Reversing Adversarial Attacks with Multiple Self Supervised Tasks

Matt Lawhon & Gustave Ducrest
Columbia University
New York, NY, 10025
{mwl2131,gd2528}@columbia.edu

Abstract

Adversarial attacks continue to exhibit the ability to slightly perturb images such that state-of-the-art models fail common vision tasks [20]. Recent work shows that adversarial attacks have difficulty attacking multiple supervised tasks at once, suggesting the use of multi-task learning to provide robustness to adversarial attacks [15] and that self-supervised tasks are a promising method to repair images as contrastive loss is also attacked by adversarial noise [14]. In this paper we propose integrating these two approaches by repairing images using a multi-self-supervised task learning approach. We find statistically significant improvements in classification accuracy over baseline benchmark robustly trained models [2].

1 Introduction

Deep learning architectures achieve state of the art and often superhuman performance across a wide variety of vision tasks [4]. Despite this, as first noticed in 2014, they remain vulnerable to *adversarial attacks* [20]. An adversarial attack generally refers to fact that we can often find, for a given image x, classifier f, true label l, norm parameter p (normally 1, 2 or ∞) and small ϵ that

$$\exists x_a : \quad f(x + x_a) \neq l, \|x_a\|_p < \epsilon$$

where for sufficiently small ϵ, the difference between x and $x + x_a$ is imperceptible to the human eye. This results in unpredictable behavior in edge-cases, contrived examples and examples unrepresented in training data. These adversarial examples are also transferable across model architectures and training sets. 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, it is difficult to provide guarantees for all possible attack methods. 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 training methods. Recent works have found that both multitask learning and self-supervised learning methods hold promise for resisting adversarial attacks.

In this paper, we integrate a multitask learning approach into *Mao et al.'s* method of reversing adversarial attacks via self-supervised learning. In doing so, we combine the inherent representation learning that occurs in learning self-supervised tasks and in solving multiple tasks. Thus this technique holds the promise of improving the already effective technique of reversing attacks using a single self-supervised learning task. As *Mao et al.* notes, an advantage of this strategy of separating the defense strategy from the visual representation is the ability to work with any variety of attacks that violate natural image manifolds, and work with existing classifiers and defense methods [14].

Preprint. Under review.

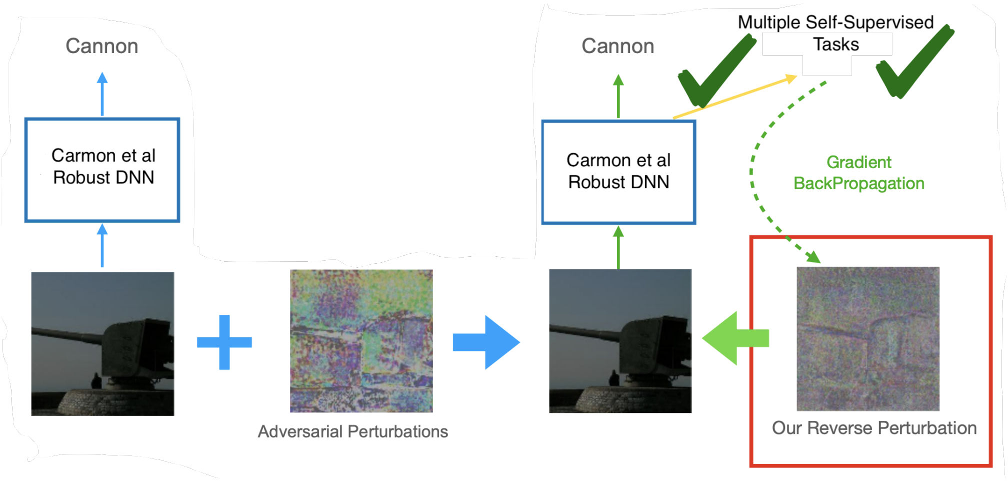


Figure 1: In this figure we demonstrated our approach visually. Given a cannon image with added adversarial perturbations, we find some small reverse perturbation via PGD to minimize multi-self-supervised learning task loss and recover the inherent structure in the image. Adding this reverse perturbation to our adversarially perturbed image, we are able to recover a correct classification on the cannon image [14]

This approach yields impressive results that match its theoretical basis. On the CIFAR-10 dataset we measured a 6% improvement on classification accuracy in the PGD attacked image case with negligible loss in accuracy with unattacked images classified on the reverse attacked image. This is an exciting result given that the baseline is *Carmon et al.'s* benchmark model trained by unsupervised robust training methods [2]. We also demonstrated a 3% improvement compared to the state of the art presented by *Mao et al.*. This presents a minor but exciting improvement over the results presented by *Mao et al.'s* self-supervised task reversal approach.

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 [20]. This result was quickly verified to be a fundamental feature of deep learning architectures, rather than a problem that is unique to vision problems (that being said, higher dimensional input space problems do struggle with this more than low dimensional input space problems) [6]. Notable early adversarial attack methods include the Fast Gradient Sign Method (FGSM) [6], and Projected Gradient Descent (PGD) [12], in which perform attempt to maximize the loss function of a classification network within the local ϵ-neighborhood of a particular example. Further methods address the failures of specific architectures and training methods that claim to provide robustness to adversarial attacks, like those that use obfuscated gradients [1] or assume black box access [16]. 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 in [14].

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 train deep learning architectures on unsupervised tasks including rotation prediction [5], inpainting [19], and contrastive predictive coding [18]. Training networks on these tasks yield representations that have rich structure and are useful in solving a number of downstream supervised tasks. Self-supervised learning has emerged as one of the dominant tools used in production machine learning systems as a result of the utility of learned representations. This paper incorporates all these self-supervised learning tasks.

Mao et al.'s paper on reversing adversarial attacks via natural supervision [14] 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 losses for self-supervised tasks. The paper's experiments showed the effectiveness creating reverse adversarial perturbations that minimize contrastive loss. This form of perturbation aims to mitigate classification damage caused by an adversarial attack, by implicitly strengthening the inherent structure in the image. This approach forms the basis, both in theory and in practice via codebase, of our approach. Our paper uses the results from contrastive learning as the inspiration for using other self-supervised tasks to reverse adversarial attacks.

2.3 Multitask Learning

In multitask learning, we attempt to learn multiple related tasks at once using a shared architecture. Heuristically, this attempts to leverage the fact that much of the representational information needed to solve related tasks is shared [3]. Further, these representations and their relatedness is discovered in a unsupervised manner. In theory, this means that a true-representational inductive knowledge bias is learned in a multitask learning approach.

In leveraging multitask learning to enhance robustness, we are building directly off *Mao et al.'s* paper on strengthening robustness with multitask learning [15]. They provide theoretical and empirical results concerning multitask learning's ability to enhance robustness of single-task and multi-task attacks. Intuitively, this results from the notion that in increasing output dimensionality (where each output dimension corresponds to an individual task), we improve the robustness of the model, because perturbations needed to attack multiple tasks cancel each other out. However, a model's robustness has an inverse relationship with the covariance of the model's different tasks: similar tasks do not increase robustness as much as unrelated tasks. In this work, we incorporate these theoretical and empirical observations by leveraging multiple self-supervised tasks to repair potentially attacked images, instead of one.

3 Our Approach

In our approach we attempt to improve *Mao et al.'s* results by migrating their self-supervised learning based image repair to a multi-self-supervised task learning based image repair approach. In this way, our method consists of a reversal method for adversarial attacks and combines self-supervised and multitask learning based representation learning.

3.1 Attack Model

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

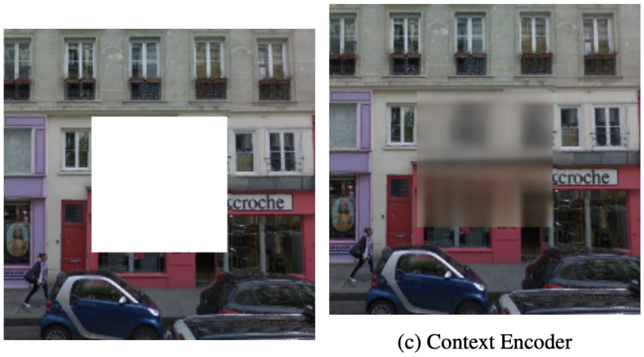
$$x_a = \arg\max_{x_a} \mathcal{L}_c(x + x_a, y), \|x_a\|_p < \epsilon$$

3.2 Reverse Model

We observe that often x_a is designed in such a way as to disrupt the inherent structure of the resultant image $x + x_a$. We can thus seek to repair this inherent structure by minimizing the loss associated with a multi-self-supervised task loss function $\mathcal{L}_m(x) = \sum_k \alpha_k \mathcal{L}_k(x))$. In our experiments, we use $\alpha_k = 1$ for all k. Thus to reverse an attack for a given input image x' (which may or may not have been attacked), classifier F, norm parameter p (∞ here) and small ϵ_r, we find reverse vector r,

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

After finding a minimal r we can recover robust classifications by classifying on $x' + r$. This optimization, like the attack optimization, is non-convex. As a result, we use the state of the art method, Progressive Gradient Descent (PGD), to find optimal r [12]. Because $\mathcal{L}_m(x)$ is a sum of multiple objective functions, a whitebox attack will have to balance objectives, thus reducing



(a) Input context

(c) Context Encoder ($L2$ loss)

Figure 2: In (a) we present a sample input for the inpainting task network to paint and in (c) we present a sample output from an appropriately trained network using L_2 loss [19]

its effectiveness as an attacker [15]. As seen in the equation above, this approach to finding r is independent of F, and is thus portable to any classification architecture, or even any supervised task. This is significant because while our repair model needs to be trained only once - separately of the classification model - robustness training for supervised models, the most common solution to adversarial attacks, occurs during training, thus adding to training time.

The multi-self-supervised task loss function is a sum of three self-supervised loss functions, $\mathcal{L}_s(x), \mathcal{L}_i(x), \mathcal{L}_r(x)$, explained in detail in 3.2.1, 3.2.2, 3.2.3. In implementing this in a multi-task paradigm, we choose some r that minimizes loss across a variety of self-supervised tasks, providing a robust guarantee that r recovers internal structure to increase the likelihood of having an accurate classification on $x' + r$. Since representations learned by all of these tasks individually have been shown to be useful in downstream supervised tasks, such as classification, and attacks on the classification task transfer to contrastive learning, a useful self-supervised representation learning task, we use their loss functions to reverse attacks on classification. *Mao et al.* [15] notes that not only the number of tasks in a model, but also its diversity, increases its adversarial robustness. To this end, we use 3 tasks that learn representations with 3 different invariances: invariance to random augmentations ($\mathcal{L}_s(x)$), invariance to masking ($\mathcal{L}_i(x)$), and rotation invariance ($\mathcal{L}_r(x)$). While these representations are all useful for classification and other supervised tasks, they are all trained on different architectures and have vastly different invariances, thus should be relatively independent and difficult to attack simultaneously in a white-box attack.

3.2.1 Contrastive Loss: $\mathcal{L}_s(x)$

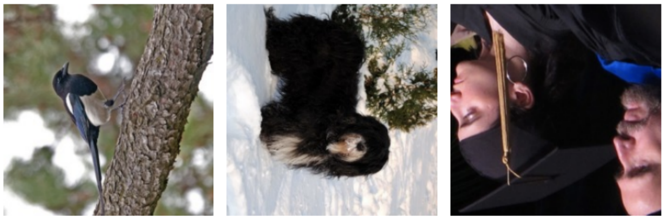
To use the contrastive learning task introduced by LeCun et al. [7], we train a ResNet [8] branch to minimize the latent space distance to copies of the same image under various transformations, and maximize the distance of non-matching image pairs, thus learning an augmentation-invariant representation of images. More formally, we define our contrastive loss for an image x as

$$\mathcal{L}_s(x) = -\mathbb{E}_{i,j}\left[y_{i,j}\log\left(\frac{\exp(z_i z_j^T/\tau)}{\sum_k \exp(z_i z_k^T/\tau)}\right)\right]$$

Where z_i is a possible result from transforming x,z_j is a random transformed image, and $y_{i,j}$ is 1 if z_i and z_j originate from the same source image 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.

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

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 [10]. An example of the inpainting task and the solution possible with an appropriately trained network is shown in 2.



90° rotation 270° rotation 180° rotation

Figure 3: In the rotation self-supervised task, we train a model to identify how much, if at all, a given image x has been rotated [5]

In our approach we use the network and approach proposed by *Pathak et al.* [19]. For a given whole image x, and context encoder-decoder F, we denote the output of F on x as $F(x)$. 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 x as

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

Intuitively speaking, we may interpret this as the L_2 norm of the difference between the true inpainted section of x, and the predicted inpainted section of the image returned by $F((1 - \hat{M}) \odot x)$ (predicted using all pixels but the pixels masked by $\hat{M}$, thus giving the $(1 - \hat{M})$ term). For full disclosure we note that as clearly demonstrated in 2 and as can be verified empirically, the L_2 loss method results in blurry solutions. This is believed to occur because blurry solutions present lower risk to arbitrarily increase mean pixel-wise error with bad predictions [19]. This can be solved using adversarial loss methods, however we don't implement such methods in our solution because we believe that little of the representational information required for the goal task of repairing images is contained in fine-grained pixel differences. This is, however, a promising direction for future exploration.

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

In the image rotation task, we train a deep convolutional neural network to predict the rotation angle of an input image x 3. We use the experimental setup given in *Gidaris et al.'s* paper [5]. Given a convolutional neural network F with learnable parameters θ designed to predict image rotations, where $F^k(x)$ denotes the probability of x having been transformed by rotation labeled k, and x^k denotes x transformed by rotation labeled k we define the loss:

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

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 $\mathcal{L}_r(x, \theta)$ pushes $F^k(x^k|\theta) \to 1$ as desired. This task is effective at learning representations of images as can be seen from that fact that in the CIFAR-10 dataset with $K = 4$, it was able to achieve over 90% accurate rotation predictions.

4 Results

For time considerations, we used pre-trained models for each SSL task and the baseline classification model [21][13]. We evaluate the classification task on the CIFAR-10 dataset [9]. We set $\tau = 0.2$ for the contrastive learning task, and trained all self-supervised tasks on 200 epochs with batch size of 100 and the Adam optimizer with a learning rate of 0.001. We used the PGD attack with a perturbation bound of 8/255 to create adversarial inputs. We build upon the state of the art adversarial trained network presented by *Carmon et al.* [2].

Over our (albeit limited sample size) subset of the CIFAR-10 dataset we measure a 6% improvement on classification accuracy in the PGD attacked image case with negligible loss in accuracy with

Table 1: Results on 200 image subset of CIFAR 10 classified with *Carmon et al.'s* SoA robust trained DNN network [2]

Experiment Parameters		
Attack	Reversal	Accuracy (SE ≈ 0.03)
-	-	0.91
-	MTL SSL	0.87
-	SSL	0.89
PGD	-	0.59
PGD	MTL SSL	0.65
PGD	SSL	0.62

unattacked images classified on the reverse attacked image. We note that while negligible, our method has a slightly lower accuracy than the baseline repair using single-task contrastive loss on clean inputs. We see that our method also beats the baseline repair method by 3%.

5 Conclusion and Future Work

In this paper, we propose a multi-SSL (MSSL) task framework to repair adversarial inputs that can be carried over to different model architectures and supervised tasks without retraining. When tested on the CIFAR-10 dataset, our method increases classification accuracy against PGD attacks by 6%, and beats single-task repair with constrastive learning by 3%.

Given the somewhat limited scope of our experimenting due to time and limited computing resources, we propose many avenues for future research that involve further exploring the possibilities within the same scope of reversing adversarial attacks via multi-self-supervised task learning. These possibilities include

- Extending experimenting to test the following (we note that the our project's codebase is configured to run this attack, but we did not run this due to time constraints):
  - Multi-task-aware attacks vs Multi-task repair.
  - Contrastive-aware attacks vs Multi-task repair (does this repair do better than just contrastive learning?)
  - Multi-task-aware attacks vs no repair (does this attack do worse because of balancing?)
- Experimenting with which tasks are most and least helpful in repairing images by varying α_k
- Using random missing patches in the inpainting task, in contrast to missing centers, so that contexts on all parts of the image are well represented
- Experimenting our method's robustness against different attack methods and testing its portability with other tasks and architectures
- Experimenting with the upper bound on robustness by expanding our multitask framework to include other self-supervised tasks like colorization [11], jigsaw puzzle solving [17], and more
- Implementing different proposed architectures and loss functions for included self-supervised tasks and seeing if the architecture is as important as the task/diversity of tasks
- Experimenting on different datasets spanning different class sizes and standard resolutions
- Adding a shared backbone for all self-supervised tasks for a true multi-task learning approach

Though these results are limited in nature, these findings suggest that MSSL image repair is a promising new direction for research in mitigating adversarial attacks and generally improving adversarial robustness.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. *Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples*. 2018. arXiv: `1802.00420 [cs.LG]`.

[2] Yair Carmon et al. *Unlabeled Data Improves Adversarial Robustness*. 2019. arXiv: `1905.13736 [stat.ML]`.

[3] Rich Caruana. "Multitask Learning". In: *Machine Learning* 28.1 (July 1, 1997), pp. 41–75. ISSN: 1573-0565. DOI: `10.1023/A:1007379606734`. URL: `https://doi.org/10.1023/A:1007379606734` (visited on 12/19/2021).

[4] Francesco Croce and Matthias Hein. *Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks*. 2020. arXiv: `2003.01690 [cs.LG]`.

[5] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. *Unsupervised Representation Learning by Predicting Image Rotations*. 2018. arXiv: `1803.07728 [cs.CV]`.

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: `1412.6572 [stat.ML]`.

[7] Raia Hadsell, Sumit Chopra, and Yann LeCun. "Dimensionality Reduction by Learning an Invariant Mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*. IEEE Computer Society, 2006, pp. 1735–1742. DOI: `10.1109/CVPR.2006.100`. URL: `https://doi.org/10.1109/CVPR.2006.100`.

[8] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: `1512.03385 [cs.CV]`.

[9] Alex Krizhevsky. *CIFAR-10 and CIFAR-100 datasets*. URL: `https://www.cs.toronto.edu/~kriz/cifar.html` (visited on 12/19/2021).

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: `https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[11] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. *Colorization as a Proxy Task for Visual Understanding*. 2017. arXiv: `1703.04044 [cs.CV]`.

[12] Aleksander Madry et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. arXiv: `1706.06083 [stat.ML]`.

[13] Chengzhi Mao. *Adversarial Attacks are Reversible via Natural Supervision*. original-date: 2021-03-24T02:22:53Z. Dec. 19, 2021. URL: `https://github.com/cvlab-columbia/SelfSupDefense` (visited on 12/21/2021).

[14] Chengzhi Mao et al. *Adversarial Attacks are Reversible with Natural Supervision*. 2021. arXiv: `2103.14222 [cs.CV]`.

[15] Chengzhi Mao et al. *Multitask Learning Strengthens Adversarial Robustness*. 2020. arXiv: `2007.07236 [cs.CV]`.

[16] Nina Narodytska and Shiva Prasad Kasiviswanathan. *Simple Black-Box Adversarial Perturbations for Deep Networks*. 2016. arXiv: `1612.06299 [cs.LG]`.

[17] Mehdi Noroozi and Paolo Favaro. *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*. 2017. arXiv: `1603.09246 [cs.CV]`.

[18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*. 2019. arXiv: `1807.03748 [cs.LG]`.

[19] Deepak Pathak et al. *Context Encoders: Feature Learning by Inpainting*. 2016. arXiv: `1604.07379 [cs.CV]`.

[20] Christian Szegedy et al. *Intriguing properties of neural networks*. 2014. arXiv: `1312.6199 [cs.CV]`.

[21] Wilson Yan. *Introduction*. original-date: 2020-03-26T03:36:49Z. Dec. 3, 2021. URL: `https://github.com/wilson1yan/cs294-158-ssl` (visited on 12/21/2021).