Paper

- **Title:** Explaining and Harnessing Adversarial Examples
- **Authors:** Ian Goodfellow, Jonathan Shlens, Christian Szegedy
- **arXiv link:** https://arxiv.org/abs/1412.6572

TL;DR

Adversarial examples are a limitation of models that are linear and easy to train, they do not exist due to overfitting. Adversarial perturbations generalize across different models. Fast Gradient Sign Method(FGSM) can generate perturbed examples easily and can be used for regularization. Radial Bias Function (RBF) networks resist against adversarial perturbations.

The Linear Explanation

Let η be the perturbation added to the image x i.e. $\widetilde{x} = x + \eta$. The constraint on η is $||\eta||_\infty < \epsilon$, where ϵ is so small that a data storage would discard due to limited precision. So, x and $\widetilde{x}$ should have the same response from a classifier. Consider,

$$w^T \widetilde{x} = w^T x + w^T \eta$$

The increase in the activation is $w^t \eta$. To maximize the increase, consider $\eta = \epsilon \cdot sign(w)$. If w has n dimensions and the average magnitude of it's element is m, then the activation increases by ϵmn. As n increases, $||\eta||_\infty$ does not increase, but the activation increases linearly. Using this linear growth, one can make many infinitesimal changes to the input that add up to one large change to the output.

Fast Gradient Sign Method

Let θ be the parameters of a model, x the input, y the target of the input and $J(\theta, x, y)$ be the cost used to train the network.

$$\eta \;=\; \epsilon sign(\nabla_x J(\theta, x, y))$$

This perturbation is added to the input image and is called the Fast Gradient Sign Method. It is very easy to calculate this using backprop.

Adversarial training as Regularization

The authors noted that training with an adversarial objective function based on the fast gradient sign method was an effective regularizer:

$$\widetilde{j}(\theta, x, y) = \alpha J(\theta, x, y) + (1 - \alpha) J(\theta, x + \epsilon sign(\nabla_x J(\theta, x, y)), y)$$

The model trained this way shows some resistance to adversarial images. Consider 2 identical models, one(A) trained on adversarial objective and the other one(B) normally. A is more resistant to adversarial images generated from B than B is to adversarial images generated from A.

RBF networks

Shallow RBF networks with

$$p(y = 1|x) = exp((x - \mu)^t \beta(x - \mu))$$

are only able to confidently predict that the positive class is present in the vicinity of μ. Elsewhere, they default to predicting the class is absent, or have low-confidence predictions. RBF networks are naturally immune to adversarial examples, in the sense that they have low confidence on adversarial inputs.

But RBF units are not invariant to any significant transformations so they cannot generalize very well.

Conclusion

- Adversarial examples can be explained as a property of high-dimensional dot products. They are a result of models being too linear, rather than too nonlinear.
- The generalization of adversarial examples across different models can be explained as a result of different models learning similar functions when trained to perform the same task.
- The direction of perturbation, rather than the specific point in space, matters most.
- FGSM is a simple method to fool very deep networks.
- RBF networks are resistant to adversarial examples.
- Models that are easy to optimize are easy to perturb.