CS 378 Lecture 20

Today
- Attention in seq2seq models
- Neural machine translation

Recap seq2seq models

\[
P(t_i | \bar{s}, t_1, \ldots, t_{i-1})
\]

Training Assume everything correct through \(i-1\), maximize \(\log\) prob of word \(t_i\)

loss = \[ \sum - \log P(t_i | \bar{s}, t_1, \ldots, t_{i-1}) \]
Announcements
- A4 due
- A5 out tonight (due in 1 week)

Problems with seq2seq models

1. Model repeats itself in a loop

Je fais ... \(\Rightarrow\) I make a desk a desk
a desk a desk...

Why didn't this happen in phrase-based?
We had a notion of coverage!

RNN doesn't track "progress"

2. Fixed vocabulary

Elle est allée à Pont-de-Buis \(\Rightarrow\) She went to UNK
③ Bad at long sentences

LSTMs have fixed hidden state, 50+ time steps is hard

Attention

I make a desk EOS

Je fais un bureau sos I make a desk

Requires modifying LSTM to take 2 inputs

Suppose it was always a word-by-word translation in order
Decoder:

\[ \text{I} \]

\[ \text{SOS Je} \]

RNN just needs to map Je to I.

Can learn this more easily with fewer params than a normal seq2seq.

(Could even delete encoder)

This is too rigid. Instead we want the decoder to softly pick where it looks in the input.

Attention: distribution over input positions

\[ \text{Je fais un bureau} \]
prediction

"loop over all inputs i) compute this score for $X_i$, softmax"

$\alpha = \text{softmax}_i \left( \frac{h_i^T W x_i} {c_i} \right)$

$\sum_i \alpha_i x_i$

$P(t, l | s) = \text{softmax} \left( \text{vocab} \left[ \sum_{i=1}^{l} \frac{h_i}{c_i} \right] \right)$

$c_i \approx \text{vec} \text{emcct}$

Do this at every timestep!

Training: with backprop
Details:

- Many ways to set this up

\[ \alpha = \text{softmax} \left( f(\tilde{h}_i, \bar{x}_i) \right) \]

\[ f: \text{dot product}, \quad \tilde{h}_i^T W \bar{x}_i \]

one-layer NN

Overall idea: form context vector \( \bar{c} \) that captures input directly
Poll: $x_2$

$\text{Je vais le faire}$

? look back at the verb

↑ look back at position 2 in the (index) source

I am going to do it

① am: look at vais
② look at Je (subject)
look at other stuff?
(other dependencies of the verb)

③ Encoder tracks: content

Decoder timestep 1: model wants to attend to enc. timestep 1