Adrian Kukla

CS7643 HW2

Question 7

a)

**Key contributions**

Paper shows that deep neural networks can perfectly fit random labels.

Shows that explicit regularization like dropout or weight decay are neither necessary nor sufficient for explaining generalization. The paper suggests that is it **not** likely that regularization techniques are the main reason for generalization.

Shows that any depth 2-layer networks of linear size with ReLU activations can represent any labelling of the training data (any function) which shows that deep neural nets have very high capacity for learning.

Shows that SGD implicitly regularizes solutions which leads to generalization. SGD even with unchanged parameters can optimize weights to fit random patterns perfectly even when there is no relationship between labels and the associated images.

**Strengths**

Authors of the paper refer to works of multiple other authors/papers to provide some context.

Work of authors challenges conventional wisdom rather than conforming to the status quo – which can spur more research and innovations.

Study does conduct robust experiments on a number of different types of neural nets with and without regularization to evaluate their hypothesis. Combines empirical findings with ground neural network theory.

**Weaknesses**

No explanation about why some networks generalize better than others.

Doesn't offer any solutions or techniques to improve understanding of generalization.

Doesn't offer alternatives to classical and established measures.

b)

Personal takeaways

My understanding was that overfitting tends to occur when number of parameters is higher than number of samples – however this paper provides a counter to my previous understanding. Moreover, I thought that explicit regularization techniques like dropout or weight decay is necessary for generalization, whereas the paper shows that they're not. Changing the model architecture appropriately can vastly improve generalization without use of explicit regularization terms.

Interesting that the paper ruled out common measures like VC dimensions, Rademacher complexity as potential explanations for generalization performance. The paper suggests that

new theoretical frameworks should be explored which are outside of the commonly accepted measures.

Moreover, potential future research would be to evaluate what common properties of models that were trained by SGD to analyse how these models generalize well without the need for explicit regularization.