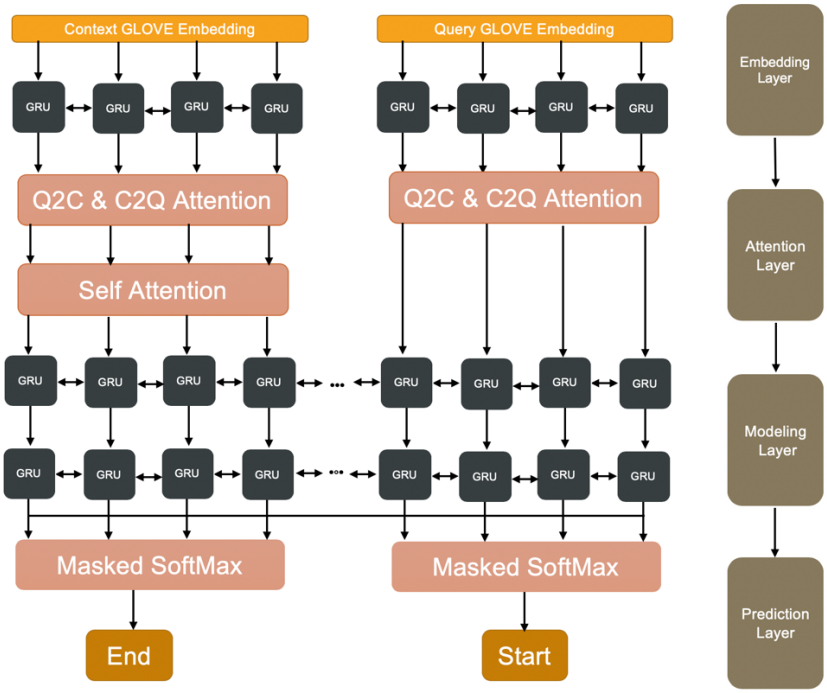
Sepehr Ardebilianfard: model, result analysis, presentation
Adam Lareza: EDA, results analysis, model
LIGN 167 – Final Project

More Attention for BiDAF

Introduction:

Machine question answering is an important evaluation method of language models as it shows how well machines can comprehend human natural language. This importance led to the creation of SQuAD [1] dataset for question and answering oriented evaluations of language models. This project uses SQuAD 2.0 dataset for QA with our language model consisting of a BiDAF [2] architecture as our baseline with added self-attention to improve performance on SQuAD 2.0 dataset. Further improvements were made to the efficiency of the model by switching over to GRU networks and AdaBound optimizer for reduced loss.

Model:



Baseline model was implemented using BiDAF model which is short for Bidirectional Attention flow where attention is calculated using context and question regarding the context. We added a self-attention layer of context on context to further improve performance.

Embedding Layer:

Embedding layer uses pretrained word embedding model Glove[7] to map each word to a vector space for encoding layer. Original BiDAF model uses both character embeddings and word embeddings to calculate attention however, due to time constraints our model only implemented

word embeddings. Word embedding is then passed on to a two-layer Highway network [6] to further refine the embeddings. For word embedding $v_1, \ldots, v_K \in N$ and learnable matrix parameters of W_{Proj} to get $h_i = W_{Proj}\, v_i \in D$ with dimensionality of H. h_i is given to Highway Network for to refine our embeddings.

$$g = \sigma(W_g h_i + b_g)$$

$$t = ReLU(W_t h_i + b_t)$$

$$h_i' = g \circ t\,(1 - g) \circ h_i$$

Where $W_g, W_t \in R^{H \times H}$ and $b_g, b_t \in R^H$ are learnable parameters. g is used for gate and t for transform of each hidden layer. This transformation is done twice with separate learnable parameters.

Encoding Layer:
Encoding layer uses bidirectional LSTM to encode time related attention between each timestep of embedding layer's output. Result is LSTM's concatenated forward and backward hidden states at each time step.

$$h_{i,fwd} = LSTM(h'_{i-1}, hi) \in R^H$$

$$h'_{i,rev} = LSTM(h'_{i+1}, hi) \in R^H$$

$$h_i' = \left[h'_{i,fwd} ; h'_{i,rev} \right] \in 2H^{2H}$$

Attention Layer:
Main part of BiDAF model, is the BiDirection Attention where attention flows both ways from question to context and context to question. For hidden states of context $c_1, \ldots, c_N$ and hidden states of question $q_1, \ldots, q_M$ we compute the similarity matrix s_{ij} for each pair of (c_i, q_j). The similarity matrix uses dropout layers to refine the pair of hidden states and calculate their similarity of c_i and q_j.

$$S_{ij} = w_{sim}^T [c_i : q_j ; c_i \cdot q_j]$$

Where w_{sim}^T is has dimensionality of 6H as a weight vector and $c_i \cdot q_j$ is the elementwise product of question and context.
For Context to Question attention (C2Q), probability distribution of context to question similarity vector as $\bar{S}$ by taking the row-wise SoftMax of S. $\bar{S}$ is used for taking the weighted sums of question hidden states q_j to get C2Q attention as a_i.

$$\bar{S}_{i,:} = softmax(S_{i,:}) \in R^M \ \forall i \in \{1, \ldots, N\}$$

$$a_i = \sum_{j=1}^{M} \bar{S}_{i,j}\, q_j \in R^{2H} \ \forall i \in \{1, \ldots, N\}$$

Now for Question to Context attention (Q2C) we compute the probability distribution $\bar{\bar{S}}$ by calculating column-wise SoftMax of S to obtain attention distribution of question over context

per column. After multiplying $\bar{S}$ by $\bar{\bar{S}}$ to get row wise distribution of (Q2C) question on context as S'. We take the weight sums of context hidden states c_j to get Q2C attention as b_i.

$$\bar{\bar{S}}_{:,j} = softmax(\bar{S}_{:,j}) \in R \ \forall j \in \{1, \dots, M\}$$

$$S' = \bar{\bar{S}}\bar{S} \, R^{NxN}$$

$$b_i = \sum_{j=1}^{N} S'_{i,j} \, C_j \in R^{2H} \ \forall i \in \{1, \dots, N\}$$

To create the BiDirectional attention g_i for each location $i \in \{1, \dots, N\}$ in context, we combine context hidden state with C2Q attention a_i and Q2C attention b_i.

$$g_i = [ci; ai; ci \cdot ai; ci \cdot bi] \in R^{8H} \ \forall i \in \{1, \dots, N\}$$

Self Attention:
Self attention is used to further refine the context on its question awareness as g_i has limited idea of what cues context provides to question. To solve this problem self attention is created to relate evidence from the whole passage with respect to current context word and the question. Self attention was implemented by going through linear transformation to change dimensions, none-linear activation functions, and dropout layers, and GRU encoder to compute attention scores which are then used to find the similarity score using Trilinear attention between context and context. Where W_a and W_b are weight matrices and v is weight vector of encoded context.

$$e_j^i = v^T \tanh (W_a v_j + W_b v_i)$$

$$a_i' = softmax(e^i)$$

$$a_i^t = \sum_{j=1}^{N} a'^i_j v_j$$

$$h_i = GRU(v_i; a_i) \ \forall i \in \{1, \dots, N\}$$

Modeling Layer:
After self attention was combined with our BiDirectional attention, we pass self attention's output to a two-layer LSTM encoder where temporal relationship of combined attention is captured as it is further refined.

Prediction Layer:
Finally, using our encoded self attention outputs $m_1, \dots, m_N$ as M and attention outputs $g_1, \dots, g_N$ as G where we apply Bidirectional LSTM to m_i to get m_i'. The result is used to produce distribution probability of p_{start} and p_{end} using G and M.

$$p_{start} = softmax(W_{start}[G; M])$$

$$p_{end} = softmax(W_{end}[G; M'])$$

Loss:

Loss is computed by evaluating the model by calculating the cross-entropy or negative likelihood for start and end location for start $i \in \{1, \ldots, N\}$ and end with $j \in \{1, \ldots, N\}$.

$$loss = -\log p_{start}(i) - \log p_{end}(j)$$

Evaluation:

$$EM = N_{exact\_match}/N_{total\_predictions}$$

$$precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

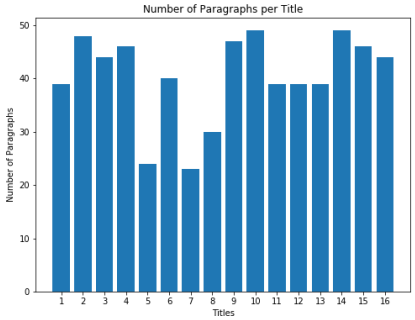
Dataset:

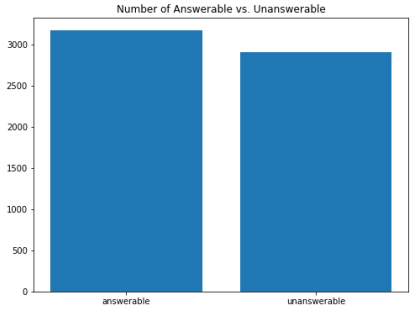
SQuAD 2.0 was created based of the original SQuAD data set which is a Question Answering dataset created by Stanford University and extracted from Wikipedia articles. This dataset consists of 150,000 questions with 50,000 questions which are unanswerable. With every question there's is a segment of text provided which is used as the context to every question. The improvement over the original SQuAD dataset is the unanswerable questions as it adds another layer of evaluation to for the model showing if model is capable of understanding that the context does not provide enough information to answer the question. Furthermore, Squad dataset was used and Glove Pretrained embedding network was used to do contextual embedding. The code for preprocessing the data and embedding using Glove was used from Stanford [8] class. The dataset is divided into train and dev each including 129,941 and 6078 examples respectively.

Example:

- **Question:** What general religious belief did the nations that received Huguenot refugees have in common?
- **Context:** The bulk of Huguenot émigrés relocated to Protestant European nations such as England, Wales, Scotland, Denmark, Sweden, Switzerland, the Dutch Republic, the Electorate of Brandenburg and Electorate of the Palatinate in the Holy Roman Empire, the Duchy of Prussia, the Channel Islands, and Ireland. They also spread beyond Europe to the Dutch Cape Colony in South Africa, the Dutch East Indies, the Caribbean, and several of the English colonies of North America, and Quebec, where they were accepted and allowed to worship freely.
- **Answer:** Protestant
- **Prediction:** Protestant

- **Question:** Who was Iqbal a critic of?
- **Context:** While studying law and philosophy in England and Germany, Iqbal became a member of the London branch of the All India Muslim League. He came back to Lahore in 1908. While dividing his time between law practice and philosophical poetry, Iqbal had remained active in the Muslim League. He did not support Indian involvement in World War I and remained in close touch with Muslim political leaders such as Muhammad Ali Johar and Muhammad Ali Jinnah. He was a critic of the mainstream Indian nationalist and secularist Indian National Congress. Iqbal's seven English lectures were published by Oxford University press in 1934 in a book titled The Reconstruction of Religious Thought in Islam. These lectures dwell on the role of Islam as a religion as well as a political and legal philosophy in the modern age.
- **Answer:** the mainstream Indian nationalist and secularist Indian National Congress
- **Prediction:** mainstream Indian nationalist and secularist Indian National Congress

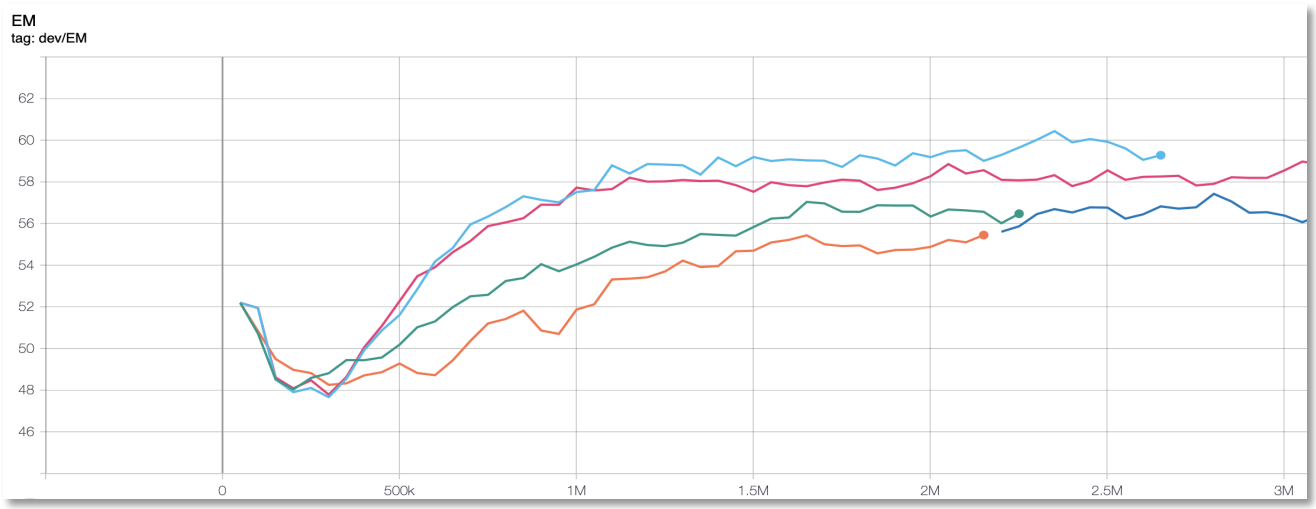


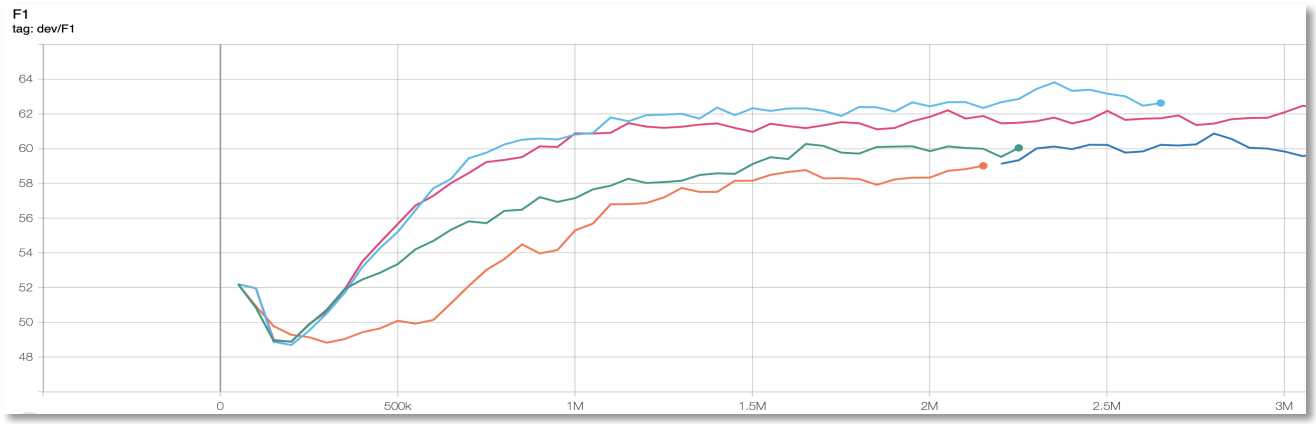


Results:

To get a better understanding of our results we experimented with multiple Optimizers as well as learning schedules to get the best results against our baseline trained model. Baseline model used an AdaDelta optimizer and we experimented with the newly discovered optimizer AdaBound, however the result showed that more time was needed for further finetuning the model and due to the lack of time the performance over the use of AdaDelta was not substantial. However, adding a learning rate scheduler which decreased the learning rate by .05 every 5 epochs gave us much better performance against the baseline with self attention added to baseline architecture.

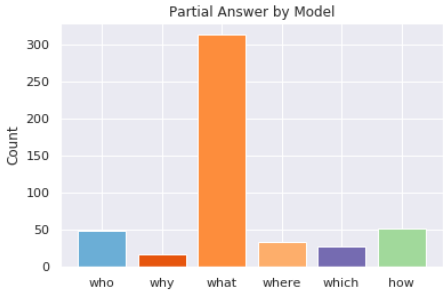
Model	F1	EM
Baseline + Self Attention + StepLr	63.83	60.44
Baseline	61.4	58.01
Baseline + Self Attention + AdaBound	60.88	57.44
Baseline + Self Attention	60.78	57.55

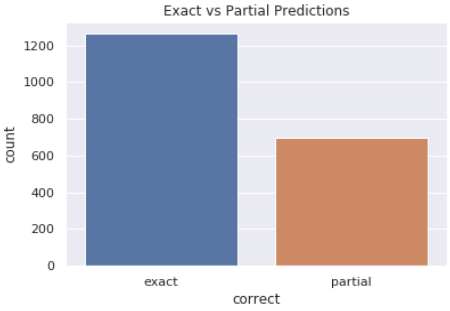


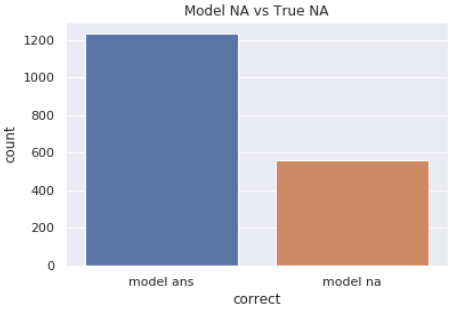


Analysis:

An interesting analysis on the data set was done using our models results where questions were divided based on the subject in question which were “who”, “why”, “what”, “which”, and “how”. The results showed that model had good performance answering questions which used “What” for the query showing its performance when a question is specific. However, when question asked “why” the model had a really hard time predicting the answer as “why” questions require deep understanding of context and perhaps could be the difference maker between distinguishing machines vs humans. Based on the metrics to evaluate the model, it the graph shows how the model performed predicting answers which exactly matched the correct answer vs partially answering the question which means not all of the answer was included in the prediction by the model.







Conclusion:

BiDAF model with self attention showed the improvements over baseline, however over the past few years with the introduction of transformers such as Bert and Elmo which has a similar concept to self attention and attention flow for context and question have significantly improved QA model’s performance. This has caused non-transformer based model improvements to be significantly lower than transformer based models. However, implementing this model allowed us to understand how a data set such as SQuAD are vital to evaluate language models with distinct metrics. Furthermore understanding how RNN’s work and the performance gains of LSTM’s and efficiency of GRU networks over traditional RNNs was very important.

References:

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016.

[2] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.

[3] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[4] A. Vaswani & N. Shazeer & N. Parmar & J. Uszkoreit & L. Jones & A. N. Gomez & L. Kaiser & I. Polosukhin. 2017 Attention is all you need. In Neural Information Processing Systems (NIPS)

[5] Mingchen Li, Gendong Zhang, Zixuan Zhou. 2018 Improved BiDAF with Self-Attention

[6] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[7] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[8] Stanford CS224n: Natural Language Processing with Deep Learning