Thomás Rodrigues Crespo

Title: Trustworthy machine learning: bridging adversarial and noise robustness of classifiers via smoothed analysis

Abstract:

Despite its ever-growing use, the sensitivity to adversarial attacks and random noise is a significant drawback of neural networks, posing a challenge to its deployment in safety-critical areas, such as medical diagnosis, and low-cost devices. Consequently, understanding and quantifying their robustness has attracted much attention. Concerning quantification, in one extreme, the worst-case approach gives a region in the input space that is safe against any adversarial perturbation, that is, a worst-case region. On the other extreme, the average-case approach describes robustness against random perturbations. While the former can yield too pessimistic certifications, the latter often fails to give a tight guarantee of robustness. Studies have attempted to bridge these two extremes, among them, Randomized Smoothing became prominent by certifying a worst-case region of a classifier subjected to input noise. In its original form, used in quantification of image classification robustness, the radius of the region certified by Randomized Smoothing scales with the input noise standard deviation, and requires an estimate of the correct classification probability with confidence intervals. This quantification suffers from a trade-off: for small variance one needs an exponentially larger sample size to obtain valid certifications, which is impractical, while for large variance there is a drop in the classifier's generalization.

In Chapter 2, inspired by the smoothed analysis of algorithmic complexity, which bridges the worst-case and average-case analyses of algorithms, we provide a novel theoretical framework for robustness analysis of classifiers, which we name Smoothed Robustness Analysis. We first present it in its general form, then demonstrate how to use it to obtain the worst-case, average-case, and Randomized Smoothing analyses as special cases.

In Chapter 3, we use the framework to propose a novel robustness analysis based on the classification margin, i.e., the difference between the largest incorrect and correct outputs. This approach works even in the small noise regime and thus provides a more confident robustness certification than Randomized Smoothing. To validate the approach, we evaluate the robustness of Lipschitz constrained fully connected and convolutional neural networks on the MNIST and CIFAR-10 datasets, respectively, by maximizing the deterministic margin, the expected margin, and the Randomized Smoothing certified radii, and we find that it indeed improves both adversarial and noise robustness.

In the experiments from Chapter 3 we used closed form estimates of the expected margin and Randomized Smoothing certified radii as losses, which are valid when using orthogonal

layers. However, the computational overhead of orthogonal layers hinders them impractical to larger neural networks architecture, required for more complex datasets. In Chapter 4, we discuss how to overcome these limitations and propose a practical approach for expected margin maximization.

Among safety-critical applications that benefit from robust classifiers are healthcare related ones, in which the wellbeing of individuals depends on the classifier's performance. In Chapter 5, we tackled the problem of detecting pain in Japanese macaques (Macaca fuscata) via single frame facial features. Due to their competitive behavior, macaques often hide any signs of weakness, making it difficult for veterinarians to know when medical intervention is needed. Despite the small sample size, with only 21 individuals, we found the finetuned ResNet50 able to generalize relatively well to individuals not presented during training, with best accuracy of 64%, and best precision and recall of pain classification, the most safety-critical class, of 61% and 69%, respectively.