Ivan Miller. Reading Summary (Dimpled Manifold Model). 12.16.2022

1.Describe the framework given here for reasoning about adversarial examples that cause trained neural networks to make classification errors.

The authors proposed a new way of thinking about training neural networks and introduced a framework outlining how the decision boundary is changing during training. In the paper the authors demonstrated that training of a deep neural network is divided into two distinct phases: 1) they called the first phase “clinging process” in which the initially randomly oriented decision boundary gets very close to the low dimensional image manifold of all the training examples. 2) second phase is called “dimpling phase” which creates shallow bulges in the decision boundary that move the decision boundary to the correct side of the training examples.
Additionally “below” and “above” natural images from the image manifold there are other images that are being recognized by the network as the legitimate image classes even though they may not look like the original classes. Within that framework an adversarial examples could be created by going vertically a tiny distance ε towards the decision boundary and then continuing another ε on the other side (ε to reach the decision boundary, and another ε to get high confidence in the opposite label of the example, 2ε in total).

2.Describe one of the testable hypotheses that the framework produces.

One of the hypotheses made in the paper is that the decision boundary of a neural network tends to be very close to the image manifold. It is achieved by developing a large derivative in the vertical direction, which allows the network to bend the decision boundary more gently around the image manifold. That, in turn, makes the decision boundary simpler: making shallower dimples that pass on the correct sides of neighboring training examples of opposite classes, and makes it easier for the network to achieve higher accuracy.

3.Explain how this framework claims to resolve a mystery about the success of adversarial examples against trained neural networks.

The authors explain the mystery about success of adversarial examples against trained neural networks by comparing decision boundaries of a neural network trained on original images (N1) with the neural net that went through adversarial training on perturbed data with incorrect labels (N2). The image manifold for N2 is being shifted further away [a distance of 2ε, see #1] in the same direction as the N1 decision boundary, meaning that the adversarial training has the effect of deepening the dimples of the decision boundary. It means that if we were to compare the decision boundaries between N1 and N2 networks to each other, they would look quite similar, which explains the good accuracy of N2 being able to correctly classify original images.

4.In light of this, what do you think might be a promising path forward toward training neural networks that will be more resilient to adversarial examples without sacrificing too much accuracy on natural examples?

It may sound silly but I’d be curious to see if adding dropout layers in the different places of the network's architecture would help against adversarial examples so I would start with experiments around that. Frankly, I’m not quite sure I could fully grasp all the benefits of this new framework of thinking about training of the neural networks, but given the ease of pushing an example to the other side of the decision boundary I would probably focus on implementing some sort of system that will be checking the input data before training the network.