GNR-638 (Paper Review - CVPR 2017)
Xception: Deep Learning with Depth Wise Separable Convolutions

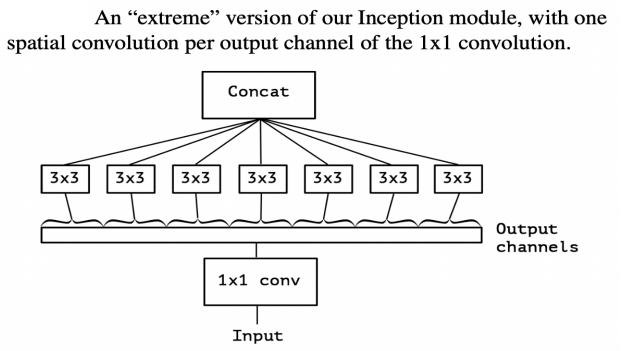
Author: Francois Chollet

Motivation Behind Choosing this Paper

We wanted to study the architecture with arguably the coolest name. We also wanted to find out what made Xception Net so extreme and a better model architecture than the already awesome Inception family of networks.

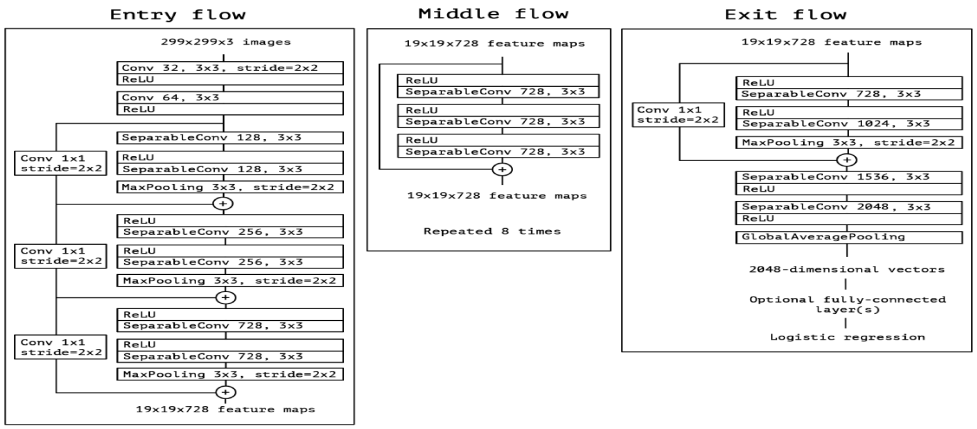
Details

“Xception” which means “Extreme Inception,” is a model which improves upon the Inception V3 model. It is a convolutional neural network architecture based entirely on depthwise separable convolution layers. It in effect entirely decouples/separates the mapping of cross-channel correlations and spatial correlations in the feature maps of convolutional neural networks.



The Xception architecture is a linear stack of modified depthwise separable convolution layers with residual connections. The modified depthwise separable convolution is the pointwise convolution followed by a depthwise convolution (different from the normal depthwise separable convolutions where the depthwise Conv operation precedes the pointwise Conv operation.)

Analysis



As in the figure on the side, SeparableConv is the modified depthwise separable convolution. We can see that SeparableConvs are treated as Inception Modules and placed throughout the whole deep learning architecture.

And there are residual (or shortcut/skip) connections, originally proposed by ResNet, placed for all flows.

Experiments Conducted

Two datasets are tested, ImageNet and JFT. ImageNet dataset has 1000 classes, Xception outperforms the VGGNet, ResNet, and Inception-v3. It is also noted that in terms of error rate, the relative improvement isn’t tiny (6.78% improvement).

To evaluate using the model trained on JFT, an auxiliary dataset was used, FastEval14k(A dataset of 14,000 images with dense annotations from about 6,00 classes; 36.6 labels per image on average.). Xception slightly outperforms Inception V3 on the ImageNet dataset which has 1000 classes (which Inception V3 was designed for) and significantly outperforms Inception V3 on a larger image classification dataset comprising 350 million images and 17,000 classes.

Further Improvements

If we use L2-norm regularization, then the accuracy of the model can be increased.

A dense architecture with residual steps between one layer to multiple layers down the model may improve the accuracy by helping create correlations between initial (simpler) features learned and the more complex features learned by Xception modules after it.

The author says that there is no reason to believe that depthwise separable convolutions are optimal. It may be that intermediate points on the spectrum, between regular inception modules and depthwise separable convolutions, hold further advantages. So further research on this by experimentally finding out the optimal extent to which spatial and cross-channel correlations must be decoupled to get the best results.

References:

https://arxiv.org/pdf/1610.02357.pdf

Paper Reviewed by:

Yash Mailapalli (200050160)

P Balasubramanian (200050103)