Adversarial Pixel Restoration as Pretext Task for Transferable Perturbations

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

(hashmat.malik,shahina.kunhimon,muzzammal.naseer,salman.khan,fahad.khan)@mbzu.ac.ae

Paper-ID: 546

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.

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.

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

\[
\max \mathcal{L}_{\text{max}} = |\mathcal{L}(T(x', y)) - \mathcal{L}(x)| \\
\mathcal{L} = |\mathcal{L}(x)|
\]

where \( x \) is the adversarial image, \( T \) 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.

\[
\min \mathcal{L}_{\text{min}} = |\mathcal{L}(T(x', y)) - \mathcal{L}(T(x))|
\]

\[
\mathcal{L}_{\text{feature}} = \mathcal{L}_{\text{encoder}}
\]

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.

Attention to salient regions

Conclusion

- We show the benefits of unsupervised adversarial training to learn transferable adversarial perturbations.
- Our adversarial training method reduces overfitting 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).

Quantitative Results

<table>
<thead>
<tr>
<th>Transformation Method</th>
<th>Ours</th>
<th>CoCo</th>
<th>Jigsaw</th>
<th>CoCo+</th>
<th>Jigsaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rotation</td>
<td>29.4</td>
<td>29.4</td>
<td>36.5</td>
<td>29.8</td>
<td>36.5</td>
</tr>
<tr>
<td>Translation</td>
<td>58.8</td>
<td>52.0</td>
<td>56.0</td>
<td>52.0</td>
<td>56.0</td>
</tr>
<tr>
<td>Jitter</td>
<td>28.0</td>
<td>28.0</td>
<td>28.0</td>
<td>28.0</td>
<td>28.0</td>
</tr>
</tbody>
</table>

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

Transferability to object detector (DETR) based on mIoU is evaluated on COCO validation set.

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