Pay Attention: Reading Comprehension on SQuAD
Sajana Weerawardena

Overview
- Reading Comprehension has been an old goal of General AI
- Stanford Question Answering Dataset (SQuAD) is ground breaking because it provides a large dataset with realistic content (100,000+ Question/Answer pairs and 500+ contexts)
- Goal: Implement and understand a model for SQuAD
- Implementation: BiDAF without a Char-CNN

Models
- Inputs:
  - Context words \( w_1, \ldots, w_n \) and question words \( q_1, \ldots, q_m \)
- Output: start and end index of answer in context.
- Baseline
  - Encoder: 2 LSTMs with Dropout
  - Attention: Bidirectional Attention with a modified similarity function:
    \[ S_{ij} = \text{sim}(w_i, q_j) \in \mathbb{R} \]
  - Decoder: fully connected layer that feeds into pair of softmax activations.
- BiDAF
  - Encoder: 2 LSTMs with Dropout
  - Attention: Bidirectional Attention with original similarity function:
    \[ S_{ij} = a^n \text{sim}(u_i, q_j) \in \mathbb{R} \]
  - Modelling: 2 LSTMs with Dropout
  - Final: fully connected layer that feeds into softmax activations

Results
- Train Loss, EM, F1: EM of 83.90, F1 of 92.53
- Dev Loss, EM, F1: EM of 51, F1 of 66

Results in Context
- It took far longer to train the BiDAF
- The BiDAF EM was over 15 points above the baseline: F1 was 20 points over
- This large difference in train and dev: overfitting in the model.
- However, when we look at the train loss and the dev loss, at their closest they were just 0.3 apart
- EM and F1 scores not having a close correspondence with the loss
- the leading model has scored 83.87% EM and 89.73% F1: long way to go

Sample Analysis
- Analysed 70 samples from dev set (0.0064% of Dev set), looking at question types. 58% 42% EM
- Question and Answer Types, from both original paper (Rajpurkar et al (2016)) and my samples
- There is a correspondence between question types and answer types.
- The model performed differently on what/how questions vs why/where/who problems.

What/How Questions:
- 15 EM, 14 Wrong (of 29). Average F1: 55
- Had Off by a Couple Word errors:
  - Intuitively captured the important information
  - Different examples had different ideas of the best length answers (make it harder for the model to learn the pattern)
- Potential Solution: Condition End index on Start Index
- When/Where/Who
- Did better: 90 EM, 63 EM and 73 EM respectively.
- Did well on lexical variation (governed vs run) drawing from the word embeddings
- Did extremely well when the question is just the answer paraphrased into a declarative form