Participation 5

wangxinshi47

March 2023

As discussed in the guest lecture by Rado, neural networks are particularly vulnerable to adversarial attacks with minimal perturbation.

1. Why do traditional neural networks have this vulnerability?
   (NOTE: this should be a high-level analysis of the nature of neural networks; no need to dive into mathematical proofs)

   One reason is they are designed to optimize for accuracy on the training data, rather than robustness to perturbations in the input data. The objective function such as cross entropy does not take robustness into consideration.

   Another reason is Neural Network might choose a high variance in the context of bias-variance decomposition.They might also learn specific features that are not robust to changes in the input data.

2. Overfitting is the phenomena where a model learns details about the data that are correlated with the training samples but not the whole distribution. Underfitting is a phenomena where a model doesn't learn enough features about the training data. How does the idea of robustness compare with these phenomena?

   Robustness is used to measure how a model performs under the presence of noise or adversarial data.

   Usually a Neural Network that overfits is more vulnerable to adversarial attack. It might learn some noise from the data and misinterpret the pattern. A Neural Network that underfits might be less vulnerable to adversarial attack as it captures the big picture and learns less from the noise.

3. What thoughts do you have on the material covered in the guest lecture?

I like this guest lecture. I taught me an optimization problem for finding the adversarial sample. I had never heard of that algorithm before.