Towards Better Accuracy and Robustness with Localized Adversarial Training

Eitan Rothberg
Computer Science and Engineering
Ohio State University
Columbus, Ohio 43210

Tingting Chen, Hao Ji
Computer Science
California State Polytechnic University
Pomona, California 91768

Abstract

As technology and society grow increasingly dependent on computer vision, it becomes important to make sure that these technologies are secure. However, even today's state-of-the-art classifiers are easily fooled by carefully manipulated images. The only solutions that have increased robustness against these manipulated images have come at the expense of accuracy on natural inputs. In this work, we propose a new training technique, localized adversarial training, that results in more accurate classification of both both natural and adversarial images by as much as 6.5% and 99.7%, respectively.

Introduction

Since the advent of machine learning to the field of computer vision, image classification software has surpassed human capabilities and enabled a slew of new technologies including facial recognition authentication, self-driving cars, and smart security cameras (Akhtar and Mian 2018). However, a unique challenge threatens these technologies: the existence of images which appear normal to humans, but reliably fool image classifiers (Szegedy et al. 2014). Because convolutional neural networks (CNNs) tend to focus on minor and easily manipulated details, attacks such as the Fast Gradient Sign Method (FGSM), discovered by Goodfellow et al. (Goodfellow, Shlens, and Szegedy 2015), and its iterative counterpart, Projected Gradient Descent (PGD) from Kurakin et al. (Kurakin, Goodfellow, and Bengio 2017), have been able to reliably generate adversarial examples by reverse engineering the training process. In the same way a classifier's weights are updated to minimize its loss during training, an adversarial attack updates a particular image to maximize the classifier's error on that image. Adversarial training, the process of including adversarial examples in the training set, is one popular defense technique. However, while this technique has been shown to improve the robustness of a classifier against adversarial examples, the inclusion of adversarial images in the training process weakens classifiers' accuracy on natural, unaltered images (Tsipras et al. 2018) (Su et al. 2018). Building classifiers that maintain state-of-the-art accuracy on both natural and adversarial examples is a key challenge in image classification, as a solution would provide not only defense, but also insight into the nature of CNNs (Tsipras et al. 2018). This work outlines the beginnings of a simple but effective solution: *including images with only adversarial backgrounds in the training set.* We successfully implemented this strategy with the MNIST dataset, creating a model that outperforms a traditional classifier by 6.5% on natural inputs and at least 65.3% on all attempted adversarial inputs.

Localized Adversarial Training

In this work, we train a classifier on images where only the backgrounds are set adversarial in a careful way, in order to improve robustness against adversarial attacks without sacrificing too much accuracy on natural inputs.

Algorithm 1 Localized Adversarial Training

```
$n$ is a CNN; $\epsilon$ is the maximum value that any pixel may legally change; $attack$ describes which pixels may legally change
repeat for each minibatch $B$ in training data
    repeat for each image in $B$
        $x \leftarrow image$
        $\lambda \leftarrow$ PGD_Attack $(x, n)$    ▷ Noise generated by PGD attack
        $x' \leftarrow x + \lambda$
        $epmatrix$ is initialized
        $epmatrix \leftarrow$Localize $(epmatrix, \epsilon, attack)$    ▷ Changes are localized
        $x'$ clipped to within $x$+$epmatrix$, $x$-$epmatrix$
        replace original image in $B$ with $x'$
    until every image in batch is altered
    Train $n$ on updated batch $B$
until training is complete
```

As described in Algorithm 1, localized adversarial training iterates through small batches of the training data, making each image in each batch adversarial. For each image, a PGD attack generates and adds adversarial perturbations. Then, the attack is localized by creating a matrix of equal size to the image, denoted as $epmatrix$, where each value is ϵ if the corresponding pixel is altered, or zero otherwise. (In the next Section, we will describe in detail the different localized $attacks$ we explored.) The adversarial image is then clipped, to ensure that its distance from the original image at any given pixel is no more than that pixel's corresponding value in the epsilon matrix (meaning that higher values of ϵ allow more visible changes at those pixels). Finally, once all the images in a batch are adversarial, the neural network is updated and trained to recognize those images correctly.

Procedure

To test the robustness and accuracy of locally adversarially trained models, we train 5 CNN classifiers to recognize handwritten digits from the MNIST dataset. Each model is trained with two convolutional layers, a fully connected layer, and an output layer, and undergoes 100,000 steps of training. The first model is a ”natural” model, which is trained on unaltered MNIST images. The next four models are trained on four different types of adversarial examples generated by the following four attacks, which are also illustrated by the four columns in Figure 1. Each attack is iterated over 100 steps, and ϵ is set to .3 for every pixel that is allowed to change.

- A ”standard” PGD attack (Column 1 below) following Madry et al.'s 2017 implementation (Madry et al. 2017)
- A general ”background” attack (Column 2) which leaves the middle 36 pixels unaltered
- An ”exact mask” attack (Column 3), which does not alter any pixel belonging to the digit itself
- A ”broad mask” attack (Column 4), where both the pixels belonging to the digit and every pixel directly adjacent remain unchanged

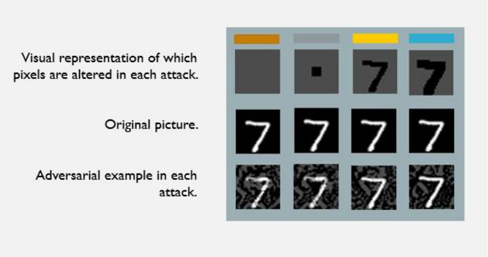
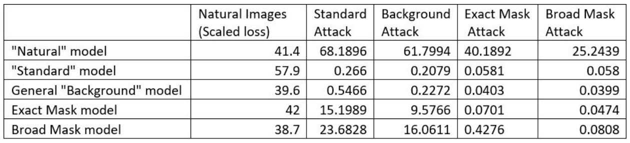

Visual representation of which pixels are altered in each attack.
Original picture.
Adversarial example in each attack.

Figure 1: One standard and three localized PGD attacks

Results

Evaluating each of the five models on each of the four attacks (plus natural inputs), yields a total of 25 evaluations (Figure 2). The exact mask attack (which does not alter pixels in the digit itself) is quite effective, generating images which were identified correctly by the natural model only .08% of the time (compared to 99.19% on natural inputs). This highlights the unnecessary sensitivity of CNN classifiers to background changes. The experiment also confirms the trade-off between accuracy and robustness for standard adversarial training: while the model trained with the standard attack suffers at least 99% less loss than a naturally trained model on various adversarial inputs, it suffers 39.8% more loss on natural inputs.

For localized adversarial training, the model trained on the broad mask attack outperformed the standard model on both types of inputs (on natural inputs by 6.5% and on adversarial inputs by between 65.3% and 99.7%.) The model trained with the ”general background” attack had a similar advantage, outperforming the natural model by 4.4% on natural inputs and at least 99% on all adversarial inputs. Figure 2 describes the loss of each model when tested on natural inputs and each kind of adversarial input. Only the models which underwent localized adversarial training outperformed the ”natural” model on all inputs.



	Natural Images (Scaled loss)	Standard Attack	Background Attack	Exact Mask Attack	Broad Mask Attack
"Natural" model	41.4	68.1896	61.7994	40.1892	25.2439
"Standard" model	57.9	0.266	0.2079	0.0581	0.058
General "Background" model	39.6	0.5466	0.2272	0.0403	0.0399
Exact Mask model	42	15.1989	9.5766	0.0701	0.0474
Broad Mask model	38.7	23.6828	16.0611	0.4276	0.0808

Figure 2: The cross entropy loss of each model on natural inputs and each type of adversarial input. Loss on natural inputs is scaled by 1000.

Conclusions

We implemented the first localized form of adversarial training, improving the robustness against adversarial examples while maintaining state-of-the-art accuracy on natural inputs. We plan to extend our work on larger sets of more complex images, to verify its effectiveness.

Acknowledgement

This work was supported in part by NSF grant CNS-1758017.

References

Akhtar, N., and Mian, A. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *CoRR* abs/1801.00553.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2017. Adversarial examples in the physical world.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks.

Su, D.; Zhang, H.; Chen, H.; Yi, J.; Chen, P.-Y.; and Gao, Y. 2018. Is robustness the cost of accuracy? – a comprehensive study on the robustness of 18 deep image classification models.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2018. There is no free lunch in adversarial robustness (but there are unexpected benefits).