Nguyen Thai Ngoc
Summary: Intriguing Properties of Adversarial Training at Scale. (ICLR 2020)

Ciang Xie
Johns Hopkins University

Alan Yuille
Johns Hopkins University

1 Contribution

The paper provides the first rigorous study on diagnosing elements of large-scale adversarial training on ImageNet. It provides two intriguing properties:

First, Batch Normalization (BN) prevents networks from obtaining strong robustness in adversarial training. It relates to the two-domain hypothesis about the distribution of clean images and adversarial images.

Second, adversarial training needs deeper networks to archive higher adversarial robustness. The networks ranges from ResNet-152 to ResNet-638.

2 Related work

Adversarial examples [1] are created by adding human imperceptible perturbation to clean data. It causes the vulnerability of machine learning models. To archive the adversarial robustness, many papers investigates different kinds of adversarial attacks and defenses.

In [2], the authors generated adversarial examples by using Fast Gradient Sign Method (FGSM) and trained models by augmenting them with clean data. The adversarially trained model is robust against adversarial attacks generated by FGSM. The stronger attack was proposed by using a saddle point optimization problem [3]. It is a multi-step variant of FGSM, called Projected Gradient Descent (PGD). The PGD is considered as the strongest first order attack. And the model under such adversarial training is resistant to a wide range of attacks.

Many papers used different techniques to obtain the stronger robustness. [4] used ensemble adversarial training, augmenting training data with adversarial examples crafted on other pre-trained models. This method boosts robustness to black-box attacks. By contrast, it is more vulnerable to white-box attacks. Adversarial logit pairing (APL) [5] explores the logit pairing from two images to be similar. This method improves white-box accuracy on ImageNet. The trade-off on robustness to multiple perturbations (l_p-bounded and spatial perturbation) is explored in [7]. And the feature denoising block is added to convolution neural networks to improve adversarial robustness [8].

3 Results and future work

3.1 Main results

The paper addresses various aspects on evaluation of adversarial robustness [6] for ”deep” neural networks on ImageNet. First observation is the effect of the portion of clean images used for training. Removing clean images from training data boosts adversarial robustness by 18.3% against PGD-2000. ALP technique also outperforms the baseline model [2] by 2.1%.

Batch normalization plays an important role in adversarial robustness. Two-domain hypothesis, clean images and adversarial images may come from two different distributions, suggests two kinds of normalization, BN_{clean} for clean images and BN_{adv} for adversarial images. Enforcing different BNs archives much stronger robustness in different training strategies: 100% adv$+0\%$ clean, 100% adv $+$ 1000% clean, 100% adv $+$ 100% clean, ALP.

Moreover, increasing depth of neural networks significantly improves the model robustness. Their deepest model ResNet-638 outperforms the shallow network ResNet-152 with PGD-30 for training by 4.5% from 42.2% to 46.7% accuracy against PGD-2000.

3.2 Discussion and future work

The paper gives the first outlook on adversarial training at large-scale, ImageNet with ”deep” neural network ResNet. The previous papers [1, 2, 3, 4, 5, 7] consider only small dataset MNIST, CIFAR or shallow neural networks with ImageNet. Moreover, the authors used different batch normalizations for clean and adversarial images which takes advantages of mixture distributions of clean and adversarial images (two-domain hypothesis).

However, taking very ”deep” neural network like ResNet-638 makes the method more expensive and difficult to generalize. Even the theoretical studies argue that robust adversarial models need much more complex classifiers [9]. I think different neural network architectures and regularization methods need to be developed to get robust models.

While the adversarial models are more robust against attacks, it is still vulnerable on clean data. The ResNet-152 trained only with PGD-30 attacks gets 62.1% accuracy for clean images comparing with 77% accuracy in normal training. It is a big gap between adversarial and normal training that one need to find a way to narrow it. [12] proposed Adversarial Propagation (AdvProp) which exploits the complementarity between clean images and their corresponding adversarial examples. The EfficientNet-B8 archives 85.5% top-1 accuracy on ImageNet without any extra data by using the AdvProv. This motivates to find network architectures and regularization methods which make machine learning models more robust against adversarial attacks and more accurate on clean data.

4 Neurocat and toolbox Aidkit

Neurocat is a startup AI company focusing on robustness, comprehensibility, functionality and privacy of machine learning models. Two papers [10, 11] express the deep understanding about adversarial examples of neurocat's members.

Aidkit is an AI quality toolbox designed by Neurocat to evaluate the robustness and comprehensibility of state of the art machine learning models. The paper about adversarial training at large scale will help Adkit to develop techniques to analyse and improve machine learning models.

Deep learning has been successful in computer vision, natural language processing tasks. It has many applications, ranging from autonomous driving over game intelligence to the health care. However, it still lacks rigorous mathematical understanding. I am interested in mathematical aspects in deep neural networks and applying it to improve machine learning models.

Theory of deep learning could be expressed in 4 fundamental elements which I had learned during Summer school on Mathematics of Deep Learning in Berlin last year. The first element is Expressivity which gives the power of the network architecture in representing functions. The second one is Learning process where optimization techniques and regularization methods are used to get reasonable results. Generalization answers the questions why deep neural networks perform well on the test data which does not belong to training set and the impact of the depth in neural networks. The last element is Interpretability which investigates on which components of the input contribute the most and why neural networks could reach a certain decision.

My passion for AI applications and my background in deep learning might be a good fit in Neurocat. I could contribute some aspects about deep understanding of machine learning models which might be useful for toolboxes of Neurocat.

References

[1] Szegedy, C., Zarenba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv: 1312.6199*.

[2] Goodfellow, I., Shlens J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples, In *ICLR'15*.

[3] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Toward deep learning models resistant to adversarial attacks, In *ICLR'18*.

[4] Tramer, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2018). Ensemble adversarial training: attacks and defenses, In *ICLR'18*.

[5] Kannan, H., Kurakin, A., and Goodfellow, I. (2018). Adversarial logit training. *arXiv preprint arXiv: 1803.06373*.

[6] Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. (2019). On evaluation adversarial robustness. *arXiv preprint arXiv: 1902.06705*.

[7] Tramer, F., and Boneh, D. (2019). Adversarial training and robustness for multiple perturbations. In *NeurIPS'19*.

[8] Xie, C., Wu, Y., Maaten, L., Yuille, A., and Kaiming, H. (2019). Feature denoising for improving adversarial robustness. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2019): 501-509.*

[9] Nakkiran, P. (2019). Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv: 1901.00532*.

[10] Assion, F., Schlicht, P., Greßner, F., Grünther, W., Hüger, F., Schmidt, N., and Rasheed, U. (2019). The attack generator: A systematic approach towards constructing adversarial attacks, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2019): 1370-1379.*

[11] Mickisch, D., Assion, F., Schlicht, P., Greßner, F., Grünther, and Motta, M. (2020). Understanding the decision boundary of deep neural networks: An emperical study. *arXiv preprint arXiv: 2002.01810*.

[12] Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A., and Quoc, L. (2020). Adversarial examples improve image recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2020): 819-828.*