Derek Lee Deep Learning Fall 2020

Professor Curro Quiz #11

1. Pick a paper here: https://proceedings.neurips.cc/paper/2020. Write something about it. Think about if you were writing a quiz, what might you want people to engage with?

https://proceedings.neurips.cc/paper/2020/file/9b8619251a19057cff70779273e95aa6-Paper.pdf

This paper challenges the idea that normalization is an essential component to training a neural network. With their unnormalized RescaleNet, the authors achieve results that are either equivalent or slightly better than the corresponding normalized models. The model offers a solution to the exploding variance problem. The authors explain that the variance doubles at each residual block, and thus "increases exponentially with the number of residual blocks". To deal with this problem, the authors introduce two parameters, α and β. These hyperparameters are defined by $\alpha^2 + \beta^2 = 1$. These hyperparameters are used to alter the residual block:

$$x_k = \alpha_k x_{k-1} + \beta_k \mathrm{F}_k(x_{k-1}).$$

With this change, the variance remains stable at each residual block. Specifically, $Var[x_k] = Var[x_{k-1}]$. The authors define these parameters in terms of a single hyperparameter, c.

$$\alpha_k = \sqrt{\frac{k-1+c}{k+c}} \text{ and } \beta_k = \frac{1}{\sqrt{k+c}}$$

The paper also introduces a novel method for Bias Initialization. In models with normalization, "the preceding linear layer usually has no bias since it may be cancelled by the mean reduction operation of normalization." Since RescaleNet does not use normalization, it is necessary to properly initialize the bias term. The convention is to apply bias after the matrix multiplication, $\boldsymbol{y} = \boldsymbol{W}(\boldsymbol{x}) + \boldsymbol{b}$ and to initialize the bias to 0. The authors propose applying the bias before the matrix multiplication, $\boldsymbol{y} = \boldsymbol{W}(\boldsymbol{x} + \boldsymbol{b})$, and initialize the bias as the negative of the mean of the first mini-batch of data.

This model achieves results that are marginally better than normalized models. The results of this paper demonstrate that unnormalized models are viable. It offers evidence against some of the common theories for why normalization is effective. This paper serves as a theoretical contribution to the study of normalization.