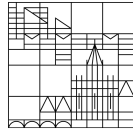


Universität
Konstanz



Bundesamt
für Sicherheit in der
Informationstechnik

Untersuchung & Entwicklung von Ansätzen zur Detektion von Poisoning-Angriffen

Master-Arbeit

Lukas Schulth
`lukas.schulth@uni.kn`

1. Oktober 2021

Erstkorrektor, Zweitkorektor, Betreuer, Fachbereich Mathematik und Statistik, Masterarbeit zur Erlangung des Titels Master of Science (M.Sc.)
vgl. mit Titelblatt Exposé(!)

Abstract

english

Zusammenfassung

deutsch

Keywords— one, two, three, four

Abbildungsverzeichnis

1	(Optischer) Vergleich von korruptem Datenpunkt und berechneter Heatmap.	14
2	Ergebnis des spektralen Clusterings unter Verwendung der Euklidischen Distanz und k=10 Nachbarn	29
3	Auswahl der relevantesten Pixel (bis zu 99% der Gesamtmasse) zweier Heatmaps	30
4	Auswahl der relevantesten Pixel (bis zu 50% der Gesamtmasse) zweier Heatmaps	31
5	Einbettung des Baryzentrums mithilfe von Multidimensionaler Skalierung(MDS) bei der Wahl von 99% der Gesamtmasse	32

Tabellenverzeichnis

1	Für einen Poisoning-Angriff interessante Klassen und die zugehörige Anzahl an Bildern.	8
2	Qualität der Angriffe auf das Inception v3-Netz mit Stickern Seitenlänge 2 und 3 Pixel bei unterschiedlich großen Anteilen an korrupten Daten	33
3	Fehler für Testproblem 1 zum Endzeitpunkt $T = 2.0$	33
4	Konvergenz im Ort für das Testproblem 4, $T = 2.0$	34

Listings

1	python-interner Aufbau einer BatchConv Schicht	7
2	Verfügbare Schichten und Aktivierungsfunktionen	14
3	Implementierte Regeln fhvilshoj	15
4	Kleines Netzwerk	31
5	Einfachere Version von Inception v3	32
6	Reversed Model incv3	35

Algorithmenverzeichnis

1	Foo bar	6
2	Berechnung der GW_{ε} Baryzentren	27
3	Algorithm caption	28

Inhaltsverzeichnis

1	Einführung	6
2	Neuronale Netzwerke	6
2.0.1	CNNS	7
2.0.2	Besondere Schichten	7
2.0.3	Inception v3	7
2.0.4	VGG16	8
2.1	Datensatz	8
3	Poisoning-Angriffe	9
3.1	Standard Poisoning-Angriffe	9
3.2	Label-konsistente Poisoning-Angriffe	9
3.3	Verteidigungen	9
4	Erklärbare KI	9
4.1	Lokale Methoden	10
4.2	Globale Methoden	10
5	Layer-wise Relevance Propagation	10
5.1	Idee	10
5.2	Behandlung von biases	12
5.3	Beispiel an einem kleinen Netzwerk	12
5.4	Deep Taylor Decomposition	12
5.4.1	Taylor Decomposition	12
5.4.2	Deep Taylor Decomposition	13
5.5	Verschiedene Verfahren	13
5.6	Eigenschaften	13
5.7	Behandlung besonderer Schichten	13
5.7.1	BatchNorm2D	13
5.8	LRP für Deep Neural Nets/Composite LRP	13
5.9	Verarbeitung der Heatmaps	13
5.10	Implementierung	13
5.10.1	Tensorflow	13
5.10.2	pytorch	13
6	Detektion von Poisoning-Angriffen basierend auf LRP	15
6.1	Idee	15
6.2	k-means / k-means++ -Clustering	16
6.3	Spektrales Clustering	16
6.4	Anwendung auf unterschiedliche Poisoning-Angriffe	16
6.5	Verwendete Distanzen & Approximationen	17
6.5.1	Histogramme & Maße	17
6.5.2	Euklidische Distanz	18
6.5.3	Gromov-Hausdorff-Distanz	18
6.5.4	Optimaler Transport (Monge Formulierung)	19
6.5.5	Optimaler Transport nach Kantorovich	19
6.5.6	Monge-Kantorovitch equivalence	20
6.5.7	Metrische Eigenschaften	20

6.5.8	Duale Formulierung	21
6.5.9	Gromov-Wasserstein-Divergenz	21
6.5.10	Regularisierungen	22
6.5.11	Entropisch Regularisierte Gromov-Wasserstein-Distanz	22
6.5.12	Berechnung der Lösung: Sinkhorn	23
6.5.13	Verallgemeinerung	24
6.5.14	Wasserstein Baryzentren	25
6.5.15	Numerische Approximationen	27
7	(Numerische) Ergebnisse/Vergleich mit anderen Verfahren	27
7.0.1	Standard Poisoning-Angriffe	27
7.0.2	Label-konsistente Poisoning-Angriffe	28
7.1	Activation Clustering	28
7.2	Räumliche Transformationen	28
8	Weitere mögliche Schritte	29
9	Zusammenfassung	29
A	Verwendete Netzwerke	31
A.1	Net	31
B	Parameter für Training und Einlesen der Daten	37
C	Datensätze	37
D	Programmcode	37
E	Notizen	37

Algorithm 1 Foo bar

...

1 Einführung

Pweave¹A Complete List of All (arXiv) Adversarial Example Papers ²

In sicherheitskritischen Anwendungsgebieten ist die Erklärung für das Zustandekommen einer Entscheidung genauso wichtig wie die Entscheidung selbst [BBM⁺16].

Clustering auf Datenpunkten direkt(50 Prozent = raten), Clustering auf Aktivierungen gut geeigneter Netzwerkschichten. Clustering auf den Heatmaps der verdächtigen Klasse.

Clustering auf unterschiedlichen Repräsentationen der Bilder:

- Clustering direkt auf den Bildern
- Clustering auf den Activations einer Netzwerkschicht(Im Paper [CCB⁺18] wird die vorletzte Schicht benutzt)
- Clustering auf den Heatmaps

In Abschnitt 2 geben wir eine kurze Einführung in Neuronale Netzwerke und stellen die untersuchten Modelle vor. Abschnitt 3 führt in die unterschiedlichen Möglichkeiten eines Poisoning-Angriffs auf Neuronale Netzwerke ein. Abschnitt 4 gibt eine kurze Übersicht über den Bereich der Erklärbaren Künstlichen Intelligenz, wobei ein Beispiel eines Verfahrens, die sogenannte Layer-wise Relevanz Propagation ausführlich in Abschnitt 5 vorgestellt wird. Kern der Arbeit bildet Abschnitt 6, wo wir zu Beginn die grundlegenden Bestandteile des Algorithmus zur Detektion von Poisoning-Angriffen auf Neuronale Netzwerke erklären, bevor die experimentellen Ergebnisse in Unterabschnitt 6.4 ausführen. Ein Vergleich mit anderen Detektionsverfahren wird in Abschnitt 7 durchgeführt.

2 Neuronale Netzwerke

Wir betrachten ein Neuronales Netzwerk (NN), dass die Funktion $f_{\theta} : \mathbb{R}^n \rightarrow \mathbb{R}$, mit $\theta = (w_{il}, b_{il})$ beschreibt. i: Schicht l: Neuron in der Schicht w: Gewichte b: Bias g: nichtlineare Aktivierungsfunktion

Pre-Activations(lokal, global): $z_{ij} = x_i * w_{ij}$ $z_j = \sum_i z_{ij} + b_j$

Vorschrift/aktivierungen: $x_j = g(z_{ij})$

Training;testing, Validation, Forward pass, backward pass SGD erklärt im Einführungsteil von [IS15]

fehlende Interpretierbarkeit

ReLUs in den meisten Netzwerken

¹<https://mpastell.com/pweave/>²<https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html>

Definition Klasse
Supervised vs Unsupervised
Dimensionality reduction and Visualisation

2.0.1 CNNS

Idee, Abstraktion, high level, low level features, bekannte Netzwerke

Ausführliche Einführung stanford Kurs [?]. Unterschied zu FC layers: It is worth noting that the only difference between FC and CONV layers is that the neurons in the CONV layer are connected only to a local region in the input, and that many of the neurons in a CONV volume share parameters. However, the neurons in both layers still compute dot products, so their functional form is identical. Therefore, it turns out that it's possible to convert between FC and CONV layers

Starting with LeNet-5 [10], convolutional neural networks (CNN) have typically had a standard structure – stacked convolutional layers (optionally followed by contrast normalization and max-pooling) are followed by one or more fully-connected layers. Variants of this basic design are prevalent in the image classification literature and have yielded the best results to-date on MNIST, CIFAR and most notably on the ImageNet classification challenge [9, 21]. For larger datasets such as ImageNet, the recent trend has been to increase the number of layers [12] and layer size [21, 14], while using dropout [7] to address the problem of overfitting. [SLJ⁺15]

Softmax am Ende für Transformation in Probabilities.

Netzwerk im Netzwerk [?, SLJ⁺15]

2.0.2 Besondere Schichten

Promotion Sebastian Lapuschkin

- *BatchConv* besteht aus

```
1 nn.Conv2d(in_channels=in_channels, out_channels=
  out_channels, **kwargs)
2 nn.BatchNorm2d(num_features=out_channels)
3 nn.ReLU()
4
```

Listing 1: python-interner Aufbau einer BatchConv Schicht

in genau dieser Reihenfolge Bem.: Nur für BatchNorm2d müsste man LRP implementieren, für Conv2d funktioniert das bereits.

Batch Normalization³

2.0.3 Inception v3

Filter, In Klassischen feed forward Netzen wird Output der vorherigen layer ist input der nächsten layer

Jetzt: Inception Block: Previous layer input, 4 operations in parallel, concatenation, 1x1 conv -> lower dimension -> less computational cost

³<https://arxiv.org/pdf/1502.03167.pdf>

Intermediate classifiers: kommt aus multitask learning. Eigentlich eine Möglichkeit gegen vanishing gradients

2.0.4 VGG16

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper Very Deep Convolutional Networks for Large-Scale Image Recognition. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. It was one of the famous model submitted to ILSVRC-2014. It makes the improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another. VGG16 was trained for weeks and was using NVIDIA Titan Black GP's.

2.1 Datensatz

GTSRB⁴

Für die Poisoning-Angriffe auf verschiedene neuronale Netzwerke benutzen wir den Datensatz German Traffic Sign Recognition Benchmark 1. Dieser besteht aus 52.001 Bildern von Verkehrsschildern aus 43 verschiedenen Kategorien der Pixelgröße 32×32 . Etwa 75 Prozent der Bilder wird für das Training, die anderen 25 Prozent für das Testen benutzt. Der Datensatz wurde ursprünglich in einem Wettbewerb auf der International Joint Conference on Neural Networks (IJCNN) im Jahr 2011 benutzt. Die Bilder sind aus einer Videosequenz herausgeschnitten. Deshalb befinden sich in einer Klasse jeweils immer mehrere Bilder desselben Verkehrsschildes zu unterschiedlichen Zeitpunkten. Aufnahmen desselben Verkehrsschildes kommen nicht übergreifend in Training-, Validierung- oder Testdatensatz vor. Verkehrsschild

Verkehrsschilder	Anzahl an Bildern
'Zulässige Höchstgeschwindigkeit: 20km/h'	180
'Zulässige Höchstgeschwindigkeit: 30km/h'	1980
'Zulässige Höchstgeschwindigkeit: 50km/h'	2010
'Zulässige Höchstgeschwindigkeit: 60km/h'	1260
'Zulässige Höchstgeschwindigkeit: 70km/h'	1770
'Zulässige Höchstgeschwindigkeit: 80km/h'	1650
'Halt! Vorfahrt gewähren'	690

Tabelle 1: Für einen Poisoning-Angriff interessante Klassen und die zugehörige Anzahl an Bildern.

In Tabelle ?? sind einige Klassen der Verkehrsschilder und deren Anzahl im Datensatz aufgelistet, die für einen Poisoning-Angriff interessant sein könnten. Die Anzahl der Schilder 'Halt! Vorfahrt gewähren'-Schilder im Trainingssatz beträgt etwa 690 Aufnahmen. Diese wurden von insgesamt nur 24 verschiedenen 'Halt!

⁴https://benchmark.ini.rub.de/gtsrb_dataset.html

Vorfahrt gewähren'-Schildern aufgenommen. Da beim Erstellen der korruptierten Daten auch immer das Bild aus der angegriffenen Klasse in die Zielklasse verschoben wird, wird die Anzahl der in der Ursprungs-kategorie verbleibenden Daten abhängig vom Anteil an korruptierten Daten kleiner. Wir werden uns deshalb im Folgenden mit Angriffen auf die Klasse 'Zulässige Höchstgeschwindigkeit: 50 km/h' beschäftigen, da sie die höchste Anzahl an Daten aufweist.

3 Poisoning-Angriffe

3.1 Standard Poisoning-Angriffe

Wir wollen Schilder der Klasse 50kmh absichtlich falsch als 80kmh. Wir wählen diese beiden Klassen aufgrund der Größe beider Klassen (s. Aufstellung in Praktikumsbericht). Stoppschildklasse ist wohl vergleichsweise ziemlich klein.

Dazu fügen wir auf den 50er Schildern einen Sticker ein und ändern das Label auf 80.

Für die Bewertung, wie erfolgreich ein Angriff war, fügen wir in jedem Bild der 50er Klasse im Testdatensatz einen Sticker ein und messen, wie groß der Anteil der 50er Schilder ist, die als 80er Schild klassifiziert werden.

CH- und Backdoor-Artefakte:

In [AWN⁺19] wird wie folgt zwischen Clever Hans- und Backdoor-Artefakten unterschieden. In beiden Fällen wird rechts oben im Bild ein grauer 3x3 Sticker eingefügt. Bei CH geschieht dies bei 25% der `airplane`-Klasse. Bei Backdoor-Artefakten werden 10% aller Bilder korruptiert. Im zweiten Fall wird das entsprechende Label abgeändert. Dies entspricht dann einem Standard- bzw. Clean-Label-Poisoning-Angriff. (Wie gut funktioniert der CLPA/CH hier ohne die Bilder vorher schlechter zu machen? TODO: Vergleich mit [TTM19]). In [AWN⁺19], Kapitel 2.1 wird auch auf die Methode der Spektralen Signatur [TLM18] eingegangen, die zur Detektion genutzt wird. Diese eignet sich wohl sehr gut für die Backdoor-Attacks, aber nur schlecht für die CH-Artefakte.

3.2 Label-konsistente Poisoning-Angriffe

3.3 Verteidigungen

4 Erklärbare KI

Erklärbarkeit vs. Interpretierbarkeit, youtube talk?!

4.1 Lokale Methoden

4.2 Globale Methoden

5 Layer-wise Relevance Propagation

5.1 Idee

Die Layer-wise Relevance Propagation (LRP) wird in [BBM⁺15] erstmalig vorgestellt. Die Idee besteht darin, einen Zusammenhang zwischen der Ausgabe eines Klassifikators $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^+$ und der Eingabe x herzustellen. Dabei wird eine definiert, die über gewisse Eigenschaften eingeschränkt wird. Die Autoren bezeichnen die Herangehensweise hier selbst als heuristisch und liefern in ?? eine Verallgemeinerung des Konzepts, die gleichzeitig die mathematische Grundlage bildet.

Wir betrachten eine nicht-negative Funktion $f : \mathbb{R}^d \rightarrow \mathbb{R}^+$. Im Bereich der Bild-Klassifizierung ist die Eingabe $x \in \mathbb{R}^d$ ein Bild, das wir als Menge von Pixelwerten $x = \{x_p\}$ auffassen können. Dabei beschreibt der Index p einen genauen Pixelpunkt. Während für schwarz-weiß Bilder $x_p \in \mathbb{R}$ gilt, gilt im Fall von RGB-Bildern $x_p \in \mathbb{R}^3$ für die einzelnen Farbkanäle Rot, Grün und Blau. Die Funktion $f(x)$ ist ein Maß dafür, wie präsent ein oder mehrere Objekte in der Eingabe/im Eingabebild vorhanden sind. Ein Funktionswert $f(x) = 0$ beschreibt die Abwesenheit. Gilt andererseits $f(x) > 0$, so wird die Präsenz mit einem gewissen Grad an Sicherheit oder eine gewisse Menge zum Ausdruck gebracht.

Mit Hilfe der LRP soll nun jedem Pixel p im Eingabebild eine Relevanz $R_p(x)$ zugeordnet werden, die für jedes Pixel x_p angibt, mit welcher Größe es für das Entstehen einer Entscheidung $f(x)$ verantwortlich ist. Die Relevanz eines jeden Pixels wird dabei in einer Heatmap $R(x) = \{R_p(x)\}$ zusammengefasst.

Die Heatmap besitzt dieselbe Größewie x und kann als Bildvisualisiert werden.

Wir definieren die folgenden Eigenschaften:

Definition 5.1.1. Eine Heatmap $R(x)$ heißt konservativ, falls gilt:

$$\forall x : f(x) = \sum_p R_p(x), \quad (5.1)$$

d.h. die Summe der im Pixelraum zugeordneten Relevanz entspricht der durch das Modell erkannten Relevanz.

Definition 5.1.2. Eine Heatmap $R(x)$ heißt positiv, falls gilt:

$$\forall x, p : R_p(x) \geq 0, \quad (5.2)$$

d.h. alle einzelnen Relevanzen einer Heatmap sind nicht-negativ.

Die erste Eigenschaft verlangt, dass die umverteilte Gesamtrelevanz der Relevanz entspricht, mit der ein Objekt im Eingabebild durch die Funktion $f(x)$ erkannt wurde. Die zweite Eigenschaft beschreibt, dass keine zwei Pixel eine gegensätzliche Aussage über die Existenz eines Objektes treffen können. Beide Definitionen zusammen ergeben die Definition einer *konsistenten* Heatmap:

Definition 5.1.3. Eine Heatmap $R(x)$ heißt konsistent, falls sie konservativ und positiv ist, d.h. Definition 5.1.1 und Definition 5.1.2 gelten.

Für eine konsistente Heatmap gilt dann $(f(x) = 0 \Rightarrow R(x) = 0)$, d.h. die Abwesenheit eines Objektes hat zwangsläufig auch die Abwesenheit jeglicher Relevanz in der Eingabe zur Folge, eine Kompensation durch positive und negative Relevanzen ist folglich nicht möglich.

Bemerkung 5.1.4. Die geforderten Eigenschaften an eine Heatmap definieren diese nicht eindeutig. Es sind also mehrere Abbildungen möglich, die die genannten Forderungen erfüllen. Beispiele dafür sind eine natürliche Zerlegung und Taylor-Zerlegungen [MLB⁺ 17a].

Die LRP liefert nun ein Konzept, mit dem eine Zerlegung

$$f(x) = \sum_d R_d \quad (5.3)$$

bestimmt werden kann.

TODO: Summenabfolge von layer zu layer einfügen

Wir gehen nun davon aus, dass die Funktion f ein NN repräsentiert, dass aus mehreren Schichten mit mehreren Neuronen pro Schicht und dazwischengeschalteten nicht-linearen Aktivierungsfunktionen aufgebaut ist. Die erste Schicht ist die Eingabe-Schicht, bestehend aus den Pixeln eines Bildes. Die letzte Schicht ist die reellwertige Ausgabe von f . Die l -te Schicht ist durch einen Vektor $z = (z_d^l)_{d=1}^{V(l)}$ der Dimension $V(l)$ dargestellt. Sei also eine Relevanz $R_d(l+1)$ für jede Dimension $z_d^{(l+1)}$ des Vektors z in der Schicht $l+1$ gegeben. Die Idee besteht nun darin, eine Relevanz $R_d^{(l)}$ für jede Dimension $z_d^{(l)}$ des Vektors z in der Schicht l zu finden, die einen Schritt näher an der Eingabeschicht liegt, sodass die folgende Abfolge von Gleichungen gilt:

$$f(x) = \dots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \dots = \sum_d R_d^{(1)}. \quad (5.4)$$

Für diese Funktion benötigen wir eine Regel, mit der die Relevanz eines Neurons einer höheren Schicht $R_j^{(l+1)}$ auf ein Neuron einer benachbarten, näher an der Eingabeschicht liegendes Neuron, übertragen werden kann. Die Übertragung der Relevanz zwischen zwei solchen Neuronen wird mit $R_{i \leftarrow j}$ bezeichnet. Auch hier muss die übertragene Relevanz erhalten bleiben. Es wird also gefordert:

$$\sum_i R_{i \leftarrow j}^{(l,l+1)} = R_j^{(l+1)}. \quad (5.5)$$

D.h. die gesamte Relevanz eines Neurons der Schicht $l+1$ verteilt sich komplett auf alle Neuronen der Schicht l . Im Falle eines linearen NN $f(x) = \sum_i z_{ij}$ mit der Relevanz $R_j = f(x)$ ist eine Zerlegung gegeben durch $R_{i \leftarrow j} = z_{ij}$. Im allgemeineren Fall ist die Neuronenaktivierung x_j eine nicht-lineare Funktion abhängig von z_j . Für die beiden Aktivierungsfunktionen $\tanh(x)$ und $\text{ReLU}(x)$ - beide monoton wachsend mit $g(0) = 0$ - bieten die Vor-Aktivierungen noch immer ein sinnvolles Maß für den relativen Beitrag eines Neurons x_i zu R_j (müsste das nicht umgekehrt sein, die INdizes?!?!).

Eine erste Mögliche Relevanz-Zerlegung, basierend auf dem Verhältnis zwischen lokalen und globalen Vor-Aktivierung, ist gegeben durch:

$$R_{i \leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j} \cdot R_j^{(l+1)}. \quad (5.6)$$

Für diese Relevanzen $R_{i \leftarrow j}$ gilt die Erhaltungseigenschaft 5.4, denn:

$$\sum_i R_{i \leftarrow j}^{(l,l+1)} = R_j^{l+1} \cdot \left(1 - \frac{b_j}{z_j}\right). \quad (5.7)$$

Dabei steht der rechte Faktor für die Relevanz, die durch den Bias-Term absorbiert wird. Falls notwendig, kann die verbleibende Bias-relevanz auf jedes Neuron x_i verteilt werden(?s.Abschnitt über Biases Promotion, S.Lapuschkin).

Diese Regel wird in der Liteartur als LRP-0 bezeichnet. Ein Nachteil dieser ist, dass die Relevanzen $R_{i \leftarrow j}$ für kleine globalen Voraktivierung z_j beliebig große Werte annehmen können.

Um dies zu verhindern, wird in der LRP- ε -Regel ein vorher festgelegter Parameter $\varepsilon > 0$ eingeführt:

$$R_{i \leftarrow j}^{(l,l+1)} = \begin{cases} \frac{z_{ij}}{z_j + \varepsilon} \cdot R_j^{(l+1)}, & z_j \geq 0 \\ \frac{z_{ij}}{z_j - \varepsilon} \cdot R_j^{(l+1)}, & z_j < 0 \end{cases} \quad (5.8)$$

In [BBM⁺15] wird die Layer-wise Relevance Propagation erstmalig vorgestellt. Zudem wird eine Taylor Zerlegung präsentiert, die eine Approximation der LRP darstellt.

Hier⁵ werden einige Bereiche vorgestellt, in denen LRP angewendet wurde.

5.2 Behandlung von biases

5.3 Beispiel an einem kleinen Netzwerk

5.4 Deep Taylor Decomposition

Mathematischer Hintergrund für LRP.LRP als Spezialfall von DTD

5.4.1 Taylor Decomposition

We will assume that the function $f(x)$ is implemented by a deep neural network, composed of multiple layers of representation, where each layer is composed of a set of neurons. Each neuron performs on its input an elementary computation consisting of a linear projection followed by a nonlinear activation function. Deep neural networks derive their high representational power from the interconnection of a large number of these neurons, each of them, realizing a small distinct subfunction.

Laut [MLB⁺17b] ist die in [BBM⁺15] vorgestellte Layer-wise Relevance Propagation eher heuristisch. In diesem Paper wird nun eine solide theoretische Grundlage geliefert.

DTD liefert den mathematischen Hintergrund für LRP

⁵<https://towardsdatascience.com/indepth-layer-wise-relevance-propagation-340f95deb1ea>

Simple Taylor decomposition. Finde rootpoints, sodass Erhaltungseigenschaft erhalten bleibt.

Simple Taylor in Practice: funktioniert in der Praxis nicht wirklich. Viel Noise meistens positive Relevanz

Relevanz Propagation: Heatmaps look much cleaner

Simple Taylor:- root point hard to find -gradient shattering. Gradient loses its informative structure in big layer nets

Use Taylor Decomposition to explain LRP from layer to layer

5.4.2 Deep Taylor Decomposition

LRP in verschiedenen Anwendungsgebieten [MBL⁺19], 10.2. In diesem Paper: LRP-0 schlechter als LRP- ε schlechter als LRP- γ schlechter als Composite-LRP.

5.5 Verschiedene Verfahren

5.6 Eigenschaften

Beweise in DTD Paper

- Numerische Stabilität
- Konsistenz (mit Linearer Abbildung)
- Erhaltung der Relevanz

5.7 Behandlung besonderer Schichten

5.7.1 BatchNorm2D

5.8 LRP für Deep Neural Nets/Composite LRP

5.9 Verarbeitung der Heatmaps

Aktuell benutzte default colormap ist Option D (Viridis)⁶

- Wertebereich
- Interpretation
- Skalen
- Normalisierung

5.10 Implementierung

5.10.1 Tensorflow

5.10.2 pytorch

Allgemeines Tutorial:⁷

pytorch-LRP für VGG16 wird vorgestellt.

⁶<https://bids.github.io/colormap/>

⁷<https://git.tu-berlin.de/gmontavon/lrp-tutorial>



(a) Verkehrsschild der Klasse 'Höchstgeschwindigkeit: 50km/h' versehen mit einem 3x3 Sticker und dem Label 'Höchstgeschwindigkeit: 80km/h' (b) Zugehörige Heatmap bezüglich der Klasse 'Höchstgeschwindigkeit: 80km/h'

Abbildung 1: (Optischer) Vergleich von korruptem Datenpunkt und berechneter Heatmap.

GiorgioML⁸:

Alternative pytorch-Implementierung basierend auf Tensorflow paper.

moboehle⁹:

Der code entstand im Rahmen der Forschungsarbeit [BEWR19], in der eine Alzheimer-Feststellung aufgrund von Bilddaten(scans?) vorgenommen wird. Framework leicht anpassbar. Benutzt pytorch hooks. Unterstützte Netzwerkschichten¹⁰:

```

1 torch.nn.BatchNorm1d,
2 torch.nn.BatchNorm2d
3 torch.nn.BatchNorm3d,
4 torch.nn.ReLU,
5 torch.nn.ELU,
6 Flatten,
7 torch.nn.Dropout,
8 torch.nn.Dropout2d,
9 torch.nn.Dropout3d,
10 torch.nn.Softmax,
11 torch.nn.LogSoftmax,
12 torch.nn.Sigmoid
13
14
```

Listing 2: Verfügbare Schichten und Aktivierungsfunktionen

fhvilshoj¹¹:

LRP für linear und Convolutional layers

- Die Klassen

torch.nn.Sequential, torch.nn.Linear und torch.nn.Conv2d werden erweitert, um autograd für die Berechnung der Relevanzen zu berechnen.

⁸<https://giorgiomorales.github.io/Layer-wise-Relevance-Propagation-in-Pytorch/>

⁹<https://github.com/moboehle/Pytorch-LRP>

¹⁰https://github.com/moboehle/Pytorch-LRP/blob/master/inverter_util.py

¹¹<https://github.com/fhvilshoj/TorchLRP>

- Ausgabe der Relevanzen von Zwischenschichten ist möglich
- : Implementierte Regeln: epsilon Regeln mit $\epsilon=1e-1$, gamma-regel mit $\gamma=1e-1$. alphabeta-Regel mit $\alpha_1\beta_0$ und $\alpha_2\beta_1$
- Netz muss hier umgeschrieben werden, sodass die Anwendung des Algorithmus möglich wird.

```

1  conv2d = {
2
3  "gradient":          F.conv2d,
4  "epsilon":          Conv2DEpsilon.apply,
5  "gamma":            Conv2DGamma.apply,
6  "gamma+epsilon":    Conv2DGammaEpsilon.apply,
7  "alpha1beta0":      Conv2DAlpha1Beta0.apply,
8  "alpha2beta1":      Conv2DAlpha2Beta1.apply,
9  "patternattribution": Conv2DPatternAttribution.apply,
10 "patternnet":       Conv2DPatternNet.apply,
11 }
12
13

```

Listing 3: Implementierte Regeln fhvilshoj

Zennit:¹² Zennit (Zennit explains neural networks in torch)

- Modell wird mithilfe eines Canonizers so aufbereitet, dass LRP möglich wird
- Backward pass wird modifiziert, um Heatmaps zu erhalten.
- VGG- und ResNet-Beispiel

6 Detektion von Poisoning-Angriffen basierend auf LRP

6.1 Idee

Die Idee zur Detektion von Poisoning-Angriffen besteht aus den folgenden Schritten:

- Berechnung der Heatmaps mit Hilfe der LRP
- Berechnung einer Distanzmatrix basierend auf L^2 - oder GMW-Distanz
- Spektrale Relevanzanalyse (Bestimmung der verschiedenen Cluster innerhalb einer Klasse)

Bemerkung: Anstatt das Clustering nur auf den Heatmaps durchzuführen, könnten die LRP-Ausgaben und/oder Aktivierungen bestimmter Netzwerkschichten hinzugenommen werden.

The theory of optimal transport generalizes that intuition in the case where, instead of moving only one item at a time, one is concerned with the problem of moving simultaneously several items (or a continuous distribution thereof) from one configuration onto another. [com19]

¹²<https://github.com/chr5tphr/zennit>

6.2 k-means / k-means++ -Clustering

Beispiel-Implementierung¹³

Baryzentrische Koordinaten¹⁴

6.3 Spektrales Clustering

Wir folgen [VL07]. Gegeben: Datenpunkte x_i, \dots, x_n sowie eine Größe $s = s_{ij} \in \mathbb{R}^+$, die einen paarweisen Zusammenhang der einzelnen Punkte beschreiben.

Ziel: Aufteilen der Punkte in verschiedene Cluster, sodass sich Punkte innerhalb eines Clusters ähnlich bezüglich s sind.

Alternative Repräsentation der Daten mithilfe eines Ähnlichkeitsgraphen $G = (V, E)$ möglich.

Umformulierung des Clustering-Problems mithilfe des Ähnlichkeitsgraphen: Finde Partitionierung des Graphen, sodass die Kanten-Gewichte innerhalb einer Gruppe niedrig (niedriges Gesamtgewicht?) und außerhalb einer Gruppe groß sind.

Graph-Notationen:

Verschiedene Konstruktionsmöglichkeiten von Ähnlichkeitsgraphen:

- ε -Nachbarschaft-Graph
- kNN-Graph
- fully connected graph

6.4 Anwendung auf unterschiedliche Poisoning-Angriffe

Berechnung der Relevanzen:

Wir berechnen die Relevanzen jedes einzelnen Eingabebildes klassenweise, d.h. besitzt eine Eingabe das Label y , so berechnen auf einem trainierten Netzwerk, für jeden Pixelwert der Eingabe, wie relevant dieser für die Ausgabe $f(x) = y$ ist.

Wir summieren über die Farboxen des Bildes, um einzelne Relevanzen pro Pixelpunkt zu erhalten.

Für die Berechnung der Relevanzen benutzen wir eine modifizierte Version des im Rahmen von [BEWR19] entstandenen Programmcodes¹⁵.

Vorverarbeitung der Relevanzen:

In [LWB⁺19] wird anschließend ein Sum-Pooling auf die Relevanzen angewendet, um eine Dimensionsreduktion zu erhalten. Wie in [AMN⁺19] verzichten wir auf eine weitere Dimensionsreduktion, da wir nur relativ kleine Relevanzen der Größe 32×32 verarbeiten.

¹³<https://towardsdatascience.com/k-means-implementation-in-python-and-spark-856e7eb5fe9b>

¹⁴https://de.wikipedia.org/wiki/Baryzentrische_Koordinaten

¹⁵<https://github.com/moboehle/Pytorch-LRP>

Für $\text{eps}=5\text{e-}2$ liegen beide barycentren identisch weit weg. Probiere nun $\text{eps}=5\text{e-}3$

Berechnung der Distanzen und Aufstellen einer Affinitätsmatrix:

Wir berechnen zunächst eine Distanzmatrix, die die paarweisen Distanzen aller Heatmaps einer Klasse enthält.

Für die Berechnung der euklidischen Distanz betrachten wir Heatmaps x, y der Größe 32×32 als Elemente $x, y \in \mathbb{R}^{32 \times 32}$. Die Distanz lässt sich dann wie in Unterabschnitt 6.5.2 berechnen.

Die Gromov-Wasserstein-Distanz lässt sich wie in [PCS16] angegeben berechnen.

Barycenters Definition und Vergleich zum euklidischen Raum [AC11]
In einer Affinitätsmatrix oder Ähnlichkeitsmatrix sind die

Berechnung Spektralen Einbettung:

Dimensionsreduktion vor dem Clustering ?!

In [CCB⁺18] wird beispielsweise eine Dimensionsreduktion mit PCA durchgeführt.

k-Means-Clustering:

Bemerkung 6.4.1. In [AMN⁺19] Kapitel '2.3. Fisher Discriminant Analysis for Clever Hans identification' wird ein Verfahren vorgestellt, mit dem verdächtige Klassen identifiziert werden können. Für diese würde man anschließend das obige Verfahren durchführen

6.5 Verwendete Distanzen & Approximationen

Um die Struktur innerhalb einer Klasse zu analysieren, benötigen wir eine Metrik. Anhand dieser wird abhängig von den Heatmaps einer Klasse eine Affinitätsmatrix berechnet, die dann anschließend zur Berechnung der Spektralen Einbettung als wichtigster Schritt von SpRAY verwendet wird. Wir wollen dazu die im Folgenden vorgestellten Metriken verwenden.

Wie in [AMN⁺19] summieren wir über die Farbkanäle, um einen einzelnen Relevanzwert pro Pixelpunkt zu erhalten. Wir benötigen also eine Metrik zur Berechnung der Distanz zwischen 32×32 großen Heatmaps.

Wir normalisieren die Relevanzen zusätzlich auf das Intervall $[0, 1]$.

6.5.1 Histogramme & Maße

Definition 6.5.1 (Histogramm). Als Histogramm (oder: Wahrscheinlichkeitsvektor) bezeichnen wir ein Element $\mathbf{a} \in \Sigma_n$, das zum folgenden Wahrscheinlichkeits-Simplex gehört:

$$\Sigma_n := \{\mathbf{a} \in \mathbb{R}_+^n \mid \sum_{i=1}^n a_i = 1\}$$

Einführung W-Theorie am KIT ¹⁶

Definition 6.5.2 (Kopplungen zwischen Histogrammen). *Seien die beiden Histogramme $p \in \Sigma_{N_1}$ und $q \in \Sigma_{N_2}$ gegeben. Als Menge der Kopplungen zwischen beiden Histogrammen definieren wir*

$$\mathcal{C}_{p,q} := \{T \in (\mathbb{R}_+)^{N_1 \times N_2} \mid T\mathbb{K}_{N_2} = p, T^\top \mathbb{K}_{N_1} = q\},$$

wobei $\mathbb{K} := (1, \dots, 1)^\top \in \mathbb{R}^N$ gilt.

Definition 6.5.3 (Distanz). *Eine Distanz*

Definition 6.5.4. *Eine Metrik*

[Metrik]

6.5.2 Euklidische Distanz

Für den Fall der euklidischen Distanz schreiben wir die Relevanzwerte pro Pixel als Vektor und berechnen die Distanz zweier Heatmaps x und y wie folgt¹⁷:

$$d_{x,y} = \sqrt{\sum_{i=1}^{32 \times 32} (x_i - y_i)^2}.$$

6.5.3 Gromov-Hausdorff-Distanz

Definition 6.5.5 (Hausdorff-Distanz). *Sei (M, d) ein metrischer Raum. Für Seien $X, Y \subset (M, d)$ definieren wir die Hausdorff-Distanz $d_H(X, Y)$ als*

$$d_H(X, Y) = \max\left\{\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y)\right\}. \quad (6.1)$$

Definition 6.5.6 (Metrischer Maßraum). *Ein metrischer Maßraum ist ein Tripel (X, d_X, μ_X) , wobei (X, d_X) ein metrischer Raum und μ_X ein borelsches W -Maß auf X ist.*

Definition 6.5.7 (Correspondance). *content...*

Definition 6.5.8 (Kopplung). *Seien*

Seien (X, d_X) und (Y, d_Y) zwei metrische Räume. Wir betrachten im Folgenden die Abbildung

$$\Gamma_{X,Y} : (X \times Y) \times (X \times Y) \rightarrow \mathbb{R}^+, \quad (6.2)$$

gegeben durch

$$\Gamma_{X,Y}(x, y, x', y') := |d_X(x, x') - d_Y(y, y')|.$$

Definition 6.5.9 (Gromov-Hausdorff-Distanz). *Für die metrischen Räume (X, d_X) und (Y, d_Y) ist die Gromov-Hausdorff-Distanz definiert als*

$$d_{\mathcal{GH}} = \frac{1}{2} \inf_R \|\Gamma_{X,Y}\|_{L^\infty(R \times R)}. \quad (6.3)$$

¹⁶<https://www.math.kit.edu/stoch/~henze/media/wt-ss15-henze-handout.pdf>

¹⁷<https://paulrohan.medium.com/euclidean-distance-and-normalization-of-a-vector-76f7a97abd9>

Definition 6.5.10 (Entropischer Optimaler Transport). *Wir definieren den n -dimensionalen Zufalls-Simplex als $\Sigma_n := \{a \in \mathbb{R}_+^n : \sum_{i=1}^n a_i = 1\}$. Ein Element $a \in \Sigma_n$ bezeichnen wir als Histogramm oder Zufallsvektor.*

Definition 6.5.11 (Diskretes Maß).

6.5.4 Optimaler Transport (Monge Formulierung)

Bemerkung 6.5.12. *Problem zwischen diskreten Maßen*

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) : T_{\#}\alpha = \beta \right\} \quad (6.4)$$

Bemerkung 6.5.13 (Fehlende Eindeutigkeit).

Definition 6.5.14 (Push-forward Operator). $\beta(B) = \alpha(\{x \in \mathcal{X} : T(x) \in B\}) = \alpha(T^{-1}(B))$

Wir können das Monge Problem wie folgt auf den Fall zweier beliebiger W-Maße (α, β) erweitern.

Definition 6.5.15 (Monge Problem zwischen beliebigen Maßen). *Seien α und β zwei W-Maße mit Support auf den Räumen \mathcal{X} bzw. \mathcal{Y} , die durch $T : \mathcal{X} \rightarrow \mathcal{Y}$ verknüpft(? Formulierung) sind. Dann ist das Monge Problem gegeben durch*

$$\min_T \left\{ \int_{\mathcal{X}} c(x, T(x)) d\alpha(x) : T_{\#}\alpha = \beta \right\}. \quad (6.5)$$

Die Bedingung $T_{\#}\alpha = \beta$ bedeutet, dass die Abbildung T die gesamte Masse mithilfe des Push-forward Operators von α auf β schiebt.

6.5.5 Optimaler Transport nach Kantorovich

Das Optimal Transport Problem nach Kantorovich gehört zu den typischen Optimal Transport Problemen. Es stellt einer Relaxierung der Formulierung von Gaspard Monge[1781], bei der nun auch das Aufteilen von Masse (mass splitting) zulässig ist. In Kantorovichs Formulierung ist eine Kopplung (oder: Transportabbildung) \mathbf{T} gesucht, die die Kosten, die bei der Verschiebung eines diskreten Maßes \mathbf{a} auf ein anderes diskretes Maß \mathbf{b} bezüglich der Kosten $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ entstehen, minimiert. Damit \mathbf{T} eine Transportabbildung ist, muss $\mathbf{T} \in \Gamma(\mathbf{a}, \mathbf{b}) = \{\mathbf{T} \geq \mathbf{0}, \mathbf{T}\mathbf{1}_{n_2} = \mathbf{a}, \mathbf{T}^T \mathbf{1}_{n_1} = \mathbf{b}\}$ gelten.

Für den Fall, dass die Grundkosten eine Metrik darstellen, ist auch die optimale Lösung des Optimal Transport Problems wieder eine Metrik [?] und definiert die *Wasserstein Distanz*. Das OT Problem ist definiert als

$$W_M(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{T}, \mathbf{M} \rangle, \quad (6.6)$$

wobei $\langle \mathbf{T}, \mathbf{M} \rangle = \sum_{ij} t_{ij} m_{ij}$ gilt.

which is a linear program. The optimization problem above is often adapted to include a regularization term for the transport plan \mathbf{T} , such as entropic regularization (Cuturi, 2013) or squared L2. For the entropic regularized OT problem, one

may use the Sinkhorn Knopp algorithm (or variants), or stochastic optimization algorithms. POT has a simple syntax to solve these problems (see Sample 1)

Die Menge der Matrizen $\Pi(\mathbf{a}, \mathbf{b})$ ist beschränkt und durch $n + m$ Gleichungen gegeben und damit ein konvexes Polytop (die konvexe Hülle einer endlichen Menge von Matrizen). Zudem ist die Formulierung von Kantorovich im Unterschied zu Monges Formulierung immer symmetrisch in dem Sinne, dass $\mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b})$ genau dann gilt, wenn $\mathbf{P}^\top \in \Pi(\mathbf{b}, \mathbf{a})$.

Kantorovichs Optimal Transport Problem lässt sich nun schreiben als

$$L_C(\mathbf{a}, \mathbf{b}) := \min_{\mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle := \sum_{i,j} \mathbf{C}_{i,j} \mathbf{P}_{i,j} \quad (6.7)$$

Diese Problem ist ein lineares Problem. Für diese Art von Problemen ist die optimale Lösung nicht notwendigerweise eindeutig. In fact Kantorovich is considered as the inventor of linear programming.¹⁸

EXISTENZ & Eindeutigkeit => vgl. Kapitel 3 in [com19]

Definition 6.5.16 (Formulierung für beliebige Maße). *Im Fall beliebiger Maße betrachten wir die Kopplungen $\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$, die die gemeinsame Verteilung auf dem Produktraum $\mathcal{X} \times \mathcal{Y}$ ist. Im diskreten Fall verlangen wir, dass das Produktmaß die Form $\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_j)}$ besitzt. Im Allgemeinen Fall wird die Massenerhaltung als Randbedingung an die gemeinsame W-Verteilung geschrieben:*

$$\Pi(\alpha, \beta) := \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : P_{\mathcal{X}\#} \pi = \alpha \text{ und } P_{\mathcal{Y}\#} \pi = \beta\}, \quad (6.8)$$

wobei $P_{\mathcal{X}\#}$ und $P_{\mathcal{Y}\#}$ die Push-forward Operatoren der Projektionen $P_{\mathcal{X}}(x, y) = x$ und $P_{\mathcal{Y}}(x, y) = y$ sind. Nach Theorem 6.5.14 sind diese Randbedingungen äquivalent zu den Bedingungen $\pi(A \times \mathcal{Y}) = \alpha(A)$ und $\pi(\mathcal{X} \times B) = \beta(B)$ für die Mengen $A \subset \mathcal{X}$ und $B \subset \mathcal{Y}$. Als Verallgemeinerung erhalten wir dann

$$\min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y). \quad (6.9)$$

Problem Gleichung 6.9 ist ein unendlich-dimensionales lineares Programm über einem Raum von Maßen. Falls $(\mathcal{X}, \mathcal{Y})$ kompakt und c stetig ist, existiert immer eine Lösung.

Beispiel 6.5.17. *Beispiele von Kopplungen*

6.5.6 Monge-Kantorovitch equivalence

The proof of Brenier theorem 1 (detailed in Section 5.3) to prove the existence of a Monge map actually studies Kantorovitch relaxation, and proves that this relaxation is tight in the sense that it has the same cost as Monge problem.¹⁹

6.5.7 Metrische Eigenschaften

Optimaler Transport definiert eine Distanz zwischen Histogrammen und W-Maßen, sofern die Kostenmatrix gewisse Eigenschaften erfüllt. Optimal Transport kann dabei als naheliegende Idee verstanden werden, um Distanzen zwischen Punkten auf Distanzen zwischen Histogrammen oder Maßen zu verallgemeinern.

¹⁸OTNotes_campriede.pdf

¹⁹CourseOT_cuturi

Definition 6.5.18 (p-Wasserstein-Distanz auf Σ_n). Sei $n = m$ und für $p \geq 1$ gelte $C = D^p = (D_{i,j}^p)_{i,j} \in \mathbb{R}^{n \times n}$, wobei $D \in \mathbb{R}_+^{n \times n}$ eine Distanz ist.

Dann definiert

$$W_p(\mathbf{a}, \mathbf{b}) := L_{D^p}(\mathbf{a}, \mathbf{b})^{1/p} \quad (6.10)$$

die p -Wasserstein-Distanz auf Σ_n

Lemma 6.5.19. W_p ist eine Distanz, d.h. W_p ist symmetrisch, positiv, es gilt $W_p(a, b) = 0$ gdw. $a = b$ und die Dreiecksungleichung

$$\forall \mathbf{a}, \mathbf{b}, \mathbf{c} \in \Sigma_n : W_p(\mathbf{a}, \mathbf{c}) \leq W_p(\mathbf{a}, \mathbf{b}) + W_p(\mathbf{b}, \mathbf{c}). \quad (6.11)$$

Beweis. Inhalt... □

Bemerkung 6.5.20 (Der Fall $0 < p \leq 1$). Für $0 < p \leq 1$ ist auch D^p eine Distanz. Damit ist für $p \leq 1$ $W_p(\mathbf{a}, \mathbf{b})^p$ eine Distanz auf dem Simplex.

6.5.8 Duale Formulierung

Kurzer Überblick hier²⁰

The Kantorovich problem (2.11) is a constrained convex minimization problem, and as such, it can be naturally paired with a so-called dual problem, which is a constrained concave maximization problem. The following fundamental proposition explains the relationship between the primal and dual problems.

6.5.9 Gromov-Wasserstein-Divergenz

Fast Computation of Wasserstein Barycenters²¹

Definition 6.5.21 (Divergenz). Sei S der Raum aller Wahrscheinlichkeitsverteilungen mit gemeinsamem Support. Dann bezeichnet die Divergenz auf S eine Funktion $D(\cdot || \cdot) : S \times S \rightarrow \mathbb{R}$, für die gilt:

1. $D(p || q) \geq 0$ f.a. $p, q \in S$

2. $D(p || q) = 0$ gdw. $p = q$.

Definition 6.5.22 (Entropie). Für $T \in \mathbb{R}_+^{N \times N}$ definieren wir die Entropie als

$$H(T) := - \sum_{i,j=1}^N T_{i,j} (\log(T_{i,j}) - 1). \quad (6.12)$$

Definition 6.5.23 (Tensor-Matrix-Multiplikation). Für einen Tensor $\mathcal{L} = (\mathcal{L}_{i,j,k,l})_{i,j,k,l}$ und eine Matrix $(T_{i,j})_{i,j}$ definieren wir die Tensor-Matrix-Multiplikation als

$$\mathcal{L} \otimes T := \left(\sum_{k,l} \mathcal{L}_{i,j,k,l} T_{k,l} \right)_{i,j}. \quad (6.13)$$

²⁰<https://arxiv.org/pdf/1609.04767.pdf>

²¹<https://arxiv.org/pdf/1310.4375.pdf>

6.5.10 Regularisierungen

Numerical Methods: Cuturi's Entropy Regularised Approach. Arguably the biggest development (at least in recent years) in the computation of optimal transport distances was due to Cuturi's entropy regularised approach. The idea is to use entropy to regularise the distance, then some simple rearrangements reveal this is a Kullback-Liebler divergence. Standard methods, e.g. Sinkhorn's algorithm, can then be used to find minimizers of the entropy regularised distance.²²

6.5.11 Entropisch Regularisierte Gromov-Wasserstein-Distanz

Entropic Regularization of Optimal Transport [com19] Mehrere Möglichkeiten einer Regularisierung der GW-Distanz²³:

- Entropic regularization [Cuturi, 2013]
- Group Lasso [Courty et al., 2016a]
- KL, Itakura Saito, β -divergences, [Dessein et al., 2016]

Optimal Transport: Regularization and Applications ²⁴ Python Optimal Transport Toolbox^{25,26}

Kapitel 2.1 im Paper: Vergleich von Histogrammen auf demselben metrischen Raum.

Lemma 6.5.24. $L_C(\mathbf{a}, \mathbf{b}) := \min_{P \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P})$

ist ein ε -streng konvexes Problem und besitzt deshalb eine eindeutige optimale Lösung

Beweis. todo. □

Proposition 6.5.25 (Konvergenz in ε). Die eindeutige Lösung P_ε von Theorem 6.5.24 konvergiert gegen die optimale Lösung mit maximaler Entropie innerhalb der Menge aller optimalen Lösungen des Kantorovich Problems, d.h.

$$P_\varepsilon \xrightarrow{\varepsilon \rightarrow 0} \arg \min_P \{ -H(\mathbf{P}) : \mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b}), \langle \mathbf{P}, \mathbf{C} \rangle = L_C(\mathbf{a}, \mathbf{b}) \}, \quad (6.14)$$

d.h. es gilt

$$L_C^\varepsilon(\mathbf{a}, \mathbf{b}) \xrightarrow{\varepsilon \rightarrow 0} L_C(\mathbf{a}, \mathbf{b}) \quad (6.15)$$

Zudem gilt

$$P_\varepsilon \xrightarrow{\varepsilon \rightarrow \infty} \mathbf{a} \otimes \mathbf{b} = \mathbf{a} \mathbf{b}^\top = (\mathbf{a}_i \mathbf{b}_j)_{i,j}. \quad (6.16)$$

Beweis. todo:Abend. s. COT, S.59 □

²²https://www.damtp.cam.ac.uk/research/cia/files/teaching/Optimal_Transport_Syllabus.pdf

²³<https://www.youtube.com/watch?v=cPVMHWF8fmE&t=2532s>

²⁴<https://www.otra2020.com/schedule>

²⁵<https://pythonot.github.io/quickstart.html>

²⁶https://pythonot.github.io/auto_examples/gromov/plot_gromov.html

Bemerkung 6.5.26. Gleichung 6.14 zeigt, dass die Lösung für kleine ε gegen die Optimale Transport Kopplung mit maximaler Entropie konvergiert. Im Gegensatz dazu, bedeutet Gleichung 6.16, die Lösung für große Regularisierungsparameter konvergiert gegen die Kopplung mit maximaler Entropie zwischen zwei gegebenen Randverteilungen \mathbf{a} und \mathbf{b} .

Bemerkung 6.5.27. A key insight is that, as ε increases, the optimal coupling becomes less and less sparse (in the sense of having entries larger than a prescribed threshold), which in turn has the effect of both accelerating computational algorithms (as we study in Abschnitt 4.2) and leading to faster statistical convergence (as shown in Abschnitt 8.5)

Wir definieren die Kullback-Leibler-Divergenz zwischen Kopplungen als

$$KL(\mathbf{P}|\mathbf{K}) := \sum_{i,j} \mathbf{P}_{i,j} \log \left(\frac{\mathbf{P}_{i,j}}{\mathbf{K}_{i,j}} \right) - \mathbf{P}_{i,j} + \mathbf{K}_{i,j}. \quad (6.17)$$

Damit ist die eindeutige Lösung \mathbf{P}_ε von Theorem 6.5.24 eine Projektion des zur Kostenmatrix \mathbf{C} gehörigen Gibbs-Kernels $\mathbf{K}_{i,j} := e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}}$ auf $\Pi(\mathbf{a}, \mathbf{b})$.

Mit der obigen Definition erhalten wir

$$\mathbf{P}_\varepsilon = \text{Proj}_{\Pi(\mathbf{a}, \mathbf{b})}^{KL}(\mathbf{K}) := \arg \min_{\mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b})} KL(\mathbf{P}|\mathbf{K}). \quad (6.18)$$

6.5.12 Berechnung der Lösung: Sinkhorn

In diesem Kapitel sehen wir, dass die Lösung des regularisierten Problems eine besondere Form besitzt, die wir über $m + n$ Variablen parametrisieren können.

Proposition 6.5.28. Die Lösung des regularisierten Problems Theorem 6.5.24 besitzt die Form

$$\forall (i, j) \in \{1, \dots, n\} \times \{1, \dots, m\} : \mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j \quad (6.19)$$

für die beiden (unbekannten) Variablen $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$.

Beweis. Wir führen für jede der beiden Nebenbedingungen die dualen Variablen $\mathbf{f} \in \mathbb{R}^n$ und $\mathbf{g} \in \mathbb{R}^m$ ein. Für die Lagrange-Funktion zu Theorem 6.5.24 erhalten wir damit:

$$\mathcal{L}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon \mathbf{H}(\mathbf{P}) - \langle \mathbf{f}, \mathbf{P} \mathbf{1}_m - \mathbf{a} \rangle - \langle \mathbf{g}, \mathbf{P}^\top \mathbf{1}_n - \mathbf{b} \rangle. \quad (6.20)$$

Mit der Optimalitätsbedingung erster Ordnung ergibt sich

$$\frac{\partial \mathcal{L}(\mathbf{P}, \mathbf{f}, \mathbf{g})}{\partial \mathbf{P}_{i,j}} = \mathbf{C}_{i,j} + \varepsilon \log(\mathbf{P}_{i,j}) - \mathbf{f}_i - \mathbf{g}_j = 0, \quad (6.21)$$

womit wir für eine optimale Kopplung \mathbf{P} für das regularisierte Problem den Ausdruck $\mathbf{P}_{i,j} = e^{\mathbf{f}_i/\varepsilon} e^{-\mathbf{C}_{i,j}/\varepsilon} e^{\mathbf{g}_j/\varepsilon}$, der in die gewünschte Form umgeschrieben werden kann. \square

Bemerkung 6.5.29 (Algorithmus von Sinkhorn). ...

Bemerkung 6.5.30 (Allgemeine Formulierung des entropisch regularisierten Problems für beliebige Maße).

- Benutze Relative Entropie als Verallgemeinerung der diskreten Kullback-Leibler-Divergenz
- Referenzmaß ist unbedeutend, lediglich er Support hat eine Bedeutung.

6.5.13 Verallgemeinerung

Der Vergleich zwischen Ähnlichkeits- bzw. Distanzmatrizen ist schwierig, da diese die innere Struktur eines Datensatzes beschreiben, die unabhängig von Rotationen und Translationen ist. Es existiert keine kanonische Ordnung der Reihen und Spalten. Verallgemeinerung auf beliebige Matrizen C , d.h. diese Distanzmatrizen müssen nicht notwendigerweise positiv sein und die Dreiecksungleichung erfüllen.

Definiere die verallgemeinerte Gromov-Wasserstein-Distanz (vGWD) wie folgt:

Definition 6.5.31 (Verallgemeinerte Gromov-Wasserstein-Distanz). *Seien zwei gewichtete Ähnlichkeitsmatrizen $(C, p) \in \mathbb{R}^{N_1 \times N_1} \times \Sigma_{N_1}$ und $(\bar{C}, q) \in \mathbb{R}^{N_2 \times N_2} \times \Sigma_{N_2}$ gegeben. Sei T eine Kopplung zwischen den beiden Räumen, auf denen die Matrizen C und \bar{C} definiert sind. Sei L eine Fehlerfunktion. Dann definieren wir die verallgemeinerte Gromov-Wasserstein-Distanz als*

$$GW(C, \bar{C}, p, q) := \min_{T \in \mathcal{C}_{p,q}} \varepsilon_{C, \bar{C}}(T), \quad (6.22)$$

wobei gilt $\varepsilon_{C, \bar{C}}(T) := \sum_{i,j,k,l} L(C_{i,k}, \bar{C}_{j,l}) T_{i,j} T_{k,l}$.

Häufig verwendete Fehlerfunktionen sind die quadratische Fehlerfunktion $L(a, b) = L_2(a, b) := \frac{1}{2}|a - b|^2$ und die Kullback-Leibler-Divergenz $L(a, b) = KL(a|b) := a \log(a/b) - a + b$.

Diese Definition der Gromov-Wasserstein-Distanz verallgemeinert die Version in [PCS16], da nun beliebige Fehlerfunktionen betrachtet werden.

Für $L = L_2$ zeigt Memoli, 2011, dass $GW^{1/2}$ eine Distanz auf dem Raum metrischer Maßräume modulo Maß-erhaltender Isometrien (? , besser zitieren -> vollständiges Resultat angeben) definiert.

Durch die Definition

$$\mathcal{L}(C, \bar{C}) := (L(C_{i,k}, \bar{C}_{j,l}))_{i,j,k,l} \quad (6.23)$$

erhalten wir

$$\epsilon_{C, \bar{C}}(T) = \langle \mathcal{L}(C, \bar{C}) \otimes T, T \rangle \quad (6.24)$$

gilt.

Mit der folgenden Proposition ergibt sich eine effiziente Berechnung von $\mathcal{L}(C, \bar{C}) \otimes T$ für eine bestimmte Klasse von Verlustfunktionen L :

Proposition 6.5.32. *Die Verlustfunktion L lasse sich schreiben als*

$$L(a, b) = f_1(a) + f_2(b) - h_1(a)h_2(b) \quad (6.25)$$

für $f_1, f_2, h_1, h_2 : \mathbb{R} \rightarrow \mathbb{R}$.

Dann gilt für $T \in \mathcal{C}_{p,q}$: $\mathcal{L}(C, \bar{C}) \otimes T = c_{C, \bar{C}} - h_1(C)Th_2(\bar{C})^T$,

wobei $c_{C,\bar{C}} := f_1(C)p\mathbf{1}_{N_2}^T + \mathbf{1}_{N_1}q^T f_2(\bar{C})^T$ unabhängig von T ist.

Beweis. todo. □

Wir betrachten nun die regularisierte Version der Gromow-Wasserstein-Diskrepanz 6.22 und definieren:

Definition 6.5.33. Für C, \bar{C}, p, q , wie oben, definieren wir die entropisch regularisierte Gromow-Wasserstein-Diskrepanz als

$$GW_\varepsilon(C, \bar{C}, p, q) := \min_{T \in \mathcal{C}_{p,q}} \varepsilon_{C,\bar{C}}(T) - \varepsilon H(T). \quad (6.26)$$

Wir erhalten damit ein nicht-konvexes Optimierungsproblem. Für dessen Lösung benutzen wir ein projiziertes Gradienten-Verfahren, bei dem sowohl die Schrittweite als auch die Projektion bezüglich der KL-Metrik berechnet werden.

Die Iterationen sind gegeben durch

$$T \leftarrow \text{Proj}_{\mathcal{C}_{p,q}}^{KL} \left(T \odot e^{-\tau(\nabla \varepsilon_{C,\bar{C}}(T) - \varepsilon H(T))} \right), \quad (6.27)$$

wobei die Schrittweite $\tau > 0$ und die KL-Projektion einer beliebigen Matrix K gegeben ist durch:

$$\text{Proj}_{\mathcal{C}_{p,q}}^{KL}(K) := \arg \min_{T' \in \mathcal{C}_{p,q}} KL(T'|K). \quad (6.28)$$

Proposition 6.5.34. Für den Fall $\tau = 1/\varepsilon$ erhalten wir die Iterationsvorschrift

$$T \leftarrow \mathcal{T}(\mathcal{L}(C, \bar{C}) \otimes T, p, q). \quad (6.29)$$

Beweis. □

Bemerkung 6.5.35. Die Iterationsvorschrift 6.29 definiert einen einfachen Algorithmus, der in jedem Update von T eine Sinkhorn-Projektion benötigt.

Bemerkung 6.5.36 (Konvergenz).

Bemerkung 6.5.37. content...

Bemerkung 6.5.38 (Wahl von ε). In [Cut13] werden verschiedene Werte für ε angegeben. Diese sind $\varepsilon = 0.02, 0.1, 1.0$

Im zugehörigen Beispiel²⁷ von POT wird $\varepsilon = 0.0005$ verwendet.

Für ε klein werden die Ergebnisse besser während der Rechenaufwand steigt. welche Resultate existieren bezüglich der Approximationsgüte abhängig von ε ?

6.5.14 Wasserstein Baryzentren

In diesem Kapitel wollen wir uns mit dem "Mittelwert" befassen, der elementar für das kmeans-Clustering ist. Auch hier soll der Mittelwert auf die abstrakte Ebene von gewichteten Distanzmatrizen angehoben werden. Dazu definieren wir im Folgenden sogenannte Wasserstein-Baryzentren und folgen [PCS16] für die Lösung des entstehenden Optimierungsproblems. Gute Einführung²⁸

²⁷https://pythonot.github.io/auto_examples/gromov/plot_gromov.html

²⁸<https://hal.archives-ouvertes.fr/hal-00637399/document>

Definition 6.5.39 (Gromov-Wasserstein Baryzentrum). *Seien die gewichteten Distanzmatrizen $C_s)_{s=1}^S$, mit $C_s \in \mathbb{R}^{N_s \times N_s}$ und den zugehörigen Histogrammen $(p_s)_s$ gegeben.*

Dann ist das Gromov-Wasserstein Baryzentrum definiert durch

$$\min_{C \in \mathbb{R}^{N \times N}} \sum_s \lambda_s \text{GW}_\varepsilon(C, C_s, p, p_s). \quad (6.30)$$

Bemerkung 6.5.40 (Darstellung des Baryzentrums). *20 Um das berechnete Bary-Zentrum C wieder zu visualisieren, kann C als Distanzmatrix wieder in den zwei-dimensionalen Raum eingebettet werden. Dies kann beispielsweise durch Multi-dimensionale Skalierung erreicht werden²⁹.*

Bemerkung 6.5.41. *Wir gehen im Folgenden davon aus, dass sowohl die Histogramme $(p_s)_s$ als auch das Histogramm p bekannt sind. Die Größe (N, N) des gesuchten Bary-Zentrums muss ebenfalls vorher festgelegt werden. Eine Erweiterung auf den Fall, dass auch p unbekannt sein sollte und damit als Optimierungsparameter aufgefasst wird, ist leicht möglich [PCS16].*

Wir können das Bary-Zentrum mithilfe von Kopplungen umformulieren als

$$\min_{C, (T_s)_s} \sum_s \lambda_s (\varepsilon_{C, C_s}(T_s) - \varepsilon H(T_s)), \quad (6.31)$$

unter den Nebenbedingungen: $\forall s : T_s \in \mathcal{C}_{p, p_s} \subset \mathbb{R}_+^{N \times N_s}$. Für den Fall, dass L bezüglich der ersten Variable konvex ist, ist dieses Problem konvex bezüglich C . Bezüglich $(T_s)_s$ ist diese Problem quadratisch aber nicht notwendigerweise positiv.

Das Problem Gleichung 6.31 kann durch eine Block-Koordinaten-Relaxierung gelöst werden. Dabei wird iterativ abwechselnd bezüglich den Kopplungen $(T_s)_s$ und der Metrik C minimiert.

Minimierung bezüglich $(T_s)_s$.

Minimierung bezüglich C . Sei (T_s) gegeben. Dann lautet, die Minimierung bezüglich C :

$$\min_C \sum_s \lambda_s \langle \mathcal{L}(C, C_s) \otimes T, T \rangle. \quad (6.32)$$

Mit der folgenden Proposition erhalten wir für eine Klasse von Verlustfunktionen eine Lösung in geschlossener Form.

Proposition 6.5.42. *Sei L eine Verlustfunktion, die die Bedingung Gleichung 6.25 erfüllt. Sei f'_1/h'_1 invertierbar.*

Dann lässt sich die Lösung zu Gleichung 6.32 schreiben als:

$$C = \left(\frac{f'_1}{h'_1} \right)^{-1} \left(\frac{\sum_s \lambda_s T_s^\top h_2(C_s) T_s}{pp^\top} \right), \quad (6.33)$$

wobei die Normalisierung $\lambda_s = 1$ gilt.

Beweis. Inhalt...

□

²⁹https://pythonot.github.io/auto_examples/gromov/plot_gromov_barycenter.html

Anhand von Gleichung 6.33 wird die folgende Interpretation deutlich/möglich:
Für jedes $s \in S$ ist $T_s^\top h_2(C_s) T_s$ eine wiedereingeordnete Matrix, wobei T_s als Optimal Transport-Kopplung von Zeilen und Spalten der Distanzmatrix C_s fungiert. Über diese Matrizen wird anschließend gemittelt, wobei die Art der Mittelung von der Verlustfunktion L abhängig ist.

Für den Fall $L = L_2$ wird aus dem Update Gleichung 6.33 die folgende Vorschrift:

$$C \leftarrow \frac{1}{pp^\top} \sum_s \lambda_s T_s^\top C_s T_s. \quad (6.34)$$

Proposition 6.5.43. *Sei $L = L_2$ und $(C_s)_s$ positiv semi-definit f.a. $s \in S$. Dann sind die Iterierten C ebenfalls positiv semi-definit.*

Für den Fall $L = KL$ ergibt sich die folgende Verfahrensvorschrift:

$$C \leftarrow \left(\frac{1}{pp^\top} \sum_s \lambda_s T_s^\top \log(C_s) T_s \right). \quad (6.35)$$

Algorithm 2 Berechnung der $GW_{-\varepsilon}$ Baryzentren

Input: $(C_s, p_s), p$. Initialisiere C .

Output: C .

if some condition is true **then**
 do some processing
else if some other condition is true **then**
 do some different processing
else
 do the default actions
end if

Pseudocode.

Häufige Verwendungen(s. Einführung hier ³⁰). zB. Shape interpolation

Wir haben in diesem Kapitel gesehen, wie mithilfe des Wasserstein-Baryzentrum eine Art Mittelwert für Wahrscheinlichkeitsverteilungen definiert werden kann.

6.5.15 Numerische Approximationen

3

7 (Numerische) Ergebnisse/Vergleich mit anderen Verfahren

7.0.1 Standard Poisoning-Angriffe

=> Did save Model - incv320epochsnormalized - at epoch: 19 => [19] loss: 0.125, accuracy: 96.514=> Accuracy of validation Dataset: 98.491

³⁰<https://arxiv.org/pdf/2102.01752.pdf>

Absolute values of first 30 eigenvalues of L_{sym} with 10-nearest neighbours

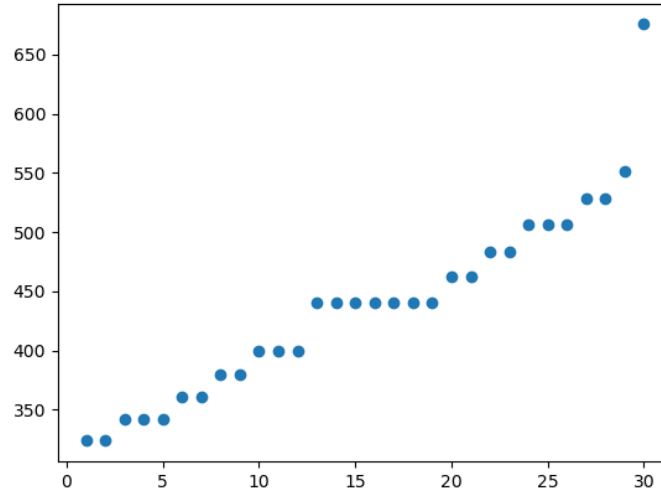


Abbildung 2: Ergebnis des spektralen Clusterings unter Verwendung der Euklidischen Distanz und $k=10$ Nachbarn

8 Weitere mögliche Schritte

- Untersuchung der Detektionsqualität in Abhängigkeit von ε
- Automatische Platzierung des Auslösers an fest gewählter Position auf dem Verkehrsschild anstatt zufälligem Platzieren in einem Fenster mit vorher festgelegter Größe. In [GDGG17] wird Faster-RCNN (F-RCNN) zur Klassifikation des LISA-Datensatzes³¹ benutzt. Es ist die Aufgabe, die Verkehrsschilder in die 3 Superklassen Stoppschild, Geschwindigkeitsbegrenzung und Warnschild einzuteilen. Der Datensatz enthält zudem die BoundingBoxen, sodass der Auslöser genauer angebracht werden kann.
- Verbesserte Version der Layer-wise Relevance Propagation
- Untersuchung anderer Verfahren, die die Interpretierbarkeit ermöglichen, beispielsweise: VisualBackProp: efficient visualization of CNNs³²
- Vergleich mit Cifar-10/Cifar-100 Datensatz³³³⁴

9 Zusammenfassung

Beobachtung: Je größer die Netzwerke sind, desto leichter lassen sich Poisoning-Angriffe realisieren.

³¹<http://cvrr.ucsd.edu/LISA/lisa-traffic-sign-dataset.html>

³²<https://arxiv.org/abs/1611.05418>

³³<https://www.cs.toronto.edu/~kriz/cifar.html>

³⁴<https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>

Heatmap und erzeugte Punktwolke für threshold = 0.99

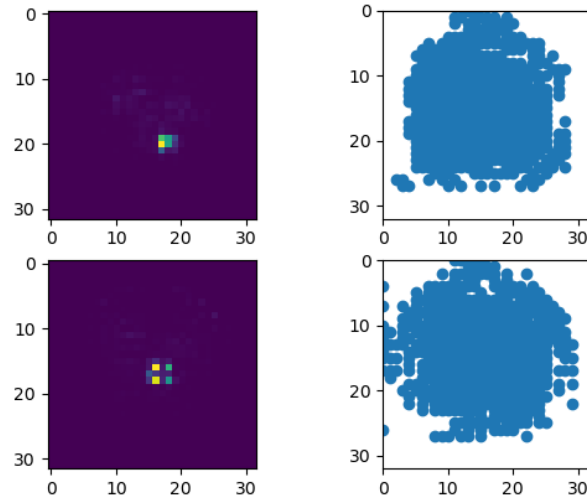


Abbildung 3: Auswahl der relevantesten Pixel (bis zu 99% der Gesamtmasse) zweier Heatmaps

Heatmap und erzeugte Punktwolke für threshold = 0.5

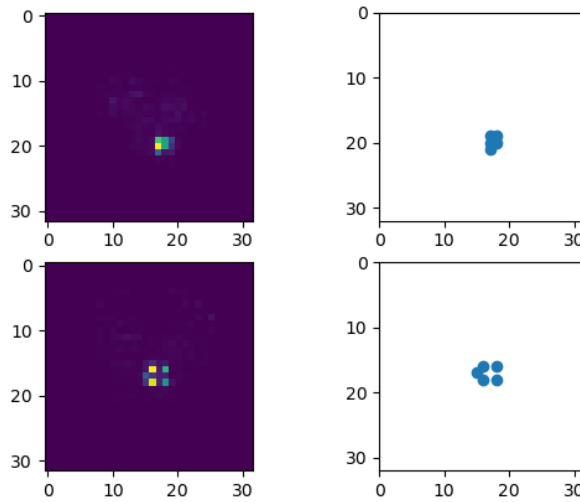


Abbildung 4: Auswahl der relevantesten Pixel (bis zu 50% der Gesamtmasse) zweier Heatmaps

A Verwendete Netzwerke

A.1 Net

```

1  class Net(nn.Module):
2
3
4  def __init__(self, ):
5      super(Net, self).__init__()
6      self.size = 64 * 4 * 4
7      self.conv1 = nn.Conv2d(in_channels=3, out_channels=12,
8                              kernel_size=5, padding=2)
9      self.pool = nn.MaxPool2d(kernel_size=2, stride=2)
10     self.conv1_in = nn.InstanceNorm2d(12)
11     self.conv2 = nn.Conv2d(in_channels=12, out_channels=32,
12                             kernel_size=5, padding=2)
13
14     self.conv2_bn = nn.BatchNorm2d(32)
15
16     self.conv3 = nn.Conv2d(in_channels=32, out_channels=64,
17                             kernel_size=5, padding=2)
18
19     self.fc1 = nn.Linear(self.size, 256)
20     self.fc1_bn = nn.BatchNorm1d(256)
21     self.fc2 = nn.Linear(256, 128)
22     self.fc3 = nn.Linear(128, 43)

```

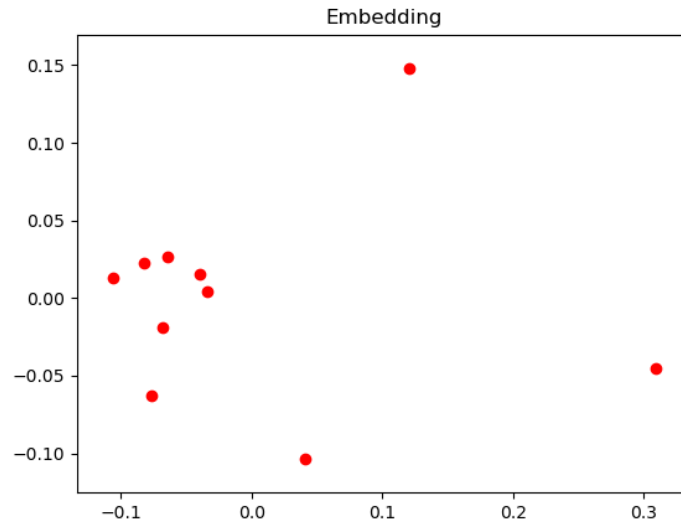


Abbildung 5: Einbettung des Baryzentrums mithilfe von Multidimensionaler Skalierung(MDS) bei der Wahl von 99% der Gesamtmasse

```

21
22 def forward(self, x):
23     x = self.pool(F.relu(self.conv1_in(self.conv1(x))))
24     x = self.pool(F.relu(self.conv2_bn(self.conv2(x))))
25     x = self.pool(F.relu(self.conv3(x)))
26     x = x.view(-1, self.size)
27     x = F.relu(self.fc1_bn(self.fc1(x)))
28     x = F.dropout(x)
29     xx = F.relu(self.fc2(x))
30     x = F.dropout(xx)
31     x = self.fc3(x)
32
33     return x, xx
34
35

```

Listing 4: Kleines Netzwerk

```

1 InceptionNet3(
2     (features): Sequential(
3         (0): InceptionA(
4             (parallel_dummyA): New_parallel_chain_dummy()
5             (conv1x1): BatchConv(
6                 (conv): Conv2d(3, 64, kernel_size=(1, 1), stride=(1, 1))
7                 (bn): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True,
8                     track_running_stats=True)
9                 (relu): ReLU()
10            )

```


Seitenlänge des Triggers	Prozentualer Anteil	AER	GUD	num. korruptierte Daten
s=2	0.000625	0.01333333	0.962	1
	0.00125	0.356	0.964	2
	0.0025	0.576	0.97	4
	0.005	0.99867	0.962	8
	0.01	0.92	0.962	17
	0.02	1.0	0.964	34
	0.10	1.0	0.968	291
	0.33	1.0	0.968	813
s=3	?	?	?	?
	0.15	1.0	0.961	
s=1	0.33	1.0	0.966	813

Tabelle 2: Qualität der Angriffe auf das Inception v3-Netz mit Stickern Seitenlänge 2 und 3 Pixel bei unterschiedlich großen Anteilen an korruptierten Daten

L2-Fehler				H1-Fehler		
	Gitter	P=1	P=2	Gitter	P=1	P=2
T=2.0	8×8	2.2e-15	1.6e-14	8×8	3.7e-14	1.7e-13
	16×16	1.4e-14	2.5e-13	16×16	1.0e-13	1.2e-12
	32×32	6.1e-15	1.1e-14	32×32	1.6e-14	6.0e-13

Tabelle 3: Fehler für Testproblem 1 zum Endzeitpunkt $T = 2.0$.

```

10 (parallel_dummyB): New_parallel_chain_dummy()
11 (conv5x5_1): BatchConv(
12 (conv): Conv2d(3, 48, kernel_size=(1, 1), stride=(1, 1))
13 (bn): BatchNorm2d(48, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True)
14 (relu): ReLU()
15 )
16 (conv5x5_2): BatchConv(
17 (conv): Conv2d(48, 64, kernel_size=(5, 5), stride=(1, 1),
    padding=(2, 2))
18 (bn): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True)
19 (relu): ReLU()
20 )
21 (parallel_dummyC): New_parallel_chain_dummy()
22 (conv3x3dbl_1): BatchConv(
23 (conv): Conv2d(3, 64, kernel_size=(1, 1), stride=(1, 1))
24 (bn): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True)
25 (relu): ReLU()

```

	L2-Norm			H1-Norm		
T = 2.0	Gitter	P=1	P=2	Gitter	P=1	P=2
$dt = \frac{T}{25}$	8×8	15.7645	16.2653	8×8	24.1041	24.2838
	16×16	16.1405	16.2668	16×16	24.5373	24.6831
	32×32	16.2353	16.2669	32×32	24.6467	24.6833
$dt = \frac{T}{50}$	8×8	15.8507	16.3562	8×8	24.2326	24.8138
	16×16	16.2303	16.3577	16×16	24.6699	24.8171
	32×32	16.3259	16.3578	32×32	24.7806	24.8173
$dt = \frac{T}{100}$	8×8	15.8939	16.4019	8×8	24.2972	24.8811
	16×16	16.2753	16.4035	16×16	24.7365	24.8843
	32×32	16.3714	16.4036	32×32	24.8475	24.8846
KO α (im Ort) ($dt = \frac{T}{100}$)		1.3787	2.7828		1.3760	2.5901

Tabelle 4: Konvergenz im Ort für das Testproblem 4, $T = 2.0$

```

26 )
27 (conv3x3dbl_2): BatchConv(
28 (conv): Conv2d(64, 96, kernel_size=(3, 3), stride=(1, 1),
   padding=(1, 1))
29 (bn): BatchNorm2d(96, eps=1e-05, momentum=0.1, affine=True,
   track_running_stats=True)
30 (relu): ReLU()
31 )
32 (conv3x3dbl_3): BatchConv(
33 (conv): Conv2d(96, 96, kernel_size=(3, 3), stride=(1, 1),
   padding=(1, 1))
34 (bn): BatchNorm2d(96, eps=1e-05, momentum=0.1, affine=True,
   track_running_stats=True)
35 (relu): ReLU()
36 )
37 (parallel_dummyD): New_parallel_chain_dummy()
38 (pool): MaxPool2d(kernel_size=3, stride=1, padding=1,
   dilation=1, ceil_mode=False)
39 (pool1x1): BatchConv(
40 (conv): Conv2d(3, 32, kernel_size=(1, 1), stride=(1, 1))
41 (bn): BatchNorm2d(32, eps=1e-05, momentum=0.1, affine=True,
   track_running_stats=True)
42 (relu): ReLU()
43 )
44 (parallel_dummyE): New_parallel_chain_dummy()
45 (cat): Cat()
46 )
47 (1): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation
   =1, ceil_mode=False)
48 (2): BatchConv(
49 (conv): Conv2d(256, 256, kernel_size=(2, 2), stride=(1, 1))
50 (bn): BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True,

```

```

    track_running_stats=True)
51 (relu): ReLU()
52 )
53 (3): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation
    =1, ceil_mode=False)
54 (4): BatchConv(
55 (conv): Conv2d(256, 256, kernel_size=(2, 2), stride=(1, 1))
56 (bn): BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True)
57 (relu): ReLU()
58 )
59 (5): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation
    =1, ceil_mode=False)
60 (6): BatchConv(
61 (conv): Conv2d(256, 256, kernel_size=(2, 2), stride=(1, 1))
62 (bn): BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True)
63 (relu): ReLU()
64 )
65 (7): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation
    =1, ceil_mode=False)
66 )
67 (classifiers): Sequential(
68 (0): Linear(in_features=256, out_features=256, bias=True)
69 (1): ReLU(inplace=True)
70 (2): Dropout(p=0.5, inplace=False)
71 (3): Linear(in_features=256, out_features=128, bias=True)
72 (4): ReLU(inplace=True)
73 (5): Dropout(p=0.5, inplace=False)
74 (6): Linear(in_features=128, out_features=43, bias=True)
75 )
76 )
77

```

Listing 5: Einfachere Version von Inception v3

Aufbau dieses Netzwerkes: 1. Inception-Modul 2. [pool1, batchConv1, pool2, batchConv2, pool3, batchConv3, pool4] 3. Drei Lineare Schichten mit ReLu und Dropout dazwischen

Im Unterschied zum offiziellen Inception Netz(v1v2v3) gibt es in dieser vereinfachten Versionn keinen stemäus convs, es geht direkt mit InceptionA los.

Wie ähnlich sind sich InceptionA(hier) und das offizielle InceptionA-Modul?

```

1
2 [Linear(in_features=128, out_features=43, bias=True),
3 Dropout(p=0.5, inplace=False),
4 ReLU(inplace=True),
5 Linear(in_features=256, out_features=128, bias=True),
6 Dropout(p=0.5, inplace=False),
7 ReLU(inplace=True),
8 Linear(in_features=256, out_features=256, bias=True),
9 MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1,
    ceil_mode=False),
10 ReLU(),

```

```
11 BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True), Conv2d(256, 256, kernel_size=(2,
    2), stride=(1, 1)),
12 MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1,
    ceil_mode=False),
13 ReLU(),
14 BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True), Conv2d(256, 256, kernel_size=(2,
    2), stride=(1, 1)),
15 MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1,
    ceil_mode=False),
16 ReLU(), BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True
    , track_running_stats=True),
17 Conv2d(256, 256, kernel_size=(2, 2), stride=(1, 1)),
18 MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1,
    ceil_mode=False),
19
20 [[Conv2d(3, 64, kernel_size=(1, 1), stride=(1, 1)),
21 BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True), ReLU()],
22
23 [Conv2d(3, 48, kernel_size=(1, 1), stride=(1, 1)),
24 BatchNorm2d(48, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True), ReLU(), Conv2d(48, 64,
    kernel_size=(5, 5), stride=(1, 1), padding=(2, 2)),
    BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True), ReLU()],
25
26 [Conv2d(3, 64, kernel_size=(1, 1), stride=(1, 1)),
27 BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True), ReLU(),
28 Conv2d(64, 96, kernel_size=(3, 3), stride=(1, 1), padding=(1,
    1)),
29 BatchNorm2d(96, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True), ReLU(),
30 Conv2d(96, 96, kernel_size=(3, 3), stride=(1, 1), padding=(1,
    1)),
31 BatchNorm2d(96, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True), ReLU()],
32
33 [MaxPool2d(kernel_size=3, stride=1, padding=1, dilation=1,
    ceil_mode=False), Conv2d(3, 32, kernel_size=(1, 1), stride
    =(1, 1)),
34 BatchNorm2d(32, eps=1e-05, momentum=0.1, affine=True,
    track_running_stats=True),
35 ReLU()],
36
37 [Cat()]]
38
```

Listing 6: Reversed Model incv3

B Parameter für Training und Einlesen der Daten

Die in [AWN⁺20] gewählten Parameter wären ein guter Ausgangspunkt. Wir trainieren die Netzwerke über maximal 100 Epochen und benutzen *early stopping* mit einer *patience* = 20. Die verwendete Implementierung ist eine modifizierte Version von Bjarte Mehus Sunde³⁵, die wiederum auf PyTorch Ignite³⁶ basiert.

C Datensätze

GTSRB³⁷ Datensatz Splitting (Train; Val, test)

D Programmcode

Der vollständige Programmcode ist verfügbar unter <https://github.com/lukasschulth/MA-Detection-of-Poisoning-Attacks>

E Notizen

registered spaces³⁸ Barycenters in the Wasserstein Space³⁹

³⁵<https://github.com/Bjarten/early-stopping-pytorch>

³⁶https://github.com/pytorch/ignite/blob/master/ignite/handlers/early_stopping.pyt

³⁷https://benchmark.ini.rub.de/gtsrb_dataset.html

³⁸<https://arxiv.org/pdf/1809.06422.pdf>

³⁹<https://arxiv.org/pdf/1809.06422.pdf>

Literatur

- [AC11] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
- [AMN⁺19] 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. *arXiv preprint arXiv:1912.11425*, 2019.
- [AWN⁺19] 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, 2019.
- [AWN⁺20] 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. 2020.
- [BBM⁺15] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.
- [BBM⁺16] Alexander Binder, Sebastian Bach, Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for deep neural network architectures. In Kuinam J. Kim and Nikolai Joukov, editors, *Information Science and Applications (ICISA) 2016*, pages 913–922, Singapore, 2016. Springer Singapore.
- [BEWR19] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer’s disease classification. *Frontiers in aging neuroscience*, 11:194, 2019.
- [CCB⁺18] 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. *arXiv preprint arXiv:1811.03728*, 2018.
- [com19] Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.
- [Cut13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *arXiv preprint arXiv:1306.0895*, 2013.
- [GDGG17] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.
- [IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

- [LWB⁺19] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.
- [MBL⁺19] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.
- [MLB⁺17a] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.
- [MLB⁺17b] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.
- [PCS16] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.
- [SLJ⁺15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [TLM18] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *arXiv preprint arXiv:1811.00636*, 2018.
- [TTM19] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks, 2019.
- [VL07] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.