Explaining and Harnessing Adversarial Examples

Ravi Raju

12-16-2018

1 Motivation

This is a continuation of why deep models fail against adverserial examples. These are transferable across model architectures. There are many speculations about what cause these cases such as nonlinearity of deep models or overfitting. However, in this paper, it was shown that the problem is the linear nature of the classifer. Goodfellow et al. created the Fast Gradient Sign Method (FGSM) as a way to quickly create adverserial examples using the sign of the gradients to perturb the image.

2 Solution

The reason why these models behave linearly is that they operate in the linear regions of the activation functions. This is because functions that are in the linear part of the activation are easy to optimize. Look at Section 3 in the paper for an explanation: Goodfellow makes it very clear that $w^T\eta$ can grow with size n, the dimensionality. The authors suggest using an adversarial objective function based on FGSM.

3 Results

Even with the FGSM regularizer, the model still misclassified 17.9% on a maxout architecture. The best regularization effect was an additive perturbation at the input layer.

4 Discussion/Takeaway

They discuss RBF networks that have the form, $p(y = 1|x) = exp((x-\mu)^T\beta(x-\mu))$, which are notably resistant to adverserial examples but do not generalize well. Another important takeaway is from Figure 4. This is highlighting that the previous proposal that adverserial examples exist in fine pockets is not the case. Rather, they are ubiquitous in the realm of linear subspaces. The direction

need only have positive dot product with the gradient of the cost function, and ϵ need only be large enough.