Adversarial Spheres

J. Gilmer[1], L. Metz[1], F. Faghri[2], S.S. Schoenholz[1], M. Raghu[1], M. Wattenberg[1], I. Goodfellow[1]

[1]Google Brain [2]University of Toronto

arXiv:1801.02774

Reviewed by : Bill Zhang
University of Virginia
https://qdata.github.io/deep2Read/

Outline

Introduction

Basic Premise and Motivation

- Many standard image models correctly classify randomly chosen images, but they are usually visually similar to an incorrectly classified image
- Hypothesize that this behavior is a natural result of the high dimensional nature of data manifold
- To investigate, study classification between two high dimensional spheres

Concentric Spheres Dataset

- Data distribution is two concentric spheres in d dimensions
  - Generate random $x \in \mathbb{R}^d$ with $||x||_2$ either 1 or R with equal probability and target y
  - If $||x||_2 = 1, y = 0$; if $||x||_2 = R$, $y = 1$
- Key advantages of concentric spheres
  - Probability density of data $p(x)$ is well defined and uniform across x; can sample uniformly by taking $z \sim (\vec{0}, I)$ and setting $x = z/||z||_2$ or $x = Rz/||z||_2$
  - There is a theoretical max margin boundary which perfectly separates two classes, the sphere with radius $(R+1)/2$
  - Can create machine learning models which can learn a decision boundary to separate the two spheres
  - Difficulty can be controlled by varying d and R
- R was arbitrarily set to 1.3, model trained online ($N = \infty$) and with fixed training set size N

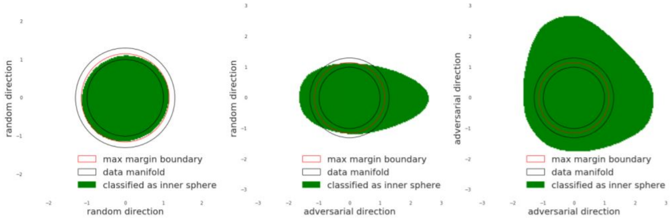
Adversarial Examples for Deep ReLU

- Experiment 1: Set $d = 500$, train a 2 hidden layer ReLU with 1000 hidden units, train with minibatch SGD on sigmoid cross entropy loss, use Adam optimizer; Online training with batch size 50 and 1 million training points
  - Evaluate on 10 million uniform samples from each sphere: no errors, so error rate is unknown with only a statistical upper bound
  - Despite lack of error, can find adversarial errors on data manifold using gradient descent (manifold attack)
  - Worst-case example: reiterate attack until convergence (not around starting point); NN example: Terminate attack on first misclassfication
  - These errors are typically close to randomly sampled points on sphere; the L2 distance is around 0.18 compared to average distance between 2 random points, 1.41

Adversarial Examples for Deep ReLU

Visualization

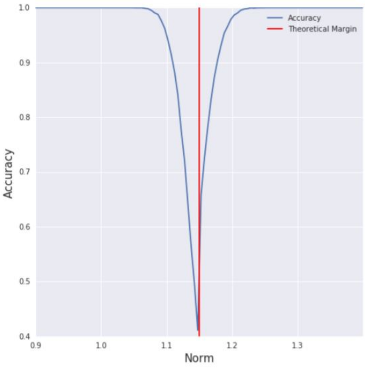
- Visualize decision boundary by taking 2d projections of 500 dimensional space; model naturally interpolates between two spheres
- Take projections of random, one basis in worst-case adversarial example, two basis of separate worst-case examples
- This only occurs when spheres are high dimensional; highest dimension without error is around $d = 60$

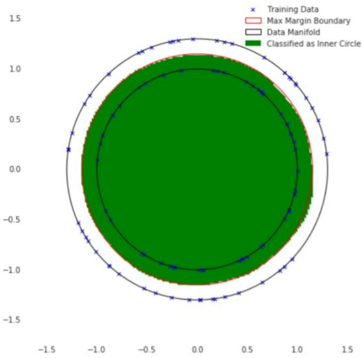


Adversarial Examples for Deep ReLU

Visualization

- Plot accuracy as points approach decision boundary; although no errors are made far from boundary, adversarial examples can be found as far as 0.6 and 2.4 norm
- Also show manifold for $d = 2$; no errors in classification





- Want to test if adversarial errors are off of the data manifold
- Traditional attacks start with input x and target $\hat{y}$ and finds an input $\hat{x}$ which maximizes $P(\hat{y}, \hat{x})$ given the constraint $||x - \hat{x}|| < \epsilon$
- Instead, use constraint $||\hat{x}||_2 = ||x||_2$ to ensure that adversarial example is of same class as starting point
- Solve this constraint problem using PGD, except when projecting, except on projection step normalize $||x||_2$ by projecting back onto the sphere; this makes it so that $p(x) = p(x_{adv})$

Simple Network Analysis

- Difficult to reason about ReLU decision boundary, so study a simpler model, "the quadratic network"
- Single hidden layer where pointwise non-linearity $\sigma(x) = x^2$; no bias in hidden layer
- Output sums hidden activations, multiplies by scalar, and adds bias
- With hidden dimension h, there are $dh + 2$ trainable parameters
- Logit is of following form where $W_1 \in \mathbb{R}^{hxd}$, $\vec{1}$ is a column vector of h 1s, w and b are learned scalars

$$\hat{y}(x) = w\vec{1}^T(W_1 x)^2 + b$$

Simple Network Analysis

- Through derivations, arrive at alternate form for logit where α_i are scalars which depend on model parameters and $\vec{z}$ is a rotation of input $\vec{x}$

$$\hat{y}(x) = \Sigma_{i=1}^{d} \alpha_i z_i^2 - 1$$

- Decision boundary is where $\Sigma_{i=1}^{d} \alpha_i z_i^2 = 1$, a d dimensional ellipsoid
  - $\alpha_i > 1 \Rightarrow$ errors on inner sphere
  - $\alpha_i < 1/R^2 \Rightarrow$ errors on outer sphere
  - Model has perfect accuracy iff all $\alpha_i \in [1/R^2, 1]$

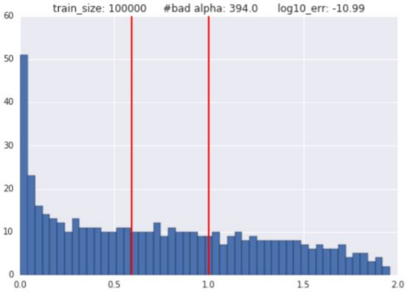
Simple Network Analysis

- Train quadratic network with $h = 1000$
  - With online training, model has perfect accuracy
  - If we have $N = 10^6$ points from $p(x)$ as training set, model has empirically low error rate (no errors from 10 million randomly sampled tests), but there are adversarial examples: 394 of 500 learned α_i are not in range
- Use CLT to estimate error of network from α_i to be around 10^{-11}
- Next, augment previous setup with all α_i within range and non-zero gradients
  - As model is trained, worst case loss increases, average case loss decreases
  - Reflects how training objective does not directly measure accuracy and also how high dimensional data may have divergent losses

Simple Model Analysis

Visualization

- Left: Distribution of α_i for $N = 10^6$
- Right: Training curves of model with perfect initialization



Simple Model Analysis

CLT Approximation

- Suppose z is chosen from inner sphere, then we want to compute the probability that $\Sigma_{i=1}^{d} \alpha_i z_i^2 > 1$
- Generate z uniformly on inner sphere by picking $u_i \sim N(0,1)$ and let $z_i = u_i / ||u||$
- Previous equation can be rewritten

$$\frac{1}{||u||} \Sigma_{i=1}^{d} \alpha_i u_i^2 > 1$$

$$\Sigma_{i=1}^{d} \alpha_i u_i^2 > \Sigma_{i=1}^{d} u_i^2$$

$$\Sigma_{i=1}^{d} (\alpha_i - 1) u_i^2 > 0$$

Simple Model Analysis

CLT Approximation

- Let $X = \Sigma_{i=1}^{d}(\alpha_i - 1)u_i^2$: if d sufficiently large, can use CLT to conclude that $X \sim N(\mu, \sigma^2)$
- Can compute μ since $E[u_i^2] = \sigma_{u_i}^2 = 1$

$$\mu = E[x] = \Sigma_{i=1}^{d}(\alpha_i - 1)$$

- Can compute σ^2 too

$$\sigma^2 = Var[X] = 2\Sigma_{i=1}^{d}(\alpha_i - 1)^2$$

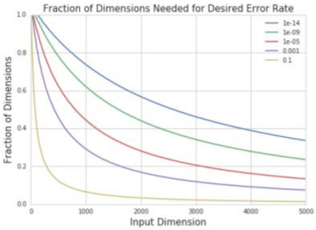
- Therefore,

$$P(X > 0) = P(\sigma Z + \mu > 0) = P(Z > -\frac{\mu}{\sigma}) = 1 - \Phi(-\frac{\mu}{\sigma})$$

Simple Model Analysis

CLT Approximation

- As long as $E[\alpha_i] \approx (1 + R^{-2})/2$ and variance is not too large, model will be extremely accurate
- Flexibility with choices of α_i increases with dimension
- Using approximation, plot fraction of dimension needed to achieve target error rate (0.5 fraction when $d = 2000$ implies 1000 hidden nodes); model size to get 0 error may be siginificantly larger than size to get small error



Local Adversarial Examples

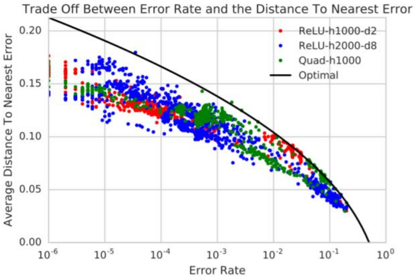
Theorem

- Attempt to explain why local adversarial examples exist for sphere dataset; do not attempt to relate sphere data to natural image manifolds
- Define terms:
  - S_0 is sphere of radius 1 in d dimensions
  - $E \subseteq S_0$ is set of all misclassified points by some model
  - For $x \in S_0$, let $d(x, E)$ denote the L2 distance between x and nearest point in E
  - Let $d(E) = E_{x \sim S_0} d(x, E)$
  - Let $\mu(E)$ denote E as a fraction of S_0
- Theorem: Consider any model trained on sphere dataset. Let $p \in [0.5, 1.0)$ be accuracy of model on inner sphere and E be the set of misclassified points ($\mu(E) = 1 - p$). Then, $d(E) = O(\Phi^{-1}(p)/d)$.

Local Adversarial Examples

Theorem Implications

- Links probability of error with average error distance independent of model
- Any model which misclassifies a small constant fraction of the sphere must have errors close to randomly sampled points
- There exists a optimal tradeoff between generalization accuracy and average distance to nearest error; train on 2 ReLU and 1 Quadratic model to test



Summary

- Concentric spheres dataset exhibit similar phenomenon to natural images: most randomly selected points are correctly classified but are close to a misclassified point
- Explain phenomenon for spheres by proving a theoretical tradeoff between error rate and average distance to nearest error of a model; show that variety of architectures match this bound
- Theorem reduces question from "why are there adversarial examples?" to "why is there a small amount of classification error?"; unclear whether this would hold for natural images as well
- Raises question of whether it is possible to solve adversarial problem given limited data; network size required to create perfect model may be significantly larger than what is needed to achieve small classification error

References

- https://arxiv.org/pdf/1801.02774.pdf