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, Δx and Δy respectively. How many floating point multiplies are required in computing the convolution on the batch (without any activation function)?

Solution:

$$BXY\ \Delta X\ \Delta Y\ IJ$$

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.

Solution:

$$L_{\ell+1}[b, x, y, t, j] = \sum_{\Delta x, \Delta y, \Delta t, i} W[\Delta x, \Delta y, \Delta t, i, j] L_\ell[b, sx + \Delta x, sy + \Delta y, st + \Delta t, i]$$

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 $0 \leq s \leq k$. The image pyramid $I_s[b, x, y, i]$ for $0 \leq s \leq k$ is defined by the following equations.

$$\begin{aligned} I_0[b, x, y, c] &= I[b, x, y, c] \\ I_{s+1}[b, x, y, c] &= \frac{1}{4} \left(\begin{array}{l} I_s[b, 2x, 2y, c] + I_s[b, 2x+1, 2y, c] \\ +I_s[b, 2x, 2y+1, c] + I_s[b, 2x+1, 2y+1, c] \end{array} \right) \end{aligned}$$

We want to compute a set of layers $L_{\ell,s}[b, x, y, i]$ where s is the scale and ℓ is the level of processing with $\ell + s \leq k$ and where $L_{\ell,s}[b, x, y, i]$ has spatial dimensions

$2^{k-\ell-s}$ (increasing either the processing level or the scale reduces the spatial dimentions by a factor of 2). First we set

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

Give an equation for a linear threshold unit to compute $L_{\ell+1,s}[b, x, y, j]$ from $L_{\ell,s}[b, x, y, j]$ **and** $L_{\ell,s+1}[b, x, y, j]$. Use parameters $W_{\ell+1,\leftarrow}[\Delta x, \Delta y, i, j]$ for the dependence of $L_{\ell+1,s}$ on $L_{\ell,s+1}$ and parameters $W_{\ell+1,\uparrow}[\Delta x, \Delta y, i, j]$ for the dependence of $L_{\ell+1,s}$ on $L_{\ell,s}$. Use $B_{\ell+1}[j]$ for the threshold. Note that these parameters do not depend on s — they are scale invariant.

Solution:

$$L_{\ell+1,s}[b, x, y, j] = \sigma \left(\begin{array}{ll} & \sum_{\Delta x, \Delta y, i} W_{\ell+1,\leftarrow}[\Delta x, \Delta y, i] L_{\ell,s+1}[b, x + \Delta x, y + \Delta y, i, j] \\ + & \sum_{\Delta x, \Delta y, i} W_{\ell+1,\uparrow}\ [\Delta x, \Delta y, i]\ L_{\ell,s}\ [b, 2x + \Delta x, 2y + \Delta y, i, j] \\ + & B[j] \end{array} \right)$$

Note: I am not aware of this architecture in the literature. A somewhat related architecture is "Feature Pyramid Networks" arxiv 1612.03144.

Problem 4. This problem is on initialization. Consider a single unit defined by

$$u = f\left(\left(\sum_{i=1}^{N} W[i]x[i]\right) - B\right).$$

where B is initialized to zero and f is an activation function such as a sigmoid or ReLU. The vector x is a random variable determined by a random draw of a training example. Assume that the components of x are independent and that each component has zero mean and unit variance. Suppose that we initialize each weight in W from a distribution with zero mean and variance σ and that is symmetric about zero — (the probability that $w[i] = z$ equals the probability that $w[i] = -z$). For example, $x[i]$ might be distributed as a zero-mean unit-variance Gaussian. Consider $y = \sum_i W[i]x[i]$ as a random variable defined by the distribution on x and the independent random distribution on W. Recall that the variance σ^2 of a sum of independent random variables is the sum of the variances and the variance of a product of zero mean independent random variables is the product of the variances.

(a) What value of the initialization variance σ for $W[i]$ gives zero mean and unit variance for y if the vectors $w[I]$ and $x[I]$ have dimension d? Show your derivation.

Solution: Let σ^2 be the variance of $x[i]$. We then have that the variance of $\sum_i W[i]x[i]$ is $\sum_i \sigma^2 = d\sigma^2$. Setting $d\sigma^2$ equal to 1 gives

$$\sigma = \frac{1}{\sqrt{d}}$$

(b) For a sigmoid activation function what is the mean of u.

Solution: We are given that the probability that $W[i] = z$ is the same as the probability of $w[i] = -z$. This implies that for a given value of $x[i]$ we have that the probability that $w[i]x[i] = z$ equals the probability that $w[i]x[i] = -z$. This further implies that, for a given value y, the probability that $\sum_i w[i]z[i] = y$ equals the probability that $\sum_i w[i]x[i] = -y$. So the input to the sigmoid is distributed symmetrically about 0. Since the sigmoid function is itself symmetric about 0, we get that the expected value of the output of the sigmoid is its value at zero which is 1/2.

(c) For a sigmoid activation function is the variance of u larger than, equal to, or smaller than the variance of y?

Solution: The variance is smaller. To show this it suffices to show that the slope of the sigmoid function is everywhere less than 1. The slope is largest at the input zero. The sigmoid function is

$$f(z) = \frac{1}{1 + e^{-y}}$$

The slope is

$$f'(y) = \frac{e^{-y}}{(1 + e^{-y})^2}$$

which equals 1/4 at $y = 0$.

(d) What is the largest possible variance of the output of a sigmoid?

Solution: The larest valriance occurs when $y = \infty$ with probability 1/2 and $y = -\infty$ with probability 1/2 ;-). Tn this case $f(y)$ is 0 with probability 1/2 and 1 with probability 1/2. Which gives a variance of 1/4.

Problem 5. This problem is on the initialization of ResNet filters. Consider the following residual skip connection where $R_{\ell+1}$ is computed with an $N \times N$ filter.

$$\begin{array}{l} \text{for } b, x, y, j\Delta x, \Delta y, j' \\ \qquad R_{\ell+1}[b, x, y, j] \quad \texttt{+=} \quad W_{\ell+1}[\Delta x, \Delta y, j', j]\ L_\ell[b, x + \Delta x, y + \Delta y, j'] \end{array}$$

$$\begin{array}{l} \text{for } b, x, y, j \\ \qquad R_{\ell+1}[b, x, y, j] \quad \texttt{-=} \quad B_{\ell+1}[j] \end{array}$$

$$\begin{array}{l} \text{for } b, x, y, j \\ \qquad L_{\ell+1}[b, x, y, j] \quad = \quad L_\ell[b, x, y, j] + R_{\ell+1}[b, x, y, j] \end{array}$$

Here we have omitted an activation function that would be present in practice. This omission allows an analysis that seems to provide insight into the more complex case with activations.
Assume that $L_0[b, x, y, j]$ is computed from the input in some unspecified way such that $L_0[b, x, y, j]$ has unit variance. Assume that the values $L_\ell[b, x, y, j]$ and $R_{\ell+1}[b, x, y, j]$ are all independent. Suppose that each weight $W_\ell[\Delta x, \Delta y, j, j']$ is drawn independently at random from a distribution with zero mean and variance σ_W. Recall that the variance σ^2 of a sum of independent random is the sum of the variances and the variance of a product of independent random variables is the product of the variances.

(a) Give an expression for the variance σ_ℓ^2 of $L_{\ell+1}[b, x, y, j]$ as a function of ℓ, the filter dimension $D = \Delta X = \Delta Y$, the feature dimension J, and the weight variance σ_W^2.

Solution: Assuming everything is independent we have

$$\begin{aligned} \sigma_{\ell+1}^2 &= \sigma_\ell^2 + D^2 J \sigma_w^2 \sigma_\ell^2 \\ &= \sigma_\ell^2 (1 + D^2 J \sigma_w^2) \end{aligned}$$

This gives

$$\sigma_\ell^2 = (1 + D^2 J \sigma_w^2)^\ell$$

(b) Using $(1 + \epsilon)^N \approx e^{\epsilon N}$ solve for the value of σ_W such that $\sigma_L^2 = 2$.

Solution:

$$\begin{aligned} \sigma_L^2 &= (1 + D^2 J \sigma_w^2)^L \\ &\approx e^{L D^2 J \sigma_W^2} \end{aligned}$$

setting

$$e^{L D^2 J \sigma_W^2} = 2$$

gives

$$\sigma_w \approx \sqrt{\frac{\ln 2}{L D^2 J}}$$

(c) Assuming $L_L.\text{grad}[b, x, y, j]$ has unit variance, and that all components of $L_\ell.\text{grad}[b, x, y, j]$ and $R_\ell.\text{grad}[b, x, y, j]$ are independent, give an expression for the variance $\sigma_{\ell,\text{grad}}^2$ of the components of $L_\ell.\text{grad}[b, x, y, j]$ as a function of ℓ, D, J and σ_W.

Solution: We have

For $b, x, y, j, \Delta x, \Delta y, j'$

$$L_\ell.\text{grad}[b, x + \Delta x, y + \Delta y, j'] \texttt{ += } W_{\ell+1}[\Delta x, \Delta y, j, j']L_{\ell+1}.\text{grad}[b, x, y, j]$$

which gives

$$\sigma^2_{\ell,\text{grad}} = (1 + D^2 J \sigma_W^2)^{L-\ell}$$