Analysis of Simplicity Blas in Neural Networks

Shivaram
BE20B032

Simplicity Bias

- Neural networks learn simple models over complex predictors.
- Reduces robustness since low margin predictors are chosen over the large margin predictors.
- The network might have higher confidence, even if complex features contradict the simple feature.
- Extreme SB can hurt Robustness and generalization.

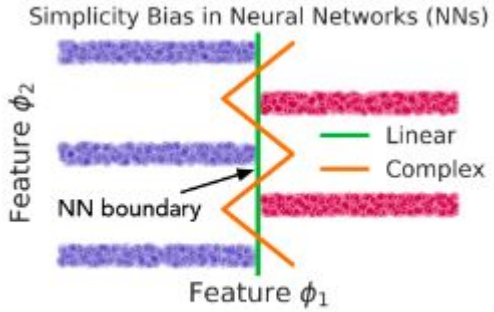

Figure 1: Simple vs. complex features

Experiments

- **Simplicity Bias in Deeper Networks:** The “Pitfalls of Simplicity Bias” paper talks about the 1-hidden layer network, in this experiment we could test how depth plays a role in deeper nets
- **Role of dimension of data in Simplicity Bias:** In the paper they have 2 features and have added gaussian noise for all other dimensions, we could test out how the weight on feature depends on the dimensions of the data.
- **Number of epochs and Simplicity Bias**
- **Can dropout layer reduce simplicity Bias?**
- **Testing out the effectiveness of ensembling.**

References:

1) The Pitfalls of Simplicity Bias in Neural Networks. Shah et al. NeurIPS 2020.
2) SGD on Neural Networks Learns Functions of Increasing Complexity