Benchmarking Robustness of Self-Supervised Learning Across Diverse Downstream Tasks

Antoni Kowalczuk[1], Jan Dubiński[2,3], Atiyeh Ashari Ghomi[4], Yi Sui[4], George Stein[4], Jiapeng Wu[4], Jesse C. Cresswell[4], Franziska Boenisch[1], Adam Dziedzic[1]

[1]CISPA Helmholtz Center for Information Security, [2]Warsaw University of Technology, [3]IDEAS NCBR, [4]Layer 6 AI

Motivation

- Self-supervised learning (SSL) vision encoders provide unprecedented performance across downstream tasks.
- Their robustness against adversarial examples on tasks beyond image classification is vastly under-explored.
- Current state-of-the-art (SOTA) robust fine-tuning fails to address that threat.

Contributions

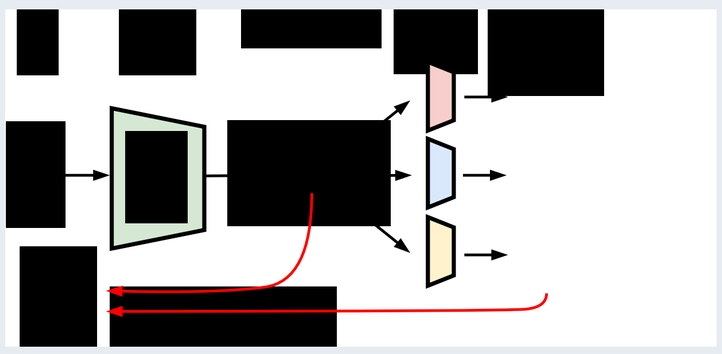
- Benchmarking robustness of SSL Encoders on classification, semantic segmentation, and depth estimation.
- Evaluation of attacks operating in the embedding space.
- Performance analysis of SOTA DeACL robust fine-tuning method against adversarial examples.

Method

We generate adversarial examples using PGD, defined as

$$\delta \leftarrow \delta + \eta \cdot \text{sign}\left(\nabla_{\delta} L\left(f_{\text{task}}(x+\delta), y\right)\right),$$

where δ $[-\ ,\ ]$ is the adversarial perturbation, and L is the optimization objective.



We target downstream tasks using the following attacks:

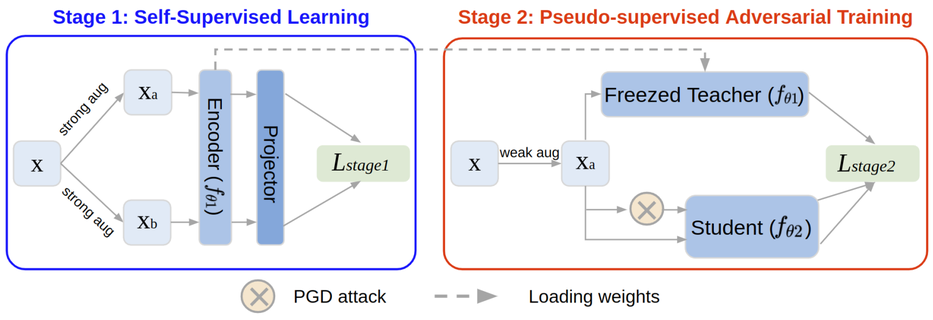
1. *EmbedAttack*: Task-agnostic
$$L = \left\| f_{\text{emb}}(x+\delta) - f_{\text{emb}}(x) \right\|_2^2.$$
2. *PGD*: Classification
$$L = \text{CELoss}(f_{\text{clf}}(x+\delta), y).$$
3. *SegPGD*: Semantic Segmentation
$$L = \text{CELoss}_{\text{pixelwise}}(f_{\text{seg}}(x+\delta), y).$$
4. *DepthPGD*: Depth Estimation
$$L = L_{\text{depth}}(f_{\text{depth}}(x+\delta), y).$$

Robust fine-tuning

We evaluate Decoupled Adversarial Contrastive Learning (DeACL) fine-tuning. The training objective is as follows:

$$L(f_R, f) = CosSim(f_R(x), f(x)) + \gamma CosSim(f_R(x_{\text{adv}}), f_R(x)),$$

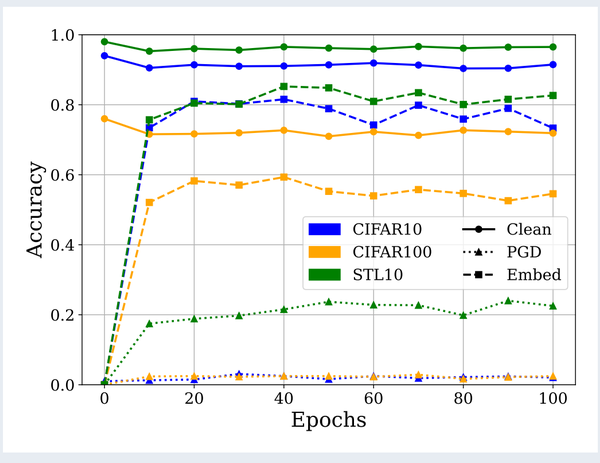
where f_R is a robust version of an encoder f, $\gamma = 2$, and x_{adv} is an adversarial example obtained using *EmbedAttack* with cosine similarity as the optimization objective.



Experimental results

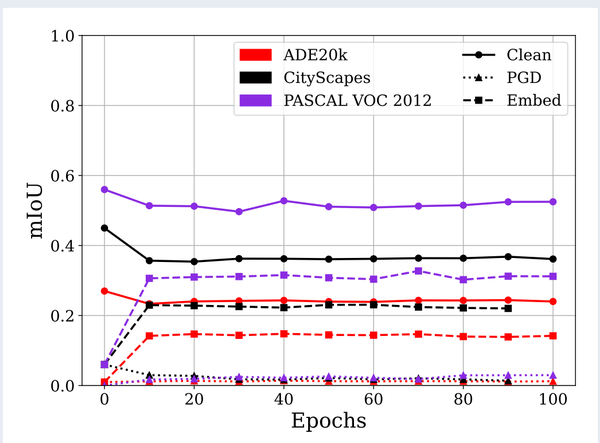
The following are the results on DINO SSL encoders, across various downstream tasks, tested for clean and robust performance under various adversarial attacks. On the left, in the plot, is the robustness of DINO-v1 ViT-B/16, evaluated after multiple epochs of DeACL fine-tuning. On the right, in the table, is the robustness comparison between different DINO versions, across different datasets.

Classification



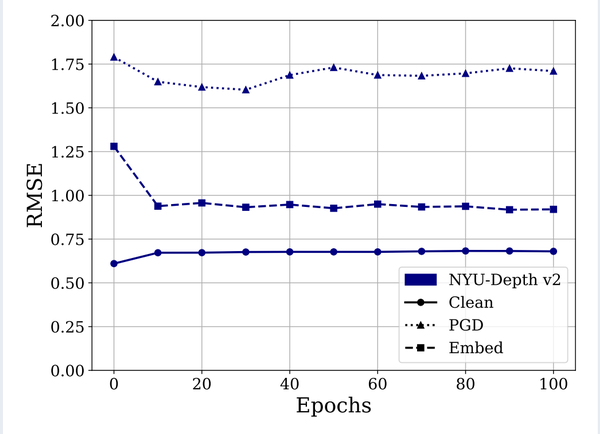
Dataset	SSL Framework	Encoder Type	*Clean* Accuracy↑	*EmbedAttack* Accuracy↑	*PGD* Accuracy↑
CIFAR10	DINO v2 ViT-S/14	Standard	0.94	0.01	0.00
CIFAR10	DINO v2 ViT-B/14	Standard	0.98	0.04	0.00
CIFAR10	DINO v1 ViT-B/16	Standard	0.94	0.01	0.00
CIFAR10	DINO v1 ViT-B/16	DeACL	0.91	0.73	0.02
CIFAR100	DINO v2 ViT-S/14	Standard	0.82	0.00	0.00
CIFAR100	DINO v2 ViT-B/14	Standard	0.86	0.00	0.00
CIFAR100	DINO v1 ViT-B/16	Standard	0.76	0.00	0.00
CIFAR100	DINO v1 ViT-B/16	DeACL	0.72	0.55	0.03
STL10	DINO v2 ViT-S/14	Standard	0.98	0.06	0.00
STL10	DINO v2 ViT-B/14	Standard	0.99	0.20	0.00
STL10	DINO v1 ViT-B/16	Standard	0.98	0.00	0.00
STL10	DINO v1 ViT-B/16	DeACL	0.97	0.83	0.23

Semantic Segmentation



Dataset	SSL Framework	Encoder Type	*Clean* mIoU↑	*EmbedAttack* mIoU↑	*SegPGD* mIoU↑
ADE20k	DINOv2 ViT-S/14	Standard	0.42	0.01	0.01
ADE20k	DINOv2 ViT-B/14	Standard	0.45	0.00	0.01
ADE20k	DINOv1 ViT-B/16	Standard	0.27	0.01	0.01
ADE20k	DINOv1 ViT-B/16	DeACL	0.24	0.14	0.01
CityScapes	DINOv2 ViT-S/14	Standard	0.65	0.02	0.01
CityScapes	DINOv2 ViT-B/14	Standard	0.68	0.03	0.00
CityScapes	DINOv1 ViT-B/16	Standard	0.45	0.06	0.06
CityScapes	DINOv1 ViT-B/16	DeACL	0.36	0.31	0.03
PASCAL VOC 2012	DINOv2 ViT-S/14	Standard	0.83	0.00	0.01
PASCAL VOC 2012	DINOv2 ViT-B/14	Standard	0.83	0.00	0.01
PASCAL VOC 2012	DINOv1 ViT-B/16	Standard	0.56	0.06	0.00
PASCAL VOC 2012	DINOv1 ViT-B/16	DeACL	0.51	0.30	0.02

Depth Estimation



SSL Framework	Encoder Type	*Clean* RMSE↓	*EmbedAttack* RMSE↓	*DepthPGD* RMSE↓
DINO v2 ViT-S/14	Standard	0.49	1.54	2.60
DINO v2 ViT-B/14	Standard	0.46	1.29	2.74
DINO v1 ViT-B/16	Standard	0.61	1.28	1.79
DINO v1 ViT-B/16	DeACL	0.68	0.92	1.71

For DeACL, we observe a lack of improvement of robustness under more potent, downstream *PGD* attacks. We note that our *EmbedAttack* performs on-par with downstream attacks in the clean setting, except for the Depth Estimation task.