USING MULTIPLE SELF-SUPERVISED TASKS IMPROVES MODEL ROBUSTNESS

Matthew Lawhon, Chengzhi Mao & Junfeng Yang
Department of Computer Science
Columbia University
New York, NY 10027, USA
matthew.lawhon@columbia.edu, {mcz, junfeng}@cs.columbia.edu

ABSTRACT

Deep networks achieve state-of-the-art performance on computer vision tasks, yet they fail under adversarial attacks that are imperceptible to humans. In this paper, we propose a novel defense that can dynamically adapt the input using the intrinsic structure from multiple self-supervised tasks. By simultaneously using many self-supervised tasks, our defense avoids over-fitting the adapted image to one specific self-supervised task and restores more intrinsic structure in the image compared to a single self-supervised task approach. Our approach further improves robustness and clean accuracy significantly compared to the state-of-the-art single task self-supervised defense. Our work is the first to connect multiple self-supervised tasks to robustness, and suggests that we can achieve better robustness with more intrinsic signal from visual data.

1 INTRODUCTION

Deep learning architectures achieve state-of-the-art and often superhuman performance across a wide variety of vision tasks (Croce & Hein, 2020). Despite this, as first noticed in 2014, they remain vulnerable to *adversarial attacks*: visually imperceptible perturbations that cause easily classifiable images to be misclassified with high confidence by otherwise state-of-the-art machine learning algorithms (Szegedy et al., 2014). This results in unpredictable behavior in edge-cases, contrived examples and examples unrepresented in training data. The inability to address this sufficiently is a leading hurdle to deploying deep learning solutions to human safety and well-being critical applications like autonomous transportation and health-care.

Though there is a large line of research into *adversarial training*, how we can train networks to resist adversarial attacks, training time defenses can be very computationally expensive and it is difficult to provide guarantees for all possible attack methods (Tramèr et al., 2018). Empirically, *unrestricted white box attacks*, in which an adversary has complete, unrestricted access to the network it is trying to corrupt, have been found very difficult to resist via adversarial training. Mao et al. (2021) propose to adapt to adversarial attacks at test time by restoring the performance of a selected self-supervised task, however, the adaption method significantly reduces the clean accuracy after adaptation because it over-fits to a single self-supervised task.

In this paper, we propose to mitigate the over-fitting problem present in Mao et al. (2021)'s reversal method by using multi-task learning. Our key insight is that different self-supervised tasks, none of which require labels, capture different aspects of the intrinsic structure of images. By restoring the performance of attacked images on multiple self-supervised tasks, we can recover a larger set of features that have been corrupted by an adversarial attack. In addition, Mao et al. (2020) show that it is harder to simultaneously attack multiple tasks, suggesting that this methodology is also resistant to an adaptive attacker who has full access to our reversal methodology and attempts to optimize their attack knowing our reversal methodology in place.

Using three self-supervised tasks, our approach yields statistically significant improvement upon the state-of-the-art defense. On the CIFAR-10 dataset using Carmon et al. (2019)'s robustly trained baseline model, we achieved a 1.1% improvement in classification accuracy of unattacked images,

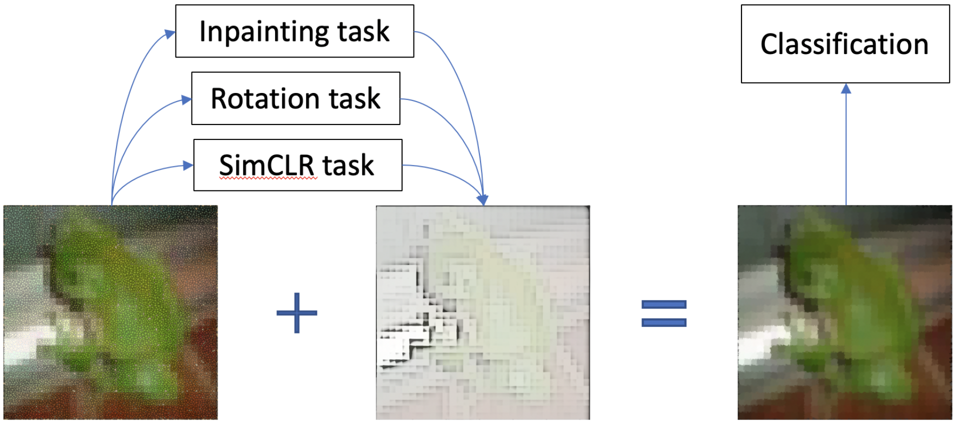

Figure 1: Given a possibly attacked image, in our approach we find some small reverse perturbation to minimize multiple self-supervised learning task losses and recover structure in the image. Adding this reverse perturbation to our image, we are able to improve classification accuracy on attacked images with minimal disruption to classification accuracy on unattacked images.

and a 0.3% improvement in classification accuracy of PGD attacked images compared with the state-of-the-art results from Mao et al. (2021). Our work suggests that deep computer vision models' robustness can be improved by ensuring that they leverage the rich intrinsic structure of image data.

2 Related Work

2.1 Adversarial Attacks

Early work in this field demonstrated the susceptibility of neural networks trained to solve computer vision tasks to human-imperceptible amounts of noise that would result in high-confidence misclassifications (Szegedy et al., 2014). Notable early adversarial attack methods include the Fast Gradient Sign Method (FGSM) from Goodfellow et al. (2015), and Projected Gradient Descent (PGD) from Madry et al. (2018), in which the attacker attempts to maximize the loss function of a classification network within the local ϵ-neighborhood of a particular example. In this paper, we assume white box access to the network and reversal procedures, and attempt to optimize attacker performance using the same minimax optimization setup presented initially by Mao et al. (2021).

2.2 Self-supervised Learning

Image data contains rich intrinsic structure that can be used in learning image representations. In Self-supervised learning for images we use deep architectures to learn unsupervised tasks including canonical examples like rotation prediction, inpainting and contrastive predictive coding (Gidaris et al., 2018; **?**; Pathak et al., 2016; Chen et al., 2020). Mao et al. (2021)'s first noted the transferability of adversarial attacks on classification, a supervised task, to contrastive learning. While we can't repair an attacked image by minimizing the classification loss since we don't have a ground truth label, we can still minimize loss for self-supervised tasks. They show the effectiveness of reversing adversarial attacks by minimizing contrastive loss thereby implicitly strengthening the inherent structure in the image.

2.3 Multitask Learning

In multitask learning, we attempt to learn multiple related tasks at once using a shared architecture. Heuristically, this leverages the fact that much of the representational information needed to solve related tasks is shared (Caruana, 1997). Mao et al. (2020) provide theoretical and empirical results concerning multitask learning's ability to enhance robustness of single-task and multi-task attacks. In this work, we incorporate these theoretical and empirical observations by leveraging multiple self-supervised tasks to repair potentially attacked images, instead of one.

Algorithm 1 Multi-task Learning Reverse Attack

```
Input: Potentially attacked image $x$, step size $\eta$, number of iterations $K$, a classifier $F$, reverse
    attack bound $\epsilon_v$, contrastive loss $\mathcal{L}_s$, roation loss $\mathcal{L}_r$ and inpainting loss $\mathcal{L}_i$.
Output: Class prediction $\hat{y}$
$x' \leftarrow x + n$, where $n$ is the initial random noise
for $k = 1, ..., K$ do
    $L = \mathcal{L}_s(\boldsymbol{x}) + \mathcal{L}_r(\boldsymbol{x}) + \mathcal{L}_i(\boldsymbol{x})$
    $\boldsymbol{x}' \leftarrow \boldsymbol{x}' - \eta(\mathcal{L}_s(\boldsymbol{x})/L)\nabla_{\boldsymbol{x}}\mathcal{L}_s(\boldsymbol{x})$
    if $k \mod 2 = 0$ then
        $\boldsymbol{x}' \leftarrow \boldsymbol{x}' - \eta(\mathcal{L}_r(\boldsymbol{x})/L)\nabla_{\boldsymbol{x}}\mathcal{L}_r(\boldsymbol{x})$
    else
        $\boldsymbol{x}' \leftarrow \boldsymbol{x}' - \eta(\mathcal{L}_i(\boldsymbol{x})/L)\nabla_{\boldsymbol{x}}\mathcal{L}_i(\boldsymbol{x})$
    end if
    $\boldsymbol{x}' \leftarrow \Pi_{(\boldsymbol{x}, \epsilon_v)}\boldsymbol{x}'$ which projects the image back into the bounded region.
    $\boldsymbol{x} \leftarrow \boldsymbol{x}'$
end for
Predict the final output by $\hat{y} = F(\boldsymbol{x}')$
```

3 Our Approach

In our approach we improve Mao et al. (2021)'s results by migrating their self-supervised learning based image repair to a multi-task learning approach using multiple self-supervised tasks. In this paper we experimented with SimCLR, Inpainting and Rotation prediction tasks, but note that this is a somewhat arbitrary and preliminary choice that would benefit from additional investigation in future work (Gidaris et al., 2018; Pathak et al., 2016; Chen et al., 2020).

3.1 Attack Model

We use a standard attack model in which for a given image $\boldsymbol{x}$, classifier F and its loss function $\mathcal{L}_c$ (we use cross-entropy loss, defined as $\mathcal{L}_c(\boldsymbol{x}, y) = H(F(\boldsymbol{x}), y)$), norm parameter p (∞ here) and small ϵ, the attacker searches for an adversarial perturbation $\boldsymbol{x}_a$ where

$$\boldsymbol{x}_a = \arg\max_{\boldsymbol{x}_a} \mathcal{L}_c(\boldsymbol{x} + \boldsymbol{x}_a, y) : \|\boldsymbol{x}_a\|_p < \epsilon$$

3.2 Reverse Model

We observe that often $\boldsymbol{x}_a$ is designed in such a way as to disrupt the inherent structure of the resultant image $\boldsymbol{x} + \boldsymbol{x}_a$. We can thus seek to repair this inherent structure by minimizing the loss associated with our multiple self-supervised task loss function $\mathcal{L}_m(\boldsymbol{x})$, defined as a weighted sum of three self-supervised task loss functions $\mathcal{L}_s(\boldsymbol{x}), \mathcal{L}_i(\boldsymbol{x}), \mathcal{L}_r(\boldsymbol{x})$, explained in detail in the following sections. To reverse an attack for a given input image $\boldsymbol{x}'$ (which may or may not have been attacked), classifier F, norm parameter p (∞ here) and small ϵ_r, we find reverse vector $\boldsymbol{r}$,

$$\boldsymbol{r} = \arg\min_{\boldsymbol{r}} \mathcal{L}_m(\boldsymbol{x}' + \boldsymbol{r}) : \|\boldsymbol{r}\|_p < \epsilon$$

After finding a minimal $\boldsymbol{r}$ via PGD we can recover robust classifications by classifying on $\boldsymbol{x}' + \boldsymbol{r}$ (Madry et al., 2018). See 1 for the full reversal algorithm. Because $\mathcal{L}_m(\boldsymbol{x})$ is a sum of multiple objective functions, a whitebox attack will have to balance objectives, thus reducing its effectiveness as an attacker (Mao et al., 2020). As seen in the equation above, this approach to finding $\boldsymbol{r}$ is independent of F, and is thus compatible with any classification architecture or supervised task.

3.2.1 Contrastive Loss: $\mathcal{L}_s$

In the SimCLR contrastive learning task introduced by Chen et al. (2020), we train a ResNet branch to minimize the latent space distance to copies of the same image under various transformations, and maximize the distance of non-matching image pairs, thus learning an augmentation-invariant

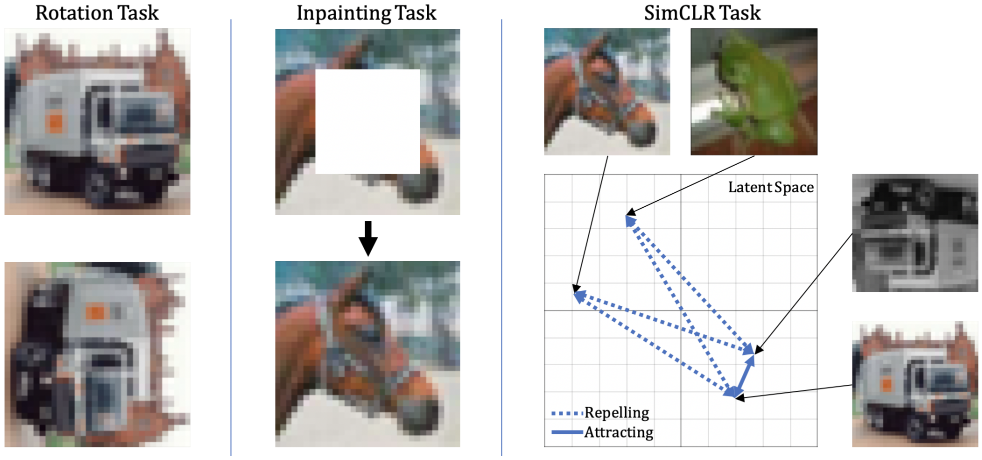

Figure 2: On the left see the rotation task, in which a network predicts the rotation of an image (lower left - 90 degree rotation, upper left - no rotation) (Gidaris et al., 2018). In the middle see the inpainting task in which given the upper middle image, a network attempts to predict the lower middle image (Pathak et al., 2016). On the right see the SimCLR task, in which a network tries to map augmented images (Crop, gray-scale and rotation shown here) from the same class closer to one another in a latent space than images from different classes (Chen et al., 2020)

representation of images (He et al., 2016). More formally, we define our contrastive loss for an image $\boldsymbol{x}$ as

$$\mathcal{L}_s(\boldsymbol{x}) = -\mathbb{E}_{i,j}\left[y_{i,j} \log \left(\frac{\exp(\boldsymbol{z}_i \boldsymbol{z}_j^T/\tau)}{\sum_k \exp(\boldsymbol{z}_i \boldsymbol{z}_k^T/\tau)} \right) \right]$$

Where $\boldsymbol{z}_i$ is a possible result from transforming $\boldsymbol{x}$,$\boldsymbol{z}_j$ is a random transformed image, and $y_{i,j}$ is 1 if $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ originate from the same source image $\boldsymbol{x}$, and 0 otherwise. τ is a hyperparameter. For our augmentations, we sequentially applied random cropping and color jittering, and also applied a horizontal flip and/or grayscale filter at random to each image. See 2 for an example.

3.2.2 Inpainting Task Loss: $\mathcal{L}_i$

In the inpainting task, we train an encoder-decoder deep network to fill in the missing centers of images, using encoders and decoder structures derived from the AlexNet architecture (Krizhevsky et al., 2012; Pathak et al., 2016). For a given whole image $\boldsymbol{x}$, and context encoder-decoder F, we denote the output of F on $\boldsymbol{x}$ as $F(\boldsymbol{x})$. We define $\hat{M}$ to be a binary mask indicating dropped pixels via 1 and 0 otherwise. Letting $\odot$ define the element-wise product operation, we define the inpainting task loss on a given image $\boldsymbol{x}$ as

$$\mathcal{L}_i(\boldsymbol{x}) = \|\hat{M} \odot (\boldsymbol{x} - F((1-\hat{M}) \odot \boldsymbol{x}))\|_2^2$$

Intuitively speaking, we may interpret this as the L_2 norm of the difference between the true inpainted section of $\boldsymbol{x}$, and the predicted inpainted section of the image returned by $F((1-\hat{M}) \odot \boldsymbol{x})$ (predicted using all pixels but the pixels masked by $\hat{M}$, thus giving the $(1-\hat{M})$ term). See 2 for an example.

3.2.3 Rotation Task Loss: $\mathcal{L}_r$

In the image rotation task, we train a deep convolutional neural network to predict the rotation angle of an input image $\boldsymbol{x}$. We use the experimental setup given in Gidaris et al. (2018). Given a convolutional neural network F with learnable parameters θ designed to predict image rotations, where $F^k(\boldsymbol{x})$ denotes the probability of $\boldsymbol{x}$ having been transformed by rotation labeled k, and $\boldsymbol{x}^k$

Classification Accuracy		
Reversal Method	Attacked Input	Clean Input
None	63.9%	89.7%
SSL	65.3%	86.6%
MTL	**65.6%**	**87.7%**

Table 1: Using Carmon et al. (2019)'s robustly trained baseline model on CIFAR-10, we see a 1.1% improvement in classification accuracy of clean images, and a 0.3% improvement in classification accuracy of PGD attacked images compared with the state-of-the-art results from Mao et al. (2021)

denotes $\boldsymbol{x}$ transformed by rotation labeled k we define the loss:

$$\mathcal{L}_r(\boldsymbol{x}, \theta) = -\frac{1}{K}\sum_{k=1}^{K} \log(F^k(\boldsymbol{x}^k|\theta))$$

Intuitively, we treat this as a classification problem with cross-entropy loss and take the average loss across a set of K different rotations, trying to maximize the probability of the correct rotation for each k. Minimizing $\mathcal{L}_r(\boldsymbol{x}, \theta)$ pushes $F^k(\boldsymbol{x}^k|\theta) \to 1$ as desired. See 2 for an example.

4 RESULTS

In line with the literature, we set our attack bounds $\epsilon = 8$ (Madry et al., 2018). In configuring reversal parameters we continue Mao et al. (2021)'s experiments and similarly find that the defense aware attacker can do no better than standard PGD attack, and that a reversal with $\epsilon = 8$ maximizes attacked input classification accuracy. Our results can are listed in 4 and our code can be found at `https://github.com/mattlawhon/SelfSupDefense/tree/one-by-one`.

4.1 ANALYSIS

Using three self-supervised tasks, our approach yields significant improvement upon the state-of-the-art defense. On the CIFAR-10 dataset using Carmon et al. (2019)'s robustly trained baseline model, we achieved a 1.1% improvement in classification accuracy of unattacked images, and a 0.3% improvement in classification accuracy of PGD attacked images compared with the state-of-the-art results from Mao et al. (2021). Given an observed standard error of at most 0.4% in the estimation of the approaches' true accuracy, this indicates statistically significant improvement in the classification accuracy of clean input with over 95% confidence. Our work suggests that deep computer vision models' robustness can be improved by ensuring that they leverage the rich intrinsic structure of image data.

We note that this procedure increases computational cost because it requires the calculation of two gradients rather than one per iteration of the PGD reversal procedure. Further, we use different backbones for each self-supervised task, which improves accuracy of each task at the cost of performance. This may very well be an acceptable trade-off for better performance on clean input, though exploring this trade-off is an interesting direction for future research.

5 CONCLUSIONS

Though we show statistically significant improvements in overfitting over baseline, there is much additional work to be done in this domain because the scope of the experiments presented here is limited. Particularly exciting avenues for expanding this body of work include: experimenting with other self-supervised tasks, different datasets and adding a shared backbone for all self-supervised tasks for a true multi-task learning approach. Further, we note that this approach broadly furthers the finding that both multi-task learning and self-supervised learning hold promise to increasing adversarial robustness.

5.1 ACKNOWLEDGEMENTS

We would like to thank Gustave Ducrest for the productive discussions.

REFERENCES

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Chengzhi Mao, Amogh Gupta, Vikram Nitin, Baishakhi Ray, Shuran Song, Junfeng Yang, and Carl Vondrick. Multitask learning strengthens adversarial robustness. In *European Conference on Computer Vision*, pp. 158–174. Springer, 2020.

Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 661–671, 2021.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.