DS-GA 1008: Deep Learning, Spring 2019
Homework Assignment 2

Chengze Zuo | cz1565@nyu.edu

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

Part I

1. Fundamentals

1.1. Convolution

Table 1 depicts two matrices. The one on the left represents a 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: Suppose the dimension of the input image is F, the kernel size is K, the size of zero-padding is P, and the stride size is S. The corresponding output size D after the convolution would be (suppose that width and height are both for the input image and the kernel):

$$D = (F-K+2P)/S+1$$

D = (5-3+0)/1+1 = 3. Therefore, the dimensionality of the output would be 3 x 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: As discussed in the part(a), the general formula would be:

$$O = (I-K+2P)/S+1 \text{ (same formula holds for height)}$$

(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: The output of the first neuron in the activation map would be element-wise multiplication between the sub-regeion matrix $\begin{array}{|c|c|c|}\hline 4 & 5 & 2 \\ \hline 3 & 3 & 2 \\ \hline 4 & 3 & 4 \\ \hline \end{array}$ and the kernel $\begin{array}{|c|c|c|}\hline 4 & 3 & 3 \\ \hline 5 & 5 & 5 \\ \hline 2 & 4 & 3 \\ \hline \end{array}$, leading to 109. After we slide the kernel through the entire input image , the output $\boldsymbol{C}$ would be

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 (w.r.t. the input image) backpropagated out of this layer.

$$\boldsymbol{A} = \begin{array}{|c|c|c|c|c|}\hline 4 & 5 & 2 & 2 & 1 \\ \hline 3 & 3 & 2 & 2 & 4 \\ \hline 4 & 3 & 4 & 1 & 1 \\ \hline 5 & 1 & 4 & 1 & 2 \\ \hline 5 & 1 & 3 & 1 & 4 \\ \hline \end{array} \qquad \boldsymbol{B} = \begin{array}{|c|c|c|}\hline 4 & 3 & 3 \\ \hline 5 & 5 & 5 \\ \hline 2 & 4 & 3 \\ \hline \end{array}$$

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

Hint: 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 $i, j \in \{1, \dots, 5\}$. The chain rule should help!

Answer: According to the chain rule, $\frac{\partial E}{\partial \boldsymbol{A}_{ij}} = \frac{\partial E}{\partial \boldsymbol{C}} \cdot \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{A}_{ij}}$, and for now, let us focus on solving $\frac{\partial E}{\partial \boldsymbol{A}_{11}}$. We can expand in it as:

$$\frac{\partial E}{\partial \boldsymbol{A}_{11}} = \frac{\partial E}{\partial \boldsymbol{C_{11}}} \cdot \frac{\partial \boldsymbol{C_{11}}}{\partial \boldsymbol{A}_{11}} + \frac{\partial E}{\partial \boldsymbol{C_{12}}} \cdot \frac{\partial \boldsymbol{C_{12}}}{\partial \boldsymbol{A}_{11}} + \frac{\partial E}{\partial \boldsymbol{C_{13}}} \cdot \frac{\partial \boldsymbol{C_{13}}}{\partial \boldsymbol{A}_{11}} + \frac{\partial E}{\partial \boldsymbol{C_{21}}} \cdot \frac{\partial \boldsymbol{C_{21}}}{\partial \boldsymbol{A}_{11}} + \frac{\partial E}{\partial \boldsymbol{C_{22}}} \cdot \frac{\partial \boldsymbol{C_{22}}}{\partial \boldsymbol{A}_{11}} + \frac{\partial E}{\partial \boldsymbol{C_{23}}} \cdot \frac{\partial \boldsymbol{C_{23}}}{\partial \boldsymbol{A}_{11}} + \frac{\partial E}{\partial \boldsymbol{C_{31}}} \cdot \frac{\partial \boldsymbol{C_{31}}}{\partial \boldsymbol{A}_{11}} + \frac{\partial E}{\partial \boldsymbol{C_{32}}} \cdot \frac{\partial \boldsymbol{C_{32}}}{\partial \boldsymbol{A}_{11}} + \frac{\partial E}{\partial \boldsymbol{C_{33}}} \cdot \frac{\partial \boldsymbol{C_{33}}}{\partial \boldsymbol{A}_{11}}$$

$$= \frac{\partial E}{\partial \boldsymbol{C_{11}}} \cdot \boldsymbol{B_{11}} + \frac{\partial E}{\partial \boldsymbol{C_{12}}} \cdot 0 + \frac{\partial E}{\partial \boldsymbol{C_{13}}} \cdot 0 + \frac{\partial E}{\partial \boldsymbol{C_{21}}} \cdot 0 + \frac{\partial E}{\partial \boldsymbol{C_{22}}} \cdot 0 + \frac{\partial E}{\partial \boldsymbol{C_{23}}} \cdot 0 + \frac{\partial E}{\partial \boldsymbol{C_{31}}} \cdot 0 + \frac{\partial E}{\partial \boldsymbol{C_{32}}} \cdot 0 + \frac{\partial E}{\partial \boldsymbol{C_{33}}} \cdot 0$$

$$= 1 \cdot 4 = 4$$

Based on this strategy, we are able to infer that:

$$\frac{\partial E}{\partial \boldsymbol{A_{12}}} = 3 + 4 = 7$$

$$\frac{\partial E}{\partial \boldsymbol{A_{13}}} = 3 + 3 + 4 = 10$$

$$\frac{\partial E}{\partial \boldsymbol{A_{14}}} = 3 + 3 = 6$$

$$\frac{\partial E}{\partial \boldsymbol{A_{15}}} = 3$$

$$\frac{\partial E}{\partial \boldsymbol{A_{21}}} = 5 + 4 = 9$$

$$\frac{\partial E}{\partial \boldsymbol{A_{22}}} = 5+5+3+4 = 17$$
$$\frac{\partial E}{\partial \boldsymbol{A_{23}}} = 5+5+5+3+3+4 = 25$$
$$\frac{\partial E}{\partial \boldsymbol{A_{24}}} = 5+5+3+3 = 16$$
$$\frac{\partial E}{\partial \boldsymbol{A_{25}}} = 5+3 = 8$$
$$\frac{\partial E}{\partial \boldsymbol{A_{31}}} = 4+5+2 = 11$$
$$\frac{\partial E}{\partial \boldsymbol{A_{32}}} = 4+5+2+3+5+4 = 23$$
$$\frac{\partial E}{\partial \boldsymbol{A_{33}}} = \sum \boldsymbol{B_{ij}} = 34$$
$$\frac{\partial E}{\partial \boldsymbol{A_{34}}} = 3+5+3+3+5+4 = 23$$
$$\frac{\partial E}{\partial \boldsymbol{A_{35}}} = 3+5+3 = 11$$
$$\frac{\partial E}{\partial \boldsymbol{A_{41}}} = 2+5 = 7$$
$$\frac{\partial E}{\partial \boldsymbol{A_{42}}} = 5+5+2+4 = 16$$
$$\frac{\partial E}{\partial \boldsymbol{A_{43}}} = 5+5+5+3+4+2 = 24$$
$$\frac{\partial E}{\partial \boldsymbol{A_{44}}} = 4+3+5+5 = 17$$
$$\frac{\partial E}{\partial \boldsymbol{A_{45}}} = 3+5 = 8$$
$$\frac{\partial E}{\partial \boldsymbol{A_{51}}} = 2$$
$$\frac{\partial E}{\partial \boldsymbol{A_{52}}} = 2+4 = 6$$
$$\frac{\partial E}{\partial \boldsymbol{A_{53}}} = 2+4+3 = 9$$
$$\frac{\partial E}{\partial \boldsymbol{A_{54}}} = 3+4 = 7$$
$$\frac{\partial E}{\partial \boldsymbol{A_{55}}} = 3$$

Therefore, the result of the gradient w.r.t the input image would be:

$$\frac{\partial \boldsymbol{E}}{\partial \boldsymbol{A}} = \begin{array}{|c|c|c|c|c|} \hline 4 & 7 & 10 & 6 & 3 \\ \hline 9 & 17 & 25 & 16 & 8 \\ \hline 11 & 23 & 34 & 23 & 11 \\ \hline 7 & 16 & 24 & 17 & 8 \\ \hline 2 & 6 & 9 & 7 & 3 \\ \hline \end{array}$$

.

1.2. Pooling

The 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 on what they do.

Answer: MaxPool2d, MaxUnpool2d(partial inverse of the MaxPool2d, can specify a different output size), AvgPool2d, FractionalMaxPool2d, LPPool2d, AdaptiveMaxPool2d, AdaptiveAvgPool2d

(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: First, denote that the kernel size as kH, kW, each of which represents the length along height and width. Then denote that the zero-padding as P and that the stride as sH and sW. Then we could derive formulas for $\boldsymbol{Y}^k_{i,j}$ for MaxPool2d, AvgPool2d and LPPool2d modules respectively.

For MaxPool2d module: $\boldsymbol{Y}^k_{i,j} = \max\limits_{m=0,1\ldots,kH-1} \max\limits_{n=0,1\ldots,kW-1} \boldsymbol{X}^k[\boldsymbol{S}^k_{i\times sH+m,j\times sW+n}]$
For AvgPool2d module: $\boldsymbol{Y}^k_{i,j} = \frac{1}{kH\times kW} \sum_{m=0}^{kH-1} \sum_{n=0}^{kW-1} \boldsymbol{X}^k[\boldsymbol{S}^k_{i\times sH+m,j\times sW+n}]$
For LPPool2d module: Su $\boldsymbol{Y}^k_{i,j} = \sqrt[p]{\sum\limits_{x\in\boldsymbol{X}^k[\boldsymbol{S}^k_{i\times sH+m,j\times sW+n}]} x^p}$ (one would get Sum pooling if p = 1, or woule get Max pooling if p = ∞)

Note that the zero-padding P is actually included in the input. If zero padding P is non-zero, then the input is implicitly zero-padded on both sides for P number of points.

(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: After applyling a max-pooling with kernel size of 2 and stride of 1 to $\boldsymbol{C}$, the result would be:

109	92
110	85

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

Answer: For the LP-pooling, if p goes to ∞, the formula $\boldsymbol{Y}^k_{i,j} = \sqrt[p]{\sum_{x\in \boldsymbol{X}^k[\boldsymbol{S}^k_{i\times sH+m,j\times sW+n}]} x^p}$ would be the max element in sub-region matrix $\boldsymbol{X}^k[\boldsymbol{S}^k_{i\times sH+m,j\times sW+n}]$. From the intuitive perspective, as p goes the ∞, the largest element would gain more weight/influence in the summation, and the difference between this largest element and other smaller elements would become larger, therefore, the result of the overall equation would approximate to that largest element. From the mathematical perspective, the result of LP-pooling as p going to ∞ would equal to computing the limit of this exponential series, where $\lim_{p\to\infty}(a_1^p + a_2^p + \ldots + a_m^p)^{1/p} = \max(a_1, a_2, \ldots, a_m)$, which is indeed the max-pooling.

If p = 1, the above formula is actually the sum of all the elements in the $\boldsymbol{X}^k[\boldsymbol{S}^k_{i\times sH+m,j\times sW+n}]$, which is proportional to the average pooling. Therefore we could express the average pooling as $\frac{1}{kH\times kW}$ times the result of LP-pooling where p equals to 1 (which one might also call as "Sum Pooling").