CS 371N Lecture 11

Transformers, Transformer Language Modeling

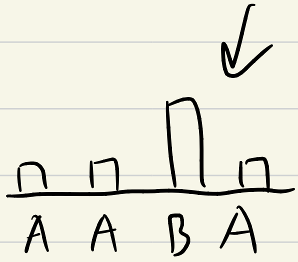
Announcements

- A3 due in 9 days

<u>Recap</u> Attention over a sequence of n tokens with embeddings $e_1 \dots e_n$

① Form keys $= W^K e_i$

query $= q$



② Scores $s_i = k_i^T q$

③ Attention weights (probs) $\alpha = \text{softmax}(s)$

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

$\wedge$ we'll throw a matrix here later

Today

- Self-attention recap
- Exercises
- Multi-head self-attention
- Transformers
- Language modeling

Self-attention

Idea: all words are now keys and queries simultaneously

$$\left.\begin{array}{ll} E: \text{seq len} \times d \text{ matrix} & \\ W^K: d \times d & K = E(W^K)^T \end{array}\right\} \text{same as before}$$

$$Q: \text{seq len} \times d \qquad (Q = E(W^Q)^T)$$

scores

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

len x len

$A = \text{softmax}(S)$ by rows

distribution A_i for each word's query q_i

Ex $\quad A = [1 \ \ 0] \quad B = [0 \ \ 1]$

A B ← sequence

"boosted" identify

$$E = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{matrix} \leftarrow A \\ \leftarrow B \end{matrix} \qquad W^K = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

$$Q = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \leftarrow$$

(Greg says) "find Bs"

$$K = E(W^K)^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

Last step:

$$\text{Output} = AE$$

"weighted sum" of E

$$S = QK^T$$

$$= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} = \begin{bmatrix} 0 & 10 \\ 0 & 10 \end{bmatrix}$$

$$A = \text{softmax}(S) = \begin{bmatrix} 0 & 10 \\ 0 & 10 \end{bmatrix} \begin{array}{l} \rightarrow \text{sm} \rightarrow \\ \rightarrow \text{sm} \rightarrow \end{array} \begin{array}{cc} A & B \\ 0 & 0.999 \\ 0 & 0.999 \end{array}$$

- Big K made our probs. peaked
- Q had B for each row $\Rightarrow$ prob on B

A A B A

$$W^K = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

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

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

$$K = E(W^K)^T = \begin{bmatrix} 10 & 0 \\ 10 & 0 \\ 0 & 10 \\ 10 & 0 \end{bmatrix}$$

$$S = \begin{bmatrix} 0 & 0 & 10 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 10 & 0 \end{bmatrix}$$

4x4

Softmax

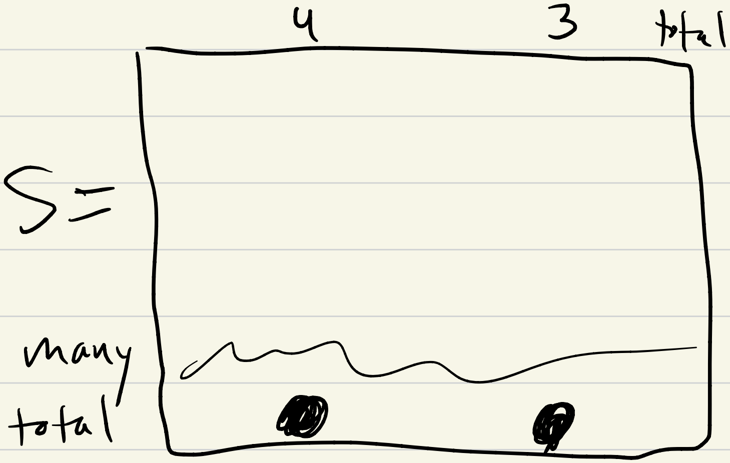
→ high prob on B

S: for word i, how much does it "attend" to word j

Mary had 4 apples. Jane had 3.
How many total? ___

12 words

12x12 matrix



result vector

$$\sum_{i=1}^{seqlen} A_{total,i} \cdot E_i$$