Ryan Filgas
CVDL Week 8 Summary

The paper “Explaining and Harnessing Adversarial Examples” starts off by summarizing a problem. The problem in question is that several machine learning models including neural networks will misclassify slightly modified adversarial examples with high accuracy and this is still a very relevant problem still being researched. Adversarial training is possible to help mitigate this, but the best way is to switch from non-linear models. The adversarial examples in question are also so close to the originals they aren’t distinguishable by the human eye.

The idea of linear perturbation of non-inear models in short is that by making a bunch of small changes in the image it can add up to one big change that matches more closely the weights of a different classification. Scientists were able to fool state of the art algorithms on popular datasets by applying simple algorithms to the input. While this is alarming, deep networks do have some capacity to resist adversarial examples when trained on them, however there are still serious flaws. RBF networks are more resistant to this as they are non linear. In summary adversarial examples can be explained as a property of high dimensional dot products, a problem difficult for linear models to solve.