Contextual Question Answering with Improved Embedding Models

George He
Department of Computer Science
Stanford University
georgehe@stanford.edu

Abstract

Recent advancements in natural language processing have lead to tremendous improvements in short-dialog question answering, where a context and information from previous conversations if ignored and . While datasets such as SQUAD 2.0 (Rajpurkar et al., 2018) have addressed scenarios consisting of independent questions on a known passage, where the answer to every question is either unanswerable or a segment from a corresponding reading passage.

In this paper, we evaluate the performance of various models on the QUaC (Choi et al., 2018) dataset and challenge, which considers the issue where individuals do not know the answers to their questions prior to asking them, and subsequently lessens the role of string matching.

1 Introduction

The Question Answering in Context (QuAC) is a dataset for modeling, understanding, and participating in information seeking dialog. When the challenge was first introduced, the baseline BiDAF model with GloVe embeddings(Pennington et al., 2014) offer many opportunities for improvement.

Simple changes to word representations through improved contextual word embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) offer straightforward opportunities to improve the performance of the existing BiDAF model, and have been shown to improve state of the art performance on a gamut of natural language process tasks, include SQUAD 2.0 and the CoQA (Reddy et al., 2018) challenge, which measures the ability of machines to understand a text passage and answer a series of interconnected questions that appear in a conversation.

CoQA and QuAC differ primarily in information made avaialable to the questioner. In CoQA, the contextual dialogues are driven by questions asked by an individual with access to the evidence passage that the machine model also has access to. In QuAC, the contextual dialogues are driven by questions asked by an individual without access to the evidence passages, resulting in often more open-ended and unanswerable questions.

Due to the above observations, we hypothesize that application of BERT and ELMo embeddings in the BiDAF model will improve representation, and modification of the architecture to support attention models such as FlowQA (Huang et al., 2018) which improve scores. Combined, the two models should achieve near-SOTA for the QUAC model.

2 Previous Work

2.1 GloVe: Global Vectors for Word Representation(Pennington et al., 2014)

Realizing a need to create a model that learns information in a more global context, GloVe(Pennington et al., 2014) was developed to create word vectors that capture meaning in vector space while also taking advantage of global count statistics instead of only local information. GloVe trained on aggregate training data based on a co-occurrence matrix of certain word, and trained word vectors so their differences predict co-occurrence ratios GloVe weights the loss based on word frequency.

This model further improved the state of the art from the existing word2vec model through incorporating a more global context during the training routine, rather than the limited local context employed during the word2vec training process.

Figure 1: Co-occurance Matrix Example (Pennington et al., 2014)

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

2.2 Deep contextualized word representations (ELMo)(Peters et al., 2018)

Embeddings from Language Models (ELMo)(Peters et al., 2018) representations are a context dependent word embedding, where the embedding for each word in the input sentence depends on the context. ELMo follows a stacked bi-directional LSTM model with a novel final output layer computed as a linear combination of all the stacked LSTM layers. This final layer differs from other previously existing works at the time of its publishing because previous works only incorporated the output of the final LSTM layer in its final output layer, which resulted in information loss. The intuition is that lower level LSTM layers are able to detect lower level language constructs such as part of speech, while the higher level LSTM layers are able to determine higher level constructs.

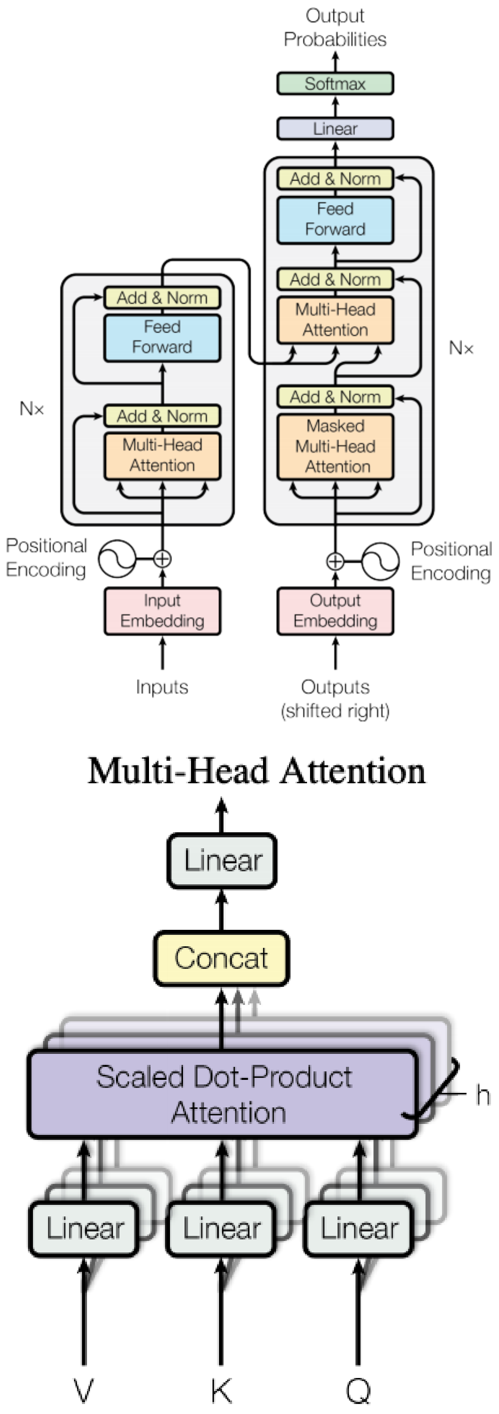
2.3 Attention Is All You Need (Vaswani et al., 2017)

As encoder-decoder (Sutskever et al., 2014) introduced to transform input sequences to new sequences of arbitrary lengths for tasks such as question answering and sentence completion. These models quickly ran into issues of runtime, due to the sequential nature of recurrent models, as well as degraded performance with deeper recurrences.

Positional encoding and multihead attention, introduced in this paper, provided a mechanism for sequence transduction models to overcome some of the issues with computation accuracy and time associated with sequence length due to the use of recurrent architectures (LSTMs/GRUs). This is done through using a attention-mechanism that looks at an input sequence and decides at each step which sections of the input sequence are important.

These transformer models have been successfully applied in BERT(?), which is based on a multi-layer bidirectional Transformer(Ashish Vaswani, 2017). BERT has been trained on plain text for masked word prediction and next sentence prediction tasks, and has been applied to different natural language understanding tasks through fine-tuning the final output layer for each task after initiating weights using a pre-trained BERT model.

Figure 2: Attention Model (Vaswani et al., 2017)



2.4 BiDAF (Seo et al., 2016)

The BiDAF model is created with a number of sequential CNN and RNN layers, followed by attention flow and transformation layers as shown in diagram 3.

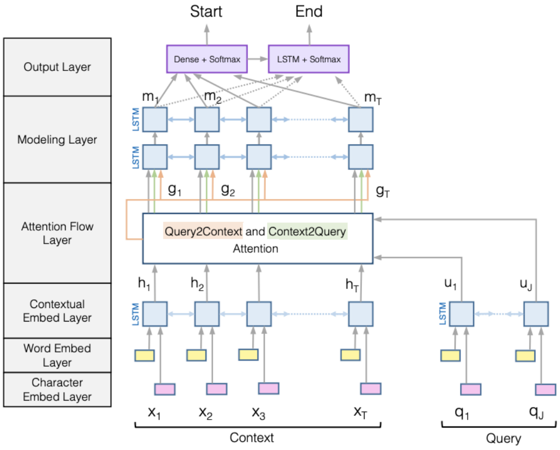


Figure 3: BiDaF(Seo et al., 2016) architecture

The first layers consist of a character embedding layer, which maps each word to a vector space using character-level CNNs.

The character embedding layer is then followed by a word embedding layer, which maps each word to a vector through using a pre-trained word embedding model. In practice, this can be a representation from GloVe, Bert, ELMo, or any pre-trained embedding. The character embedding and word embedding layers are then concatenated and fed into the following layer.

Next, a contextual embedding layer comprised of a bidirectional RNN (for this paper, we will use a GRU), which allows flow of the context from the nearby words to improve the embedded model.

The next critical layer is the attention flow layer, which applies an attention transformation on the query and context vectors output from the contextual embedding layer.

The flow of attention is bidirectional, which allows for us to create a context to query and a query to context representation of the attention flow.

The contextual embedding layer and both produced attention layers are combined through a predefined layer (often a perception) to produce a context aware feature vector for each word.

A modeling layer then uses a series of stacked bidirectional RNN layers to scan the context and feed the context into an output layer, which can be reshaped to be task specific.

2.5 SQuAD 1.0 (Rajpurkar et al., 2016) and SQuAD 2.0(Rajpurkar et al., 2018)

In the SQuAD 1.0 challenge, Rajpurkar et al presented a reading comprehension dataset consisting of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question was a segment of text (span) from the corresponding reading passage. This challenge, however, lacked the complexity that would mirror real world scenarios. Models such as BERT were able to surpass human performance in terms of both exact match and F1 metrics on this dataset, with an emsembled BERT model achieving a 87.433 EM score, and a 93.160 F1 score compared to human performance with a 82.304 EM score and a 91.221 F1 score.

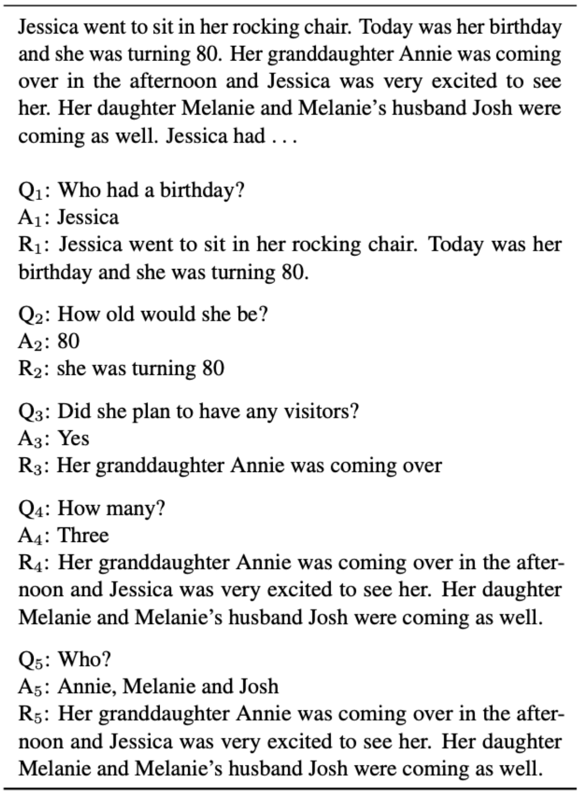
Subsequently, the SQuAD 2.0 challenge expanded the coverage and depth of the SquAD 2.0 model by combining the existing SQuAD reading comprehension data with over 50,000 unanswerable questions. While a more complex challenge, ensembled BERT models were able to quickly approach/exceed human performance as well.

2.6 CoQa (Reddy et al., 2018)

The Conversational Question Answering(CoQa) challenge was developed to address the contextual. In the CoQa challenge, the conversational questions and answers are free-form text with their corresponding evidence highlighted in the passage in figure 4.

2.7 QuAC (Choi et al., 2018)

QuAC introduces a dataset collected in an interactive setting where two crowd workers play the roles of teacher and student. The student asks the teacher questions related to a passage that only the teacher has access to. Students only had access to the passage heading. In this case, wikipedia passages were used as reference texts for the teacher. This type of dataset is not new, but the novelty of the QuAC dataset is in the new scenarios that it introduces. The student does not know the reference task before asking them. Student questions can be open ended and unanswerable by the text, or dependent on the previous questions. As mentioned by the paper, this results in a dataset that is more representative of real user queries. As structured in the dataset, the dialogues continue until either 12 questions are answered, one of the individuals decides to end the interaction, or more than two



Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

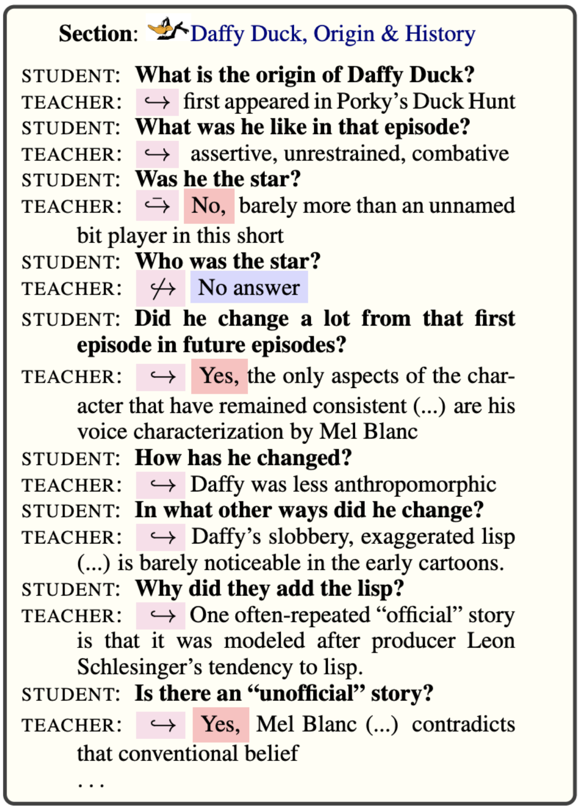
Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 4: CoQa(Reddy et al., 2018) example

Figure 5: QuAC Dialog Example (Choi et al., 2018)



Section: Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
. . .

unanswered questions are asked.

In this paper, authors experimented with different architectures, wherein BiDaF with no context was able to achieve an F1 score of 50.2 and BiDaF with additional (two) contexts was able to achieve F1 scores of 60.6. Currently, the top model achieves an F1 score of 67.8, while human performance has a F1 score of 81.1 - indicating room for improvement.

2.8 FlowQA (Huang et al., 2018)

FlowQA introduces a novel FLOW mechanism that encodes the conversation history. The Flow mechanisms feed the model with the entire hidden representations generated during the process of answering previous questions.

These hidden representations capture related information, such as phrases and facts in the context at each stage and are propagated to different stages of the question answering interaction through an integration-flow layer.

3 Technical Overview

3.1 Data

For this dataset, we will be utilizing the publicly provided QuAC dataset consisting of over 100k question dialogues. The dialogues are created by pairs of crowd-sourced workers consisting of a student who asks questions and a teach who has access to a passage, and provides answers. By doing so, QuAC introduces question formulation methods not found in existing datasets, resulting in usually more open ended and longer answers.

For the validation dataset, as opposed to the training dataset, four additional annotations per question were collected by reusing past dialog questions - resulting in higher f1 scores on the validation than on the training dataset since the F1 score is taken as the maximum of all reference texts.

As a point of reference, the measured human performance was approximately 74.7 F1. We will only be comparing results against performance on the validation and test datasets.

3.2 Metrics

As each model learns, we will be using negative log likelihood loss as a metric for loss. The negative log likelihood loss functions is used to compute the loss when a model outputs a probability for each class. It utilizes a softmax function to compute the probability and can be computed as shown in figure 6.

$$\mathrm{loss}(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right) = -x[class] + \log\left(\sum_j \exp(x[j])\right)$$

Figure 6: NLL Loss

However we will be comparing a number of different metrics for this dataset, including both qualitative and quantitative reviews not completely aligned with the NLL loss. F1 score (a measure of accuracy comprised of the precision and recall) will serve as a basis for computing quantitative performance of different models. We will also be evaluating convergence properties of different models over time by comparing their validation F1 scores.

For quantitative evaluations, we will compare performance of different models on difficult dialog settings, and reviewing characteristics and potential flaws in the responses of each model.

3.3 General Reasoning

While each of the prior models have introduced novel improvements on top of the QuAC challenge, it seems that BERT and ELMo embeddings have provided improvements to architectures relying on word embeddings to encode input word representations.

The contextual flow presented in the FlowQA model also begins to incorporate non-recurrent methods of transferring contextual knowledge from prior dialog, and offer better representations of the conversation than a normal attention/recurrent information transfer mechanism.

Combining the two, we hope to see much better improvements in the scoring function.

4 Experiments

As mentioned above, please note that for the validation dataset, as opposed to the training dataset, four additional annotations per question were collected by reusing past dialog questions - resulting in higher f1 scores on the validation than on the training dataset since the F1 score is taken as the maximum of all reference texts.

We experiment with varying embedding schemas, comparing the BiDAF++ model using the baseline word embedding representation from GloVe. We build upon the AllenNLP implementation of the bidaf dialog QA model as noted in the original QuAC paper(Choi et al., 2018), utilizing the BiDAF++ dialog question answering model with BERT and ELMo embeddings to observe improvements in performance resulting from contextualized representations.

4.1 BiDAF Models

We utilize the BiDAF(Seo et al., 2016) dialog question answering model published and open sources by AllenNLP as a baseline for our experiments.

During the training process, we used GRUs in the RNN encoder layers, with a stochastic gradient descent based optimization scheme with learning rate of 0.01 and a momentum of 0.9. These values were chosen based off of findings in the QUaC paper.

In the GloVe word embedding based BiDAF++ model, we utilize pretrained GloVe(Pennington et al., 2014) embeddings with 300 output dimensions (840B.300d). These embeddings have been prertrained on a common crawl of 840 billion tokens.

In the ELMo word embedding based BiDaF++ model, we utilize pretrained embedding model and weights

`ELMo_2x4096_512_2048cnn_2xhighway`

, resulting in 1024 output dimensions.

In the BERT model embedding based BiDaF++, we utilize the pretrained bert-large-cased model, resulting in 1024 vectorized units in the output dimensions.

Between the three models, we compare performance on the training and validation datasets in terms of F1 score, NLL loss and dialog specific criteria. Based off the publically available information about the dataset, we will compare the performance on model in the following areas:

- yes/no/not a binary answer
- predicting if there is a follow up questions or not
- span (start, end, exact match performance)

4.2 FlowQA Model

In the FlowQA model, we compare results relative to the BiDAF++ model. In the FlowQA model, we will only compare base performance using ELMo embeddings to observe how modifications of architecture, rather than embedded representations of the questions, offer improvements in performance.

5 Results

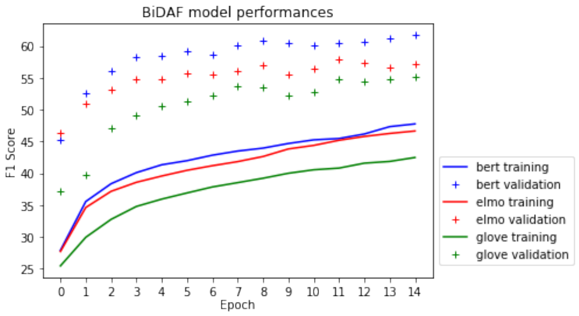
5.1 BiDAF Models

Training consisted of 15 epochs using a stochastic gradient descent optimizer with learning rate of 0.01 and momentum of 0.9, on a reduce on plateau learning rate scheduler. Specific configuration can be found in the associated project code.

Models trained using GloVe embeddings has performed on par with the reporting F1 score in the original QuAC paper - with a validation f1 score of approximately 54, whereas the ELMo model has a validation f1 score of approximately 57 and the BERT model has a validation f1 score of approximately 61.8.

During training, each epoch took approximately 2 hours to train on a V100 GPU, totalling roughly 90 hours of training for the three models. We use the best observed training and validation values to compare performance of the individual models in table 4.

Figure 7: Training BiDAF++ f1 scores



Model	Train F1	Val F1
BIDAF+GloVe	42.2	55.64
BIDAF+ELMo	46.6	57.9
BIDAF+BERT	49.2	61.8

Table 1: BiDAF++ Model Performance

It can be seen that BERT embeddings have led to the greatest improvements in the BiDAF++ dialog question answering model, with the BERT based BiDAF++ model achieving a validation F1 score of 61.8 compared to the 55.64 observed using GloVe.

To get a better understanding of how the contextualized embeddings are able to better convey information, we compare performance of the models through conducting detailed analysis of the different models applied in different contexts, as shown in table 2 and table 3.

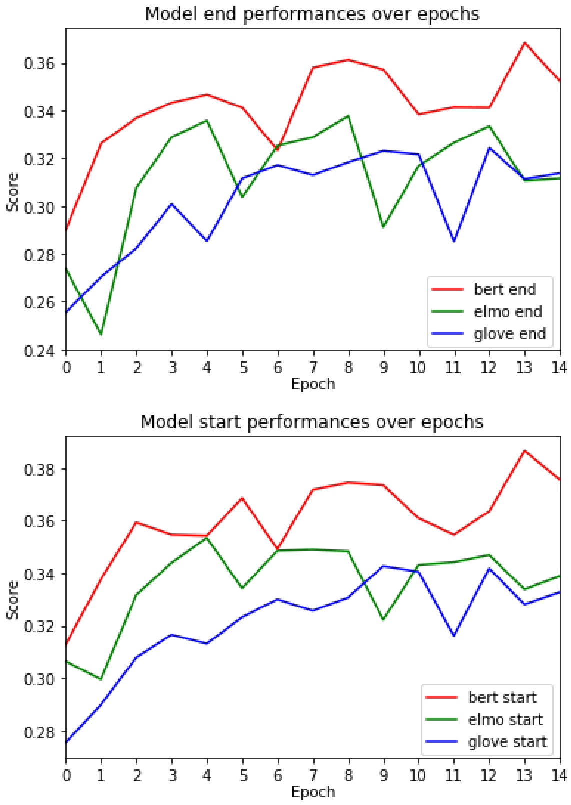
Embedding	yesno	followup	span
GloVe	0.852	0.608	0.263
ELMo	0.864	0.610	0.273
BERT	0.868	0.620	0.289

Table 2: BiDAF++ Model Performance

Embedding	end span	start span
GloVe	0.343	0.327
ELMo	0.353	0.338
BERT	0.386	0.368

Table 3: BiDAF++ Model Span Performance

Figure 8: Span Performance



Careful comparison of the GloVe, ELMo, and BERT embeddings applied in the BiDaF++ question answering model reveals that BERT embeddings show the greatest improvements in the area of span detection, when compared to the other two models. Notably, while ELMo and GloVe embedding based models have comparable span start and end performance, BERT excels at detecting correct associations between the question and reference texts.

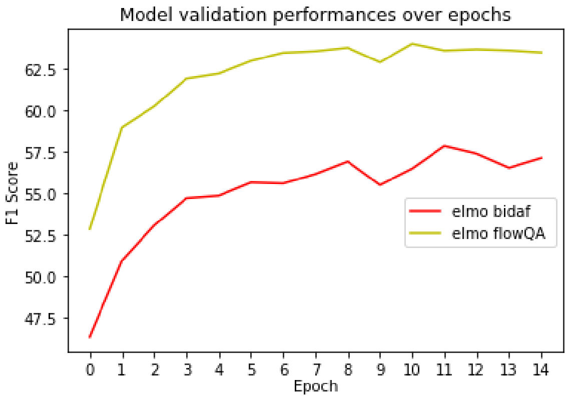
5.2 FlowQA Model

We observe that the FlowQA model outperforms its BiDAF counterparts on the validation dataset. During the training process, we used the open source implementation of FlowQA as noted by Huang et al.(Huang et al., 2018). As expected, the introduction of flow allows for context to be better communicated, and leads to an improved model capable of learning many more representations than

Model	Val F1
BIDAF+ELMo	57.9
FlowQA+ELMo	64.0

Table 4: ELMo Embedding Model Performance

Figure 9: FlowQA vs BiDaF++ Comparison



6 Conclusions and Future Work

Through the experimental results, we have highlighted unique areas in which improved contextualized embeddings offer the most improvement, in context of the QuAC dataset.

We see that embedding quality and contextual representations offer significant improvements in quantitative scores of the models, but recent improvements in contextual representations still leave much to be desired (as observed in the difference between state of the art models and human performance in the QuAC dataset.

Future work in expanding the flow model to expose dense flow representations to mimic models such as Densenet(Huang et al., 2016), wherein flow is deeply and directly connected between dialogue contexts, as well through applications of transformers rather than long-range RNN layers can bring a more global and contextualized view of dialogue into question answer models to address the QuAC challenge.

7 Acknowledgements

We would like to thank the cs224u teaching staff and project mentor Jayadev Bhaskaran for providing resources and academic references during the development of this project.

8 Authorship Statement

George conceived of the idea to compare embedding and contextualized performance, conducted research into existing models, and performed experiments building upon identified open source models. In addition, George analyzed the experimental results and wrote the project paper.

References

Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. 2017. Attention is all you need. https://arxiv.org/abs/1706.03762.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac : Question answering in context. *CoRR*, abs/1808.07036.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2016. Densely connected convolutional networks. *CoRR*, abs/1608.06993.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. Flowqa: Grasping flow in history for conversational machine comprehension. *CoRR*, abs/1810.06683.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

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

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *CoRR*, abs/1808.07042.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

9 Supplementary Materials

For access to project code and other relevant files, please contact the paper author.