Compact Fully-Connected Layer Computation

Dat Huynh

January 16, 2021

A fully-connected layer has the form:

$$y = f(Wx + b), \tag{1}$$

where W, b are learnable linear transformation and bias vector, respectively, and $f(\cdot)$ is an element-wise activation function such as ReLU. Let assume we have N input dimension $x \in R^N$. If we break it down into each output dimension j, we would have:

$$y_j = f\Big(\sum_{i=1}^{N} W_{i,j}x_i + b_j\Big), \tag{2}$$

which requires 2 steps of multiplication and addition. To perform this computation in a single multiplication step, we can augment x by 1 and extend the shape of W appropriately. It can be shown that $f(W[x|1])$ is:

$$f\Big(\sum_{i=1}^{N} W_{i,j}x_i + W_{N+1,j}x_{N+1}\Big) = f\Big(\sum_{i=1}^{N} W_{i,j}x_i + W_{N+1,j}\Big) \tag{3}$$

for each output dimension j. Notice that $x_{N+1} = 1$ since we augment x by 1. If we look closely, this is equivalent to y_j as last term $W_{N+1,j}$ can be considered as b_j since they are both independent of x. Thus, $f(W[x|1])$ can be considered as a compact way to compute $f(Wx + b)$