CS 378 Lecture 16: Transformers

Announcements

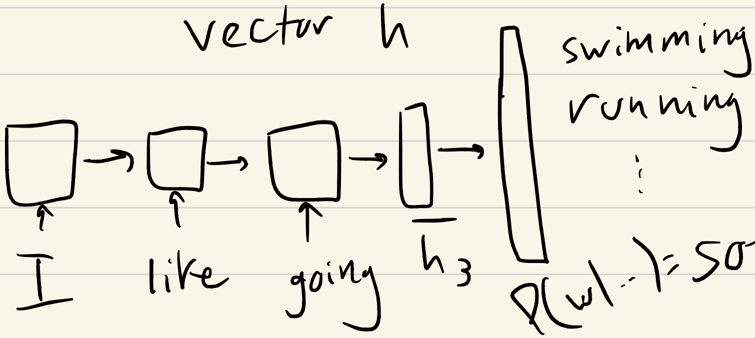
- Midterm back
- A4 out
- Custom FPs due Thurs

Recap Language modeling

$$P(\bar{w}) = \prod_{i=1}^{n} P(w_i \mid w_1 \ldots w_{i-1})$$

N-grams: $P(w_i \mid w_{i-n+1}, \ldots, w_{i-1})$

RNNs: encode whole sequence into vector h

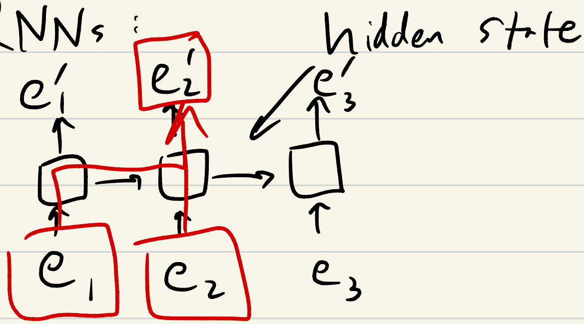


$P(w \mid \ldots) = \text{softmax}(W\bar{h}_3)$

<u>Today</u> Transformers:
- Attention
- Self-attention
- Details: masking and position encoding
- Transformer architecture

Transformer Abstraction

RNNs:



$e_2' = \text{function}(e_1, e_2)$

$$(e_1', e_2', e_3') = \text{RNN}(e_1, e_2, e_3)$$

hidden state at time i is a contextualized embedding of e_i

e_5'

I'm scared of bats

e_3'

I swing bats

Transformer: layer that contextualizes words based on other words in the sequence

same "API" as RNN

$$(e_1', e_2', e_3') = \text{Transformer}(e_1, e_2, e_3)$$

Running example:

Suppose we have seqs of 0s and 1s

00000 if all 0s $\Rightarrow$ ends in 0

01101 if any 1 $\Rightarrow$ ends in 1

$\underbrace{1000}_{\text{OR}}\boxed{1}$

RNN won't do well on

10000 . . . . — 1

(100 0s)

info needs to travel through 100 cells

Attention: allow us to attend to certain elements of the context (we want to find 1s)

Keys, a query

Keys: embedded versions of the sequence.

Assume: $0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ one-hot embs.

keys $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$

0 0 1 0

query: what we want to find

$q = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ want to find 1s

Attention

① Compute score for each key

$s_i = k_i^T q$ dot product

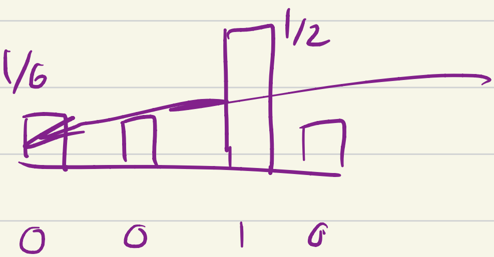
s: 0 0 1 0

0 0 1 0 query = 1 = $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$

② softmax the scores to get probs.

$\bar{\alpha} = \text{softmax}(\bar{s})$

Assume e=3



$$\frac{e^0}{e^0 + e^0 + e^1 + e^0} = \frac{1}{6}$$

③ Compute output value

$$\text{result} = \sum_{i=1}^{n} \alpha_i e_i$$

$$= \frac{1}{6}\begin{bmatrix}1\\0\end{bmatrix} + \frac{1}{6}\begin{bmatrix}1\\0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0\\1\end{bmatrix} + \frac{1}{6}\begin{bmatrix}1\\0\end{bmatrix}$$

$$= \begin{bmatrix}1/2\\1/2\end{bmatrix}$$

(no attn) average of all 4 vectors $= \begin{bmatrix}3/4\\1/4\end{bmatrix}$

What if we had the seq

0 0 00 and $q = \begin{bmatrix}0\\1\end{bmatrix}$ (looking for 1)

attns: 1/4 1/4 1/4 1/4

result: $\begin{bmatrix}1\\0\end{bmatrix}$

Compare: with 1 vs no 1

Problem: long seq $\Rightarrow$ attn not very peaked?

Modify the keys

Before: $k_i = e_i$ (embedding)

Now: $k_i = W^K e_i$

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

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

$k_i^T q$

seq: 0 0 1 0

Score: 0 0 10 0

attns: - - 0.999 - $\dfrac{e^{10}}{1+1+e^{10}+1}$

Formulas for attention:

dot product: $k_i^T q$

"linear" attn: $k_i^T W q$

Can view it as $(W^T k_i)^T q$

or $k_i^T (W q)$

W either affects k or q

In reality: W^K, W^Q both

$(k_i^T W^K)(W^Q q)$

Self-attention:

every word is a key and a query simultaneously

($d = 2$, emb. dim)

Q: seq len $\times$ d

K: seq len $\times$ d

we want to find ls

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

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

(in general: these differ)

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

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

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

Scores

$$\underbrace{S}_{\text{len} \times \text{len}} = \underbrace{Q}_{\text{len} \times d} \underbrace{K^T}_{d \times \text{len}}$$

$S_{ij} = q_i$ (ith row of Q) $\cdot K_j$ (jth row of K) scores for query 1

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ & & & \\ & & & \end{bmatrix}$$

attns for query 1

$$A = \mathrm{softmax}(S) = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ & & & \\ & & & \end{bmatrix}$$

<u>Example</u> Let's take 0 1 as the sequence

① $W^Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ (identity) $\quad W^K = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$

emb for word 1

$Q = \mathrm{Embs}\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$ for word 2

$\cdot\ (W^Q)^T$ identity $= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

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

"word at posn 1 is looking for 0s"

"word at posn 2 is looking for 1s"

$$K: \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$ boosted E

score for $q_1 + k_1$

$q_1 + k_2$

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

$$A = \text{softmax}(S) = \begin{bmatrix} 0.999 & 0 \\ 0 & 0.999 \end{bmatrix}$$

sequence: 0 1

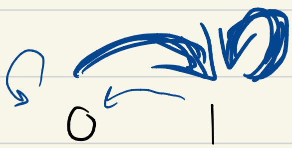
(2) $W^Q = [\ ?\ ]$ $\quad E(W^Q)^T$

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

k scales

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

$$A = \begin{bmatrix} 0 & 0.999 \\ 0 & 0.999 \end{bmatrix}$$



attns - embs

Output: AE

more params

(In Transformer paper: $W^O AE$)

In paper: $\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

scaling

$E(W^V)^T$

more params