CS 371 N Lecture 11
Transformers, Transformer Language Modeling

Announcements

- A3 out

Recap Attention: places a probability distribution over a sequence of n tokens with embeddings $e_1 \ldots e_n$

Simplified version:

① Form keys $k_i = W^K e_i$

$$K = E(W^K)^T$$

query q $\qquad\qquad\qquad s \ [0\ 0\ 1\ 0]$

$$\downarrow$$

② Compute scores $s_i = k_i^T q \quad \alpha \ [\tfrac{1}{6}\ \tfrac{1}{6}\ \tfrac{1}{2}\ \tfrac{1}{6}]$

③ Compute attn weights $\alpha = \text{softmax}(s)$

④ Result (output) $= \sum \alpha_i e_i$

Self-attention

E now gives rise to q_i and k_i for each word

E: seq len $\times d$

W^K: $d \times d$ matrix $\quad K = E(W^K)^T$

W^Q: $d \times d$ matrix $\quad Q = E(W^Q)^T$

K, Q: seq len $\times d$

$S = QK^T \qquad S_{ij} = q_i \cdot k_j$

Suppose $E = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad W^Q = W^K = I$

$$S = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

matrix of similarities

$$A = \text{softmax}(S) = \begin{bmatrix} 3/10 & 3/10 & 1/10 & 3/10 \\ - & - & - & - \\ 1/6 & 1/6 & 1/2 & 1/6 \\ - & - & - & - \end{bmatrix}$$

W^v produces "values"

Last step:

$$\text{Output} = A\left(\underbrace{E(W^v)^T}\right)$$

$$(\text{seqlen} \times \text{seqlen}) \cdot (\text{seq len} \times d) \cdot (d \times d)$$

$$\text{Out} = \text{seq len} \times d$$

A takes a weighted sum of values according to attention weights at each position

First row of output $= \frac{3}{10} \cdot v_1 + \frac{3}{10} \cdot v_2 + \frac{1}{10} \cdot v_3 + \frac{3}{10} \cdot v_4$

Third row $= 1/6 \cdot v_1 + 1/6\, v_2 + \frac{1}{2} v_3 + \frac{1}{6} v_4$