Rethinking Robust Contrastive Learning from the Adversarial Perspective

Fatemeh Ghofrani, Mehdi Yaghouti, Pooyan Jamshidi

University of South Carolina

Motivation

The study focuses on three key research questions:

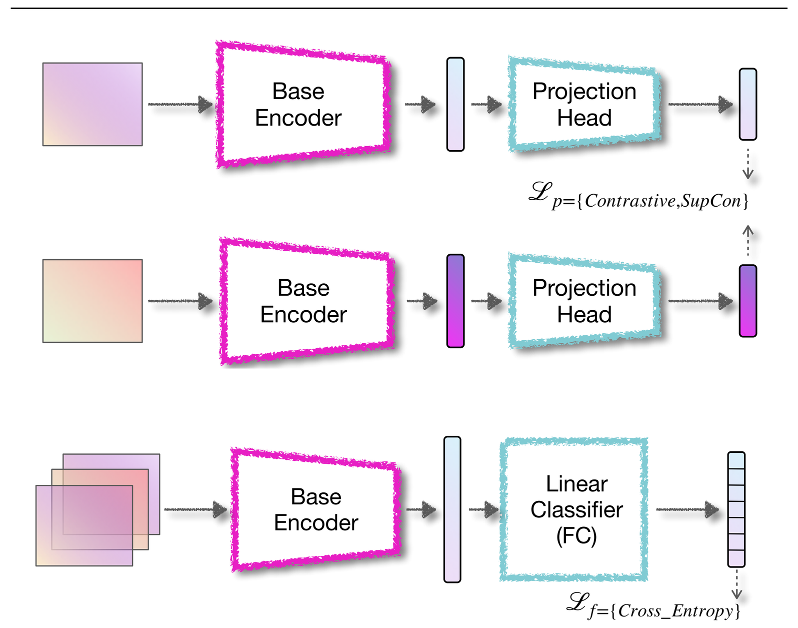
- RQ1: Is there anything special about the robustness of contrastive learning representations?
- RQ2: How does the incorporation of label information affect the robustness of contrastive learning representations?
- RQ3: How does adversarial training affect the learned representations in supervised and contrastive learning?

Methodology

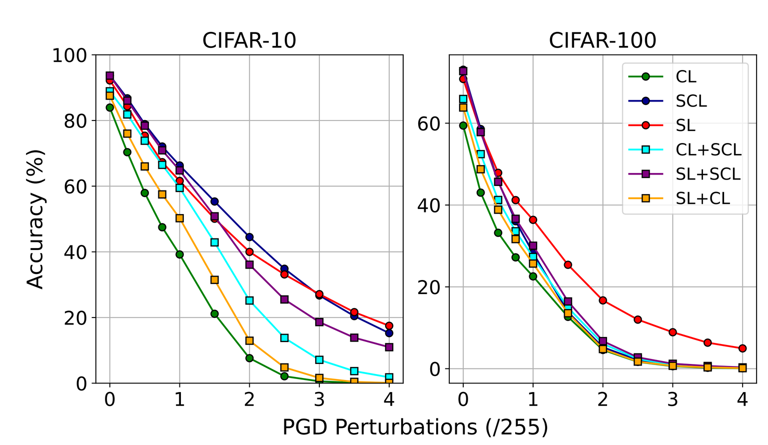
- Representation Learning schemes:
  - Contrastive Learning (CL)
  - Supervised Contrastive Learning (SCL)
  - Supervised Learning (SL)
- Training Scenarios:

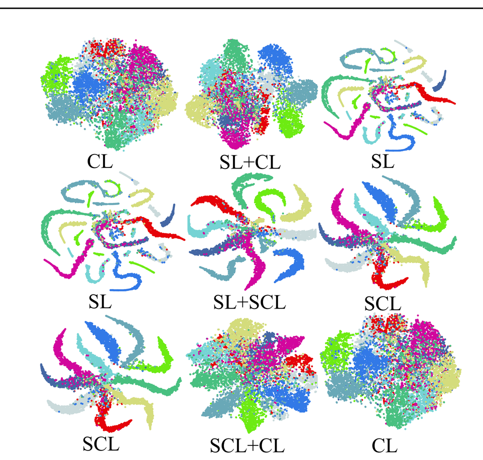
Scenarios	Pretraining Phase	Finetuning Phase
ST	Standard Training	Standard Training (with fixed θ_b)
AT	Adversarial Training	Standard Training (with fixed θ_b)
Partial-AT	Adversarial Training	Partial Adversarial Training (with fixed θ_b)
Full-AT	Adversarial Training	Full Adversarial Training

Training Process



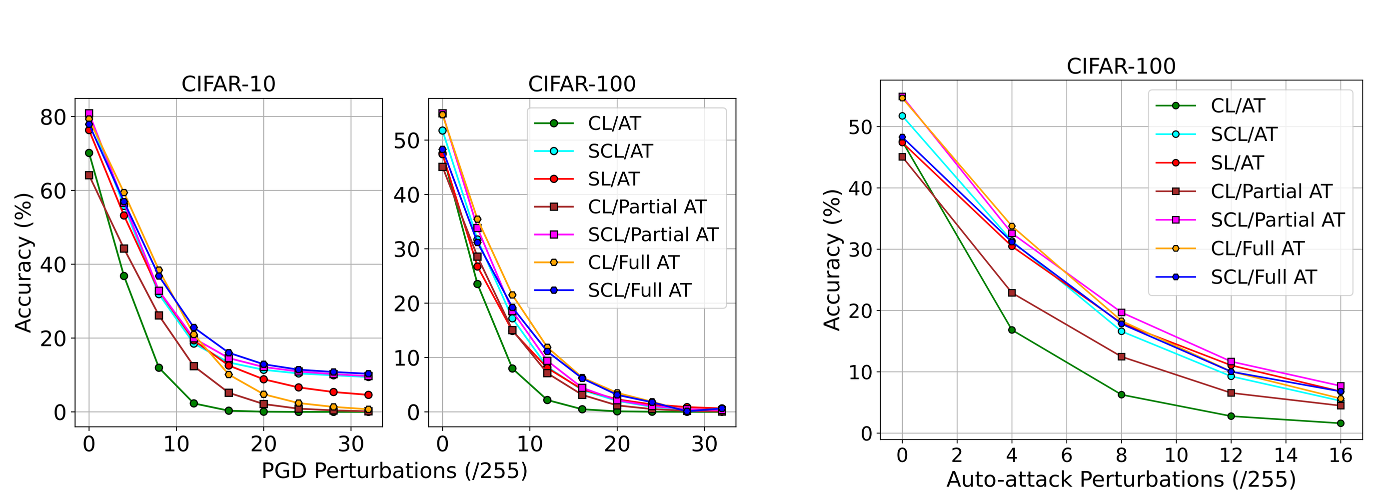
Incorporating label information into CL enhances the robustness of representations



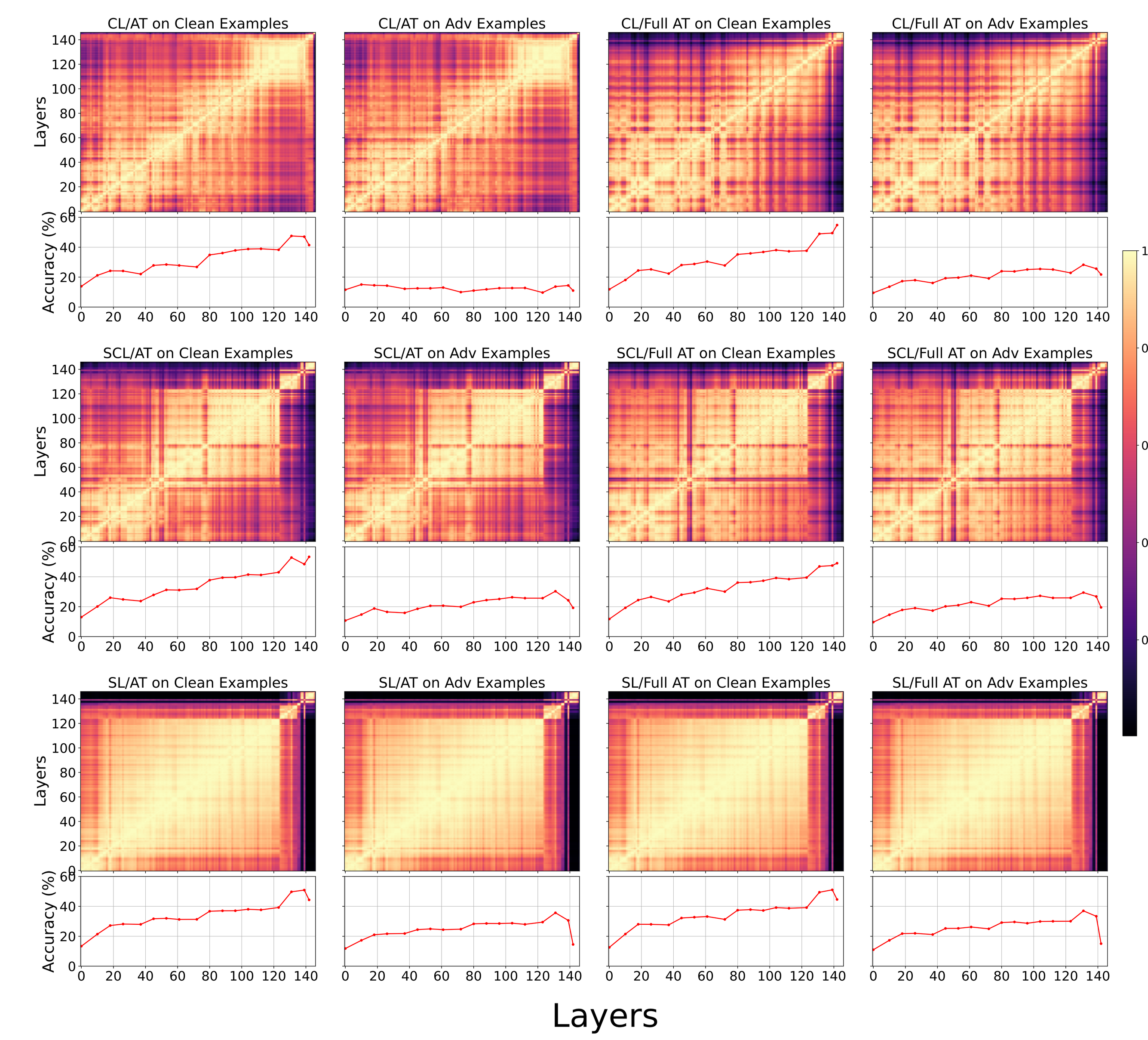


- Both SCL and SL exhibit clearer class boundaries compared to CL.
- Incorporating label information in the semi-supervised learning schemes (SL+CL and SCL+CL) enhances the separation of classes, indicating increased robustness against adversarial perturbations.

Adversarial Training: Direct Comparison

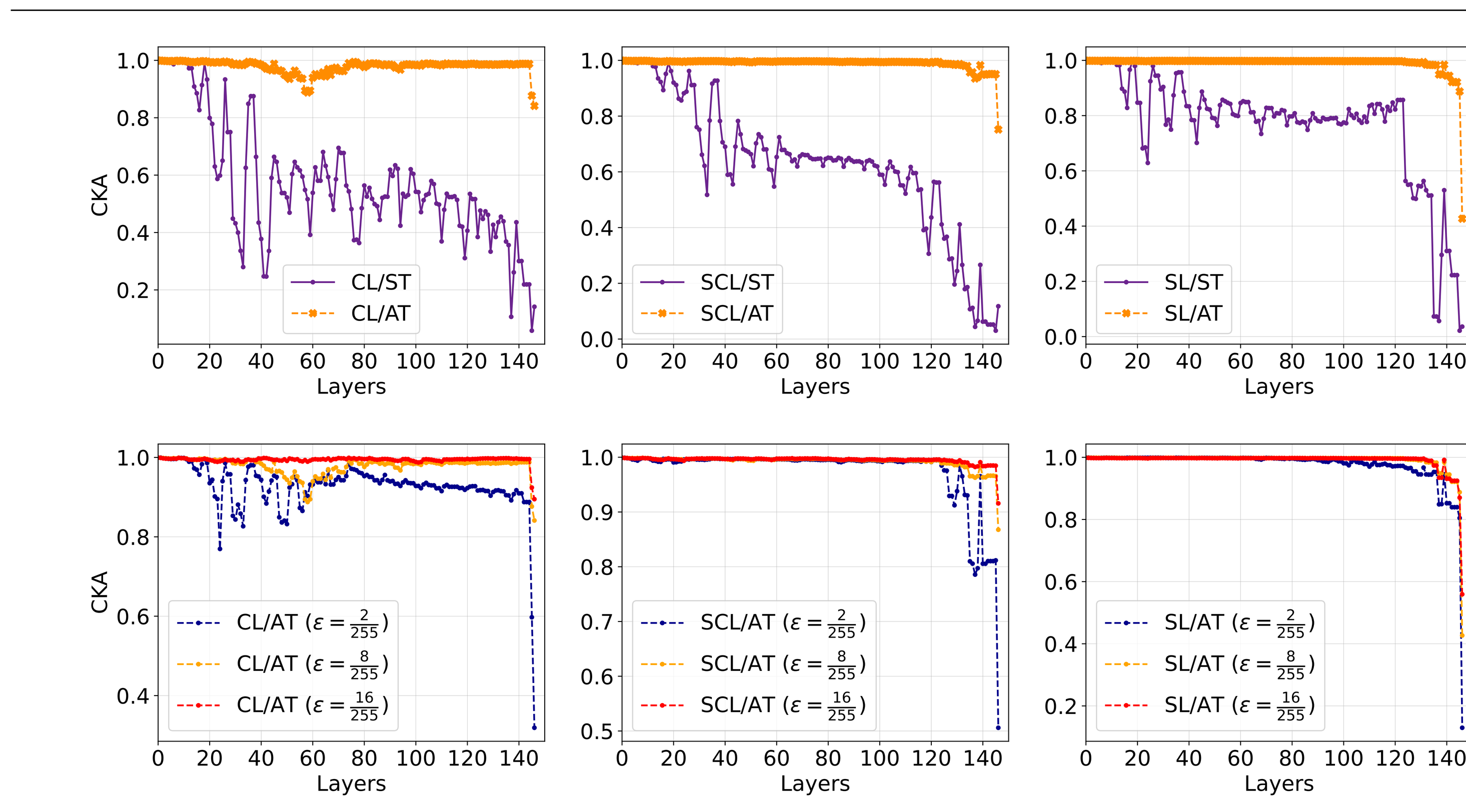


Adversarially trained networks exhibit significant similarities between adversarial and clean representations



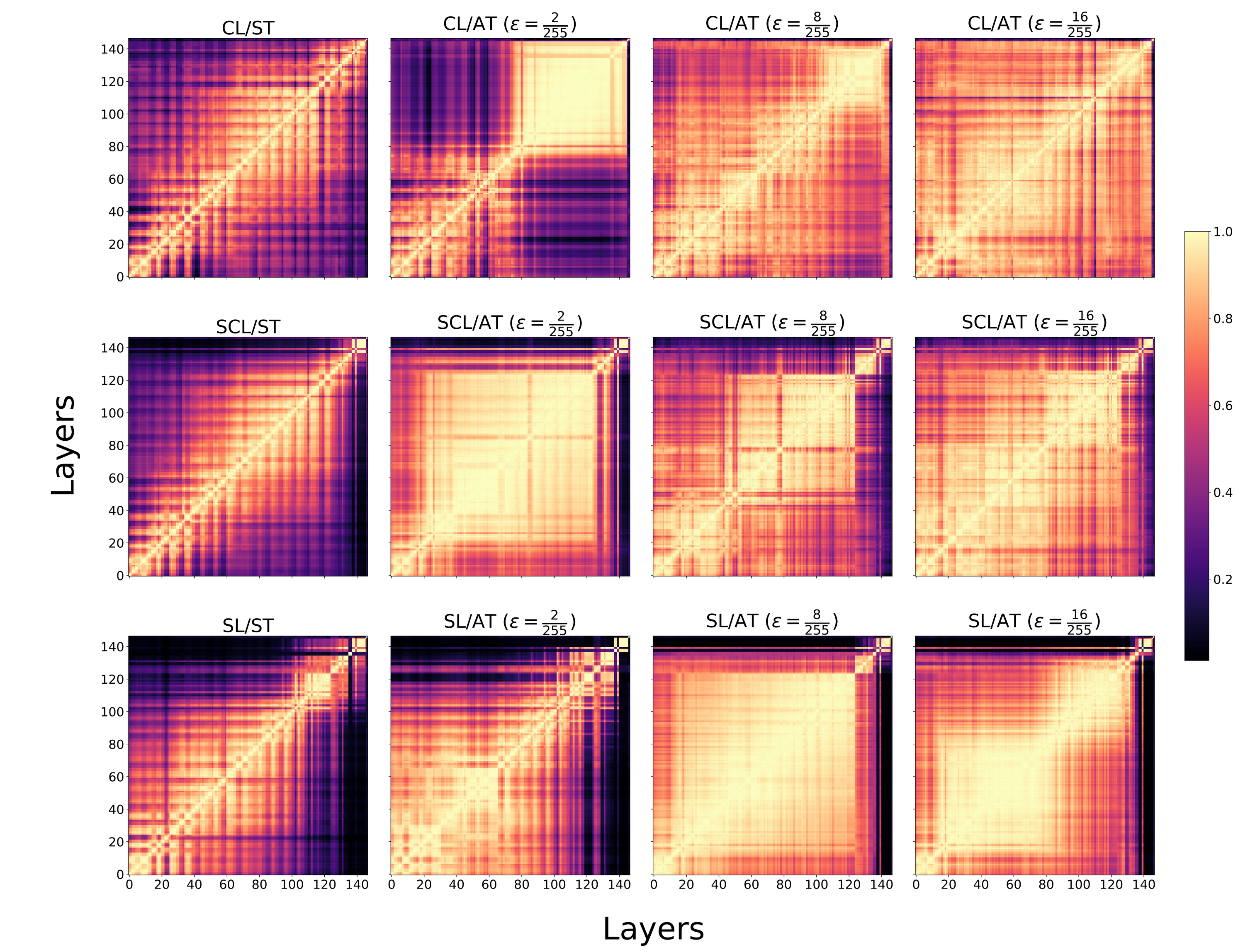
- Networks trained through adversarial training exhibit significant similarities between adversarial and clean representations.
- Full AT significantly enhances long-range similarities and improves both standard and adversarial accuracy in CL.
- Slight differences in representations and performance are observed in the SCL and SL under AT and Full AT scenarios.

Increasing the similarity between adversarial and clean representations improves robustness

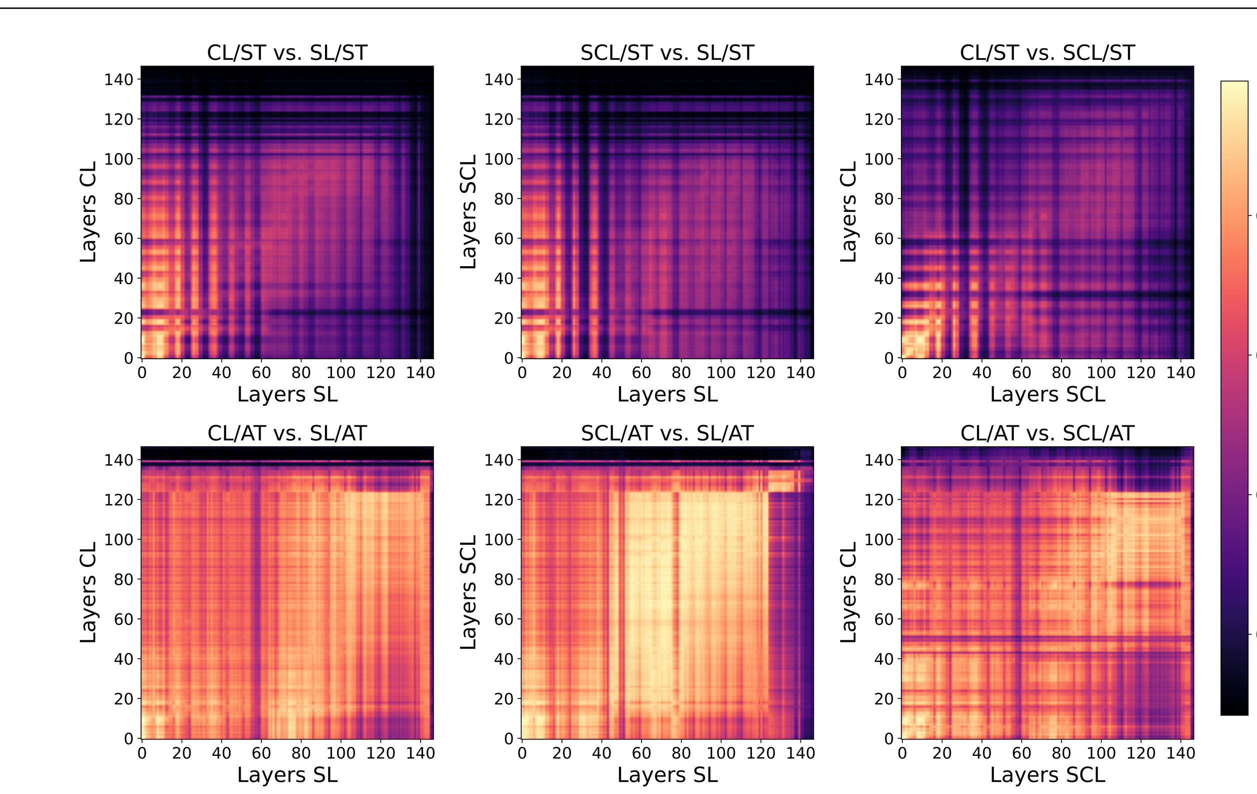


- Comparing clean and adversarial representations in different layers of the model reveals significant dissimilarity in standard-trained networks.
- Adversarial training reduces this divergence, leading to similar representations for clean and adversarial examples in robust networks.

Adversarial training promotes the emergence of long-range similarities between layers



Adversarial training promotes similarity in adversarial representations across various learning schemes



Summary

- **CL without labels is less robust than other learning schemes** in standard training, but incorporating supervised cross-entropy or supervised contrastive loss enhances robustness by utilizing label information.
- **Full adversarial fine-tuning enhances the robustness of representations learned by CL**, but it is ineffective in SCL or standard SL schemes.
- **Adversarial training promotes the convergence of representations towards a universal set**, leading to the increased similarity between adversarial and clean representations and improved robustness, particularly at the network's last layers.