MA 598 MACHINE LEARNING SEMINAR
SUMMARY 4

ID 5107

4. Deep Residual Learning for Image Recognition

What is their primary result? The authors (He, Zhang, Ren, and Sun) address the degradation problem faced by very deep neural networks by introducing layers which optimize the residual mapping.

Why is this important? The more layers a neural network has the better it appears to get at image classification tasks. Adding too many layers, however, comes at a cost and such network suffers from overfitting, exploding/vanishing gradient, and especially degradation. The question of exploding/vanishing gradients has been addressed; the problem is ameliorated by adding a normalized initialization layer as well as an intermediate normalization layer. The authors

The authors address the last of these problems—degradation (the loss in accuracy of the neural network at its task). They propose a solution to the degradation problem by introducing a deep residual learning framework.

What are their key ideas? If $\mathcal{H}$ is the desired underlying mapping, let $\mathcal{F}(\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \mathbf{x}$. The authors hypothesize that it is easier for the network to optimize the residual mapping $\mathcal{F}$ by recasting the mapping problem from $\mathcal{H}$ to $\mathcal{F}(\mathbf{x}) + \mathbf{x}$. This is especially obvious in the case that $\mathcal{H}$ is the identity mapping, as the nonlinear layers can better push $\mathcal{F}$ to 0 than $\mathcal{H}$ to the identity.

What are the limitations, either in performance or applicability? The authors employ their framework on both a 110-layer network and a 1202-layer network, and notice a drop in testing accuracy which they attribute to overfitting. The authors hope to combine their framework with stronger regularization in the future to diminish the effects of overfitting.

What's the architecture? The authors implement two architectures. One is based on a plain network inspired by VGG nets, and the other is a residual network with shortcut connections. The most interesting aspects of the network architectures follow.

For the plain network, the authors propose two simple design rules: (i) for the same output feature map size, the layers have the same number of filters; and (ii) if the feature map size is halved, the number of filters is doubled to preserve the time complexity of each layer.

The residual architecture employs two strategies when the dimension of the network increases. The first is perform the identity mapping on the shortcut connections, but pad the entries with extra zeros to account for the change in dimension. The second

---

Date: January 25, 2019.

strategy is to use a projection on the shortcut connections, i.e., use a projection matrix W_s in the equation $\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + W_s\mathbf{x}$ to match the dimensions of the network.

How did they train and evaluate it? The networks are trained on the ImageNet 2012 classification dataset, primarily. They are also tested on PASCAL VOC 2007 and 2012 and COCO.