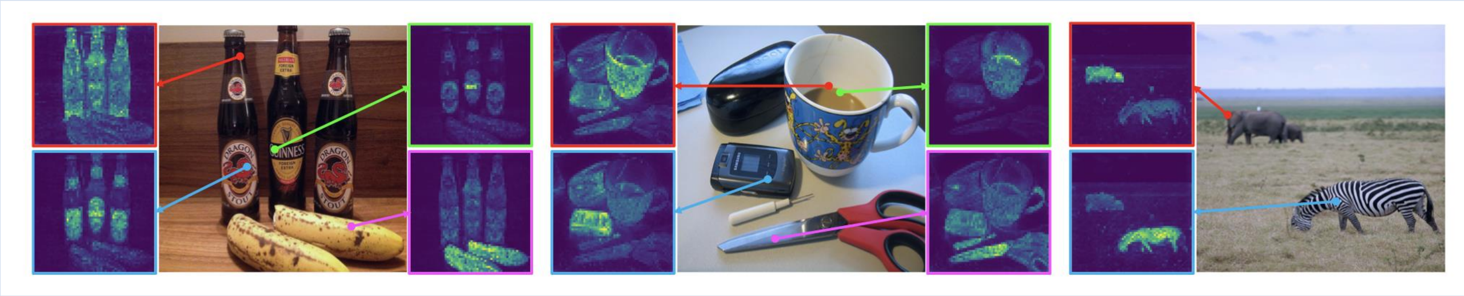
Exploring Adversarial Attacks and Defenses in Vision Transformers trained with DINO

Read the paper

Javier Rando, Nasib Naimi, Thomas Baumann, Max Mathys
ETH Zurich

1 Introduction

Self-supervised Vision Transformers trained with DINO learn human interpretable features[1].



Are they more robust to Adversarial Attacks that preserve input similarity?

2 DINO ViTs Are Not More Robust to Adv. Attacks

		FGSM			PGD			C&W	Clean
		$\epsilon = 0.001$	$\epsilon = 0.03$	$\epsilon = 0.1$	$\epsilon = 0.001$	$\epsilon = 0.03$	$\epsilon = 0.1$	$c = 50$	
DINO	ViT-S/16	52.4%	0.9%	1.1%	49.6%	0.0%	0.0%	0.2%	76.8%
	ViT-B/16	**58.9%**	1.8%	1.5%	**56.8%**	0.0%	0.0%	0.4%	77.9%
Superv.	ViT-B/16	55.1%	**17.3%**	14.5%	47.7%	**0.7%**	**0.1%**	0.8%	**80.2%**
ResNet-50		47.8%	8.0%	**24.3%**	43.9%	0.1%	0.0%	**7.2%**	75.7%

Accuracy of ViT models trained using DINO and supervised (Superv.) learning against different white-box adversarial attacks. Metrics are computed on the whole ImageNet validation set. Last column represents accuracy on the original samples from which adversarial images were generated.

3 Self-Supervision Increases Attack Transferability

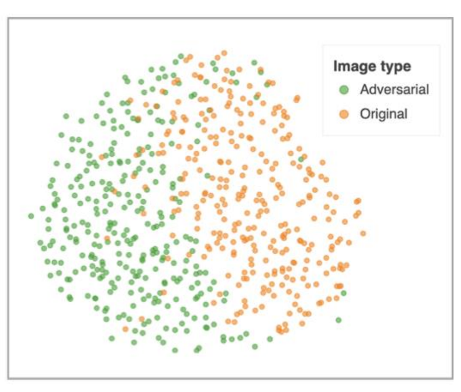
Self-supervision increases attacks transferability among ViTs and CNNs

Evaluated on (columns); *Crafted on* (rows)

	ViT-S (D)	ViT-B (D)	ViT-B (S)	ResNet-50
ViT-S (D)	0.0%	12.5%	41.1%	51.3%
ViT-B (D)	5.2%	0.0%	31.4%	48.5%
ViT-B (S)	47.1%	47.8%	0.8%	59.7%
ResNet-50	65.2%	68.4%	74.9%	0.1%

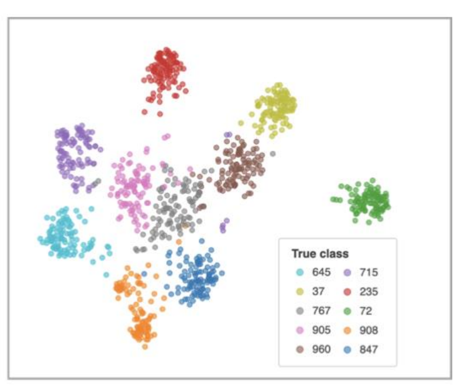
Classification accuracy of adv. samples transferred across architectures. All attacks were crafted using PGD ($\epsilon = 0.03$). Rows represent generation setups and columns, the network used for evaluation. Computed on all validation images from ImageNet-1k. (S) and (D) indicate Supervised and DINO training respectively.

4 Latent Space Analysis



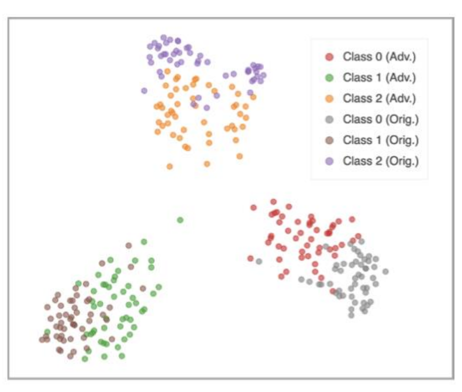
(a) Random images and their adversarial perturbations. Color shows whether a samples is adversarial or original.

Adversarial inputs are linearly separable in the latent space.



(b) Original and adversarial images from 10 random classes. Color represents true label.

They stay close to original samples from their true class.



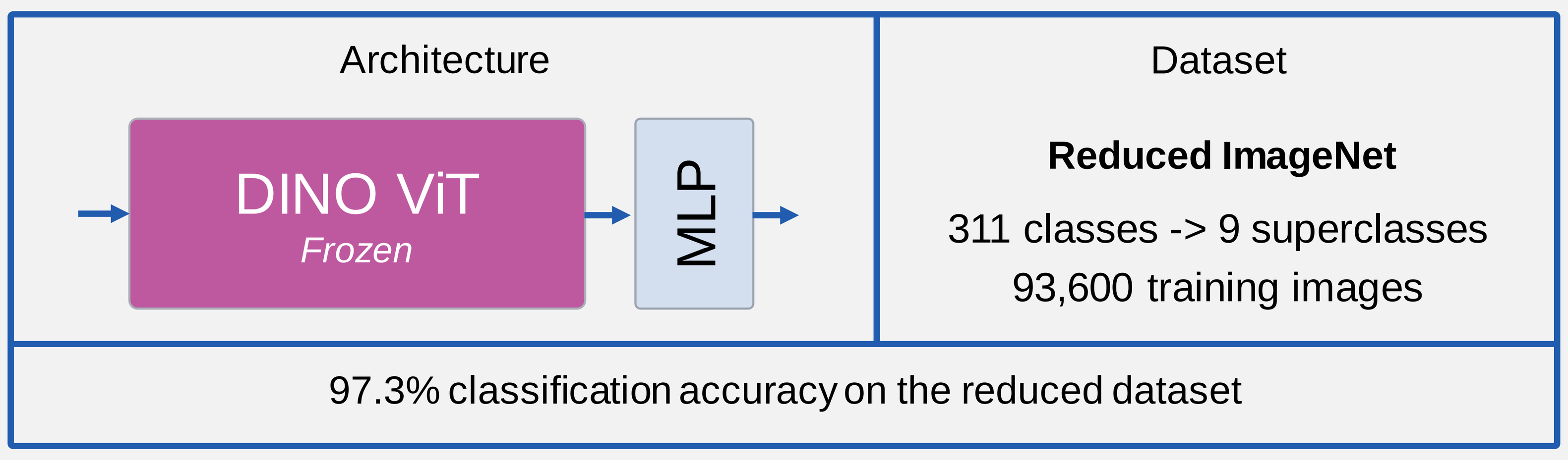
(c) Adversarial and original images from 3 random classes. Color indicates true class, and if it is an attack.

However, they remain separable within classes.

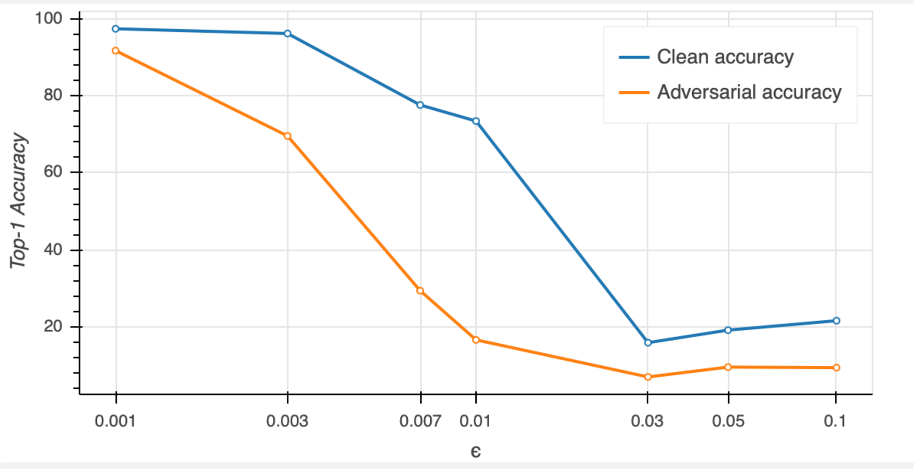
The latent space may comprise enough information to linearly separate adversarial samples without retraining the ViT.

5 Defenses Under Limited Computational Power

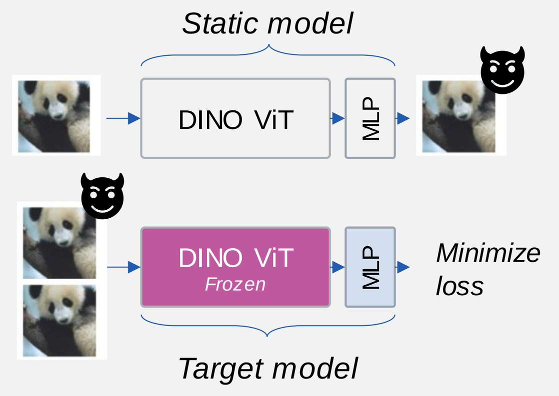
Does it have enough representational power to increase robustness only through fine-tuning?



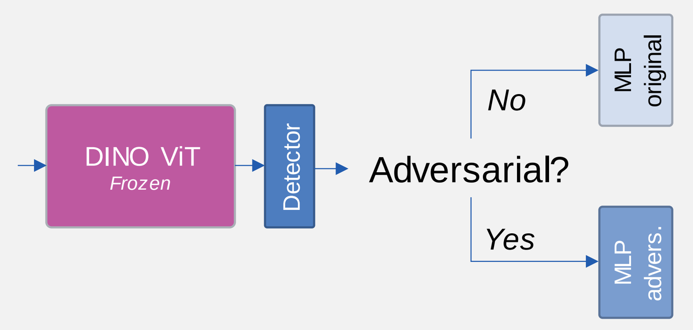
Adversarial Training[2] collapses for large perturbations if the ViT is frozen.



Ensemble Adv. Training[3]



Ensemble Specialized Networks



Strategy	ϵ	Clean	PGD			FGSM			C&W
			$\epsilon = 0.001$	$\epsilon = 0.03$	$\epsilon = 0.1$	$\epsilon = 0.001$	$\epsilon = 0.03$	$\epsilon = 0.1$	$c = 50$
No defense	-	97.3%	89.2%	0.0%	0.0%	90.3%	1.5%	3.1%	0.1%
Ensemble AT	0.001	97.3%	92.6%	0.3%	0.1%	93.2%	13.8%	12.3%	55.4%
Ensemble AT	0.03	93.3%	91.6%	98.7%	97.5%	91.9%	**89.8%**	**76.1%**	**86.6%**
Ensemble AT	0.1	95.3%	91.9%	90.1%	98.7%	91.9%	85.5%	75.4%	80.0%
Specialized Net.	0.001	96.9%	**92.9%**	1.1%	1.2%	**93.3%**	26.6%	20.6%	71.0%
Specialized Net.	0.03	96.9%	88.6%	**99.5%**	99.2%	89.7%	84.3%	70.7%	5.7%
Specialized Net.	0.1	**97.4%**	89.2%	62.6%	**99.6%**	90.3%	72.1%	66.7%	0.4%

Accuracy for Ensemble Adversarial Training (AT) and Ensemble of Specialized Networks compared to a baseline classifier trained on the (reduced) ImageNet dataset.

5 Conclusion

- DINO provides **no additional robustness** against white-box adversarial attacks.
- It **increases attacks transferability** among ViTs and CNNs.
- Several defenses may provide **robustness against black-box attacks under limited computational power** only through fine-tuning.

References

1. Caron, M., et al. *Emerging properties in self-supervised vision transformers.* 2021.
2. Madry, A., et al. *Towards deep learning models resistant to adversarial attacks.* 2017.
3. Tramèr, F., et al. *Ensemble adversarial training: Attacks and defenses.* 2017.