Neighborhoods of Adversarial Examples

Brian Bell : University of Arizona

February 3, 2023

Contents

Part I
Persistence

1 Artificial Neural Networks ANNs

2 Adversarial Attacks

3 Defining adversarial attacks mathematically

4 Persistence metric for Adversarial Attacks

5 Persistence results

6 Introduction

Artificial Neural Networks and other optimization-based general function approximation models are the core of modern machine learning [Prakash et al.(2018)Prakash, Moran, Garber, DiLillo, and Storer]. These models have dominated competitions in image processing, optical character recognition, object detection, video classification, natural language processing, and many other fields [Schmidhuber(2015)]. All such modern models are trained via gradient-based optimization, e.g. Stochastic Gradient Descent (SGD) with gradients computed via back propagation. [Goodfellow et al.(2013)Goodfellow, Bulatov, Ibarz, Arnou
Although the performance of these models is practically miraculous within the training and testing context for which they are designed, they have a few intriguing properties. It was discovered in 2013 [Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus] that images can be generated which apparently trick such models in a classification context in difficult-to-control ways [Khoury and Hadfield-Menell(2018)]. The intent of this research is to investigate these *adversarial examples* in a mathematical context and use them to study pertinent properties of the learning models from which they arise.

7 Artificial Neural Networks (ANNs)

The history of Neural Networks begins in the field of Theoretical Neuropsychology with a much-cited paper by McCulloch and Pitts in which they describe the mechanics of cognition in the context of computation [McCulloch and Pitts(1943)]. This initial framework for computational cognition did not include a notion for learning, but the following decade brought the concept of learning (as optimization) and many simple NNs (linear regression models applied to computational cognition). The perceptron, the most granular element of a neural network, was proposed in another much-cited paper by Rosenblatt in 1958 [Rosenblatt(1958)], and multilevel (deep) networks were proposed by 1965 in a paper by Ivakhnenko and Lapa [Ivakhnenko and Lapa(1965)].

By the 1960s, these neural network models became disassociated from the cutting-edge of cognitive science, and interest had shifted to their application in modeling and industrial computation. The hardware limitations of the time served as a significant barrier to wider application and the concept of the "neural network" was sometimes treated as a solution looking for a problem. Compounding these limitations was a significant roadblock published by Minsky and Papert in 1969: A proof that basic perceptrons could not encode exclusive-or [Minsky and Papert(1969)]. As a result, interest in developing neural network theory waned. The next necessary step in the development of modern neural network models was an advance that would allow them to be trained efficiently with computing power available. Learning methods required a gradient, and the technique necessary for computing gradients of large-scale multi-parameter models was apparently proposed in a 1970 in a Finnish masters thesis [Linnainmaa(1970)]. Techniques from control theory were applied to develop a means of propagating error backward through models which could be described as directed graphs. The idea was applied to neural networks by a Harvard student Paul Werbos[Werbos(1974)] and refined in later publications.

The final essential puzzle piece for neural network models was to take advantage of their layered structure, which would allow backpropagation computations at a given layer to be done in parallel. This key insight, indeed the core of much of modern computing, was a description of parallel and distributed processing in the context of cognition by Rumelhard and McClelland in 1986 [McClelland et al.(1986)McClelland, Rumelhart, Group, et al.] with an astonishing 22,453 citations (a number that grows nearly every day). With these pieces in place, the world was ready for someone to finally apply neural network models to a relevant problem. In 1989, Yann LeCun and a group at Bell Labs managed to do just that. LeCun refined backpropagation into the form it is used today [LeCun et al.(1989)LeCun, Boser, Denker, Henderson, Howard, Hubbard, and Jackel], invented Convolutional Neural Networks 7.2 [LeCun et al.(1995)LeCun, Bengio, et al.], and by 1998, he had worked with his team to implement what has become the industry standard for banks to recognize hand-written numbers on checks [LeCun et al.(1998)LeCun, Bottou, Bengio, Haffner, et al.].

Starting in the 2000s, neural networks have blown up in scale and application, so it's harder to keep track of the discrete historical developments. Most of the progress in terms of performance and application has come from contests (e.g: `http://image-net.org/challenges/LSVRC/`) in which a bounty (or publicity) is offered and labs around the world compete to build a network which solves the competition's problem. Schimdhuber's group [Schmidhuber(2015)], and a similar group at Google have dominated many of these competitions. The cutting-edge today is represented by networks like Inception v4 designed by Google for image classification which contains approximately 43 million parameters [Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Early versions of this network took 1-2 million dollars worth of compute-time to train. ANNs now appear in nearly every industry from devices which use ANNs to intelligently adapt their performance, to the sciences which rely on ANNs to eliminate tedious sorting and identification of data that previously had to be relegated to humans.

7.1 Structure

In this section we give a mathematical description of artificial neural networks.

A *neuron* is a nonlinear operator that takes input in $\mathbb{R}^n$ to $\mathbb{R}$, historically designed to emulate the activation characteristics of an organic neuron. A collection of neurons that are connected via a (usually directed) graph structure are known as an *Artificial Neural Network (ANN)*.

The fundamental building blocks of most ANNs are artificial neurons which we will refer to as *perceptrons*.

Definition 7.1. *A **Perceptron** is a function $P_{\vec{w}} : \mathbb{R}^n \to \mathbb{R}$ which has* weights *$\vec{w} \in \mathbb{R}^n$ corresponding with each element of an input vector $\vec{x} \in \mathbb{R}^n$ and a bias $b \in \mathbb{R}$:*

$$P_{\vec{w}}(\vec{x}) = f((\langle \vec{w}, \vec{x} \rangle + b)$$

$$P_{\vec{w}}(\vec{x}) = f\left(b + \sum_{i=1}^{n} w_i x_i\right)$$

*where $f : \mathbb{R} \to \mathbb{R}$ is continuous. The function f is called the **activation function** for P.*

The only nonlinearity in P_w is contained in f. If f is chosen to be linear, then P will be a linear operator. Although this has the advantage of simplicity, linear operators do not perform well on nonlinear problems like classification. For this reason, activation functions are generally chosen to be nonlinear. Historically, heaviside functions were used for activation, later replaced with sigmoids [Malik and Perona(1990)] for their smoothness, switching structure, and convenient compactification of the output from each perceptron. It was recently discovered that a simpler nonlinear function, the *Rectified Linear Unit (ReLU)* works as well or better in most neural-network-type applications [Glorot et al.(2011)Glorot, Bordes, and Bengio] and additionally training algorithms on ReLU activated networks converge faster [Nair and Hinton(2010)].

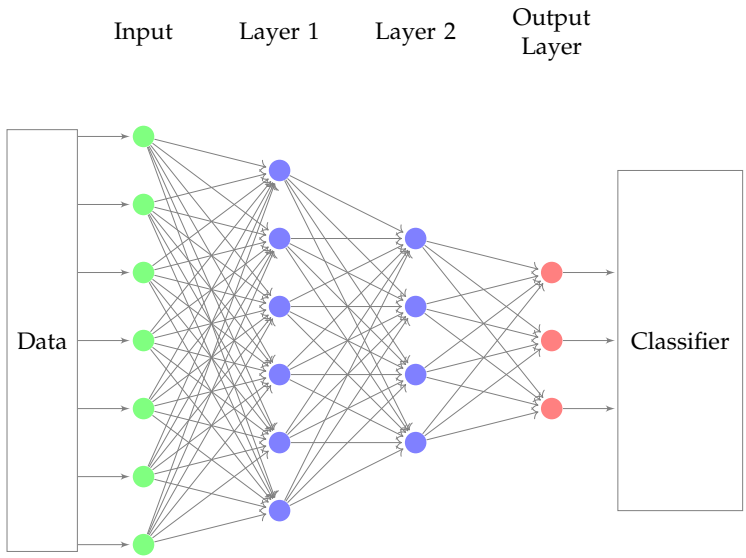
Definition 7.2. *The Rectified Linear Unit (ReLU) function is*

$$ReLU(x) = \begin{cases} 0, & x \leq 0; \\ x, & x > 0, \end{cases}$$

The single nonlinearity of this activation function at $x = 0$ is sufficient to guarantee existence of ϵ approximation of smooth functions by an ANN composed of sufficiently numerous perceptrons connected by ReLU [Petersen and Voigtlaender(2018)]. In addition, ReLU is convex, which enables efficient numerical approximation of smooth functions in shallow networks [Li and Yuan(2017)].

In general ANNs must not be cyclic and, for convenience, are often arranged into independent layers. An early roadblock for neural networks was a proof by Minsky [Minsky and Papert(1969)] that single layers of perceptrons could not encode exclusive-or. Depth, the number of layers in a neural network, is a key factor in its ability to approximate complicated functions including exclusive-or [Kak(1993)]. For this reason, modern ANNs are usually composed of many layers (3-100). The most common instance of a neural network model is a fully connected *feed forward (FF)* configuration. In this configuration data enters as an input layer which is fed into each of the nodes in the first layer of neurons. Output of the first layer is fed into each of the nodes in the second layer, and so on until the output of the final layer is fed into an output filter which generates the final result of the neural network.

In this example of a FF network, an input vector in $\mathbb{R}^7$ is mapped to a an output in $\mathbb{R}^3$ which is fed into a classifier. Each blue circle represents a perceptron with the ReLU activation function.



The output of this ANN is fed into a classifier. To complete this example, we can define the most common classifier, Softmax:

Definition 7.3. *Softmax (or the normalized exponential) is the function given by*

$$s : \mathbb{R}^n \to [0,1]^n$$

$$s_j(\vec{x}) = \frac{e^{x_j}}{\sum_{k=1}^{n} e^{x_k}}$$

Definition 7.4. *We can define a classifier which picks the class corresponding with the largest output element from Softmax:*

$$\textit{(Output Classification)}\ c_s(\vec{x}) = argmax_i s_i(\vec{x})$$

During training, the output $y \in \mathbb{R}^n$ from a network can thus be compressed using softmax into $[0,1]^n$ as a surrogate for probability for each possible class or directly into the classes which we can represent as the simplex for the vertices of $[0,1]^n$.

[Bishop(2006)]

7.2 Convolutional Neural Networks (CNNs)

Another common type of neural network which is a component in many modern applications including one in the experiments to follow are Convolutional Neural Networks (CNNs). CNNs are fundamentally composed of perceptrons, but each layer is not fully connected to the next. Instead, layers are arranged spatially and overlapping groups of perceptrons are independently connected to the nodes of the next layer, usually with a nonlinear filter that computes the maximum of all of the incoming nodes to a new node. This structure has been shown to be very effective on problems with spatial information [LeCun et al.(1995)LeCun, Bengio, et al.].

8 Training ANNs

Neural networks consist of a very large number of perceptrons with many parameters. Directly solving the system implied by these parameters and the empirical risk minimization prob-

lem defined below would be difficult, so we must use a modular approach which takes advantage of the simple and regular structure of ANNs.

A breakthrough came with the application of techniques derived from control theory to ANNs in the late 1980s [Rumelhart et al.(1986)Rumelhart, Hinton, and Williams], dubbed back-propagation. This technique was refined into its modern form in the thesis and continuing work of Yann LeCun [LeCun et al.(1988)LeCun, Touresky, Hinton, and Sejnowski]. In this method, error is propagated backward taking advantage of the directed structure of the network to compute a gradient for each parameter defining it. Because modern ANNs are usually separated into discrete layers, gradients can be computed in parallel for all perceptrons at the same depth of the network [Bishop(2006)]. Leveraging modern GPUs and parallel computing technologies, these gradients can be computed very quickly. There are a number of important considerations in training. We discuss a few in the following sections.

8.1 Selection of the Training Set

The first step in training an ANN is the selection of a training set. ANNs fundamentally are universal function approximators: Given a set of input data and corresponding output data, they approximate a mapping from one to the other. Performance is dependent on how well the phenomenon we hope to model is represented by the training data. The training data must consist of a set of inputs (e.g., images) and a set of outputs (e.g., labels) which contain sufficient examples to characterize the intended model. In a way, this is how we pose a question to the neural network. One must always ask whether the question we wish to pose is well-expressed by the training data we have available.

The most important attributes of a training dataset are the number of samples it contains and its density near where the model will be making predictions. According to conventional wisdom, training a neural network with K parameters will be very challenging if there are fewer than K training samples available. The modular structure of ANNs can be combined with regularization of the weights to overcome these limitations [Liu and Deng(2015)]. In general, we will denote a training set by (X, Y) where X is an indexed set of inputs and Y is a corresponding indexed set of labels.

8.2 Selecting a Loss Function

Once we have selected a set of training data (both inputs and outputs), we must decide how we will evaluate the match between the ANNs output and the defined outputs from the training dataset – we will quantify the deviation of the ANN compared with the given correspondence as a Loss. In general *loss functions* are nonzero functions which compare an output y against a ground-truth $\hat{y}$. Generally they have the property that an ideal outcome would have a loss of 0.

One commonly used loss function for classification is known as Cross-Entropy Loss:

Definition 8.1. *The Cross-Entropy Loss comparing two possible outputs is* $L(y, \hat{y}) = -\sum_i y_i \log \hat{y}_i$.

Other commonly used loss functions include L^1 loss (also referred to as Mean Absolute Error (MAE)), L^2 loss (often referred to as Mean-Squared-Error (MSE)), and Hinge Loss (also known as SVM loss).

To set up the optimization, the loss for each training example must be aggregated. Generally, ANN training is conducted via Empirical Risk Minimization where Empirical Risk is defined for a given loss function L as follows:

Definition 8.2. *Given a loss function L, the Empirical Risk over a training dataset (X, Y) of size N is*

$$R_{emp}(P_{\vec{w}}(x) = \frac{1}{N} \sum_{(x,y)\in(X,Y)} L(P_{\vec{w}}(x)), y).$$

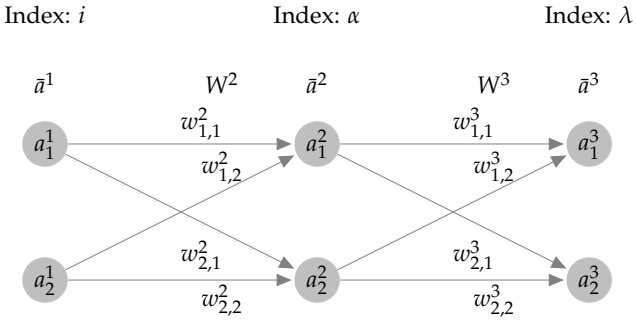
We seek parameters $\vec{w}$ which will minimize $R_{\text{emp}}(P_w(x))$. This will be done with gradient-based optimization.

8.3 Computation of Gradient via Backpropagation

Since it is relevant to the optimization being performed, we will briefly discuss the computation of gradients via backpropagation. For this discussion, we will introduce a small subset of a neural network in detail. In general, terms will be indexed as follows:

$$x^{[\text{layer}]}_{[\text{node in layer}],\,[\text{node in previous layer}]}$$

When the second subscript is omitted, the subscript will only index the node in the current layer to which this element belongs.



In this diagram, the W^l are matrices composed of the weights indexed as above. Given an activation function for layer n A^n and its element-wise application to a vector $\bar{A}^n$, we can now write the output $\bar{a}_n$ for any layer of an arbitrary ANN in two ways [Krause(2020)]. Recursively, we can define

$$a^n_\lambda = A^n(\sum_\alpha w^n_{\alpha,\lambda} a^{n-1}_\alpha) \tag{1}$$

We can also write the matrix form of this recursion for every node in the layer:

$$\bar{a}^n = \bar{A}^n(W^n(\bar{a}^{n-1})) \tag{2}$$

The matrix form makes it easier to write out a closed form for the output of the neural network.

$$\bar{a}^n = \bar{A}^n(W^n(\bar{A}^{n-1}(W^{n-1}(\cdots(\bar{A}^2(W^2\bar{a}^1))\cdots)))) \tag{3}$$

Now, given a loss function $L = \sum_i \ell_i(a^n_i)$ where each ℓ_i is a loss function on the i^{th} element of the output, we wish to compute the derivatives $\dfrac{\partial L}{\partial_{w^l_{i,j}}}$ for every l, i, and j which compose the gradient ∇L. Using the diagram above, we can compute this directly for each weight using chain rule:

$$\frac{\partial L}{\partial w^3_{\lambda,\alpha}} = \frac{\partial L}{\partial a^3_\lambda}\frac{\partial a^3_\lambda}{\partial w^3_{\lambda,\alpha}} = \sum_{\lambda=1}^{n} \ell'_\lambda(a^3_\lambda)A'^3(\sum_{\alpha=1}^{n} w^3_{\alpha,\lambda}a^2_\alpha)a^2_\alpha$$

Many of the terms of this gradient (e.g. the activations a_i^n and the sums $\sum_i w_{i,j}^n a_i$) are computed during forward propagation when using the network to generate output. We will store such values during the forward pass and use a backward pass to fill in the rest of the gradient. Furthermore, notice that ℓ'_λ and A'^n are well understood functions whose derivatives can be computed analytically almost everywhere. We can see that all of the partials will be of the form $\frac{\partial L}{\partial w_{n,i}^l} = \delta_n^l a_i^l$ where δ_n^l will contain terms which are either pre-computed or can be computed analytically. Conveniently, we can define this error signal recursively:

$$\delta_n^l = A'^l(a_n^l) \sum_{i=1}^{n} w_{i,n}^{l+1} \delta_i^{l+1}$$

In matrix form, we have

$$\bar{\delta}^l = \bar{A}'^l(W^l \bar{a}^l) \odot ((W^{l+1})^T \bar{\delta}^{l+1}$$

Where $\odot$ signifies element-wise multiplication.

Then we can compute the gradient with respect to each layer's matrix W^l as an outer product:

$$\nabla_{W^l} L = \bar{\delta}^l \bar{a}^{(l-1)T}$$

Since this recursion for layer n only requires information from layer $n+1$, this allows us to propagate the error signals that we compute backwards through the network.

8.4 Optimization of Weights

Given a set of training input data and a method for computing gradients, our ultimate goal is to iteratively run our training-data through the network, updating weights gradually according to the gradients computed by backpropagation. In general, we start with some default arrangement of the weights and choose a step size η for gradient descent. Then for each weight, in each iteration of the learning algorithm, we apply a correction so that

$$w'_{i',j',k'} = w_{i',j',k'} - \eta \frac{\partial E(Y, \hat{Y})}{\partial w_{i',j',k'}}$$

In this case, the step size (learning rate) η is fixed throughout training. Numerical computation of the gradient requires first evaluating the network forward by computing the output for a given input. The value of every node in the network is saved and these values are used to weight the error as it is propagated backward through the network. Once the gradient is computed, the weights are adjusted according to the step defined above. This process is repeated until convergence is attained to within a tolerance. It should be clear from the number of terms in this calculation that the initial guess and step size can have significant effect on the eventual trained weights. Due to lack of a guarantee for general convexity, poor guesses for such a large number of parameters can lead to gradients blowing up or down [Bishop(2006)]. . Due to nonlinearity and the plenitude of local minima in the loss function, classic gradient descent usually does not perform well during ANN training.

By far the most common technique for training the weights of neural networks adds noise in the form of random re-orderings of the training data to the general optimization process and is known as stochastic gradient descent.

Definition 8.3. *Stochastic Gradient Descent (SGD)*

Given an ANN $N : \mathbb{R}^n \to C$, an initial set of weights for this network $\vec{w}_0$ (usually a small random perturbation from 0), a set of training data X with labels Y, and a learning rate η, the algorithm is as follows:

Batch Stochastic Gradient Descent

$w = w_0$

while $E(\hat{Y}, P_w(X))$ (cumulative loss) is still improving **do** ▷ (the stopping condition may require that the weight change by less than ε for some number of iterations or could be a fixed number of steps)

Randomly shuffle (X, Y)

Draw a small batch $(\hat{X}, \hat{Y}) \subset (X, Y)$

$w \leftarrow w - \eta \left(\sum_{(x,y) \in (\hat{X}, \hat{Y})} \nabla L(P_w(\hat{x}), \hat{y}) \right)$

end while

Stochastic gradient descent achieves a smoothing effect on the gradient optimization by only sampling a subset of the training data for each iteration. Miraculously, this smoothing effect not only often achieves faster convergence, the result also generalizes better than solutions using deterministic gradient methods [Hardt et al.(2015)Hardt, Recht, and Singer]. It is for this reason that SGD has been adopted as the de facto standard among ANN training applications.

9 Problem Setting: Adversarial Attacks

In general *Adversarial Attacks* against models are small perturbations from natural data which significantly perturb model output. Szegedy et al. [Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goo realized that the same computational tools used to train ANNs could be used to generate attacks that would confuse them. Their approach was to define a loss function relating the output of the ANN for a given initial image to a target adversarial output plus the L^2-norm of the input and use backpropagation to compute gradients – not on the weights of the neural network, but on just the input layer to the network. The solution to this optimization problem, efficiently approximated by a gradient-based optimizer, would be a slightly perturbed natural input with a highly perturbed output. Their experimental results are striking:

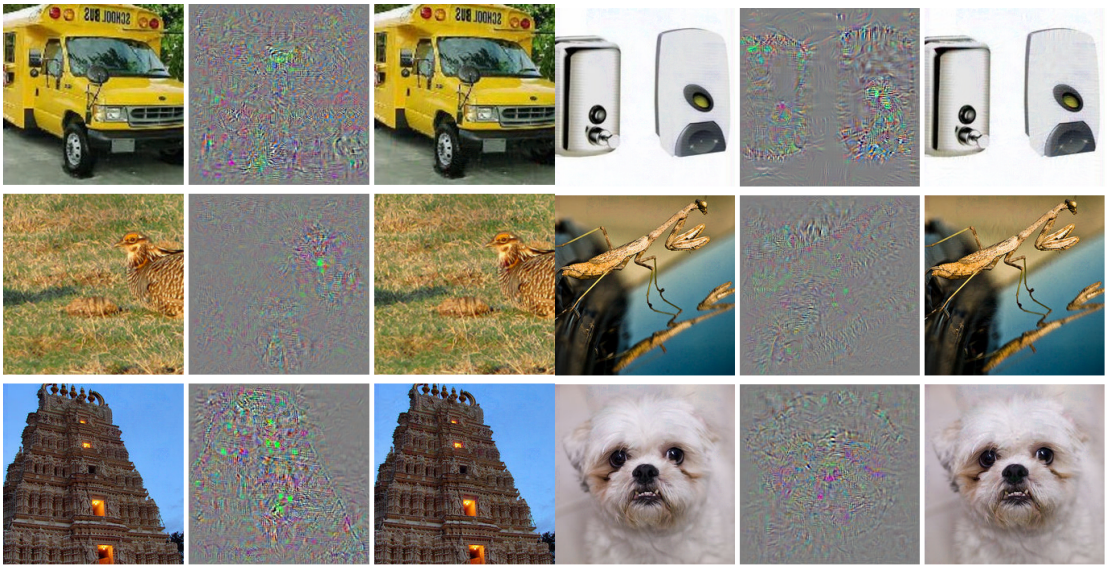


Figure 1: Natural Images are in columns 1 and 4, Adversarial images are in columns 3 and 6, and the difference between them (magnified by a factor of 10) is in columns 2 and 5. All images in columns 3 and 6 are classified by AlexNet as "Ostrich" [Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus]

9.1 Data Sets

The Dataset used above is known as ImageNet – a large set of labeled images varying in size originally compiled for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). This dataset and its many subsets has become a standard for image classification and feature identification experiments. In the experiments that follow, ImageNet will be featured alongside the Modified National Institute of Standards and Technology (MNIST) dataset which is a database of hand written digits often used to develop image processing and character recognition systems. This dataset is much lower resolution than ImageNet and is therefore experiments run much more quickly on it and require less complex input/output.

9.2 L-BFGS minimizing distortion

Szegedy et al. took advantage of the tools they had on hand for training neural networks to set up a box-constrained optimization problem whose approximated solution generates these targeted misclassifications.

Let $f : \mathbb{R}^m \to \{1, ..., k\}$ be a classifier and assume f has an associated continuous loss function denoted by $\text{loss}_f : \mathbb{R}^m \times \{1, ..., k\} \to \mathbb{R}^+$ and l a target adversarial .
Minimize $\|r\|_2$ subject to:

1. $f(x + r) = l$
2. $x + r \in [0, 1]^m$

The solution is approximated with L-BFGS (see Appendix **??**) as implemented in Pytorch or Keras. This technique yields examples that are close to their original counterparts in the L^2 sense.

9.2.1 L-BFGS: Mnist

The following examples are prepared by implementing the above technique via pytorch on images from the Mnist dataset with FC200-200-10, a neural network with 2 hidden layers with 200 nodes each:

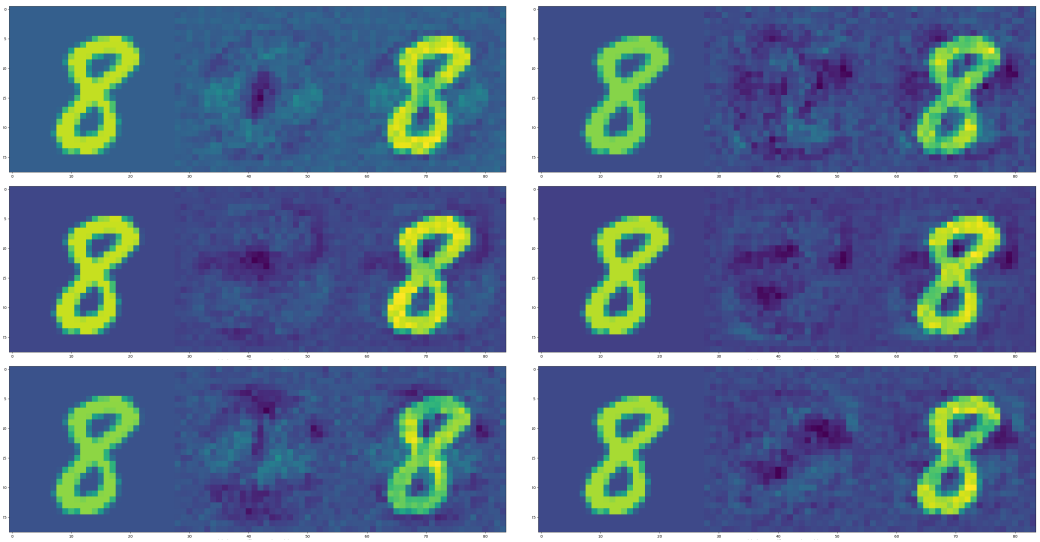


Figure 2: Original images on the left, Perturbation is in the middle, Adversarial Image (total of Original with Perturbation) is on the right. Column 1 shows an original 8 being perturbed to adversarial classes 0, 2, and 4. Column 2 shows adversarial classes 1, 3, and 5

Borrowing a metric from Szegedy et al to compare the magnitude of these distortions, we will define

Definition 9.1. *Distortion is the L^2 norm of the difference between an original image and a perturbed image, divided by the square root of the number of pixels in the image:*

$$\sqrt{\frac{\sum_i \hat{(}x_i - x_i)^2}{n}}$$

Distortion is L^2 magnitude normalized by the square-root of the number of dimensions so that values can be compared for modeling problems with differing numbers of dimensions.

The 900 examples generated for the network above had an average distortion of 0.089 with the following distribution of distortions, given in figure 3.

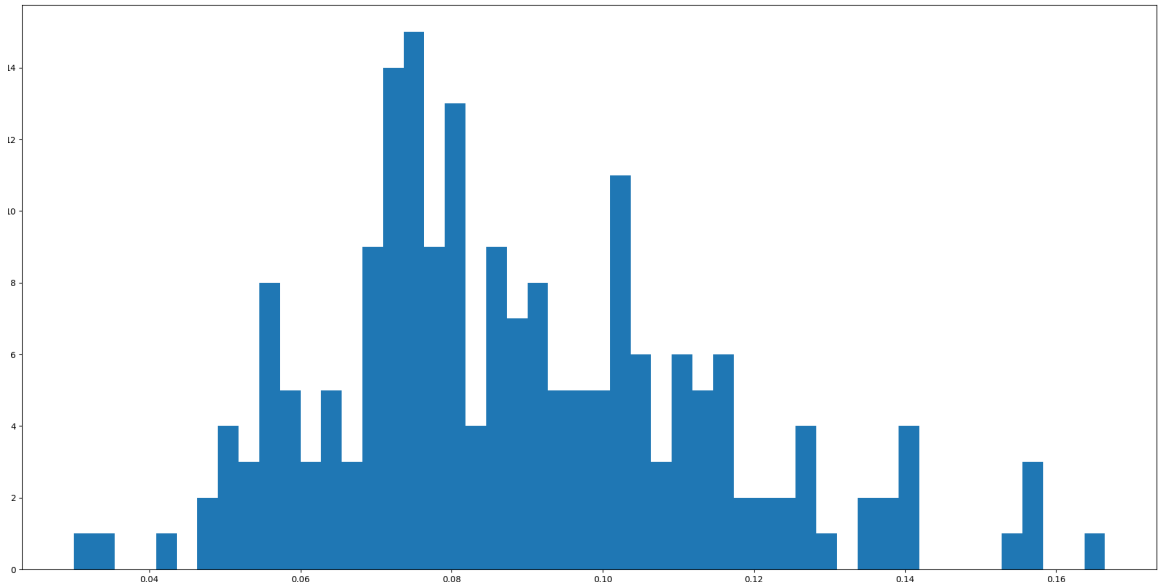


Figure 3: A histogram of the distortion measured for each of 900 adversarial examples generated using L-BFGS against the FC-200-200-10 network on Mnist. Mean distortion is 0.089.

9.2.2 L-BFGS: ImageNet

We also tried to replicate [Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergu results on ImageNet. Attacking VGG16, a well known model from the ILSVRC-2014 competition [Simonyan and Zisserman(2014)], on ImageNet images with the same technique generates the examples in figure 4:

Figure 4: Original images on the left, Perturbation (magnified by a factor of 100) by is in the middle, Adversarial Image (total of Original with Perturbation) is on the right.

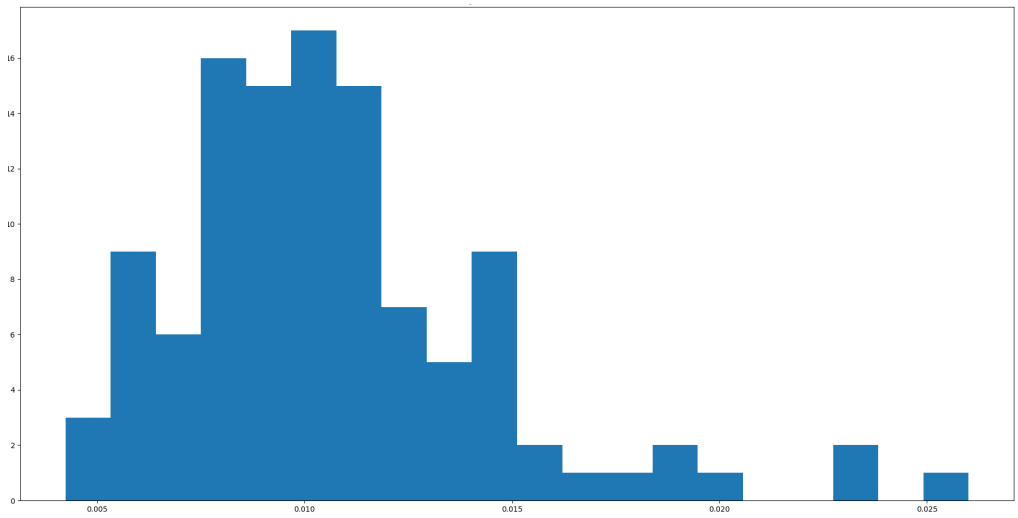


Figure 5: A histogram of the distortion measured for each of 112 adversarial examples generated using L-BFGS against the VGG16 network on ImageNet images with mean distortion 0.0107

9.2.3 Fast Gradient Sign Method (FGSM)

We also implemented a single step attack process uses the gradient of the loss function L with respect to the image to find the adversarial perturbation [Goodfellow et al.(2014)Goodfellow, Shlens, and Szegedy]. for given ε, the modified image $\hat{x}$ is computed as

$$\hat{x} = x + \epsilon \text{sign}(\nabla L(P_w(x), x)) \tag{4}$$

This method is simpler and much faster to compute than the L-BFGS technique described above, but produces adversarial examples less reliably and with generally larger distortion. Performance was similar but inferior to the Iterative Gradient Sign Method summarized below.

9.2.4 Iterative Gradient Sign Method (IGSM)

In [Kurakin et al.(2016)Kurakin, Goodfellow, and Bengio] an iterative application of FGSM was proposed. After each iteration, the image is clipped to a εL_∞ neighborhood of the original. Let $x'_0 = x$, then after m iterations, the adversarial image obtained is:

$$x'_{m+1} = \text{Clip}_{x,\epsilon}\left\{x'_m + \alpha \times \text{sign}(\nabla \ell(F(x'_m), x'_m))\right\} \tag{5}$$

This method is faster than L-BFGS and more reliable than FGSM but still produces examples with greater distortion than L-BFGS.

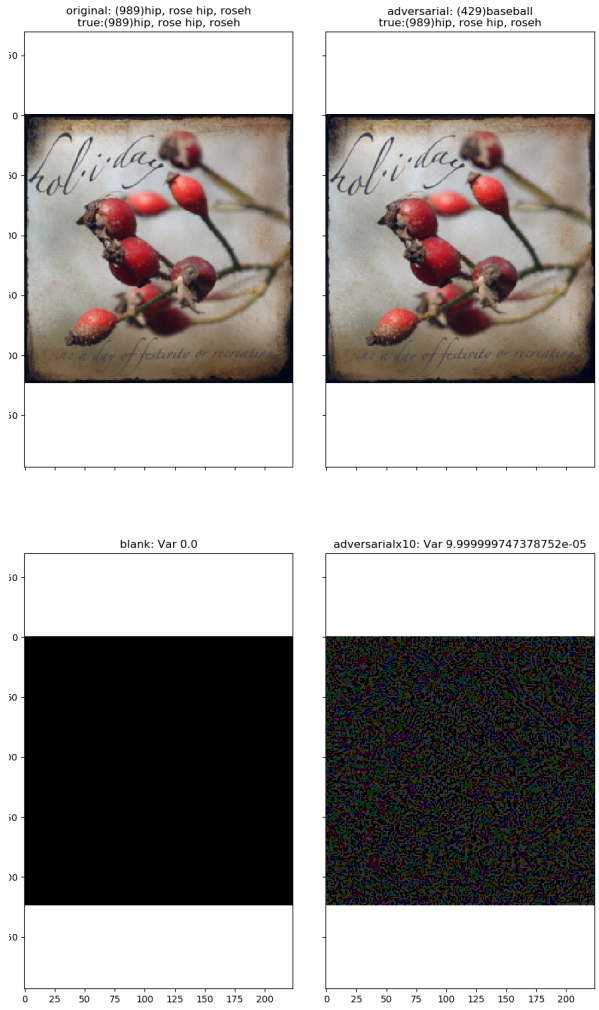


Figure 6: adversarial example generated against VGG16 (ImageNet) with IGSM. Original Image on the left, adversarial image and added noise (ratio of variance adversarial noise/original image: 0.0000999) on the right.

9.3 Other Attacks

The following attack techniques are also prevalent in the literature but have not been replicated in these experiments.

9.3.1 Jacobian-based Saliency Map Attack (JSMA)

Another attack noted by [Papernot et al.(2015)Papernot, McDaniel, Jha, Fredrikson, Celik, and Swami] estimates the *saliency map*, a rating for each of the input features (e.g. each pixel) on how influential it is for causing the model to predict a particular class with respect to the model output [Wiyatno and Xu(2018)]. This attack modifies the pixels that are most salient. This is a targeted attack, and saliency is designed to find the pixel which increases the classifier's output for the target class while tending to decrease the output for other classes.

9.3.2 Deep Fool (DFool)

A technique proposed by [Moosavi-Dezfooli et al.(2015)Moosavi-Dezfooli, Fawzi, and Frossard] to generate an un-targeted iterative attack. This method approximates the classifier as a linear decision boundary and then finds the smallest perturbation needed to cross that boundary. This attack minimizes L_2 norm with respect to to the original image.

9.3.3 Carlini & Wagner (C&W)

In [Carlini and Wagner(2016)] an adversarial attack is proposed which updates the loss function such that it jointly minimizes L_p and a custom differentiable loss function based on unnormalized outputs of the classifier (*logits*). Let Z_k denote the logits of a model for a given class k, and κ a margin parameter. Then C&W tries to minimize:

$$||x - \hat{x}||_p + c * max\left(Z_k(\hat{x}_y) - max\{Z_k(\hat{x}) : k \neq y\}, -\kappa\right) \tag{6}$$

10 Exploring Stability under classifications

In this section we define and experimentally demonstrate a notion of stability of classification under a given classification model.

10.1 Definitions

To build up to a definition for stability, we develop definitions and terms to refer to adversarial examples without relying on subjective characteristics like human vision.Let X denote a space of possible data and C denote a set of classes,

Definition 10.1. *Consider a point $x \in X$ with corresponding class $c \in C$ and a classifier $\mathcal{C} : X \to C$. We say that x admits an $(\varepsilon, d)-$*adversarial example *on $\mathcal{C}$ if there exists a point $\hat{x}$ such that $d(x, \hat{x}) < \varepsilon$ and $\mathcal{C}(\hat{x}) \neq c$.*

This definition refers to the most general case of intentional mis-classification. The adversarial class can also be explicitly targeted:

Definition 10.2. *Consider a point $x \in X$ with corresponding class $c \in C$ and a classifier $\mathcal{C} : X \to C$. We say that x admits an $(\varepsilon, d, c_t)-$*targeted adversarial example *on $\mathcal{C}$ if there exists a point $\hat{x}$ such that $d(x, \hat{x}) < \varepsilon$ and $\mathcal{C}(\hat{x}) = c_t$.*

These definitions rely on a metric d which is commonly L^2, but can be other metrics. The following definition complements the definition for adversarial examples by providing a criteria for the local stability of adversarial examples:

Definition 10.3. *x is (γ, σ)-*stable *if $\mathbb{P}[\mathcal{C}(x') = \mathcal{C}(x)] \geq \gamma$ when $x' \sim \rho = N(X, \sigma^2 I)$ (i.e. x' is drawn from a Gaussian with variance σ and mean x).*

We see that

$$\mathbb{P}[\mathcal{C}(x') = \mathcal{C}(x)] = \int \chi_{\mathcal{C}^{-1}(\mathcal{C}(x))}(x') d\rho(x') = \rho(\mathcal{C}^{-1}\mathcal{C}(x))$$

In the case of images drawn from $\mathbb{R}^n$, we can write this integral precisely

$$\frac{1}{\sigma(2\pi)^{n/2}} \int_{\mathbb{R}^n} \chi_{\mathcal{C}^{-1}(\mathcal{C}(x))}(y) e^{\frac{|x-y|^2}{2\sigma^2}} d\vec{y}$$

This integral is approximated by taking N samples $x_k \sim \rho$ and computing

$$\frac{|x_k : \mathcal{C}(x_k) = \mathcal{C}(x)|}{N}$$

In our experiments, γ is fixed and σ is adjusted to determine the σ for which an image is $\gamma - \sigma$ stable. This is accomplished by a bracketing procedure, first expanding the search space bounds σ_{min} and σ_{max} so that the average number of samples at these bounds are below (respectively above) γ. This search space is then sequentially bisected to identify a σ corresponding with γ. The algorithm is presented in Appendix **??**

10.2 Stability Experiments

In the following experiments, adversarial examples are generated using IGSM (9.2.4) and L-BFGS methods (9.2.2). The neighborhoods around them are sampled using Gaussians with varying σ. Separately the bracketing algorithm described above is used to identify σs corresponding with $\gamma = 0.7$. The resultant σ statistic identifies examples as being $(\gamma = 0.7, \sigma)$ stable meaning that 70% of the samples drawn from a Gaussian with standard deviation σ are classified the same as the original class by the model.

10.2.1 FGSM Neighborhood sampling

In these initial experiments, the neighborhood around normal and adversarial examples are explored via sampling. An Mnist image x with classification 1 is selected, adversarial noise x_a is prepared using IGSM 5 for each target other than 1. The neighborhoods around each such image are examined by generating 1000 Gaussian samples at 1000 equally spaced σs.

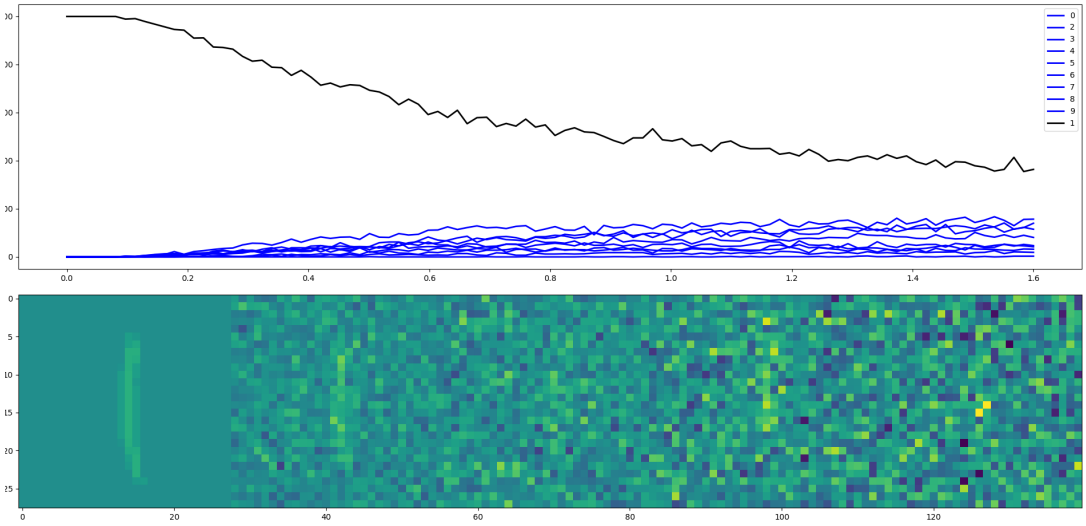


Figure 7: Plotting frequency of each class in samples with increasing variance around natural image. Original Image Class is shown as a black curve. Bottom shows example sample images.

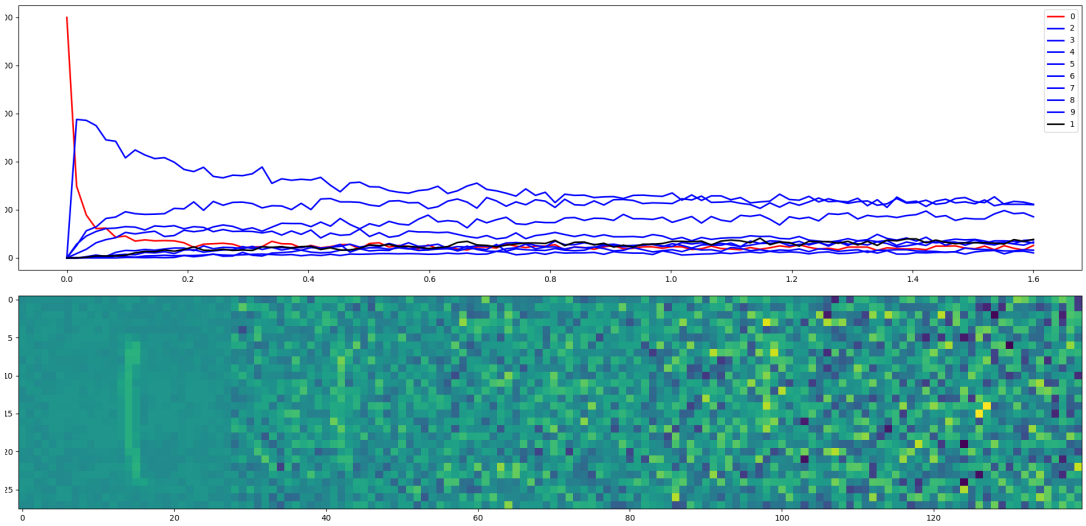


Figure 8: Plotting frequency of each class in samples with increasing variance around adversarial image. Adversarial class is shown as a red curve. Bottom shows example sample images.

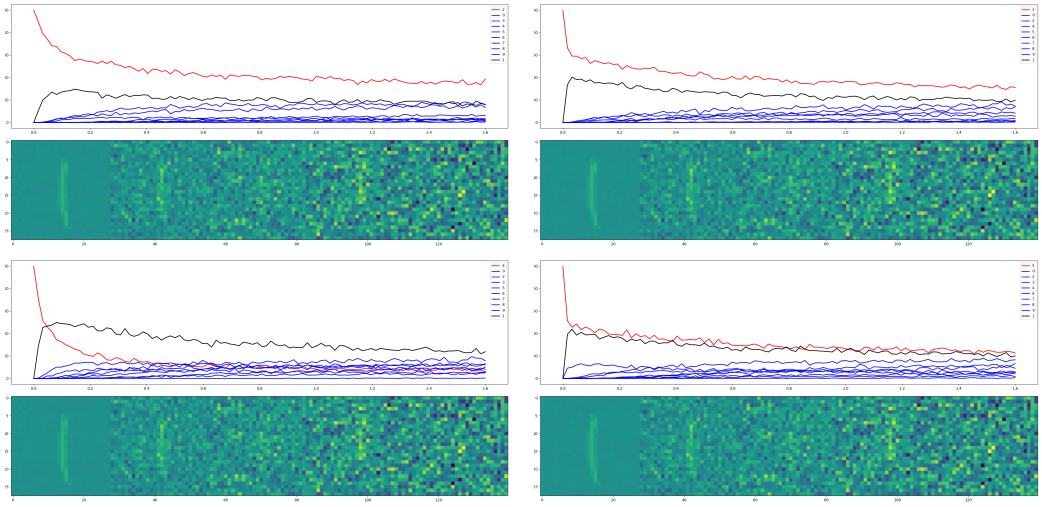


Figure 9: Plotting frequency of each class in samples with increasing variance around adversarial images. Original Class is shown as a black curve, adversarial in red, all others are blue. Bottom shows example sample images.

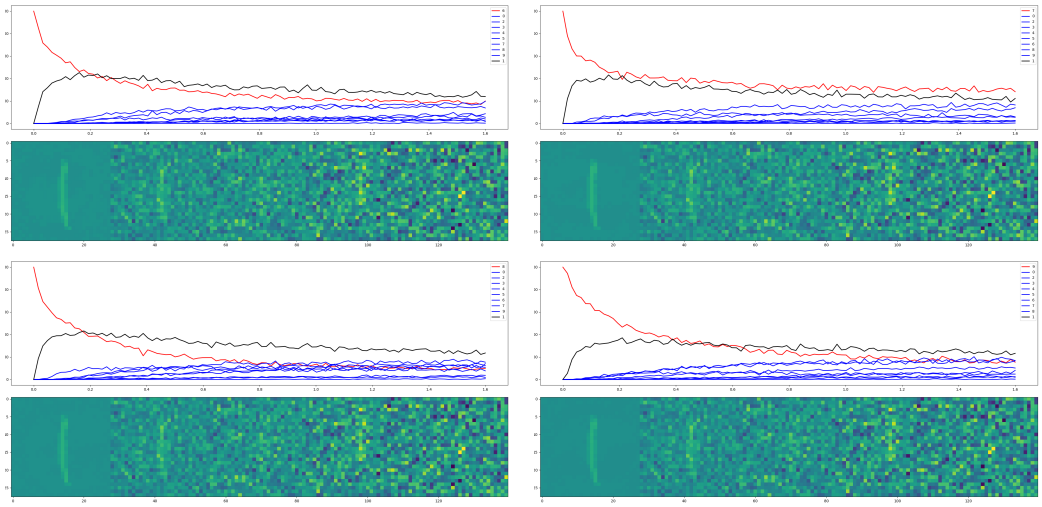


Figure 10: Plotting frequency of each class in samples with increasing variance around adversarial images. Original Class is shown as a black curve, adversarial in red, all others are blue. Bottom shows example sample images.

Attack examples are generated using IGSM (5) from each of 200 Mnist images to each possible target other than the original class (9 total targets) for a total of 1800 adversarial examples. 1800 natural images are selected with the same classes as the 1800 adversarial examples. The images are processed using the bracketing algorithm to identify σs corresponding with $\gamma = 0.7$ and the resulting sigmas are aggregated into the following histogram:

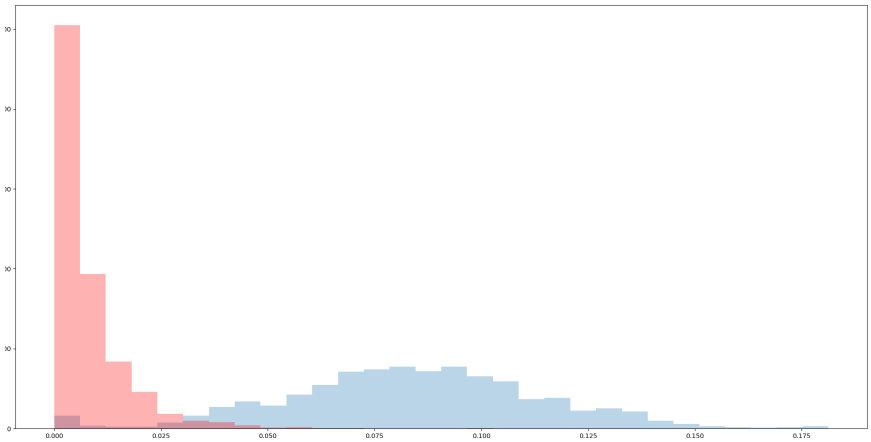


Figure 11: Histogram of σ for $(\gamma = 0.7, \sigma)$ stability of Adversarial Examples (Red) and Natural Examples (Blue)

In this experiment, the adversarial images (red in the histogram above) were much less stable than the natural images. In this way, 98% of the adversarial images could be identified by their $\gamma - \sigma$ stability characteristics.

10.2.2 Sampling Analysis of ImageNet and FGSM)

Similar to the above analysis for Mnist, an image x is selected with original class "Rose-Hip", an adversarial perturbation x_a is prepared so that the sum $x + x_a$ has class "Baseball", and 50 Gaussian samples are generated for each of 1000 evenly spaced σ values. These samples are added to the above images and the classifications are colored via a heatmap:

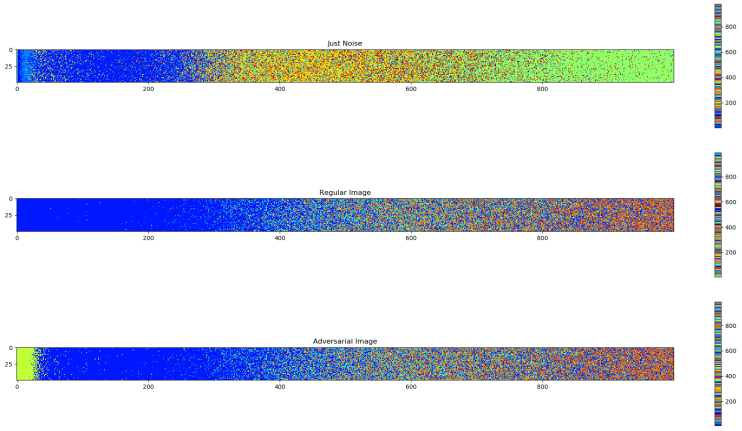


Figure 12: Top heatmap indicates Gaussian samples added to a pure black image,
Middle heatmap indicates Gaussian samples added to x (the original image),
Bottom heatmap indicates Gaussian samples added to $x + x_a$ (the adversarial image).

50 such examples are generated and bracketing is used to determine the σ for $(\gamma = 0.7, \sigma)$ stability in the following histogram. The adversarial example is denoted with a red dot.

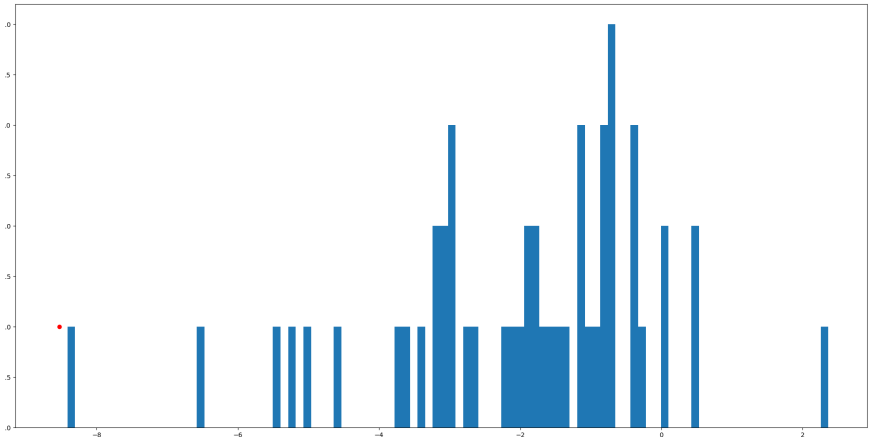


Figure 13: Histogram of σ for $(\gamma = 0.7, \sigma)$ stability of ImageNet examples with class "Baseball." Red dot on the left indicates the same σ measurement for an adversarial perturbation of a "Rose-Hip" with new class "Baseball" 6

Much like the above case for Mnist, IGSM examples generated on ImageNet were much less $\gamma - \sigma$ stable than natural images, again about 98% of adversarial images could be distinguished by their $\gamma - \sigma$ stability. It should be noted that due to computational intensity of processing ImageNet samples, this dataset is significantly smaller than that used for the Mnist experiments.

10.3 L-BFGS Neighborhood Sampling

In the following experiments, we will use the slower but more reliable L-BFGS technique from [Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus] to prepare adversarial examples and will recreate results from the paper relating distortion with network robustness. The same networks and examples will then be subjected to $\gamma - \sigma$ stability analysis for $\gamma = 0.7$ to determine any correspondence of stability with network accuracy and average distortion.

10.3.1 Re-examining Szegedy Results with sampling

The following is a recreation of Table 1 from [Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goo
in which networks of differing complexity are trained and attacked using L-BFGS. In addition to recreating the results from Szegedy et al. the table contains columns for $(\gamma = 0.7, \sigma)$ stability of natural and adversarial examples for each network. The networks listed are FC10-4, a fully connected network with no hidden layers that that takes the input vector $x \in \mathbb{R}^{784}$ to an output vector of $y \in \mathbb{R}^{10}$ with the $\lambda * \|\vec{w}\|_2 / N$ where $\lambda = 10^{-4}$ and N is the number of parameters in $\vec{w}$) added as regularization to the objective function during training. FC10-2 is the same except with $\lambda = 10^{-2}$ and FC10-0 has $\lambda = 1$ (a very large coefficient for regularization). FC100-100-10 and FC200-200-10 are networks with 2 hidden layers with regularization added for each layer of perceptrons with the λ for each layer equal to $10^{-5}, 10^{-5}$, and 10^{-6}. Training for these networks is conducted with a fixed number of epochs.

Network	Test Acc	Avg Distortion	Adversarial $(\gamma = 0.7, \sigma)$	Natural $(\gamma = 0.7, \sigma)$
FC10-4	92.09	0.123	1.68	0.93
FC10-2	90.77	0.178	4.25	1.37
FC10-0	86.89	0.278	12.22	1.92
FC100-100-10	97.31	0.086	0.56	0.65
FC200-200-10	97.61	0.087	0.56	0.73

Table 1: Recreation of Table 1 from Szegedy et al. using pytorch. New columns show σ values which achieved $(\gamma = 0.7, \sigma)$ stability for Adversarial and Natural examples.

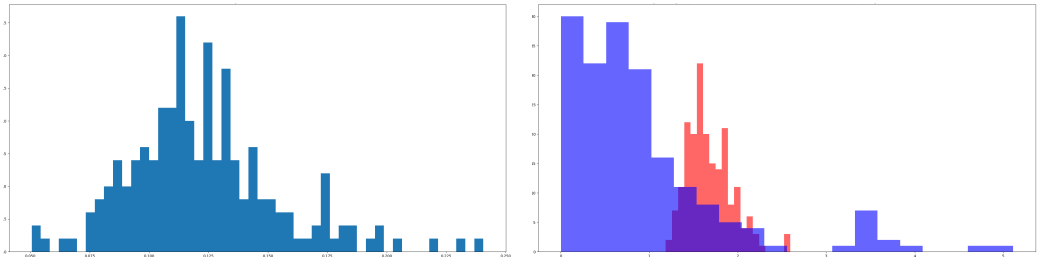


Figure 14: Distortion and σ histograms for FC-10-4 in Table 1

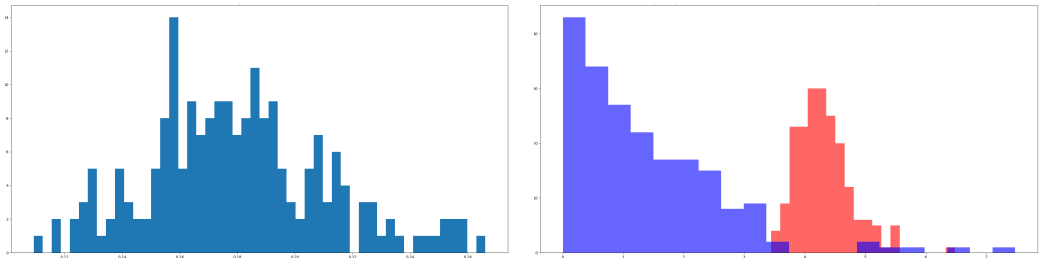


Figure 15: Distortion and σ histograms for FC10-2 in Table 1

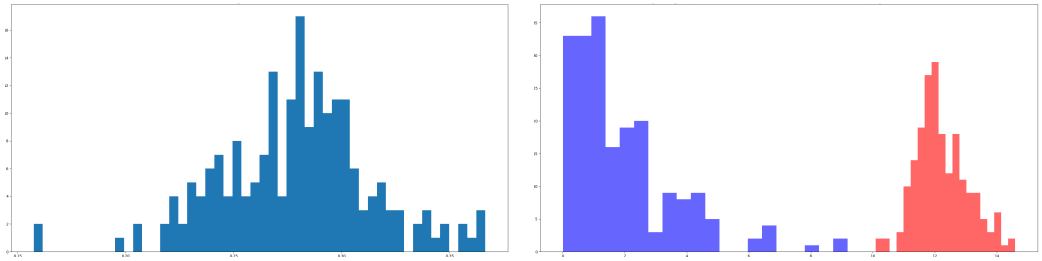


Figure 16: Distortion and σ histograms for FC10-0 in 1

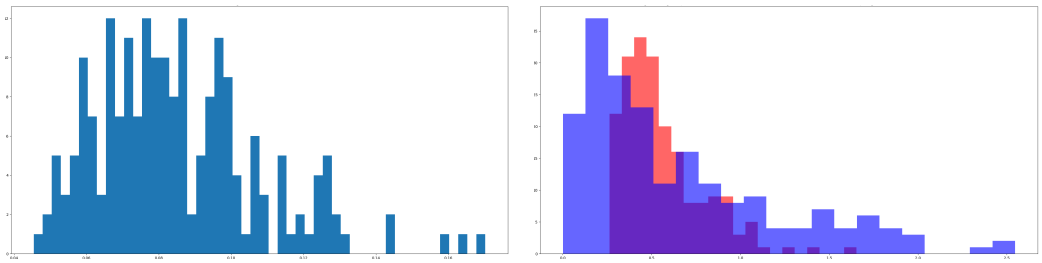


Figure 17: Distortion and σ histograms for FC100-100-10 in 1

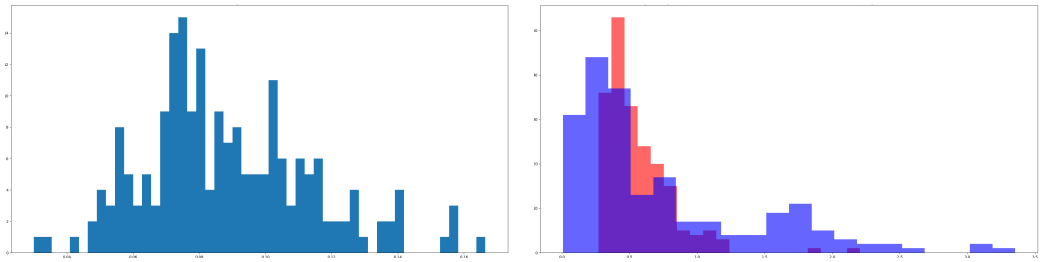


Figure 18: Distortion and σ histograms for FC200-200-10 in 1

10.4 Discussion of Experimental Results

In these results, we first notice that adversarial attacks generated with IGSM against Mnist are less stable than natural examples from these data sets. This difference is even more pronounced

when comparing a adversarial image with other image samples from its target class. This lower stability of adversarial examples is even more pronounced among the attacks prepared via IGSM against VGG16 on ImageNet. Using this technique, we can distinguish adversarial examples prepared by these techniques from natural examples using stability as a test.

In the next section of experiments, Table 1 from [Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, G was reconstructed and augmented with stability measurements. In this table, of the 5 networks trained on the Mnist dataset, the more complicated networks achieved higher accuracy but were defeated by adversarial examples with much less distortion than the simpler adversarial examples. We have added columns indicating the stability of natural and adversarial examples in these models. Adversarial examples generated for more complicated models were much less stable (some less stable than the corresponding natural examples). Adversarial examples generated for the less complicated networks were vastly more stable than natural examples. These results are summarized in the histograms Figure 14 through 18.

Part II
Decision Boundaries

11 decision boundary definitions

12 The Argmax Function

A central issue when writing classifiers is mapping from continuous outputs or probabilities to discrete sets of classes. Frequently argmax type functions are used to accomplish this mapping. To discuss decision boundaries, we must precisely define argmax and some of its properties.

In practice, argmax is not strictly a function, but rather a mapping from the set of outputs or activations from another model into the power set of a discrete set of classes:

$$\text{argmax} : \mathbb{R}^k \to \mathcal{P}(C) \tag{7}$$

Defined this way, we cannot necessarily consider arg max to be a function in general as the singleton outputs of argmax overlap in an undefined way with other sets from the power set. However, if we restrict our domain carefully, we can identify certain properties.

Restricting to only the pre-image of the singletons, it should be clear that argmax is continuous.

Indeed, restricted to the pre-image of any set in the power-set, argmax is continuous.

Further, we can directly prove that the pre-image of an individual singleton is open. Observe that for any point whose image is a singleton, one element of the domain vector must exceed the others by $\varepsilon > 0$. We shall use the ℓ^1 metric for distance, and thus if we restrict ourselves to a ball of radius ε, then all elements inside this ball will have that element still larger than the rest and thus map to the same singleton under argmax. Since the union of infinitely many open sets is open in $\mathbb{R}^k$, the union of all singleton pre-images is an open set. Conveniently this also provides proof that the union of all of the non-singleton sets in $\mathcal{P}(C)$ is a closed set. We will call this closed set the argmax Decision Boundary.

Note : there are ways argmax can be forced to break ties, i.e. by ordering the

Questions:

Is the decision boundary connected

13 Defining Decision Boundaries

13.1 Complement Definition

A point x is in the *decision interior* D'_f for a classifier $f : \mathbb{R}^N - > \mathcal{C}$ if there exists $\delta > 0$ such that $\forall \epsilon < \delta, |f(B_\epsilon(x))| = 1$.

The *decision boundary* of a classifier f is the closure of the complement of the decision interior $\overline{\{x : x \notin D'_f\}}$.

13.2 Constructive Definition

A point x is on the *decision boundary* D of a classifier f if $\forall \epsilon > 0, |f(B_\epsilon(x))| \geq 2$.

A point x is a *binary decision point* if $\exists \delta > 0$ such that $\forall \epsilon > 0$ if $\epsilon < \delta$, then $|f(B_\epsilon(x))| = 2$.

A point x is on the *K-decision boundary* D^K of a classifier f if $\exists \delta > 0$ such that $\forall \epsilon > 0$ if $\epsilon < \delta$, then $|f(B_\epsilon(x))| = K$.

13.3 Level Set Definition

The decision boundary D of a probability valued function F is a union of all level sets $L_{c_1,c_2,a}$ defined by two classes c_1 and c_2 and a constant a which satisfy two properties: First, given $x \in L_{c_1,c_2,a}$, $F(x)_{c_1} = F(x)_{c_2} = a$ where a is a constant also defining the level set. Second, for all $c \notin \{c_1, c_2\}$, we have $a > F(x)_c$.

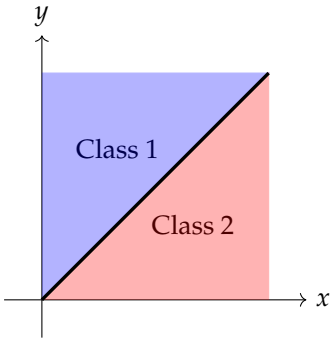
13.4 Images of Decision Boundaries

We can consider a classification workflow f from a data space X (e.g. $\mathbb{R}^N$) using a model F mapping the data space X to a probability space Y (e.g. $[0,1]^K$ where K is the number of classes) and using argmax to convert continuous representations in Y to discrete classes in the set of classes $\mathbb{C}$.

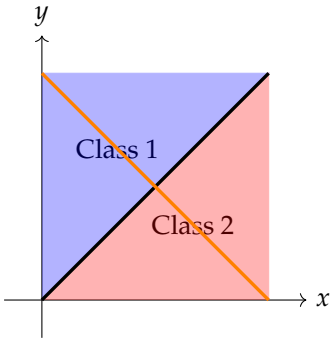
$$X \xrightarrow{F} Y \xrightarrow{\text{argmax}} \mathcal{C} \tag{8}$$

In this way we define a usual discrete classifier $f = F \circ \text{argmax}$

The decision boundary D is defined on X, however we may look at the image of the decision boundary under F. For example, if there are only two classes, then $Y = [0,1]^2$ and the decision boundary can be nicely visualized by the black line below:



Furthermore, if the output of F are *probabilities* which add to one, then all points of x will map to the orange line:



We note that the point $(0.5, 0.5)$ is therefore the only point on the decision boundary for probability valued F. We may generalize to higher dimensions where all probability valued models F will map into the the plane $x + y + z + \cdots = 1$ in Y and the decision boundary will be partitioned into $K - 1$ components, where the K-decision boundary is the intersection of this plane with the *centroid* line $x = y = z = \cdots$ and the 2-decision boundaries become planes intersecting at the same line.

14 exploring boundary curvature with Random Walks

To analyze decision boundary curvature, we will project samples of points onto the decision boundary and then use Singular Value Decomposition to analyze the *projected points*. In general, this process will involve first selecting two images x_1 and x_2 for which $C(x_1) \neq C(x_2)$. A point x_b for which $C(x_b)$ is ambiguous between $C(x_1)$ and $C(x_2)$. The resulting sample X will be projected to the decision boundary by either computing a loss function that is minimized when each of the classes $C(X)$ flip from $C(x_1)$ to $C(x_2)$ or vice versa respectively and performing gradient descent, or by interpolating to the parent point of opposite class (so for $x \in X$, if $C(x) = C(x_1)$, we will interpolate from x to x_2).

Once a projection has been found, we will take the singular value decomposition (SVD) of this sample and examine it for degeneracy.

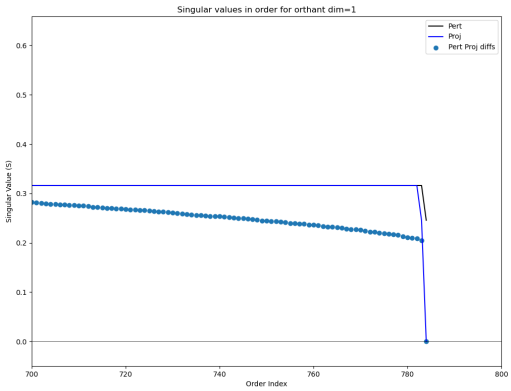
Initial comparison of SVDs of decision boundary points yields a single dominant singular value which corresponds with the content of the original image on the boundary. All random noise appears to be mostly orthogonal to this.

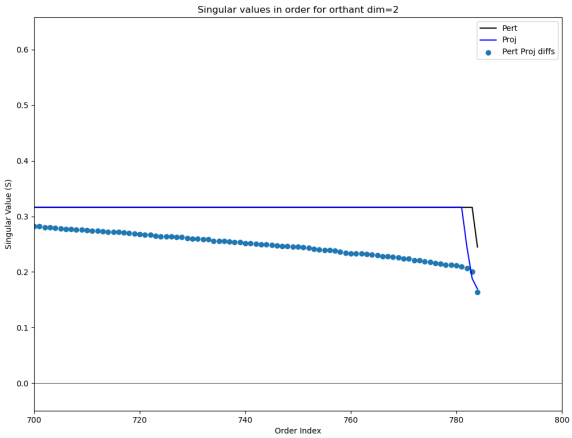
Once this vector is removed, the remaining signal is simply the adversarial attack information.

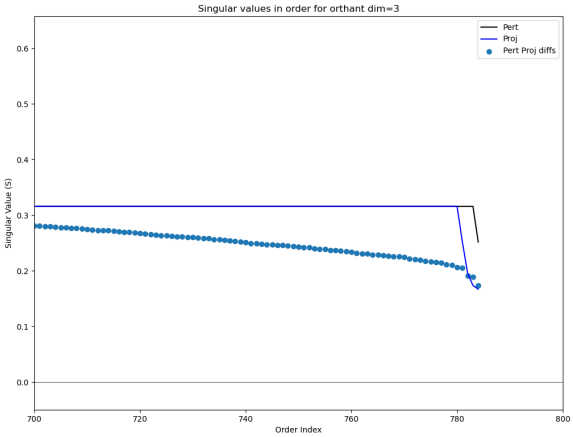
To get a set of singular vectors which to not emphasize the image content, we take random differences among the sampled images and take the SVD of those random differences.

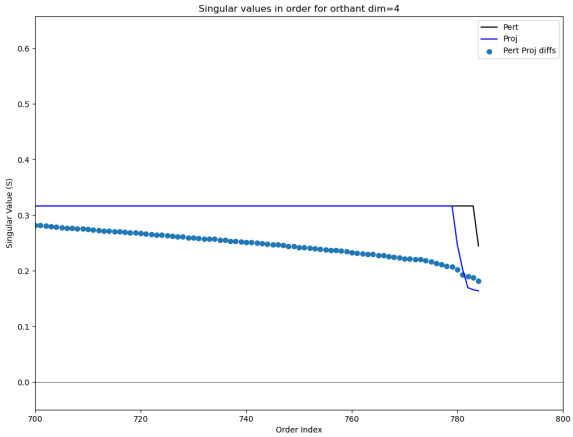
In these first 10 images, this procedure is carried out on the orthant, where samples are generated with mean 0 and are projected onto orthants with increasing numbers of dimensions. We can see a very clear dropped set of singular values smaller than the rest, which indicate the number of degenerate dimensions. In each of these cases, the dropped singular values match the dimension of the orthant.

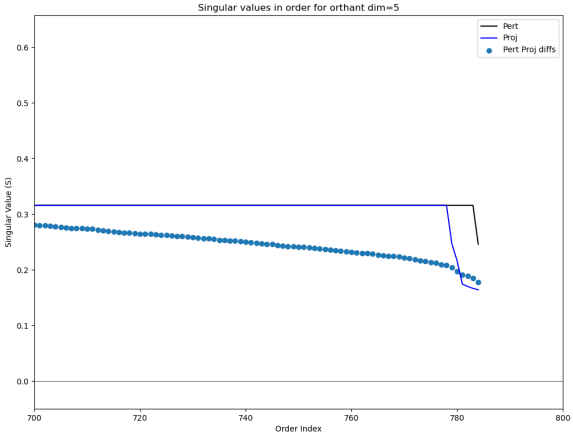
Experiment : A valid experiment here is to measure the difference between the images and the projection to get – in this case – normal vectors to the orthant for each image sampled. The SVD of these normals can be computed to determine the number of dimensions in the projecetion operation. This same procedure can be caried out later in the real practical image projections, although in that case orthogonality is not guaranteed.

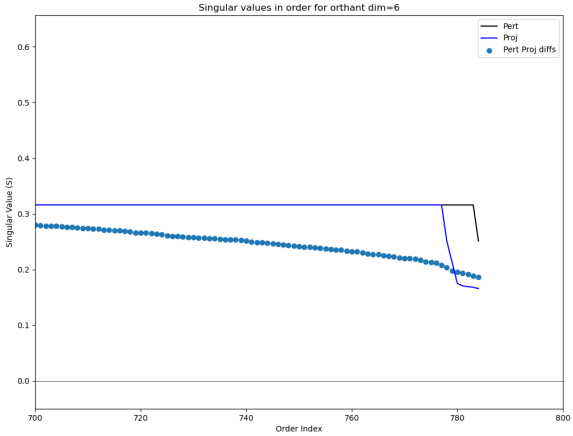

Singular values in order for orthant dim=1
Pert
Proj
Pert Proj diffs
Singular Value (S)
Order Index

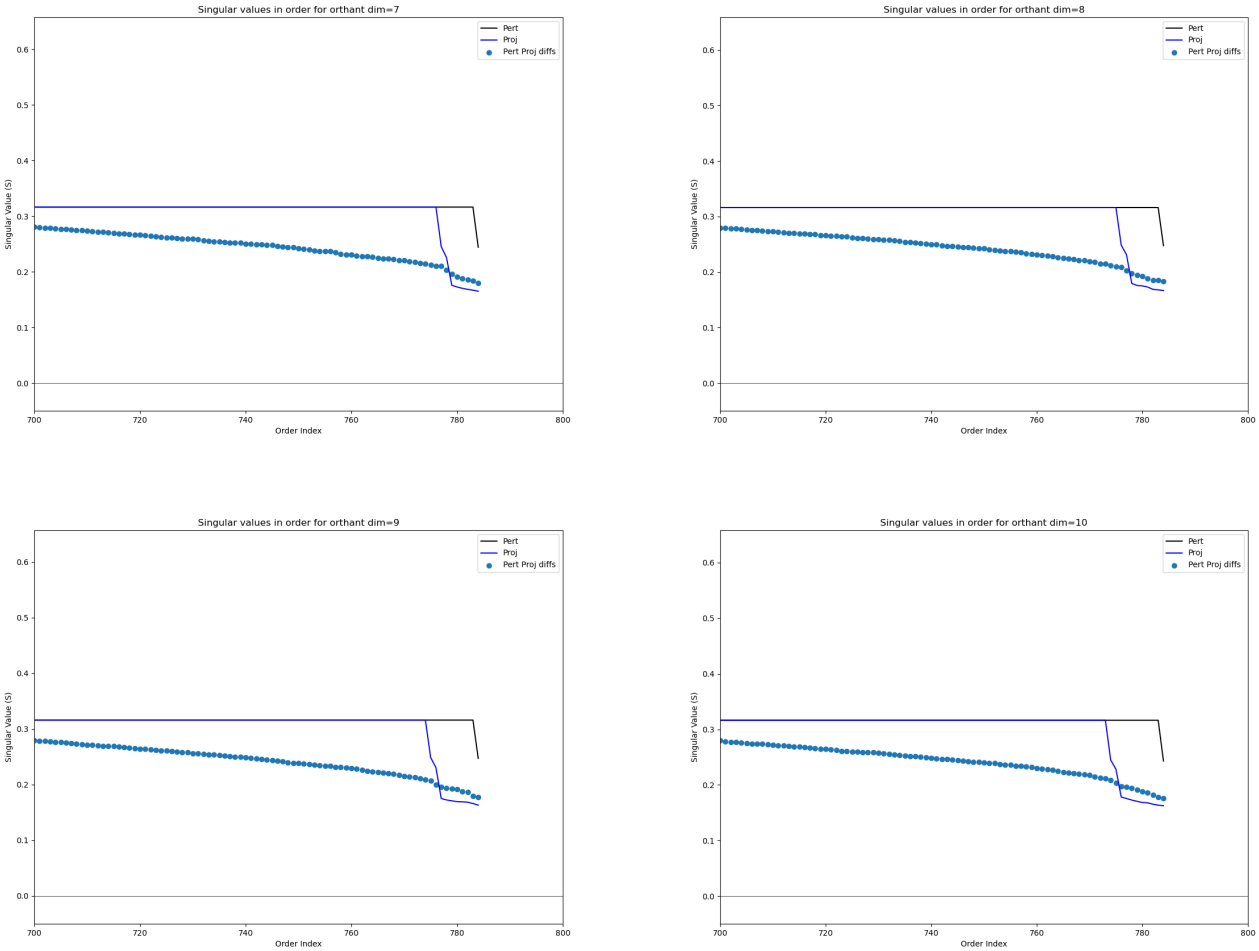

Singular values in order for orthant dim=2
Pert
Proj
Pert Proj diffs
Singular Value (S)
Order Index


Singular values in order for orthant dim=3
Pert
Proj
Pert Proj diffs
Singular Value (S)
Order Index


Singular values in order for orthant dim=4
Pert
Proj
Pert Proj diffs
Singular Value (S)
Order Index


Singular values in order for orthant dim=5
Pert
Proj
Pert Proj diffs
Singular Value (S)
Order Index

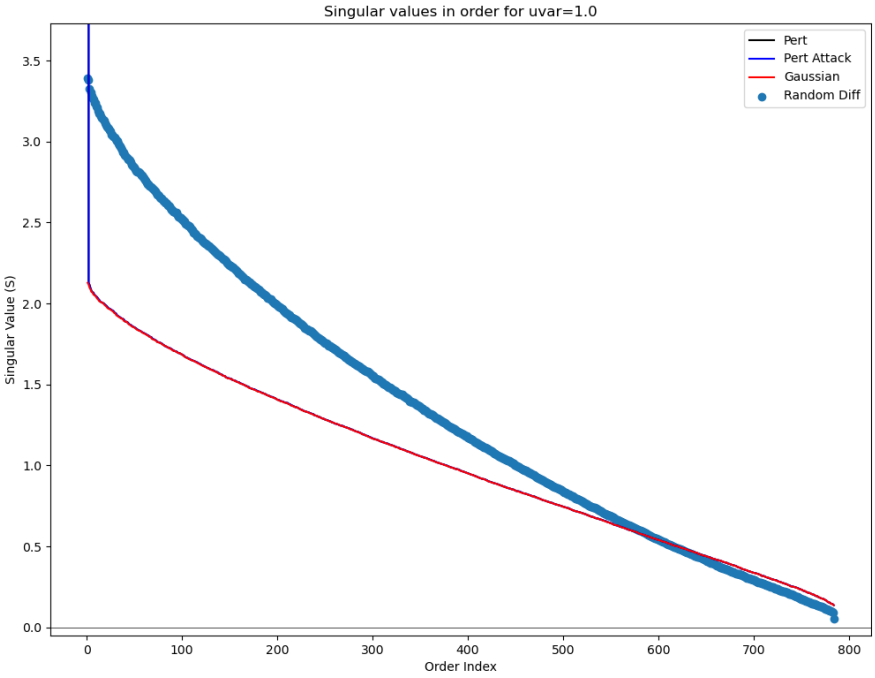

Singular values in order for orthant dim=6
Pert
Proj
Pert Proj diffs
Singular Value (S)
Order Index



There is a question of how best to measure "normal" vectors for each projection. The most direct computation is to take a sample with small variance around each point on the decision boundary and solve least squares of each of these samples. We have previously observed that *most* of the decision boundary is locally planar.

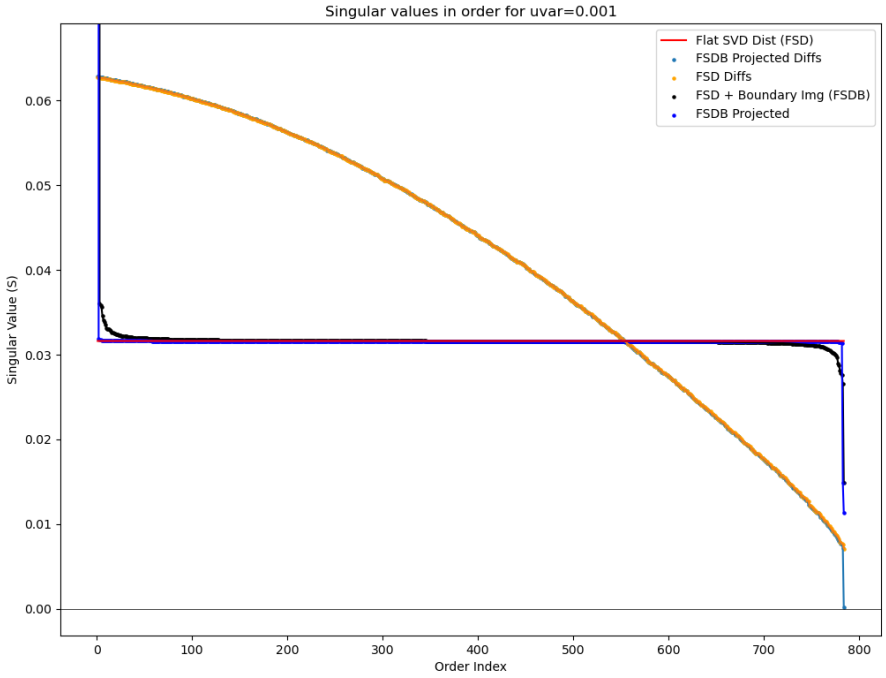
It remains to compare these computed normals with other known quantities, e.g. the gradient computed with respect to adversarial loss functions and the difference between each sampled image and its resultant projection. In addition, it remains to compare the normals of multiple nearby images to determine at what radius of sample the decision boundary begins to show curvature.

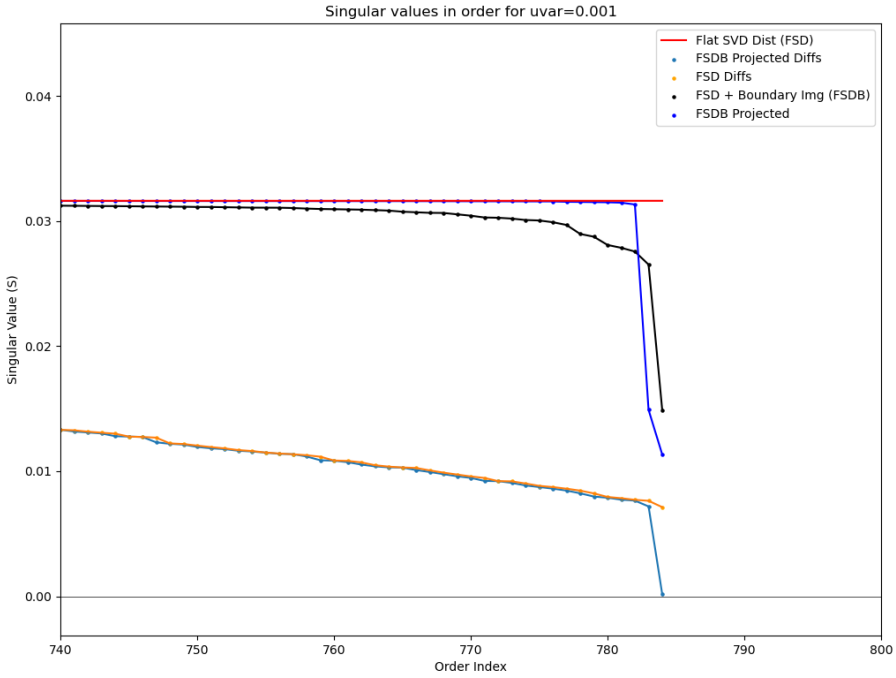
The following image shows singular values in order for a sample around a decision boundary image in MNIST generated by interpolating from a test image to an adversarial image generated from it. This plot is dominated by the natural distribution of singular values for a gaussian and also by the original image information which can be seen as a huge singular value on the far left. We will address both of these factors in following plots.



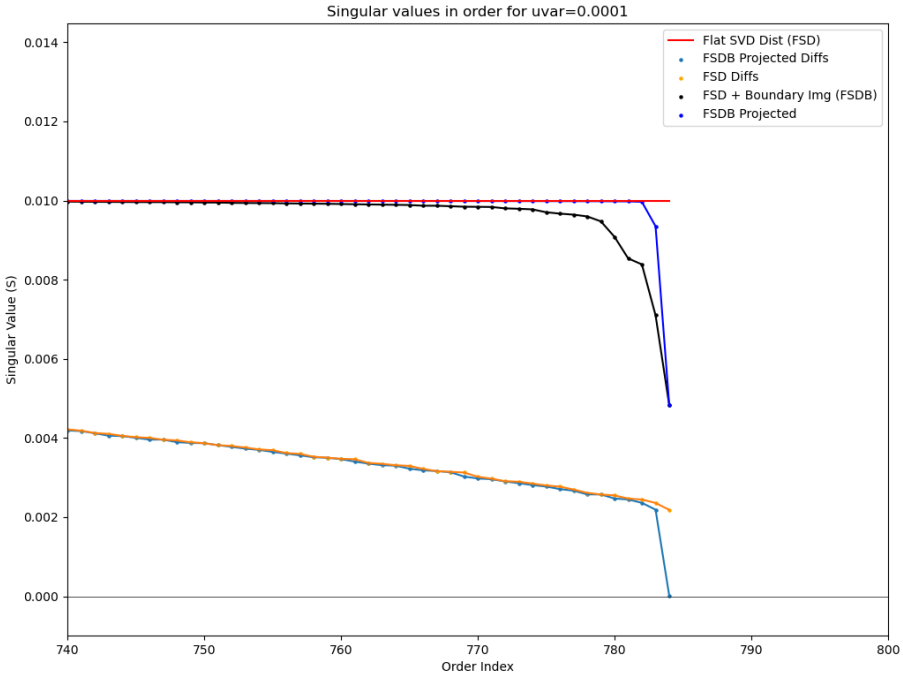
The following plots are generated with a new distribution to replace the gaussian. This "Flat SVD" distribution is generated by first taking a gaussian sample representable by a matrix X, then taking the SVD of this gaussian sample $U\Sigma V = X$, replacing Σ by $|\Sigma|_2 I$. This distribution has the property that its SVD is flat, while maintaining the Frobenius norm of X. This helps to highlight degeneracy in singular values.

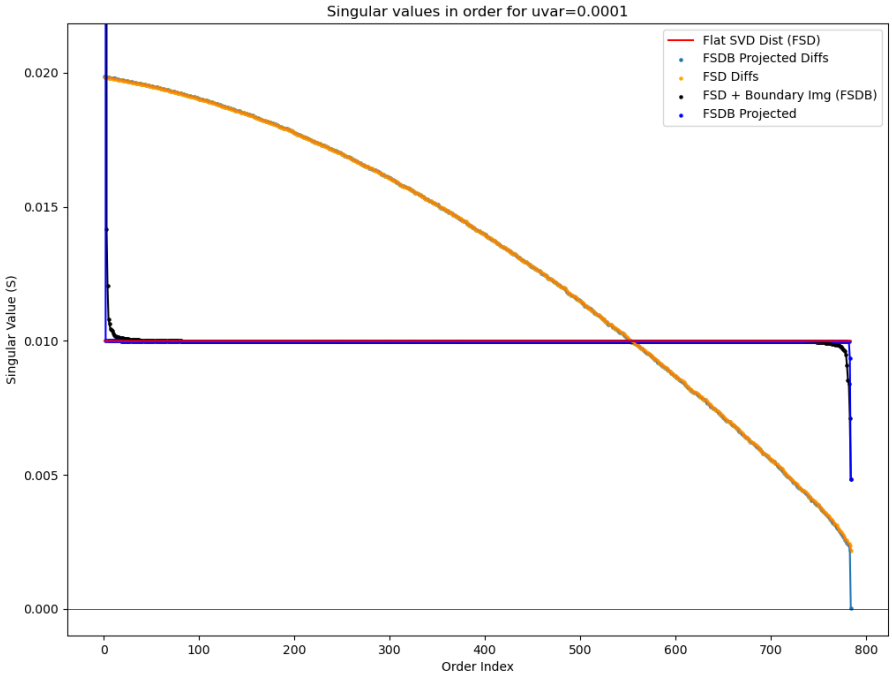
These first two plots are generated by sampling around a decision boundary image found by interpolating between two test images (rather than a test image and an adversarial example generated from it). We note that the SVD contains two degernate values. These seem to correpond with the two original images of the interpolation.


Singular values in order for uvar=0.001
Flat SVD Dist (FSD)
FSDB Projected Diffs
FSD Diffs
FSD + Boundary Img (FSDB)
FSDB Projected
Singular Value (S)
Order Index
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
100
200
300
400
500
600
700
800



These last two images are generated with the flattened distribution

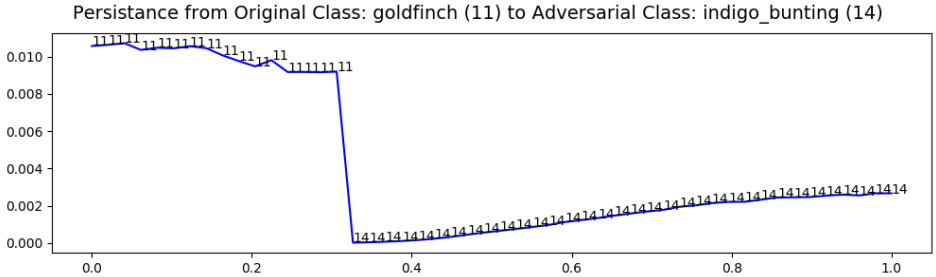

Singular values in order for uvar=0.0001
Flat SVD Dist (FSD)
FSDB Projected Diffs
FSD Diffs
FSD + Boundary Img (FSDB)
FSDB Projected
Singular Value (S)
Order Index
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
740
750
760
770
780
790
800



What these experiments tell us is that most of the time when we cross a decision oundary in the MNIST Dataset, we are crossing at a plane, however if we examine a slightly larger ball around a given point, it will intersect the decision boundary at many other places.

15 Re-Examining Persistence

A movivating picture in this research is an image generated while interpolating persistence in the beginning of this research.



The next objective is to examine this particular case with our random walk and projection Tools.

We also wish to augment these tools to include

15.1 in probability space

15.2 in image space

16 define orthants

17 skewness

18 sampling decision boundaries and analyzing dimensionality with PCA

19 neural network attack gradients versus decision boundaries

20 skewed orthant recreates persistence picture (?).

21 measuring skewness of ANNs

22 Relate skewness with dimpled manifold and features not bugs papers

Part III
Model Geometry

23 Neural Networks are Gaussian Processes

With Dropout on, the interpolant from one class to another will go into a variety of other classes. If you make a histogram of the locations where these boundary crossings occur, that will show a gaussian.

insert figure(s)

24 prove path kernel result in context of differential flow of gradients on neural network. using forward euler approx of grad flow.

25 ** look for sample in weight space and look for gradients

that are pointing toward the final point versus wanting a different direction. Then dot product those with the training direction.

26 training gradients are smooth

27 robust network types : regularized, michael's pca, Soft Nearest Neighbor Loss (SNNL) and adversarially trained.

28 high dimensional arcs are very similar to chords

29 linear interpolated model parameters from random to trained state yield robust models

29.1 For Mnist Inner Products in weight space matter more than distances

– does this generalize to ImNet?

30 define robustness in terms of skew versus orthogonal

31 define robustness in terms of attack perturbation magnitude

32 Model Skewness and decision_boundaries.

turn observations into are they a definition, a theorem, or a discussion
decision_boundary crossing

References

[Bishop(2006)] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

[Carlini and Wagner(2016)] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *arXiv:1608.04644 [cs]*, Aug. 2016. URL http://arxiv.org/abs/1608.04644. arXiv: 1608.04644.

[Glorot et al.(2011)Glorot, Bordes, and Bengio] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[Goodfellow et al.(2013)Goodfellow, Bulatov, Ibarz, Arnoud, and Shet] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks, 2013.

[Goodfellow et al.(2014)Goodfellow, Shlens, and Szegedy] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv:1412.6572 [cs, stat]*, Dec. 2014. URL http://arxiv.org/abs/1412.6572. arXiv: 1412.6572.

[Hardt et al.(2015)Hardt, Recht, and Singer] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. *CoRR*, abs/1509.01240, 2015. URL http://arxiv.org/abs/1509.01240.

[Ivakhnenko and Lapa(1965)] A. G. Ivakhnenko and V. G. Lapa. *Cybernetic predicting devices*. CCM Information Corporation, 1965.

[Kak(1993)] S. Kak. On training feedforward neural networks. *Pramana*, 40(1):35–42, 1993.

[Khoury and Hadfield-Menell(2018)] M. Khoury and D. Hadfield-Menell. On the geometry of adversarial examples. *CoRR*, abs/1811.00525, 2018. URL http://arxiv.org/abs/1811.00525.

[Krause(2020)] A. Krause. Introduction to machine learning, August 2020.

[Kurakin et al.(2016)Kurakin, Goodfellow, and Bengio] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv:1607.02533 [cs, stat]*, July 2016. URL http://arxiv.org/abs/1607.02533. arXiv: 1607.02533.

[LeCun et al.(1988)LeCun, Touresky, Hinton, and Sejnowski] Y. LeCun, D. Touresky, G. Hinton, and T. Sejnowski. A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school*, volume 1, pages 21–28. CMU, Pittsburgh, Pa: Morgan Kaufmann, 1988.

[LeCun et al.(1989)LeCun, Boser, Denker, Henderson, Howard, Hubbard, and Jackel] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[LeCun et al.(1995)LeCun, Bengio, et al.] Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[LeCun et al.(1998)LeCun, Bottou, Bengio, Haffner, et al.] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Li and Yuan(2017)] Y. Li and Y. Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

[Linnainmaa(1970)] S. Linnainmaa. The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), Univ. Helsinki*, pages 6–7, 1970.

[Liu and Deng(2015)] S. Liu and W. Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, pages 730–734. IEEE, 2015.

[Malik and Perona(1990)] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *JOSA A*, 7(5):923–932, 1990.

[McClelland et al.(1986)McClelland, Rumelhart, Group, et al.] J. L. McClelland, D. E. Rumelhart, P. R. Group, et al. Parallel distributed processing. *Explorations in the Microstructure of Cognition*, 2:216–271, 1986.

[McCulloch and Pitts(1943)] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[Minsky and Papert(1969)] M. Minsky and S. Papert. Perceptrons: Anlntroduction to computational geometry. *MITPress, Cambridge, Massachusetts*, 1969.

[Moosavi-Dezfooli et al.(2015)Moosavi-Dezfooli, Fawzi, and Frossard] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. DeepFool: a simple and accurate method to fool deep neural networks. *arXiv:1511.04599 [cs]*, Nov. 2015. URL http://arxiv.org/abs/1511.04599. arXiv: 1511.04599.

[Nair and Hinton(2010)] V. Nair and G. E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. pages 807–814, 2010.

[Papernot et al.(2015)Papernot, McDaniel, Jha, Fredrikson, Celik, and Swami] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The Limitations of Deep Learning in Adversarial Settings. *arXiv:1511.07528 [cs, stat]*, Nov. 2015. URL http://arxiv.org/abs/1511.07528. arXiv: 1511.07528.

[Petersen and Voigtlaender(2018)] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

[Prakash et al.(2018)Prakash, Moran, Garber, DiLillo, and Storer] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. A. Storer. Deflecting adversarial attacks with pixel deflection. *CoRR*, abs/1801.08926, 2018. URL http://arxiv.org/abs/1801.08926.

[Rosenblatt(1958)] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[Rumelhart et al.(1986)Rumelhart, Hinton, and Williams] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[Schmidhuber(2015)] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85 – 117, 2015. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2014.09.003. URL http://www.sciencedirect.com/science/article/pii/S0893608014002135.

[Simonyan and Zisserman(2014)] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Szegedy et al.(2013)Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. URL http://arxiv.org/abs/1312.6199.

[Werbos(1974)] P. J. Werbos. Beyond regression: New tools for prediction and analysis in the be havioral sciences/'phd diss., harvard uni versity. werbos, paul j. 1988. *Generalization of back propagation with application to a recurrent gas market method," Neural Networks*, 1(4):339–356, 1974.

[Wiyatno and Xu(2018)] R. Wiyatno and A. Xu. Maximal jacobian-based saliency map attack. *CoRR*, abs/1808.07945, 2018. URL http://arxiv.org/abs/1808.07945.