Exploring Adversarial Attacks and Defenses in Vision Transformers trained with DINO

Javier Rando [* 1]   **Nasib Naimi** [* 1]   **Thomas Baumann** [* 1]   **Max Mathys** [1]

Abstract

This work conducts the first analysis on the robustness against adversarial attacks on self-supervised Vision Transformers trained using DINO. First, we evaluate whether features learned through self-supervision are more robust to adversarial attacks than those emerging from supervised learning. Then, we present properties arising for attacks in the latent space. Finally, we evaluate whether three well-known defense strategies can increase adversarial robustness in downstream tasks by only fine-tuning the classification head to provide robustness even in view of limited compute resources. These defense strategies are: Adversarial Training, Ensemble Adversarial Training and Ensemble of Specialized Networks.

1. Introduction

Adversarial attacks (Szegedy et al., 2013) are one of the major challenges faced by current machine learning models. These attacks perturb inputs to provoke incorrect predictions while preserving input similarity, even on state-of-the-art architectures (Goodfellow et al., 2014). They are especially dangerous in safety-critical applications. Effectively defending against them is an open problem that seems to be in odds with the accuracy of the models (Roth et al., 2019).

Meanwhile, advances in Computer Vision result in new architectures that are able to increase accuracy under existing benchmarks. Recently, Vision Transformers (ViTs) (Dosovitskiy et al., 2020) were presented as a successful application to vision of the foundational Transformer (Vaswani et al., 2017) used extensively in Natural Language Processing (NLP) tasks. However, this first ViT architecture re-

quired labelled data for training, imposing an important bottleneck for scalability. Caron et al. introduced a successful self-supervised training scheme, DINO (Caron et al., 2021). The authors claim that self-attention in models trained using DINO aligns with human interpretable features.

This motivates our evaluation of whether the representation learnt through DINO yields increased robustness against adversarial attacks that preserve input similarity. This work analyzes for the first time the effect that adversarial attacks have on self-supervised Vision Transformers trained using DINO. We first show that these ViTs are vulnerable in light of different adversarial attacks. Then, we conduct an in-depth analysis of how the attacks perturb the latent space generated by the Transformer encoder to provoke misclassification. Using these insights, we discuss potential efficient defense strategies that do not require retraining the Transformer backbone in scenarios where compute resources are limited. Overall, our contributions are as follow[1]:

- First evaluation of self-supervised DINO ViTs against different adversarial attacks, and analysis of properties emerging in their latent space. We also show that self-supervision increases attack transferability between ViTs and convolutional networks.

- Drawing on the mentioned properties, we motivate and empirically validate the suitability of the latent space of ViTs for precise detection and classification of adversarial inputs.

- Comparison of three defense strategies under the assumption of limited compute resources: Adversarial Training, Ensemble Adversarial Training and Ensemble of Specialized Networks.

2. Related Work

2.1. Vision Transformers

Transformers (Vaswani et al., 2017) were first presented in the field of Natural Language Processing (NLP) as a novel architecture relying entirely on self-attention mechanisms. They can be efficiently pre-trained on unlabeled

---

[*]Equal contribution [1]ETH Zurich, Switzerland. Correspondence to: Javier Rando <jrando@student.ethz.ch>, Nasib Naimi <nnaimi@student.ethz.ch>, Thomas Baumann <thobauma@student.ethz.ch>.

---

[1]Our code can be accessed on GitHub: https://github.com/thobauma/AADefDINO

data allowing for extracting knowledge from huge datasets without human annotations, e.g. web corpora. The resulting knowledge is then leveraged to solve specific downstream tasks by means of supervised fine-tuning (Brown et al., 2020). The same architecture was then transferred to Computer Vision. In a first attempt, supervised learning was presented as the only way to build so called Vision Transformers (Dosovitskiy et al., 2020). Although they outperformed state-of-the-art architectures in most vision tasks (Khan et al., 2021), they required huge amounts of labeled data and computational power. Later, a successful self-supervised training framework, DINO, was presented to avoid the need for labeled data and exploit the advantages that turned Transformers into the leading architecture for NLP (Caron et al., 2021). Moreover, experimental results show that self-attention of Transformers trained using DINO contain explicit information about semantic segmentation of an image and align with human interpretable features. After DINO, other self-supervised training frameworks have been proposed such as Masked Autoencoders (He et al., 2021) and BEiT (Bao et al., 2021).

Different architectures have been proposed for Vision Transformers by tuning the number of parameters, or by creating "hybrid" models combining Transformers and CNNs (Dosovitskiy et al., 2020). In this work, we will only consider ViTs relying completely on self-attention, and follow the nomenclature provided by Caron et al., i.e. ViT-S/16 refers to the "small" (S) network size (21M parameters) with input patches of size 16x16 (Caron et al., 2021).

2.2. Adversarial Attacks

Adversarial attacks are artificially crafted inputs designed to fool machine learning models (Szegedy et al., 2013). Two main types of attacks can be found in the literature: white-box and black-box attacks. In white-box attacks, the attacker has full access to the model including its parameters and loss. The adversary may maximize the loss with respect to the target class for each input to achieve misclassification. Some attacks within this category are FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2017) and C&W (Carlini & Wagner, 2017b). On the other hand, under black-box attacks, the adversary has no information about the model but can only observe inputs and corresponding predictions (Papernot et al., 2017; Chen et al., 2017).

Previous work has shown supervised ViTs to be more robust than traditional convolutional classifiers against adversarial attacks (Aldahdooh et al., 2021; Shao et al., 2021). Moreover, attacks crafted on supervised ViTs exhibit worse transferability across architectures (Naseer et al., 2022).

Self-supervised ViTs trained with DINO have only been evaluated on transfer attacks for segmentation tasks (Naseer et al., 2022). However, their robustness have not been as-

sessed for classification tasks nor compared with different architectures.

2.2.1. DEFENSES AGAINST ADVERSARIAL ATTACKS

Numerous defense strategies have been explored for different attacks and architectures in Computer Vision (Ren et al., 2020). We group them into four main categories:

Adversarial training (Szegedy et al., 2013; Madry et al., 2017). When training the network, optimization becomes a $min\text{-}max$ problem. For each input x, we find the perturbation x' that maximizes the loss in the l_p ball of radius ϵ around x. Then, we try to find the optimal parameters (θ^*) by minimizing the loss for x' (see Equation 1). In practice, the search for x' can be approximated efficiently with PGD (Madry et al., 2017). Adversarial training increases resistance to attacks at the cost of accuracy on original samples and expensive optimization (Roth et al., 2019). Subsequent work by Tramer et al. shows that this form of adversarial training converges to a degenerate global minimum. They propose *Ensemble Adversarial Training* as an effective alternative that augments training data with attacks generated on different networks (Tramèr et al., 2017).

$$\theta^* = \arg \min_{\theta} \max_{||r||_p \leq \epsilon} loss(f_\theta(x + r), y) \qquad (1)$$

Preprocessing. Preprocessing input images may remove perturbations induced by attackers and lead to more robust predictions (Guo et al., 2017; Buckman et al., 2018; Yang et al., 2019). JPEG compression has been shown to be an efficient defense, but it is not sufficient as the strength of the attack increases (Dziugaite et al., 2016; Das et al., 2017).

Post-hoc detectors. Detect whether an input sample was adversarial (Feinman et al., 2017). For example, using output logits (Roth et al., 2019; Mosca et al., 2022) or SHAP values (Fidel et al., 2020). These systems provide an alert if an attack is detected, but generally do not find the true class.

Ensemble models. Combine multiple models for a given task. It is motivated by the fact that an attack may fool one architecture but not every model in the ensemble. Abbasi et al.(Abbasi & Gagné, 2017) makes use of specialist classifiers trained on the classes most prone to misclassification.

However, if an attacker gets to know a defense, *adaptitve attacks* may be crafted to surpass them (Carlini & Wagner, 2017a; Tramèr et al., 2017).

3. Adversarial Attacks on Self-Supervised ViT

As motivated previously, the goal of this section is to evaluate whether features learnt through DINO are more robust against adversarial perturbations that preserve input similarity. Although adversarial accuracy of supervised Vision Transformers has already been assessed in previous works
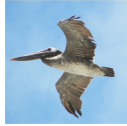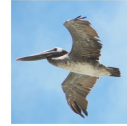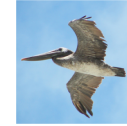
Clean	FGSM			PGD			C&W
	$\epsilon = 0.001$	$\epsilon = 0.03$	$\epsilon = 0.1$	$\epsilon = 0.001$	$\epsilon = 0.03$	$\epsilon = 0.1$	$c = 50$
Pelican	Pelican	Kite	Kite	Pelican	Kite	Kite	Albatross

Table 1. Adversarial perturbations generated on a validation sample using each of the attacks reported in Table 2 on ViT-S/16 trained with DINO. True label is "Pelican". Last row indicates predicted class after the attack. Perturbations include important artifacts as ϵ grows.

		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%

Table 2. 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. ResNet-50 is included as a benchmark for convolutional architectures.

(Aldahdooh et al., 2021; Mahmood et al., 2021), there are no existing benchmarks for self-supervised ViTs trained using DINO. Moreover, a comparison between these two training schemes has not been done yet. In view of fairness, we produce results using the same Transformer architecture. We consider the supervised ViT-B/16 HuggingFace implementation (Wolf et al., 2019), and its equivalent self-supervised ViT-B/16 taken from the DINO official repository[2]. Both have 85M parameters and are fine-tuned for classification on ImageNet-1k (Russakovsky et al., 2015). We also include a smaller self-supervised ViT-S/16 (21M parameters) to measure robustness across different architecture sizes, and ResNet-50 (He et al., 2016) as benchmark for CNNs.

We evaluate all four models on the same random split containing 3000 samples from ImageNet validation set. Their corresponding adversarial images are crafted using white-box attacks. We test against FGSM (L_∞), PGD (L_∞) and C&W (L_2) using the implementation provided by (Kim, 2020). Attack parameters are chosen through visual exploration (see Table 1). $\epsilon = 0.001$ is a subtle attack that does not introduce artifacts in the images. $\epsilon = 0.03$ yields small artifacts, and finally, $\epsilon = 0.1$ as an upperbound with pronounced perturbations. Accuracy on each of the attacks is reported in Table 2. Results show that ViTs trained using DINO do not exhibit a significant difference in terms of robustness to these white-box attacks compared to their supervised counterparts.

[2]https://github.com/facebookresearch/dino

	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%

Table 3. Classification accuracy of adversarial 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. Patch size (16) has been omitted for clarity. (S) and (D) indicate Supervised and DINO training respectively.

3.1. Adversarial Attacks Transferability

Another well-known property from adversarial attacks is their transferability across architectures (Goodfellow et al., 2014; Szegedy et al., 2013; Demontis et al., 2019). We compare how attacks generated on supervised and self-supervised ViTs, and ResNet-50 transfer among them. Our results in Table 3 show that DINO ViTs are equally vulnerable to such adversaries, especially when another DINO ViT is used as source. Overall, ViTs are more robust against attacks generated on ResNet-50 than against those from transformers.

Our results reproduce the limited transferability between supervised ViTs and convolutional models presented by (Naseer et al., 2022). However, they also reveal that self-supervision increases transferability in both directions.
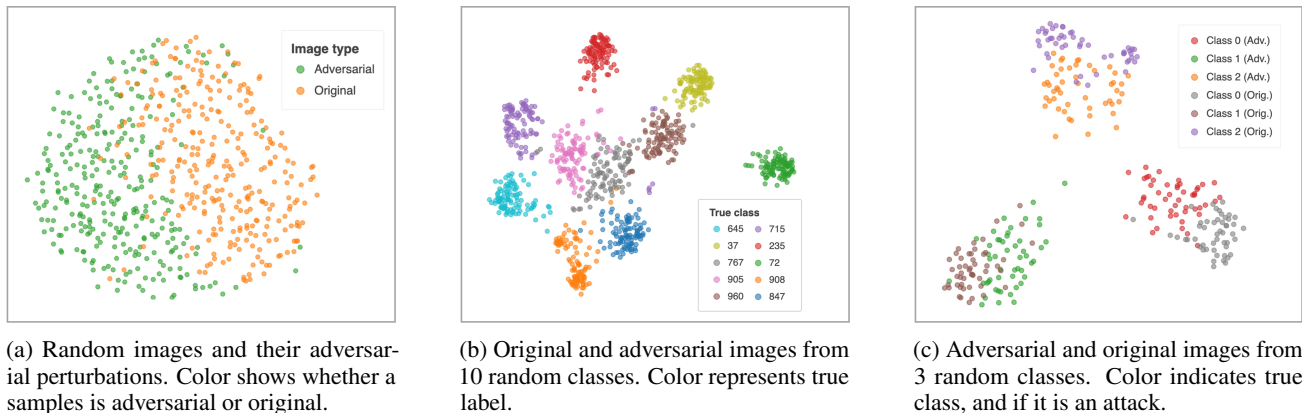
(a) Random images and their adversarial perturbations. Color shows whether a samples is adversarial or original.

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

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

Figure 1. UMAP visualization (*canberra* distance) of the output space generated by ViT-S/16 on ImageNet validation images and their adversarial perturbations generated using PGD (L_∞ and $\epsilon = 0.03$).

3.2. Latent Space for Adversarial Attacks

This section motivates how attacks affect DINO ViTs by analyzing the latent representation for adversarial inputs which would then be used for classification.

For the experiments, we use a ViT-S/16 trained using DINO. The latent space for this model, following its original implementation, is constructed by concatenating the `[CLS]` embedding from the last 4 layers of the ViT encoder. This results in a 1536 dimensional representation for each input image. The adversarial images are crafted using PGD with L_∞ and $\epsilon = 0.03$. As shown in Table 2, this attack has 100% success rate in the selected architecture, i.e. all adversarial images are incorrectly classified by the model.

We use UMAP (Sainburg et al., 2021) and different random samples of ImageNet validation set to interpret this high-dimensional space (see Figure 1). We identify three relevant properties. (i) Adversarial inputs can be separated from original images in this space (see Figure 1a). (ii) Adversarial images stay close to original samples in their true class, i.e., the class predicted before perturbations (see Figure 1b). (iii) Although adversarial inputs are close to original, they remain separable within the clusters (see Figure 1c).

These properties show that the latent space may comprise enough information to linearly separate adversarial samples without retraining the ViT. This is crucial for setups in which computational resources are limited and will motivate our *Ensemble of Specialized Networks*.

4. Adversarial Defense Strategies

We present three defense strategies for a *target model* using a self-supervised DINO ViT as backbone, and optimized for classification on ImageNet (Deng et al., 2009). Since training on the whole ImageNet dataset is known to be a computationally expensive problem, we limit ourselves to a restricted subset used in previous works (Tsipras et al., 2018; Engstrom et al., 2019b). It comprises semantically similar images into 9 super-classes (see Table 6). The reduced dataset is then balanced so that all super-classes contain the same number of samples, resulting in 93,600 training and 3,600 validation images.

All things considered, we create our *target model* by training a linear classification layer on top of a ViT-S/16 encoder using the 9-class reduced dataset. It obtains 97.3% classification accuracy in this task. We will use this architecture and reduced dataset throughout the remaining sections.

4.1. Adversarial Training

Assuming adversarial training on the whole Transformer architecture is a very costly process, our goal is to test whether it is possible to perform adversarial training just on the classification head while keeping the ViT frozen to reduce complexity.

In the foundational paper for adversarial training, it was already pointed out that capacity in the network is very important to achieve high accuracy (Madry et al., 2017). We found out that a linear classification head didn't generalize properly, so we consider a fully connected neural network with 3 hidden layers (2048, 1024, 512) and ReLU activation.

We replicate adversarial training as presented by Madry et al. by optimizing Equation 1. Maximization of the loss is approximated with PGD since it has been shown to generalize best (Madry et al., 2017). The attack is always crafted to fool the latest and most robust version of the classifier during training. Figure 2 plots top-1 accuracy on clean and adversarial samples after training the classification head for 5 epochs. Adversarial Training can be achieved by only fine-tuning the classification head for low values of ϵ, but collapses as it increases.
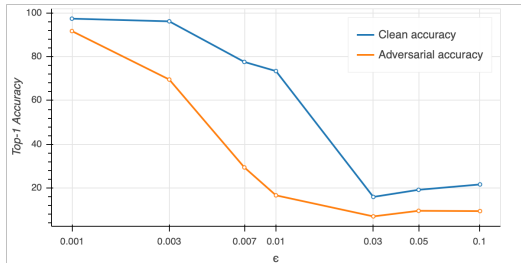
Figure 2. Linear-log plot for top-1 accuracy on validation set after performing PGD adversarial training for different values of ϵ.

and attacks can improve generalization.

(a) Original and adversarial images colored according to their true class.

(b) Original images colored according to their true class and adversarial in grey.

Figure 3. UMAP visualization (*canberra* distance) of training images and their adversarial perturbation using PGD (L_∞ and $\epsilon = 0.03$) during adversarial training. Figure (a) represents all images according to their true class and Figure (b) shows which of those points correspond to adversarial images.
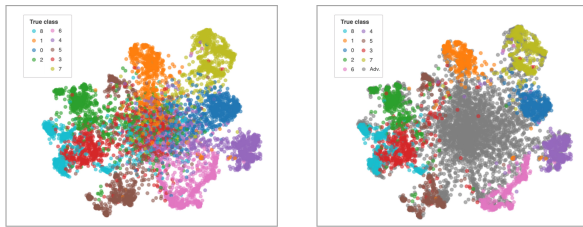
Finally, we motivate this collapse. Figure 3 depicts the latent representation generated by the ViT for adversarial (PGD with $\epsilon = 0.03$) and clean samples. This space is then used as input for classification. It shows how the attack was able to generate samples that are mapped to a region which is no longer separable by the classifier. This is most likely due to the limitations that freezing the Transformer imposes on our representational power.

4.2. Ensemble Adversarial Training

As Tramer et al. showed in previous work, vanilla adversarial training may converge to a degenerate global minimum. Thus, they propose Ensemble Adversarial Training which consists in augmenting training data with perturbations transferred from a different network to decouple adversarial generation (Tramèr et al., 2017). In our setup, we consider attacks crafted on a surrogate of the *target model* solving the same task. Thus, our robust classification layer will be optimized on an augmented dataset containing both clean and adversarial images generated on a static surrogate; i.e. the model generating adversarial samples does not change as training progresses. The main difference with vanilla adversarial training is that attacks are no longer crafted to fool the last iteration of the trained classifier. For each batch, we uniformly select between clean images or their adversarial counterparts. We use PGD to generate the augmented data using the surrogate.

Table 4 (rows 2-4) reports the performance obtained by the resulting classifiers on attacks crafted on the *target model* and validation images. Overall, the model trained on data augmented with PGD and $\epsilon = 0.03$ exhibits the most robust generalization to attackers of different natures and strengths, with an average accuracy of 87.3%. This comes at the cost of only 4% accuracy on clean samples. Then, we also test whether these defenses are robust against attacks crafted on a larger ViT/B-16. Whereas the *target model* achieved 7.7% on such attack, Ensemble Adversarial Training results in a model which correctly classifies 66% of the adversarial samples (see rows 1-4 in Table 5). We leave as future work how combining perturbations coming from varied networks

4.3. Ensemble of Specialized Networks

Drawing from Section 3.2, we motivate a potential defense that leverages the separability of adversarial inputs in the latent space generated by self-supervised DINO Vision Transformers. As a first step, we build a *post-hoc detector* that determines whether a given sample was perturbed by an adversary or not. As input, it takes the image latent representation generated by the ViT encoder, i.e. the same input as the classification head. We use a linear layer that maps this 1536 dimensional vector into 2 classes: adversarial and original. Then, we train an *adversarial classifier* specialized in correctly labelling perturbed inputs.

These two steps can be achieved with high accuracy, validating the property of adversarial inputs being separable in the latent space introduced in Section 3. Although we just report end-to-end performance of the ensemble, detailed intermediate results can be found in Appendix C.

Finally, since both clean and adversarial inputs can be precisely classified by their respective *specialized classification heads*, we ensemble them by means of the post-hoc detector to build a robust end-to-end model. At inference time, we use the adversarial classifier whenever the post-hoc detector outputs more than 50% probability of a sample being adversarial, and the original classifier trained on (reduced) ImageNet otherwise. We evaluate the performance of the ensemble pipeline on adversarial samples generated with various attacks on the *target model* (ViT/S-16), as well as on ViT-B/16 to test against transfer attacks. The last three rows in Table 4 report the performance of ensembles using a post-hoc detector and adversarial classifier trained on the same attack. However, an extensive analysis of all possible combinations is included in the Appendix (see Table 9). Accuracy of over 70% is obtained for all attacks generated using PGD and FGSM, but drops for C&W when compared

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%

Table 4. Accuracy for Ensemble Adversarial Training (AT) and Ensemble of Specialized Networks compared to a baseline classifier trained on the (reduced) ImageNet dataset. All defenses are built using PGD with the reported values for ϵ. Columns indicate evaluation black-box attacks crafted on the *target model* without defense.

Strategy	ϵ	Accuracy
No defense	-	7.7%
Ensemble AT	0.001	13.7%
Ensemble AT	0.03	**66.0%**
Ensemble AT	0.1	51.6%
Specialized Net.	0.001	25.6%
Specialized Net.	0.03	43.3%
Specialized Net.	0.1	18.1%

Table 5. Accuracy for Ensemble Adversarial Training (AT) and Ensemble of Specialized Networks trained on different values of ϵ against transfer attacks crafted on DINO ViT/B-16 using PGD ($\epsilon = 0.03$).

with Ensemble Adversarial Training. This accuracy across the attacks comes at a minimal cost on the accuracy for the unperturbed samples (decreasing at most 0.4%). Lastly, although the ensemble improves on the clean classifier for samples perturbed on ViT/B-16, its performance falls behind Ensemble Adversarial Training (see last 3 rows in Table 5)

5. Conclusion

We have shown that features emerging from self-supervision in Vision Transformers using DINO do not bear a significant difference in terms of robustness to adversarial attacks compared to their equivalent supervised architectures and CNNs. Our results also reveal that self-supervision increases transferability of attacks between ViTs and CNNs. Furthermore, our analysis of the latent space generated by the Transformer provides the first insights into how such features are perturbed by adversarial attacks.

Building on these insights, we evaluate the suitability of the latent space for different defense strategies that do not require retraining the ViT backbone. This is especially relevant for scenarios with limited computational resources.

First, we showed that vanilla Adversarial Training can be successful for small perturbations, but may require unfreezing of the Transformer with increasing strength of the attack. Then, we used Ensemble Adversarial Training to augment training data with perturbations transferred from a surrogate. Finally, we built an ensemble network in which a post-hoc detector determines if an input is adversarial and according to its prediction, a specialized head for clean or adversarial inputs is used for classification.

Of these strategies, we report that Ensemble Adversarial Training results in the most robust behaviour across different attackers and at a small cost on accuracy for clean samples. It is able to increase average accuracy against adversaries from 24.0% to 87.3% at the cost of only 4% accuracy on clean samples.

6. Limitations & Future Work

Limited Dataset. Explore how the performance of such defenses generalize to the entire ImageNet.

Adversarial training. It remains unexplored if unfreezing some layers in the encoder could lead to successful AT for larger perturbations.

Ensemble Adversarial Training. Analyze how to optimally augment training data to increase generalization.

Further self-supervision strategies. Assess if these properties and defenses generalize to ViTs trained using other self-supervision strategies.

Adaptive attacks. Explore if our defenses can be surpassed by crafting a specific adaptive attack.

Latent space. We build the latent space as presented in the original DINO work. Future work may explore the impact of incorporating information from more layers.

Acknowledgements

References

Abbasi, M. and Gagné, C. Robustness to adversarial examples through an ensemble of specialists, 2017.

Aldahdooh, A., Hamidouche, W., and Deforges, O. Reveal of vision transformers robustness against adversarial attacks, 2021.

Bao, H., Dong, L., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018.

Carlini, N. and Wagner, D. *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*, pp. 3–14. 2017a. ISBN 9781450352024.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017b. doi: 10.1109/SP.2017.49.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. *ZOO: Zeroth Order Optimization Based Black-Box Attacks to Deep Neural Networks without Training Substitute Models*, pp. 15–26. 2017. ISBN 9781450352024.

Das, N., Shanbhogue, M., Chen, S.-T., Hohman, F., Chen, L., Kounavis, M. E., and Chau, D. H. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., and Roli, F. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX security symposium (USENIX security 19)*, pp. 321–338, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dziugaite, G. K., Ghahramani, Z., and Roy, D. M. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., and Tsipras, D. Robustness (python library), 2019a.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., and Madry, A. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019b.

Feinman, R., Curtin, R. R., Shintre, S., and Gardner, A. B. Detecting adversarial samples from artifacts, 2017.

Fidel, G., Bitton, R., and Shabtai, A. When explainability meets adversarial learning: Detecting adversarial examples using shap signatures. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. 2014.

Guo, C., Rana, M., Cisse, M., and Van Der Maaten, L. Countering adversarial images using input transformations. 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. Transformers in vision: A survey, 2021.

Kim, H. Torchattacks: A pytorch repository for adversarial attacks. 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. 2017.

Mahmood, K., Mahmood, R., and van Dijk, M. On the robustness of vision transformers to adversarial examples, 2021.

Mosca, E., Agarwal, S., Rando, J., and Groh, G. "That is a suspicious reaction!": Interpreting logits variation to detect NLP adversarial attacks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7806–7816. Association for Computational Linguistics, May 2022.

Naseer, M., Ranasinghe, K., Khan, S., Khan, F., and Porikli, F. On improving adversarial transferability of vision transformers. In *International Conference on Learning Representations*, 2022.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning, 2017.

Ren, K., Zheng, T., Qin, Z., and Liu, X. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020. ISSN 2095-8099. doi: https://doi.org/10.1016/j.eng.2019.12.012.

Roth, K., Kilcher, Y., and Hofmann, T. The odds are odd: A statistical test for detecting adversarial examples. In *International Conference on Machine Learning*, pp. 5498–5507. PMLR, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Sainburg, T., McInnes, L., and Gentner, T. Q. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907, 2021.

Shao, R., Shi, Z., Yi, J., Chen, P.-Y., and Hsieh, C.-J. On the adversarial robustness of vision transformers, 2021.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks, 2013.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Yang, Y., Zhang, G., Katabi, D., and Xu, Z. Me-net: Towards effective adversarial robustness with matrix estimation, 2019.

A. Setup

All the experiments presented in this paper have been performed using the Euler cluster at ETH Zürich. More specifically, our working environment used 4 CPU cores with 10240MB of RAM each, and 1 GPU (GeForceRTX2080Ti).

B. Dataset

For our experiments we limit ourselves to a restricted dataset considered in previous works (Tsipras et al., 2018; Engstrom et al., 2019b). It comprises 311 semantically similar classes into 9 super-classes. The reduced dataset is then balanced so that all super-classes contain the same number of images. This results in 93,600 training images and 3600 for validation. The following table displays each of the resulting super-classes and their child classes in the original ImageNet Dataset. Library by Engstrom et al. was used to create the reduced dataset (Engstrom et al., 2019a).

Super-class	ImageNet classes (inclusive)
Dog	151 to 268
Cat	281 to 285
Frog	30 to 32
Turtle	33 to 37
Bird	80 to 100
Primate	365 to 382
Fish	389 to 397
Crab	118 to 121
Insect	300 to 319

Table 6. Classes from ImageNet-1k used to generate our restricted dataset.

C. Ensemble detailed metrics

	PGD			FGSM			C&W
	$\epsilon = 0.001$	$\epsilon = 0.03$	$\epsilon = 0.1$	$\epsilon = 0.001$	$\epsilon = 0.03$	$\epsilon = 0.1$	$c = 50$
$\epsilon = 0.001$	58.2%	73.375%	74.9%	59.0%	75.708%	75.8%	72.5%
$\epsilon = 0.03$	50.2%	99.5%	99.6%	50.2%	99.6%	99.7%	56.5%
$\epsilon = 0.1$	50.0%	81.4%	99.9%	50.0%	96.3%	99.9%	50.3%

Table 7. Post-hoc detection accuracy. Rows represent ϵ used in PGD for training. The columns show the different datasets used for evaluation.

	$\epsilon = 0.001$	$\epsilon = 0.03$	$\epsilon = 0.1$	**Clean**
$\epsilon = 0.001$	**93.58**%	1.08%	1.22%	96.78%
$\epsilon = 0.03$	2.22%	**99.86**%	99.25%	0.61%
$\epsilon = 0.1$	1.75%	99.50%	**99.64%**	0.31%
Clean	89.17%	0.00%	0.08%	**97.53%**

Table 8. Accuracy for different linear classifiers trained on adversarial datasets. All adversarial datasets were generated using PGD, where the ϵ was varied. Rows show the data used for training the classifier and columns represent the datasets used for evaluation.

Post-Hoc Classifier	ImageNet Classifier	Clean	PGD $\epsilon = 0.001$	PGD $\epsilon = 0.03$	PGD $\epsilon = 0.1$	FGSM $\epsilon = 0.001$	FGSM $\epsilon = 0.03$	FGSM $\epsilon = 0.1$	C&W 50
$\epsilon = 0.001$	$\epsilon = 0.001$	96.89%	92.94%	1.08%	1.22%	93.33%	26.58%	20.61%	71.03%
$\epsilon = 0.001$	$\epsilon = 0.03$	51.81%	35.58%	94.89%	97.39%	34.11%	84.08%	70.67%	29.36%
$\epsilon = 0.001$	$\epsilon = 0.1$	51.64%	35.08%	94.56%	97.78%	33.67%	76.83%	66.67%	16.89%
$\epsilon = 0.03$	$\epsilon = 0.001$	97.58%	89.36%	1.08%	1.22%	90.47%	26.58%	20.61%	9.91%
$\epsilon = 0.03$	$\epsilon = 0.03$	96.92%	88.58%	99.50%	99.19%	89.67%	84.28%	70.67%	5.67%
$\epsilon = 0.03$	$\epsilon = 0.1$	96.92%	88.53%	99.14%	99.58%	89.64%	76.94%	66.67%	3.83%
$\epsilon = 0.1$	$\epsilon = 0.001$	97.53%	89.22%	0.89%	1.22%	90.36%	23.72%	20.61%	0.64%
$\epsilon = 0.1$	$\epsilon = 0.03$	97.42%	89.17%	62.83%	99.22%	90.31%	78.33%	70.67%	0.50%
$\epsilon = 0.1$	$\epsilon = 0.1$	97.42%	89.17%	62.58%	99.61%	90.31%	72.14%	66.67%	0.42 %

Table 9. Accuracy for all possible combinations of post-hoc classifiers and adversarial classification heads to build an Ensemble. They are evaluated on all available adversarial attacks.

Post-Hoc Classifier	ImageNet Classifier	ViT/B-16
$\epsilon = 0.001$	$\epsilon = 0.001$	25.61%
$\epsilon = 0.001$	$\epsilon = 0.03$	44.08%
$\epsilon = 0.001$	$\epsilon = 0.1$	34.92%
$\epsilon = 0.03$	$\epsilon = 0.001$	24.06%
$\epsilon = 0.03$	$\epsilon = 0.03$	43.28%
$\epsilon = 0.03$	$\epsilon = 0.1$	34.69%
$\epsilon = 0.1$	$\epsilon = 0.001$	11.72%
$\epsilon = 0.1$	$\epsilon = 0.03$	19.42%
$\epsilon = 0.1$	$\epsilon = 0.1$	18.14%

Table 10. Accuracy against attacks crafted on ViT/B-16 using PGD ($\epsilon = 0.03$) for all possible Ensemble models.