Adversarial examples and data-augmentation
Are modern ML and DL systems susceptible to attack?

About Google Brain

As one part of the overall Google AI effort, the Google Brain team works to advance the state of the art in artificial intelligence by research and systems engineering. A research oriented arm of one of the most recognized companies in the world, it has a long-term research agenda of making machines intelligent, and collaborates with partners to use the results of cutting edge research to improve people’s lives.

Problem Statement

Deep learning can potentially revolutionize several fields in science and technology. Two of the main concerns are its interpretability and reliability. The topic of adversarial examples is right at the heart of both of these concerns. Adversarial examples are samples that are very slightly perturbed, in a particular way, to get the machine learning algorithm to change its prediction from a correct class to an incorrect class. For example, on image recognition tasks, the accuracy of neural networks can be reduced to zero by perturbing the test set samples by a very small amount in an adversarial fashion. These perturbations can be so small that a human cannot perceive them, which raises the question: are machine learning algorithms reaching a non-trivial level of understanding to perform well on challenging tasks, or are they just memorizing? Furthermore, adversarial examples raise safety concerns: can the decision-making agents of the future that use machine learning be fooled easily by adversarial perturbations?

Data Resources

Students will focus on common generally available image datasets.

- MNIST
- CiFAR-10
- CiFAR-100
- ImageNet

Project Goals
● Successfully train deep neural networks that can classify digits in the MNIST dataset or pictures in the CIFAR10 dataset.
● Create adversarial examples and show that the trained networks are susceptible to these examples.
● Consider the effect of recent data-augmentation methods, like mixup, on adversarial sensitivity. Mixup trains a network on linear combinations of input images and the corresponding linear combinations of one-hot labels.
  ○ Specifically, if mixup is applied to only images of certain pairs of classes, does adversarial sensitivity decrease only for those classes or all classes?
● Implement adversarial training to reduce adversarial sensitivity. Then investigate whether mixup with adversarial training (network is trained on linear combinations of adversarial examples) can improve results on adversarial examples.
● Construct simulated models to investigate cross-technique and cross-technique transferability of adversarial samples
● Use adversarial samples to construct black-box attacks on third-party classification systems

References