Research Plan Summary: Contrastive Adversarial Learning

Ziqi Oliver Zhang, ziqizh@umich.edu

Summary of the Proposal

Deep neural networks, despite their success in various computer vision tasks, are susceptible to adversarial examples. Recent work suggests that the brittleness of the models may come from the use of non-robust features, which can be easily leveraged in adversarial examples. Therefore, to improve the robustness of a model, we expect to train the model that learns representations from only robust features. In this work, we are going to first design a supervised representation learning framework trained on adversarial examples, and use the novel representation training method to build a robust classifier.

Background

Recent findings show that learning good representations may benefit the robustness of models in two ways. First, good representations should be mainly learned from robust features, and thus adversarial examples should cause a limited change in the representation space. Second, the corresponding classifier should produce a large margin, so that the model is not sensitive to the perturbations within a specific boundary ϵ.

Adversarial training can be seen as a practical approach to mitigate the influence of non-robust features during training. However, even the adversarially-trained models can only achieve 54.33% adversarial accuracy on CIFAR10 dataset under PGD-20 attack. One reason behind the brittleness of the deep neural networks may be the widely used cross-entropy loss. Some works have explored shortcomings with this loss, including the possibility of poor margins that lead to a lack of robustness and reduced generalization power. The contrastive learning framework is one way to mitigate this issue. The contrastive learning method imposes that normalized embeddings from the same class are closer together than embeddings from different classes, potentially improving robustness.

Supervised contrastive training includes two parts. The first part is to train a representation network that separates different classes in the representation space, and then apply transfer learning to this network to train a classifier. This method achieves a new state of the art accuracy for natural training and improves robustness compared to other cross-entropy based natural training methods.

However, how will the contrastive learning framework benefit the robustness is still an open question. The first challenge to design a adversarial training framework that is compatible to the contrastive learning framework. We can't merely apply existing adversarial training methodology, because there is no label to attack in the representation training stage. The second challenge is to increase the margin of the classifier. The previous findings discuss the margin mostly empirically, but we need a deterministic way to expand the margin so that it can be used as a means of defense.

In this work, we assume that 1) large margin and robust features make a good representation, and 2) a good representation can improve the robustness of a model. Thus, with these two assumptions, I will design an adversarial training methodology to increase the network robustness that 1) uses adversarial examples to reduce the effect of non-robust features, and 2) leverages the contrastive learning framework to maximize the margin of the classifier.

Timeline

- Start Project (Aug. - Nov.)
- CVPR Submission (Nov. 16)
- Presentation (Before Dec. 10)