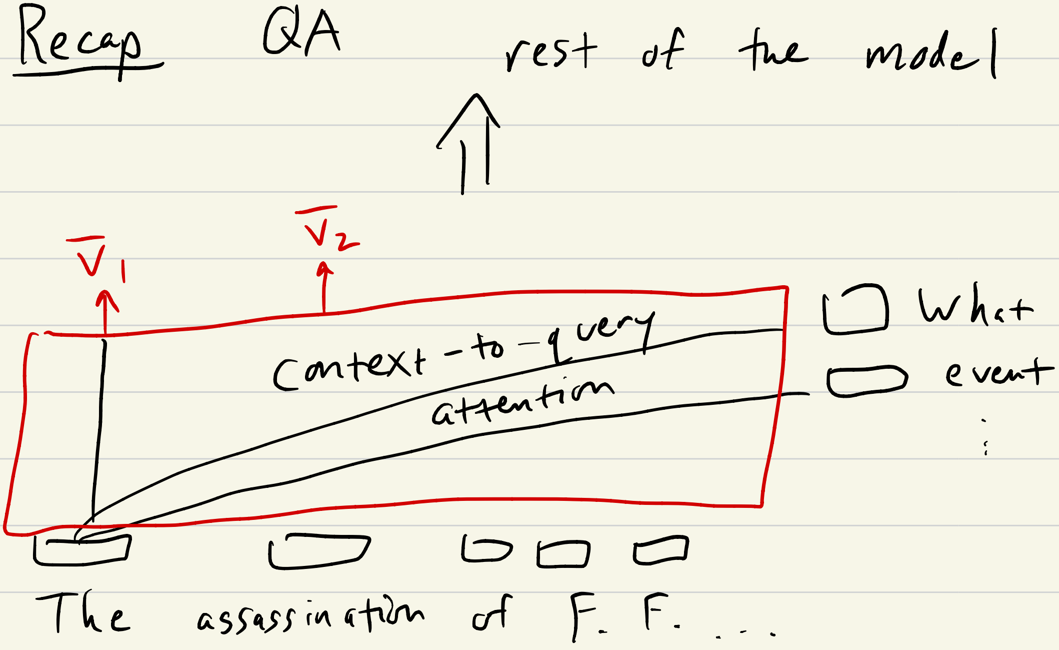
CS 378 Lecture 24

Today

1. Self-attention for language modeling
2. Transformers
3. BERT
4. Analysis + results of BERT

Recap QA



ELMo: train a RNN LM on lots of data, use it to produce "contextualized" embeddings

---

Announcements

- FP: updated models
- A4 back soon

Self-attention

Lang modeling: $P(\overline{w}) = P(w_1)\, P(w_2 \mid w_1)\, P(w_3 \mid w_1 w_2) \cdots$

n-grams: look at past n-1 words only

RNNs: look at everything, but they can forget stuff

In October, people in the US celebrate ______

Halloween

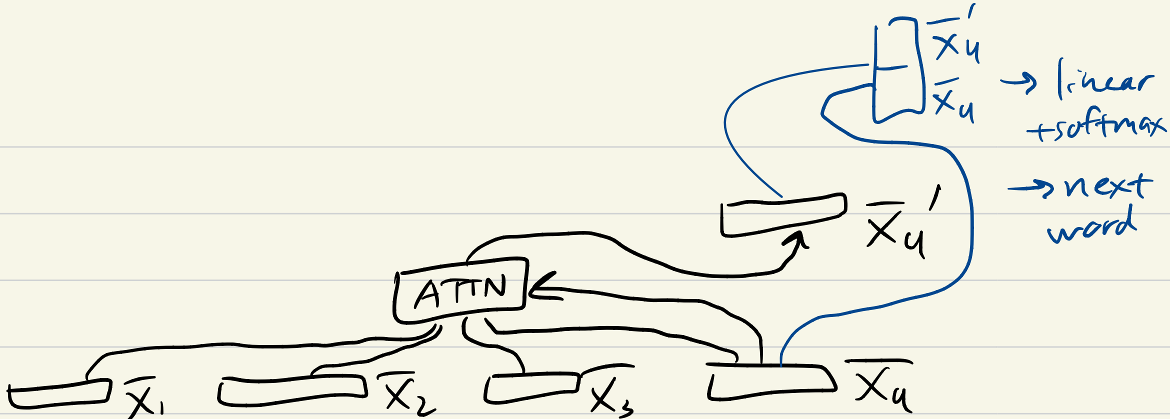
Predicting the next word requires looking back a long way, but sparsely

$\boxed{\text{Alice}}$ really likes to go to the movies with me. She likes horror movies. I'm good friends with ______.

her

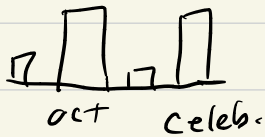
Alice

Self-attention: look back at the sequence so far to predict the next word



In October people celebrate ______

$$\alpha_4 = \text{softmax}_i(\bar{x}_4^T W \bar{x}_i)$$

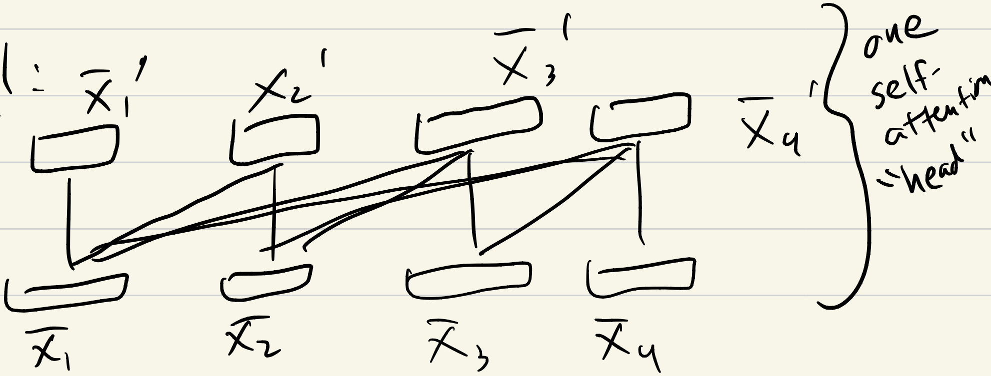


$\bar{x}_4$ "key"

$\bar{x}_1 \ldots \bar{x}_4$ "values" the attention is over

$$\bar{x}_4' = \sum_i \bar{\alpha}_{4,i} \bar{x}_i$$

General:



Follows same abstraction as RNN:

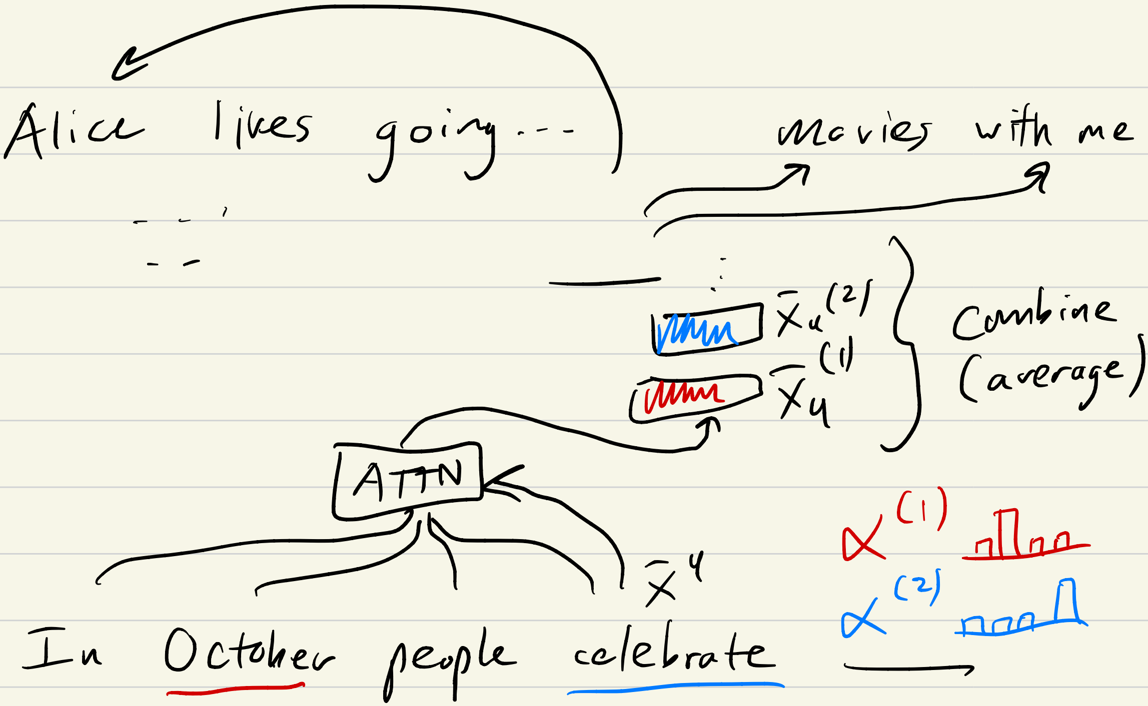
sequence of vectors $\bar{x}_1, \ldots, \bar{x}_n$
$\Rightarrow$ new sequence of vectors where
$\bar{x}_i'$ "knows about" $\bar{x}_1, \ldots, \bar{x}_{i-1}$

Advantages: easy access to past words
parallelizable

Disadvantages: not as powerful as LSTMs
(so far)

We want to look back at lots of
things in the context

Multi-head self-attention: K "heads"
which each do an attn computation



$$\alpha_u^{(k)} = \text{softmax}_i\left(\bar{x}_u^T W^{(k)} \bar{x}_i\right)$$

$$\bar{x}_u'^{(k)} = \sum_i \alpha_{u,i}^{(k)} V^{(k)} \bar{x}_i$$

new param matrix

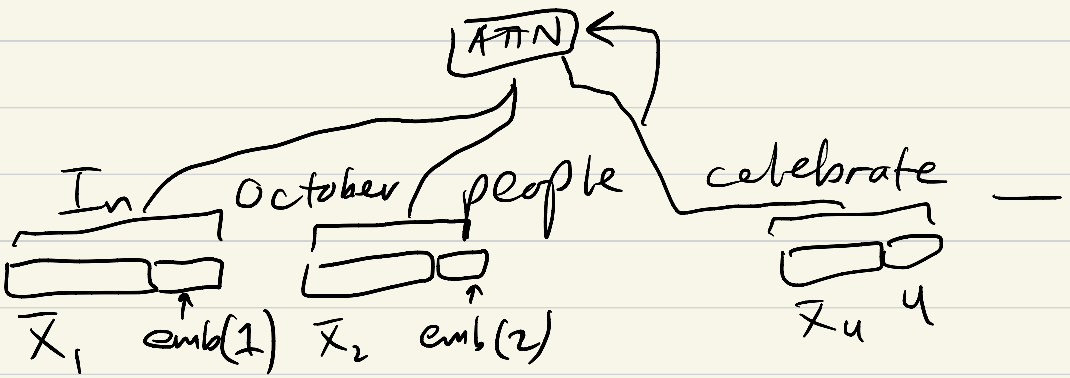
$k = 1 \ldots K$, do independent copies of the computation

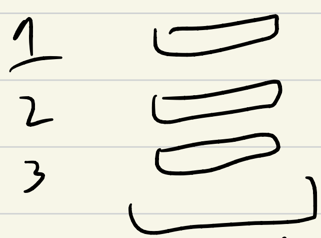
$(W^{(k)}, V^{(k)})$ is a head

Positional encoding

Attention doesn't know the order of the words

Solution: encode position into $\bar{x}_i$

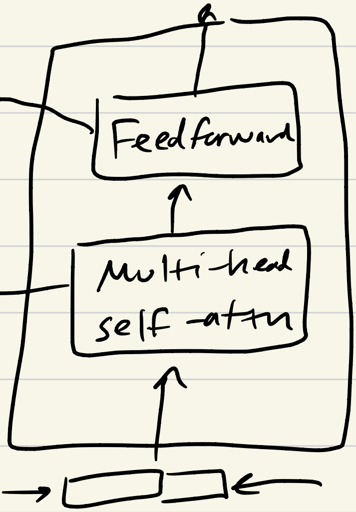




50-dim embs, trained with the rest of the model

Transformer

more params



many layers

K=16 heads

16 x (W, V) matrices

word embs

posn embs