Understanding Multiple Question-Answering Models and Its Generalizability

Liming Jiang
University of Massachusetts Lowell
liming_jiang@student.uml.edu

Ariel Pena-Martinez
University of Massachusetts Lowell
ariel_penamartinez@student.uml.edu

Abstract

Context understanding and the extraction of information from it are very important NLP tasks, but very complex at the same time. In this paper, we aimed to test the new QuAIL dataset [8] with our baseline model and from that, we started building our models while concatenating QuAIL and RACE [9] as a means of data augmentation in order to gain better accuracy in tasks like reading comprehension. We implemented a Bi-Directional attention flow (BiDAF) [10] baseline model. And proposed a self-attention layer on top of it. The results show the data augmentation can help with the accuracy in the validation set. Surprisingly, it turns out BiDAF is not a good model for multiple-choice question answering tasks although it was the state-of-the-art model for question answering tasks without given choices. Our model makes slightly better improvements on the baseline performance, and data augmentation on the RACE dataset also slightly improves accuracy score.

1 Introduction

Even though NLP is making a lot of progress, there are still barriers preventing the complete understanding of human language. Reading comprehension, one of the most challenging NLP tasks, is the focus of this project. The topic diversity and the focus on specific types of reasoning of existing datasets is one of the main factors that reading comprehension is struggling with. QuAIL [8] was created to balance these types of reasoning and rule them all while handling a certain amount of unanswered questions as well. We aim to train a model and increase the accuracy based on baseline models previously tested with QuAIL. The trained model takes the input of a paragraph or sentence, a question related to that paragraph, and corresponding multiple choices with only one correct answer within. The output would be the index of the correct answer to that question.

For this task, we implemented a baseline model BiDAF [10]. And proposed a self-attention layer on top of it in order to gain some improvements. To our best knowledge, nobody has ever tried BiDAF on QuAIL before, and in this study, we attempt to find out if the BiDAF and its variation (our proposal) can achieve decent accuracy on QuAIL.

2 Related work

After game-changer BERT was proposed, it became a new state-of-the-art and nearly conquered every single task in NLP. Models afterwards try to make changes on BERT and surpass its performance.

Zhang et al. proposed a retrospective reader (Retro-Reader) to better address unanswerable questions. This model integrates two stages of reading and verification strategies: 1) sketchy reading that briefly investigates the overall interactions of passage and question and yield an initial judgment; 2) intensive reading that verifies the answer and gives the final prediction. The

model is tested on SQuAD and NewsQA and beat the baseline ALBERT significantly [2].

The joint team of Google and TTIC proposed a light version of BERT which can scale much better and established a new state-of-the-art result on SQuAD benchmarks [3].

To incorporate explicit syntactic constraints, Zhang et al. proposed a syntax-guided network (SG-Net) for better linguistic inspired word representation. The proposed SG-Net helps achieve substantial improvement compared to strong baselines [4].

Liu et al. studied the hyperparameter configuration of BERT. It turns out BERT was significantly undertrained. With various design choices, their best well-tuned model can achieve state-of-the-art results on three well-known datasets [5].

To overcome the disadvantage of BERT, Yang et al. proposed a generalized autoregressive pretraining method (XLNet) which could also learn from bi-directional context. XLNet outperforms BERT under 20 tasks, in which question answering dataset like SQuAD is included [6].

Zhang's team proposed a semantic-aware version BERT for better language representation. They consider incorporating structured semantic-information and their model can easily absorb contextual semantics. They achieve state-of-the-are results on ten reading comprehension datasets [7].

BERT is the current state-of-the-art model implemented on the QuAIL dataset according to its leaderboard [1] with 59.8 overall accuracy. On the other hand, Megatron-BERT (ensemble)[11] and ALBERT + DUMA (ensemble)[12] are the models leading RACE leaderboard[2] with 90.9 and 89.8 respectively.

3 Methodology

3.1 Problem formulation

The task is given many paragraphs. Each paragraph (or context) c_i is paired with several questions $\{q_{ij}\}$ which are given four candidate choices $\{a_{ij1},\ a_{ij2},\ a_{ij3},\ a_{ij4}\}$. The goal is to pick the right choice for each question q_{ij}. Essentially this is a supervised learning task, and ground truth labels for each question are given beforehand.

3.2 Baseline model

The baseline model is a simplified version of the Bi-Directional Attention Flow (BiDAF) model [10]. Contrary to the original model, the baseline model only considers word-level embeddings for the inputs, it does not include a character-level embedding layer as it is in Figure 1.

In brief, the model is composed of the following layers:

- Embedding layer (layers.Embedding): inputs are embedded using GloVe pre-trained word vectors, then projected and passed through a two-layer Highway network [13]. Given an input vector $h_i \in R^H$, a one-way highway network computes:

$$g = \sigma(W_g h_i + b_g) \in R^H$$
$$t = ReLU(W_t h_i + b_t) \in R^H$$
$$h_i' = g \odot t + (1-g) \odot h \in R^H$$

  We use a two-layer highway network to transform each hidden vector h_i, which means the model applies the above transformation twice.
- Encoder layer (layers.RNNEncoder): output of the embedding layer is passed through a bidirectional LSTM to allow the model to incorporate temporal dependencies between time steps of the embedding layers output. The encoded output is the RNNs hidden state at each position:

$$h_i', fwd = LSTM(h_{i-1}', h_i) \in R^H$$
$$h_i', rev = LSTM(h_{i+1}', h_i) \in R^H$$
$$h_i' = [h_i', fwd; h_i', rev] \in R^{2H}$$

[1] http://text-machine.cs.uml.edu/lab2/projects/quail/

[2] http://www.qizhexie.com/data/RACE_leaderboard.html

Note: h_i' is of dimension $2H$, as it is the concatenation of forward and backward hidden states at timestep i.

- Attention layer (layers.BiDAFAttention): The core part of the BiDAF model is the BiDirectional Attention Flow layer. The main idea is that attention should flow both ways from the context to the question and from the question to the context. First it computes the similarity matrix $S \in R^{NxM}$, which contains a similarity score S_{ij} for each pair (c_i, q_j) of context (c_i) and question (q_j) hidden states. Next it performs Context-to-Question (C2Q) Attention and Question-to-Context(Q2C) Attention. Lastly, for each context location $i \in 1, \ldots\ldots, N$ it obtains the output g_i of the Bidirectional Attention Flow Layer by combining the context hidden state c_i, the C2Q attention output a_i, and the Q2C attention output b_i

$$g_i = [c_i;\ a_i;\ c_i \cdot a_i;\ c_i \cdot b_i] \in \{1, \ldots\ldots, N\}$$

where $\cdot$ represents elementwise multiplication.

- Modeling layer (layers.RNNEncoder): The modeling layer is tasked with refining the sequence of vectors after the attention layer. It integrates the temporal information between context representations conditioned on the question. Similar to the Encoder layer, it uses a two-layer bidirectional LSTM and the last state of hidden state is adopted as layer output.
- Output layer (layers.BiDAFOutput): given the last hidden state from the previous layer, the output layer uses a linear layer to project the hidden size to the number of choices (which is four). It is then projected and passed through a SoftMax to get the probability of each choice being the correct answer.

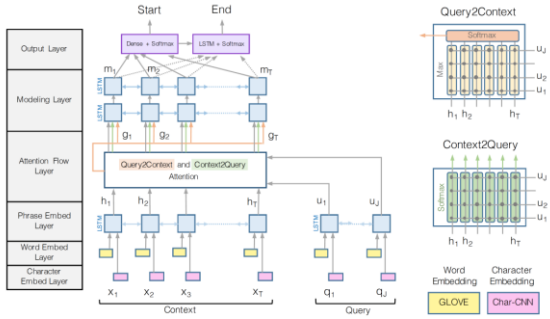

Figure 1: BiDAF Architecture

3.3 Proposed model

The difference of our proposal to the baseline model is an additional self-attention layer.

- Self-attention layer (layers.Selfatt): The proposed architecture self-attention bi-directional attention flow model(SA-BiDAF for short) has an additional self-attention layer on top of a previous attention layer (layers.BiDAFAttention). The purpose of this self-attention layer is to capture the global interactions within the output of the last bi-directional attention layer. The output of the self-attention layer are the aggregations of these interactions and attention scores. The steps to calculate the outputs of the self-attention layer are not of much difference to the regular attention mechanism. The difference is in the self-attention layer at i^{th} iteration, the attention scores to the i^{th} input are calculated on the same input sequence. Then a softmax function is followed to get attention values. Finally, the output at i^{th} position is obtained based on i^{th} input and its corresponding attention values.

After flowing through the self-attention layer, the inputs would also go through the modeling layer and output layer like the baseline model.

3.4 Datasets

Our models are trained on two datasets, QuAIL and RACE and a concatenation of them.

QuAIL is the first multi-domain text comprehension challenge that is balanced and annotated for 9 types of verbal reasoning. It has a new corpus of 800 texts, 300-500 tokens per text handpicked to ensure coherence, and 15K multi-choice questions (18 per text). QuAIL aims to show the extent to which current models can generalize over different domains and reasoning strategies and handle questions that can be answered with the information in a given text, unanswerable questions and questions that require extra world knowledge.[8]

Model	*Text-based questions*				*World knowledge questions*				Unanswe-rable questions	All questions
	Temporal order	Corefe-rence	Causa-lity	Factual	Subsequent state	Event duration	Entity properties	Belief states		
LongChoice	36.3	32.3	46.8	35.9	29.5	33.6	35.0	30.9	12.2	35.6
AvgCos	13.6	5.9	28.2	6.3	20.0	7.7	27.7	21.8	**65.9**	22.1
LSTM	37.0	32.4	38.5	20.2	36.8	43.6	30.8	34.7	51.8	37.2
PMI	42.5	48.3	57.8	**57.5**	32.9	37.0	33.7	37.5	23.3	41.8
IR	27.9	30.0	42.5	30.8	29.6	35.4	27.5	32.0	28.8	32.4
TriAN	**55.5**	**53.1**	60.1	55.0	47.5	56.9	45.8	43.3	65.0	54.7
BERT	52.9	46.2	**67.1**	55.8	**56.7**	**63.8**	**48.8**	**55.0**	54.2	**55.9**

Accuracy of baseline models on QuAIL by question types

These are some examples of baseline model accuracy from the Text-Machine Lab team tested on 15% of the full QuAIL dataset.

RACE is a dataset for benchmark evaluation of methods in the reading comprehension task. Collected from the English exams for middle and high school Chinese students in the age range between 12 to 18, RACE consists of near 28,000 passages and near 100,000 questions generated by human experts (English instructors) and covers a variety of topics which are carefully designed for evaluating the students' ability in understanding and reasoning. In particular, the proportion of questions that requires reasoning is much larger in RACE than that in other benchmark datasets for reading comprehension, and there is a significant gap between the performance of the state-of-the-art models (43%) and the ceiling human performance (95%). Some desirable properties of RACE include the broad coverage of domains/styles and the richness in the question format. Most importantly, it requires substantially more reasoning to do well on RACE than on other datasets, as there is a significant gap between the performance of state-of-the-art machine comprehension models and that of the human. [9]

3.5 Experimental setup

For dataset preparing, we have two setups: the QuAIL standalone, or the combination of QuAIL and RACE. For the QuAIL standalone, we only use training data and validation data from QuAIL, which has 10246 and 2164 respectively. For the combination of QuAIL and RACE, we use training data from QuAIL and RACE and the validation data from QuAIL only. The quantities for training and validation are 98112, 2164 respectively.

The original implementation of BiDAF is for reading comprehension tasks without choices, and the context and question are encoded separately. For this study, we further encode the questions together with their corresponding choices (Q&C) like the following format: "$< q >$ where is the author from? $< 0 >$ US $< 1 >$ China $< 2 >$ France $< 3 >$ England", in which "$< q >$" is the label for the starting position of the question and "$< 0 >$","$< 1 >$","$< 2 >$","$< 3 >$" are sequential labels for the starting points of different choices.

The pre-trained embeddings GloVe with 50 embedding dimensions is adopted as the fixed length vectors to represent tokens in context and Q&C. And the vector is configured to be flexible to update itself.

For the experiment with data augmentation, the fixed paragraph length is 1392, and the fixed length for Q&C is 349. For the ones without data augmentation, the fixed paragraph length is 494, and the fixed length for Q&C is 104. The difference comes from the length difference of paragraph between QuAIL and RACE.

For other parameters, we set 64 as batch size, 80 as hidden size. And Adadelta is adopted as the optimization algorithm, with learning rate 0.5. And the number of training epochs is set to be 20.

Setup	Dataset(s)	Accuracy
BiDAF Baseline	QuAIL	0.234
BiDAF Baseline	QuAIL+RACE	0.255
SA-BiDAF	QuAIL	0.251
SA-BiDAF	QuAIL+RACE	**0.264**

Table 2. Experiment Results

The results of four different setups on the same validation set are documented in Table 1. Overall speaking, the accuracy scores are not promising. We further investigated others' work [14] and confirmed that BiDAF is not an ideal model (no better than random guess) for reading comprehension with multiple choices, although it does well at finding the starting and ending position of the relevant answer to a selected question in a given context. Nevertheless, our proposal improved the baseline by 3%, which proves the efficiency of the proposed self-attention layer. And the additional training data from the RACE dataset can further augment the performance of both BiDAF and SA-BiDAF.

4 Discussion

In this study, a modified version of the BiDAF model is implemented on the QuAIL dataset. It is found that BiDAF and its variation may not be ideal models for the multiple-choice question answering task as these models do not significantly outperform the random guess. However, we proved that a concatenation of QuAIL and RACE could be a promising improvement in accuracy while implementing other models.

There are several things we can do for further improvements. First, we have encoded questions and their candidate choices together. However, there may exist better ways to encode the inputs. Three-way attention could be a promising direction, in which choices are separated out from their questions. And a new layer of choices-to-context attention could be added and concatenated with the existing attention layer. Also, due to limited computation power (single GPU) and constrained time, the ideal setting of hyperparameters are not fully explored. Another future work includes implementation of BERT with the concatenation of QuAIL and RACE and any other state-of-the-art model for QA. And to further study the strength and drawbacks of our model and to compare to other benchmarks in the QuAIL leaderboard, more detailed accuracy scores on different question types could be reported.

5 References

[1] Welbl, J.; Liu, N. F.; Gardner, M. (2017) Crowdsourcing Multiple Choice Science Questions. arXiv:1707.06209v1

[2] Zhang, Z.; Yang, J.; Zhao, H. (2020) Retrospective Reader for Machine Reading Comprehension. arXiv:2001.09694v1

[3] Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. (2020) ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS. arXiv:1909.11942v6

[4] Zhang, Z.; Wu, Y.; Zhou, J.; Duan S.; Zhao, H.; Wang, R. (2019) SG-Net: Syntax-Guided Machine Reading Comprehension. arXiv:1908.05147

[5] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692v1

[6] Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q. V. (2020) XLNet: Generalized Autoregressive Pre-Training for Language Understanding. arXiv:1906.08237v2

[7] Zhang, Z.; Wu, Y.; Zhao, H.; Li, Z.; Zhang, S.; Zhou, Xi; Zhou Xiang (2020) Semantics-aware BERT for Language Understanding. arXiv:1909.02209v3

[8] Rogers, A.; Kovaleva, O.; Downey, M.; Rumshisky, A. (2019) Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks. AAAI

[9] Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; Hovy, E. (2017) RACE: Large-scale ReAding Comprehension Dataset From Examinations. arXiv:1704.04683v5

[10] Seo, M.; Kembhavi, M.; Farhadi, A.; Hajishirzi, H. (2017) Bidirectional attention flow for machine comprehension. arXiv:1611.01603

[11] Shoeybi, M.; Patwary, M.; Puri, R.; LeGresley, P.; Casper, J.; Catanzaro, B. (2020) Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv: 1909.08053v1

[12] Zhu, P.; Zhao, H.; Li, X. (2020) Dual Multi-head Co-attention for Multi-choice Reading Comprehension. arXiv:2001.09415v4

[13] Srivastava, R.K.; Greff, K.; & Schmidhuber, J. (2015) Highway Networks. arXiv:1505.00387

[14] Pirtoaca, G.; Rebedea, T.; Ruseti, S. (2019) Improving Retrieval-Based Question Answering with Deep Inference Models. arXiv:1812.02971v2