Announcements

- A4, A5 grading underway
- FP check-in May 1

Recap Reading comprehension: Attentive reader

\[ P(\text{start} | p, q) = \text{softmax}(\bar{p}_i W_{\text{start}} q) \]
\[ P(\text{end} | p, q) = \ldots \]

Input: doc and question
Output: distributions over start and end points of answer
Today

1. Two improvements to Attentive Reader
2. SQuAD state-of-the-art
3. Pre-training with ELMo

Improvements to AR

$p, q$ encoders more complex

1. “context to query” attention

Intuition: reducing the question to $\overline{q}$ throws out info
→ hard to represent entities in a fixed vector

\[
\overline{x}^{(i)} = \text{softmax}(\overline{x} : \overline{q}_i)
\]

Event

\[
\overline{v}^{(i)} = \sum_j \alpha_j^{(i)} \overline{q}_j
\]

Attention over question for each passage word
\( \overline{X}^{(i)} \) : distribution over \( \overline{q}_1, \ldots, \overline{q}_m \) for the \( i \)th passage token

\( \overline{V}^{(i)} \) : representation of the question "tailored" to this passage token

Suppose \( \overline{X}_8 = \text{WW1} \)

\( \overline{q}_7 = \text{WW1} \)

\( \overline{V}^{(8)} \approx \text{WW1} \approx \overline{X}^{(8)} \)

We can see at this point that the question has matched the passage

By injecting into about \( \overline{q} \) earlier, the model can do better at finding the right spots.
(2) question encoding modification

Attentive reader

\[ \overline{q_1} \rightarrow \overline{q_2} \rightarrow \ldots \rightarrow \overline{q} \rightarrow \text{use in bilinear layers} \]

What event

Switch to a more "bag of words" like representation

\[ b_j = \text{softmax}_j (\overline{W} \cdot \overline{q_j}) \]

dist over \( q \) tokens

trainable param vector

\[ \overline{q} = \sum b_j \overline{q_j} \]

looks like attm

but uses a fixed "key" \( \overline{w} \)

"span attention"