Question Answering on SQuAD Dataset

Zihuan Diao, Junjie Dong, Jiaxing Geng
{diaozh, junjied, jg755} @ stanford.edu

Stanford | ENGINEERING

Overview

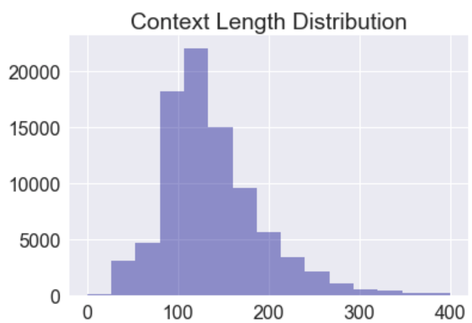
Motivation
Machine comprehension (MC), answering a question based on a given context, has gained great interests in recent years. Researchers have made significant progress applying deep learning approaches, especially variations of recurrent neural networks and attention mechanisms, to close the gap between machine performance and human performance. In this project, we build an end-to-end neural network model to tackle this problem.

Problem Definition
Given a sequence of context words $W^C = [w_1^C, \cdots, w_T^C]$ and question words $W^Q = [w_1^Q, \cdots, w_J^Q]$, find a sub-span of the context (a_s to a_e) that answers the question.

Context: In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail…
Question: What causes precipitation to fall?
Answer: gravity

Dataset

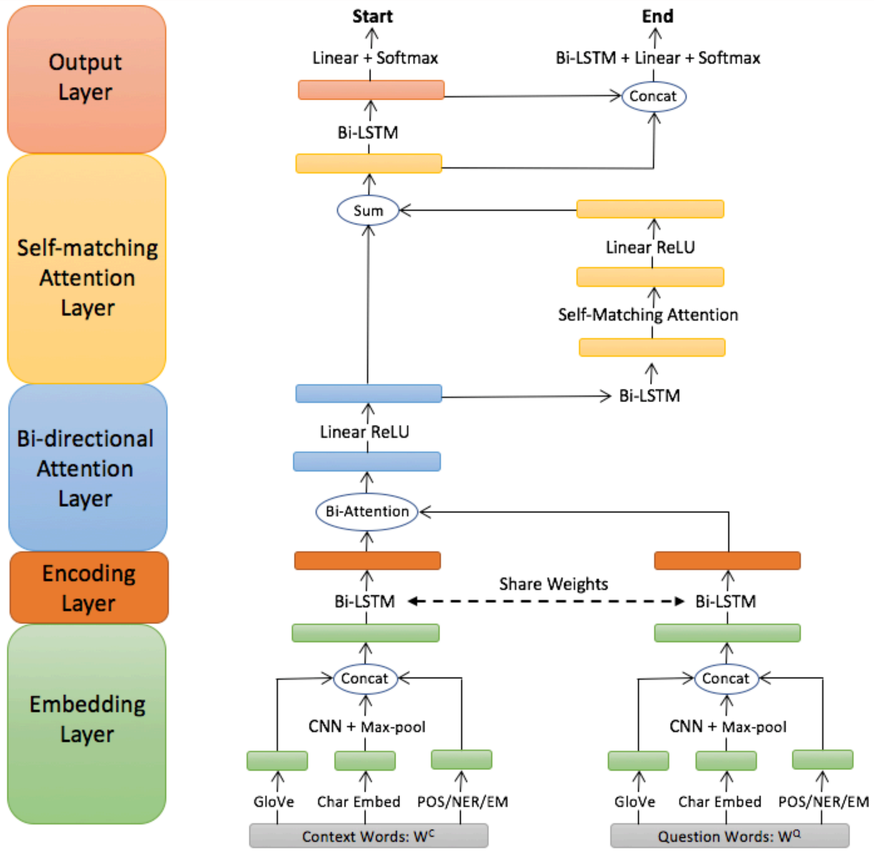
We use the Stanford Question Answering Dataset (SQuAD), which consists of more than 100,000 questions posed by crowdworkers on Wikipedia articles. It is randomly partitioned into a training set (80%), a development set (10%), and a test set (10%).



Question Distribution

Type	%
What	43.25
Who	9.42
How	9.16
When	6.20
Which	4.78
Where	3.77
Why	1.37
(Others)	22.05

Model Architecture



Embedding Layer: Use 300-dimensional GloVe word embeddings, character CNN with 100 filters of various sizes, and POS/NER features extracted using pre-trained models
Encoding Layer: Two bi-directional LSTM networks (weights are shared between the context and question networks)
Bi-directional Attention: Similarity matrix $S_{ij} = w_s^T \cdot [h_t; u_j; h_t \circ u_j]$
Context-to-query: $a_t = softmax(S_{t:})$; $\widetilde{\boldsymbol{u_t}} = \sum_j a_{tj} \cdot u_j$
Query-to-context: b $= softmax(\max_{col}(S))$; $\widetilde{\boldsymbol{h_t}} = \sum_t b_t \cdot h_t$
Self-matching Attention: Similarity matrix $S_{ij} = w_m^T \cdot [g_i; g_j; g_i \circ g_j]$
Attention vector $m_t = \sum_i a_{ti} \cdot g_i$, where $a_t = softmax(S_{t:})$
Decoding Algorithm: Filter sub-spans with $a_s > a_e$ or $a_e > a_s + 13$ then choose the sub-span with the maximum confidence score
Ensemble Method: Choose answer with highest sum of confidence scores ($p_{start} * p_{end}$)

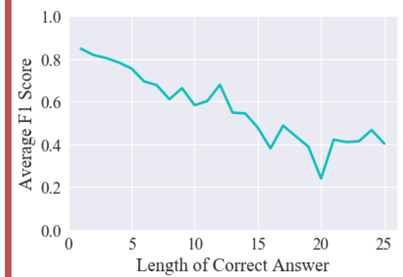
Results

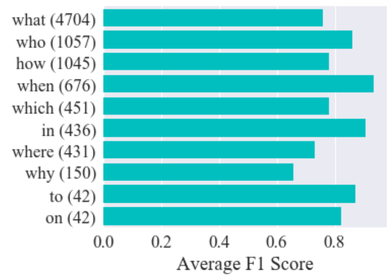
Evaluation: We evaluate our models using the F1 and EM (Exact Match) metrics on the dev and test sets. The table below compares our results to state-of-the-art models.

	Dev Set	Test Set
Single Model	**F1 / EM**	**F1 / EM**
Logistic Regression Baseline	51.0 / 40.0	51.0 / 40.4
Dynamic Chunk Reader	71.2 / 62.5	71.0 / 62.5
BiDAF	77.3 / 67.7	77.3 / 68.0
R-NET	79.5 / 71.1	79.7 / 71.3
Stochastic Answer Networks	84.1 / 76.2	84.4 / 76.8
QANet	- / -	87.8 / 80.9
Our Model	**79.3 / 69.6**	**79.9 / 70.7**
Ensemble Model	**F1 / EM**	**F1 / EM**
BiDAF	80.7 / 72.6	81.1 / 73.3
R-NET	82.8 / 75.6	82.9 / 75.9
Stochastic Answer Networks	85.9 / 78.6	86.5 / 79.6
QANet	- / -	89.0 / 82.7
Hybrid AoA Reader	- / -	89.3 / 82.5
Our Model	**80.8 / 72.4**	**81.9 / 73.8**
Human Performance	- / -	91.2 / 82.3

Analysis

Error Analysis
The model is good at finding relevant context information that is useful for answering the questions, but struggles on questions that require deeper logical reasoning and modeling longer-term dependencies.





Future Work
1. Multiple hops memory network
2. Deep contextualized character embedding such as ELMo
3. Better gating mechanisms to control information flow