Adversarially Enhanced Supervised Contrastive Learning: Balancing Robustness and Performance

Abinav Chari *(Project Lead)* Josh Landsman *(Project Member)* Kevin Zhao *(Project Member)*

Abstract

We propose a modification to supervised contrastive learning by incorporating adversarial examples as positive pairs. This study will focus on exploring the impact of this approach on adversarial robustness and model performance.

1 Motivation

Adversarial attacks pose a challenge to the reliability and robustness of machine learning models, especially in critical applications such as healthcare, autonomous vehicles, and cybersecurity. This problem is both theoretical, as it explores the limits of model robustness, and practical, as adversarial robustness is essential for deploying machine learning systems in real-world scenarios to deter threat actors. By modifying supervised contrastive learning to include adversarial examples, we aim to understand and potentially mitigate these vulnerabilities, contributing to more robust and secure AI systems.

2 State of the Art

Several works have explored contrastive learning and adversarial robustness, which form the foundation of our study:

Comprehensive Survey on Contrastive Learning. The survey by Haigen Hu et al. (2024) outlines the principles and advancements of contrastive learning, emphasizing its role in clustering intra-class features and separating inter-class features in the latent space. Challenges such as the selection of positive and negative pairs and effective data augmentation are discussed. The survey highlights supervised contrastive learning (SCL) as a promising extension that improves robustness and performance in various applications [1].

Enhancing Adversarial Robustness Through Supervised Contrastive Learning. Wang et al. (2024) present a framework combining supervised contrastive loss with margin-based contrastive loss to improve adversarial robustness. The method aligns intra-class embeddings while separating inter-class embeddings, resulting in structured feature spaces resistant to adversarial perturbations. Experiments on CIFAR-100 demonstrate significant improvements in robustness under adversarial attacks [2].

3 Problem Statement

We aim to explore two related problems: first, we would like to explore different methods of generating adversarial examples during training (i.e., without access to a fully-trained model). One preliminary idea to find this adversarial example is to find a sample within an ε-ball of our original sample which maximizes unsupervised contrastive loss. This option allows us to avoid using a model to generate examples, which is a desirable property. Our second problem is to test the performance and adversarial robustness of training a deep neural network by contrastive loss with positive pairs consisting of adversarial examples. We would like to compare this model to the SCL model described above as well as a baseline deep neural network which does not use contrastive loss, such as ResNet-18.

4 Timeline and Plan

To begin, we plan on performing more extensive research into the area, exploring pre-existing papers, techniques, and results. Afterward, we can then solidify our knowledge, starting point, and exploratory goals. Following this, we will implement and experiment with generating adversarial examples as outlined in the problem statement, testing its results along the way. Then, we will implement contrastive learning with adversarial examples, building on the pipeline from the previous phase to incorporate the generated adversarial examples as positive pairs in contrastive learning. Finally, we will evaluate the results of the training in comparison to baseline models and investigate their adversarial robustness and defense against

attacks.

Following our experiments, we will spend the last few weeks writing our work into a paper and creating the final project presentation.

References

[1] HU, H., WANG, X., ZHANG, Y., CHEN, Q., AND GUAN, Q. A comprehensive survey on contrastive learning. *Neurocomputing 610* (2024), 128645.

[2] WANG, L., NAYYEM, N., AND RAKIN, A. Enhancing adversarial robustness of deep neural networks through supervised contrastive learning, 2024.