Talk Structure "Risks of AI: Adversarial Examples"

von Pascal Müller

Ich habe das Konzept noch einmal überdacht und denke ich werde nicht so sehr in die Details gehen wie ich ursprünglich vorhatte (wir werden mehrere Vorträge an einem Tag hören (Zuhörer erschöpfen irgendwann) und ich habe nur 20 Minuten für den Vortrag). Stattdessen will ich mich darauf konzentrieren die Zuhörer abzuholen und ihnen nur einen Vorgeschmack auf das geben, was in der schriftlichen Ausarbeitung stehen wird. Insbesondere werde ich versuchen die Stolpersteine Backpropagation und "viel Code am Bildschirm" zu umschiffen, insofern wahrscheinlich auch keine Live Demo.

0. Einstieg

Zuerst einmal ist mir aufgefallen, dass es in keinem anderen Vortrag um Risiken von Machine Learning Modellen geht. Neben Adverserial Examples gibt es z.B. noch Data Poisoning (dem Modell misgelabelte Daten geben) und Data Extraction (Datenpunkte aus dem Originaldatenset aus dem Modell wiederherstellen). Ich möchte meinen Vortrag gerne damit beginnen eine Übersicht über diese Risiken zu geben um dem ganzen mehr Kontext zu geben.

Als nächstes werde ich auf den Begriff "Adverserial Examples" eingehen
Übersetzung: Adverserial Examples = "Widersprüchliches Beispiel"
Der Widerspruch ist: Mensch sagt es ist Klasse A, aber Netzwerk sagt es ist Klasse B
Der Begriff bezieht sich meist auf modifizierte echte Datenpunkte (klassisches Beispiel sind Bilder), ohne diese Modifikation gäbe es auch den Widerspruch nicht (Netzwerk würde auch Klasse A sagen)

1. Paper #1: Intriguing properties of neural networks https://arxiv.org/pdf/1312.6199.pdf

Das ist nach meinem Wissen (https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html) das erste Paper, das den Begriff "Adverserial Examples" benutzt.

In diesem Paper wurde gezeigt, dass man mit bestimmten (für den Menschen) unsichtbaren Veränderungen an einem Bild die vom Netzwerk vorhergesagte Klasse ändern kann. Das haben Sie mit Bildern aus dem MNIST-Datensatz (Handgeschriebene Zahlen - 10 Klassen) und dem Imagenet-Datensatz (Fotos - 1000 Klassen) getan.

In dem Vortrag möchte ich auf die Ergebnisse des Papers eingehen, insbesondere auch auf die "Cross model generalization" (Black-Box transferability) und "Cross training-set generalization". Diese beiden Eigenschaften zu überprüfen wäre eine Interessante Implementierungsaufgabe, aber ich habe das Gefühl, dass sie dem Vortrag nicht viel hinzufügen würde. Vielleicht finde ich einen Weg, in der schriftlichen Ausarbeitung etwas in der Art zu machen.

In der Conclusion nennen die Forscher eine offene Frage:
"Indeed, if the network can generalize well, how can it be confused by these adversarial negatives, which are indistinguishable from the regular examples? Possible explanation is that the set of adversarial negatives is of extremely low probability, and thus is never (or rarely) observed in the test set, yet it is dense (much like the rational numbers), and so it is found near every virtually every test case."

Ich finde diese Frage fasst gut die "Merkwürdigkeit" von Adverserial Examples zusammen. Insofern möchte ich gerne an dieser Stelle vom Vortrag den Zuhörern einen Zugang zu der Frage vermitteln. Soweit ich das im Moment weiß, gibt es keine zufriedenstellende Antwort auf diese Frage, aber wenn ich in meiner weiteren Arbeit an dem Vortrag auf eine solche Antwort (oder Ansätze) stoße, kommt diese hier rein.

2. Paper #2: Adversarial Examples: Attacks and Defenses for Deep Learning https://arxiv.org/pdf/1712.07107.pdf

Dieses Paper gibt einen guten Überblick über verschiedene Geschmacksrichtungen von Adverserial Examples und Countermeasures.

Ich möchte hier einen Überblick über die wichtigsten Attacks/Defenses geben (also im Prinzip eine Synopsis des Papers. Die Methoden (z.B. Fast Gradient Sign Method) werde ich mir für die schriftliche Ausarbeitung aufsparen.

Folgende Anwendungsfelder für Adverserial Exaples möchte ich ansprechen:
- Object Classifciation (war ja schon die ganze Zeit in 1. das Thema)
- Face Recognition (sehr relevant wegen Impersonation / Surveilance Evasion)
- NLP (nicht sicher wie gut sich der Text dann noch liest, fliegt vielleicht raus wenn ich mir die Quellen genau anschaue)
- Malware/Spam Detection

Außerdem Countermeasures:
- Adverserial (Re)training
- Input Reconstruction (Autoencoder der aus Adverserial Examples "normale" Bilder macht)
- Network Verification (Beweis, das in einem gewissen Beriech um ein Bild mit kleinen Veränderungen kein Adverserial Example zu finden ist. Das schaue ich mir noch mal genauer an, aber sieht nett aus, passt vor allem auch gut zum Logikteil unserer Veranstaltung)
- Ensembling Defenses (mehr/verschiedene Netzwerke)

In Conclusion sagen sie: "Almost all defenses are shown to be effective only for part of attacks."
Wie dieser Satz genau auf die oben genannten Techniken zutrifft, werde ich mir noch erarbeiten müssen

3. ?
Je nachdem wieviel Zeit bleibt würde ich auch gerne noch den Zusammenhang zu Attribution herstellen. Attribution soll die Bereiche des Inputs zeigen, wegen denen

das Netzwerk eine Entscheidung getroffen hat und Adverserial Examples setzen z.T. genau an diesen Bereichen im Input an um die Entscheidung des Netwerks möglichst stark zu beeinflussen. Das könnte vor allem Leuten, die schon einmal eine Machine-Learning-Veranstaltung besucht haben einen weiteren interessanten Blickwinkel auf Adverserial Examples geben.