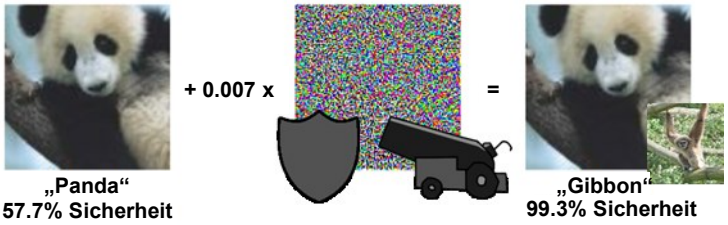
Hauptseminar

Thema: Adversarial Attacken auf Neuronale Netze und Verteidigungen

In [1] wurde gezeigt, dass man Neuronale Netze leicht täuschen kann, sodass sie das falsche Ergebnis liefern und eine vermeintlich hohe Sicherheit angeben. In [2] wurde vorgestellt, wie man Neuronale Netze mittels Adversarial Training zu einem gewissen Maße robust gegenüber diesen Angriffen machen kann. Seitdem ist ein Wettrüsten zwischen immer neuen Verteidigungen und Angriffen zu beobachten. Beispielsweise wurde auf der ICML 2018 ein Angriff [4] vorgestellt, der 7 Verteidigungen robust umgehen kann, die nur drei Monate zuvor auf der ICLR vorgestellt wurden. Nur die in [5] vorgestellte Verteidigung konnte zu einem gewissen Grad standhalten.

Aufgabenstellung:

- Kurze Zusammenfassung der Grundlagen zum Thema aus [1, 2]
- Aufarbeiten der in [3] und [4] beschriebenen Attacken
- Aufarbeiten der in [5] und [6] vorgestellten Verteidigungen
- Systematisierung der State-of-the-Art-Ansätze für Attacken und Verteidigungen
- Vorstellung des Themas im Rahmen einer Abschlusspräsentation



Beispiel für einfache Täuschung. Bildquelle: [2]

Geeignet für:

Bachelor- / Masterstudiengänge

Themengebiet / Schwerpunkte:

Machine Learning, Neuronale Netze

Erforderliche Vorkenntnisse:

Guter Abschluss der Vorlesung „Neuroinformatik“

Zu verwendende Literatur:

[1] Szegedy et al.: **Intriguing properties of neural networks**. ICLR, 2014 (Link)
[2] Goodfellow et al.: **Explaining and Harnessing Adversarial Examples**. ICLR, 2015 (Link)
[3] Carlini et al.: **Towards evaluating the robustness of neural networks**. SP, 2017 (Link)
[4] Athalye et al.: **Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples**. ICML, 2018 (Link)
[5] Madry et al.: **Towards deep learning models resistant to adversarial attacks**. ICLR, 2018 (Link)
[6] Tramer et al.: **Ensemble adversarial training: Attacks and defenses**. ICLR, 2018 (Link)

- Elektronische Literaturdatenbank des FG NI&KR mit Recherchemöglichkeiten
- Elektronische Konferenzproceedings-Datenbank des FG NI&KR
- IEEE Recherchesystem www.ieeexplore.ieee.org (nur aus dem Uni-Netz bzw. via VPN)
- Google Scholar scholar.google.com
- Microsoft Academic Search academic.research.microsoft.com
- Proceedings der relevanten Konferenzen (NIPS, ICML, ICLR, IJCNN, WCCI, ICANN, CVPR, ICCV, ECCV, BMVC, AVSS, ICPR, ICIP, …)

Betreuer: Dipl.-Inf. Markus Eisenbach (Markus.Eisenbach@tu-ilmenau.de)
Betr. Hochschullehrer: Prof. Dr. H.M. Groß
Bearbeiter: **Johannes Strauch**