Announcements

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

Recap

The assassination of Franz...

passage/context

question

as have diff lengths

When batching

\[ \text{question-mask} \]
ELMo

John visited Madagascar

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 LMs revisited

\[ P(\mathbf{w}) = \prod P(\mathbf{w}_i | \mathbf{w}_{i-1}, \ldots, \mathbf{w}_1) \]

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, October → Halloween

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

Emily really likes to go to the movies.

She

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

\[ W^u: P(w_5|w_4..w_1) = \text{FFNN}(\bar{x}_u) \]

\[ (\text{multiply by } \text{voc size } \times \text{hidden matrix}) \ldots \]

\[ \bar{x}_u \text{ is computed using self-attention} \]

\[ \bar{x}_u \text{ "key"}, \bar{x}_1, ..., \bar{x}_y \text{ "values"} \]

\[ \bar{\alpha}^{(u)} = \text{softmax} \left( \frac{\bar{x}_u^T W \bar{x}_1}{\bar{x}_u^T W \bar{x}_2} \right) \text{ dist over } 1, 2, 3, y \]

\[ \bar{x}_y' = \sum_{i} \bar{\alpha}_i^{(u)} V \bar{x}_i \quad W, V \text{ params} \]

Self-attention: \( \bar{x}_u \) attends to \( \bar{x}_1, ..., \bar{x}_y \)

each word "informs itself about its context"
Properties:

- Each word can look at all prior words directly
- Behaves like LSTM layer

Downside: $O(n^2)$ ops

BUT parallelizable
Attention "heads"

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

In October, people in the US celebrate

\[ \alpha(i, k) = \text{softmax}_i \left( \frac{X_j T W^{(k)} X_i}{Q, K} \right) \]

j: "last word"/key
k: index of head 1...3
i: "loop var" over context 1...j

\[ X'(k) = \sum_i \alpha(i, k) V^{(k)} X_i \]

W(k), V(k) K (x2) matrices w/ diff params for each head
Transformer

MHSA

stacked multi-head self-attention + feedforward layers

Transformer: 6+ layers

MHSA

FFNN

MHSA

FFNN

MHSA

Positional encoding

Emily wants to talk to John. —
John wants to talk to Emily. —
look the same

Solution: Emily wants to

append posn embedding trainable embedding layer