TTIC 31230 Fundamentals of Deep Learning, winter 2019

Highways Problems

Problem 1. This problem is on initialization (slide 23 of lecture 4). Consider a single unit defined by $u = Wx + b$ where u and v are vectors, W is a $d_u \times d_x$ matrix and b is a bias vector initialized to zero. Assume that each component of x has zero mean and unit variance. Suppose that we initialize each weight in W from a zero mean Gaussian distribution with variannce σ. Consider u as a random variable defined by the distribution on x and the independent random distribution on W.

a) What value of the initialization variance σ gives zero mean and unit variance for each component of u? Show your derivation.

b) Consider $u' = \text{relu}(Wx + b) - b'$. If we want u' to have zero mean and unit variance will this increase or decrease the variance of the random distribution used to initialize W? Explain your answer.

Problem 2. The equations defining a UGRNN are given below.

$$\begin{aligned} h^{t+1} &= f^t \odot h^t + (1 - f^t) \odot d^t \\ f^t &= \sigma(W^f[x^t, h^t] + b^b) \\ d^t &= \tanh(W^d[x^t, h^t] + b^d) \end{aligned}$$

a) Rewrite these as an equivalent set of equations with the vector concatenations replaced with a pair of matrix multiplications where W^f is replaced by two matrices $W^{f,x}$ and $W^{f,h}$ and similarly for W^d.

b) Translate the equations from part (a) into explicit index notation with explicit summations including the batch index.

c) Give explicit index expression for the backward method for h^{t+1}. Your equations should compute additions to f^t.grad, h^t.grad and d^t.grad.

Problem 3. This problem is on the initialization of Resnet filters. Consider the following residual skip connection where the diversion is a convolution with an $N \times N$ filter.

$$x^{\ell+1} = x^\ell + \text{Conv}(W^\ell, x^\ell)$$

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 of adding a relu activation around the convolution.

Consider a network of L layers of this equation, all of which are done stride 1 so that the image dimensions are preserved and with a different weight matrix W^ℓ at each layer. Assume that each image x^ℓ has C channels (ignore the fact that input images have only three channels). Assume that each channel of each pixel of each image is independent of the other channels and assume that each channel

value of each pixel of the input image is drawn independently with zero mean and unit variance. Also suppose that the weights of the matrices W^ℓ are each drawn at random from a Gaussian distribution with zero mean and variance σ_W. Also assume that the the two terms in the sum of the residual skip equation above are independent (although they aren't). Just assume everything is independent and 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 $\sigma^2_{\ell+1}$ of each channel value at layer $\ell+1$ as a function of the variance σ^2_ℓ at layer ℓ and the weight variance σ_W.

b) Assume that the input layer x^0 has independent channel values each with variance 1. Give an expresison for the variance σ^2_ℓ directly as a function of σ_w and ℓ.

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

d) Assuming top level gradient x^L.grad has zero mean and unit variance, and that all components of $x^{\ell+1}$.grad are independent, give an expression for the variance $\sigma^2_{\ell,\text{grad}}$ of the components of x^ℓ.grad as a function of ℓ and σ_W.

e) Consider the limit of $\sigma_W \to 0$. Give an explicit index expression for the limit as $\sigma_W \to 0$ of W^ℓ.grad. Your expression should be in terms of x^L.grad. If we add an activation function, does W^ℓ.grad have a well defined limit as $\sigma_W \to 0$?