Allow me to answer : Machine Comprehension with Bi-Directional Attention Flow

Abstract

Machine Comprehension and Question Answering is the task of answering a question based on a given context. The task needs to be carried out by the computer and this paper introduces to an end-to-end model for question answering. The model combines the idea of Bi-Directional Attention Flow (BiDAF) network along with high performing bidirectional Gated Recurrent Units to achieve a F1 score of 39.73 which is comparable to the performance of state of the art models on the SQuAD dataset.

1. Introduction

With the improvements in the state of the art models for Natural Language Processing, the task of machine comprehension for question answering has gained widespread importance. The task involves the machine answering a given question based on the provided context. The answers are a sub phrase of the context with the model predicting the starting and the ending word of the chosen sub phrase.

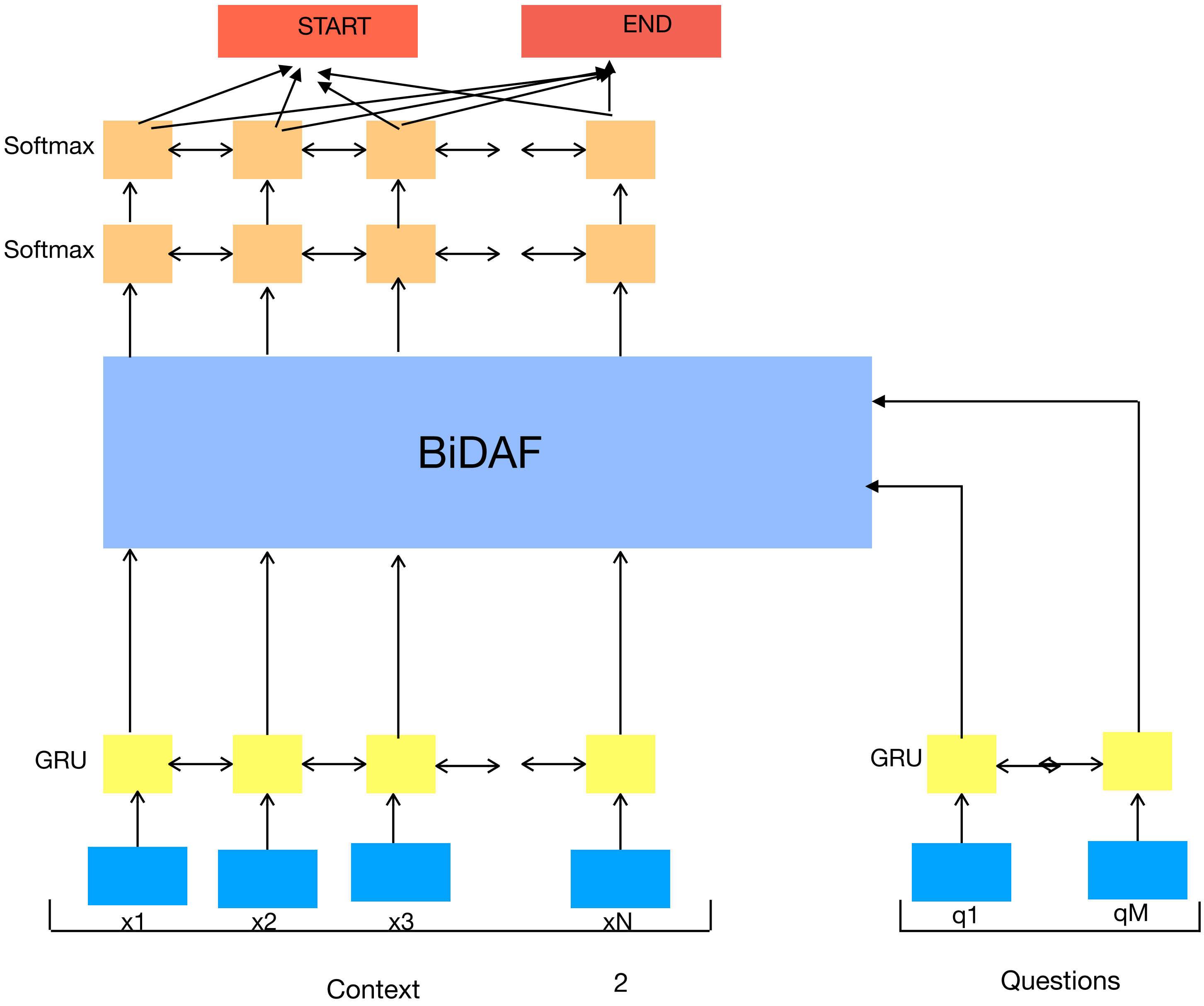
The task of Machine Comprehension for question answering has gained significant research attention with the publication of the SQuAD dataset .[1] The dataset has inspired significant number of research papers int eh field of question-answers and continuously poses a challenge for the researchers to beat the models leading the leaderboard. F1/EM score has been used to compare our model with the baseline model and for complete evaluation the results of the leaderboards have also been charted.

This paper presents our model for the task of machine comprehension for question answering on the SQuAD dataset. The results have been achieved by incorporating the high performing Bi-Directional Attention Flow (BiDAF) model as a substitute for the attention layer in the baseline model and the F1/EM scores have been compared and listed. Another improvement from the baseline models is the use of the bidirectional Gated Recurrent Units instead of vanilla RNNs in our encoder layer. They have provided some improvement over the baseline model as well.

2. Related Work

There has been significant work in this area over the past few year with the publication of the dataset. Some of the notable high performing models have been highlighted across the paper and certain inspiration has been gathered from these works. Dynamic co-attention networks proposed by Xiong et al. [2016] were one of the first models to achieve impressive performance of this task, they proposed the co-attention layer which attended to both the question and the context simultaneously.[2] Chen et al. [2016] used simple bilinear term for computing the attention weights in the same model to drastically improve the accuracy. Wang et al. [2017] introduced many techniques like self matching layers and pointer networks based output layers.

3. Model



3.1 Contextual Embedding Layer

The Context is represented by a d-dimensional vector $x_1, \ldots, x_n \in \mathbb{R}^d$which are the word embeddings for the Context. Similarly the Context is represented by a sequence of d-dimensional word embeddings $q_1, \ldots, q_n \in \mathbb{R}^d$. These are fixed pre-trained GloVe embeddings.[3]

3.2 Encoding Layer

The embeddings are then fed into bi-directional GRU layer which acts as the encoder for eh context and the question embeddings. It outputs a sequence of hidden states for both the context as well as the questions embeddings as follows:

$$\{\overrightarrow{c_1}, \overleftarrow{c_2}, \ldots, \overrightarrow{c_N}, \overleftarrow{c_N}\} = biGRU(\{x_1, \ldots, x_N\})$$
$$\{\overrightarrow{q_1}, \overleftarrow{q_2}, \ldots, \overrightarrow{q_M}, \overleftarrow{q_M}\} = biGRU(\{y_1, \ldots, y_M\})$$

The bidirectional GRUs produce a sequence of forward hidden states ($\overrightarrow{c_i} \in \mathbb{R}^h$ for the context and $\overrightarrow{q_j} \in \mathbb{R}^h$ for the question where h is the number of hidden GRU units) and a sequence of backward hidden states ($\overleftarrow{c_i} \in \mathbb{R}^h$ and $\overleftarrow{q_j} \in \mathbb{R}^h$). The bidirectional GRU concatenates these states to produce context hidden states c_i and question hidden states q_j which are give as follows:

$$c_i = [\overrightarrow{c_i}, \overleftarrow{c_i}] \in \mathbb{R}^{2h} \qquad \forall i \in \{1,\ldots,N\}$$
$$q_j = [\overrightarrow{q_j}, \overleftarrow{q_j}] \in \mathbb{R}^{2h} \qquad \forall j \in \{1,\ldots,M\}$$

3.3 Attention Layer

3.3.1 Basic Attention Flow

The basic attention flow layer applies a dot product between the context hidden states c_i and question hidden states q_j. For context hidden state the attention distribution $\alpha^i \in \mathbb{R}^M$ is computed as follows :

$$e^i = [c_i^T q_1, \ldots, c_i^T q_M] \in \mathbb{R}^M$$
$$\alpha^i = softmax(e^i) \in \mathbb{R}^M$$

The attention output is a linear combination of the attention distribution and the question hidden states for each context hidden state and the output is then given by:

$$a_i = \Sigma_{j=1}^{M} \alpha_j^i q_j \in \mathbb{R}^{2h}$$

The attention outputs are then concatenated to the context hidden state to produce the blended representation b_i:

$$b_i = [c_i; a_i] \in \mathbb{R}^{4h} \quad \forall i \{1,\ldots,N\}$$

3.3.2 BiDirectional Attention Flow

The BiDAF is a high performing SQuAD model.[4] The core part of the model dictates that the attention flows in both the directions. Hence the context attends to the questions as well as the questions attend to the context. The BiDAF model has been described as follows :

$c_1, \ldots, c_N \in \mathbb{R}^{2h}$ are context hidden states.
$q_1, \ldots, q_N \in \mathbb{R}^{2h}$ are question hidden states.

We define a similarity matrix $S \in \mathbb{R}^{N \times M}$ which contains the similarity score S_{ij} for each pair (c_i, q_j) of the context and the hidden states.

$$S_{ij} = w_{sim}^T [c_i; q_j; c_i \odot q_j] \in \mathbb{R}$$

$w_{sim} \in \mathbb{R}^{6h}$ is the weight vector.

Context-to-Question (C2Q) Attention

The attention distribution as described in the basic model is given by :

$$\alpha^i = softmax(S_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1,\ldots,N\}$$

The attention distribution is linearly distributed over the question hidden states to gove the attention outputs as follows :

$$a_i = \Sigma_{j=1}^{M} \alpha_j^i q_j \in \mathbb{R}^{2h} \quad \forall i \in \{1\ldots,N\}$$

Question-to-Context (Q2C) Attention

For each context location $i \in \{1,\ldots,N\}$ we take the maximum of corresponding row of similarity matrix :

$$m_i = max_j S_{ij} \in \mathbb{R} \qquad \forall i \in \{1,\ldots,N\}$$

Softmax over the resulting vector $m \in \mathbb{R}^N$ gives an attention distribution $\beta \in \mathbb{R}^N$ over the context locations

$$\beta = softmax(m) \in \mathbb{R}^N$$

The Q2C output is calculated as the linear combination of teh attention distribution and the context hidden states :

$$c' = \Sigma_{i=1}^{N} \beta_i c_i \in \mathbb{R}^{2h}$$

Now, for each context location $i \in \{1,\ldots,N\}$ we obtain the blended representation of the the output as :

$$b_i = [c_i; a_i; c_i \odot a_i; c_i \odot c'] \in \mathbb{R}^{8h} \qquad \forall i \in \{1,\ldots,N\}$$

Hence the output of this layer we obtain $\mathbb{R}^{8h}$ matrix.

3.4 Output Layer

Each of the blended representations b_i are fed through fully connected layer followed by a ReLU non-linearity:

$$b_i' = ReLU(W_{FC} b_i + v_{FC}) \in \mathbb{R}^h \qquad \forall i \in \{1,\ldots,N\}$$

where $W_{FC} \in \mathbb{R}^{h \times 4h}$ and $v_{FC} \in \mathbb{R}^h$ are weight and bias vector.

A score is then assigned to each context location i as follows :

$$logits_i^{start} = w_{start}^T b_i' + u_{start} \in \mathbb{R} \forall i \in \{1,\ldots,N\}$$

where $w_{start} \in \mathbb{R}^h \;\; and \;\; u_{start} \in \mathbb{R}$ are the weight vector and the bias term respectively.

Finally the softmax is applied to obtain the probability distribution p^{start} over the context locations $\{1,\ldots,N\}$:

$$p^{start} = softmax(logits^{start}) \in \mathbb{R}^N$$

The probability of p^{end} is computed in the same way.

3.5 Loss and Prediction

The loss function is the sum of cross-entropy loss for the start and the end locations. If the predicted start and the end locations are $i_{start} \in \{1, \ldots, N\}$ and $i_{end} \in \{1, \ldots, N\}$ respectively, then loss for a single example is :

$$loss = -\log p^{start}(i_{start}) - \log p^{end}(i_{end})$$

The prediction for simply taken as :

$$l^{start} = argmax_{i=1}^{N} p_i^{start}$$
$$l^{end} = argmax_{i=1}^{N} p_i^{end}$$

the loss is then minimized according to the Adam Optimizer across training average.

4. Experiments

4.1 Dataset

Stanford Question Answering Dataset (SQuAD), Rajpurkar et al. [2016] is a reading comprehension dataset, consisting of of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a span form the corresponding reading passage with a fixed starting and ending word. The question might be unanswerable. The dataset has 100,000+ question- answer pairs taken from 500+ articles.

4.2 Implementation Details

Both the Baseline as well as the improved models have been trained on Tensorflow-1.4 with python 2.7 on a MacBook Air 2015 hardware machine. The training has been done for 2000 training pairs for both the models and the results have been contrasted for the same. The visualizations show an improving trend in the F1/EM score and it has been suggested an ideal iteration training number to be around 15k pairs. The machine that could be used for the purpose is Azure NV6.

4.3 Hyperparameters

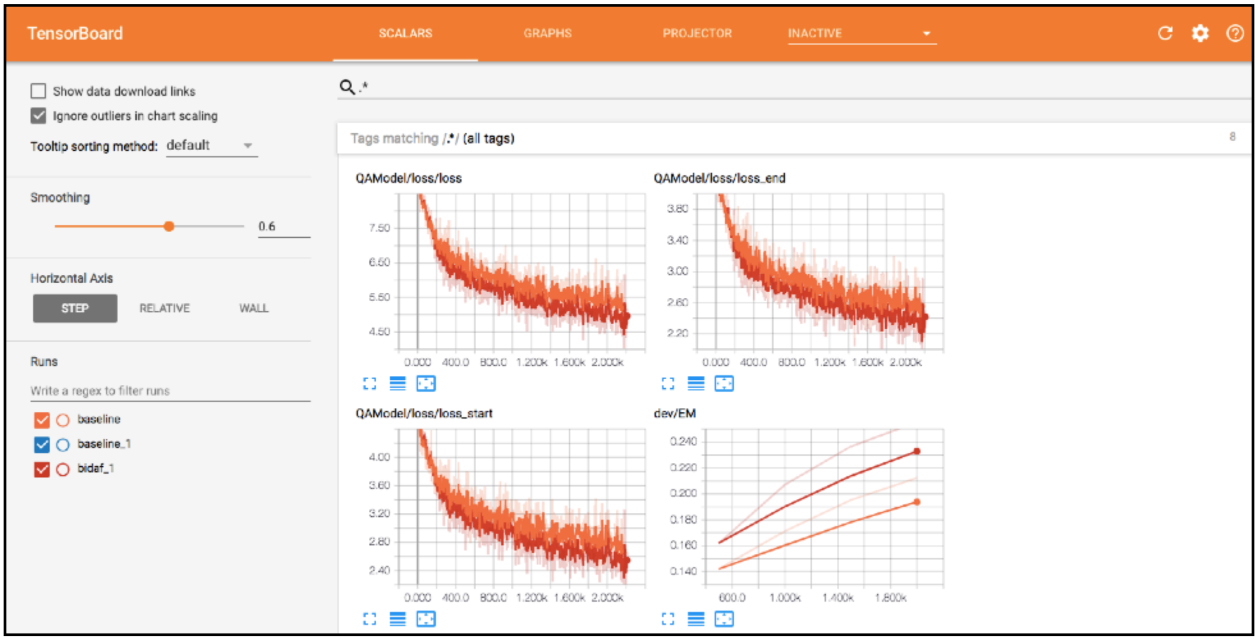
The hyperparameter tuning has been done using the Development set. The Cross Entropy loss was used between the predicted start and the end probabilities of the answer and the true span for training the complete network. The Adam Optimizer has been used with a learning rate of 0.001. A Gradient clipping norm of 5. 0 has been used to avoid the vanishing gradient problem. The dropout rate has been taken to be 0.15 which are the fraction of units randomly dropped on non-recurrent connections.

The batch size of 100, hidden size of 200, context size of 600, question size of 30 and embedding size of 100 have been taken for the training purpose.

4.4 Results

The results for the baseline model and our model have been compared as follows:

Model	F1	EM
Baseline model with simple attention layer	32.84	23.66
BiDAF model (Described model)	**39.73**	**28.67**



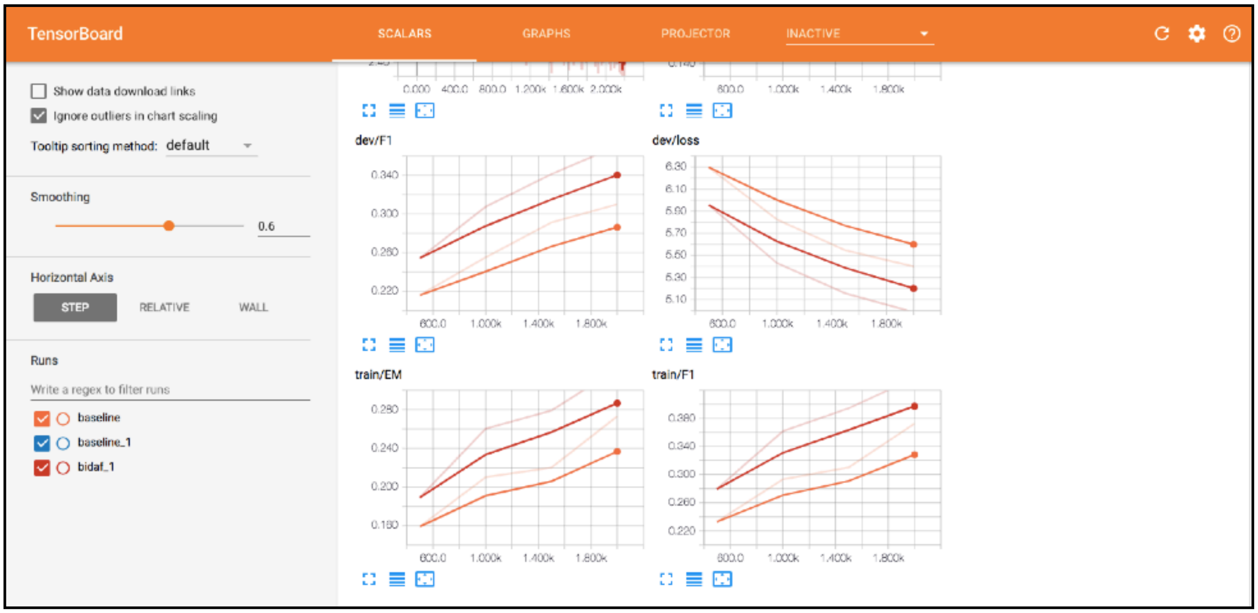
□ Show data download links
☑ Ignore outliers in chart scaling
Tooltip sorting method: default

Smoothing 0.6

Horizontal Axis
STEP RELATIVE WALL

Runs
Write a regex to filter runs
☑ ○ baseline
☑ ○ baseline_1
☑ ○ bidaf_1

.*
Tags matching /.*/ (all tags) 8



TensorBoard
SCALARS
GRAPHS
PROJECTOR
INACTIVE

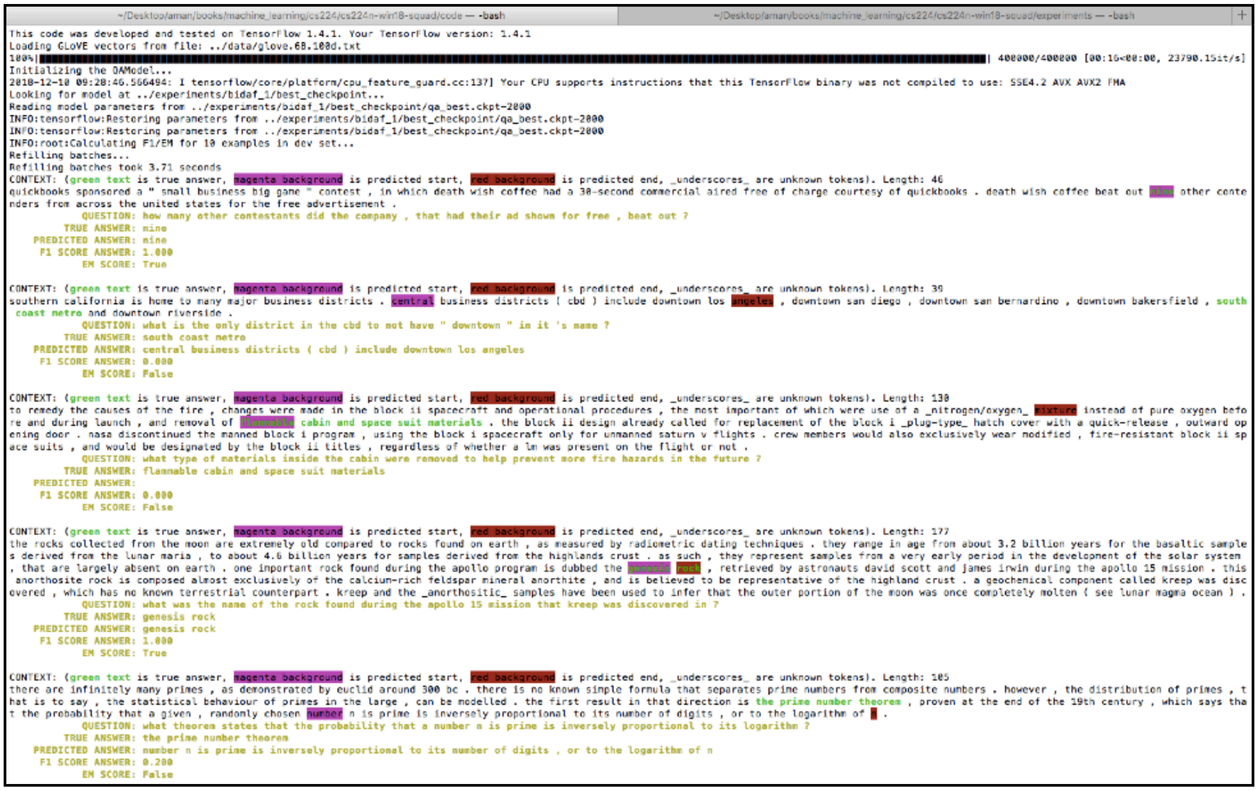
□ Show data download links
☑ Ignore outliers in chart scaling
Tooltip sorting method: default

Smoothing 0.6

Horizontal Axis
STEP RELATIVE WALL

Runs
Write a regex to filter runs
☑ ○ baseline
☑ ○ baseline_1
☑ ○ bidaf_1

0.000 400.0 800.0 1.200k 1.600k 2.000k
600.0 1.000k 1.400k 1.800k
dev/F1
0.340
0.300
0.260
0.220
600.0 1.000k 1.400k 1.800k
dev/loss
6.30
6.10
5.90
5.70
5.50
5.30
5.10
600.0 1.000k 1.400k 1.800k
train/EM
0.280
0.240
0.200
0.160
600.0 1.000k 1.400k 1.800k
train/F1
0.380
0.340
0.300
0.260
0.220
600.0 1.000k 1.400k 1.800k


~/Desktop/aman/books/machine_learning/cs224/cs224n-win18-squad/code — -bash
~/Desktop/aman/books/machine_learning/cs224/cs224n-win18-squad/experiments — -bash

```
This code was developed and tested on TensorFlow 1.4.1. Your TensorFlow version: 1.4.1
Loading GLoVE vectors from file: ../data/glove.6B.100d.txt
100%|████████████████████████████████████████████████████████████████████████████████| 400000/400000 [00:16<00:00, 23790.15it/s]
Initializing the QAModel...
2018-12-10 09:28:46.566494: I tensorflow/core/platform/cpu_feature_guard.cc:137] Your CPU supports instructions that this TensorFlow binary was not compiled to use: SSE4.2 AVX AVX2 FMA
Looking for model at ../experiments/bidaf_1/best_checkpoint...
Reading model parameters from ../experiments/bidaf_1/best_checkpoint/qa_best.ckpt-2000
INFO:tensorflow:Restoring parameters from ../experiments/bidaf_1/best_checkpoint/qa_best.ckpt-2000
INFO:tensorflow:Restoring parameters from ../experiments/bidaf_1/best_checkpoint/qa_best.ckpt-2000
INFO:root:Calculating F1/EM for 10 examples in dev set...
Refilling batches...
Refilling batches took 3.71 seconds
CONTEXT: (green text is true answer, magenta background is predicted start, red background is predicted end, _underscores_ are unknown tokens). Length: 46
quickbooks sponsored a " small business big game " contest , in which death wish coffee had a 30-second commercial aired free of charge courtesy of quickbooks . death wish coffee beat out nine other conte
nders from across the united states for the free advertisement .
          QUESTION: how many other contestants did the company , that had their ad shown for free , beat out ?
       TRUE ANSWER: nine
  PREDICTED ANSWER: nine
   F1 SCORE ANSWER: 1.000
          EM SCORE: True

CONTEXT: (green text is true answer, magenta background is predicted start, red background is predicted end, _underscores_ are unknown tokens). Length: 39
southern california is home to many major business districts . central business districts ( cbd ) include downtown los angeles , downtown san diego , downtown san bernardino , downtown bakersfield , south
 coast metro and downtown riverside .
          QUESTION: what is the only district in the cbd to not have " downtown " in it 's name ?
       TRUE ANSWER: south coast metro
  PREDICTED ANSWER: central business districts ( cbd ) include downtown los angeles
   F1 SCORE ANSWER: 0.000
          EM SCORE: False

CONTEXT: (green text is true answer, magenta background is predicted start, red background is predicted end, _underscores_ are unknown tokens). Length: 130
to remedy the causes of the fire , changes were made in the block ii spacecraft and operational procedures , the most important of which were use of a _nitrogen/oxygen_ mixture instead of pure oxygen befo
re and during launch , and removal of flammable cabin and space suit materials . the block ii design already called for replacement of the block i _plug-type_ hatch cover with a quick-release , outward op
ening door . nasa discontinued the manned block i program , using the block i spacecraft only for unmanned saturn v flights . crew members would also exclusively wear modified , fire-resistant block ii sp
ace suits , and would be designated by the block ii titles , regardless of whether a lm was present on the flight or not .
          QUESTION: what type of materials inside the cabin were removed to help prevent more fire hazards in the future ?
       TRUE ANSWER: flammable cabin and space suit materials
  PREDICTED ANSWER:
   F1 SCORE ANSWER: 0.000
          EM SCORE: False

CONTEXT: (green text is true answer, magenta background is predicted start, red background is predicted end, _underscores_ are unknown tokens). Length: 177
the rocks collected from the moon are extremely old compared to rocks found on earth , as measured by radiometric dating techniques . they range in age from about 3.2 billion years for the basaltic sample
s derived from the lunar maria , to about 4.6 billion years for samples derived from the highlands crust . as such , they represent samples from a very early period in the development of the solar system
, that are largely absent on earth . one important rock found during the apollo program is dubbed the genesis rock , retrieved by astronauts david scott and james irwin during the apollo 15 mission . this
 anorthosite rock is composed almost exclusively of the calcium-rich feldspar mineral anorthite , and is believed to be representative of the highland crust . a geochemical component called kreep was disc
overed , which has no known terrestrial counterpart . kreep and the _anorthositic_ samples have been used to infer that the outer portion of the moon was once completely molten ( see lunar magma ocean ) .
          QUESTION: what was the name of the rock found during the apollo 15 mission that kreep was discovered in ?
       TRUE ANSWER: genesis rock
  PREDICTED ANSWER: genesis rock
   F1 SCORE ANSWER: 1.000
          EM SCORE: True

CONTEXT: (green text is true answer, magenta background is predicted start, red background is predicted end, _underscores_ are unknown tokens). Length: 105
there are infinitely many primes , as demonstrated by euclid around 300 bc . there is no known simple formula that separates prime numbers from composite numbers . however , the distribution of primes , t
hat is to say , the statistical behaviour of primes in the large , can be modelled . the first result in that direction is the prime number theorem , proven at the end of the 19th century , which says tha
t the probability that a given , randomly chosen number n is prime is inversely proportional to its number of digits , or to the logarithm of n .
          QUESTION: what theorem states that the probability that a number n is prime is inversely proportional to its logarithm ?
       TRUE ANSWER: the prime number theorem
  PREDICTED ANSWER: number n is prime is inversely proportional to its number of digits , or to the logarithm of n
   F1 SCORE ANSWER: 0.200
          EM SCORE: False
```

5. Conclusion and Future Work

It has been therefore proved that an improvement in the attention layer of the baseline model has a significant improvement in the model results with the F1 score being 29.73 and the EM score being 28.67.

However there are several other improvements that could further improve the results. Firstly, a character level embedding could be performed at the contextual layer using the Char-CNN model. Moreover the attentional layer could be complemented with self-attention and co-attention models to improve the ensemble average. Lastly, through hyper parameter tuning could be performed to further improve the results.

References

1. Rajpurkar : https://rajpurkar.github.io/SQuAD-explorer/
2. Anand Dhoot, Anchor Gupta : Iterative reasoning with bi-directional flow for machine comprehension
3. Jeffrey Pennington, Richard Socher, Christopher D. Manning GloVe: Global Vectors for Word Representation.
4. Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hananneh Hajishirzi : Bi-Directional Attention Flow for Machine Comprehension.