EE6427

Q (i) output after the convolution layer?

$$A = \begin{bmatrix} 4 & 0 & 1 \\ 4 & 0 & 2 \\ 0 & 2 & 2 \end{bmatrix} \qquad F = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$ zero padding 1

stride 2.

sigmod $\sigma(x) = \dfrac{1}{1+e^{-x}}$

2x2 max pooling stride 2.

Solution ① zero padding

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 1 & 0 \\ 0 & 4 & 0 & 2 & 0 \\ 0 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

② convolution

$$F = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 4 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} = 4 \times 0 + 4 \times 0 = 0$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & \enclose{circle}{1} & 0 \\ 0 & \enclose{circle}{2} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 2 & \enclose{circle}{0} & -2 \\ 1 & \enclose{circle}{0} & -1 \end{bmatrix} = 1\times 0 + 2\times 0 = 0$$

$$\begin{bmatrix} 0 & \enclose{circle}{4} & 0 \\ 0 & 0 & \enclose{circle}{2} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & \enclose{circle}{0} & -1 \\ 2 & 0 & \enclose{circle}{-2} \\ 1 & 0 & -1 \end{bmatrix} = 4\times 0 + 2\times(-2) = -4$$

$$\begin{bmatrix} 0 & 2 & 0 \\ \enclose{circle}{2} & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ \enclose{circle}{2} & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} = 2\times 2 = 4. \quad \text{output} = \begin{bmatrix} 0 & 0 \\ -4 & 4 \end{bmatrix}$$

(ii)Q effect of F?

Solution

The effect of filter F is that it computes the horizontal gradient information which reflects the edge information

水平方向的数量变化率

(iii)Q: output after activation

Solution

$$e(x) = \frac{1}{1+e^{-x}}$$

$$\begin{bmatrix} 0 & 0 \\ -4 & 4 \end{bmatrix} \frac{1}{1+e^{-x}} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{1+e^{4}} & \frac{1}{1+e^{-4}} \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 \\ 0.018 & 0.982 \end{bmatrix}$$

(iv) Q output after max pooling layer

Solution Max = 0.9820

(V) Q: parameterz ?

Solution 100x100x3 → 6 channel

filter 3x3x3 parameter

number of filte: 6

total $3 \times 3 \times 3 \times 6 = 162$

(b) Answer : $\boxed{\text{Transformer}}$

Justification

classifying vide clips requires modeling temporal dependencies and relation across frames.

① Transformer are designed to handle sequential data and excel at capturing long-range dependencies through self-attention mechanisms. They can effectively model the relationships between all pairs of frames in the video, enabling a comprehensive understanding of the temporal dynamics essential for genre classification

② Compared to Vanilla RNNs, which process sequences sequentially and may struggle with long-term dependencies due to vanishing gradients. However, transformers process all position in sequence simultaneously and can better capture global context

③ CNNs are less suited for modeling temporal sequences.