# CS2109S Tutorial 8

AY 25/26 Sem 1 — github/omgeta

A. 1. $Y = \begin{bmatrix} -0.2 & 0 & 0.2 & 1.1 \\ -0.3 & 0.1 & 0.1 & 1.2 \\ -0.2 & 0.1 & 0.1 & 0.3 \\ -1.2 & 0 & 0.5 & 1 \end{bmatrix}$. This should be to detect the strength of right edges.

2. Max-Pool $= \begin{bmatrix} 0.1 & 1.2 \\ 0.1 & 1 \end{bmatrix}$ to keep strongest local feature

   Average-Pool $= \begin{bmatrix} -0.1 & 0.65 \\ -0.325 & 0.475 \end{bmatrix}$ to capture overall feature

3. Size $= \lfloor (224 - 11)/4 \rfloor + 1 = 54$

4. $(B, 96, 54, 54)$. This provides higher throughput and steadier gradients.

B. 1. First Layer: $5 \times 5$, Second Layer; $7 \times 7$

2. Larger receptive fields can improve broadness of context but lose local specific data.

C. 1. Many-to-one

2. $h^{[1]} = (W^{[xh]})^\top x_1 + (W^{[hh]})^\top h^{[0]} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

   $h^{[2]} = (W^{[xh]})^\top x_2 + (W^{[hh]})^\top h^{[1]} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

3. $\hat{y}^{[2]} = softmax((W^{[hy]})^\top h^{[2]}) = \begin{bmatrix} 0.0418 \\ 0.1135 \\ 0.8390 \\ 0.0057 \end{bmatrix}$ so output is "coding"

4. $h^{[t]}$ depends on $h^{[t-1]}, x^{[t]}$ so order can change the trajectory of hidden values and hence the final output.

5. Vector size can get large for large vocabularies, and there is no relation between words (e.g. love and like)

D. 1. Yes, but it would be less sensitive to global order

2. Preprocess with CNN to get a feature map, which is then read row-by-row as a sequence.