TECHNISCHE UNIVERSITÄT DARMSTADT

Abschlussbericht
Teil I: Kurzbericht

Verbundprojekt:	Forschungsbereich Polizei und Nachrichtendienste: Einsatz von KI zur Früherkennung von Straftaten
Akronym:	KISTRA
Teilvorhaben:	Sichere Tiefe Bildverarbeitung
Förderkennzeichen:	13N15343
Laufzeit des Vorhabens:	01.07.2020 bis 31.12.2023
Datum:	15.04.2024
Ausführende Stelle:	Technische Universität Darmstadt Artificial Intelligence and Machine Learning Lab Altes Hauptgebäude, Room 074, Hochschulstrasse 1 64289 Darmstadt
Ansprechpartner:	Prof. Dr. Kristian Kersting kersting@cs.tu-darmstadt.de Tel.: +49 6151 16 24411

Datum: 15. April 2024
Autoren: Dominik Hintersdorf, Lukas Struppek, Kristian Kersting
Institution: Technische Universität Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

Kurzbericht

Im Folgenden werden die Ergebnisse des Forschungsprojektes *Einsatz von KI zur Früherkennung von Straftaten (KISTRA)*, Teilvorhaben *Sichere Tiefe Bildverarbeitung* zusammenfassend dargestellt.

1.1 Aufgabenstellung

Im Rahmen von AP3 wurden Angriffe auf die Privatsphäre von Modellen bzw. deren Trainingsdaten untersucht, namentlich Modellinversionsangriffe (Model Inversion Attacks) und Membership Inference Attacks. Modellinversionsangriffe versuchen, Trainingsdaten aus einem bereits trainierten Modell mithilfe von generativen Modellen zu extrahieren bzw. rekonstruieren. Membership Inference Attacks hingegen versuchen aus einer vorhandenen Datenmenge diejenigen Datenpunkte zu identifizieren, die für das Training eines bestimmten Modells verwendet worden sind. Ziel der Arbeitspakete war eine umfassende Untersuchung der Angriffsszenarien hinsichtlich möglicher Einflussfaktoren und Szenarien sowie deren allgemeine Effektivität.

1.2 Ablauf des Vorhabens

Zu Beginn des Vorhabens wurde der Fokus zunächst auf die Untersuchung bestehender Membership Inference Attacks gelegt (AP3.4). Hierzu wurden existierende Angriffe aus der Literatur in verschiedenen Szenarien hinsichtlich ihrer Aussagekraft und Zuverlässigkeit untersucht und beurteilt. Dabei wurden diverse Einflussfaktoren auf den Erfolg der Angriffe untersucht. Es konnte festgestellt werden, dass Angriffe basierend auf den Vorhersagewahrscheinlichkeiten von Neuronalen Netzen dazu tendieren, eine hohe Rate an falsch-positiven Vorhersagen zu generieren. Dies bedeutet, dass eine große Anzahl an ungesehenen Daten fälschlicherweise als Teil der Trainingsdaten identifiziert werden. Grund hierfür ist die Überkonfidenz tiefer neuronaler Netze bei ihren Vorhersagen. Eine verlässliche Vorhersage der Zugehörigkeit von Trainingsdaten ist daher unter realistischen Gesichtspunkten häufig kaum möglich. Weiterhin konnte festgestellt werden, dass die Kalibrierung von Modellen, d.h. eine Anpassung der Konfidenz der Modelle hinsichtlich der tatsächlichen Vorhersagegenauigkeit, zu einer Stärkung der Angriffe beiträgt. Erlangen Modelle die Fähigkeit zu erkennen, wann sie eine Eingabe nicht zuverlässig verarbeiten können, so werden Membership Inference Attacks auf diesen Modellen begünstigt. Eine höhere Konfidenz eines Modells in seine Vorhersage erlaubt zugleich Rückschlüsse, ob das Modell eine gewisse Eingabe bereits während des Trainings gesehen hat. Im weiteren Verlauf des Projekts wurde die Betrachtung von Membership Inference Attacks erweitert auf multimodale Modelle und der Frage, ob ein Modell bereits Daten einer gewissen Person während des Trainings gesehen hat. In diesem Setting konnte mit sehr hoher Genauigkeit die Zugehörigkeit zu den Trainingsdaten vorhergesagt werden. Eine darauf aufbauende Arbeit liefert eine wirksame Verteidigungsstrategie basierend auf einem Fine-Tuning Ansatz, um das Modell anzupassen und ungewünschte Namen und Bezeichner herauszulöschen.

Nach der initialen Untersuchung von Membership Inference Attacks wurde in Kooperation mit der ZITiS die Untersuchung von Modellinversionsangriffen gestartet. Hierzu wurde zunächst ein neuer Angriffsalgorithmus entwickelt, um Limitierungen bestehender Angriffe aufzuheben. Der entwickelte Angriff

Datum: 15. April 2024
Autoren: Dominik Hintersdorf, Lukas Struppek, Kristian Kersting
Institution: Technische Universität Darmstadt

erlaubt eine flexiblere Untersuchung der Extrahierung von Trainingsdaten, ohne dass für jedes neue Zielmodell ein separates generatives Modell trainiert werden muss. Im Vergleich mit existierenden Ansätzen liefert der entwickelte Ansatz deutlich bessere und höherwertige Resultate, sodass eine starke Baseline zur Untersuchung vorliegt. Im Rahmen der Untersuchung wurden mögliche Einflussfaktoren auf den Erfolg der Angriffe und die Extrahierung von Trainingsdaten untersucht. Hierzu zählt insbesondere der Einfluss von Modell Regularisierung während des Trainings mithilfe von sogenanntem Label Smoothing. Es konnte gezeigt werden, dass durch den Einsatz von Label Smoothing, welches die Performance und die Kalibrierung eines Modells verbessert, zugleich der Anteil an extrahierbaren Informationen eines Modells steigt. Dieser Effekt lässt sich jedoch umkehren, sodass durch die Anwendung von Label Smoothing mit einem negativen Smoothing Faktor der Anteil an extrahierbaren Informationen reduzieren lässt. Im Vergleich mit existierenden Verteidigungsmaßnahmen liefert der vorgestellte Ansatz einen deutlich besseren Trade-Off zwischen der Performance und der Datenextrahierbarkeit eines Modells.

Gegen Ende des Projekts wurde der Fokus der Untersuchungen verstärkt auf multi-modale Modelle gelegt, um den aktuellen KI-Trends gerecht zu werden. In diesem Rahmen wurde zunächst die Robustheit von Modellen hinsichtlich der Kodierung von Eingabedaten untersucht mit dem Ergebnis, dass multi-modale Modelle durch die Verwendung nicht-lateinischer Zeichen eine starke Verzerrung hinsichtlich der kulturellen Repräsentation erfahren. In einer weiteren Arbeit konnte gezeigt werden, dass die Verwendung von öffentlich verfügbaren, vortrainierten Modellen ein großes Sicherheitsrisiko hinsichtlich möglicher Modellmanipulationen aufweisen und geheime Funktionalitäten enthalten können, die zu ernsthaften Sicherheitsschwächen führen können.

Weitere Untersuchungen befassten sich mit möglichen Verteidigungen gegenüber Inferenzangriffen auf multi-modale Modelle. Hierzu wurde basierend auf unseren vorangegangenen Untersuchungen zur Extrahierung sensitiver Information eine Verteidigungsstrategie entwickelt, die es ermöglicht, einzelne Konzepte aus dem Modell zu löschen, beispielsweise die Assoziation von Namen mit den zugehörigen Bildern.

TECHNISCHE UNIVERSITÄT DARMSTADT

Abschlussbericht
Teil II: Eingehende Darstellung

Verbundprojekt:	Forschungsbereich Polizei und Nachrichtendienste: Einsatz von KI zur Früherkennung von Straftaten
Akronym:	KISTRA
Teilvorhaben:	Sichere Tiefe Bildverarbeitung (SiTiBi)
Förderkennzeichen:	13N15343
Laufzeit des Vorhabens:	01.07.2020 bis 31.12.2023
Datum:	15.04.2024
Ausführende Stelle:	Technische Universität Darmstadt Artificial Intelligence and Machine Learning Lab Altes Hauptgebäude, Room 074, Hochschulstrasse 1 64289 Darmstadt
Ansprechpartner:	Prof. Dr. Kristian Kersting kersting@cs.tu-darmstadt.de Tel.: +49 6151 16 24411

TECHNISCHE UNIVERSITÄT DARMSTADT

Inhaltsverzeichnis

1 Durchgeführte Arbeiten

Im Folgenden werden die durchgeführten Forschungsarbeiten im Vergleich zur ursprünglichen Vorhabenbeschreibung ausführlich dargestellt.

Die durchgeführten Arbeiten erfüllen alle in der Vorhabensbeschreibung genannten Ziele und Anforderungen. Um dem Aufkommen von multimodalen und generativen KI-Modellen sowie der im allgemeinen schnellen Entwicklungen Rechnung zu tragen, wurden zusätzliche Arbeiten im Rahmen der Robustheit und Stabilität aktueller KI-Modelle durchgeführt. Im Folgenden werden jeweils die in der Teilvorhabenbeschreibung definierten Ziele benannt sowie die Ergebnisse der Arbeiten in diesem Kontext vorgestellt. Aufgrund des begrenzten Umfangs dieses Berichtes wird für eine umfassendere Darstellung der Resultate auf die zugehörigen Publikationen verwiesen, welche alle frei zugänglich im Internet verfügbar sind. Zudem sind zugehörige Demonstratoren zusammen mit dem Quellcode öffentlich verfügbar zur Reproduzierbarkeit sowie Erweiterbarkeit der durchgeführten Untersuchungen. Die Verweise auf den Quellcode sind in den entsprechenden Publikationen vorhanden.

1.1 AP 3.3: Modellinversion- und stabilität

Im Rahmen von AP3.3 wurden diverse Untersuchung von Modellinversionsangriffen durchgeführt. Die ausführlichen Ergebnisse sind in den Publikationen *(Struppek et al. "Plug & Play Attacks: Towards Robust and Flexible Model Inversion". ICML 2022)* und *(Struppek et al., "Be Careful What You Smooth For: Label Smoothing Can Be a Privacy Shield but Also a Catalyst for Model Inversion Attacks". ICLR 2024)* publiziert. Im Folgenden werden die wichtigsten Erkenntnisse der beiden Arbeiten sowie darüber hinaus vorgestellt.

1.1.1 Ziele von AP 3.3: Modellinversion- und stabilität

AP 3.3 verfolgt das Ziel, die technischen Voraussetzungen für polizeiliche Früherkennung unter dem Einsatz von KI zu überprüfen. Es wird in bestimmten Fällen notwendig sein, sensible Daten zum Training zu verwenden. Deshalb ist eine nähere Untersuchung der technischen Voraussetzungen, welche hierfür einzuhalten sind, angebracht, um die geltenden Bestimmungen (etwa in Bezug auf Geheimschutz) einzuhalten. Zu diesem Zweck werden Angriffsmethoden auf gängige Architekturen betrachtet. Bei Sichere Tiefe Bildverarbeitung (**SiTiBi**) soll der Fokus auf den Aspekt der „Generative Model Inversion Attack" gelegt werden.

1.1.2 Entwicklung eines neuen Modellinversionsangriffs

Im Bereich der tiefen neuronalen Netze werden zwar ständig neue Maßstäbe gesetzt, aber die kritischen Fragen des Datenschutzes und der Sicherheit bleiben vergleichsweise unbeachtet. Benutzer gehen oft fälschlicherweise davon aus, dass die während des Modelltrainings gelernten Informationen sicher in den Gewichten des Modells gekapselt bleiben. Dieser Irrglaube stellt jedoch ein erhebliches Risiko dar,

insbesondere in Szenarien wie der Gesichtserkennung zur Entsperrung von Mobilgeräten oder der Zahlungsfreigabe, bei denen die Weitergabe von Modellinformationen zu schwerwiegenden Verletzungen der Privatsphäre und der Sicherheit des Einzelnen führen kann. Bestehende Modellinversionsangriffe (Model Inversion Attacks) stützen sich überwiegend auf Generative Adversarial Networks (GANs), leiden jedoch unter verschiedenen Einschränkungen, darunter die Anfälligkeit für Verteilungsänderungen, die Abhängigkeit von bestimmten Zielmodellen und einer limitierten Evaluation auf Bilder mit geringer Auflösung. Es besteht daher ein dringender Bedarf an der Entwicklung effizienterer und robusterer Angriffstechniken zur verlässlichen Evaluation von Modellen zu forschen. Im Rahmen von KISTRA haben wir an der TUD, in Kooperation mit der ZITiS, daher die sogenannten "Plug & Play-Attacks" entwickelt mit dem Ziel, die genannten Einschränkungen bestehender Angriffe zu überwinden und eine umfassendere Evaluation von Inversionsangriffen zu ermöglichen.

Der entwickelte Angriff erweitert bestehende Angriffstechniken hinsichtlich verschiedener Kriterien. Um das Problem des Verschwindens von Gradienten, sogenannter Vanishing Gradients, während der Optimierung anzugehen, verwenden wir eine neuartige Loss-Funktion basierend auf der Poincaré-Distanz anstelle des üblichen Cross-Entropy Losses vor. Diese innovative Loss-Funktion hilft, den Optimierungsprozess zu stabilisieren und eine effektive Modellinversion zu gewährleisten. Zusätzlich integrieren wir zufällige Bild-Transformationen in den Optimierungsprozess, um Overfitting vorzubeugen und die Robustheit der extrahierten Merkmale zu verbessern. Durch die Einführung von Variabilität während des Inversionsprozesses können die Angriffe vielfältigere und realistischere synthetische Bilder generieren.

Aus einer Menge an Angriffsresultaten wird mithilfe eines neuartigen Auswahlprozesses aus den generierten Bildern eine Teilmenge an robusten Resultaten herausgefiltert. Durch die Priorisierung von Bildern basierend auf ihrer Robustheit stellt der Ansatz sicher, dass die generierten Bilder die Merkmale der Zielklasse aus den privaten Trainingsdaten des Modells verlässlich widerspiegeln. Im Gegensatz zu früheren Ansätzen, die stark von spezifischen Zielmodellen und Bildvorlagen abhängig waren, zielen die Plug & Play-Angriffe darauf ab, die Abhängigkeiten zwischen dem generativen Modell (GAN) und den Zielmodellen zu lockern. Diese erhöhte Flexibilität ermöglicht eine effizientere und vielseitigere Modellumkehr in verschiedenen Szenarien und Datensätzen. Hierfür verwendet der Ansatz öffentlich verfügbare, vortrainierte GANs für die Angriffe, was den Bedarf an zusätzlichem Training und Rechenressourcen reduziert. Diese Nutzung vortrainierter Modelle optimiert den Umkehrprozess und erhöht dessen Praktikabilität.

Diese Kernkomponenten tragen gemeinsam zur Wirksamkeit, Robustheit und Flexibilität der entwickelten *Plug & Play Model Inversion Attacks* bei und machen sie zu einem bedeutenden Fortschritt auf dem Gebiet der Model Inversion Attacks auf tiefe neuronale Netzwerke.

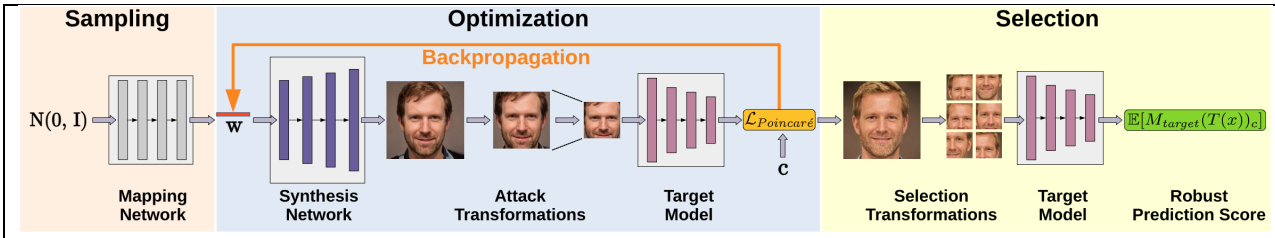


Abbildung 1: Überblick über die Angriffspipeline unserer Plug & Play Angriffe. Zunächst werden initiale latente Vektoren gesampled, aus welchen das GAN entsprechende Bilder generieren kann. Diese latenten Vektoren werden iterativ angepasst mithilfe der Poincaré Loss Funktion. Aus den optimierten Bildern wird zuletzt eine Teilmenge an robusten Bildern als Angriffsresultate selektiert.

1.1.3 Untersuchung der Extrahierbarkeit von Trainingsdaten

Mithilfe der vorgestellten *Plug & Play Angriffe* wurden diverse Einflussfaktoren auf die Extrahierbarkeit von Trainingsdaten. Hierzu wurde das Face Recognition Setting gewählt, das de-faktor Standard Setting zur Evaluation von Modellinversionsangriffen in der Literatur. Die Bewertung der Angriffsresultate basiert auf verschiedenen Metriken:

- Angriffsgenauigkeit (Acc@1): Beschreibt den Anteil an Angriffsresultaten, die von einem separaten Evaluationsmodell korrekt klassifiziert werden. Höhere Werte weisen auf stärkere Angriffe hin.
- FaceNet Distance: Misst die visuelle Ähnlichkeit zwischen Angriffsresultaten und Bildern aus den Trainingsdaten. Kleinere Distanzen weisen auf stärkere Angriffe hin.
- Evaluation Distance: Misst ebenfalls die visuelle Ähnlichkeit. Auch hier weisen kleinere Distanzen auf stärkere Angriffe hin.
- Frechet Inception Distance (FID): Misst die Verteilungsähnlichkeiten zwischen den Trainingsdaten und den Angriffsresultaten. Kleinere Werte weisen auf stärkere Angriffe hin.

Im Rahmen von KISTRA wurden diverse Einflussfaktoren auf den Erfolg von Modellinversionsangriffen untersucht. Im Folgenden wird ein Teil davon vorgestellt, eine ausführlichere Untersuchung der genannten sowie zusätzlicher Einflussfaktoren findet sich in unseren Publikationen.

1.) **Größe des Trainingsdatensatzes:** Es wurden zwei Datensätze mit verschiedenen Größen untersucht, einmal FaceScrub mit 530 Identitäten und CelebA mit 1000 Identitäten. Trainiert wurde jeweils ein ResNet-152. Die Angriffe liefern auf beiden Datensätzen sehr gute Ergebnisse, trotz der unterschiedlichen Anzahl an Klassen. Zugleich sind die Angriffe auf dem FaceScrub Datensatz stärker, allerdings bietet dieser auch qualitativ hochwertigere Daten. Unsere Ergebnisse zeigen, dass die Datensatz Größe keine allzu große Rolle auf den Erfolg der Angriffe hat, solange eine gewisse Mindestanzahl an Trainingsdaten vorhanden ist. Für eine kleine Anzahl an Trainingsdaten sinkt die Performanz des Modells und in gleichem Maße die Qualität der extrahierten Trainingsdaten.

Datensatz	↑ Acc@1	↓ FaceNet Distance	↓ Evaluation Distance	↓ FID
FaceScrub	92.73%	0.7163	123.25	46.69
CelebA	80.61%	0.7362	312.58	40.43

Tabelle 1: Die Attack-Metriken unserer entwickelten Plug & Play Attacke auf verschiedenen Datensätzen. Attackiert wurde ein ResNet-152.

2.) Einfluss der Modellarchitektur: Es wurden Modelle diverser Modellarchitekturen trainiert, insbesondere Modelle basierend auf ResNets, DenseNets, ResNeSt und ResNeXt. Hierzu wurden erneut verschiedene Modelle auf FaceScrub und CelebA trainiert und die Wirksamkeit der Angriffe untersucht. Insgesamt konnten keine signifikanten Unterschiede zwischen den Modellarchitekturen und Größen festgestellt werden.

Architektur	↑ Test Acc	↑ Acc@1	↓ FaceNet Distance	↓ Evaluation Distance	↓ FID
ResNet-18	94.22%	95.48%	0.6867	112.24	45.82
ResNet-152	93.74%	92.73%	0.7163	123.25	46.69
DenseNet-161	94.22%	91.49%	0.7083	123.41	46.92
ResNeSt-101	95.38%	93.95%	0.7199	119.79	46.30
ResNeXt-50	95.25%	94.97%	0.6977	119.51	41.61

Tabelle 2: Angriffsergebnisse der Plug & Play Attacke auf verschiedenen Architekturen. Wie zu sehen ist, hat die Wahl der Architektur nur geringen Einfluss auf den Erflog der Attacke.

3.) Aufgabe der Modelle: Im Rahmen von KISTRA wurden Klassifikationsmodelle hinsichtlich ihrer Privatsphäre untersucht. Dieses stellt das übliche Setting für Model Inversion Attacks dar. Eine Erweiterung der entwickelten Plug & Play Angriffe auf andere Modelle, z.B., Modelle basierend auf Ähnlichkeitslernen, ist möglich. Eine solche Anpassung unseres Angriffes für self-supervised Learning Methoden wurde bei Huang et al. (2024) untersucht.

4.) Verteilung individueller Klassen im Trainingsdatensatz: Die Verteilung einzelner Klassen hat in unseren Experimenten keinen signifikanten Einfluss auf die Ergebnisse gehabt. Unsere Angriffe haben sowohl unabhängig vom Geschlecht, dem Alter sowie der Anzahl und Varianz der Samples zuverlässig charakteristische Merkmale einzelner Identitäten rekonstruieren können.

5.) Einfluss des Datenformats: Im Rahmen unserer Experimente konnte kein signifikanter Einfluss der Wahl des Datenformates festgestellt werden.

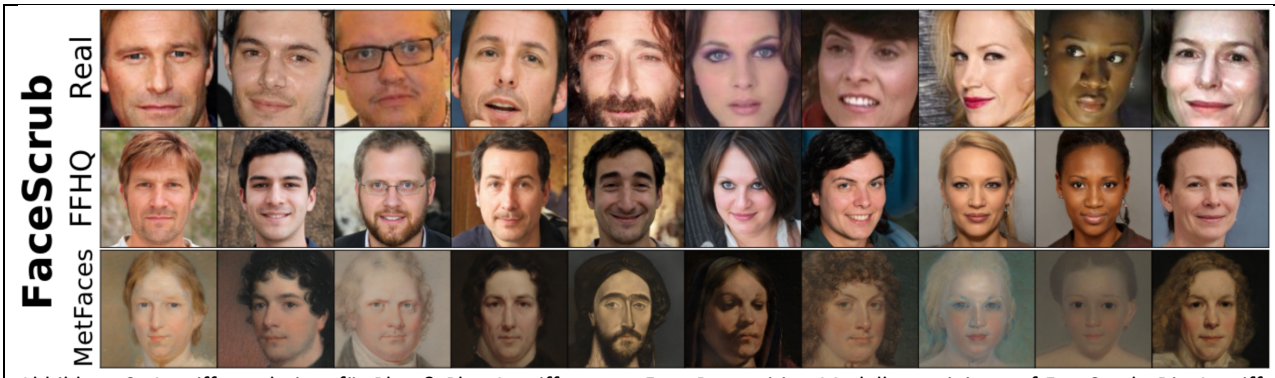
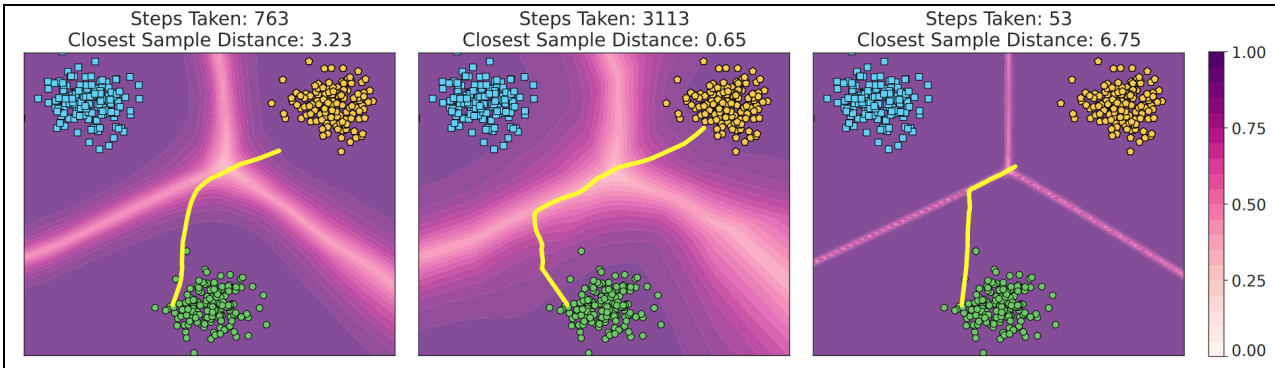


Abbildung 2: Angriffsergebnisse für Plug & Play Angriffe gegen Face Recognition Modellen trainiert auf FaceScrub. Die Angriffe wurden mit StyleGAN2 Modell durchgeführt, die auf Portraitfotos (FFHQ) und Gemälden (MetFaces) trainiert wurden. Die rekonstruierten Bilder extrahieren erfolgreich charakteristische und sensitive Features wie Geschlecht, Haarfarbe, etc. aus den Modellen.

1.1.4 Untersuchung des Einflusses von Regularisierungsmethoden auf die Extrahierbarkeit von Trainingsdaten



(a) Hard Labels ($\alpha = 0.0$) (b) Positive LS ($\alpha = 0.05$) (c) Negative LS ($\alpha = -0.05$)

Abbildung 3: Ein Beispiel zum Einfluss von Label Smoothing auf die Kalibrierung von Modellen und Extrahierbarkeit von Trainingsdaten. Die Hintergrundfarbe illustriert die Konfidenz des Modells, dunklere Farben demonstrieren höhere Konfidenz. Die gelben Linien symbolisieren die Trajektorie eines Modelinversionsangriffs ausgehend von einem Sample der grünen Klasse hinzu zur orangen Klasse. Angriffe auf Modellen trainiert mit positivem Label Smoothing führen den Angriff deutlich näher an die Trainingsdaten heran.

Trainings-Regularisierungsmethoden wie Label Smoothing werden eingesetzt, um eine Überanpassung zu verhindern und die Generalisierungsfähigkeit von Modellen des maschinellen Lernens zu verbessern. Label Smoothing zielt speziell auf das Problem des übermäßigen Vertrauens in die Vorhersagen neuronaler Netze ab. Label Smoothing funktioniert, indem die Ziel-Labels während des Trainings modifiziert werden. Anstatt der wahren Klasse eine harte 0 oder 1 zuzuweisen, werden diese Extremwerte durch geglättete Wahrscheinlichkeiten ersetzt. Wenn das wahre Label für eine bestimmte Stichprobe beispielsweise die Klasse "Katze" ist, könnte die Label-Glättung der Klasse "Katze" einen etwas niedrigeren Wert zuweisen (z. B. 0,9) und die verbleibende Wahrscheinlichkeitsmasse auf die anderen Klassen verteilen, anstatt der Klasse "Katze" ein Label von 1 und allen anderen Klassen 0 zuzuweisen.

Der Grundgedanke hinter Label Smoothing ist, dass das Modell daran gehindert wird, zu viel Vertrauen in seine Vorhersagen zu gewinnen. Wenn ein Modell mit herkömmlichen “harten” Bezeichnungen trainiert

wird, kann es lernen, bestimmten Klassen sehr hohe Wahrscheinlichkeiten zuzuweisen, selbst wenn die Beweise, die diese Vorhersagen unterstützen, schwach sind. Dies kann zu einer Überanpassung führen, bei der das Modell bei den Trainingsdaten gut abschneidet, aber bei den ungesehenen Daten schlecht abschneidet. Durch die Verwendung von Label Smoothing wird das Modell dazu angeregt, besser kalibrierte Wahrscheinlichkeitsschätzungen zu produzieren. Dies kann dazu beitragen, die Generalisierung zu verbessern, insbesondere in Szenarien, in denen die Trainingsdaten verrauscht oder begrenzt sind.

Wir haben uns Label Smoothing im Kontext von Modellinversionsangriffen genauer angeschaut und festgestellt, dass diese sowohl die Extrahierbarkeit von Trainingtsdaten begünstigen, wie auch verhindern können. Abbildung 3 illustriert den Einfluss von Label Smoothing auf das Modelltraining. Positives Label Smoothing verbessert die Kalibrierung des Modells und führt dazu, dass dieses hohe Prediction Scores nur für Eingabedaten vorhersagt, die nahe an der Trainingsverteilung liegen. Zugleich führt dies dazu, dass ein Modellinversionangriff leichter die Trainingsdaten rekonstruieren kann, da diese in einem Hochkonfidenten Raum liegen. Umgekehrt führt ein Training mit einem negativen Smoothing Faktor zu einem überkonfidenten Modell, welches für nahezu alle Inputs hohe Predictions Scores vorhersagt. Zugleich liefert in diesem Szenario ein Modellinversionsangriff keine sinnvollen Ergebnisse, sondern bleibt dicht an der Entscheidungsgrenze stehen. Eine Rekonstruktion von Trainingsdaten ist hier nicht möglich.

Übertragen auf das Face Recogntion Setting ergibt sich ein ähnliches Bild: Die Angriffe sind stärker auf Modellen, welche mit positivem Label Smoothing trainiert wurden. In diesem Fall können charakteristische Features der einzelnen Klassen zuverlässiger extrahiert werden. Umgekehrt schützt negatives Label Smoothing des Modells vor dem Angriff und reduziert den Erfolg der Angriffe merklich. Insbesondere im Vergleich mit existierenden Verteidigungsmaßnahmen erzielt negatives Label Smoothing einen deutlich besseren Utility-Privacy Tradeoff.

Modell	**↑ Test Acc**	**↑ Acc@1**	**↓ FaceNet Distance**	**↓ Evaluation Distance**	**↓ FID**
Standard	94.9%	94.3%	0.71	124.30	40.88
Positive Label Smoothing	97.4%	95.20	0.6343	107.36	43.33
Negative Label Smoothing	91.5%	14.34%	1.2320	239.02	59.38

Tabelle 3: Einfluss von positivem und negativem Label Smoothing auf den Erfolg der Model Inversion Attacke.

Abbildung 4: Angriffsresultate für Modelle trainiert ohne, positivem und negativem Label Smoothing. Während die Resultate für Modelle mit positivem Label Smoothing sich verbessern, d.h., charakteristische Features besser extrahiert werden können, kehrt negatives Label Smoothing diesen Effekt um und verteidigt das Modell gegen den Angriff.

1.1.5 Weitere Untersuchungen der Modellstabilität von tiefen neuronalen Netzen gegenüber Manipulationen

Weitere Untersuchungen im Rahmen von KISTRA haben sich mit der allgemeinen Modellstabilität von tiefen neuronalen Netzen hinsichtlich verschiedener Angriffsvektoren beschäftigt. Zu nennen ist hier die Untersuchung der Anfälligkeit von Text Encoder Modellen bezüglich Backdoor Attacks. Solche Angriffe haben zum Ziel, geheime Funktionalitäten in Modelle einzubauen und bei der Aktivierung mithilfe eines Triggers diese auszuführen. Es konnte gezeigt werden, dass innerhalb weniger Minuten sich solche geheimen Funktionalitäten in vortrainierte Modelle einbauen lassen, ohne die ursprüngliche Funktionalität zu stören. Im Kontext von textgesteuerter Bildgenerierung kann dies dazu führen, dass unerwünschte Bildinhalte erzeugt werden können, z.B. Gewalt- oder Propagandamaterial. Aber auch in anderen Anwendungen, z.B. Image Retrieval haben sich solche Angriffe als sehr zuverlässig erwiesen. Eine umfangreiche Darstellung der Resultate wird in *(Struppek et al., "Rickrolling the Artist: Injecting Backdoors into Text Encoders for Text-to-Image Synthesis". ICCV 2023)* geliefert.

Eine andere Untersuchung beschäftigt sich mit der Stabilität von Perceptual Hashing Verfahren. Diese Modelle berechnen für jeden Input einen digitalen Fingerabdruck zur Re-identifizierung von bekanntem Bildmaterial. Beispielsweise in der Strafverfolgung können solche Verfahren verwendet werden, um bekanntes strafbares Material auf Endgeräten zu detektieren, ohne dass dieses Material explizit außerhalb des Gerätes untersucht werden muss. Grundsätzlich wird ein tiefes neuronales Netz verwendet, um für jeden Input charakteristische Features zu extrahieren und darauf basierend einen Fingerabdruck zu berechnen. Im Rahmen unserer Untersuchungen konnte gezeigt werden, dass solche Systeme wenig robust sind und sehr anfällig auf kleinste Bildänderungen reagieren. Durch die Manipulation weniger Pixel in einem Bild kann die Detektion verhindert werden. Weiterhin können Bilder dahingehend manipuliert werden, dass die berechneten Fingerabdrücke einen nahezu beliebigen Wert annehmen können. Da entsprechende Systeme vielfach in verschiedenen Regierungen, inkl. der Europäischen Union, diskutiert werden, liefern unsere Ergebnisse wichtige Entscheidungsgrundlagen zur Einführung eines solchen Systems. Umfassende Details zur

Untersuchung sind verfügbar in der Publikation *(Struppek et al., "Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash". FAccT 2021)*.

1.2 AP 3.4: Datenzuordnung in Trainingsdaten

Im Rahmen von AP3.4 wurden diverse Untersuchung von Membership Inference Attacks durchgeführt. Die ausführlichen Ergebnisse sind in den Publikationen *(Hintersdorf et al. "To Trust or Not To Trust Prediction Scores for Membership Inference Attacks". IJCAI 2022)* und *(Hintersdorf et al., "Does CLIP Know My Face?". JAIR 2024)* publiziert. Im Folgenden werden die wichtigsten Erkenntnisse der Arbeiten vorgestellt.

1.2.1 Ziele von AP 3.4: Datenzuordnung in Trainingsdaten

In AP 3.4 setzen wir das Ziel aus AP 3.3, die technischen Voraussetzungen für polizeiliche Früherkennung unter dem Einsatz von KI zu überprüfen, fort. Im Gegensatz zu AP 3.3 werden in AP 4.3 zu diesem Zweck „Membership Inference" auf gängige Architekturen betrachtet werden.

1.2.2 Untersuchung des Aussagegehalts existierende Membership Inference Angriffe

In einer Zeit, in der Deep-Learning Modelle in verschiedenen Bereichen eingesetzt werden und große Mengen an Daten verarbeiten, ist es entscheidend sicherzustellen, dass diese Daten ordnungsgemäß geschützt und nicht missbraucht werden. Bei sogenannten Membership Inference Attacks versucht der Angreifer herauszufinden, welche Daten zum Trainieren eines gegebenen Netzes verwendet wurden. Um zu verstehen, warum eine erfolgreiche Membership Inference Attacke eine Verletzung der Privatsphäre zur Folge hat, stelle man sich vor, dass ein Modell auf den Patientendaten eines bestimmten Krankenhauses trainiert wurde. Der Angreifer hat Zugriff auf das trainierte Modell und hat Datenpunkte, für welche er entscheiden möchte, ob diese zum Trainieren des Modells verwendet wurden oder nicht. Wenn der Angreifer nun erfolgreich eine Membership Inference Attacke durchführt, wird nicht nur preisgegeben, dass diese Daten zum Trainieren des Modells verwendet wurde, sondern gleichzeitig auch, dass diese Person in der Vergangenheit Patient in diesem Krankenhaus war und womöglich sogar welche Krankheit diese Person hatte. Konfidenzbasierte Membership Inference Attacken verwenden die Konfidenzwerte der neuronalen Netze, um zwischen Trainingsdaten von Nicht-Trainingsdaten zu unterscheiden. Die Intuition hierbei ist, dass Beispiele, die zum Trainieren verwendet wurden, einen höheren Konfidenzwert haben als Beispiele, die nicht zum Trainieren verwendet wurden und ein Angreifer somit herausfinden kann, welche Daten zum Trainieren verwendet wurden.

Im Zuge von KISTRA haben wir verschiedene Membership Inference Attacken analysiert und die Resultate unserer Untersuchung im Folgenden zusammengefasst. Weiterführende und detailliertere Resultate finden sich in unserer Publikation.

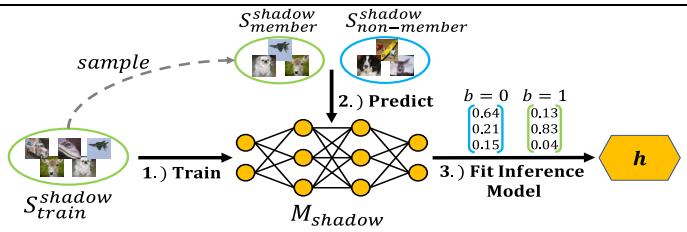


Abbildung 5: Vorbereitungsprozess einer Membership Inference Attacke. Der Angreifer benutzt einen Teil des Shadow Sets zum Trainieren eines Shadow Modells, um daraufhin ein Modell h für die Attacke zu trainieren.

Um einen best-möglichen Angreifer zu simulieren und somit eine obere Schranke für die Angreifbarkeit des Modells abzuschätzen wird üblicherweise angenommen, dass der Angreifer Zugriff auf Daten aus der gleichen Verteilung wie die Trainingsdaten hat. Dieser Datensatz wird in der Regel „Shadow Datensatz“ genannt. Diese Daten verwendet der Angreifer, um die Attacke zu kalibrieren. Hierzu wird, wie in Abbildung 5 zu sehen, aus dem Shadow Datensatz ein Shadow Trainingsset entnommen, mit dem der Angreifer ein eigenes „Shadow Modell“ trainiert, welches nach dem Training dem anzugreifenden Zielmodell sehr ähnlich ist. Da für dieses Modell bekannt ist, welche Trainingspunkte zum Trainieren des Modells verwendet wurden, kann der Angreifer ein Attack Modell *h* trainieren und die Attacke somit kalibrieren, um entsprechende Grenzwerte für die Identifizierung von Trainingsdaten festzulegen. Um die Membership Inference Attacke durchzuführen wird, wie in Abbildung 6 zu sehen, das Attack Model *h* verwendet, um anhand der Ausgabe des Zielnetzes zu klassifizieren, ob die Datenpunkte zum Trainieren verwendet wurden oder nicht.

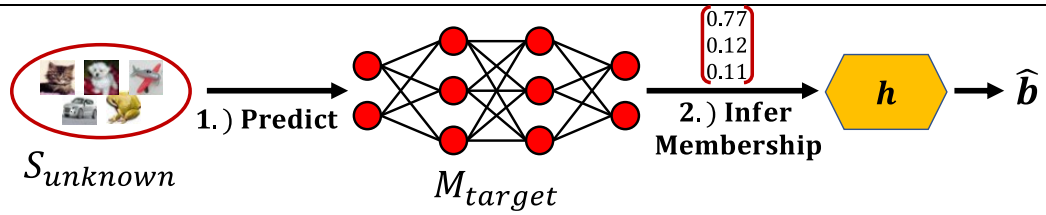


Abbildung 6: Durchführung einer Membership Inference Attacke auf ein Zielnetz. Der Angreifer benutzt das vorher trainierte Angriffsmodell *h* um die Attacke durchzuführen und eine Aussage über die Zugehörigkeit der Datenpunkte zu den Trainingsdaten zu treffen.

In unserer Arbeit haben wir drei der aktuellen Konfidenz-basierten Membership Inference Attacken auf verschiedenen Zielarchitekturen untersucht. Die untersuchten Attacken verwenden die Top-3 Konfidenzwerte, die maximale Konfidenz und die Entropie der Ausgabe des Zielnetzes, um eine Vorhersage über die Zugehörigkeit zu den Trainingsdaten zu treffen. Als Zielarchitekturen wurden ein simples Convolutional Neural Network (SalemCNN), ein ResNet-18 und ein EfficientNetB0 untersucht. Im Rahmen von KISTRA wurden verschiedene Einflussfaktoren auf den Erfolg der Membership Inference Attacken untersucht. Im Folgenden wird ein Teil davon vorgestellt, eine ausführlichere Untersuchung der genannten sowie zusätzlicher Einflussfaktoren findet sich in unseren Publikationen.

1.) **<u>Robustheit der Attacken:</u>** Wie in Tabelle 4 und Tabelle 5 zu sehen ist, erreichen alle drei der untersuchten Attacken ein zu vorheriges Arbeiten vergleichbar hohen AUROC Wert. Dies gilt sowohl für die Zielmodelle, welche auf CIFAR-10 trainiert wurden, als auch für das ResNet-50 welches auf Stanford Dogs trainiert wurde. Allerdings ist zugleich bei allen untersuchten Modellen und Datensätzen die Falsch-

Positiv-Rate (FPR) der Attacken hoch. Dies zeigt, dass die Angreifbarkeit von Membership Inference Attacken in vorherigen Arbeiten deutlich überschätzt wurde. Durch die hohe FPR könnte ein Angreifer schwer zwischen einer true-positive und einer false-positive Vorhersage unterscheiden, was zeigt, dass die Effektivität der Angriffe in einem realistischen Scenario deutlich geringer ist als zuvor angenommen.

	SalemCNN	**ResNet-18**	**EfficientNetB0**
↑ Train Accuracy	100%	100%	99.03%
↑ Test Accuracy	59.04%	69.38%	71.06%
↑ Entropy AUROC	70.94%	76.50%	66.67%
↓ Entropy FPR	46.60%	44.76%	50.35%
↑ Max. Score AUROC	72.03%	77.50%	66.58%
↓ Max. Score FPR	46.40%	44.76%	50%
↑ Top-3 Scores AUROC	71.57%	77.14%	66.61%
↓ Top-3 Scores FPR	60.04%	55.52%	53.40%

Tabelle 44: Training Accuracy, Test Accuracy, die Area Under the ROC (AUROC) und die False-Positive Rate (FPR) für 3 verschiedene Konfidenz-basierte Attacken auf 3 Zielarchitekturen welche auf CIFAR-10 trainiert wurden.

	Train Accuracy	**Test Accuracy**	**Entropy AUROC**	**Entropy FPR**	**Max. Score AUROC**	**Max. Score FPR**	**Top-3 Scores AUROC**	**Top-3 Scores FPR**
ResNet-50	98.48%	59.69%	78.22%	39.36%	78.12%	38.87%	78.29%	42.35%

Tabelle 55: Training Accuracy, Test Accuracy, die Area Under the ROC (AUROC) und die False-Positive Rate (FPR) für 3 verschiedene Konfidenz-basierte Attacken auf einem ResNet-50 welches auf Stanford Dogs trainiert wurde.

2.) **Positiver Einfluss von Überkonfidenz auf Attacken:** Tiefe neuronale Netze sind oft überkonfident, was dazu führt, dass die Ausgabewerte der Modelle nicht der wahren Wahrscheinlichkeit der korrekten Vorhersage entsprechen. Stattdessen sind die Ausgabewerte für einzelne Klassen zu hoch. Da Membership Inference Attacken die Ausgabewerte verwenden, um eine Aussage über die Zugehörigkeit zu den Trainingsdaten zu treffen, ist dies ein Grund für die hohen falsch-positiv Raten. Wie in Tabelle 6 zu sehen ist, sind die durchschnittlichen maximalen Vorhersagewerte (MMPS) von falsch-positiven Vorhersagen deutlich höher als die von richtig-negativ vorhergesagten Membership Inference Attacken. Dies lässt den Schluss zu, dass die Überkonfidenz der Netze einen negativen Einfluss auf konfidenzbasierte Membership Inference Attacken hat. Zudem scheinen die Attacken ihre Vorhersage ausschließlich auf dem maximalen Konfidenzwert zu basieren, da es keinen wahrnehmbaren Unterschied zwischen den False-Positive MMPS-Werten der Top-3 Scores und der Max. Scores Attacke gibt.

Dataset	Attacke	False-Positive MMPS	True-Negative MMPS
Stanford Dogs	Entropy	0.9984	0.7565
	Max. Score	0.9985	0.7580
	Top-3 Scores	0.9979	0.7486
Fake Dogs	Entropy	0.9977	0.7700
	Max. Score	0.9979	0.7724
	Top-3 Scores	0.9971	0.7648
AFHQ Cats	Entropy	0.9972	0.7205
	Max. Score	0.9972	0.7208
	Top-3 Scores	0.9959	0.7137

Tabelle 66: Mean Maximum Prediction Scores (MMPS) eines auf Stanford Dogs trainierten ResNet-50 auf verschiedenen Datensätzen. Fake Dogs entspricht synthetisch generierten Hundebildern während AFHQ Cats realen Katzenbildern entsprechen.

3.) **Negativer Einfluss von Kalibrierung auf Attacken:** Um die Überkonfidenz von neuronalen Netzen zu minimieren, werden sogenannte Kalibrierungsmethoden verwendet. In unserer Arbeit haben wir die Auswirkung zwei der gängigsten Verfahren zur Kalibrierung von neuronalen Netzen namens Laplace Approximation (LA) und Label Smoothing (LS) auf Membership Inference Attacken untersucht. Wie in Abbildung 7 zu sehen ist, reduzieren beide Kalibrierungsmethoden die FPR der Attacken deutlich. Die FPR vor der Kalibrierung wird durch transparente Balken dargestellt, während die FPR nach der Kalibrierung durch opake Balken dargestellt wird. Dies lässt darauf schließen, dass durch die Kalibrierung der Modelle die Anfälligkeit gegenüber Membership Inference Attacken steigt. Da kalibrierte Modelle niedrigere Konfidenzwerte bei Datenpunkten produzieren, die stärker von der Verteilung der Trainingsdaten abweichen, wurden durch die Kalibrierung zusätzliche Informationen über die Trainingsdatenverteilung preisgegeben.

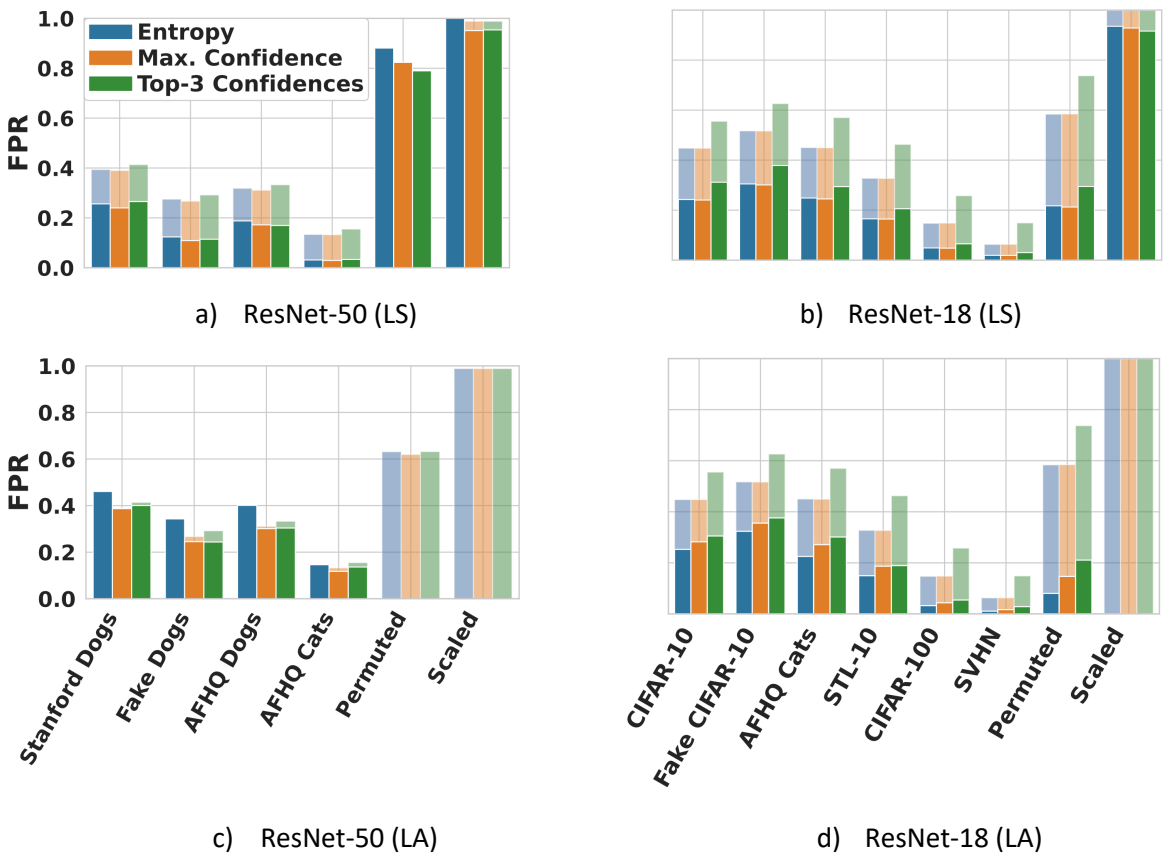
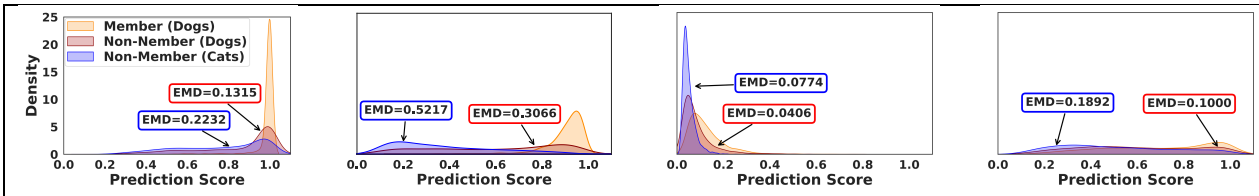


Abbildung 7: False-Positive Rate (FPR) verschiedener Modelle. Das ResNet-50 wurde auf Stanford Dogs trainiert, während das ResNet-18 auf CIFAR-10 trainiert wurde. Die transparenten Balken stellen die Werte vor der Kalibrierung der Modelle dar, während die opaken Balken die FPR-Werte nach der Kalibrierung angeben.

4.) **Gegensätzliche Effekte von Verteidigungen und Kalibrierung:** Während Kalibrierungsmethoden versuchen den Informationsgehalt der Vorhersagewerte eines Modells zu maximieren, versuchen Verteidigungen gegen Membership Inference Attacken den Informationsgehalt zu minimieren. Wir haben Temperature Scaling und L2-Regularisierung als zwei der aktuellen Verteidigungen gegen Membership Inference Attacken untersucht. Wie in Abbildung 8 zu sehen ist, haben Verteidigungen einen gegensätzlichen Effekt zu Kalibrierungsmethoden. Während Kalibrierungsmethoden den Unterschied in den Verteilungen der Vorhersagewerte zwischen Trainings- und Nicht-Trainingsdaten vergrößern, gleichen Verteilungen von Trainings- und Nicht-Trainingsdaten an. Wenn die Verteilungen unterscheidbar sind, sind somit auch die Membership Inference Attacken deutlich effektiver. Andererseits sind die Attacken weitaus weniger effektiv, wenn die Verteilungen von Trainings- und Nicht-Trainingsdaten kaum unterscheidbar sind.



a) ResNet-50 b) ResNet-50 (LS) c) ResNet-50 (Temperature) d) ResNet-50 (L2)

Abbildung 8: Die Verteilung der Vorhersagewerte eines ResNet-50 mit angewendeter Kalibrierung (LS) und zwei angewendeter Verteidigungen (Temperature und L2). Während Kalibrierungsmethoden die Verteilungen stärker differenzierbar machen, gleichen Verteidigungen die Verteilungen der Vorhersagewerte aneinander an.

Im Rahmen von der in KISTRA durchgeführten Arbeiten wurden wie bereits gesehen diverse Einflussfaktoren auf den Erfolg von Membership Inference Attacken untersucht. Im Folgenden wird ein Teil davon erneut zusammengefasst. Eine ausführlichere Untersuchung der genannten sowie zusätzlicher Einflussfaktoren findet sich in unseren Publikationen.

1) **Größe des Trainingsdatensatzes:** Es wurde zwei Datensätze mit unterschiedlicher Größe untersucht. Zum einen wurden Modelle angegriffen, welche auf dem CIFAR-10 Datensatz trainiert wurden. Dieser Datensatz beinhaltet 10 Klassen von verschiedenen Bildern und hat insgesamt 50 000 Trainingsbilder. Als zweiter Datensatz wurde der Stanford Dogs Datensatz untersucht welcher Bilder von 120 verschiedenen Hunderassen beinhaltet und insgesamt 20 580 Trainingsbilder hat. In unseren Experimenten war zu beobachten, dass die Attacken auf dem Stanford Dogs Datensatz effektiver sind und eine höhere AUROC haben als die Modelle, welche auf dem CIFAR-10 Datensatz trainiert wurden. Dies lässt den Schluss zu, dass Modelle, welche auf größeren Datensätze trainiert wurden, tendenziell schwieriger anzugreifen sind. Gleichzeitig hat der Stanford Dogs Datensatz 12-malso viele Klassen wie der CIFAR-10 Datensatz. Dies spricht dafür, dass eine größere Anzahl an Klassen die Angreifbarkeit der Modelle verstärkt.
2) **Verteilung individueller Klassen im Trainingsdatensatz:** Die Verteilung einzelner Klassen hat in unseren Experimenten keinen signifikanten Einfluss auf die Ergebnisse gehabt. Die untersuchten Angriffe haben unabhängig von der Anzahl und Varianz der Samples die gleichen Vorhersagen getroffen.
3) **Einfluss der Modellarchitektur:** Es wurden Modelle diverser Modellarchitekturen trainiert, insbesondere Modelle basierend auf ResNets, klassischen Convolutional Neural Networks und EfficientNets. Hierzu wurden erneut verschiedene Modelle auf dem CIFAR-10 Datensatz trainiert und die Wirksamkeit der Angriffe untersucht. Insgesamt konnten keine signifikanten Unterschiede zwischen den Modellarchitekturen und Modellgrößen festgestellt werden.
4) **Einfluss des Datenformats:** Im Rahmen unserer Experimente konnte kein signifikanter Einfluss der Wahl des Datenformates festgestellt werden.

1.2.3 Identitätsinferenz Angriffe in multimodalen Modellen

Während in klassischen Membership Inference Attacken der Angreifer versucht die Zugehörigkeit eines bestimmten Datenpunktes zu bestimmen, ist oftmals die viel entscheidendere Frage, ob generell Daten einer bestimmten Person zum Trainieren eines Modells verwendet wurden. Daher haben wir zusätzlich zu Klassifikationsmodellen auch Membership Inference Attacken auf einem CLIP (Contrastive Language Image

Pre-Training) Modell untersucht, welches durch Ähnlichkeitslernen trainiert wird. In der durchgeführten Untersuchung haben wir eine neue Attacke namens Identity Inference Attack (IDIA) entwickelt, welche in Abbildung 9 zu sehen ist. Ziel der IDIA ist es herauszufinden, ob Daten einer bestimmten Person verwendet wurden, um ein gegebenes CLIP-Modell zu trainieren.

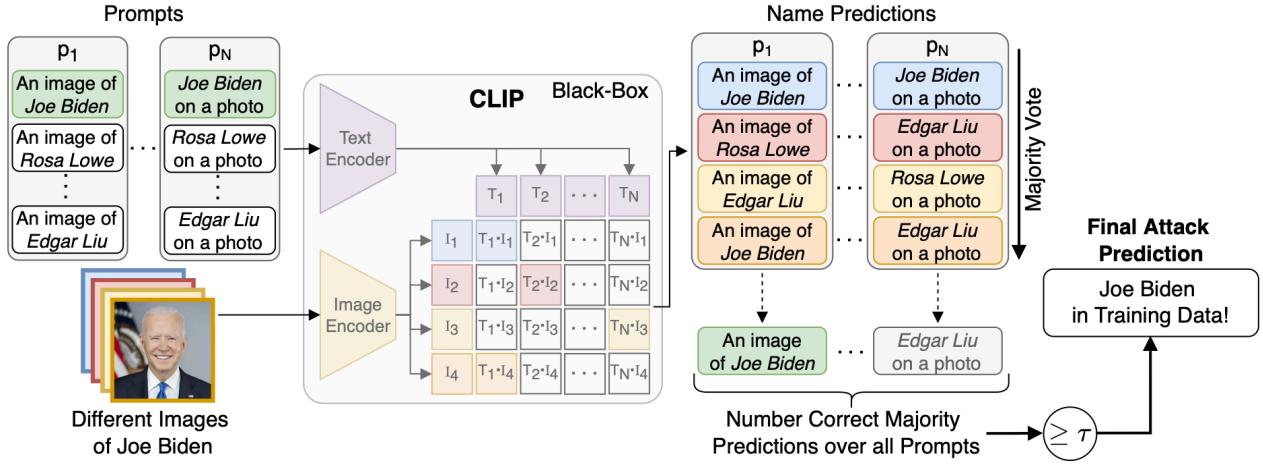
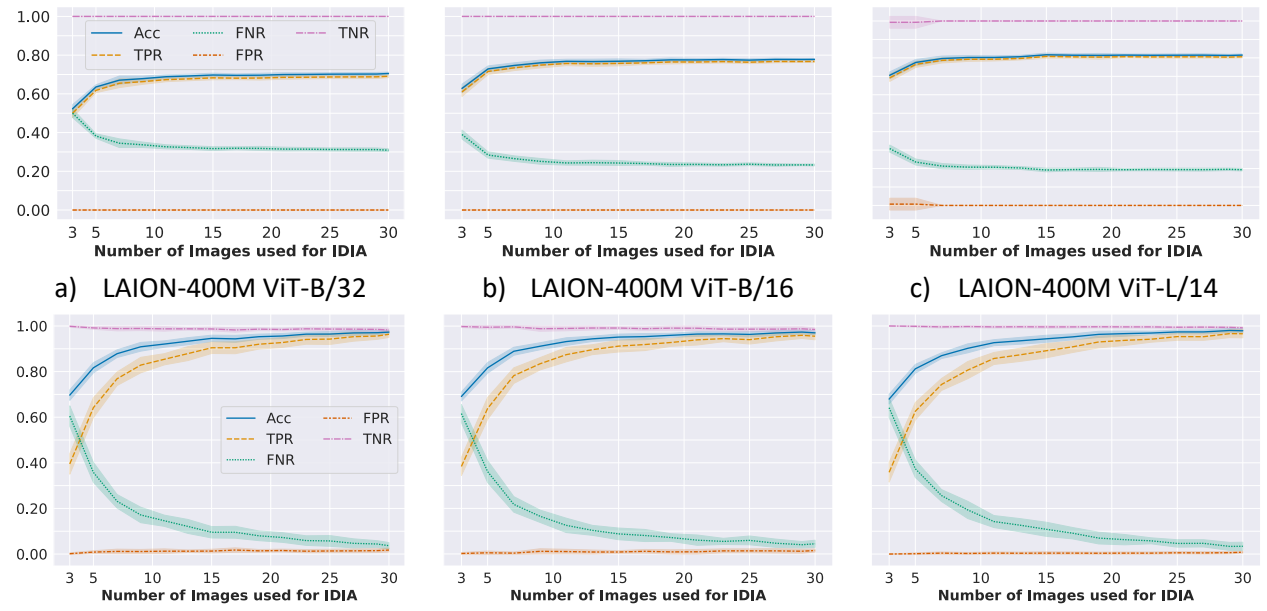


Abbildung 9: Schematische Darstellung unserer Identity Inference Attack. Ausgehend von verschiedenen Bildern und dem Namen einer Person wird CLIP mit den Bildern und mehreren Prompt Templates mit möglichen Namen abgefragt. Nach Erhalt der Vorhersagen wird für jeden der Prompts der Name, der für die meisten Bilder abgeleitet wurde, als vorausgesagter Name verwendet. Wenn die Anzahl der korrekten Vorhersagen über alle Eingabeaufforderungen größer oder gleich Tau ist, wird angenommen, dass die Person in den Trainingsdaten enthalten ist.

Um unsere entwickelte IDIA zu evaluieren wurden mehrere Experimente auf zwei großen Datensätzen durchgeführt. Einer der Datensätze ist der LAION-400M Datensatz, welcher aus 400 Millionen Bild-Text-Paaren besteht. Der zweite Datensatz, der zur Evaluation der Attacke verwendet wurde, ist der Conceptual Captions (CC3M) Datensatz, welcher aus 3 Millionen Bild-Text-Paaren besteht. CLIP besteht aus einem Text- und einem Bild-Encoder. Da es verschiedene CLIP-Modelle unterschiedlicher Größe gibt und es auch möglich ist unterschiedliche Architekturen für den Bilde-Encoder zu verwenden, wurden die Experimente mehrmals mit unterschiedlichen Modellgrößen und Bild-Encoder Architekturen durchgeführt.
Um die Attacke zu evaluieren, wurde der LAION-400M Datensatz nach berühmten Personen durchsucht und analysiert, wie oft diese im Datensatz vorkommen. Anschließend wurden die Personen als Ziel ausgewählt, welche maximal 300-mal im Trainingsdatensatz vorkommen. Da der CC3M Datensatz anonymisiert ist, werden für die Evaluationen auf diesem Datensatz kontrolliert Namen und Bilder von Promis hinzugefügt.

Die Resultate der Attacke auf dem LAION-400M und dem CC3M Datensatz mit unterschiedlicher Anzahl an Bildern, welche für die Attacke verwendet werden, sind in Abbildung 10 zu sehen. Generell ist festzuhalten, dass die IDIA keine falsch-positiven Vorhersagen trifft und gleichzeitig eine hohe True-Positive Rate hat. Dies führt dazu, dass Angreifer ein hohes Vertrauen in die Vorhersagen der Attacke haben können und mit sehr hoher Genauigkeit bestimmen kann, ob Bilder einer Person zum Trainieren des Modells verwendet wurden.

Für weitere und detailliertere Resultate verweisen wir auf unsere zugehörige Publikation (Hintersdorf et al., "*Does CLIP Know My Face?*". JAIR 2024).



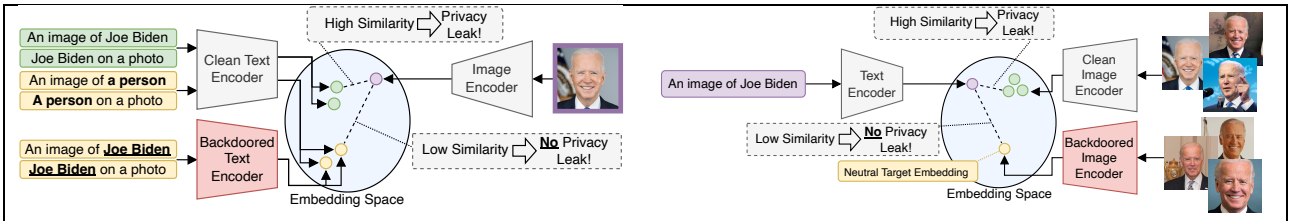
a) LAION-400M ViT-B/32 b) LAION-400M ViT-B/16 c) LAION-400M ViT-L/14

d) CC3M ResNet-50 e) CC3M ResNet-50x4 f) CC3M ViT-B/32

Abbildung 10: Resultate der Identity Inference Attacke auf Modelle welche auf LAION-400M und CC3M trainiert wurden. Wie zu sehen ist, ist ein Angreifer in der Lage zuverlässig vorherzusagen, ob die Daten eines Individuums zum Trainieren eines CLIP-Modells verwendet wurden.

1.2.4 Verteidigungen gegen Identitätsinferenz Angriffe in multimodalen Modellen

Große Modelle wie CLIP, welche oft auf Millionen von Datenpunkten trainiert werden, lernen wie im vorherigen gezeigt sensible Daten. Um sich gegen Attacken wie die Identity Inference Attack zu verteidigen beziehungsweise die Privatsphäre der Nutzer zu schützen, ist es daher wichtig personenbezogene Daten aus dem Modell zu entfernen.



a) Schematische Darstellung unserer Verteidigung für Textmodelle

b) Schematische Darstellung unserer Verteidigung für Bildmodelle

Abbildung 11: Backdoors können verwendet werden, um Einbettungen für das Unlearning umzuwandeln. Beide Abbildungen zeigen das Konzept des Einsatzes von Backdoor-Angriffen zum Unlearnen, ein Ansatz, der sowohl auf Text- als auch auf Bildmodelle anwendbar ist. Bei Textmodellen kann der Name auf einen neutralen Begriff wie ``eine Person'' abgebildet werden, während bei Bildmodellen die Gesichtseinbettung auf eine neutrale Zieleinbettung wie die durchschnittliche Gesichtseinbettung umgeschrieben werden kann.

Im Zuge von KISTRA haben wir einen Ansatz zum Löschen von privaten Informationen untersucht. Dieser Ansatz basiert auf sogenannten Backdoor Attacken, welche das Ziel haben ein vordefiniertes Verhalten in einem Modell während des Trainings zu verstecken. Der Angreifer kann anschließend dieses vordefinierte Verhalten durch einen Auslöser abrufen. In unserer entwickelten Verteidigung wird diese Methode jedoch

verwendet, um Informationen aus dem Modell zu entfernen. Wie in Abbildung 11 zu sehen ist, benutzen wir Backdoors, um Namen und Gesichter aus dem Text- und Bild-Encoder eines CLIP-Modells zu entfernen. Hierzu wird der Name der Person als Backdoor Trigger verwendet und einem neutralen Begriff, wie z.B. “Person”, zugeordnet. Das Resultat ist, dass das Modell in Zukunft den Namen als diesen neutralen Begriff interpretiert.
Bei einem Bild-Encoder kann eine ähnliche Methode angewendet werden. Hier wird anstatt eines Namens, das Gesicht einer Person als Auslöser für die Backdoor verwendet und auf eine neutrale Repräsentation im Einbettungsraum projiziert. Um die Verteidigung zu evaluieren, haben wir Personen aus dem Modell entfernt, für welche die Identity Inference Attack (IDIA) deren Zugehörigkeit zu den Trainingsdaten korrekt vorhergesagt hat. Anschließend haben wir die personenbezogenen Daten von diesen Individuen aus dem Modell entfernt und die IDIA erneut auf dem verteidigten Modell angewendet.

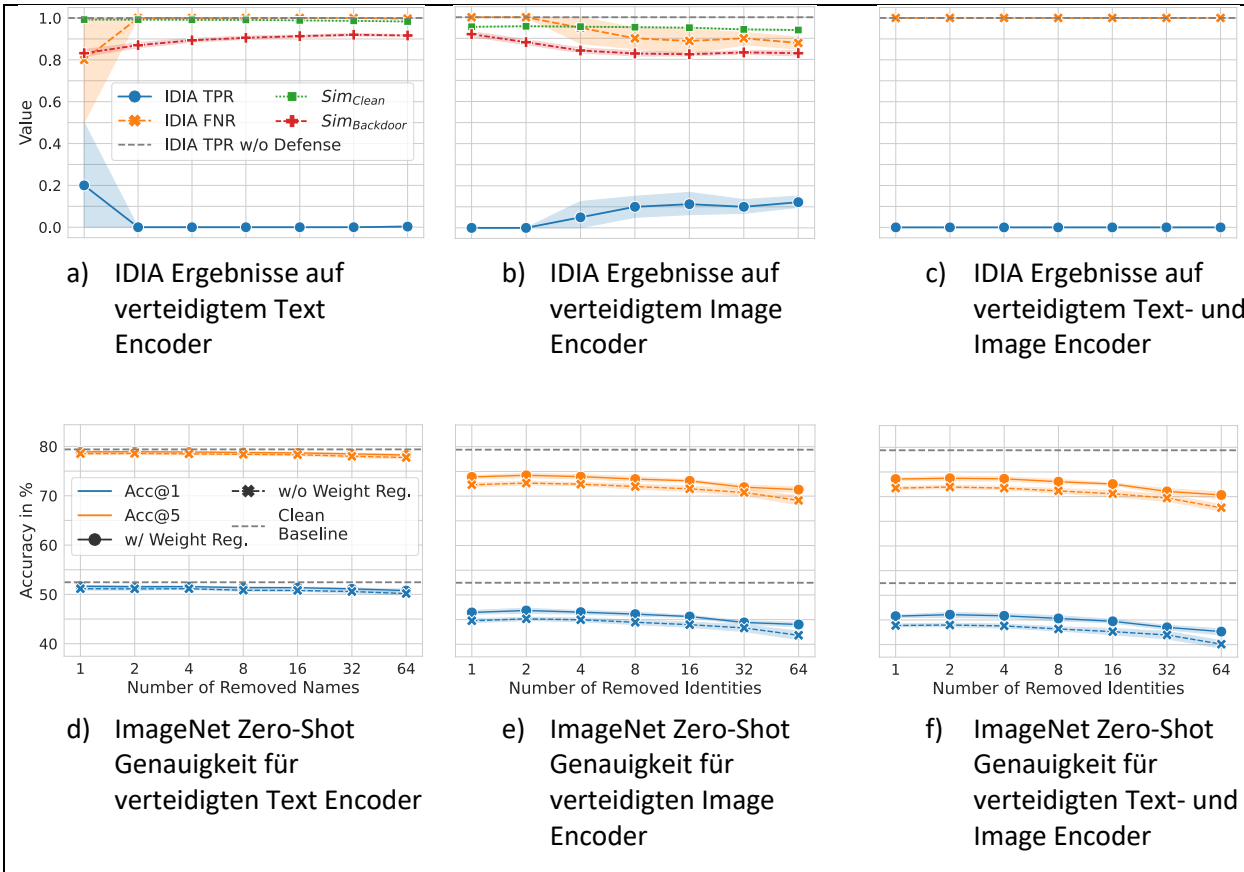


Abbildung 12: Die Verteidigung durch Backdoors ist erfolgreich darin die personenbezogenen Daten von Individuen aus dem Text- und dem Bild-Encoder zu entfernen. Nach anwenden der Verteidigung ist die Identity Inference Attack nicht mehr erfolgreich.

Die Ergebnisse dieser Untersuchung sind in Abbildung 12 zu sehen. Wie zu sehen ist, ist die Verteidigung gegen die Identity Inference Attacke sehr effektiv. Nach dem Anwenden unserer Verteidigung ist die IDIA für die Personen, deren Daten aus dem Modell entfernt wurden, nicht mehr erfolgreich. Auch wenn unser entwickelter Ansatz für den Text Encoder sehr effektiv ist, um personenbezogene Daten aus dem Modell zu entfernen, scheint es auf dem Image Encoder deutlich schwieriger zu sein zuverlässig alle Gesichter der Personen aus dem Modell zu löschen. Dies kann jedoch dadurch kompensiert werden, dass man die

personenbezogenen Daten sowohl aus dem Text Encoder, als auch aus dem Image Encoder löschen kann. Gleichzeitig konnten wir zeigen, dass das Verteidigen und das damit verbundene Injizieren einer Backdoor nur einen mäßigen bis fast gar keinen Einfluss auf die Performanz des Modells hat.

Für weitere und detailliertere Resultate verweisen wir auf unsere Publikation (Hintersdorf et al., “Defending Our Privacy With Backdoors”. NeurIPS Workshop 2023).

1.3 AP5: Technische Zusammenführung, Erstellung eines Frameworks und Evaluierung

1.3.1 Ziele von AP5: Technische Zusammenführung, Erstellung eines Frameworks und Evaluierung

Ziel dieses Arbeitspaketes ist die Bereitstellung von Teildemonstratoren, die die technischen Anforderungen aus AP 2-3 zusammenführen und als Evaluierungsgrundlage dienen.

1.3.2 Bereitstellung von Teildemonstratoren

Zu allen durchgeführten Arbeiten ist entsprechender Quellcode zur Reproduzierung der Ergebnisse sowie zu Demonstrationszwecken vorhanden. Verweise auf den jeweiligen Quellcode findet sich in den zugehörigen Publikationen. Der Quellcode dient weiterhin als Teildemonstrator der verschiedenen Angriffsszenarien und liefert vortrainierte Modelle zur Demonstration. Eine Einbindung der Angriffe in Umgebungen und Frameworks aus anderen Arbeitspaketen wurde unter Absprache nicht vorgenommen. Grund sind eine inkompatible Struktur der Angriffe und vorhandener Modelle, da für jedes vorhandene Modell umfangreiche Anpassungen an den Angriffen und der Modellstruktur vorgenommen werden müssen. Es wurde daher entschieden, für die verschiedenen Szenarien unabhängige Teildemonstratoren bereitzustellen, um eine einfache und schnelle Anwendung zu gewährleisten.

2 Verwendung der Zuwendung

Gegenstand dieses Kapitels ist die nachvollziehbare Erläuterung der Verwendung der erhaltenen Zuwendung sowie des Nutzens und der Fortschritte in dem untersuchten Forschungsgebiet.

2.1 Überblick über den zahlenmäßigen Nachweis

Die wichtigsten Positionen des Projektpartners Technische Universität Darmstadt waren Personalmittel sowie Reisekosten im Rahmen von Fachkonferenzen, auf welchen die Forschungsergebnisse präsentiert wurden.

2.2 Notwendigkeit und Angemessenheit der Projektarbeiten

Die Forschung zu Modellinversionsangriffen und Membership Inference Attacks ist aus mehreren Gründen wichtig. Ein besseres Verständnis dieser Angriffe hilft Forschern und Anwendern, Schwachstellen in KI-Modellen zu identifizieren. Indem bekannt ist, wie Angreifer Modelle ausnutzen können, um sensible Informationen über Trainingsdaten oder einzelne Datenpunkte zu erhalten, können Maßnahmen ergriffen werden, um diese Risiken zu mindern und die Sicherheit von KI-Systemen zu verbessern. Weiterhin bergen Modellinversions- und Membership Inference Angriffe erhebliche Datenschutzrisiken, insbesondere in Anwendungen, bei denen sensible Daten beteiligt sind. Forschung in diesem Bereich hilft bei der Entwicklung von Techniken zum Schutz der Privatsphäre von Einzelpersonen und stellt sicher, dass sensible Informationen nicht über KI-Modelle preisgegeben werden.

Angesichts des zunehmenden Fokus auf Datenschutzbestimmungen wie der DSGVO (Datenschutz-Grundverordnung) ist es wichtig, Datenschutzrisiken im Zusammenhang mit KI-Modellen zu verstehen und zu mindern, um die Einhaltung zu gewährleisten. Forschung zu Angriffen informiert über die Entwicklung von bewährten Verfahren und Richtlinien für den Einsatz von maschinellen Lernmodellen auf eine datenschutzkonforme Weise. Es ist außerdem wichtig, die Robustheit und Zuverlässigkeit von KI-Modellen sicherzustellen, um das Vertrauen in KI-Systeme aufrechtzuerhalten. Durch die Behandlung von Schwachstellen tragen wir dazu bei, vertrauenswürdigere und widerstandsfähigere KI-Systeme aufzubauen. Auch ist der Schutz der Privatsphäre und Vertraulichkeit von Personen eine ethische Verpflichtung in der KI-Forschung und -Implementierung. Forschung in diesem Bereich hilft, das Bewusstsein für die potenziellen Risiken im Zusammenhang mit maschinellen Lernmodellen zu schärfen und ethische Praktiken in ihrer Entwicklung und Implementierung zu fördern.

Zusammenfassend ist die Durchführung von Forschung zu sogenannten Privacy Angriffen auf KI-Modelle essenziell, um die Sicherheit, Privatsphäre und Vertrauenswürdigkeit von maschinellen Lernsystemen zu verbessern und damit ihre verantwortungsvolle und ethische Nutzung in verschiedenen Anwendungen zu erleichtern.

2.3 Voraussichtlicher Nutzen

Eine wirtschaftliche Verwertung der Ergebnisse ist aus Sicht der Technischen Universität Darmstadt bisher nicht geplant. Sie wird sowohl die Erkenntnisse mit ZITIS, dem BKA und anderen Sicherheitsbehörden teilen. Durch Einspielen der Ergebnisse in das BMBF Leuchtturmprojekt SPAICER und das Nationale Forschungszentrum für angewandte Cybersicherheit ATHENE in Darmstadt entsteht direkt ein Mehrwert in Deutschland. Das Bundesamt für Sicherheit in der Informationstechnik (BSI) hat bereits in ihrem Report „Generative AI Models - Opportunities and Risks for Industry and Authorities“ auf einen Teil der von der TUD in KiSTRA durchgeführten Arbeiten verwiesen, was den Nutzen der durchgeführten Arbeiten verdeutlicht.

2.4 Relevante F&E-Ergebnisse Dritter

Im Bereich der Modellinversionsangriffe wurde in den vergangenen Jahren diverse neue Algorithmen entwickelt. Nguyen et al. (2023) sowie Yuan et al. (2023) verbessern dabei den Trainingsprozess des generativen Modells sowie die Optimierungsfunktion während des Angriffs. Weitere Arbeiten von Han et al. (2023) sowie Kahla et al. (2022) befassen sich mit Angriffsszenarien, in welchen der Angreifer nur eingeschränkten Zugriff auf das Zielmodell besitzt. Allerdings sind die vorgestellten Angriffe leidiglich auf Daten mit geringer Auflösung evaluiert und teilweise mit hohem Rechenaufwand verbunden. Die von uns im Rahmen von KISTRA entwickelten Angriffe sind weiterhin die einzigen, die sich direkt auf Modelle trainiert auf Daten mit höherer Auflösung anwenden lassen.

Der größte Fortschritt im Bereich der Membership Inference Attacks liefert ein neuartiger Angriff von Carlini et al. (2022), welcher eine worst-case Perspektive einnimmt. Das Ziel des Angriffs ist die Untersuchung, welcher Anteil an Trainingsdaten mit sehr hoher Zuverlässigkeit identifiziert werden kann bei gleichzeitig einer geringen Rate an false-positive Vorhersagen.

2.5 Veröffentlichungen im Rahmen von KISTRA

2022

- To Trust or Not To Trust Prediction Scores for Membership Inference Attacks. Dominik Hintersdorf, Lukas Struppek, Kristian Kersting. International Joint Conference on Artificial Intelligence (IJCAI)
- Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash. Lukas Struppek, Dominik Hintersdorf, Daniel Neider, Kristian Kersting. ACM Conference on Fairness, Accountability, and Transparency (FaccT)
- Investigating the Risks of Client-Side Scanning for the Use Case NeuralHash. Dominik Hintersdorf, Lukas Struppek, Daniel Neider, Kristian Kersting. Workshop on Technology and Consumer Protection.
- Plug & Play Attacks: Towards Robust and Flexible Model Inversion. Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, Kristian Kersting. International Conference on Machine Learning (ICML)

2023

- Rickrolling the Artist: Injecting Backdoors into Text Encoders for Text-to-Image Synthesis. Lukas Struppek, Dominik Hintersdorf, Kristian Kersting. International Conference on Computer Vision (ICCV)
- Leveraging Diffusion-Based Image Variations for Robust Training on Poisoned Data. Lukas Struppek, Martin B Hentschel, Clifton Poth, Dominik Hintersdorf, Kristian Kersting. Conference on Neural Information Processing Systems (NeurIPS) - Workshop on Backdoors in Deep Learning.
- Defending Our Privacy With Backdoors. Dominik Hintersdorf, Lukas Struppek, Daniel Neider, Kristian Kersting. Conference on Neural Information Processing Systems (NeurIPS) - Workshop on Backdoors in Deep Learning
- Class Attribute Inference Attacks: Inferring Sensitive Class Information by Diffusion-Based Attribute Manipulations. Lukas Struppek, Dominik Hintersdorf, Felix Friedrich, Manuel Brack, Patrick Schramowski, Kristian Kersting. Preprint
- Combining AI and AM — Improving Approximate Matching through Transformer Networks. Frieder Uhlig, Lukas Struppek, Dominik Hintersdorf, Thomas Göbel, Harald Baier, Kristian Kersting. Forensic Science International: Digital Investigation
- Balancing Transparency and Risk: The Security and Privacy Risks of Open-Source Machine Learning Models. Dominik Hintersdorf, Lukas Struppek, Kristian Kersting. AISoLA: Bridging the Gap Between AI and Reality
- Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis. Lukas Struppek, Dominik Hintersdorf, Felix Friedrich, Manuel Brack, Patrick Schramowski, Kristian Kersting. Journal of Artificial Intelligence Research (JAIR)

2024

- Does CLIP Know My Face? Dominik Hintersdorf, Lukas Struppek, Manuel Brack, Felix Friedrich, Patrick Schramowski, Kristian Kersting. Journal of Artificial Intelligence Research (JAIR)
- Be Careful What You Smooth For: Label Smoothing Can Be a Privacy Shield but Also a Catalyst for Model Inversion Attacks. Lukas Struppek, Dominik Hintersdorf, Kristian Kersting. International Conference on Learning Representations (ICLR)

2.6 Literaturverzeichnis

Carlini et al., Membership Inference Attacks From First Principles (2022). IEEE Symposium on Security and Privacy.

Han et al., Reinforcement Learning-Based Black-Box Model Inversion Attacks (2023). Conference on Computer Vision and Pattern Recognition (CVPR).

Huang et al., Inference Attacks Against Face Recognition Model without Classification Layers (2024). Preprint.

Kahla et al., Label-Only Model Inversion Attacks via Boundary Repulsion (2022). Conference on Computer Vision and Pattern Recognition (CVPR).

Nguyen et al., Re-thinking model inversion attacks against deep neural networks (2023). Conference on Computer Vision and Pattern Recognition (CVPR).

Yuan et al., Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network (2023). AAAI Conference on Artificial Intelligence (AAAI).