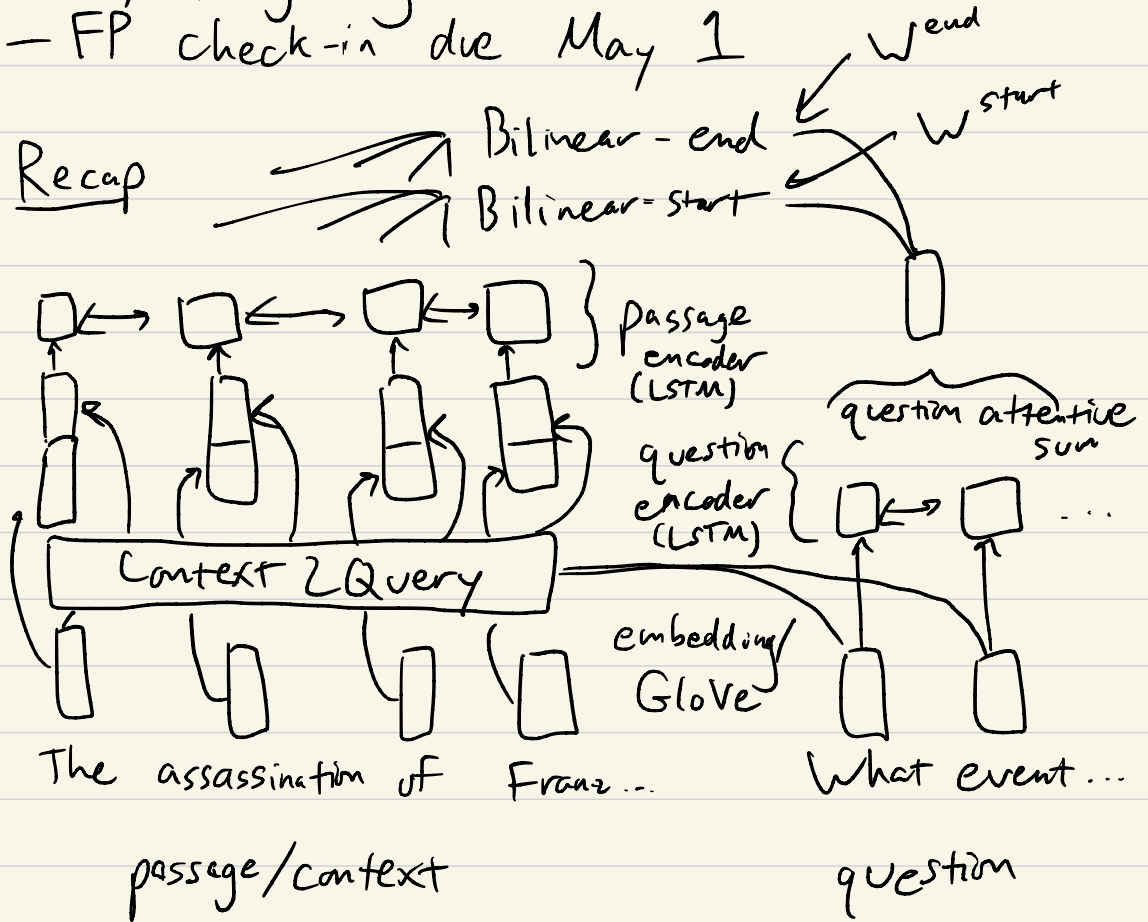


Announcements

- A4, A5 grading
- FP check-in due May 1

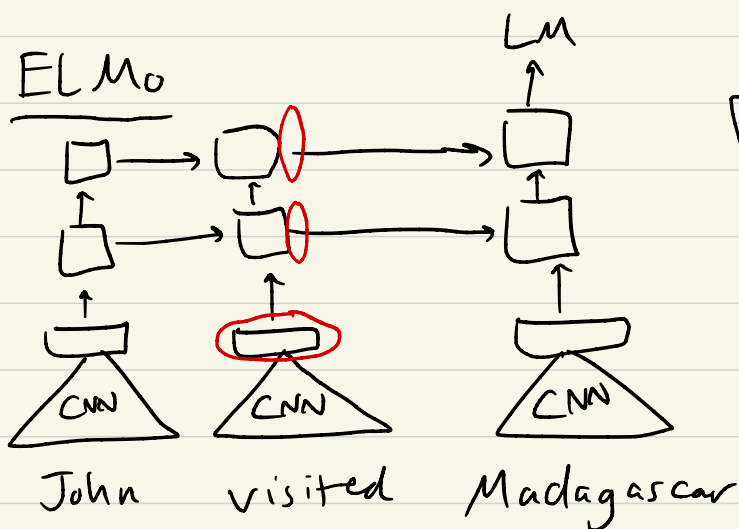
Recap



When batching q_s have diff lengths

$$\begin{matrix} \text{=====} \\ \text{=====} \\ \text{=====} \end{matrix} = \begin{bmatrix} q_{11} & q_{12} & \dots \\ \vdots & \vdots & \vdots \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & \dots \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

question-mask



Pre-trained model: train on a big unlabeled corpus, use "downstream" task

forward representation of "visited"
backward model too: concatenate repr w/ forward

Use this instead of GloVe

Today ① Transformers: self-attention
② BERT: upgraded ELMo w/ Transformers

Rest of the course: Thursday: applications
Next Tues: multilinguality
Next Thurs: ethics

Transformers LM revisited

$$P(\bar{w}) = \prod_i P(w_i | w_1, \dots, w_{i-1})$$

n-gram: look at $n-1$ prior words

RNN: look at all prev words

"weighted" towards recent ones

Ex In October, people in the US celebrate

July → Independence. Halloween

- Model needs to look at far away information, but sparingly

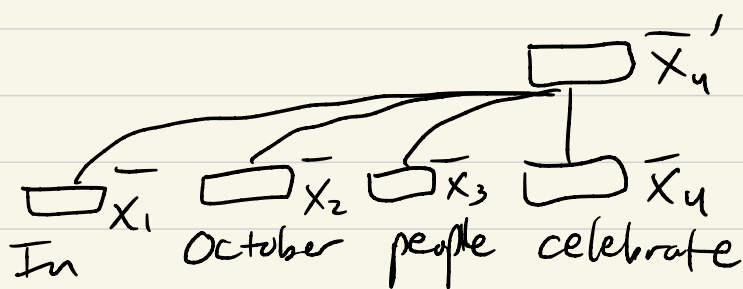
Emily really likes to go to the movies.

↑
she

Rather than modeling whole context continuously, need sparse accesses to some previous words

Transformer

$$u = P(w_5 | w_1 \dots w_4) \\ = \text{FFNN}(\bar{x}_5')$$



(multiply by
voc size \times hidden
matrix, ...)

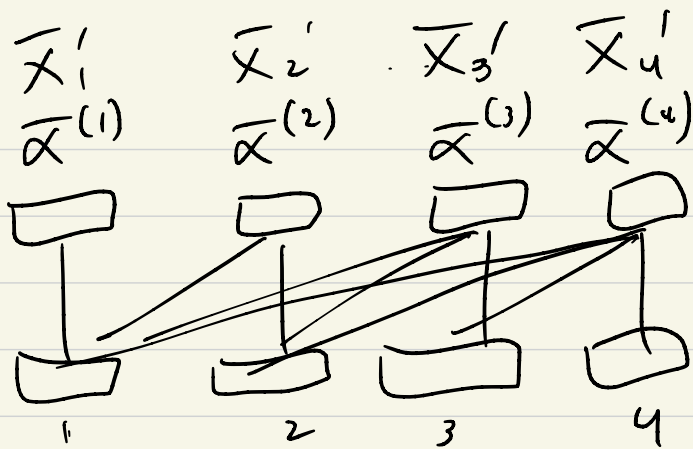
\bar{x}_5' is computed using self-attention

\bar{x}_4 "key", $\bar{x}_1, \dots, \bar{x}_4$ "values"

$$\alpha^{(u)} = \text{softmax} \begin{pmatrix} \bar{x}_4^T W \bar{x}_1 \\ \bar{x}_4^T W \bar{x}_2 \\ \vdots \end{pmatrix} \quad \text{dist over } 1, 2, 3, 4$$

$$\bar{x}_5' = \sum_i \alpha_i^{(u)} V \bar{x}_i \quad W, V \text{ params}$$

self-attention: \bar{x}_4 attends to $\bar{x}_1, \dots, \bar{x}_4$
each word "informs itself" about
its context



self-attention
layer:
each word
does a
self-attn
computation

Properties:

n vectors in \Rightarrow n vectors out

\hookrightarrow behaves like LSTM layer

Each word can look at all prior words directly

Emily did
 $\underbrace{\hspace{10em}}_{\text{attend}}$

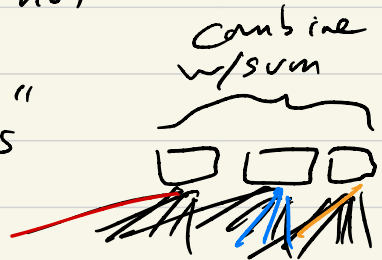
Downside: $O(n^2)$ $\underbrace{\bar{x}_i W \bar{x}_j}_{\text{floating point ops}}$
BUT parallelizable

n words
 $\times n$ attn values
 $\times k$ heads
 $\times l$ layers

Attention "heads"

Attention is often peaked, maybe
balance 2-3 things but not 10

"heads"



In October, people in the US celebrate

$$\alpha^{(j,k)} = \text{softmax}_i \left(\bar{x}_j^T W^{(k)} \bar{x}_i \right)$$

j : "last word"/key

k : index of head $1 \dots 3$

i : "loop var" over context $1 \dots j$

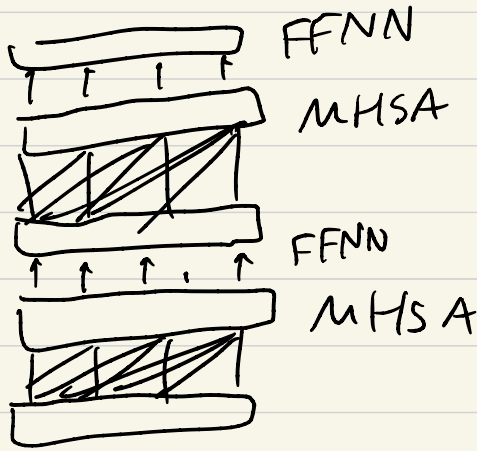
$$\bar{x}'^{(k)}_j = \sum_i \alpha^{(j,k)}_i V^{(k)} \bar{x}_i$$

$W^{(k)}, V^{(k)}$ $k \times 2$ matrices w/
 diff params for
 each head

Transformer

MHSA

stacked multi-head self-attn + feedforward layers



LSTM: 2 layers

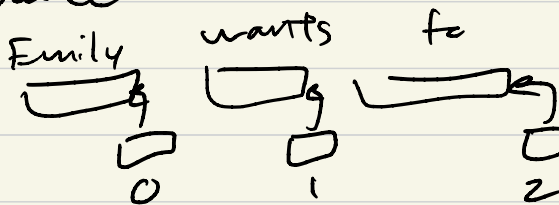
Transformer: 6+ layers

Positional encoding

Emily wants to talk to John. —

John wants to talk to Emily. —

look the same

Solution: 
append
posn embedding

trainable embedding layer