Derek Lee Deep Learning Fall 2020

Professor Curro Quiz #4

1. Explain how the main technique from "Delving Deep into Rectifiers" may alleviate a core issue in the training scheme for "Going Deeper with Convolutions." (Think about what they did to "get-it-to-work")

In "Delving Deep into Rectifiers", the authors proposed a new generalization of ReLU: the Parametric Rectified Linear Unit (PReLU). The PReLU is similar to the Leaky ReLU, except that the parameter for the negative side is learned, rather than constant. Additionally, the authors propose a "theoretically sound initialization method", which assists convergence on deep models. The authors address the nonlinearities present in ReLU and PReLU, noting that a previous strategy for initialization assumes linear activations, which do not apply in this case.

In "Going Deeper with Convolutions", the authors mentioned that because of the "relatively large depth of the network, the ability to propagate gradients back through all the layers in an effective manner was a concern." They used auxiliary classifiers connected to intermediate layers to boost the gradient signal, mitigating the dead ReLU problem. If the authors used the PReLU instead, it is possible that the auxiliary classifiers would not be needed because of the learned parameter of the PReLU.

In "Going Deeper", the authors mentioned that the main limitation in training was memory usage. The network would require one week to train to convergence. The initialization method proposed in "Delving Deep" can be used in "Going Deeper" to achieve faster convergence.

2. Explain the core effect of pre-activation from "Identity Mappings in Deep Residual Networks" compared to the original residual formulation.

The authors propose a new residual unit, one that has "a 'direct' path for propagating information". Instead of the approach taken by a normal residual unit, where $h(x_l) = x_l$ is an identity mapping and $f(y_l)$ is a ReLU, the authors propose $f(y_l)$ as an identity mapping as well, in addition to modifying the architecture of the residual unit. This creates some "nice properties. The feature x_L of any deeper unit L can be represented as the feature x_l of any shallower unit l

plus a residual function in a form of $\sum_{i=l}^{L-1} F$. This also leads to "nice backward propagation properties." The gradient "can be decomposed into two additive terms: a term… that propagates information directly without concerning any weight layers, and another term… that propagates through the weight layers. The additive term… ensures that information is directly propagated back to *any shallower unit l*.

This change eases optimization and improves regularization. On the tests used by the authors, the training loss decreases very quickly, and results in the lowest loss among all models tested. The authors speculate that Batch Normalization is primarily responsible for the regularization effect. In the original residual unit, batch normalization normalizes the signal, but the normalized signal is then added to the shortcut, resulting in an unnormalized signal. This signal is then used as the input to the next weight layer. In the new residual unit, the input to every weight layer are normalized.