Question Answering using simple model with attention

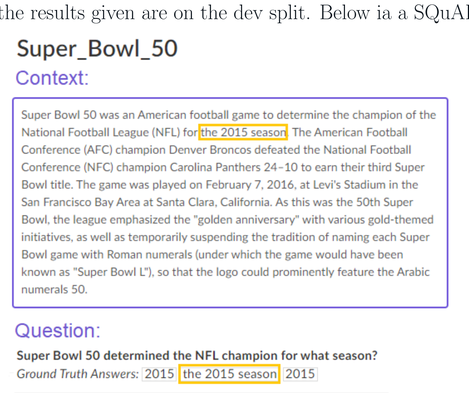
Danit Widder
IDC, Israel

Abstract

Question Answering is a reading comprehension task, consisting of answering a query about a given context paragraph, where the answer is a span within the context paragraph. It requires modeling of the given paragraph as well as the query and interactions between them, in order to output the relevant answer found in the context paragraph. In this project I've built and trained a simplified neural-network model based on BiDAF model suggested in the paper "Bidirectional Attention Flow for Machine Comprehension". The model uses a modular attention layer, in order to compare several attention schemes' results for this task. I've compared a dot-product attention, bi-linear attention and bi-directional attention flow (BiDAF) used by my module and was able to get rather good results on SQuAD v1.1 dataset, given the model's simplicity.

The Problem in More Details

Stanford's SQuAD v1.1 dataset consists of more than 100,000 pair of contexts-queries taken from Wikipedia, and their corresponding answers. The answer contained in the context, so it can be seen as a span of start and end offset in the context. SQuAD 2.0 has 50,000 more contexts-queries for which an answer for the question cannot be obtained from the context. In this project I trained my model on SQuAD 1.1 and all the results given are on the dev split. Below ia a SQuAD dataset example:



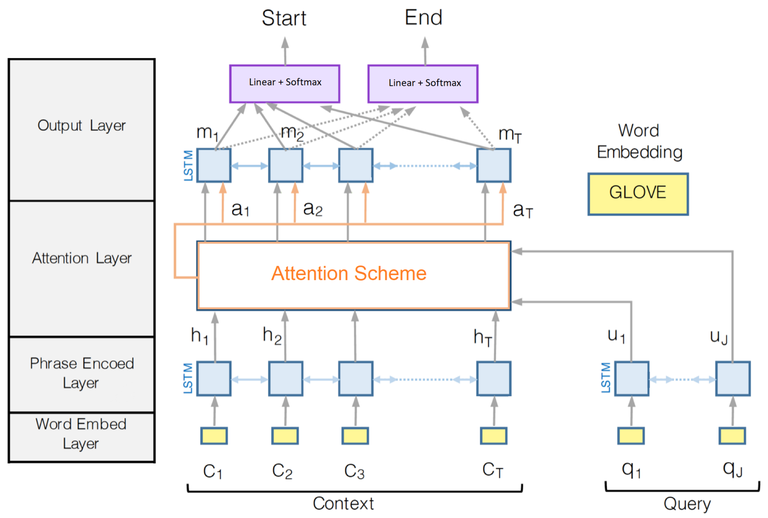
The evaluation scores considered are Exact Match (EM) as well as (F1): The measurement metrics commonly used for this task are:

Exact Match (EM): measures the percentage of predictions that match any one of the ground truth answers exactly. Exact Match is a binary measure (i.e. true/false) of whether the model output matches the ground truth answer exactly. For example, if the model answered a question with 'Einstein' but the ground truth answer was 'Albert Einstein', then the EM score is 0 for that example. This is a fairly strict metric.

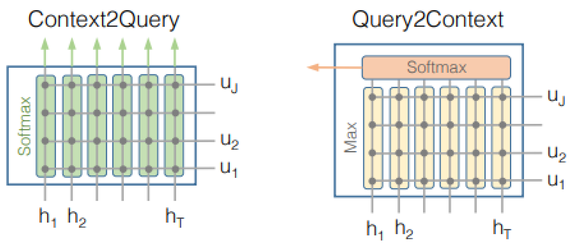
F1: measures the average overlap between the prediction and ground truth answer. We treat the prediction and ground truth as bags of tokens, and compute their F1 score (the harmonic mean of the precision and recall). We take the maximum F1 over all of the ground truth answers for a given question, and then average over all of the questions. F1 is a less strict metric: In the 'Einstein' example, the system would have 100% precision (its answer is a subset of the ground truth answer) and 50% recall (it only included one out of the two words in the ground truth output),thus a F1 score of $2 \cdot prediction \cdot recall/(precision+recall) = 2 \cdot 50 \cdot 100/(100+50) = 66.67\%$

Method

The general approach to the problem was to encode the context and the query separately, and then to combine them using an attention layer, allowing context to query and query to context attention, depending on the chosen attention scheme. The encoded context and attention output are then combined by concatenation and are passed to an output layer that predicts the start and end offset of the answer span in the given context.



The best performed attention scheme was the attention scheme described in the paper "Bidirectional Attention Flow for Machine Comprehension". The main idea of this attention scheme is that attention should flow both ways and computes both Context2Query and Query2Context attention:



Results

The table below presents the results received with Dot-Product, Bilinear and BiDAF attention schemes used with my model, compared to BiDAF model results for single model presented in the BiDAF paper, and to Human Performance on SQuAD 1.1 dataset.

	EM	F1
Dot-Product Attention:	65.31	75.00
Bilinear Attention:	66.33	76.28
BiDAF Attention:	68.57	78.56
BiDAF (single) Model:	68.0	77.3
Human Performance:	82.3	91.2

Conclusions

The main insight is that the attention layer in most Question Answering models is the combining layer of context and query and thus has a very large impact on model's results. Attention schemes which were used for sequence-to-sequence tasks can also be applied for this task, as well as attention schemes which were designed proprietorially for this task. The conclusion from the results is that in this case the proprietary designed attention (BiDAF) has achieved better results then other attention schemes.

Another result of this project is that I was able to achieve similar results on SQuAD 1.1 as were presented on BiDAF paper, using the BiDAF attention layer with my simplified model.

References

SQuAD: https://rajpurkar.github.io/SQuAD-explorer/

BiDAF: https://arxiv.org/abs/1611.01603

Stanford cs224n: http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture10-QA.pdf

https://towardsdatascience.com/nlp-building-a-question-answering-model-ed0529a68c54 Attention:

https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

https://arxiv.org/pdf/1508.04025.pdf

https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html