Incorporating data pruning into robust learning with large-scale datasets

With the ongoing advancements in AI and machine learning, an increasing number of these techniques are deployed in real-world applications. Consequently, the demand for robust and secure models is on the rise. One significant vulnerability of machine learning models is their susceptibility to adversarial attacks [1]. An adversarial example refers to an input which through the addition of a slight perturbation leads to a misclassification of the model, while a human in general can't differentiate between the adversarial input and the original input. Even modern, complex models like neural networks, are highly vulnerable to adversarial examples [3]. Adversarial robustness, in this context, represents a model's capacity to withstand such adversarial attacks. Various techniques have been proposed to improve adversarial robustness, including directly integrating adversarial examples into the model's training process. Additionally, increasing model capacity [6] and using more training data contributes to enhanced adversarial robustness [7]. However, even with large models and a considerable amount of data, the performance on the classification task for robust models still lacks behind the performance of standard models [7]. Thus, efficient training algorithms are needed to further improve adversarial robustness. In the realm of dataset pruning research, a subset of the dataset is selected for training by identifying samples that have a substantial impact on the model's performance. By deploying data pruning techniques, the total amount of training data can be lowered significantly while maintaining a comparable performance [4].

This thesis seeks to close the gap between the performance of robust networks and standard networks by applying a classical data pruning approach to large-scale synthetic datasets. For this work, we plan to employ the CIFAR10 [5] dataset, both in its original form (50,000 samples) and using a substantial amount of additional synthetic data (up to 5 million) [7]. We will train neural networks both for normal and for robust classification. For the pruning, we plan to implement the dynamic uncertainty metric as the primary pruning algorithm, which is a very recent cross-architecture approach, yielding up to 75% reduction of dataset size [4] on other large scale datasets [2]. During the literature review, another pruning algorithm might also be considered. As the calculation of pruning scores for each data sample is often expensive we further aim to develop a scaleable pruning approach. Specifically, we plan to use the scores produced by the pruning algorithm(s) to train a pruning network predicting a pruning score for each sample. Both pruning approaches will be tested against the classifier subsampling approach used in [7] and against a naive random subsampling approach as baselines.

1. Naveed Akhtar und Ajmal Mian. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. Feb. 2018. arXiv: 1801 . 00553 [cs]. (Besucht am 07. 09. 2023).
2. Jia Deng u. a. "ImageNet: A Large-Scale Hierarchical Image Database". In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Juni 2009, S. 248–255. doi: 10.1109/CVPR.2009.5206848.
3. Ian J. Goodfellow, Jonathon Shlens und Christian Szegedy. Explaining and Harnes[1]sing Adversarial Examples. März 2015. arXiv: 1412.6572 [cs, stat]. (Besucht am 07. 09. 2023).
4. Muyang He u. a. Large-Scale Dataset Pruning with Dynamic Uncertainty. Juni 2023. arXiv: 2306.05175 [cs]. (Besucht am 05. 09. 2023).
5. Alex Krizhevsky. "Learning Multiple Layers of Features from Tiny Images". In: ().
6. Aleksander Madry u. a. Towards Deep Learning Models Resistant to Adversarial At[1]tacks. Sep. 2019. arXiv: 1706.06083 [cs, stat]. (Besucht am 05. 09. 2023).
7. Zekai Wang u. a. Better Diffusion Models Further Improve Adversarial Training. Juni 2023. arXiv: 2302.04638 [cs]. (Besucht am 05. 09. 2023).