Adversarial Pixel Restoration as Pretext Task for Transferable Perturbations

Hashmat Shadab Malik, Shahina K Kunhimon, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan

{hashmat.malik,shahina.kunhimon,muzammal.naseer,salman.khan,fahad.khan}@mbzuai.ac.ae

Paper-ID: 546

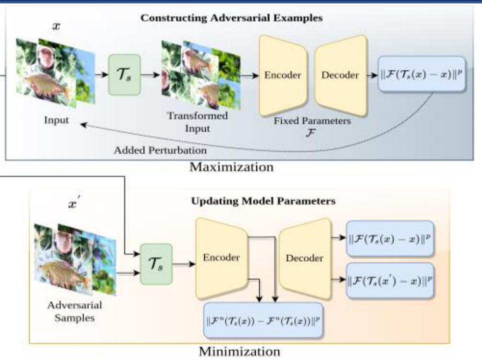
https://github.com/HashmatShadab/APR

Introduction

In the black-box setting, adversarial examples are typically created using surrogate models trained on the target model's large labeled distribution. Such attacks are thus limited by the availability of a pretrained surrogate model and label space information. Our work focuses on a stronger threat model on how to learn an effective surrogate model from the limited **unlabelled** data and then how to generate self-supervised transferable adversarial examples.

With limited samples, training surrogate models in a supervised fashion (conventional) causes severe overfitting, leading to poor adversarial transferability. We propose a Self-Supervised Adversarial Training method to find highly transferable patterns by learning over flatter loss surfaces

Self-supervised Adversarial Training



We train an autoencoder-based surrogate model via self-supervised adversarial pixel restoration to learn generalizable representations from limited data samples (≤ 20) or cross-domain samples.

In the maximization step, adversarial examples are generated by fooling model's reconstruction ability; in the minimization step, model parameters are updated using restoration objectives between adversarial and clean samples.

Self-supervised Adversarial Objectives

Maximization Objective: Adversarial examples can be generated by maximizing a self-supervised objective based on reconstructing a transformed(rotated/shuffled) image.

$$\boxed{\underset{x'}{\text{maximize}} \quad \mathcal{L}_{max} = \|\mathcal{F}(\mathcal{T}_s(\boldsymbol{x}')) - \boldsymbol{x}\|^p} \quad \boxed{x' = x + \alpha \times \nabla_x \mathcal{L}_{max}}$$

x' is the adversarial image, $\mathcal{T}_s$ is represents pixel transformation (e.g., rotation or jigsaw shuffle)

Minimization Objective: The surrogate model is trained by minimizing the reconstruction error of the adversarial and clean output with respect to the original image. Furthermore, the model's feature space is regulated by enforcing alignment between clean and adversarial features at the encoder stage.

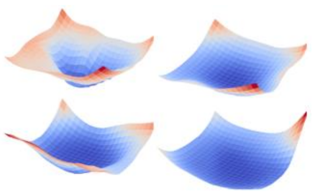
$$\mathcal{L}_{out} = \|\mathcal{F}(\mathcal{T}_s(\boldsymbol{x}')) - \boldsymbol{x}\|^p + \|\mathcal{F}(\mathcal{T}_s(\boldsymbol{x})) - \boldsymbol{x}\|^p \quad \text{and} \quad \mathcal{L}_{feature} = \|\mathcal{F}^n(\mathcal{T}_s(\boldsymbol{x}')) - \mathcal{F}^n(\mathcal{T}_s(\boldsymbol{x}))\|^p$$

$$\boxed{\mathcal{L}_{min} = \mathcal{L}_{out} + \lambda \mathcal{L}_{feature}}$$

Training Settings & Unsupervised Attack

Limited Samples: Surrogate model are trained only on the few data samples (≤ 20) on which adversarial examples need to be crafted.

Cross-Domain Samples: Due to abundance of unlabelled data, we scale our self-supervised adversarial training to large-scale datasets and then test the cross-domain transferability of our method.



The first row shows loss landscapes of surrogate models trained by reducing the reconstruction objective (**top:** rotation, **bottom:** jigsaw). The second row shows the smoother loss surfaces obtained by using our training method. This has significant effect on finding generalizable adversarial examples with better transferability.

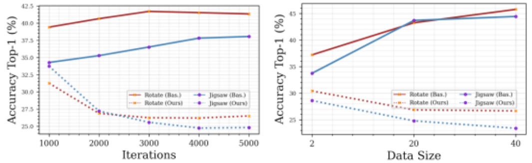
Crafting Adversarial examples: The attack objective for surrogate models trained in a self-supervised manner is based on maximizing the reconstruction error between clean and adversarial samples. Adversarial transferability is evaluated on a selected set of 5000 images from ImageNet validation set, with perturbation budget of 0.1.

Quantitative Results

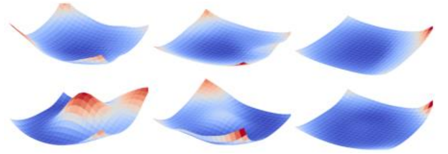
Limited Samples

Transformation	Method	VGG-19	Inc-V3	Res152	Dense121	SeNet	WRN	MNet-V2	Average
Jigsaw	[26]	31.54	50.28	46.24	42.38	59.06	51.24	25.24	43.71
	Ours	16.82	25.54	31.18	22.64	38.06	25.76	13.70	**24.81** (-18.9)
Rotation	[26]	31.14	48.14	47.40	41.26	58.20	50.72	26.00	43.27
	Ours	19.02	25.76	33.60	25.60	38.92	29.78	15.38	**26.87** (-16.4)
Prototypical	[26]	18.74	33.68	34.72	26.06	42.36	33.14	16.34	29.29
	Ours	17.02	21.48	28.66	21.06	35.04	23.56	13.06	**22.84** (-6.45)

Our attack boosts adversarial transferability across ImageNet models.



The performance of our method improves with more training iterations and data size in contrast to the baseline.



Loss landscape of surrogate models with increasing robustness strength (**α**) from left to right. The first row shows the loss surface on the clean samples, while as the second row plots the loss surface with respect to adversarial samples. It becomes harder to maximize the reconstruction error or flip decisions on the excessively smooth loss surface during attack.

Cross-Domain Samples

Transformation	Method (→)	[26]	Ours
Rotation	CoCo	28.56	**23.31**
	Paintings	27.83	**17.75**
	Comics	58.38	**24.19**
Jigsaw	CoCo	43.93	**31.28**
	Paintings	44.07	**33.42**
	Comics	67.70	**41.54**

Our method provides favorable results on cross-domain transferability averaged across ImageNet models.

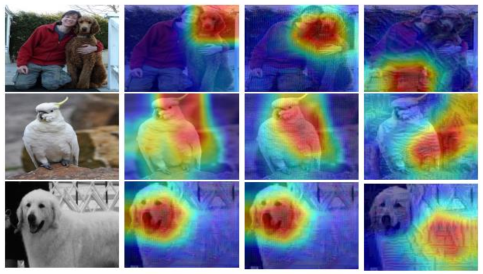
Transformation (↓)	Dataset (→)	CoCo		Paintings		Comics	
	No Attack	[26]	Ours	[26]	Ours	[26]	Ours
Rotation	39.7	19.3	**14.6**	17.2	**11.9**	34.3	**13.3**
Jigsaw	39.7	24.7	**14.5**	24.1	**14**	38	**20.8**

Transferability to object detector (DETR) based on mAP is evaluated on CoCo validation set.

Transformation (↓)	Dataset (→)	CoCo		Paintings		Comics	
	No Attack	[26]	Ours	[26]	Ours	[26]	Ours
Rotation	61.8	53.2	**48.9**	52.6	**46.9**	57.78	**47.81**
Jigsaw	61.8	53.9	**46.6**	53.2	**48.5**	58.29	**51.65**

Transferability to object segmentation (DINO) based on Jacard index is evaluated on DAVIS validation set.

Attention to salient regions



Samples Clean Maps [26] Ours

Conclusion

- We show the benefits of unsupervised adversarial training to learn transferable adversarial perturbations.
- Our adversarial training method reduces over-fitting during training and can exploit very few data samples to learn meaningful adversarial features while it can also scale to large unsupervised datasets.
- Our unsupervised attack is task independent and allows cross-domain attacks (e.g., learning surrogate on comics and transferring its perturbations to models trained on natural images).