Week 9 paper summary

Sina Mehdinia

05/29/2021

Summary

Goodfellow et al. [1] explore adversarial examples in this work. First, they show that the reason for this phenomena is not high non-linearity but linearity itself. They argue that with our current optimization and simple nonlinearities like ReLU, our networks are still highly linear and that is why our networks are predicting high confidence for perturbed examples (such as 99% error). They mathematically show that small perturbations can grow through linearity if the input has enough dimensionality. They present a method to train adversarial examples and they show that their proposed fast gradient methods can have more regularizing effect than dropouts. It is simply sign of the gradient of the loss with respect to input multiplied by an epsilon. They also show that adding noise and training or ensembling cannot help with adversarial examples. Another important fact was that many different architectures are making the same mistake in misclassifying adversarial examples. The authors showed this can be due to the fact that with current methodologies, all function approximators resemble a linear classifier (learning similar functions). This paper shows that our ML models, with current optimization methods, even though having a very good training and testing accuracy, are not maybe learning the true concepts.

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.