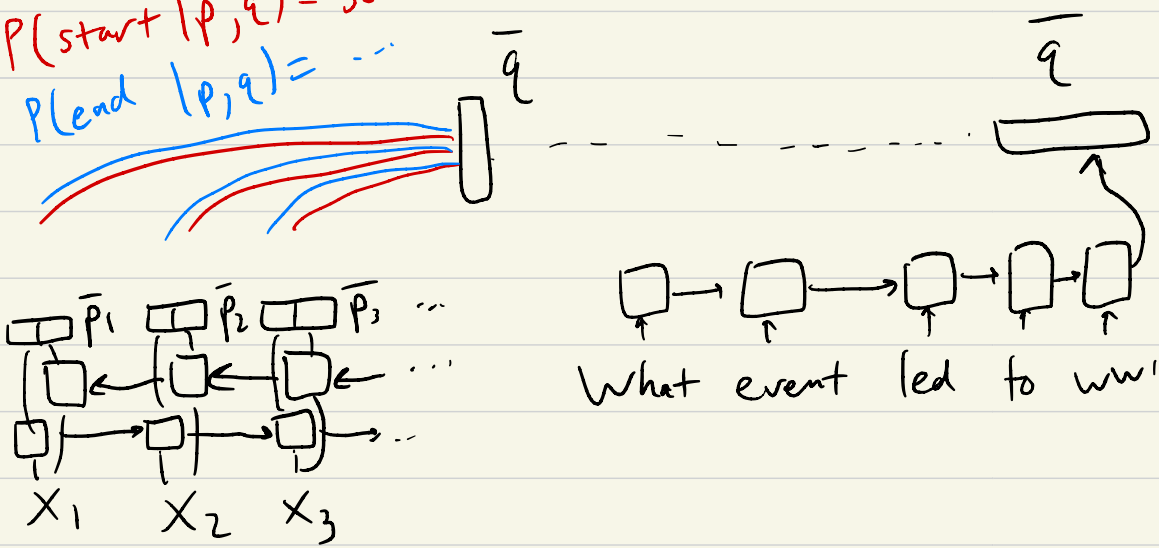
Announcements

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

Recap Reading comprehension: Attentive reader

$P(\text{start} \mid p, q) = \text{softmax}(\bar{p}_i W^{\text{start}} \bar{q})$

$P(\text{end} \mid p, q) = \ldots$



The a. of F F caused ...

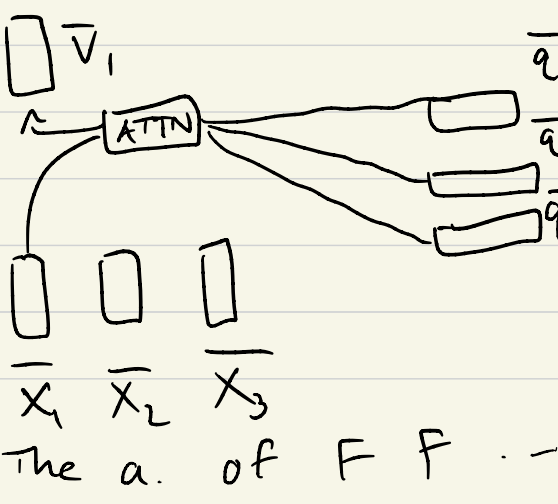
Input: doc and question
Output: distributions over start and end points of answer

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

Improvements to AR

$\bar{p}, \bar{q}$ encoders more complex

① "context to query" attention
Intuition: reducing the question to $\bar{q}$ throws out info
→ hard to represent entities in a fixed vector



$$\bar{\alpha}^{(i)} = \underset{j}{\text{softmax}}\left(\bar{x}_i \cdot \bar{q}_j\right)$$

$$\bar{v}^{(i)} = \sum_j \alpha_j^{(i)} \bar{q}_j$$

Attention over question for each passage word

indices 1 ... m

$\bar{\alpha}^{(i)}$: distribution over $\bar{q}_1 \dots \bar{q}_m$ for the ith passage token

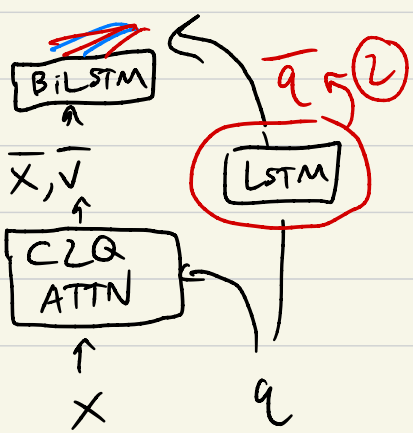
$\bar{v}^{(i)}$: representation of the question "tailored" to this passage token

Suppose $\bar{x}_8 = ww1$

$\bar{q}_7 = ww1$

$\bar{v}^{(8)} \approx ww1 \approx \bar{x}^{(8)}$

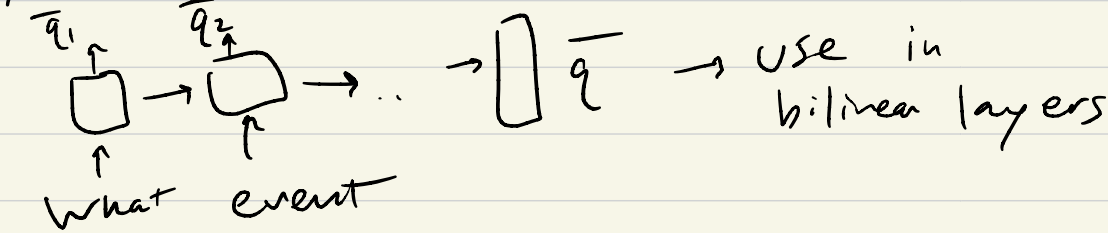
We can see at this point that the question has matched the passage



By injecting info about $\bar{q}$ earlier, the model can do better at finding the right spots

(2) question encoding modification

Attentive reader

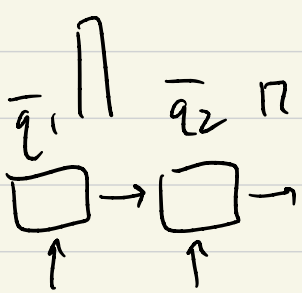


Switch to a more "bag of words" like representation

$$b_j = \text{softmax}_j\left(\bar{w} \cdot \bar{q}_j\right)$$

dist over q tokens

$\bar{w}$: trainable param vector



$$\bar{q} = \sum b_j \bar{q}_j$$

looks like attn but uses a fixed "key" $\bar{w}$

"span attention"