

TTIC 31230 Fundamentals of Deep Learning, winter 2019

CNN Problems

Problem 1. Consider convolving a filter $W[\Delta x, \Delta y, i, j]$ with thresholds $B[j]$ on a “data box” $L[b, x, y, i]$ where $B, X, Y, I, J, \Delta X, \Delta Y$ are the number of possible values for $b, x, y, i, j, \Delta x$ and Δy respectively. How many floating point multiplies are required in computing the convolution on the batch (without any activation function)?

Problem 2: Suppose that we want a video CNN producing layers of the form $L[b, x, y, t, i]$ which are the same as the layers of an image CNN but with an additional time index. Write the equation for computing $L_{\ell+1}[b, x, y, t, j]$ from the tensor $L_{\ell}[B, X, Y, T, I]$. Your filter should include an index Δt and handle a stride s applied to both space and time.

Problem 3: Images have translation invariance — a person detector must look for people at various places in the image. Translation invariance is the motivation for convolution — all places in the image are treated the same.

Images also have some degree of scale invariance — a person detector must look for people of different sizes (near the camera or far from the camera). We would like to design a deep architecture that treats all scales (sizes) the same in a manner that similar to the way CNNs treat all places the same.

Consider a batch of images $I[b, x, y, c]$ where c ranges over the three color values red, green, blue. We start by constructing an “image pyramid” $I_s[x, y, c]$. We assume that the original image $I[b, x, y, c]$ has spatial dimensions 2^k and construct images $I_s[b, x, y, c]$ with spatial dimensions 2^{k-s} for $0 \leq s \leq s_{\max} < k$. These images are defined by the following equations.

$$I_0[b, x, y, c] = I[b, x, y, c]$$

$$I_{s+1}[b, x, y, c] = \frac{1}{4} \begin{pmatrix} I_s[b, 2x, 2y, c] + I_s[b, 2x+1, 2y, c] \\ + I_s[b, 2x, 2y+1, c] + I_s[b, 2x+1, 2y+1, c] \end{pmatrix}$$

We want to compute a set of layers $L_{s,\ell}[b, x, y, i]$ where s is the scale and ℓ is the level of processing. First we set

$$L_{0,s}[b, x, y, c] = I_s[b, x, y, c].$$

The layers $L_{\ell,0}[b, x, y, i]$ can be computed using the standard CNN equations holding the scale at zero.

Give an equation for a linear threshold unit to compute $L_{\ell+1,s+1}[b, x, y, j]$ from $L_{\ell,s+1}[b, x, y, j]$ and $L_{\ell+1,s}[b, x, y, j]$. Assume that the spatial dimension of $L_{\ell,s}$ is 2^{k-s} and use an appropriate stride between $L_{\ell+1,s+1}[b, x, y, j]$ and $L_{\ell+1,s}[b, x, y, j]$. Use parameters $W_{\ell+1,\rightarrow}[\Delta x, \Delta y, i, j]$ for the dependence of $L_{\ell+1,s}$ on $L_{\ell,s}$ and parameters $W_{\ell+1,\uparrow}[\Delta x, \Delta y, i, j]$ for the dependence of $L_{\ell+1,s+1}$ on $L_{\ell+1,s}$. Use $B_{\ell}[j]$ for the threshold. Note that these parameters do not depend on s — they are scale invariant.