CS 378 Lecture 24

Today

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

Recap Q&A rest of the model

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

---

Announcements
- FP: updated models
- AU back soon

---

Self-attention

Lang modeling: \( P(w) = P(w_1) P(w_2|w_1) P(w_3|w_1,w_2) \ldots \)

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.

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

\[ \alpha_y = \text{softmax} \left( \sum_{i} x_y^T W x_i \right) \]

\( x_4 \) “key”
\( x_1 \ldots x_4 \) “values” the attention is over

\[ x_4' = \sum_{i} \alpha_{y,i} x_i \]

General: \( x_1, x_2, x_3, x_4 \)

\{ self-attention “head” \}
Follows same abstraction as RNN:

Sequence of vectors \(X_1, ..., X_n\) ⇒ new sequence of vectors where \(X_i \) "knows about" \(X_1, ..., X_{i-1}\)

**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
Alice likes going... movies with me.

In October people celebrate.

\[\alpha_{y_i}^{(k)} = \text{softmax}_i \left( \overline{x}_y^T W^{(k)} \overline{x}_i \right)\]

\[\overline{x}_y^{(k)} = \sum \alpha_{y_i}^{(k)} V^{(k)} \overline{x}_i\]

For \(k = 1\ldots K\), 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 $X_i$

In October people celebrate $x_u y$

$X, \text{emb}(1), X, \text{emb}(2)$

1
2
3

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

more params

Feedforward

Multi-headed self-attention

K = 6 heads

16 x (W, V) matrices

word embs → multi-layer feedforward

many layers

post embs