Universität Konstanz
Fachbereich Mathematik und Statistik
&
Bundesamt für Sicherheit in der
Informationstechnik

Exposé für eine Masterarbeit zum Thema:

# Untersuchung & Entwicklung von Ansätzen zur Detektion von Poisoning-Angriffen

Lukas Schulth
lukas.schulth@uni.kn

unter der Betreuung von

Herr Prof. Dr. Johannes Schropp
johannes.schropp@uni.kn

Herr Prof. Dipl.-Ing. Markus Ullmann
markus.ullmann@bsi.bund.de

Herr Dr. Christian Berghoff
christian.berghoff@bsi.bund.de

Herr Matthias Neu
matthias.neu@bsi.bund.de

19. September 2021

# Inhaltsverzeichnis

# 1 Forschungsthema

Heute gibt es kaum noch einen Bereich, in dem Anwendungen auf Basis von Künstlicher Intelligenz keine Rolle spielen, sei es in der Produktion, Werbung, Kommunikation, Biometrie oder Automotive. Viele Unternehmen nutzen KI-Systeme, etwa um präzise Nachfrageprognosen anzustellen und das Kundenverhalten exakt vorherzusagen. Auf diese Weise lassen sich beispielsweise Logistikprozesse regional anpassen. Auch im Gesundheitswesen bedient man sich spezifischer KI-Tätigkeiten wie dem Anfertigen von Prognosen auf Basis von strukturierten Daten. Hier betrifft das etwa die Bildverarbeitung: So werden Röntgenbilder als Input in ein KI-System gegeben, der Output ist eine Diagnose. Das Erfassen von Bildinhalten ist auch beim autonomen Fahren entscheidend, wo Verkehrszeichen, Bäume, Fußgänger und Radfahrer fehlerfrei erkannt werden müssen. In solch sensiblen Anwendungsfeldern wie der medizinischen Diagnostik oder in sicherheitskritischen Bereichen müssen KI-Systeme absolut zuverlässige Problemlösungsstrategien liefern [1].

Ein neuronales Netzwerk als KI-System kann als die Verkettung von linearen Funktionen und nicht-linearen Aktivierungsfunktionen verstanden werden. Zum Lebenszyklus eines KI-Systems gehören die Datenerhebung und Datenaufbereitung, der Trainingsprozess, das Testen und die Inferenz, bei der das parametrisierte Netzwerk in der Anwendung genutzt wird. Häufig sind neuronale Netzwerke klassischen Verfahren deutlich überlegen, gleichzeitig fehlt es aber an der Interpretierbarkeit und damit an der Nachvollziehbarkeit getroffener Entscheidungen.
Bei einem Entscheidungsbaum kann beispielsweise genau erklärt werden, aufgrund welcher Schwellenwerte welche Entscheidung zustande kommt. Im Falle eines neuronalen Netzwerkes liegen Millionen von Parametern vor, die genau verstanden und interpretiert werden müssen.

Genau diese Schwachstelle der fehlenden Interpretierbarkeit bietet auch die Möglichkeit, ein neuronales Netzwerk gezielt zu manipulieren. Während der Angreifer bei einem adversarialen Angriff das fertig trainierte Netz während der Inferenz angreift, findet der Poisoning-Angriff vor bzw. zeitgleich zum Training statt.
Dabei werden korrumpierte (verfälschte) Daten in den Trainingsdatensatz eingeschleust, um die Vorhersagequalität eines Netzwerkes zu verringern. Das Ziel des Angreifers ist es, im Fall eines Klassifikationsproblems beispielsweise, die Testgenauigkeit einer einzelnen Klasse zu verringern, während die Testgenauigkeit auf allen anderen Klassen gar nicht oder nur leicht verringert wird.
Ein Spezialfall sind die sogenannten Backdoor-Poisoning-Angriffe, bei denen der Angreifer eine Art Hintertür im Datensatz implementiert, die während der Inferenz ausgenutzt werden kann. Diese Angriffe können in ihrer Detektion noch erschwert werden, indem das entsprechende Label nicht zusätzlich abgeändert wird. In diesem Fall sprechen wir von Label-konsistenten Poisoning-Angriffen.

Es ist noch immer schwierig, die konzeptionell leicht umzusetzenden Poisoning-Angriffe erfolgreich zu detektieren.

## 2 Zielsetzung und Erkenntnisinteresse

Ziel ist es, einen Algorithmus zu entwickeln, der einen Backdoor-Poisoning-Angriff erkennen kann. Dieser soll mit state-of-the-art-Methoden konkurrieren können und gleichzeitig auch eine Antwort auf Label-konsistente Poisoning-Angriffe bieten. Zudem soll die Entscheidungsfindung Neuronaler Netzwerke besser verstanden werden.

Als Datensatz soll *The German Traffic Sign Recognition Benchmark* [1], bestehend aus mehr als 50.000 Bildern in über 40 verschiedenen Klassen, verwendet werden. Das Klassifikationsproblem besteht darin, die Bilder von Verkehrsschildern ihren entsprechenden Klassen zuzuordnen.

## 3 Forschungsstand und theoretische Grundlagen

Im Allgemeinen verfolgen Poisoning-Angriffe das Ziel, die Vorhersagegenauigkeit eines Netzwerkes bei der Anwendung auf nicht-korrumpierten Daten zu verringern.
Für den Fall eines Klassifikationsproblems kann zwischen gezielten und ungezielten Poisoning-Angriffen unterschieden werden. Erstere versuchen, die Vorhersagegenauigkeit nur auf einzelnen Klassen zu verringern, während die Vorhersagegenauigkeit auf allen anderen Klassen gar nicht oder nur leicht verringert wird.
Bei sogenannten Backdoor-Poisoning-Angriffen (Hintertür-Angriffen) wird zunächst wieder eine spezielle Klasse angegriffen. Dabei platziert der Angreifer einen Trigger (Auslöser) innerhalb des Datenpunktes und ändert die zugehörigen Labels ab. Während des Trainings lernt das Netzwerk nun, Datenpunkte dieser speziellen Klasse in Anwesenheit eines Triggers, aus Sicht des Angreifers korrekt falsch zu klassifizieren. In der Abwesenheit eines Triggers funktioniert das Netz wie gewünscht und die Datenpunkte werden korrekt klassifiziert. Dieses Verhalten macht es schwierig, solche Angriffe selbst durch umfangreiche Tests zu erkennen.
Im Unterschied zum Standard-Backdoor-Poisoning-Angriff, wird bei einem Label-konsistenten Poisoning-Angriff das zum Datenpunkt gehörige Label nicht zusätzlich abgeändert. Bevor ein Trigger implementiert wird, benutzt man ein zweites Neuronales Netz, um die Vorhersagegenauigkeit auf dem Datenpunkt ohne Trigger möglichst stark zu verringern, sodass sich das angegriffene Netzwerk später mehr auf den Trigger selbst anstatt auf den Datenpunkt verlässt.
Zwei mögliche Verfahren zum Erstellen eines solchen Angriffs werden in [2] vorgestellt. Bei dem dort verwendeten Trigger wird bei neun ausgewählten Pixeln

[1] `https://benchmark.ini.rub.de/gtsrb_news.html`

auf allen drei Farbkanälen eine Amplitude addiert bzw. subtrahiert.

In ersten Untersuchungen im Rahmen eines Praktikums beim Bundesamt für Sicherheit in der Informationstechnik (BSI) von Januar bis April 2020 konnten wir feststellen, dass Standard-Poisoning-Angriffe auf mehreren Inception-Netzwerken[2] sehr gut funktionieren. Ein Angriff wird mit Hilfe einer Angriffserfolgsrate bewertet. Diese gibt den prozentualen Anteil der korrumpierten Daten im Testdatensatz an, die beim Testen auch erfolgreich falsch klassifiziert werden.
Eine bereits untersuchte Gegenmaßname aus der Literatur ist das sogenannte Activation Clustering[3]. Dabei werden die Aktivierungen der vorletzten Netzwerkschicht als Merkmale der Eingabe betrachtet. Nach einer Dimensionsreduktion wird ein Clustering mit $K = 2$ Clustern durchgeführt. Dabei sollen im Idealfall ein korrumpiertes und ein nicht-korrumpiertes Cluster entstehen. Die Verteilung auf beide Cluster wird mittels eines Silhouettenkoeffizienten bewertet. Dieser ist ein Maß dafür, wie gut ein Clustering bezüglich einer Distanz zu einem Datensatz passt. Anhand eines Schwellenwertes wird die Präsenz von korrumpierten Daten ermittelt.
Eine im Anschluss mögliche Gegenmaßnahme ist das erneute Trainieren des Netzwerkes ohne den verdächtigen Anteil an Daten und der anschließende Vergleich der Testgenauigkeit. Dieser Ansatz funktioniert deutlich besser, ist jedoch mit sehr großem Aufwand verbunden.
Des Weiteren zeigte sich, dass ein kleinerer Anteil an korrumpierten Daten die Angriffserfolgsrate verringert. Besonders diejenigen Angriffe, die eine hohe, aber nicht 100-prozentigen Angriffsrate besitzen, lassen sich kaum erkennen. Zudem bedeutet eine Rate von 100 Prozent nicht unbedingt, dass der Angriff auch detektiert wird.
Bei den Label-konsisten-Angriffe konnten wir die Amplitudenstärke jedoch nicht so weit absenken, wie in [2] angegeben. Die Methode des erneuten Trainierens als Gegenmaßnahme funktioniert bei den Label-konsistenten Angriffen nicht, da die Label zum Datenpunkt passen.

In [4] wird ein Verfahren zur Detektion vorgestellt, das darauf beruht die Testdatenpunkte einer Transformation zu unterziehen, bevor diese durch das Netzwerk ausgewertet werden. Vorteil davon ist der sehr geringe Rechenaufwand.
Eine ausführliche zeitliche Entwicklung möglicher Poisoning-Angriffe und entsprechender Gegenmaßnahmen ist in [5] gegeben.

# 4 Konzept

Mit Methoden aus dem Bereich *explainable AI* (erklärbare Künstliche Intelligenz, kurz: XAI) soll die Detektion von Poisoning-Angriffen verbessert werden. XAI beschreibt Methoden, die es ermöglichen, Entscheidungen eines neuronalen Netzes besser interpretieren zu können. Diese lassen sich in lokale und globale

[2] `https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/44903.pdf`

Methoden unterteilen. Ein Beispiel für einen lokalen Ansatz ist die sogenannte „Layer-wise Relevance Propagation“ (LRP)[6, 7], die sichtbar macht, aufgrund welcher Kriterien KI-Systeme Entscheidungen treffen. Dabei werden die einzelnen Bestandteile eines Datenpunktes nach ihrer Relevanz für die Entscheidung des Netzwerkes sortiert. Zu einem Bild als Datenpunkt entsteht eine sogenannte Heatmap.
Anschließend wird eine Spektrale Relevanz-Analyse (SpRAy)[8] durchgeführt. Diese identifiziert und quantifiziert ein breites Spektrum erlernter Entscheidungsverhalten. So wird es möglich, auch in sehr großen Datensätzen unerwünschte Entscheidungen zu erkennen[9].
Im Unterschied zum Activation-Clustering, bei welchem die Aktivierungen bestimmter Netzwerkschichten als Merkmale extrahiert werden, benutzt man hier die Heatmaps.
Ausgehend von einer Metrik (Euklidische Distanz oder Gromov-Wasserstein-Distanz[10]) wird eine Affinitätsmatrix A aller paarweisen Netzwerkeingaben einer bestimmten Klasse berechnet. Durch die Berechnung der Spektralen Einbettung kann Information über die Cluster-Struktur der Eingaben einer bestimmten Klasse gewonnen werden. Dabei liefert eine Eigenwertzerlegung der zu A gehörigen Laplace-Matrix L Aufschluss über die Anzahl an verschiedenen Clustern. Das Clustering findet auf den Heatmaps bzw. einer in der Dimension reduzierten Version davon statt.
Diese Idee basiert auf [11] und könnte durch die neuere und deutlich umfangreichere Version des Papers [12] ergänzt werden. Eine Verbesserung der LRP für Convolutional Neural Networks, die sogenannte Softmax Gradient Layer-wise Relevance Propagation, wird in [13] vorgestellt.

# 5 Vorläufiger Zeitplan

Ausgehend von einer Dauer von 24 Wochen unterteilen wir die einzelnen Schritte wie folgt:

- Woche 1-3: Literaturrecherche
- Woche 4-6: Implementierung des LRP-Algorithmus
- Woche 7-10: Implementierung der Spektralen Relevanz-Analyse
- Woche 11-14: Anwendung auf verschiedene Arten von Poisoning-Angriffe
- Woche 15-16: Vergleich mit bisher untersuchten Methoden
- Woche 17-23: Aufschreiben der Ergebnisse
- Woche 24: Druck und Abgabe der Arbeit

# 6 Vorläufige Gliederung

- Einführung
- Theoretische Grundlagen
- Methoden
  - Linear Relevance Propagation
  - Spectral Relevance Analysis
- Entwurf des Algorithmus
- Anwendung auf bekannte Poisoning-Angriffe
- Ergebnisse
- Zusammenfassung/Fazit

# 7 Literatur


[1] Fraunhofer Gesellschaft, Forschung Kompakt. 1. Juli 2019. *Künstliche Intelligenz erklärbar machen - Der Blick in Neuronale Netze*. Zugriff am 21.01.2021.
`https://www.fraunhofer.de/content/dam/zv/de/presse-medien/2019/Juli/forschung-kompakt/hhi-der-blick-in-neuronale-netze.pdf`

[2] Alexander Turner, Dimitris Tsipras, Aleksander Madry. *Label-Consistent Backdoor Attacks*. 6 Dec 2019
`https://arxiv.org/abs/1912.02771`

[3] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. *Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering*. 12 Nov 2018
`arXiv:1811.03728v1`

[4] Yiming Li,Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, Shutao Xia *Rethinking the Trigger of Backdoor Attack*. 24 Jun 2020
`https://arxiv.org/abs/2004.04692`

[5] Yansong Gao, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal and Hyoungshick Kim. *Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review*. 2 Aug 2020
`https://arxiv.org/abs/2007.10760`

[6] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, W. Samek. *The LRP Toolbox for Artificial Neural Networks*. 2016
`https://jmlr.org/papers/volume17/15-618/15-618.pdf`

[7] Binder A., Bach S., Montavon G., Müller KR., Samek W. (2016) *Layer-Wise Relevance Propagation for Deep Neural Network Architectures*. In: Kim K., Joukov N. (eds) Information Science and Applications (ICISA) 2016. Lecture Notes in Electrical Engineering, vol 376. Springer, Singapore.
`https://doi.org/10.1007/978-981-10-0557-2_87`

[8] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. *Unmasking Clever Hans predictors and assessing what machines really learn*. Nature Communications, vol. 10, p. 1096, 2019.
`https://www.nature.com/articles/s41467-019-08987-4`

[9] Pressemitteilung Fraunhofer HHI, 12. März 2019. Zugriff am 21.01.2021.
`https://www.hhi.fraunhofer.de/presse-medien/nachrichten/2019/wie-intelligent-ist-kuenstliche-intelligenz.html`

[10] Facundo Mémoli. Department of Mathematics, The Ohio State University, Columbus, OH, USA. 2 Sep 2014. *The Gromov–Wasserstein Distance: A Brief Overview* `file:///tmp/mozilla_lukasschulth0/The_Gromov-Wasserstein_Distance_A_Brief_Overview.pdf`

[11] Christopher J. Anders, Talmaj Marinč, David Neumann, Wojciech Samek, Klaus-Robert Müller and Sebastian Lapuschkin. *Analyzing ImageNet with Spectral Relevance Analysis: Towards ImageNet un-Hans'ed*. 22 Dec 2019 `https://arxiv.org/abs/1912.11425v1`

[12] Christopher J. Anders, Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. *Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models*. 22 Dec 2020 `https://arxiv.org/pdf/1912.11425.pdf`

[13] Brian Kenji Iwana, Ryohei Kuroki, Seiichi Uchida, Kyushu University, Fukuoka, Japan. *Explaining Convolutional Neural Networksusing Softmax Gradient Layer-wise Relevance Propagation*. 7 Nov 2019. `https://arxiv.org/pdf/1908.04351.pdf`