Avoiding Shortcut Learning in Deep Models for Enhanced Defense Against Adversarial Attacks

Contact: Son VU son.vu@ensea.fr
(Office: D392-ENSEA)

Keywords: deep neural networks, adversarial learning, spurious features, shortcut learning, backdoor poisoning attacks, deepfake, optimization-based

Required courses: Machine learning, deep learning.
Required skill: good programming skill in Python (with pytorch or tensorflow)

Shortcut learning/spurious correlations. Despite the remarkable performance of deep learning models, recent studies [1-4] have identified their vulnerability to shortcut learning, also called spurious correlations, which are the dependencies between observed features and class labels that only hold for certain groups of training data. For example, classifying cows and camels in natural images, where 90% of cows are on grass and 90% of camels on sand, can lead a model trained with standard ERM (empirical risk minimization) to rely on the background rather than the animals. This may result in high worst-group test error, such as misclassifying cows on sand as camels. Similarly, in natural language processing, models often rely on specific words and syntactic heuristics when predicting the sentiment of a sentence or the relationship between a pair of sentences. In an especially alarming example, CNNs trained to recognize pneumonia were shown to rely on hospital-specific metal tokens in the chest X-ray scans, instead of features relevant to pneumonia [2].

Spurious correlations—often described as "correlations that do not imply causation"—have been extensively studied in recent years under various terms, including shortcut learning, group robustness, simplicity bias, and so on [4]. Understanding these correlations is crucial for enhancing AI systems to be more efficient (achieving high performance with smaller, less data-hungry models) and safer (resilient to adversarial samples, backdoor poisoning attacks, and noisy samples, for example).

Due to its importance, numerous techniques have been proposed at various stages in the pipeline of ML/DL model training to avoid spurious correlations. In this project, we are interested in learning techniques, in particular optimization-based methods.

Adversarial learning. We are also exploring adversarial learning, a critical area in machine learning that addresses the challenges posed by attacks on models. Adversarial learning involves understanding and defending against techniques that exploit vulnerabilities in machine learning systems. These attacks can subtly alter inputs to produce incorrect predictions and include adversarial examples, which directly manipulate data, backdoor attacks with hidden triggers, or even deepfakes that generate deceptive content. Effectively countering these threats is essential for improving the robustness and security of AI systems across various applications.

Shortcut learning vs. adversarial learning. Shortcut learning significantly impacts adversarial learning by affecting model vulnerability to attacks and the effectiveness of defenses [5]. This issue arises when models rely on superficial patterns or spurious correlations that attacks can exploit to manipulate predictions. Understanding this dynamic is crucial for identifying vulnerabilities and developing robust defenses. In this project, we aim to investigate how learning techniques can be optimized to counteract shortcut learning, and ***student focuses on one specific type of attack - such as adversarial samples, backdoor triggers, out-of-distribution samples or deepfakes*** [6, 7]. This allows for a diverse exploration of how different attacks can be mitigated, ultimately enhancing overall model security and robustness.

References

[1] Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. On feature learning in the presence of spurious correlations. In NeuRIPS 2022.
[2] Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In ICLR 2023.
[3] Yang, Y., Gan, E., Dziugaite, G. K., and Mirzasoleiman, B. Identifying spurious biases early in training through the lens of simplicity bias. In AISTATS 2024.
[4] Ye, W. et al. Spurious Correlations in Machine Learning: A Survey. Arxiv2024.
[5] Zhang, Y. et al. Causaladv: adversarial robustness through the lens of causality. In ICLR, 2022.

[6] L. Jezequel, N-S. Vu et al. Efficient anomaly detection with adversarial learnable tasks. preprint 2023.
[7] N. Larue, N-S. Vu, et al. SeeABLE: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In ICCV 2023