

## CS2109S Tutorial 9

AY 25/26 Sem 1 — github/omgeta

- A.
- Many-to-one; we only need binary classification for a single sequence of text
  - Many-to-many; we need to read in a sequence of data and output a sequence of data
  - One-to-many; take in non-sequential data and output sequence of data by time
- B.
- $$Q = W^q X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

$$K = W^k X = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$V = W^v X = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \end{bmatrix}$$
  - $$A = \frac{K^\top Q}{\sqrt{d_k}} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \\ 2 & 1 & 3 \end{bmatrix}$$
 and  $\alpha'_{cat} = \text{softmax}(\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}) = \begin{bmatrix} 0.248 \\ 0.248 \\ 0.503 \end{bmatrix}$  so "dog" receives highest attention when the query comes from "cat"
  - $$h_{cat} = V\alpha'_{cat} \approx \begin{bmatrix} 1.502 \\ 0.751 \end{bmatrix}$$
  - Each projection matrix has  $512 \times 64 = 32768$  weights, so in total we have  $3 \times 32768 = 98304$  weights
- C.
- Masked attention layer is needed by the decoder to look at the words already generated, to use in the context of the next word.
  - Query must come from the decoder which is asking for the next word given the currently generated French words. Key and Value must come from encoder which holds the context and information for the words to translate.
  - It cannot reference source English text, so text generated will be correct grammatically but unrelated to the original English text.
- D.
- Unique: Yes, each position has distinct encoding  
 Consistent: Yes, encoding for a position is always the same regardless of sequence length  
 Bounded: No, scales linearly with position and larger positions will have PE dominating.
  - Unique: Yes, each position has distinct encoding  
 Consistent: No, values dependent on  $T$   
 Bounded: Yes, bounded in  $(0, 1]$
  - Unique: Yes, each position has distinct encoding due to varying cosine frequencies  
 Consistent: Yes, only dependent on  $t$  and  $T_{max}$   
 Bounded: Yes, bounded in  $(\cos(1), 1)$