Introduction to Machine Learning
Problems: Convolutional Neural Networks

Prof. Sundeep Rangan

1. *Tensors.* For each of the following datasets, describe how you would represent them as tensors. Specifically, give the shape of the tensors.

   (a) A batch of 100 color images, each image is 256×256.

   (b) A batch of 40 EEG recordings. Each EEG records has 80 channels of output at a sample rate of 240 Hz for 10 seconds.

   (c) A batch of 32 videos. Each video has a frame rate of 30 frames per second and is 10 seconds long. The video is color with a resolution of 512×512.

2. *2D convolutions.* Let X and W be arrays,

$$X = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 3 & 3 & 0 \\ 0 & 3 & 3 & 3 & 0 \\ 0 & 3 & 2 & 3 & 0 \\ 0 & 3 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad W = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}.$$

   Let Z be the 2D convolution (without reversal):

$$Z[i,j] = \sum_{k_1,k_2} W[k_1,k_2] X[i+k_1, j+k_2]. \tag{1}$$

   Assume that the arrays are indexed starting at $(0,0)$.

   (a) What are the limits of the summations over k_1 and k_2 in (1)?

   (b) What is the size of the output $Z[i,j]$ if the convolution is computed only on the *valid* pixels (i.e. the pixel locations (i,j) where the summation in (1) does not exceed the boundaries of W or X).

   (c) What is the largest positive value of $Z[i,j]$ and state one pixel location (i,j) where that value occurs.

   (d) What is the largest negative value of $Z[i,j]$ and state one pixel location (i,j) where that value occurs.

   (e) Find one pixel location where $Z[i,j] = 0$.

3. *Complexity and number of parameters.* Suppose that a convolutional layer of a neural network has an input tensor $X[i,j,k]$ and computes an output via a convolution and ReLU activation,

$$Z[i,j,m] = \sum_{k_1} \sum_{k_2} \sum_{n} W[k_1,k_2,n,m] X[i+k_1, j+k_2, n] + b[m],$$
$$U[i,j,m] = \max\{0, Z[i,j,m]\}.$$

for some weight kernel $W[k_1,k_2,n,m]$ and bias $b[m]$. Suppose that X has shape (48,64,10) and W has shape (3,3,10,20). Assume the convolution is computed on the *valid* pixels.

(a) What are the shapes of Z and U?

(b) What are the number of input channels and output channels?

(c) How many multiplications must be performed to compute the convolution in that layer?

(d) If W and b are to be learned, what are the total number of trainable parameters in the layer?

4. *Back-propagation.* Suppose that a convolutional layer in some neural network is described as a linear convolution followed by a sigmoid activation,

$$Z[i,j_1,j_2,m] = \sum_{k_1} \sum_{k_2} \sum_{n} W[k_1,k_2,n,m] X[i, j_1+k_1, j_2+k_2, n] + b[m],$$
$$U[i,j_1,j_2,m] = 1/(1+\exp(-Z[i,j_1,j_2,m])).$$

where $X[i,j_1,j_2,n]$ is the input of the layer and $U[i,j_1,j_2,m]$ is the output. Suppose that during back-propagation, we have computed the gradient $\partial J/\partial U$ for some loss function J. That is, we have computed the components $\partial J/\partial U[i,j_1,j_2,m]$. Show how to compute the following:

(a) The gradient components $\partial J/\partial Z[i,j_1,j_2,m]$.

(b) The gradient components $\partial J/\partial W[k_1,k_2,n,m]$.

(c) The gradient components $\partial J/\partial X[i,j_1,j_2,n]$.

5. *Sub-sampling and pooling.* In CNNs, convolution operations are often followed by a data reduction step, typically either via *sub-sampling* or *max pooling*. The methods can be described as follows: Let $x[j]$, $j = 0,1,\ldots,N-1$ be a 1D input (say in one channel in one sample). The outputs $y[k]$ for sub-sampling and max-pooling are given by:

- *Sub-sampling* with *stride* s selects every s-th sample:

$$y[k] = x[sk], \quad k = 0,1,\ldots,\left\lfloor \frac{N-1}{s} \right\rfloor.$$

- *Max pooling* with *pool size* p and *stride* s computes,

$$y[k] = \max_{j=0,1,\ldots,p-1} x[sk+j], \quad k = 0,1,\ldots,\left\lfloor \frac{N-1}{s} \right\rfloor.$$

(a) Let $\mathbf{x}$ be the vector,

$$\mathbf{x} = [1, 2, 3, 2, 0, 10, 1, 0].$$

Find the output $\mathbf{y}$ when sub-sampling with stride $s = 2$.

(b) For the same vector $\mathbf{x}$ as in part (a), find the output of max pooling with stride $s = 2$ and pool size $p = 2$.

(c) Let $X[i, j, n]$ be a tensor of shape (B, N, C) where B is the batch size, N is the number of samples per channel and C is the number of channels. Write equations for sampling and max pooling of X if the operations are to be performed on each channel and sample. What are the output shapes?