4. In an image application, the input feature maps of spatial size $P \times Q$ with C channels and the output feature maps of spatial size $P \times Q$ with D channels of a layer of neural network are expressed as $x_{i,j,k}, 1 \leq i \leq P, 1 \leq j \leq Q, 1 \leq k \leq C$, and $y_{i,j,k}, 1 \leq i \leq P, 1 \leq j \leq Q, 1 \leq k \leq D$, respectively.

(a) Express the output feature maps $y_{i,j,k}$ in terms of the input feature maps $x_{i,j,k}$ via the scalar network parameters of a layer of fully connected neural network. What is the number of learnable parameters?

(5 Marks)

(b) Express the output feature maps in terms of the input feature maps via the scalar network parameters of a layer of spatial convolutional neural network of filter size 3×3. What is the number of learnable parameters?

(5 Marks)

(c) Express the output feature maps in terms of the input feature maps via the scalar network parameters of a layer of spatial convolutional neural network of filter size 1×1. What is the number of learnable parameters?

(5 Marks)

(d) If we use a single index for the 2D spatial position to express the input and output feature maps by $x_{i,k}, 1 \leq i \leq PQ, 1 \leq k \leq C$, and $y_{i,k}, 1 \leq i \leq PQ, 1 \leq k \leq D$, respectively, re-express the answer to part (c) without using the bias. What is the number of learnable parameters?

(5 Marks)

(e) Arrange all $x_{i,k}$ in part (d) into a $PQ \times C$ matrix, $\boldsymbol{X}$, and arrange all $y_{i,k}$ in part (d) into a $PQ \times D$ matrix, $\boldsymbol{Y}$. Re-express the answer to part (d) in the matrix format $\boldsymbol{X}$ and $\boldsymbol{Y}$. (Note that $PQ \times C$ matrix has PQ rows and C columns.) What can you conclude about the relation between a convolutional neural network and a Transformer?

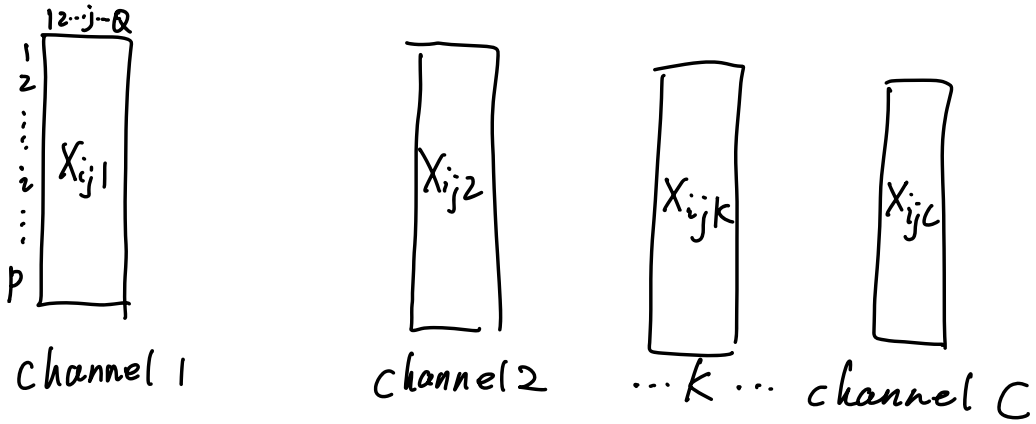
(5 Marks)

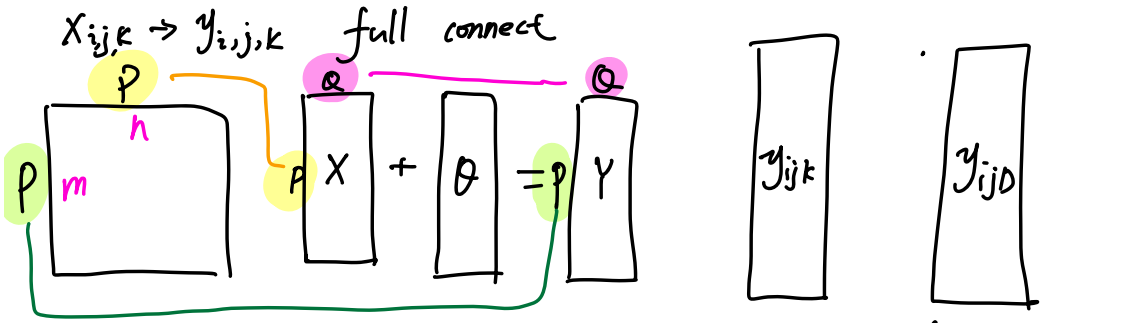
[**Hint for Question 4:** If inputs and outputs of a layer of network are expressed as $x_i, 1 \leq i \leq P$ and $y_i, 1 \leq i \leq Q$, respectively, the outputs of a layer of fully connected neural network can be expressed in terms of the inputs as $y_i = \sum_{l=1}^{P} w_{l,i} x_l + b_i, 1 \leq i \leq Q$, where $w_{l,i}, b_i$ are the scalar parameters called the weights and biases of the network, respectively.]

23-S2-Q4

(a) Q: $x_{ijk} \to y_{ijk}$? parameters?

Solution ① understand





$$\sum \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2n} & \cdots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} & \cdots & w_{mp} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & & w_{pn} & \cdots & w_{pp} \end{bmatrix} \begin{bmatrix} x_{11} & \cdots & x_{1Q} \\ x_{21} & \cdots & x_{2Q} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{iQ} \\ \vdots & \ddots & \vdots \\ x_{P1} & & x_{PQ} \end{bmatrix} + \begin{bmatrix} b_{11} & \cdots & b_{1Q} \\ b_{21} & \cdots & b_{2Q} \\ \vdots & \ddots & \vdots \\ b_{i1} & \cdots & b_{iQ} \\ \vdots & \ddots & \vdots \\ b_{P1} & \cdots & b_{PQ} \end{bmatrix}$$

$$y_{ijk} = \sum_{l=1}^{C} \sum_{m=1}^{P} \sum_{n=1}^{P} W_{mnlk}\, x_{ijl} + b_{ijk} \quad (1 \le k \le D)$$

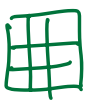
parameters $P \times P \times C \times D + P \times Q \times D$

一张W　　1张bias

C张W生成1张Y　　D张bias

生成D张Y

(b) CNN ?



$$y_{i,j,k} = \sum_{u=-1}^{1} \sum_{v=-1}^{1} \sum_{l=1}^{C} w_{u,v,l,k}\, x_{i-u,j-v,l} + b_k$$

$$3 \times 3 \times C \times D + D = 9 \times C \times D + D$$

(c) $$y_{i,j,k} = \sum_{l=1}^{C} w_{l,k}\, x_{i,j,l} + b_k$$

$$C \times D + D$$

RGB　D个特征图　bias

(d) $$y_{i,k} = \sum_{l=1}^{C} w_{l,k}\, x_{i,l}$$

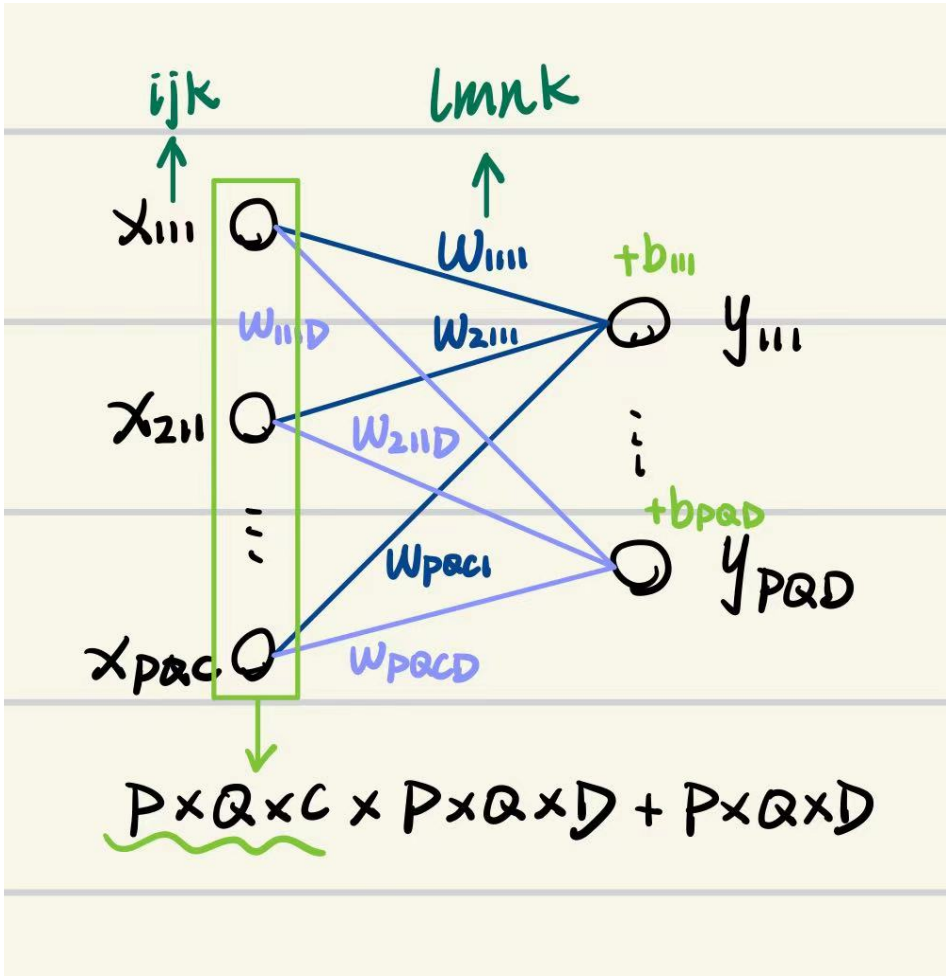
$$C \times D$$

(e) $Y = XW$

① Both architectures apply matrix multiplications to process input

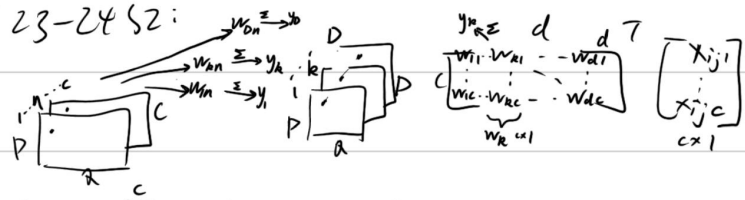
② $Y = XW$ shows the CNN perform a linear transformation on the input features at each spatial position, identical across all position.

It is similar to the linear layers used in Transformer, where input are transformed via weight matrices



7403 23-24 S2:

4):

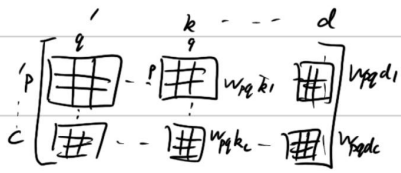


a): $y_{ijk} = \sum_{n=1}^{c} w_{kn} \cdot x_{ijn}$, W shape $c \times D$ bias: $CD+D$

$c \times D$ trainable params

b): $y_{ijk} = \sum_{n=1}^{c} \sum_{p=1}^{l} \sum_{q=1}^{l} w_{pqkn} x_{i-p,j-q,n}$ bias: $9CD + D$

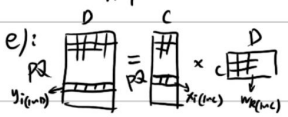
$\Rightarrow 3\times3\times c\times D = 9CD$ trainable params



c): $y_{ijk} = \sum_{n=1}^{c} w_{kn} \cdot x_{ijn}$ $\Rightarrow CD$ trainable params. bias: $CD+D$

d): $y_{ik} = \sum_{n=1}^{c} w_{kn} \cdot x_{in}$ $\Rightarrow CD$ trainable params. bias: $CD+D$

e):



$Y = XW$

$X = [\vec{x}_1^{\,c} \; \cdots \; \vec{x}_{PQ}^{\,c}]^T$ pixels of all channels

tokens ~ pixels

position ~ channel

Transformer: $X = [\vec{x}_1^{\,P} \; \cdots \; \vec{x}_N^{\,P}]^T$ tokens embeded with position

$Q = XW_Q \quad K = XW_K \quad V = XW_V$

Ani

主要区别在输出的时候是逐位置算一组权重还是逐channel算一组权重

Ani

逐位置是PQCPQD+PQD，逐chanel是PQCD+D

JIL WANG

应该说最多就是PQDPQC+PQD个参数

JIL WANG

比这个少的都是简化或者共享了参数