DS-GA 1008: Deep Learning, Spring 2020
Homework Assignment 2

```
The search for truth is more precious than its possession.
Albert Einstein (1879 - 1955)
```

1. Fundamentals

1.1. Convolution

Table 1 depicts two matrices. The one on the left represents an 5×5 single-channel image $\boldsymbol{A}$. The one on the right represents a 3×3 convolution kernel $\boldsymbol{B}$.

(a) What is the dimensionality of the output if we forward propagate the image over the given convolution kernel with no padding and stride of 1?

Answer: 3×3

(b) Give a general formula of the output width O in terms of the input width I, kernel width K, stride S, and padding P (both in the beginning and in the end). Note that the same formula holds for the height. Make sure that your answer in part (a) is consistent with your formula.

Answer:
No padding, no stride:

$$O = I - (K - 1) \tag{1}$$

Add stride:

$$O = \left\lceil \frac{I - (K - 1)}{S} \right\rceil \tag{2}$$

Finally, add padding (both sides):

$$O = \left\lceil \frac{(I + 2P) - (K - 1)}{S} \right\rceil \tag{3}$$

(c) Compute the output $\boldsymbol{C}$ of forward propagating the image over the given convolution kernel. Assume that the bias term of the convolution is zero.

Answer:

$\boldsymbol{C} =$

109	92	72
108	85	74
110	74	79

(d) Suppose the gradient backpropagated from the layers above this layer is a 3×3 matrix of all 1s. Write the value of the gradient with respect to the input image backpropagated out of this layer. That is, you are given that $\frac{\partial E}{\partial \boldsymbol{C}_{ij}} = 1$ for some scalar error E and $i, j \in \{1, 2, 3\}$. You need to compute $\frac{\partial E}{\partial \boldsymbol{A}_{ij}}$ for $i, j \in \{1, \ldots, 5\}$. The chain rule should help!

Answer:
Formula for $C[n, m]$:

$$C[n,m] = \sum_{\substack{k=1\ldots3\\p=1\ldots3}} B[k,p] * A[n-1+k, m-1+p] \tag{4}$$

Using chain rule, we obtain expression for $\frac{\partial E}{\partial A[i,j]}$:

$$\frac{\partial E}{\partial A[i,j]} = \sum_{\substack{1\leq n\leq 3\\1\leq m\leq 3\\i-1\leq n+1\leq i+1\\j-1\leq m+1\leq j+1}} \frac{\partial E}{\partial C[n,m]}\frac{\partial C[n,m]}{\partial A[i,j]} \tag{5}$$

Using formula (4) we can find $\frac{\partial C[n,m]}{\partial A[i,j]}$:
Let $n - 1 + k = i$ and $m - 1 + p = j$:

$$\frac{\partial C[n,m]}{\partial A[i,j]} = B[i+1-n, j+1-m] \tag{6}$$

Hence $\frac{\partial E}{\partial C[n,m]} = 1$

$$\frac{\partial E}{\partial A[i,j]} = \sum_{\substack{1\leq n\leq 3\\1\leq m\leq 3\\i-1\leq n+1\leq i+1\\j-1\leq m+1\leq j+1}} B[i+1-n, j+1-m] \tag{7}$$

$\boldsymbol{A} =$

4	5	2	2	1
3	3	2	2	4
4	3	4	1	1
5	1	4	1	2
5	1	3	1	4

$\boldsymbol{B} =$

4	3	3
5	5	5
2	4	3

Table 1: Image Matrix (5×5) and a convolution kernel (3×3).

DS-GA 1008: Deep Learning, Spring 2020
Homework Assignment 2

1.2. Pooling

Pooling is a technique for sub-sampling and comes in different flavors, for example max-pooling, average pooling, LP-pooling.

(a) List the `torch.nn` modules for the 2D versions of these pooling techniques and read what they do.

Answer:

1. max-pooling — `nn.MaxPool2d`
2. average pooling — `nn.AvgPool2d`
3. LP-pooling — `nn.LPPool2d`

(b) Denote the k-th input feature maps to a pooling module as $\boldsymbol{X}^k \in \mathbb{R}^{H_{\text{in}} \times W_{\text{in}}}$ where H_{in} and W_{in} represent the input height and width, respectively. Let $\boldsymbol{Y}^k \in \mathbb{R}^{H_{\text{out}} \times W_{\text{out}}}$ denote the k-th output feature map of the module where H_{out} and W_{out} represent the output height and width, respectively. Let $S^k_{i,j}$ be a list of the indexes of elements in the sub-region of X^k used for generating $\boldsymbol{Y}^k_{i,j}$, the (i,j)-th entry of $\boldsymbol{Y}^k$. Using this notation, give formulas for $\boldsymbol{Y}^k_{i,j}$ from three pooling modules.

Answer:

1. max-pooling:

$$\boldsymbol{Y}^k_{i,j} = \max_{n \in S^k_{i,j}} \boldsymbol{X}^k_n \tag{8}$$

2. average pooling:

$$\boldsymbol{Y}^k_{i,j} = \frac{1}{|S^k_{i,j}|} \sum_{n \in S^k_{i,j}} \boldsymbol{X}^k_n \tag{9}$$

3. LP-pooling

$$\boldsymbol{Y}^k_{i,j} = \sqrt[p]{\sum_{n \in S^k_{i,j}} |\boldsymbol{X}^k_n|^p} \tag{10}$$

(c) Write out the result of applying a max-pooling module with kernel size of 2 and stride of 1 to $\boldsymbol{C}$ from Part 1.1.

Answer:

$$\boldsymbol{M} = \begin{array}{|c|c|} \hline 109 & 92 \\ \hline 110 & 85 \\ \hline \end{array}$$

(d) Show how max-pooling and average pooling can be expressed in terms of LP-pooling.

Answer:

1. max-pooling — by definition LP-pooling with $p = \infty$
2. average pooling — proportional to LP-pooling with $p = 1$