Exploring the Manifolds of Adversarial Images Across Various Machine Learning Models

Motivation

Adversarial examples show glaring weaknesses in current machine learning models. A popular example shows how a self-driving car camera model can easily be fooled to think a stop sign is a speed limit sign with only 3 strips of tape placed on the stop sign.

By understanding where the adversarial images generated by different methods lie within the model's manifolds, we can better train more robust models that are more resistant to attacks.

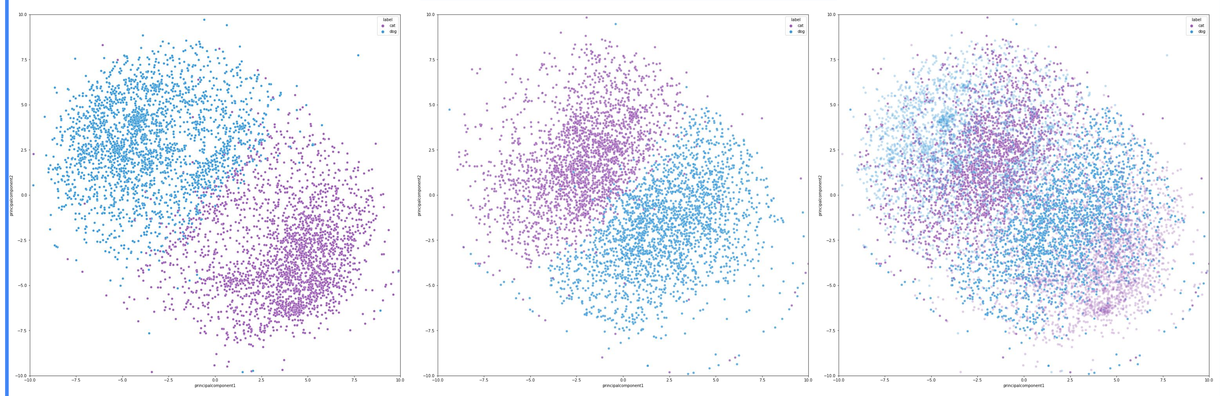
Overview

Rather than analyzing just one model, we decided to generate the manifolds of various models. Doing so would allow us to better understand the types of spaces that adversarial images exist within, rather than just the vulnerabilities within one specific model. Thus, we compiled three different models and datasets to investigate.

Process

We created adversarial images using the Basic Iterative Method (BIM) on testing images for each of the datasets. The manifolds were generated using the t-Distributed Stochastic Neighbor Embedding (TSNE) technique on the training and adversarial images together to map perceived image features into two dimensions.

Cat vs Dog Dataset



Training Images

The Dog vs Cat dataset is made up of 25,000 training images and 5,000 testing images of colored pictures that are 224x224 pixels. We used a vgg16 model configuration that totals 16 layers of CNN, Max-Pooling, and linear layers.

Adversarial Images

The above adversarial images have been perturbed and are labeled by their mistaken classification. In contrast to the other models, the perturbations are entirely indistinguishable to the human eye.

Superimposed Graphs

Image Recognition Accuracy
Regular: 99.16% | Adversarial: 3.16%

BIM Statistics
Alpha: .08 | Epsilon: .3 | Steps: 5

The adversarial manifold shows how nearly the entirety of the two classes was transposed in relative position. On the superimposed graph, it is evident that the density of adversarial images is higher where the density of training images is lower.

Authors

Thomas Cilloni, Robert Hughes, Brannan Kovachev

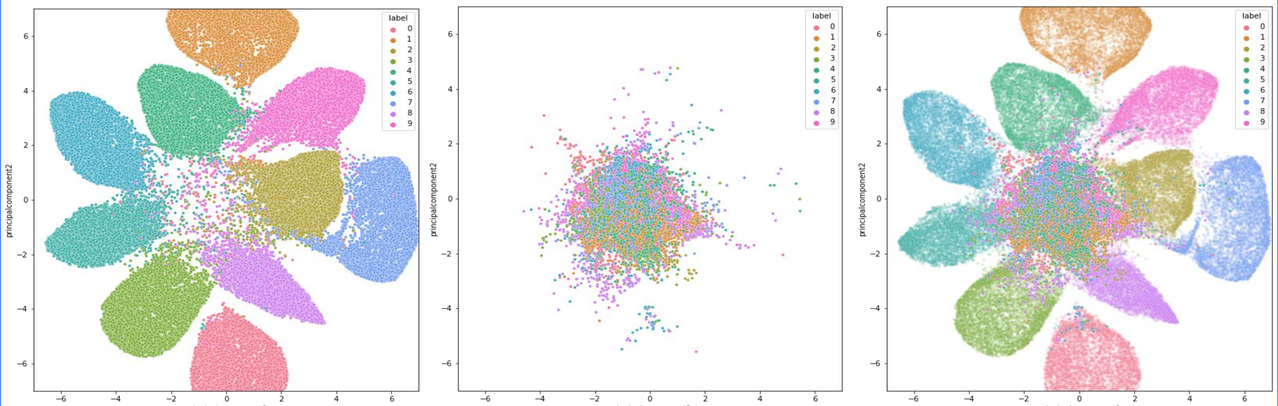
Analysis

The different models showed distinct levels of resilience to adversarial attacks. For example, the Cat vs Dog dataset was able to almost completely fool the model with little change to the images, indistinguishable to the human eye. In contrast, the MNIST dataset required much more perturbation to affect the model's accuracy, resulting in much noise. The Cifar-10 dataset was in between these two with how noticeable yet nearly irrelevant the perturbations were for a human.
The three models also depict different outcomes of adversarial images. Cifar-10 showed the greatest level of clustering with groups of adversarial images within areas previously defined as another class. MNIST best depicted how adversarial images clustered in one entirely undefined space that results in essentially random classification. The Cat vs Dog manifolds show something in between, with images clustering in undefined spaces.

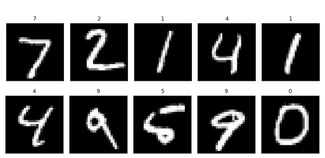
Future Work

Extending this research would mean developing techniques to correctly label more adversarial images. There have been some used in the past. However, the structure of these manifolds may reveal new strategies.

MNIST Dataset



Training Images



The MNIST dataset is made up of 60,000 training images and 10,000 testing images of hand drawn digits on a 26x26 grayscale image. The model used was a simple 5 layer model that uses CNN, dropout, and linear layers.

These manifolds depict a Multiclass Classification model that decides the digit depicted on the input image.

Adversarial Images

The above adversarial images have been perturbed and labeled with their mistaken classes. Although a human notices the difference in noise, the written number is still as comprehensible.

The generated adversarial images generally congregated towards the least well-defined section of the training images' manifold.

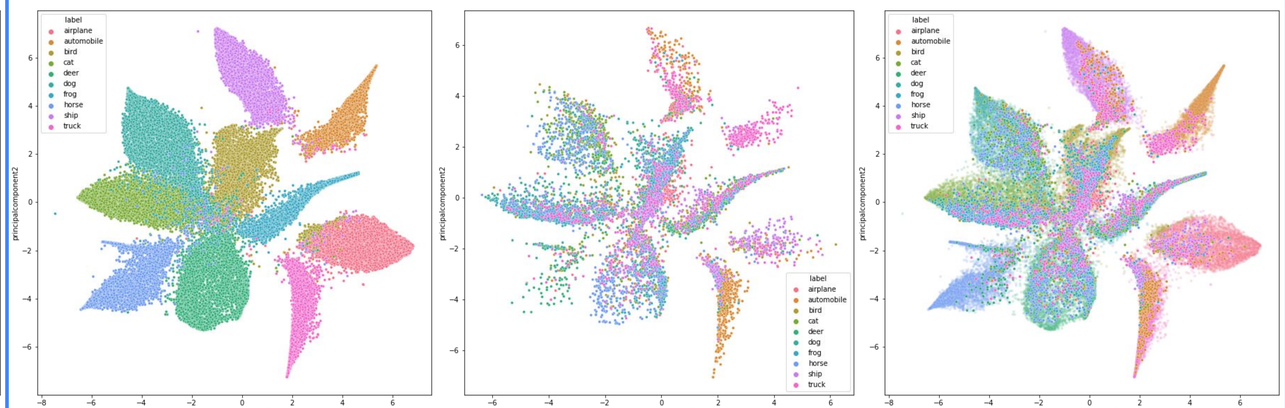
Superimposed Graphs

Image Recognition Accuracy
Regular: 98.1% | Adversarial: 11.62%

BIM Statistics
Alpha: .25 | Epsilon: .5 | Steps: 5

The adversarial images were often misclassified as the same number. A majority of adversarial images were misclassified as the number 8 followed by many misclassified as 2. This can largely be attributed to the classes 2 and 8 falling close in location to the center of the manifold where the adversarial images are grouped.

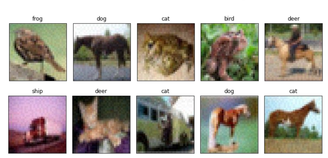
Cifar-10 Dataset



Training Images

The Cifar-10 dataset is made up of 50,000 training images and 10,000 testing images of 10 different types of objects from airplanes to frogs on a 32x32 RGB image. The model uses the ResNet9 architecture which contains 32 layers consisting of CNN, Max-Pooling, RNN, and linear layers.

Adversarial Images



These manifolds depict a Multiclass Classification model that decides what object or creature appears in the input image. The above adversarial images have been perturbed and labeled with their mistaken classes. The level of perturbation can be noticed by a human though it has practically no effect on our ability to understand the image.

Superimposed Graphs

Image Recognition Accuracy
Regular: 92.3% | Adversarial: 6.89%

BIM Statistics
Alpha: .1 | Epsilon: .25 | Steps: 5

In the superimposed graph above, we can see that the adversarial images tend to lie in areas that were previously grouped together by the model. This shows clustering of groups of adversarial images lying in their own smaller areas within those of non-adversarial images. These results best show how adversarial images form clusters within real classes.