Master Thesis

Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

A Firewall for Neural Networks: Detecting Adversarial Inputs

Neural networks are a double-edged sword. On the one hand, they are unmatched in prediction accuracy for some tasks, such as image classification. On the other hand, an adversary may manipulate predictions of neural networks by forging falsified inputs, e.g., via slight modifications of an input image [1]. Neural networks are not robust against such *adversarial inputs* and thus unsafe to deploy in some settings.

There are two ways of addressing this robustness issue. The first way is to improve the accuracy of neural networks on adversarial inputs, by designing new network architectures. So far, architectures proposed to this end achieve higher accuracy for specific types of adversarial inputs, but not in general. This motivates the second way: designing methods to *detect* adversarial inputs [2]. Figuratively speaking, such methods act as a *firewall* for neural networks. They enable machine learning applications to deal with adversarial inputs in a distinct way, e.g., by discarding them. However, most existing detection methods only work for *zero-knowledge* adversaries. Adversaries who know the type and parameters of a detector can often forge malicious inputs that go undetected [3].

The goal of this thesis is to develop a new method to detect adversarial inputs of neural networks. We hypothesize that adversarial inputs tend to activate other combinations of neurons than normal training inputs. In other words, neuron activations from adversarial inputs are *outlying* w.r.t. the activations observed during the training phase. In consequence, one should treat the detection of adversarial inputs as an outlier detection task. This leads to the following questions:

- Which definition of “outlierness” could work to detect adversarial examples? Options include probability-, clustering-, or density-based outliers, see [4].
- There can be thousands or millions of neurons in a network. The activations of these neurons form a space of much higher dimensionality than what has been studied in the outlier detection literature (e.g. [5, 6]). Is there a method that can deal with such high dimensionality?
- Is our “firewall” robust against specific types of adversarial inputs? How far does the robustness depend on the knowledge of the adversary?

This results in the following tasks:

- Review of existing approaches for detecting adversarial inputs in neural networks, in particular approaches that use neuron activations, e.g., [7].
- Implementation of a method for adversarial input detection. This method should build on neuron activations and address the challenges linked to very high dimensionality.
- Investigation of the differences between “normal” neuron activations, and activations for different types of adversarial inputs.
- Evaluation of the robustness of the new method.

To help you with this work, we offer mentoring via meetings with your advisor and access to our computing infrastructure. We expect basic knowledge in Python programming and machine learning.

[1] K. Eykholt et al. *Robust Physical-World Attacks on Deep Learning Models*. Apr. 10, 2018. arXiv: 1707.08945.

[2] S. Bulusu et al. *Anomalous Example Detection in Deep Learning: A Survey*. Feb. 19, 2021. arXiv: 2003.06979.

[3] N. Carlini and D. Wagner. “Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods”. In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. AISec ’17. New York, NY, USA, Nov. 3, 2017, pp. 3–14.

[4] C. C. Aggarwal. “Outlier Analysis”. In: *Data Mining: The Textbook*. Ed. by C. C. Aggarwal. Cham, 2015, pp. 237–263.

[5] S. Sathe and C. C. Aggarwal. “Subspace Outlier Detection in Linear Time with Randomized Hashing”. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 2016 IEEE 16th International Conference on Data Mining (ICDM). Dec. 2016, pp. 459–468.

[6] H. Trittenbach and K. Böhm. “Dimension-Based Subspace Search for Outlier Detection”. In: *International Journal of Data Science and Analytics* 7.2 (Mar. 1, 2019), pp. 87–101.

[7] F. Carrara et al. “Adversarial Examples Detection in Features Distance Spaces”. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018, pp. 0–0.

Contact: Moritz Renftle, M.Sc. Raum 338, Gebäude 50.34
moritz.renftle@kit.edu Am Fasanengarten 5
+49 721 608-44066 76131 Karlsruhe