Improving Adversarial Robustness via Adversarial attacks

- Shubhankar Joshi

Problem Statement

- Implementation of different attacks such as AutoAttack and PGD,FGSM and defense methods against them.
- The defense method mainly would be the Adversarial Predictive Normalized Maximum Likelihood and would compare it with the Adversarial Training method.

Goals and Ideas

- Goals –
  - Showing Accuracy drop with Auto Attack, PGD and FGSM
  - Implementing the Adversarial pNML defense and comparing with the baseline adversarial training.
- Ideas –
  - Using robust architectures like ResNet and WideResNet as base models
  - Test on MNIST for simplicity and CIFAR10 for complexity
  - Comparing Accuracy before and after defenses.

Plan of Action

- Train Models on MNIST and CIFAR 10 and implement the PGD,FGSM and AutoAttack.
- Adversarial Training of all models on PGD
- Implement the pNML defense and evaluate.

Adversarial Training

- Begin with a standard training phase using clean data to optimize the model for performance.
- Introduce small perturbations into clean data to craft adversarial examples.
- Mix adversarial examples with clean data to create a combined training dataset.
- Re-train the model iteratively using the combined dataset of clean and adversarial examples.
- Optimize the model to learn from both types of data to improve robustness.

Adversarial pNML

1. Taking the test input x
2. Perturbing it slightly towards a specific target label yt using a weak adversarial attack:
3. $x\_refine(x, yt) = x - \lambda \cdot sign(\nabla x\ L(w, x, yt))$
4. Passing this refined input through the pretrained model to get the predicted probability for label yt:
5. $p(yt \mid x\_refine(x, yt))$
6. So for each hypotheses :
    1. Perturbing x towards one possible label yt
    2. Getting the model's predicted probability for that targeted label

Experimental Design

- **Datasets: -**
- MNIST
- CIFAR10

CIFAR10 –
Optimizer- SGD
Resnet ADV
Epochs -100
WideResNet ADV
Epochs - 50

MNIST –
Optimizer – SGD
ADV Epochs - 40

- **Models:**
- ResNet50
  - Pre-trained on ImageNet
  - Fine-tuned on MNIST.
- WideResNet-28-10
  - Trained from scratch on MNIST and CIFAR10
- ResNet18
  - Trained from scratch on CIFAR10.

- **Attacks**:
- FGSM
  - Epsilon 0.03, 0.3
- PGD:-
  - Multi-step attack
  - epsilons: 0.03, 0.3
  - Steps: 10
- Auto Attack: -
  - APGD-CE
  - APGD-DLR

- **Defenses:**
- Adversarial Training
  - Use PGD to generate adversarial examples
- Adversarial pNML
  - Generate adversarial examples targeting each label
  - Predict label with max probability

Results - Attacks

CIFAR 10

Model	CLEAN	APGD CE	APGD D	PGD (0.03), 10	FGSM
ResNeT18	88.9%	0.0 %	0.0 %	0.0 %	3.89 %
Wide ResNet 28-10	89.39 %	12.45 %	12.19 %	0.04 %	22.5 %

MNIST

Model	CLEAN	APGD CE	APGD D	PGD (0.3), 10	FGSM
ResNeT50	99 %	0 %	0 %	22.79 %	30 .56 %
Wide ResNet 28-10	99.10 %	0 %	0 %	4.61 %	7.44 %

Adversarial Training

CIFAR 10

Model	CLEAN	APGD CE	APGD D	PGD (0.03), 10	FGSM
ResNeT18	80.5 %	41.70 %	39.84 %	43.5 %	51.37 %
Wide ResNet 28-10	81.8 %	44.72 %	42.92 %	45.7 %	53.98 %

MNIST

Model	CLEAN	APGD CE	APGD D	PGD (0.3), 10	FGSM
ResNeT50	98.80 %	71.05 %	68.10 %	93.25 %	95.32 %
Wide ResNet 28-10	98.58 %	83.74 %	82. 66 %	94.24 %	96.17 %

Adversarial pNML

CIFAR 10

Model	CLEAN	PGD (0.03), 10 λ= 0.03	FGSM
ResNeT18	79.8 %	62 %	64.3 %
Wide ResNet 28-10	81.2 %	66.75 %	68.7 %

MNIST

Model	CLEAN	PGD (0.3), 10 λ = 0.3	FGSM
ResNeT50	98.92 %	94.6 %	95.4 %
Wide ResNet 28-10	98.63%	94.97 %	95.62%

Conclusion

- Adversarial attacks like FGSM, PGD and AutoAttack significantly reduce accuracy of image classifiers
  - Up to 100% accuracy drop on unsecured ResNet18 and WideResNet models
- Defenses like adversarial training and adversarial pNML boost robustness against attacks
  - Adversarial training recovers 40-50% accuracy under PGD attack
  - pNML recovers 60-68% accuracy under FGSM and PGD attacks on CIFAR-10
- Adversarial training generates adversarial samples during training to improve resilience
- pNML perturbs inputs and predicts labels to be more robust

References

1. Pesso, Uriya & Bibas, Koby & Feder, Meir. (2021). Utilizing Adversarial Targeted Attacks to Boost Adversarial Robustness.
2. Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv (Cornell University)*. http://arxiv.org/pdf/2003.01690.pdf

THANK YOU