Review on “Explaining and Harnessing Adversarial Examples”

Authors: I. Goodfellow, J. Shlens, C. Szegedy

# Short Summary

In this paper, the authors present the hypothesis that linear behaviour in high-dimensional spaces is the cause of failure for algorithms prone to adversarial examples. This led to a methodology referred to as the “fast gradient sign method” for quickly generating adversarial examples; with this method, the authors show that adversarial training can act as a regularizer that provides benefits on top of conventional approaches such as dropout.

The intuition behind the fast gradient sign method is that many infinitesimal changes to the input cause the model to attend to these perturbations more strongly than other signals resulting in a large change in the output. The authors found that this works best by scaling the sign of the gradient of the loss function in a model. Using various scaling factors, they were able to produce up to 99.9% error rate with a high confidence 80% on MNIST using a softmax classifier. Similar tests showed that the perturbation consistently reduced classification performance, and that an alternative form of generating adversarial examples could be achieved by rotating the input by a small angle in the direction of the gradient.

Extending their analysis, the authors compared adversarial training with L1 weight decay and additive noise; in both cases, the adversarial training proved to outperform the alternatives in terms of regularization. They achieved the best result on the permutation invariant version of MNIST with an average error rate of 0.782% using different random seeds. This was achieved by using a large model and setting early stopping on the adversarial validation set error. The model, after training, showed that it was somewhat resistant to adversarial examples reducing the error rate on the adversarial set from 89.4% to 17.9%.

Additionally, in the paper, the authors found that it was better to apply perturbations to the original input rather than hidden layers contradicting some previously found results in a previous study. Furthermore, they showed that RBF networks are naturally immune to adversarial examples as they have low confidence when they are fooled. This is suspected to do with the capacity and non-linearity of RBFs.

The authors also hypothesize why adversarial examples generate the same misclassifications across networks irrespective of where they were generated. Their first proposal is that these adversarial examples exist in a contiguous regions of a 1-D subspace and hence are always misclassified. Another is that classifiers are trained to approximate functions and generalize well; correspondingly, the stability in their classification weights makes them misclassify adversarial examples similarly.

Finally, the authors provide counter-points to two hypothesis about adversarial training. First, they show that generative models do not perform better with adversarial examples because they are intrinsically trained to separate real and fake. Secondly, they show that an ensemble of methods does not improve performance on adversarial examples and thus there is not an issue with individual model quirks.

# Main Contributions

* Proposed the fast gradient sign method for generating adversarial examples
* Showed evidence to support their hypothesis of adversarial weakness coming from an over-reliance on linearity
* Showed that adversarial perturbations rely on the direction rather than location in space
* Showed that adversarial training can result in regularization
* Discussed applying perturbation to input vs hidden layers
* Showed that generative and ensemble methods do not resolve problems with adversaries

# High-Level Evaluation of Paper

This paper read more like a text-book than a research paper. However, it did present theories and support them with evidence. Furthermore, the authors did a good job in addressing alternative hypothesises and presented solid counter-examples. Overall, the paper was clearly written and easy to understand. I do wish, however, that results were presented in organized tables rather than in the body of the text. It’s harder to compare approaches with the way the evidence is presented.

Additionally, I believe that the authors presented one possible reason why models struggle with adversarial examples (overreliance on linearity). There could be other obfuscated factors that may have been missed in the scope of the paper. Furthermore, a lot of their experimentation is performed using adversarial examples generated with their method. It’s unclear to me whether their adversarial training would still be as effective if some other form of adversary was presented.