Supplementary Explanation on Attention Mechanism

1. Feature Map

   The output of a 2D-convolution layer is a tensor which can be denoted as F^{m*n*k}, where m and n are width and height of the "image" in each channel respectively and k is the number of channels. Each channel in this tensor is a potential representation of the of the original image in lower dimension.

2. Global average pooling and attention map

   There are 2 ways to implement a global average pooling (GAP). One way is taking the average of all the pixels of the image in each channel. This will convert the feature map F^{m*n*k} to a 1D array A^{k}. This layer is usually followed by a dense layer with softmax activation to predict the class. The matrix for the dense layer is W^{k*p}, where p is the number of class. Each column in this matrix indicates the importance or weight of each corresponding channel in deciding the class p. The region where the neural network is trying to look at is the region with higher weights. In order to visualize which part of the image the neural network is focusing on when deciding a specific class, the class activation map or attention map is calculated by multiplying the feature map F^{m*n*k} with the specific column c_1 in $W^{k}{}_{c1}$. The process is shown in the Figure 1.

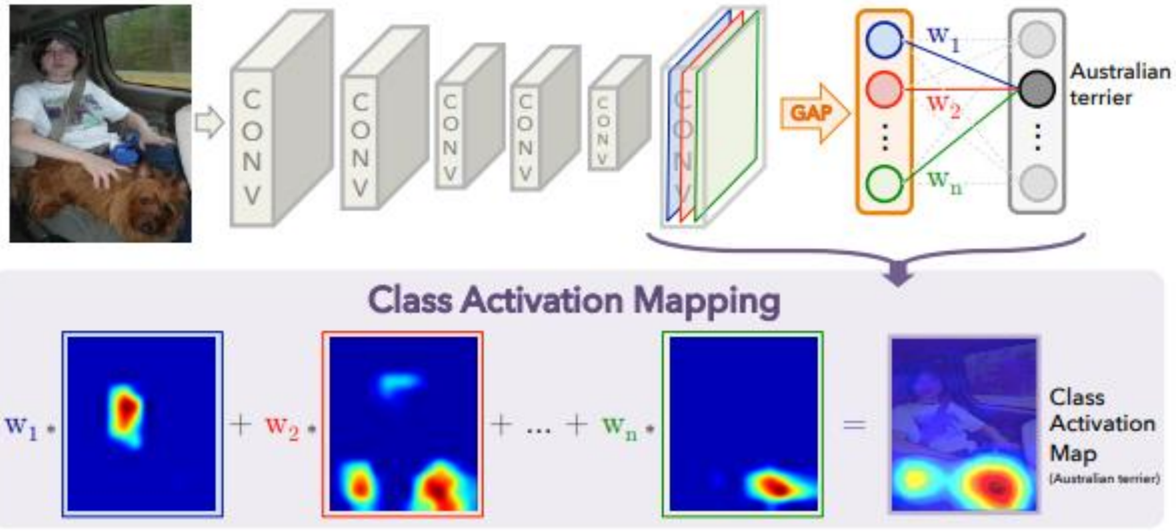

Figure 1. The class activation mapping [1]

The previous map is giving a channel-wise attention map. The other way is taking the average of the values in all the channels for each pixel in an image. This will convert the feature map F^{m*n*k} to a 2D image A^{m*n}, which gives a spatial-wise attention map. This image can then be flattened out and connected to a dense layer with softmax activation to make prediction.

Usually the attention map is created by global average pooling instead of global max pooling (GMP). The global max pooling is dominated by the maximum value on each channel, which lose some generality in representing the channel or pixels. The global average pooling, on the other hand, considers every pixel in a channel or every channel for a single pixel. In fact, in the work of Sanghyun et. al [2], they combined all these possible method for developing attention. For a feature map F^{m*n*k} of the image, the GAP and GMP are applied to every channel to get two 1D arrays. Then these arrays are passed by a fully connected layer to reduce the noise. Then the modified arrays are added together to create a modified channel attention, which is applied on the feature map to get F'^{m*n*k}. Then the GAP and GMP are applied for each pixel along all the channels to get two 2D maps. These maps are concatenated and followed by a convolution layer to create a 2D spatial attention map. This map is then used to modify the feature map F' to create F'', which is then used as input in the following layers.

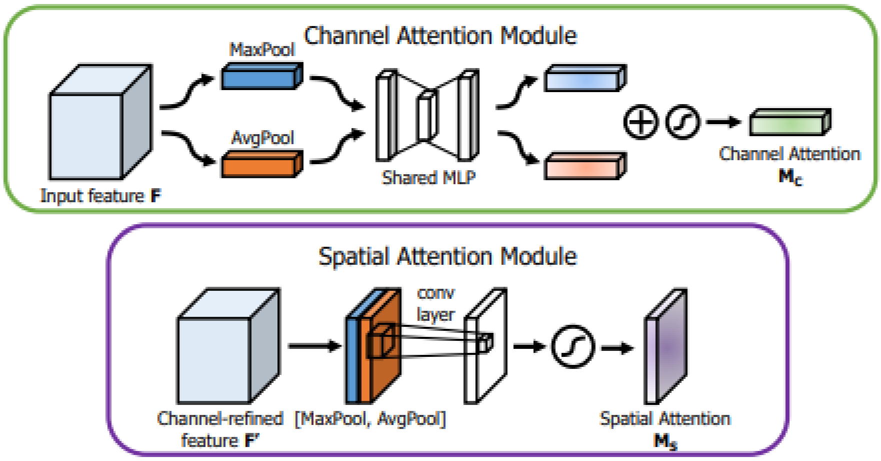

Figure 2. The channel and spatial attention maps [2]

3. Reference

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. CVPR'16 (arXiv:1512.04150, 2015).
[2] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon. CBAM: Convolutional Block Attention Module. ECCV 2018(arXiv:1807.06521)