2019). Cross-domain transferability of adversarial perturbations. In *Advances in Neural Information Processing Systems (NeurIPS).*

[126] Neve, J. and McConville, R. (2020). ImRec: Learning reciprocal preferences using images. In *ACM Conference on Recommender Systems (RecSys).*

[127] Oh, S. J., Benenson, R., Fritz, M., and Schiele, B. (2016). Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision (ECCV).*

[128] Oh, S. J., Fritz, M., and Schiele, B. (2017). Adversarial image perturbation for privacy protection a game theory perspective. In *IEEE International Conference on Computer Vision (ICCV).*

[129] Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

[130] Orekondy, T., Schiele, B., and Fritz, M. (2017). Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *IEEE International Conference on Computer Vision (ICCV)*.

[131] Orekondy, T., Schiele, B., and Fritz, M. (2019). Knockoff nets: Stealing functionality of black-box models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[132] O'Mahony, M., Hurley, N., and Silvestre, G. (2002). Promoting recommendations: An attack on collaborative filtering. In *Database and Expert Systems Applications*.

[133] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[134] Papernot, N., McDaniel, P., and Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

[135] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z., and Swami, A. (2017a). Practical black-box attacks against machine learning. In *ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*.

[136] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z., and Swami, A. (2017b). Practical black-box attacks against machine learning. In *ACM Asia Conference on Computer and Communications Security (AsiaCCS)*.

[137] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *NeurIPS Autodiff Workshop*.

[138] Peng, K.-C., Chen, T., Sadovnik, A., and Gallagher, A. (2015). A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[139] Peralta, V. (2007). Extraction and integration of movielens and imdb data. *Laboratoire Prisme, Université de Versailles, Versailles, France*.

[140] Pevny, T. and Fridrich, J. (2008). Detection of double-compression in jpeg images for applications in steganography. *IEEE Transactions on Information Forensics and Security*, 3(2):247–258.

[141] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR).

[142] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[143] Popescu, A. and Farid, H. (2005). Exposing digital forgeries by detecting traces of resampling. *IEEE Transactions on Signal Processing*, 53(2):758–767.

[144] Pouyanfar, S., Yang, Y., Chen, S.-C., Shyu, M.-L., and Iyengar, S. (2018). Multimedia big data analytics: A survey. *ACM computing surveys*, 51(1):1–34.

[145] Prévost, B., Janssen, J., Camacaro, J., and Bessega, C. (2018). Deep inventory time translation to improve recommendations for real-world retail. In *ACM Conference on Recommender Systems (RecSys)*.

[146] Radenović, F., Tolias, G., and Chum, O. (2018). Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[147] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

[148] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[149] Radiya-Dixit, E., Hong, S., Carlini, N., and Tramèr, F. (2022). Data poisoning won't save you from facial recognition. In *International Conference on Learning Representations (ICLR)*.

[150] Radlinski, F. and Craswell, N. (2017). A theoretical framework for conversational search. In *Conference on Human Information Interaction and Retrieval*.

[151] Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. (2019). On the spectral bias of neural networks. In *International Conference on Machine Learning (ICML)*.

[152] Rajabi, A., Bobba, R., Rosulek, M., Wright, C., and Feng, W. (2021). On the (im) practicality of adversarial perturbation for image privacy. In *Privacy Enhancing Technologies Symposium (PETS)*.

[153] Rammstedt, B. and John, O. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212.

[154] Raval, N., Machanavajjhala, A., and Cox, L. (2017). Protecting visual secrets using

adversarial nets. In *CVPR Workshops*.

[155] Reece, A. and Danforth, C. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):15.

[156] Regulation, P. (2016). Regulation (EU) 2016/679 of the european parliament and of the council. *Regulation (EU)*, 679:2016.

[157] Ren, J., Xu, H., Wan, Y., Ma, X., Sun, L., and Tang, J. (2023). Transferable unlearnable examples. In *International Conference on Learning Representations (ICLR)*.

[158] Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

[159] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[160] Sabour, S., Cao, Y., Faghri, F., and Fleet, D. (2016). Adversarial manipulation of deep representations. In *International Conference on Learning Representations (ICLR)*.

[161] Sachidanandan, S., Luong, R., and Joergensen, E. (2019). Designer-driven add-to-cart recommendations. In *ACM Conference on Recommender Systems (RecSys)*, pages 525–525.

[162] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[163] Sandoval-Segura, P., Singla, V., Geiping, J., Goldblum, M., Goldstein, T., and Jacobs, D. (2022). Autoregressive perturbations for data poisoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[164] Sanghi, A., Chu, H., Lambourne, J., Wang, Y., Cheng, C.-Y., Fumero, M., and Malekshan, K. (2022). CLIP-Forge: Towards zero-shot text-to-shape generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[165] Sattar, H., Krombholz, K., Pons-Moll, G., and Fritz, M. (2020). Body shape privacy in images: Understanding privacy and preventing automatic shape extraction. In *European Conference on Computer Vision (ECCV)*.

[166] Saura, J., Ribeiro-Soriano, D., and Palacios-Marqués, D. (2021). From user-generated data to data-driven innovation: A research agenda to understand user privacy in digital markets. *International Journal of Information Management*, 60:102331.

[167] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

[168] Schonberger, J. L., Radenovic, F., Chum, O., and Frahm, J.-M. (2015). From single image query to detailed 3D reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[169] Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., Cristani, M., and Lepri, B. (2017a). What your facebook profile picture reveals about your personality. In *ACM international conference on Multimedia (MM)*.

[170] Segalin, C., Cheng, D. S., and Cristani, M. (2017b). Social profiling through image understanding: Personality inference using convolutional neural networks. *Computer Vision and Image Understanding*, 156:34–50.

[171] Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. (2020). The pitfalls of simplicity bias in neural networks. In *NeurIPS Workshops*.

[172] Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., and Zhao, B. (2023). Glaze: Protecting artists from style mimicry by Text-to-Image models. In *USENIX Security Symposium*.

[173] Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., and Zhao, B. (2020). Fawkes: Protecting privacy against unauthorized deep learning models. In *USENIX Security Symposium*.

[174] Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). Cnn features off-the-shelf: An astounding baseline for recognition. In *CVPR Workshops*.

[175] Shen, J., Zhu, X., and Ma, D. (2019). TensorClog: An imperceptible poisoning attack on deep neural network applications. *IEEE Access*, 7:41498–41506.

[176] Shi, Y., Larson, M., and Hanjalic, A. (2014). Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys*, 47(1):1–45.

[177] Shin, R. and Song, D. (2017). JPEG-resistant adversarial images. In *NeurIPS Workshops*.

[178] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

[179] Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision (ICCV)*.

[180] Slokom, M., Hanjalic, A., and Larson, M. (2021). Towards user-oriented privacy for recommender system data: A personalization-based approach to gender obfuscation for user profiles. *Information Processing & Management*, 58(6):102722.

[181] Snoek, C., Worring, M., and Smeulders, A. W. (2005). Early versus late fusion in

semantic video analysis. In *ACM international conference on Multimedia (MM)*.

[182] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[183] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

[184] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*.

[185] Tang, J., Du, X., He, X., Yuan, F., Tian, Q., and Chua, T.-S. (2019). Adversarial training towards robust multimedia recommender system. *IEEE Transactions on Knowledge and Data Engineering*, 32(5):855–867.

[186] Tang, J., Wen, H., and Wang, K. (2020). Revisiting adversarially learned injection attacks against recommender systems. In *ACM Conference on Recommender Systems (RecSys)*, pages 318–327.

[187] Tao, L., Feng, L., Yi, J., Huang, S.-J., and Chen, S. (2021). Better safe than sorry: Preventing delusive adversaries with adversarial training. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[188] Tian, S., Yang, G., and Cai, Y. (2018). Detecting adversarial examples through image transformation. In *AAAI Conference On Artificial Intelligence (AAAI)*.

[189] Tolias, G., Sicre, R., and Jégou, H. (2016). Particular object retrieval with integral max-pooling of cnn activations. In *International Conference on Learning Representations (ICLR)*.

[190] Tramer, F. and Boneh, D. (2019). Adversarial training and robustness for multiple perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[191] Tramer, F., Carlini, N., Brendel, W., and Madry, A. (2020). On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems (NeurIPS)*.

[192] Van Der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

[193] van Vlijmen, D., Kolmus, A., Liu, Z., Zhao, Z., and Larson, M. (2022). Generative poisoning using random discriminators. In *ECCV Workshops*.

[194] Vivek, B., Mopuri, K. R., and Babu, R. V. (2018). Gray-box adversarial training. In *Proceedings of the European conference on computer vision (ECCV)*.

[195] Wang, J., Huang, P., Zhao, H., Zhang, Z., Zhao, B., and Lee, D. (2018). Billion-scale commodity embedding for e-commerce recommendation in alibaba. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

[196] Wang, X., Chen, H., Sun, P., Li, J., Zhang, A., Liu, W., and Jiang, N. (2023). AdvST: Generating unrestricted adversarial images via style transfer. *IEEE Transactions on Multimedia*.

[197] Wang, Z., Bovik, A., Sheikh, H. R., and Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

[198] Wang, Z., Wang, Y., and Wang, Y. (2021). Fooling adversarial training with inducing noise. *arXiv:2111.10130*.

[199] Weinsberg, U., Bhagat, S., Ioannidis, S., and Taft, N. (2012). BlurMe: Inferring and obfuscating user gender based on ratings. In *ACM Conference on Recommender Systems (RecSys)*.

[200] Wen, R., Zhao, Z., Liu, Z., Backes, M., Wang, T., and Zhang, Y. (2023). Is adversarial training really a silver bullet for mitigating data poisoning? In *International Conference on Learning Representations (ICLR)*.

[201] Wenger, E., Shan, S., Zheng, H., and Zhao, B. (2023). SoK: Anti-facial recognition technology. In *IEEE Symposium on Security and Privacy (S & P)*.

[202] Wilber, M., Shmatikov, V., and Belongie, S. (2016). Can we still avoid automatic face detection? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[203] Wu, S., Chen, S., Xie, C., and Huang, X. (2023). One-pixel shortcut: On the learning preference of deep neural networks. In *International Conference on Learning Representations (ICLR)*.

[204] Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. (2018). Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*.

[205] Xie, C. and Yuille, A. (2020). Intriguing properties of adversarial training at scale. In *International Conference on Learning Representations (ICLR)*.

[206] Xie, Y. and Richmond, D. (2018). Pre-training on grayscale ImageNet improves medical image classification. In *ECCV Workshops*.

[207] Xu, W., Evans, D., and Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security Symposium (NDSS)*.

[208] Xu, Z.-Q. J., Zhang, Y., and Xiao, Y. (2019). Training behavior of deep neural

network in frequency domain. In *International Conference on Neural Information Processing*.

[209] You, Q., Luo, J., Jin, H., and Yang, J. (2016). Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI Conference On Artificial Intelligence (AAAI)*.

[210] Yu, D., Zhang, H., Chen, W., Yin, J., and Liu, T.-Y. (2022). Availability attacks create shortcuts. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

[211] Yuan, C.-H. and Wu, S.-H. (2021). Neural tangent generalization attacks. In *International Conference on Machine Learning (ICML)*.

[212] Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., and Smola, A. (2017). Deep sets. *Advances in Neural Information Processing Systems (NeurIPS)*.

[213] Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*.

[214] Zhang, J., Ma, X., Yi, Q., Sang, J., Jiang, Y., Wang, Y., and Xu, C. (2023). Unlearnable clusters: Towards label-agnostic unlearnable examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[215] Zhao, Z., Liu, Z., and Larson, M. (2020). Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[216] Zhao, Z., Liu, Z., and Larson, M. (2021). On success and simplicity: A second look at transferable targeted attacks. *Advances in Neural Information Processing Systems (NeurIPS)*.

[217] Zhao, Z., Liu, Z., and Larson, M. (2023a). Adversarial image color transformations in explicit color filter space. *IEEE Transactions on Information Forensics and Security*.

[218] Zhao, Z., Zhang, H., Li, R., Sicre, R., Amsaleg, L., Backes, M., Li, Q., and Shen, C. (2023b). Revisiting transferable adversarial image examples: Attack categorization, evaluation guidelines, and new insights. *arXiv preprint arXiv:2310.11850*.

[219] Zheleva, E. and Getoor, L. (2009). To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *The Web Conference (WWW)*.

Research Data Management

This thesis research has been carried out under the research data management policy of the Institute for Computing and Information Sciences of Radboud University, The Netherlands.[1]

The following research datasets have been produced during this PhD research:

- Chapter 2: Liu, Z. (Radboud University); Larson, prof. dr. M. (Radboud University) (2021): Adversarial item promotion: Vulnerabilities at the core of top-n recommenders that use images to address cold start. GitHub. `https://github.com/liuzrcc/AIP`

- Chapter 3: Liu, Z. (Radboud University); Zhao, dr. Z. (Xi'an Jiaotong University); Larson, prof. dr. M. (Radboud University) (2023): Image shortcut squeezing: Countering perturbative availability poisons with compression. GitHub. `https://github.com/liuzrcc/ImageShortcutSqueezing`

- Chapter 4: Liu, Z. (Radboud University); Zhao, dr. Z. (Xi'an Jiaotong University); Larson, prof. dr. M. (Radboud University) (2024): Resisting Bag-based Attribute Profiling by Adding Items to Existing Media Profiles. GitHub. `https://github.com/liuzrcc/Resisting-Bag-based-Profiling`

- Chapter 5: Liu, Z. (Radboud University); Zhao, dr. Z. (Xi'an Jiaotong University); Larson, prof. dr. M. (Radboud University) (2019): GitHub. `https://github.com/liuzrcc/PIRE`

[1] `https://www.ru.nl/en/institute-for-computing-and-information-sciences/research`, last accessed: 2025-03-09.

Acknowledgements

I want to thank my advisor, Martha Larson, for all her support, guidance, mentorship, and encouragement. The PhD journey is a self-doubting yet slowly correcting process where I learn the advancing pattern. Unreservedly, you instruct me how to conduct solid research with perseverance. It would be impossible for me to make it without your mindful troubleshooting. I was wondering what would happen if I did not get the chance to work with you, for which I believe my perspective on science and research would be quite different from now. Thank you for patiently guiding me through this journey, step by step.

I also want to thank my postdoc advisor, Lejla Batina, for all your trust and support. All your help and professional advice made me a solid step forward as a researcher with systematic thinking. The lab journey has been invaluable, providing me with an eye-opening experience in practice. I couldn't have done it without your support. Thank you for your patience and kindness.

Many thanks to members of my manuscript committee and examination board, Prof. dr. Marco Loog, Prof. dr. Marcel Worring, and Dr. Jennifer Williams, for your availability and timely feedback, and Prof. dr. ir. Inald Lagendijk, Dr. Rajiv Ratn Shah, and Prof. dr. Lejla Batina, for your valuable participation and support.

Working with my PhD brother, Zhenyu Zhao, is my great pleasure. Back in the autumn of 2018, I couldn't tell whether it was the heated discussion or the direct sunlight that filled the office with excitement. From setting up the GPU server together to debating the smallest details, time has flown by, and I learned a lot from you.

Special thanks to my DaS colleagues Alex, Emma, Gabriel, Jacopo, Konrad, Simon, Simone, Kai, Mingliang, Nik, Norman, Ruifei, Xiaomeng, Yao, and Zaheer for sharing research ideas, neat office rooms, table football, and interesting stories with me, especially, Alex and Ruifei, for your company and help. Another special thanks to Arjen, Djoerd, Elena, TomC, TomH, and Twan, for providing me with career support, professional advice, and valuable discussions. I want to thank Nicole and Kasper for solving all my random troubles. I would also like to thank all DaS colleagues Ankur, Charlotte, Chris, David, Faegheh, Feri, Franka, Gijs, Harrie, Hideaki, Ilona, Inge, Ivan, Jesse, Koen, Jelle, Johannes, Luc, Marieke, Mirthe, Nastaran, Negin, Nicole, Perry, Roel, Wieske, and Yuliya.

Working on digital security provides me a great opportunity to learn from excellent

colleagues outside the machine learning domain. To all CESCA colleagues: Asmita, Abrahamm, Azade, Bart, Behrad, Dirk, Durba, Giacomo, Ileana, Konstantina, Léo, Lizzy, Łukasz, Niels, Parisa, Péter, Senna, Silvia, and Vahid, thank you for your support and shared fun moments. I would also like to thank Hamid, Janet, Joan, Paulus, Ronny, Shanley, Stjepan, Tom, Veelasha, Xiaoyun, and all DiS colleagues: it is a great honor to learn from you.

My Nijmegen journey started smoothly with the help of all my friends: Chang, Johannes, Jeftha, Jing, Koen, Maarten, Marta, Niels, and Patrick. Thank you for all your friendship and support. I would also like to thank my friends Baojun, Bing, Dirren, Fugang, Guodong, Henan, Hongbo, Hongling, Manel, Manxia, Mo, Jiamian, Jiangyan, Jiaxin, Jinlong, Jinshuo, Keyang, Liu, Peiliang, Sam, Shuang, Xiaojing, Xiangrong, Wei, Yang, Yidong, Yiping, Yongjun, and Yunqi, Zhi, and Zongtian for your company. Special thanks to Chuan, Haojia, Sanzheng, Weibing, and Yongzheng for always backing me up, particularly Haojia for tricky cases. Thank you, my CISPA colleagues Rui, Zheng, and Yang, for your support. Some names may have been missed unintentionally. Thank you everyone.

To my parents, my big family, my wife, and my daughter: 感谢我的父母给我的爱,支持和自由，让我能无惧未知和困难. 感谢我的大家庭,谢谢你们一直以来的关心和帮助. Thank you, Xue, my dear wife, for your support, trust, and companionship. You are a brave, strong, yet gentle wife taking great care of our small family. The courage inside you inspires me all the time, and it's my great pleasure to explore this bittersweet but unique life journey with you. I believe we are ready for new adventures and to enjoy every moment along the way. Final thanks to my dear, lovely daughter, Lexi, for bringing love, trust, and joy into my life. As the most significant change ever in our life, you have brought your mom and me countless surprises and happiness. Watching you grow up every day, from a tiny baby wrapped in a blanket to a curious, energetic, and sweet little girl with her own opinions, we are so proud of you!

Curriculum Vitae

Zhuoran Liu was born in Qinghai, China, in 1990. He obtained his Bachelor's degree in Mathematics from Jilin University in 2012 and Master's degree in Mathematics from Radboud University in 2018. In 2018, he joined the Data Science Group, Institute for Computing and Information Sciences at Radboud University as a junior researcher and later a PhD student, under the supervision of Prof. dr. Martha Larson. His PhD research is about privacy protection in multimedia with a focus on adversarial machine learning. He also works on side-channel analysis in physical security. He has published a number of research papers in conference proceedings, including ICML, NeurIPS, ICLR, NDSS, USENIX Security, CVPR, BMVC, ICMR, UMAP, ICASSP, and MMM. He regularly serves as a reviewer for conferences (ICML, ICLR, NeurIPS, CVPR, ICCV, ECCV, USENIX Security AE) and journals (IEEE TIFS, IEEE TPAMI, IEEE TIP, TMLR) and as a sub-reviewer for conferences (USENIX Security, NDSS, IEEE S&P). He was a Helmholtz visiting researcher at CISPA Helmholtz Center for Information Security in 2024. He was a poster session co-chair of ACM Multimedia 2019 and task co-organizer of Pixel Privacy at MediaEval Workshop 2018-2020. He has received the Top Reviewer Award at NeurIPS 2024. He is currently a postdoctoral researcher at CESCA lab within the Digital Security Group, working on side-channel analysis, physical security, and adversarial machine learning.

9 789465 150727

Radboud Universiteit