Adversarial Logit Pairing

H. Kannan, A. Kurakin, I. Goodfellow

Google Brain

arXiv: 1803.06373

Reviewed by : Bill Zhang
University of Virginia
https://qdata.github.io/deep2Read/

Outline

Introduction

Definitions and Models

Challenges of Defending

Methods

Logit Pairing Results

Comparison

Conclusion and Future Work

References

Introduction

Basic Premise and Motivation

- In computer vision, object recognition classifiers incorrectly recognize images that have been modified with small, imperceptible changes
- Adversarial defenses are necessary so that ML can be used in situations where attackers can attempt to interfere with systems, ML can be more useful for model-based optimization, better performance guarantees can be obtained, and better understanding of how to enforce smoothness assumptions can be obtained

Introduction

Key Contributions

- Implemented state-of-the-art adversarial training on ImageNet at unprecedented scale
- Proposed logit pairing which encourages logits of two examples to be similar
- Showed that clean logit pairing has minimal computation cost and can defend against PGD almost as well as adversarial training
- Showed that adversarial logit pairing increases accuracy when subjected to white and black-box attacks
- Showed that attacks constructed with their adversarially trained models substantially damages state-of-the-art black-box defenses

Definitions and Models

- Assume that adversaries are capable of forming attacks that consist of perturbations of limited L_∞ norm; more amenable to benchmark evaluations
- Consider both white box (attacker has full information about the model) and black box (attacker has no information about model's architecture or parameters)

Challenges of Defending

- Many defenses have been proposed, but majority have already been broken; Madry et al.'s adversarial training with PGD stands as most long-lasting
- Certified defenses provide guarantee, but total robustness still small compared to empirical results of Madry et al.
- Previously, Madry et al. not shown to scale to ImageNet
- All previous defenses for ImageNet report error rates of 99% on strong, multi-step white box attacks
- In this paper, scale Madry et al. and introduce enhanced defense which further improves over this baseline

Methods

Adversarial Training

- Madry et al. suggests that PGD is a universal first order adversary: robustness against PGD means robustness against most 1st order attacks
- Train on a mixture of clean and adversarial examples, as described by Goodfellow et al. and Kurakin et al.
- Call this defense the mixed-minibatch PGD; still use Madry et al.'s attack

$$\arg\min_{\theta} \left[\mathbb{E}_{(x,y)\in\hat{p}_{\text{data}}} \left(\max_{\delta \in S} L(\theta, x+\delta, y) \right) + \mathbb{E}_{(x,y)\in\hat{p}_{\text{data}}} \Big(L(\theta, x, y) \Big) \right]$$

Methods

Logit Pairing

- Propose logit pairing, a method to make encourage logits from two images to be similar
- For a model with inputs x which computes logits $z = f(x)$, add a loss

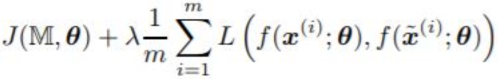
  $$\lambda L(f(x), f(x'))$$

  for a pair of images x and x'
- For this paper, used L^2 loss, but other distance metrics could also work

Methods

Adversarial Logit Pairing

- ALP matches the logits from a clean image x and its adversarial image x'; this provides more regularization than just training the model to assign the same target to both images
- For a model with parameters θ trained on a minibatch $\mathbb{M}$ of clean examples $\{x^{(1)}, \ldots, x^{(m)}\}$ and corresponding adversarial examples $\{\tilde{x}^{(1)}, \ldots, \tilde{x}^{(m)}\}$, let $f(x, \theta)$ be the function mapping inputs to logits and $J(\mathbb{M}, \theta)$ be the adversarial training loss
- Then, train on following loss



$$J(\mathbb{M}, \boldsymbol{\theta}) + \lambda \frac{1}{m} \sum_{i=1}^{m} L\left(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), f(\tilde{\boldsymbol{x}}^{(i)}; \boldsymbol{\theta})\right)$$

Methods

Clean Logit Pairing

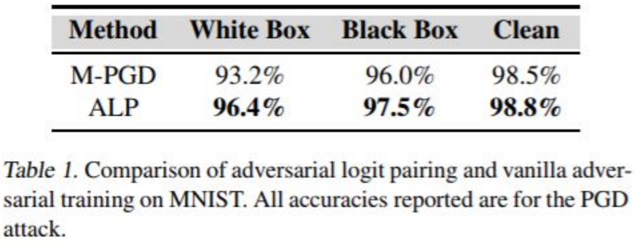
- In CLP, x and x' are two randomly chosen clean samples (which frequently have different classes)
- Wanted to use this to perform an ablation study, understanding the contribution of the pairing loss relative to the formation of clean and adversarial pairs
- Found that this method actually increases robustness in MNIST and SVHM; perhaps because model is learning to predict logits with smaller magnitude and being penalized for overconfidence
- As such, also tested logit squeezing which directly penalizes large logit norms
- Minimize the following loss:

$$J^{(\text{clean})}(\mathbb{M}, \boldsymbol{\theta}) + \lambda \frac{2}{m} \sum_{i=1}^{\frac{m}{2}} L\left(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), f(\boldsymbol{x}^{(i+\frac{m}{2})}; \boldsymbol{\theta})\right)$$

Logit Pairing Results

MNIST

- For ALP, logit pairing weight just had to be between 0.2 and 1 to see increased robustness
- Used LeNet model and same attack procedure as Madry et al.



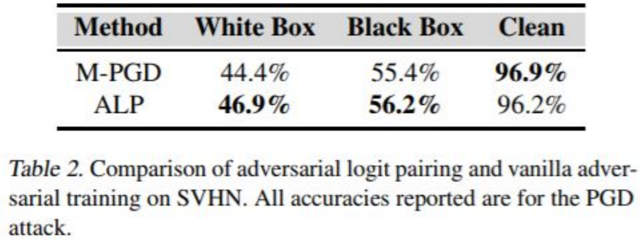
Method	White Box	Black Box	Clean
M-PGD	93.2%	96.0%	98.5%
ALP	**96.4%**	**97.5%**	**98.8%**

Table 1. Comparison of adversarial logit pairing and vanilla adversarial training on MNIST. All accuracies reported are for the PGD attack.

Logit Pairing Results

SVHM

- For ALP, logit pairing weight had to between 0.5 and 1; if too large, then same as vanilla adversarial training
- Used RevNet-9 model



Method	White Box	Black Box	Clean
M-PGD	44.4%	55.4%	**96.9%**
ALP	**46.9%**	**56.2%**	96.2%

Table 2. Comparison of adversarial logit pairing and vanilla adversarial training on SVHN. All accuracies reported are for the PGD attack.

Logit Pairing Results

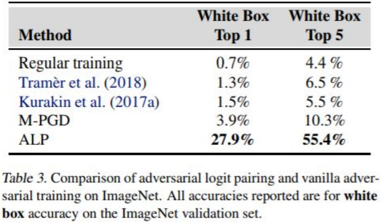
ImageNet: Motivation and Implementation

- Madry et al. demonstrated successful defenses based on multi-step noisy PGD adversarial training on MNIST and SVHM, but not ImageNet
- Total adversarial time roughly proportional to number of attack steps due to needing a full round of backpropagation at each step
- To scale up PGD to ImageNet, implemented synchronous distributed training in Tensorflow with 53 workers (50 for gradient aggregation, 3 for backup replicas)
- Used 17 parameter servers running on CPUs, large batch training, and RMSProp optimizer
- Used InceptionV3 to better compare to Kurakin et al.

Logit Pairing Results

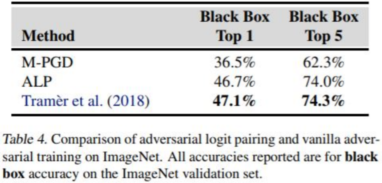
ImageNet: Results Discussion

- Used only targeted attack results since more meaningful than untargeted attacks (which can cause misclassification to a similar class)



Method	White Box Top 1	White Box Top 5
Regular training	0.7%	4.4 %
Tramèr et al. (2018)	1.3%	6.5 %
Kurakin et al. (2017a)	1.5%	5.5 %
M-PGD	3.9%	10.3%
ALP	**27.9%**	**55.4%**

Table 3. Comparison of adversarial logit pairing and vanilla adversarial training on ImageNet. All accuracies reported are for **white box** accuracy on the ImageNet validation set.



Method	Black Box Top 1	Black Box Top 5
M-PGD	36.5%	62.3%
ALP	46.7%	74.0%
Tramèr et al. (2018)	**47.1%**	**74.3%**

Table 4. Comparison of adversarial logit pairing and vanilla adversarial training on ImageNet. All accuracies reported are for **black box** accuracy on the ImageNet validation set.

Logit Pairing Results

ImageNet: Results

- Using the ALP-trained model to create a transfer attack resulted in severe decrease in accuracy in ensemble adversarial training, a recently successful defense strategy (reduce 66.6 to 47.1% accuracy)
- Possibly strong because it came from model with multi-step adversarial training
- Thus, recommend to use adversarial attacks from models adversarially trained with multi-step attacks to simulate strongest black box attacks

Logit Pairing Results

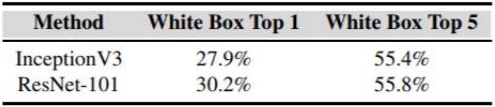
ImageNet: Discussion

- PGD can be scaled to ImageNet
- Multi-step adversarial training shows improvement over previous state-of-the-art
- ALP improves both white and black box accuracy
- ALP achieves state-of-the-art on ImageNet, showing vast improvements over previous state-of-the-art (i.e. ensemble adversarial training)
- This method provides an additional prior which regularizes model towards more accurate understanding of classes

Logit Pairing Results

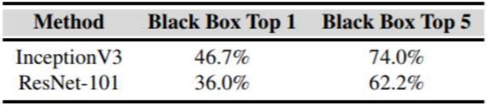
ImageNet: Architecture Comparison

- Also considered ResNet since architectures with higher capacities tend to be more robust



Method	White Box Top 1	White Box Top 5
InceptionV3	27.9%	55.4%
ResNet-101	30.2%	55.8%

Table 5. Comparison of InceptionV3 and ResNet101 on ImageNet. All accuracies reported are for **white box** accuracy on the ImageNet validation set.



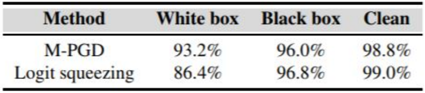
Method	Black Box Top 1	Black Box Top 5
InceptionV3	46.7%	74.0%
ResNet-101	36.0%	62.2%

Table 6. Comparison of InceptionV3 and ResNet101 on ImageNet. All accuracies reported are for **black box** accuracy on the ImageNet validation set.

Logit Pairing Results

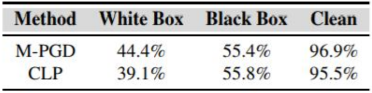
Clean Logit Pairing

- After augmenting images with Gaussian noise, applied either clean logit pairing or logit squeezing
- Logit squeezing competitive with M-PGD on MNIST, despite significantly lower computational cost
- CLP also competitive with M-PGD on SVHM, despite significantly lower computational cost
- Logit squeezing and CLP do not scale with model size, input size, or attack step number like M-PGD does



Method	White box	Black box	Clean
M-PGD	93.2%	96.0%	98.8%
Logit squeezing	86.4%	96.8%	99.0%

Table 7. Comparison of clean logit squeezing and vanilla adversarial training on MNIST. All accuracies reported are for the PGD attack.

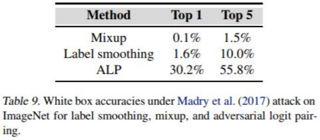


Method	White Box	Black Box	Clean
M-PGD	44.4%	55.4%	96.9%
CLP	39.1%	55.8%	95.5%

Table 8. Clean logit pairing results on SVHN.

Comparision

- Logit pairing is similar to two other approaches: label smoothing and mixup
- Label smoothing (Szegedy et al. 2016) trains classifiers using soft targets rather than hard targets: correct class is given target probability of $1 - \delta$ while the remaining δ is distributed uniformly among remaining classes
- Label smoothing shown to add small amount of robustness to adversarial examples
- Mixup (Zhang et al. 2017) trains the model on input points interpolated between training examples; shown to add some robustness as well



Method	Top 1	Top 5
Mixup	0.1%	1.5%
Label smoothing	1.6%	10.0%
ALP	30.2%	55.8%

Table 9. White box accuracies under Madry et al. (2017) attack on ImageNet for label smoothing, mixup, and adversarial logit pairing.

Comparison

- Also related to a method for semi-supervised leraning, virtual adversarial training (VAT) (Miyato et al. 2017)
- Both ALP and this method have some loss term which minimizes difference between clean and adversarial predictions; VAT, however, used the KL divergence
- ALP consistently performed better than VAT, possibly because KL divergence has the potential to suffer from saturating gradients or because KL is invariant to shift of all logits for a given example

Conclusion and Future Work

Conclusion

- Answered open question about whether or not adversarial training scales to ImageNet
- Introduced adversarial logit pairing as an extension to adversarial training which greatly improves its effectiveness
- Introduced clean logit pairing and logit squeezing, which provides a low cost way to improve adversarial robustness
- Demonstrated that ALP-trained models can generate attacks which significantly damage state-of-the-art defenses and also showed that ALP-trained models achieve state-of-the-art defense against both black and white box attacks

Conclusion and Future Work

Future Work

- Feature matching (like logit matching) may be useful to study
- Possible certification of this method
- Recognize that there may be attacks which break this defense in the future, but note that ALP can be used in conjunction with any future defenses as well (rather than only with adversarial training)

References

- https://arxiv.org/pdf/1803.06373.pdf