Hessisches Kompetenzzentrum
für Hochleistungsrechnen

Universal Adversarial Perturbations: A Review

Project Manager
Senih Avdovic

Principal Investigator
Prof. Dr. Stefan Ulbrich

Project Term
2025 - 2025

Clusters
Lichtenberg II Cluster Darmstadt

Software
PyTorch

Institute
Optimization

University
Technische Universität Darmstadt

Introduction

Machine learning has become an integral part of our society. In particular, the advancements driven by neural networks (Deep Learning) have led to major breakthroughs. However, despite these remarkable developments, neural networks still suffer from a fundamental issue. Let us consider the task of image classification, such as distinguishing between images of dogs and cats. From a well-performing model, we would expect that images which are clearly identifiable to the human eye are also correctly classified by the model. Misclassifications should ideally only occur when even a human would struggle to categorize the image correctly. One could go a step further and demand that if two images appear identical to a human observer, then the model should also assign them to the same class. And this is precisely where the problem lies. At present, it is not possible to build a neural network that satisfies this condition reliably. But it gets even more concerning: this weakness can be exploited. It is possible, using specific algorithms, to transform an image in such a way that the model's classification changes, while the image appears completely unchanged to a human viewer. This presents a serious threat, particularly in safety-critical domains. In our work, we investigated and compared two algorithms, see [1, 2] (we call them algorithm 1 and 2 from here), which compute one specific type of such transformations, also called Universal Adversarial Perturbations [1]. Running these algorithms requires significantly more computational power than a typical home PC can provide, which was the primary reason for utilizing the Lichtenberg Cluster.

Methods

We evaluated both algorithms in the context of image processing using images from the ImageNet dataset. Transformations were computed for four relatively recent and compact neural networks not used in [1, 2]. The comparison focused on two key metrics: the fooling rate of the computed transformations and the efficiency of each algorithm. Each transformation is generated using one dataset (also called the training set) and then validated on a separate dataset. A high fooling rate indicates that after applying the transformation to the validation images, a large proportion of them is misclassified - i.e., their predicted category changes compared to the original image. The efficiency of an algorithm reflects how many images are needed to compute a transformation that achieves a certain fooling rate. An algorithm with high efficiency requires fewer training images to produce an effective transformation than one with lower efficiency. Based on these two criteria, fooling rate and efficiency, we compared the performance of the two algorithms

Results

Based on our findings, we concluded that algorithm 2 is more effective when the training set is very small. However, if a larger training set is available and a high fooling rate (above 80%) is required, algorithm 1 proves to be more reliable and user-friendly. This is because algorithm 2 often requires extensive parameter tuning, which can be time-consuming. Algorithm 1 consistently produced effective transformations across all tested neural networks, provided that sufficient training data was available. In contrast, the effectiveness of transformations generated by algorithm 2 varied significantly depending on the specific neural network used, its achievable fooling rate was highly network-dependent.

Discussion

We were able to confirm the results from [1, 2] for more recent neural networks. The issue discussed in the introduction does not appear to be specific to certain architectures, but rather seems to be a fundamental problem of neural networks in general, regardless of their complexity. Ultimately, our results highlight the need for further research in this area to better understand the origin of this phenomenon.

Reference

[1] Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; Frossard, P.: “Universal adversarial perturbations,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE CVPR, Honolulu, HI: IEEE, 2017, pp. 86–94, isbn: 978-1-5386-0457-1 https://doi.org/10.1109/CVPR.2017.17

[2] Oseledets, I.; Khrulkov, V.: “Art of singular vectors and universal adver- sarial perturbations,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE CVPR, Salt Lake City, UT: IEEE, 2018, pp. 8562–8570, isbn: 978-1-5386-6420- https://doi.org/10.1109/CVPR.2018.00893

Last Update: 2026-03-31 10:54