DA-300 - Lab Based Project

A Study on Loss Functions for Adversarial Finetuning of Contrastive Models

Submitted By:

Aayan Yadav (22323001)
Shree Singhi (22125035)

Advisor

Prof. Sanjeev Kumar

Mehta Family School of Data Science and Artificial Intelligence
Indian Institute of Technology, Roorkee

May 16, 2025

Final Report

Mehta Family School of Data Science and Artificial Intelligence
Indian Institute of Technology, Roorkee

Project Title:	A Study on Loss Functions for Adversarial Finetuning of Contrastive Models
Student 1:	Aayan Yadav (22323001)
Student 2:	Shree Singhi (22125035)
Primary Supervisor:	Prof. Sanjeev Kumar

Self-supervised contrastive learning has shown remarkable success in learning visual representations without labeled data. However, these models remain vulnerable to adversarial perturbations, posing challenges in security-sensitive applications. In this work, we investigate methods to enhance the adversarial robustness of SimCLR with a ResNet-50 backbone through unsupervised adversarial fine-tuning. Building on the FARE approach from RobustCLIP, we propose a modified loss function that incorporates both adversarial alignment and clean consistency terms to balance robustness and representational fidelity. Experiments on CIFAR-10 and ImageNet provide key insights regarding generalization of adversarial training. Code is available at `https://github.com/ydvaayan/loss-unsup-adv`.

1 Introduction

Deep learning has revolutionized machine learning across diverse domains, particularly in computer vision and natural language processing. However, the remarkable achievements of these models predominantly rely on supervised learning paradigms that demand substantial quantities of labeled data, which is often expensive and time-consuming to obtain(Chen et al., 2020; Bengio et al., 2013). To address this limitation, self-supervised learning (SSL) has emerged as a powerful alternative, enabling models to learn meaningful representations from unlabeled data by generating supervisory signals from the data itself(Jing & Tian, 2020; Vincent et al., 2010; Kingma & Welling, 2014; Shwartz-Ziv et al., 2023).

Self-supervised learning encompasses a broad spectrum of techniques ranging from generative approaches like autoencoders to more recent contrastive methods. Comprehensive surveys(Shwartz-Ziv et al., 2023) highlight the rapid evolution of this field and its potential to transform how machines learn without explicit human supervision. This paradigm is particularly valuable in domains where labeled data is scarce or prohibitively expensive to acquire, such as medical imaging, where ethical considerations, privacy constraints, and requirements for expert annotation present significant challenges.

Among the various self-supervised approaches, contrastive learning methods have demonstrated exceptional effectiveness in learning visual representations (Chen et al., 2020; He et al., 2020). These methods operate by maximizing agreement between differently augmented views of the same data example (positive pairs) while pushing apart representations from different examples (negative pairs). SimCLR (Simple Framework for Contrastive Learning of Visual Representations) (Chen et al., 2020) stands out as an elegant and effective framework that achieves state-of-the-art performance without requiring specialized architectures or memory banks. Using ResNet-50 (He et al., 2015) as its backbone, SimCLR achieves 76.5% top-1 accuracy on ImageNet with linear evaluation, matching the performance of supervised counterparts and showing remarkable results in semi-supervised settings with limited labeled data. Other notable contrastive learning frameworks include Momentum Contrast (MoCo) (He et al., 2020), which constructs a dynamic dictionary using a queue and moving-averaged encoder, and Bootstrap Your Own Latent (BYOL) (Grill et al., 2020), which eliminates the need for negative pairs through an asymmetric architecture. These innovations collectively demonstrate the power of self-supervised contrastive learning in extracting rich representations from unlabeled data, significantly reducing dependence on labor-intensive annotation processes.

Despite these advances, recent research has revealed that deep neural networks, including self-supervised models, exhibit significant vulnerability to adversarial attacks-subtle perturbations to input data that are imperceptible to humans yet cause dramatic alterations in model predictions (Madry et al., 2018; Goodfellow et al., 2015; Kim et al., 2020). This susceptibility raises serious concerns regarding the deployment of these models in security-critical applications where adversaries might exploit such weaknesses. Consequently, adversarial robustness-the capacity of models to maintain performance under adversarial conditions-has become an essential research direction (Madry et al., 2018; Croce & Hein, 2020).

In supervised learning contexts, adversarial training has emerged as a leading defense mechanism, integrating adversarially perturbed examples into the training process (Madry et al., 2018). However, adapting these techniques to self-supervised learning frameworks presents unique challenges due to the absence of explicit labels (Kim et al., 2020). Recent work on Robust Contrastive Learning (RoCL) (Kim et al., 2020) has begun addressing this gap by introducing instance-wise adversarial attacks designed to confuse instance-level identities, though significant room for improvement remains. In this work, we build upon these foundations to enhance the adversarial robustness of contrastive learning frameworks, specifically focusing on SimCLR with ResNet-50 as the backbone architecture. Drawing inspiration from the recent Robust CLIP framework (Schlarmann et al., 2024), we propose an unsupervised adversarial fine-tuning approach that aims to improve model robustness while preserving clean data performance. Our key contribution is the introduction of a modified loss function that combines an adversarial alignment term with a clean consistency regularization term, effectively balancing robustness against adversarial attacks with fidelity to original representations. Through extensive experiments on CIFAR-10 and ImageNet we study the impact of this term and various hyperparameters.

2 Background

Symbol	Description
ϕ	Encoder which is getting finetuned
ϕ_{Org}	Original (pretrained) encoder
ϕ_{FT}	Finetuned encoder
x	Clean input image
z	Adversarially perturbed input
ϵ	Maximum perturbation magnitude (in ℓ_∞ norm)
α	Weighting parameter for regularization

Table 1: Notation used in report.

In recent years, self-supervised learning has emerged as a powerful paradigm for training deep neural networks without relying on labeled data. Among various self-supervised approaches, contrastive learning methods have demonstrated remarkable success in learning useful visual representations that can be transferred to downstream tasks with competitive performance compared to supervised learning (Chen et al., 2020). This section provides background on contrastive learning frameworks, specifically SimCLR, the challenges of adversarial vulnerability in these models, and the approach we adopt from the Robust CLIP framework to enhance robustness.

2.1 Contrastive Learning and SimCLR

SimCLR (Simple Framework for Contrastive Learning of Visual Representations), introduced by Chen et al. (2020), has established itself as a pioneering approach in self-supervised visual representation learning. The framework learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space.

SimCLR consists of several key components: a stochastic data augmentation module that creates multiple views of each training sample; a neural network base encoder (typically ResNet) that extracts representation vectors; a projection head that maps these representations to a space where the contrastive loss is applied; and a contrastive loss function that encourages representations from the same image to be similar, while pushing apart representations from different images.

SimCLR's simplicity and effectiveness without requiring specialized architectures or memory banks is a major advantage. The model has achieved impressive results, with a linear classifier trained on its self-supervised representations reaching 76.5% top-1 accuracy on ImageNet, matching the performance of supervised ResNet-50. When fine-tuned on just 1% of labels, SimCLR achieves 85.8% top-5 accuracy, substantially outperforming earlier approaches with far fewer labels.

2.2 ResNet-50 Architecture

ResNet-50 (He et al., 2015) is a convolutional neural network (CNN) architecture that addresses the degradation problem in deep neural networks through the use of residual connections. As networks become deeper, their accuracy tends to saturate and then degrade. ResNet solves this through "residual blocks" that allow direct flow of information across layers via skip connections, effectively mitigating the vanishing gradient problem.

In SimCLR, ResNet-50 serves as the backbone encoder, extracting 2048-dimensional feature vectors from input images before they are processed by the projection head. The architecture's depth and efficiency make it well-suited for learning rich visual representations capturing both low-level and high-level image semantics.

2.3 Adversarial Vulnerability in Representation Learning

Despite strong performance, models like SimCLR are highly vulnerable to adversarial attacks—subtle perturbations to input images that are imperceptible to humans but cause incorrect model outputs (Kim et al., 2020). This vulnerability is critical in security-sensitive applications, where adversaries might exploit such weaknesses to bypass systems or influence decisions.

Studies show that contrastive learning frameworks like SimCLR are particularly susceptible to adversarial attacks, suffering major performance degradation even with minor perturbations. This arises from the model's sensitivity to features that may be useful for contrastive tasks but lack adversarial robustness.

2.4 Adversarial Training for Robust Representations

Adversarial training, which incorporates adversarial examples into training, has emerged as a strategy to improve model robustness. In supervised learning, this involves generating adversarial examples and training the model to classify them correctly (Croce & Hein, 2020; Madry et al., 2018; Goodfellow et al., 2015).

In self-supervised contrastive learning, the absence of labels presents challenges. RoCL (Robust Contrastive Learning) (Kim et al., 2020) introduces instance-wise adversarial attacks to confuse instance-level identities, thereby improving robustness while preserving clean data performance.

2.5 Unsupervised Adversarial Fine-tuning from Robust CLIP

Schlarmann et al. (2024) introduce an unsupervised adversarial fine-tuning method that we adapt for SimCLR. Known as FARE (Fine-tuning for Adversarially Robust Embeddings), this technique enhances robustness while preserving clean input performance.

FARE fine-tunes the vision encoder so that features of adversarial inputs remain close to those of clean inputs. This is achieved with the loss function:

$$\mathcal{L}_{\text{FARE}}(\phi, x) = \max_{\|z-x\|_\infty \leq \epsilon} \|\phi(z) - \phi_{\text{Org}}(x)\|_2^2$$

This method is unsupervised i.e. does not require labels. This preserves integrity of original embeddings and creates a robust model whose robustness generalizes across downstream tasks.

We adapt FARE to SimCLR with ResNet-50 and modify it, aiming to produce representations that are both semantically rich and adversarially robust.

3 Method

3.1 Attack

We use 10 iterations of the Auto-PGD (APGD) adversarial attack with $\epsilon = \frac{8}{255}$ to generate adversarial examples. APGD (Croce & Hein, 2020) is a variant of Projected Gradient Descent (PGD) that adaptively selects the step size using a backtracking line search and incorporates momentum to improve optimization stability. At each iteration t, the adversarial example $x^{(t)}$ is updated as:

$$x^{(t+1)} = \Pi_{\mathcal{B}_\epsilon(x)} \left(x^{(t)} + \alpha_t \cdot \text{sign}(\nabla_x \mathcal{L}(f(x^{(t)}), y)) \right),$$

where $\Pi_{\mathcal{B}_\epsilon(x)}$ denotes the projection onto the ℓ_∞-ball of radius ϵ centered at the original input x, α_t is the step size at iteration t, $f(\cdot)$ is the model (in this case, a linear classifier), and $\mathcal{L}$ is the loss function, typically cross-entropy. The attack is applied using a linear classification layer as a surrogate model, which enables efficient and transferable adversarial example generation. This training strategy encourages the model to learn robust features by exposing it to strong, adaptive perturbations.

3.2 Additional Term in Loss

In this work, we build upon the loss function introduced in Schlarmann et al. (2024) by incorporating an additional regularization term aimed at preserving clean image representations. Our goal is to enhance robustness against adversarial perturbations while ensuring that the learned embeddings remain consistent for unperturbed (clean) inputs as well.

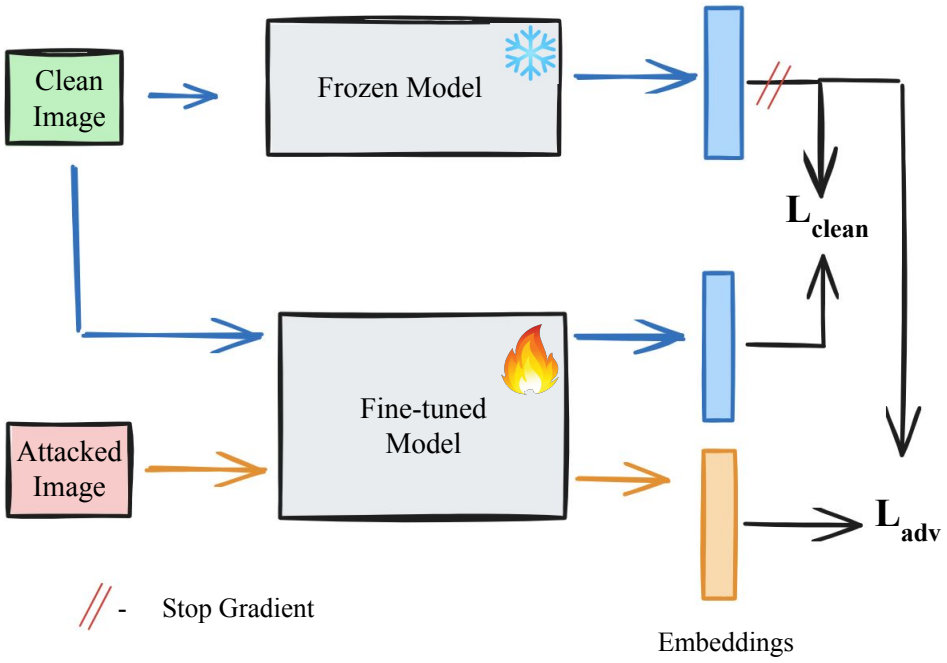

Figure 1: We try to preserve both embeddings of clean and attacked images unlike Schlarmann et al. (2024).

Adversarial Loss. We begin by adopting the adversarial alignment loss from RobustCLIP. This loss encourages the representation of an adversarially perturbed input to remain close to the original (frozen) CLIP embedding. Formally, the adversarial loss is defined as:

$$\mathcal{L}_{\text{adv}}(\phi, x) = \max_{\|z-x\|_\infty \leq \epsilon} \|\phi(z) - \phi_{\text{Org}}(x)\|_2^2$$

Here, x is the clean input image, z is an adversarially perturbed version of x constrained within an ℓ_∞ ball of radius ϵ, $\phi(z)$ denotes the current model's embedding of the adversarial input, and $\phi_{\text{Org}}(x)$ is the frozen original CLIP embedding of the clean image. This loss penalizes deviation of the adversarial feature from the original feature and forms the core of RobustCLIP's training strategy.

Clean Consistency Term. While $\mathcal{L}_{\text{adv}}$ promotes robustness by aligning adversarial features with the original CLIP features, it does not explicitly enforce consistency for clean inputs. During adversarial training, this can lead to unintended shifts in the model's output even for clean examples, as the optimization primarily focuses on the worst-case adversarial directions.

To address this, we propose an additional clean consistency term:

$$\mathcal{L}_{\text{clean}}(\phi, x) = \|\phi(x) - \phi_{\text{Org}}(x)\|_2^2$$

This term directly penalizes the deviation of the model's embedding $\phi(x)$ for the clean input from the frozen original embedding $\phi_{\text{Org}}(x)$. Including this term encourages the model to retain fidelity to the original encoder representation for unperturbed data, helping to preserve semantic integrity and avoid overfitting to adversarial examples.

Combined Loss Function. Our final training objective combines the adversarial and clean consistency losses. The combined loss is given by:

$$\mathcal{L}_{\text{Ours}}(\phi, x) = \max_{\|z-x\|_\infty \leq \epsilon} \|\phi(z) - \phi_{\text{Org}}(x)\|_2^2 + \alpha \|\phi(x) - \phi_{\text{Org}}(x)\|_2^2$$

Here, α is a weighting hyperparameter that balances the influence of the clean term relative to the adversarial loss. The first term remains identical to RobustCLIP, focusing on robustness, while the second term enforces stability on clean inputs.

Fine-Tuning Objective. To optimize the model, we minimize the combined loss over the dataset:

$$\mathcal{L}_{\text{FT}}(\phi) = \arg\min_{\phi} \sum_{i=1}^{n} \mathcal{L}_{\text{Ours}}(\phi, x_i)$$

This fine-tuning objective ensures that the model learns to resist adversarial perturbations while maintaining alignment with the original feature space on clean images.

Motivation and Benefit. By including the clean consistency term, we aim to address a common drawback in adversarial training—degradation in clean accuracy and semantic drift of representations. Our combined loss function enables the model to learn robust yet semantically faithful representations, improving generalization in both clean and adversarial settings.

In summary, while $\mathcal{L}_{\text{adv}}$ aligns with the objective of RobustCLIP in securing robustness, the proposed addition of $\mathcal{L}_{\text{clean}}$ encourages representational stability. Together, they form a loss function that balances adversarial robustness with clean feature preservation.

4 Experiments

Dataset We conducted experiments using CIFAR-10 (Krizhevsky, 2009) and ImageNet (Deng et al., 2009). CIFAR-10 consists of 60,000 32×32 color images in 10 classes, with 6,000 images per class, commonly used for evaluating image classification models in low-resolution settings. ImageNet is a large-scale dataset with over 1 million high-resolution images categorized into 1,000 classes, serving as a standard benchmark for large-scale visual recognition tasks. These datasets provide a diverse range of visual concepts and complexities, enabling thorough evaluation of model performance.
Implementation Details We use AdamW optimize with $\beta 1$ and $\beta 2$ equal to 0.9 and 0.95 respectively. We use a linear warmup for learning rate for 1400 steps followed by cosine decay. We train for 20000 steps with batch size of 256. We we use 10 iterations of APGD attack with epsilon value of 8/255. It took 36 hours for performing 20000 steps of training on 1 V100 GPU node. We use SimCLR with ResNet-50 backbone. The pretrained model on CIFAR-10 was trained for 1000 epochs with batch size of 1024 while for ImageNet the model was trained for 200 epochs with batch size of 2000.

4.1 CIFAR-10

We begin our empirical evaluation on the CIFAR-10 dataset to study the effect of our proposed clean consistency term on both clean and adversarial performance. Specifically, we investigate the impact of two key factors: learning rate and the weighting parameter α that controls the influence of the clean term in the loss function.

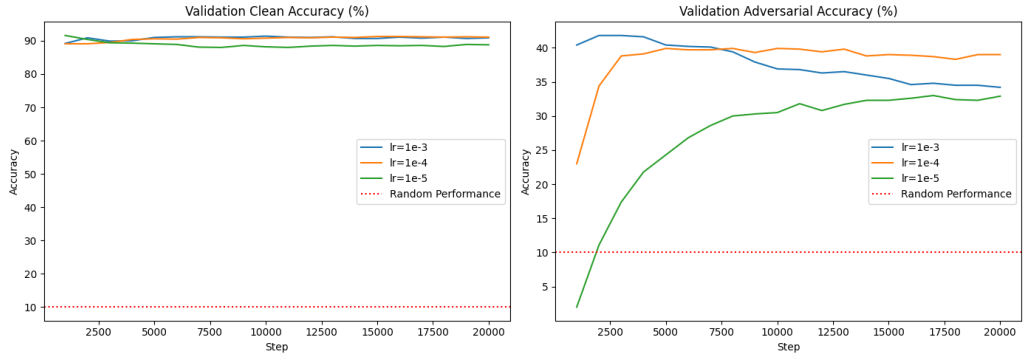

Figure 2: **Left:** Change in learning rate does not affect clean accuracy on validation set significantly. **Right:** Learning rate of 1e-4 is the best candidate with respect to adversarial accuracy among the tested learning rates. All values are logged every 1000 step. We use *alpha* 0.1 for this experiment.

Effect of Learning Rate. Figure 2 (left) shows that varying the learning rate has minimal impact on the clean validation accuracy, which remains stable across a range of values. However, as seen in Figure 2 (right), adversarial accuracy is more sensitive to the choice of learning rate. Among the tested values, a learning rate of 1×10^{-4} achieves the highest adversarial accuracy, indicating that careful tuning of the learning rate is crucial for achieving robust performance without compromising clean accuracy.

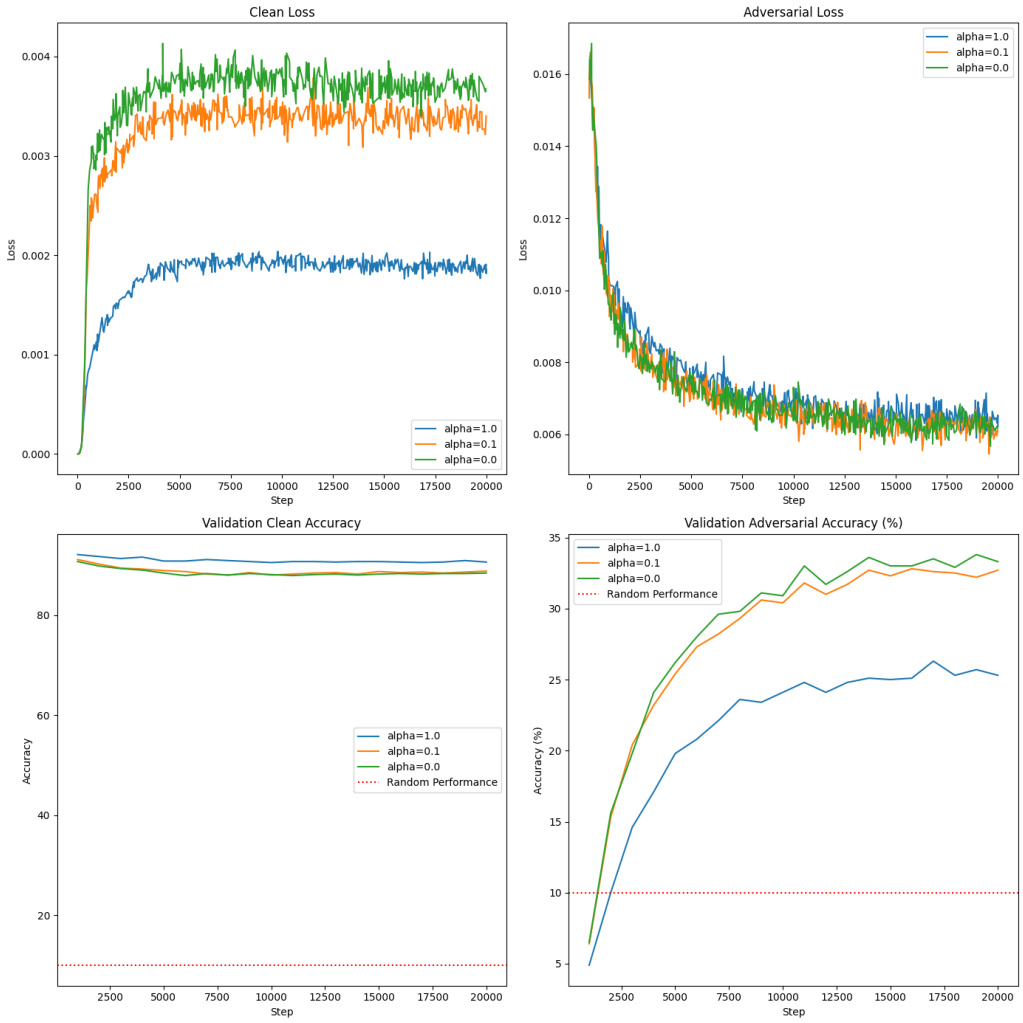

Figure 3: **Top:** Training loss logged at every step. Adversarial loss in all three cases remain similar while clean loss decreases with increase in α. Clean loss saturates while adversarial loss decreases. **Bottom:** Validation accuracy logged every 1000 step. Clean accuracy is approxmately equal and stagnated. Adversarial accuracy decreases with α.

Effect of Clean Consistency Weight α. To further evaluate our loss function, we vary the clean consistency weight α and analyze its influence during training. Figure 4 (top) presents the training losses over time. We observe that the adversarial loss remains relatively unaffected by changes in α, while the clean loss decreases more rapidly as α increases. This confirms that the clean term effectively enforces alignment with the original CLIP embeddings for clean inputs.

Figure 4 (bottom) shows the validation accuracy throughout training. Clean accuracy remains largely unchanged, suggesting that clean feature preservation does not harm clean performance. However, adversarial accuracy declines as α increases, reflecting a trade-off: placing excessive emphasis on clean consistency can reduce the model's resilience to adversarial perturbations.

Overall, these results demonstrate that the clean consistency term supports representational stability for clean images, but its influence must be carefully balanced to maintain adversarial robustness. Increasing *alpha* has same performance on train split but decreasing performance on validation split which indicates overfitting on train data. Poor generalization might also be due to poor resolution and small dataset size of CIFAR-10 images. So we conduct further experiments in ImageNet to get a clearer picture.

4.2 ImageNet-1k

To evaluate the scalability and generalizability of our approach, we conduct experiments on the ImageNet-1k dataset—a more challenging benchmark due to its high resolution and diversity.

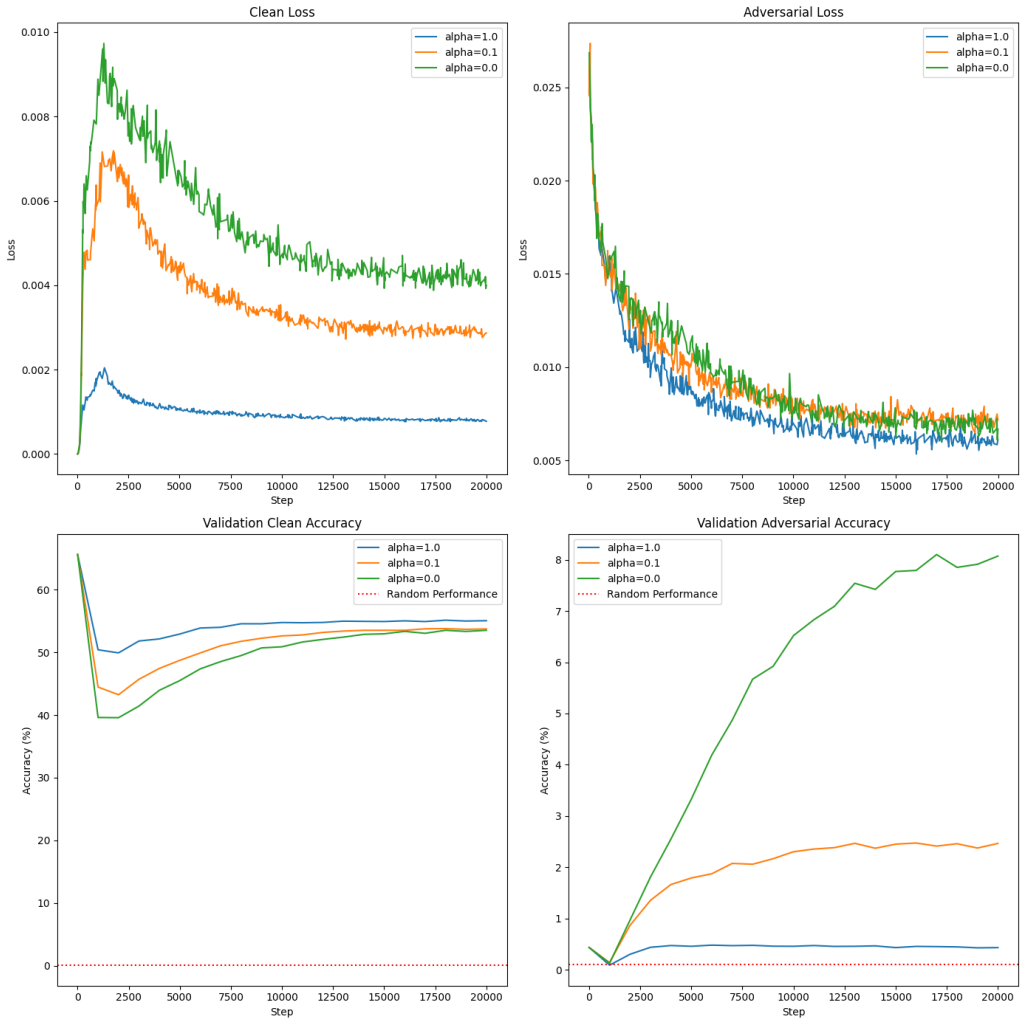

Figure 4: **Top:** Training loss logged at every step. Adversarial loss in all three cases remain similar while clean loss decreases with increase in α. **Bottom:** Validation accuracy logged every 1000 step. Clean accuracy though initially increases with α, converges as trained longer. Adversarial accuracy decreases with α.

Training Behavior. Figure 4 (top) illustrates the training loss curves for different values of α. As with CIFAR-10, the adversarial loss remains consistent across all settings, while the clean loss decreases more substantially with higher values of α. This confirms the effectiveness of the clean consistency term in preserving the original CLIP representations at scale.

Validation Accuracy. In Figure 4 (bottom), we present clean and adversarial validation accuracies. Clean accuracy initially improves with higher α, likely due to stronger enforcement of semantic consistency. However, over time, all models converge to similar clean accuracy levels. On the other hand, adversarial accuracy shows a clear negative correlation with α: higher values result in reduced robustness.

These observations align with our findings on CIFAR-10. Thus addition of extra term of clean loss restricts embedding space such that it hinders generalization.
In summary, our method demonstrates consistent trends across both small-scale and large-scale datasets. The inclusion of the clean consistency term improves the fidelity of clean embeddings, but reduces generalization ability of model. It suggests restricting embedding space near initial embeddings in L2 sense might not be the best approach

5 Conclusion

In this work, we enhance the adversarial robustness of SimCLR through an unsupervised adversarial fine-tuning approach inspired by RobustCLIP. By introducing a clean consistency term alongside the adversarial alignment loss, we encourage stability in clean embeddings while improving resilience against adversarial attacks. Our

experiments on CIFAR-10 and ImageNet reveal that the proposed method tries to effectively balances robustness and clean accuracy, but the extra term in loss function causes overfitting. These findings suggest that restricting the embedding space nearby initial embeddings in L2 sense does not improve clean accuracy significantly and also hinders generalization. We hope our findings can help guide future research in the field.

References

Bengio, Y., Courville, A., & Vincent, P. 2013, IEEE transactions on pattern analysis and machine intelligence, 35, 1798

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, in Proceedings of Machine Learning Research, Vol. 119, Proceedings of the 37th International Conference on Machine Learning, ed. H. D. III & A. Singh (PMLR), 1597–1607. https://proceedings.mlr.press/v119/chen20j.html

Croce, F., & Hein, M. 2020, in Proceedings of the 37th International Conference on Machine Learning (PMLR), 2206–2216. https://proceedings.mlr.press/v119/croce20b.html

Deng, J., Dong, W., Socher, R., et al. 2009, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 248–255, doi: 10.1109/CVPR.2009.5206848

Goodfellow, I. J., Shlens, J., & Szegedy, C. 2015, in International Conference on Learning Representations. https://arxiv.org/abs/1412.6572

Grill, J.-B., Strub, F., Altché, F., et al. 2020, in Advances in Neural Information Processing Systems, Vol. 33, 21271–21284

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. 2020, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9729–9738

He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv preprint arXiv:1512.03385. https://arxiv.org/abs/1512.03385

Jing, L., & Tian, Y. 2020, IEEE transactions on pattern analysis and machine intelligence, 43, 4037

Kim, M., Tack, J., & Hwang, S. J. 2020, in Advances in Neural Information Processing Systems, Vol. 33, 3455–3465. https://arxiv.org/abs/2006.07589

Kingma, D. P., & Welling, M. 2014, in 2nd International Conference on Learning Representations, ICLR 2014

Krizhevsky, A. 2009, Learning Multiple Layers of Features from Tiny Images, Tech. rep., University of Toronto. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. 2018, in International Conference on Learning Representations. https://arxiv.org/abs/1706.06083

Schlarmann, C., Singh, N. D., Croce, F., & Hein, M. 2024, in Proceedings of the 41st International Conference on Machine Learning (ICML). https://dblp.org/rec/conf/icml/SchlarmannSC024.html

Shwartz-Ziv, R., Balestriero, R., Kawaguchi, K., Rudner, T. G. J., & LeCun, Y. 2023, in Advances in Neural Information Processing Systems, Vol. 36

Vincent, P., Larochelle, H., Lajoie, I., et al. 2010, Journal of machine learning research, 11, 3371