Mathematical Foundation of DNN : HW 1

Jeong Min Lee

March 27, 2024

Notations

1. $\otimes$: convolution

1

Let $\omega^{(1)}, \omega^{(2)} \in \mathbb{R}^{3\times 3}$ denote the map from X to Y_1, Y_2, respectively. Then, $\omega \in \mathbb{R}^{(2\times 3\times 3)} = [\omega^{(1)}, \omega^{(2)}]$. Noting that the process of discrete convolution and being aware of where the filter maps each image pixel X_{ij}, $\omega^{(1)}, \omega^{(2)}$ can be represented as follow.

$$\omega^{(1)} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{1}$$

$$\omega^{(2)} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \tag{2}$$

For instance,

$$\begin{pmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{pmatrix} \otimes \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix} = X_{32} - X_{22} = Y_{22}$$

2

We can formalized this problem as follow.

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \otimes \begin{pmatrix} & & \\ & ? & \\ & & \end{pmatrix} = \frac{1}{9}(a+b+c+d+e+f+g+h+i)$$

This is obvious that I_k/k^2 satisfy the condition above, where I_k is k dimensional identity matrix. Note that we have to divide all summation of subimage since there are k^2 pixels that overlap the filter whose size is k.

3

This problem is obvious that $\omega = [0.299, 0.587, 0.114] \in \mathbb{R}^{3\times 1\times 1}$. As the given expression $Y_{ij} = 0.299X_{1,i,j} + 0.587X_{2,i,j} + 0.114X_{3,i,j}$ says, 1×1 convolution is a type of constant multiplicatioon. The reason we use 1×1 convolution can be elucidated by seeing the following example: Consider the 192 gray scale images $X \in \mathbb{R}^{192\times 28\times 28}$. Doing appropriate padding to conserve the output dimension, using a filter whose size is 5 and channel is 32 returns $32 \times 28 \times 28$ activation maps. The number of float multiplication calculation in this case is $28 \times 28 \times 32 \times 5 \times 192 = 120\text{M}$. However, inserting 1d convolution whose depth is 16 results in the activation map whose dimension is $16 \times 28 \times 28$.(Again, I assumed appropriate padding to preserve the dimension.) Then, applying $32 \times 5 \times 5$ makes the output activation map have $32 \times 28 \times 28$ dimension. (Note that this secondary filter is identical to the filter in the first case.) For second case, the number of computations is $28 \times 28 \times 1 \times 1 \times 192 + 28 \times 28 \times 32 \times 5 \times 5 \times 16 = 12.4\text{M}$, which is 10 times smaller than the first case. This example implies that inserting 1d convolution between the filter convolving the image reduces the time complexity of the model and can improve the model's performance.

4

CLAIM : σ can commute to max
proof : Consider $S = \{a_1, a_2, \cdots, a_n\} \in \mathbb{R}^n$. Withoug loss of generality, suppose $\max S = a_1$. Then, $\sigma(\max S) = \sigma(a_1)$. However, noting that σ is non-decreasing function, $\max\{\sigma(a_1), \cdots, \sigma(a_n)\} = \sigma(a_1)$. Otherwise, there is some a_i s.t. $\sigma(a_i) > \sigma(a_1)$. However, since $a_i \le a_1$, it contradicts to the assumption of σ. Thus, σ and max commute each other.■

Let $\mathfrak{M}_{m,n}(F)$ be the set of $m \times n$ matrices over some field F. Then, in alegebra, it is well known fact that $\mathfrak{M}_{m,n}(F)$ is isomorphic to F^{mn}. Letting $F = \mathbb{R}$, one can apply the **CLAIM** to the $X \in \mathfrak{M}_{m,n}(\mathbb{R})$. [1] Since ρ maps from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{k\times l}$, the filter size that ρ uses is $(m/k, n/l)$. Considering the feature of max pooling, $\rho = \max$ for the input whose size is identical to the filter size. Let

$$X_{i:i+m/k-1,j:j+n/l-1} = \begin{pmatrix} X_{ij} & \cdots & X_{i,j+n/l-1} \\ \vdots & \ddots & \vdots \\ X_{i+m/k-1,j} & \cdots & X_{i+m/k-1,j+n/l-1} \end{pmatrix}$$

and $y_{ij} = \max X_{i:i+m/k-1,j:j+n/l-1}$. Then, $\rho(X) = (y_{ij})$.

$$\begin{aligned} \sigma(\rho(X_{i:i+m/k-1,j:j+n/l-1})) &= \sigma\left(\max\begin{pmatrix} X_{ij} & \cdots & X_{i,j+n/l-1} \\ \vdots & \ddots & \vdots \\ X_{i+m/k-1,j} & \cdots & X_{i+m/k-1,j+n/l-1} \end{pmatrix}\right) \\ &= \max\left(\begin{pmatrix} \sigma(X_{ij}) & \cdots & \sigma(X_{i,j+n/l-1}) \\ \vdots & \ddots & \vdots \\ \sigma(X_{i+m/k-1,j}) & \cdots & \sigma(X_{i+m/k-1,j+n/l-1}) \end{pmatrix}\right) \\ &= \rho(\sigma(X_{i:i+m/k-1,j:j+n/l-1})) \\ &= \sigma(y_{ij}) \end{aligned}$$

This implies that σ and ρ can commute.

5

[1] This is corollary of the following theorem : $V \approx W \iff \dim V = \dim W$ for two vector spaces V, W.