

Self-attention

Language modeling: $P(w_i | w_1, \dots, w_{i-1})$

In October, people in the US
celebrate _____
Halloween

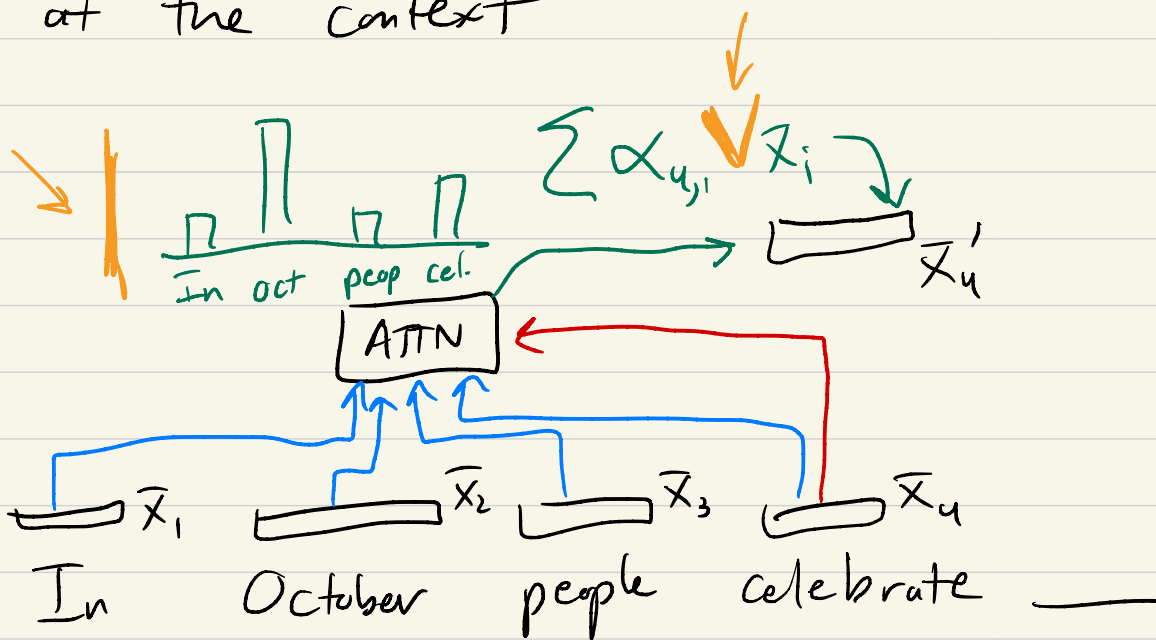
How to predict?

LSTM: "read" the context in order

Alice really likes to go to the movies.
She _____
"Hey _____" Alice

Self-attention: attention over the words in a sequence that is being predicted

Allows us to look back "sparsely" at the context

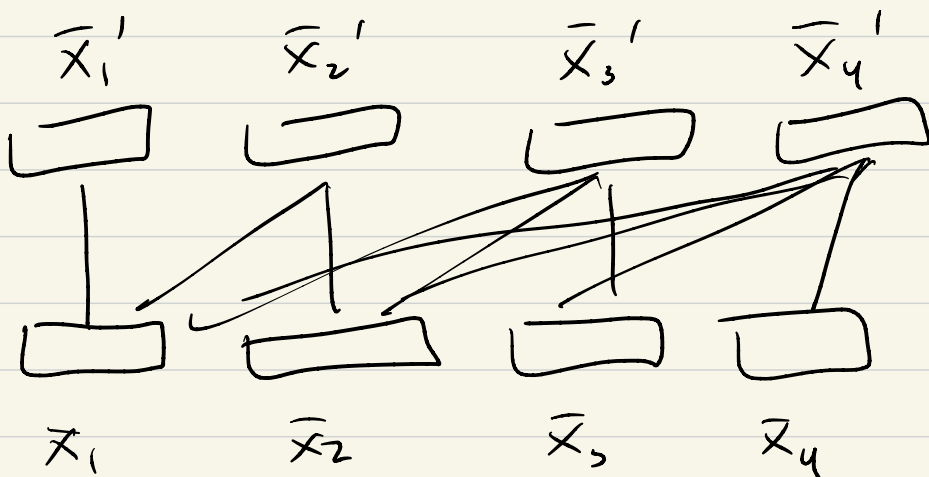


\bar{x}_4 "key" ●

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

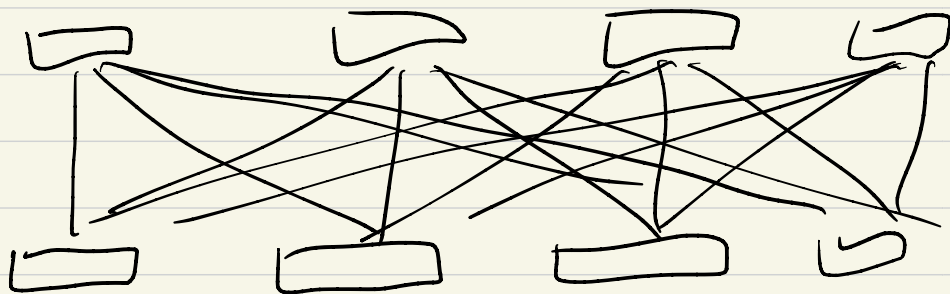
$$\bar{x}_u = \text{softmax}_i(\bar{x}_u^T W \bar{x}_{(i)})$$

one attention "head" (W, V)



"causal" self-attention
(one-directional)

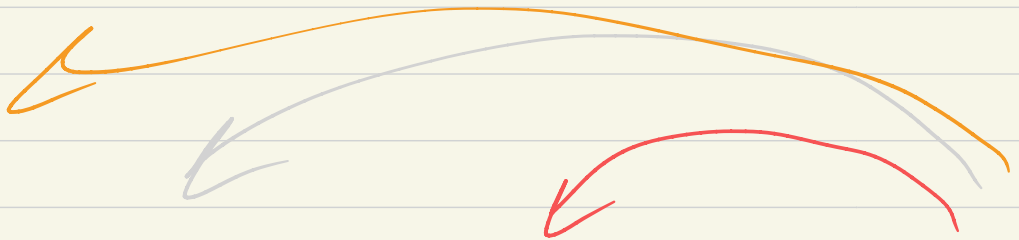
LM = GPT-x



bidirectional (full) self-attn

BERT

Multi-head self-attention: do several copies of attention in parallel



In October people celebrate _

The diagram shows three curved arrows above the text, representing parallel attention heads. The top arrow is orange and points from the end of the sentence back to the word 'October'. The middle arrow is light blue and points from the end of the sentence back to the word 'people'. The bottom arrow is red and points from the end of the sentence back to the word 'celebrate'.