Adversarial Inputs to Image Classification Neural Networks

Justin Ledford

Deep convolutional neural networks are now able to outperform humans on many image classification tasks. Although these models have superhuman performance on naturally occurring inputs, they fail to properly classify when exposed to slightly perturbed data points. Recent research has shown most of these high performing models suffer from this major flaw and that it is not a random artifact from training, but inherent across architectures. Methods have been developed to find minimal perturbations to inputs to cause neural networks to classify the input incorrectly with a given target label. These inputs are known in the literature as adversarial examples. As more and more commercial and consumer-facing systems utilize neural networks for image classification, it is important to understand the security threats to these systems and how we can protect against these attacks.

In this tutorial I provide an overview of various attack methods developed to show the weaknesses in these models, as well as defensive strategies that can be used to mitigate the problems resulting from these weaknesses. I also provide a review on neural networks and image classification to build the context necessary to understand how these attacks and defenses work. Examples of physical world applications are discussed, and finally a walkthrough of code used to generate these examples is given.