Deep Learning (SS2021)
Seminar 3

March 19, 2021

Assignment 1 (Backprop of scan) The *inclusive cumulative sum* or for brevity *scan* operation is defined as follows: Given the input vector $x \in \mathbb{R}^n$ the output $y \in \mathbb{R}^n$ has components:

$$y_i = \sum_{j \leq i} x_j.$$

Compute the backprop of scan, i.e. given $\nabla_y L$ compute $\nabla_x L$.

Assignment 2 (Sampling with Replacement)

a. Let the dataset contain n points. During an epoch, we draw a random point n times. What is the probability that point i has been drawn at least once? What is the limit of this probability as $n \to \infty$.

Hint1: Write out the probability that a point has not been drawn in n trials.

Hint2: To compute the limit use L'Hôpital's rule
(or compute e.g. with `www.wolframalpha.com`)

b. See the "Coupon collector's problem" on wikipedia. What is the expected number of epochs we need to run to have each data point being drawn at least once?

Assignment 3 (EWA and Momentum)

a. SGD with momentum is defined in pytorch as follows:

$$\begin{aligned} v_{t+1} &= \mu v_t + g_t \\ \theta_{t+1} &= \theta_t - \varepsilon v_{t+1}, \end{aligned} \tag{1}$$

where g_t is the stochastic gradient at point θ_t. Derive this algorithm by applying EWA to stochastic gradient estimates in plain SGD. How does momentum parameter μ is related to q in EWA?

b. Let g_t for $t = 1, \dots n$ be a sequence of stochastic gradients obtained for the same model parameter vector by sampling mini-batches at random but not doing any optimization steps. Consider a variant of the exponentially weighted average:

$$v_t = (1 - q_t)v_{t-1} + q_t g_t,$$

where $v_0 = 0$, $q_t = \frac{q}{1-(1-q)^t}$ and q is a constant.

b.1) Show that for any $t \geq 1$, v_t is an unbiased estimator of the true gradient. *Hint:* start by showing for $t = 1$, $t = 2$.

b.2) Consider the usual EWA with constant q:

$$\hat{v}_t = (1 - q)\hat{v}_{t-1} + q g_t.$$

Show that $\hat{v}_t/(1 - (1 - q)^t)$ is also unbiased. Does it coincide with v_t?

Inspect the Adam optimizer in pytorch and the implementation of momentum there (the relevant momentum parameter is beta1). Which EWA method is used there?

Assignment 4 (CNNs)

a. Show that convolution is equivariant to sub-pixel translations of an image, i.e. translation by a fraction of pixels with interpolation. Assume that a sub-pixel translation is implemented using a bilinear interpolation technique.
Hint: Represent the sub-pixel displacement as cross-correlation.

b. What is the size of the receptive field of one unit in the output of a fully convolutional network with the following layers without padding:
conv(5×5, stride 1, dilation 1)
conv(3×3, stride 1, dilation 2)
conv(3×3 stride 2, dilation 1),

where dilation 1 means standard convolution without holes and dilation 2 is as illustrated in the CNN lecture slide 23.