Adversarially Enhanced Supervised Contrastive Learning: Balancing Robustness and Performance

Josh Landsman Kevin Zhao Abinav Chari

Abstract

We propose a modification to supervised contrastive learning aimed at improving adversarial robustness by incorporating adversarial examples as additional positive pairs during training. By explicitly encouraging the model to align clean and adversarial representations within the same class, this approach promotes the learning of more stable and attack-resistant features. Our study investigates the impact of this strategy on both adversarial robustness and clean data performance. Through a series of controlled experiments, we analyze the trade-offs introduced by this modification and demonstrate its potential for enhancing model resilience against adversarial perturbations.

1 Introduction

Deep learning models, particularly convolutional neural networks (CNNs), have demonstrated success across numerous applications. However, they remain highly susceptible to adversarial attacks - imperceptible input perturbations that can lead to misclassifications. This vulnerability significantly undermines their deployment in safety-critical areas such as autonomous driving, healthcare, and cybersecurity.

Recent research highlights that this vulnerability often stems from weak or unstable feature representations learned by standard training objectives [7]. As a response, supervised contrastive learning (SCL) [4] has emerged as a promising model for enforcing robust and semantically aligned feature spaces. By explicitly pulling together representations of samples from the same class while pushing apart those from different classes, SCL enhances intra-class compactness and inter-class separability, producing more reliable decision boundaries under perturbations.

Furthermore, a comprehensive survey of contrastive learning techniques reveals the field's rapid advancement, with growing attention to architectural innovations, augmentation strategies, and loss functions designed to improve generalization and robustness [3]. While most contrastive learning methods have focused on unsupervised settings, recent work has demonstrated that supervised variants, especially when integrated with additional constraints such as margin-based losses, can significantly improve adversarial robustness [7].

Previous works have focused on directly attacking a model through gradient-based adversarial examples, such as Projected Gradient Descent (PGD) [5] and Fast Gradient Sign Method (FGSM) [2]. However, training directly with adversarial examples can negatively impact the utility of the target models, trading off accuracy for robustness.

In this work, we build upon this line of research by exploring a novel extension to supervised contrastive learning. Specifically, we incorporate adversarial examples as additional positive pairs during contrastive training. Our hypothesis is that directly exposing the model to adversarial variants of inputs in a structured contrastive setting encourages the formation of more resilient representations without sacrificing performance on clean data. We compare our methods with traditional Projected Gradient Descent [5] and Cross Entropy loss.

Through this adversarially augmented contrastive learning framework, we investigate the balance between robustness and accuracy, and assess its effectiveness across a range of threat models. Our contributions aim to shed light on how adversarial contrastive training can be strategically used to support the defense capabilities of deep neural networks.

2 Problem and Setup

Machine learning models are highly susceptible to adversarial examples — inputs that have been subtly perturbed to induce incorrect predictions while remaining visually or semantically unchanged to humans. This vulnerability undermines the reliability of models in high-stakes environments and highlights the need for improved training strategies that encourage both generalization and robustness.

This project investigates the problem of enhancing adversarial robustness through the lens of supervised contrastive learning. In the standard formulation, contrastive learning optimizes representations by pulling together samples from the same class (positive pairs) and pushing apart samples from different classes (negative pairs). Our modification introduces adversarially generated versions of each sample as additional positive pairs, with the hypothesis that this will improve the stability and robustness of the learned representations under adversarial conditions.

We adopt the following setup for our experiments:

- **Dataset:** Standard image classification benchmarks are used for evaluation, including Imagenet-1k
- **Model Architecture:** We utilize a ResNet-50 model pre-trained on ImageNet as a target model. For supervised contrastive loss, we use DINOv2-small for image embeddings [6].
- **Adversarial Attack:** We apply Projected Gradient Descent (PGD) to craft adversarial examples for both training and evaluation, controlling the perturbation size via an l_∞-norm constraint.
- **Training Strategy:** We attempted two approaches. (1) During training, adversarial examples are generated on-the-fly for each batch and are treated as additional positive samples within the contrastive loss function. (2) We pre-generate N adversarial examples. We then use these adversarial examples by themselves, matching positive and negative pairs based on their label, or we mix in clean training data and perform the same matching.

This setup allows us to evaluate whether integrating adversarial examples into the contrastive learning process can improve model robustness without significantly degrading clean accuracy.

3 Approach

Adversarially Augmented Contrastive Learning. Our approach is motivated by the idea that adversarial robustness can be enhanced by incorporating adversarial examples directly into the contrastive learning process. Rather than treating adversarial samples solely as adversaries during evaluation or as replacements during adversarial training, we introduce them as additional positive pairs within the supervised contrastive framework. This design encourages the model to learn class-consistent representations that are invariant to adversarial perturbations.

In traditional supervised contrastive learning, a model learns to pull together representations of samples belonging to the same class and push apart those from different classes. We extend this by including, for each clean sample, its corresponding adversarially perturbed version as an additional positive pair. This forces the model to map both clean and adversarial samples of the same class into a shared region of the embedding space, strengthening its resistance to adversarial attacks at the feature level.

Methodology. We implemented this strategy using a ResNet-50 architecture. We conducted experiments using three different training configurations:

1. **Cross-Entropy (CE) Loss:** The model was trained using the standard cross-entropy loss for classification.
2. **Supervised Contrastive (SupCon) Loss [1]:** The model was trained using the SupCon loss, which operates on feature embeddings. For our pre-trained ResNet-50 model, we apply the loss on the outputs of the layer before the final classification layer, treating it as an embedding layer.
3. **Combined Loss:** The model was first trained using the SupCon loss with adversarial examples as positive pairs. Then, the entire ResNet-50 model (both the backbone and the classification head) was fully fine-tuned using only the cross-entropy loss.

To generate SupCon adversarial examples, we implement the same attack strategy as Projected Gradient Descent [5]. However, instead of a standard cross-entropy loss between the target model's outputs and the target labels, we use a proxy embedding model, in this case DINOv2 [6], to generate embedding vectors for each input image in the clean batch. We then take the mean squared distance between the adversarial example and the original image's embedding vectors as the loss, and maximize this loss to improve the adversarial examples. Our goal is to create generalized adversaries regardless of the target model, as in many scenarios the exact target model's gradients are not available.

For SupCon loss training, given a batch of clean images, we generate adversarial examples using the Projected Gradient Descent (PGD) attack, constrained by an l_∞-norm budget. These adversarial examples are then treated as positive pairs for their corresponding clean images within the SupCon loss calculation.

Our second approach for SupCon loss training utilizes pre-generated adversarial images. For a given batch of adversaries, we compute a mask based on each example's target label, denoting whether the pair is a positive or negative pair. Using this mask, we then calculate the SupCon loss by masking the euclidean distance between each sample in the batch, then summed for each sample.

Additionally, we conducted an experiment where, after training a model with SupCon loss, we froze the ResNet-50 backbone and retrained only the classification head using cross-entropy loss on clean data. This was done to investigate whether the performance of the SupCon model could be improved by fine-tuning the classification layer.

Implementation Challenges. One of the key technical challenges we faced was the computational overhead introduced by generating adversarial examples on-the-fly for every training iteration. Initially, adversarial samples were being generated inside the custom loss function for each incoming batch. While this setup was straightforward to implement, it led to significant slowdowns during training.

To address this, we first tried designing a custom batching strategy that decoupled adversarial example generation from the loss computation and introduced better control over when adversarial samples were created and used. Specifically, instead of generating adversarial examples on every iteration, we modified the data loading pipeline to refresh adversarial samples at fixed intervals (every X iterations). When new adversarial samples were generated, they were temporarily added to the dataset and assigned as additional positive pairs for the contrastive loss. After the designated interval, the outdated adversarial examples were replaced with freshly generated ones, ensuring both diversity and efficiency. Implementing this solution required building a custom data loader capable of dynamically updating its internal sample pool, including managing the separation of clean and adversarial samples to prevent duplication or label leakage.

Ultimately, this dynamic batching approach also presented implementation challenges. Consequently, we adopted a simpler pre-generation method. This involved separating the generation and training processes entirely. We first pre-generated all adversarial examples using the *resnet_attack.py* script, and subsequently, the training script (*main.py*) loaded and utilized these pre-computed adversarial examples. This approach allowed us to maintain the benefits of adversarial supervision without incurring the prohibitive computational costs of per-iteration generation.

4 Experiments

Experimental Setup We conducted a series of experiments to evaluate the effectiveness of our adversarially augmented contrastive learning approach. We used the ImageNet-1K dataset for training and evaluation. A ResNet-50 architecture served as the backbone model. Adversarial examples were generated using both the Projected Gradient Descent (PGD) attack, with an l_∞-norm constraint to limit the perturbation size, and our SupCon attack, with the same parameters. We trained models using three different loss configurations:

- **Cross-Entropy (CE) Loss:** The standard classification loss.

- **Supervised Contrastive (SupCon) Loss:** A contrastive loss that pulls together representations of samples from the same class and pushes apart representations from different classes.

- **Combined Loss:** Applying SupCon loss and finetuning with CE loss.

We evaluated the models on both clean images and adversarial examples generated by the PGD attack. Additionally, we evaluated the models against a contrastive-based attack (SC), which aims to maximize the distance between original embeddings and adversarial embeddings.

Results and Analysis The results of our experiments are summarized in Table 4.

Model	Evaluation	
	Adversarial	Clean
Clean	0.0228	0.8149
PGD - CE Loss	0.6912	0.7222
PGD - SupCon Loss	0.1788	0.4028
PGD - Both Loss	0.6640	0.7031
SC - CE Loss	0.1188	0.7312
SC - SupCon Loss	0.1844	0.4082
SC - Both Loss	0.1188	0.7312
SC - SupCon Retrained	0.1720	0.3135

Table 1: Accuracy Experimental Results. Models are labeled as [attack strategy]-[training loss strategy]

- **Clean Accuracy:** As expected, the model trained solely with the standard Cross-Entropy (CE) loss achieved the highest accuracy on clean images (81.49%). Models trained with SupCon loss, either alone or in combination with CE loss, exhibited significantly lower clean accuracy. This suggests a potential trade-off between standard classification performance and the learning of contrastive representations.

- **Adversarial Robustness:** Out of the models trained with PGD adversaries, CE loss performed the best with 69.12% accuracy, followed by the combined loss at 66.40%. Training with just SupCon loss provided a slight improvement in adversarial robustness compared to the baseline model. However, the robustness for models trained with SupCon adversaries shows opposite trends, with the highest accuracies using SupCon loss at 18.44% followed by the retraining approach at 17.20%. Both CE and combined loss gave the same results.

- **Retrained Head:** We hypothesized that retraining the classification head of the SupCon model after contrastive training might improve clean accuracy. However, the results show that this approach failed to recover the performance of the standard CE loss model.

Analysis As expected, the baseline model performs well on clean data but fails when testing for adversarial robustness, highlighting its vulnerability to attacks. Overall, PGD + CE loss yielded the best adversarial robustness, and its relatively high accuracy on an unpoisoned dataset shows that the this method provides a good robustness-utility trade-off. However, PGD + SupCon loss significantly underperforms. Despite its minor improvements to adversarial robustness compared to the baseline model, its robustness is well under the accuracy of CE loss. It also drops the clean accuracy in half compared to the baseline model. Combining both losses together performs worse overall than just using CE loss, suggesting that our implementation of SupCon loss is not very helpful for this task and/or model scenario.

Using the SupCon adversaries for training, we see once again that CE loss retains the most utility and improves the adversarial robustness compared to the baseline model. However, unlike with the PGD adversaries, its robustness is the worst out of all of the training methods. Using just the SupCon loss with the SupCon adversaries reached the highest robustness, but at the cost of its clean utility, with similar performance to the training with PGD adversaries. Similar to before, training with both losses provides no meaningful difference.

A key observation is that the SupCon loss directly modifies the feature embeddings of the ResNet-50 model, bypassing the final classification layer. With enough updates, the model learns different embeddings that are now off sync from the unchanged classification head, which maps these embeddings to class probabilities. We hypothesized this is the cause for the significantly lower clean accuracy when training with just SupCon loss.

To investigate whether the poor performance of SupCon loss was due to a poorly trained classification head, we performed an additional experiment where we froze the ResNet-50 backbone after SupCon training and retrained only the classification head. However, as shown in Table 4, this retraining did not substantially improve either clean or adversarial accuracy. This suggests that the limitations of SupCon loss in our setup stem from the feature representations themselves, rather than solely from the classification head.

Our results indicate that while contrastive learning aims to create more robust representations, the standard SupCon loss, as implemented in our experiments, does not provide significant adversarial robustness, especially against contrastive attacks.

The vulnerability to the contrastive attack is a critical finding. It suggests that adversarial training strategies need to be carefully designed to defend against a wide range of potential attacks, including those that directly manipulate the learned embeddings.

Further investigation is needed to explore different architectures, training strategies, and loss functions to effectively combine the benefits of contrastive learning with adversarial robustness. Future work should also focus on developing more robust evaluation metrics that capture a model's resilience to diverse attack types.

Loss Score Analysis In addition to accuracy, we analyzed the loss scores of the trained models to gain further insights into training and the models' internal representations. The loss values provide a measure of how well the model's predictions align with the ground truth labels. Lower loss values generally indicate better alignment. We present the loss scores for both adversarial and clean evaluations in Table 2.

Table 2: Loss Experiment Results

Model	**Evaluation**	
	Adversarial Loss	**Clean Loss**
Clean	6.9071	6.5415
PGD - CE Loss	6.2583	6.2300
PGD - SupCon Loss	6.9067	6.9045
PGD - Both Loss	6.2822	6.2510
SC - CE Loss	6.7986	6.2117
SC - SupCon Loss	6.9063	6.9039
SC - Both Loss	6.7986	6.2117
SC - SupCon Retrained	6.9028	6.8963

Observations on Loss Scores

- **Clean Training:** The model trained only on clean images exhibits a lower loss on clean images (6.5415) compared to adversarial examples (6.9071). This is expected, as the model is optimized solely for clean image classification.
- **PGD Training:** Models trained with PGD adversarial training show relatively lower loss values compared to those trained with SC attacks. Notably, the CE Loss model maintains a consistently low loss for both adversarial (6.2583) and clean (6.2300) evaluations, further supporting its robustness to PGD attacks as observed in the accuracy results. The SupCon Loss model, on the other hand, has the highest loss values (around 6.9) in the PGD training group, indicating a struggle to fit the data, which aligns with its poor accuracy.
- **SC Training:** Models trained with SC attacks generally exhibit higher loss values on adversarial examples compared to clean examples, suggesting that these attacks pose a greater challenge to the models. Interestingly, the CE and Both Loss models under SC training have identical loss scores, hinting at similar learned representations in this adversarial setting. The SupCon loss models maintain high loss, similar to the PGD training results.
- **SupCon Retrained:** Retraining the head of SupCon model does not significantly improve the loss scores, re-

inforcing the conclusion that the feature representations learned by SupCon are less effective for this task.

Interpretation The loss score analysis complements the accuracy results. The lower loss values often correlate with higher accuracy, and the differences in loss between adversarial and clean evaluations provide insights into a model's robustness. The high loss values observed for SupCon loss models across different training scenarios further confirm that directly optimizing for contrastive representations, in our setup, leads to a less effective model in terms of classification loss and accuracy.

5 Conclusions

In this work, we investigated the use of adversarially augmented contrastive learning to improve the adversarial robustness of deep neural networks. Our experiments revealed several key findings:

- Standard Cross-Entropy (CE) loss provided the best defense from PGD adversarial attacks, achieving an adversarial accuracy close to clean accuracy.
- Supervised Contrastive (SupCon) loss, while designed to learn robust representations, performed poorly in terms of both clean accuracy and adversarial robustness against both PGD and a contrastive-based attack.
- Combining CE and SupCon losses did not consistently improve robustness over using CE loss alone.
- Retraining the classification head after SupCon training failed to recover the clean accuracy, suggesting that the limitations of SupCon loss in our setup stem from the learned feature representations rather than solely from the classification head.
- All models demonstrated significant vulnerability to a contrastive-based attack, highlighting the importance of evaluating robustness against a diverse range of attack types.

These results suggest that adversarial robustness is dependent on the training objective and the type of attack considered. The standard SupCon loss, as implemented here, does not directly optimize the classification head for clean accuracy, and it may not learn feature representations that are inherently robust to all types of adversarial perturbations.

Our study has several limitations. We focused on a unimodal task with a single architecture (ResNet-50) trained on a single dataset and a specific adversarial attack (PGD) for the primary robustness evaluation. The hyperparameter tuning, particularly for the combined loss, was limited, and the scalability of our approach to larger datasets was not explored. Furthermore, the contrastive attack we implemented requires further investigation to fully understand its implications.

Future work should address these limitations by exploring:

- Alternative contrastive learning formulations and training strategies that more effectively integrate classification objectives and promote robustness against diverse attack types.
- Methods for adapting or jointly training the classification head within a contrastive learning framework to improve clean accuracy without sacrificing robustness.
- A broader range of adversarial attacks and evaluation metrics to provide a more comprehensive assessment of model robustness.
- The scalability of adversarially robust contrastive learning to larger datasets and more complex architectures, particularly those with a well-defined latent space.
- Embedding analysis to understand why SupCon underperforms.

Ultimately, developing robust deep learning models requires a holistic approach that considers both the training objective and the evaluation methodology, with a focus on defending against a wide spectrum of potential adversarial manipulations.

6 Contributions

- **Josh Landsman:** Wrote code for PGD adversarial attack, retraining head with CE, and initial evaluation metrics. Debugged any pipeline issues during training. Trained all models and did data analysis, final paper, and final presentation creation. Presented final project.
- **Kevin Zhao:** Wrote code for PGD adversarial generation, SupCon adversarial generation, training implementation for both CE and SupCon loss, evaluation statistics on both adversarial and clean datasets, and complete pipeline for data loading and processing to training and evaluation. Assisted data analysis, final presentation creation, and final paper. Presented final project.
- **Abinav Chari:** Proposed initial idea. Attempted approach 1 for project implementation.

References

[1] Chen, W., Tian, Y., Isola, P., and Krishnan, D. Supcontrast: Supervised contrastive learning. https://github.com/HobbitLong/SupContrast, 2020. Accessed: 2025-04-18.

[2] Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples, 2015.

[3] Hu, H., Wang, X., Zhang, Y., Chen, Q., and Guan, Q. A comprehensive survey on contrastive learning. *Neurocomputing 610* (2024), 128645.

[4] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning, 2021.

[5] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks, 2019.

[6] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2024.

[7] Wang, L., Nayyem, N., and Rakin, A. Enhancing adversarial robustness of deep neural networks through supervised contrastive learning, 2024.