Dilated Residual Networks Preliminary Summary

Shishir Jakati
University of Massachusetts, Amherst
sjakati@umass.edu

Abstract

Convolutional networks have the unfortunate and debilitating property of reducing the resolution of their inputs. Dilated convolutional layers attempt to solve the problem of this input constriction by exponential increasing the receptive field of the layer. Additionally, residual networks address the problem of learning a more accurate spatial representation of an input. By incorporating the two, Dilated Residual Networks[2] (DRNs) allow networks to capture a more accurate representation of objects and their relative position in the input.

1. Residual Networks

Plain networks, as they are referred to by [1], generally do not share the output of layers with multiple layers. In a residual block, the output of the layer is fed directly into the next layer, but also into layers which are deeper than the next layer. Ultimately, these skip connections allow the network to propagate larger gradients to shallower layers and learn as quickly as the deep layers.

Formally, the deep residual network framework adds a direct "shortcut" connection between the shallow layers and the deeper layers. While the deeper layers all receive the activated output of the shallower layers directly, the deeper layers also receive the identity mapping of the shallow layer with the shortcut connection. Again, this identity mapping combats the degradation problem, by providing the residual.

2. Dilated Convolutions

The purpose of a convolution operation, in the context of a convolutional neural network is to provide some receptive field over the inputs that a particular layer receives. Usually, this receptive field is a contiguous rectangular area over the input, which then maps into another contiguous rectangular area as an output. Dilated convolutions dilate the contiguous rectangular area by introducing separations between segmented areas of input. In doing so, dilated convolutions effectively increase the receptive field by an exponential factor.

2.1. Degridding

Due to the segmented nature of the gridding artifacts may surface. This is caused partially by the use of max pooling layers in the original ResNet[1] architecture. These pooling layers generate areas within the activation map with high amplitude activations, which are then propagated to subsequent layers and exaggerate the effect of gridding. By simply removing these max pool layers, the effects of gridding are reduced.

Two separate techniques to combat gridding, as presented in the original publication, are to increase the number of convolutional layers and to remove specific residual connections within the network. When increasing the number of convolutional layers, a block of layers is added with a decreasing dilation factor. Removing residual connections discourages gridding artifacts from propagating further down a deep network.

3. Relevance to Research

In the problem of scene text recognition, individual character recognitions are difficult to obtain. This is due to problems prevalent in scene text recognition, but heightened due to the smaller "surface area" of character instances. Convolutional networks may have the problem of degradation, or a reduce spatial representation. Using DRNs may alleviate the problems of constriction by increasing the receptive field to generate a spatial representation that is more attune to classify character representations. Additionally, employing the residual network framework will allow the use of deeper networks while not invoking the degradation problem.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. 1

[2] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks, 2017. 1