WEEKLY PARTICIPATION 5: NEURAL NETWORK ROBUSTNESS

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

(a) 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)

(b) 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?

(c) What thoughts do you have on the material covered in the guest lecture?
(NOTE: This is a free-form question; your response will not impact your grade as long as a reasonable response is given)