CS231n Project Proposal

Austin Wang, Daniel Kunin, Justin Pyron

4/25/2018

For our CS231n project, we will be conducting an exploration of the space of adversarial inputs for a network. Currently, the most popular evaluation criterion for image classification models is simply classification accuracy, the proportion of test images that are classified correctly. However, this metric fails to evaluate other important aspects of a model. One very important metric for evaluating classification models that is often overlooked is robustness; for two images that are similar to each other, we expect their predicted labels to be the same. However, this is often not the case. Christian Szegedy showed in 2014 how it was possible to generate extremely small perturbations in images to cause total misclassifications by maximizing the network's prediction error. What is interesting about these perturbations is that they were found to cause adversary effects across different networks, and therefore are not unique. The goal of our project is to attempt to understand the structure and distribution of these perturbations, describe the size and geometry of the adversarial image space, and explore ways to use this information to make CNN models more robust.

Potential ideas we would like to look into include:

1. Exploring the distribution of adversarial perturbations. An interesting phenomenon is that adversarial perturbations are transferable across networks with different architectures, parameters and even training datasets. While this might imply these perturbations are network independent, it has been shown that random perturbation matrices (with similar levels of distortion) are unsuccessful at generating adversarial examples. So what is special about these perturbation matrices?

2. Exploring the space of adversarial images to gain insight into its topology (size and structure). Szegedy et al. concludes that one possible explanation for why state of the art networks might be so susceptible to adversarial attacks is that "the set of adversarial negatives is of extremely low probability and thus is never (or rarely) observed in the test set, yet is dense (much like the rational numbers) and so it is found near every [sic] virtually every test case". Combining tools from information theory and methods developed in previous papers, we hope to further our understanding of this space.

3. Exploring remedies for lack of robustness. Adversarial images often look indistinguishable from real images to the human eye. A possible explanation is that many adversarial images are essentially identical to real images, except that a small number of pixels (relative to the total number of pixels in the image) are modified. In such a scenario, adversarial images might be neutralized by first passing input images through a filtering algorithm, e.g. akin to an algorithm that takes noisy images and outputs smoothed images. Such an algorithm could be implemented utilizing a probabilistic graphical model, in which each pixel is a random variable in a large multivariate distribution.

Below is a list of papers we have found useful that will guide our project:

- Intriguing properties of neural networks (Szedgedy et al.)
- Adversarial Examples: Attacks and Defenses for Deep Learning (Yuan et al.)
- The Limitations of Deep Learning in Adversarial Settings (Papernot et al.)
- Explaining And Harnessing Adversarial Examples (Goodfellow et al.)
- Adversarial Machine Learning At Scale (Kurakin et al.)
- Boosting Adversarial Attacks with Momentum (Dong et al.)
- Space of Transferable Adversarial Examples (Tramer et al.)

We will rely primarily on MNIST and ImageNet (which are the predominantly used image datasets in papers focused on adversarial attacks). Additionally, we will use a dataset of targeted adversarial attacks provided through Kaggle's NIPS 2017: Defense Against Adversarial Attack competition. Lastly, using the CleverHans Python library, written by Ian Goodfellow, we can generate new adversarial attacks for custom networks and data sets.