Critique Report on the Paper: "Adversarial examples improve image recognition [8]"

Sumaiya Tabassum Nimi

Background Information and Purpose of the Research

Adversarial examples are those image samples that have been generated by adding perturbations to clean images in such a way that Convolutional Neural Network (CNN) models cannot classify them correctly. Besides reducing the classification accuracy of CNN models, adversarial images can be used to attack CNN based applications also. The paper [8] explores the adversarial examples from a point of view that contrasts this traditional outlook. The objective of the research is to use the adversarial examples to learn significant features that would eventually make the deep learning models more general, robust, less prone to overfitting and hence will lead to better classification performance.

Methods Proposed

Training CNN models using adversarial examples is not a novel idea. But the earlier researches, in spite of being successful on small datasets, failed to generalize well on large datasets. In [8], the mismatch of distribution between clean and adversarial examples was identified as the reason behind this failure. The authors argued that this problem of distribution mismatch harmed the batch normalization (BN) [5] calculations most, since BN assumes that the features it normalizes come from similar distribution. In order to address this issue, an auxiliary BN was incorporated into the training algorithm proposed called AdvProp. The main BN normalized the input features for mini-batch of clean images and the auxiliary one normalized those coming from mini-batch of adversarial images. During training, loss was calculated for both these mini-batches of images, each at a time, each using its corresponding BN and the network parameters were updated towards minimizing the total loss.

Strength of the Work

- The main contribution of the paper [8] was the utilization of the adversarial examples as potential regularizers, leading to enhanced generalization ability of deep learning models for classification of images. The proposed training scheme, called AdvProp, is robust and resilient to overfitting. This advantage paved the way towards development of a large network architecture called EfficientNet-B8. The network was trained using the proposed algorithm AdvProp on clean ImageNet [1] samples and corresponding adversarial samples generated using Projected Gradient Descent (PGD) [6] attacker, and achieved state-of-the-art classification performance on the validation set of ImageNet, without using any extra data for training. Whereas prior research works even used billions of extra images during training to get the best classification performance on ImageNet dataset.
- EfficientNet [7] models trained using AdvProp on ImageNet was tested on ImageNet-A [3], ImageNet-C [4] and Stylized-ImageNet [2] datasets also, that consist of challenging ImageNet samples. State-of-the-art accuracy was obtained on all these datasets also, establishing the fact that the

proposed algorithm indeed made the deep learning models more general and robust.

- The solution provided in the proposed training algorithm AdvProp to solve the issue of distribution mismatch was very intuitive and simple.
- Extensive ablation studies were done and reported.

Limitations of the Work

- The proposed training algorithm was used to train only large architectures on a huge dataset, ImageNet. Performance on smaller datasets like CIFAR-10, CIFAR-100 etc. trained using networks with limited capacity like ResNet, Densenet, MobileNet etc. was not reported. At least one such setting should have been experimented and reported.
- The proposed training algorithm is computationally very expensive.

Questions Unanswered

- In Page-6 of the paper it has been stated that "With AdvProp, we observe that smaller networks generally favor weaker attackers". But the networks on which the proposed approach was experimented were still sufficiently large compared to many of the frequently used network architectures. Hence one question naturally arises that was not answered in the paper, will any attacker even work on networks smaller than the ones on which the experiments were done?
- Applicability of the proposed training in class-incremental and few-shot learning settings had not been discussed.
- Also it was not reported if the proposed training algorithm could be used to train on real-life challenging datasets that are most often class-imbalanced and consist of inherently corrupted samples.

Suggested Future Studies

Future works can explore tweaking required to make the proposed training algorithm work better than the traditional training algorithms on networks with limited capacity, possibly in few-shot learning settings.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased

towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[3] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.

[4] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[7] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[8] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.