Error Analysis of Machine Comprehension on SQUAD 2.0

Shi Feng Pengyuan Chen Xiaohui Chen Xi Chen

Abstract

The task of Question Answering has gained prominence in the past few decades for testing the ability of machines to understand natural language. And as computational resources are more available nowadays, model like BERT, ALBERT are proposed and have reach the SOTA performance on Question Answering area. In this report, we attempt to understand and compare the existing transformer-based model on Stanford Question Answering Dataset (SQuAD) by performing quantitative as well as qualitative analysis of the results attained by each of them. We observed the visualized attention of BERT that might reflect the reason of prediction errors, which we further discuss in the qualitative analysis section.

1 Introduction

Attention was introduced in encoder-decoder architecture (Zhou et al., 2016), which focus on the important parts of input by assigning different weights to them. This mechanism has been proved useful and adopted in many kinds of models, one typical model is the transformer architecture, based on which the pre-trained model BERT was proposed(Devlin et al., 2018). A trained BERT model takes a sentence as input and outputs accordingly the vector for each word in the sentence. Also, the output vector for a word depends on the context it comes from. Later on, many changes have been made mainly based on architecture of BERT. For instance, XLNet(Yang et al., 2019), ALBERT(Lan et al., 2019) and so on. The changes are made in order to improve the performance of BERT, save the computing memory as well as speed up the training time. In our project, we adopt both BERT and ALBERT to do the prediction task on the dataset SQuAD 2.0.

The dataset we use is Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2018), which is a reading comprehension dataset. It consists of questions posed on Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, some of the questions are unanswerable. We use SQuAD 2.0 since it is a popular dataset for machine comprehension and question answering task. This QA task is always challenging since it requires a comprehensive understanding of natural languages and the ability to do further inference and reasoning (Wang and Jiang, 2016) (Seo et al., 2016) (Chen et al., 2017), we choose to implement the popular pre-trained model BERT and ALBERT to solve this problem and analyze the predicted answer mainly by visualizing the attention layer.

In the following parts, section 3 introduces the related works including transformer, BERT, XLNet, ALBERT and SQuAD; section 4 introduces the methods we use to do the prediction task in specific; section 4 is the analysis of our experiment results, mainly focused on the weights of attention heads; section 5 is the prospect of further work.

2 Related Work

2.1 Transformer architecture

In 2017, Vaswani (Vaswani et al., 2017) showed that the dominant sequence transduction models that based on complex recurrent or convolutional neural networks can be replaced with a simple and parallel network architecture, the Transformer. This structure is based solely on self-attention mechanism. With larger and larger embedding vector of words, this structure is proved to save time and efficient on several famous NLP tasks. Then in 2018 scholars adopted this structure with more new features to create computational but powerful pretrained models like BERT (Devlin et al., 2018), ELMo (Peters et al., 2018), and OpenAI GPT (Radford et al., 2018).

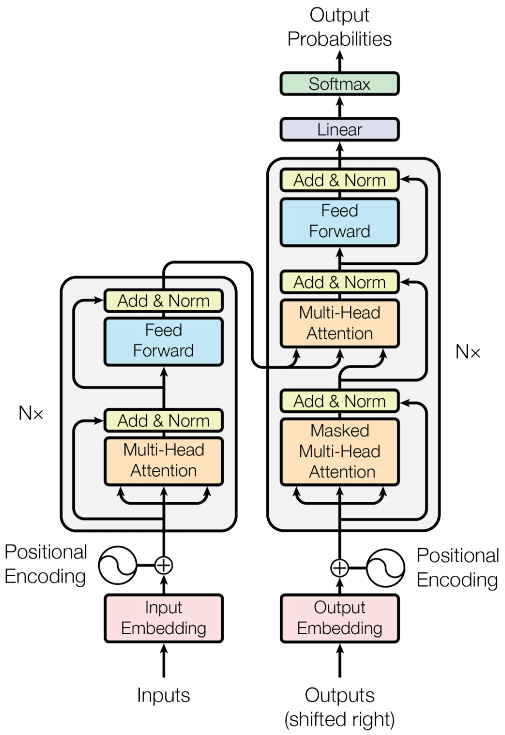


Figure 1: Transformer model architecture

2.2 BERT

BERT, stands for Bidirectional Encoder Representations from Transformers(Devlin et al., 2018), has reached state-of-the-art results in a wide variety of NLP tasks. The key technical innovation of BERT is applying the bidirectional training of transformer, and providing two novel pretraining technique name Masked LM (MLM), which allows bidirectional training in models in which it was previously impossible, and Next Sentence Prediction (NSP), which allows the model to capture and understand the relationship between two sentences. The pretraining of BERT is proceeded unsupervisedly without human interference. And a well-trained BERT is used for multiples down stream NLP tasks such as Question Answering (SQuAD), Natural Language Inference (MNLI), etc.

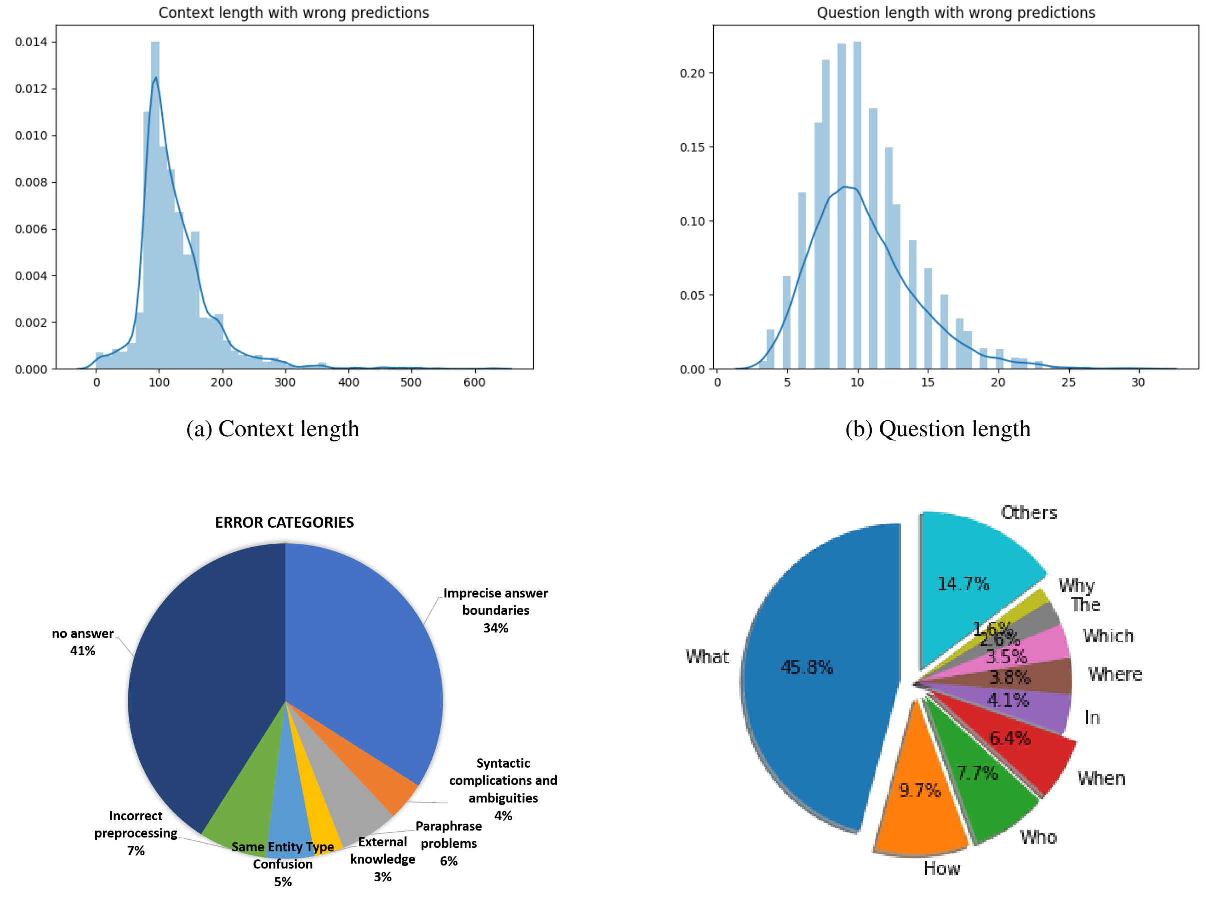
2.3 XLNet

XLNet was introduced in 2019 after BERT by the paper: “XLNet: Generalized Autoregressive Pretraining for Language Understanding” (Yang et al., 2019). It is an autoregressive(AR) language model using the Transformer-XL as the network architecture. XLNet leverages the advantages of both AR and AE, while avoiding their limitations. It adopts permutation operations on the factorization order and keeps the original sequence order to encode bidirectional context information in pretraining. Unlike AE language modeling BERT, XLNet does not have [MASK] in the pretraining, which eliminating the independence pitfalls and pretrain-finetune discrepancy. However, permutation language modeling causes slow convergence in the experiments. The authors chose to predict the last N tokens in a factorization order to reduce the optimization difficulty. To handling position embedding, XLNet proposes ”Two-Stream Self-Attention” scheme to address the idea of target-aware representations. It uses two sets of hidden representations: the content stream encodes both positional embeddings and token embeddings; the query stream encodes positional information, but token embeddings exclude the actual token the model is predicting.

2.4 ALBERT

As we know, the increase on the model size when doing pretraining could result in better performance(Devlin et al., 2018), which can be shown in the work of BERT. However, on the other side, as model size increases, the amount of parameters also increases which means more memory resources are needed. Making a trade-off between parameter increase and high performance is always hard. A Lite BERT (ALBERT) (Lan et al., 2019) was proposed to cut down the parameters and get better performance at the same time.

Compared with BERT, the improvements made by ALBERT focus on three parts. First, the factorized embedding parameterization. In BERT, the WordPiece embedding size E is tied with the hidden layer number H, which means H would increase as H increases. WordPiece embeddings are supposed to learn the context-indepdent representations while the hidden-layer embeddings are meant to learn context-dependent representation, comparing with context-independent representations, the context-dependent representations are supposed to be learned in BERT, that is to say, the increase on E is useless and redundent. So ALBERT insert a lower dimensional embedding space E into the projection from one-hot vectors, the vocabulary size V to hidden space, the hidden layer size H, which can be expressed as $O(V \times H)$ to $O(V \times E + E \times H)$, the parameters can be reduced efficiently when H and E are untied and $H \gg E$. Second, the cross-layer parameter shar-



(a) Context length

(b) Question length

(c) Qualitative error type statistics

(d) Interrogative words statistics

Figure 2: Quantitative and Qualitative Statistical Analysis: (a)X-axis is the length of each context and Y-axis is the proportion it takes; (b)X-axis is the length of each question and Y-axis is the proportion it takes; (c)Statistical results over 7 error types: imprecise answer boundaries and no answer contribute to 75% of all errors; (d)The pie chart shows that, 45.8% of all the interrogative words are ”What”, the other words each takes less than 10% of all.

ing. There are three ways of sharing parameters: only sharing parameters of feed-forward network (FFN), only sharing parameters of attention and sharing them both. All sharing method reduces the parameters most efficiently and therefore lower down the performance most, this could be compensated by increasing the size of ALBERT, which is called ALBERT-xxlarge. Third, sentence-order prediction (SOP) loss. In BERT, the next-sentence prediction (NSP) is used, which is a binary classification loss designed to improve downstream tasks. NSP conflates topic prediction and coherence prediction in one task, however, since topic prediction is easier to learn and therefore overlaps what the masked language modeling (MLM) loss has already learned, NSP shows ineffectiveness. In ALBERT, we focus more on the coherence and propose SOP loss, which avoids topic prediction. It turns out the SOP can solve the NSP task while NSP cannot solve SOP task.

2.5 Machine comprehension on Squard

Since the publication of the SQuAD data set, a large number of representative models have emerged one after another, which has greatly promoted the development of the field of machine reading comprehension. In general, since the answer of SQuAD is limited to the original text, the model only needs to determine which words in the original text are the answer, so it is a extractive QA task rather than a generative task. Most SQuAD models can be summarized into the same framework (Weissenborn et al., 2017): Embed layer, Encode layer, Interaction layer and Answer layer. The Embed layer is responsible for mapping the original text and the tokens in the question into a vector representation; the Encode layer mainly uses RNN to encode the original text and the question, so that the vector representation of each token after encoding contains the semantic information of the context;

the Interaction layer is the majority. The focus of the research work is that this layer is mainly responsible for capturing the interaction between the question and the original text, and outputting the original text representation that encodes the semantic information of the problem, that is, the original text representation of query-aware; the Answer layer is based on the original text To predict the range of answers.

3 Methods

3.1 Length of questions and context

After extracting the questions with wrong predictions, we first count the length of these questions and context.

For long length question, it's take more time to understand by human-beings and so it is with the machine. Extracting useful information from context for long question seems hard for state-of-art model.

3.2 Interrogative words

We count the Interrogative words of all the questions, we found that nearly half of them is ”What”, some other words like: ”How”, ”Who”, ”When” appear frequently as well. The pie chart Fig 2d shows the distribution intuitively.

3.3 Classification

For qualitative error analysis, we sample 100 incorrect predictions from BERT model, and classify those errors into 7 categories based on error types defined by Soumya Wadhwa(Wadhwa et al., 2018).

Imprecise answer boundaries: a model predicted span is longer or shorter than ground truth but contains all or some of the answer.

Syntactic complications and ambiguities: wrong predictions caused by the context itself could be interpreted in more than one way due to ambiguous sentence structure.

Paraphrase problems: question paraphrases certain parts of the context which leads to errors in prediction.

External knowledge: even the ground truth is contained in context, the question requires world knowledge to answer.

Same entity type confusion: the question is about an entity type. The model returns an entity that is the same type as the ground truth entity but different meaning.

Incorrect preprocessing: the model has difficulty to preprocess and embedding special characters the context which leads to wrong prediction.

No answer: no prediction generated by the model when the question should have an answer.

4 Experiments

The experiments were performed by reproduced several models performance from pretrained transformers. We use the framework huggingface to further develop our training, evaluating and analyzing codes. We use P100 graphical card, with 8 cores CPU and 32G memory from Tufts HPC to finetune BERT, ALBERT, Roberta on SQuAD 2.0 dataset for 3 epochs respectively.

4.1 Imprecise answer boundaries

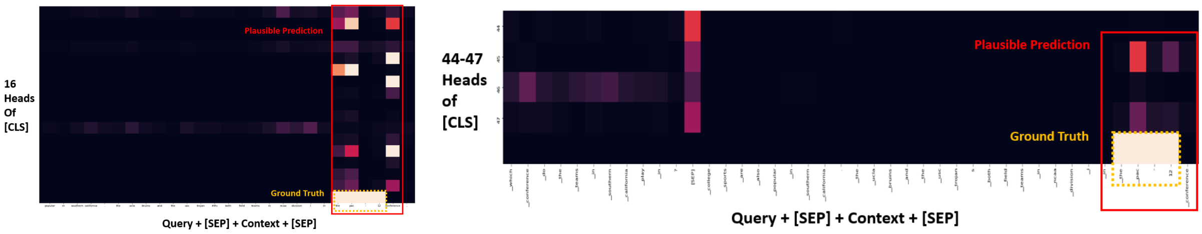
According to our statistical analysis, 34%of sample errors are falling into this category. We are interested in finding the model internal relationship for those errors. Thus, we extract and visualize the attention weights layer by layer for transformer models.

We list one example in this category for further discussion:

Example: 570614ff52bb89140068988d
Context: College sports are also popular in southern California. The UCLA Bruins and the USC Trojans both field teams in NCAA Division I in the Pac-12 Conference, and there is a longtime rivalry between the schools.
Question: “Which conference do the teams in southern California play in?”
BERT Prediction: pac 12 conference
Answer: pac12

We first implement visualization operations on BERT-large model. BERT-large model's architecture uses 16 head attention mechanisms to capture a broader range of relationships, and it also stacks 24 layers of attention to form richer representations.

In BERT embeddings, [CLS] is a special symbol added in front of every input example; [SEP] is a special separator token. The final hidden state corresponding to [CLS] token is used as the aggregate



(a) 4.1 BERT-large attention visualization

(b) 4.1 Albert-xxlargev1 attention visualization

Figure 3: Imprecise answer boundaries

sequence representation and is fed into an output layer (Devlin et al., 2018).

According to those features, we visualize attention matrices by forming the 16 heads of [CLS] token to query and context in the last layer. It provides us with a lens through how BERT forms complex representations to tackle with reading comprehension tasks.

In the visualization, the x-axis is query +[SEP] + context + [SEP], and the y-axis is 16 heads of [CLS]. We mark the ground truth at the bottom with lightest color.

The Figure 3a below is the example case 570614ff52bb89140068988d. It indicates tokens “the” “pac” “12” “conference” in context get higher attention weights than others, so the prediction is “pac 12 conference”. However, the ground truth is “pac 12” which has shorter span.

In approximately 85% of the cases in category 1, the predictions are longer than the ground truth.As expected, we find whenever a prediction is longer than the correct answer, the distribution of attention weights would be spread throughout more tokens than the ground truth; and vice versa.

We also compare attention matrices for the same examples in Albert model. Albert-xxlargev1 model’s architecture has 64 head attention mechanisms and 12 repeating layers.

Albert model prediction for this question is the same as BERT prediction.

The Figure 3b above displays 44-47 heads of attention, where the most attention weights are contained. For the uncropped version, refer to the Appendices A.

Although the token “conference” still gets some attention weights, it is significantly smaller than the weights in BERT model.

We find that Albert improves imprecise boundary errors remarkably. 14 out of 34 Category 1 errors made in BERT are fixed by Albert model. 82% of the cases show lower attention weights on wrong tokens as the example we have discussed above. These results indicate that Albert model develops more robust contextual representations to avoid boundary-based errors.

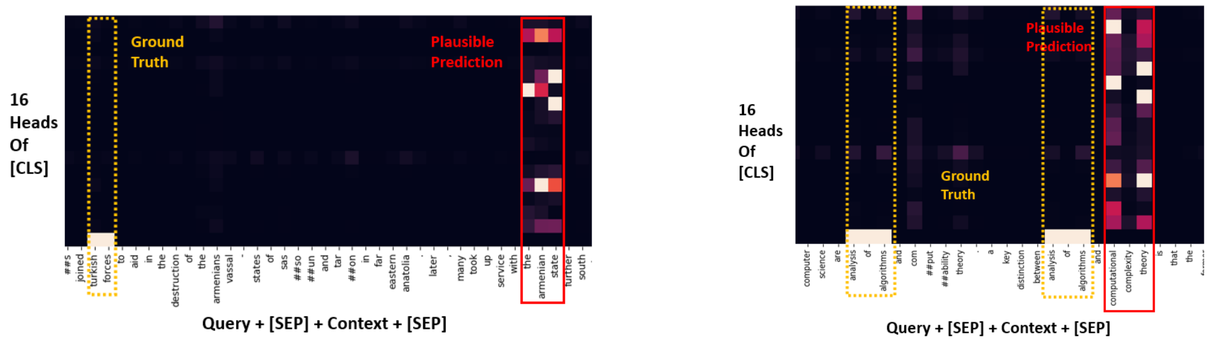
4.2 Syntactic complications and ambiguities

Analysis: As we seen in Figure 4a, most of the attention weights focus on American state, and thus the prediction is American state. For the context we can find that the model doesn’t link the mean of ”join” and ”team up with”. However, it links the mean of ”join” and ”take up service with”. There an ambiguity in this question that Norman cooperate with both Turkey force and American state. It’s even hard for a real people to distinguish the difference from them. The ground true for this wrong question is a little better than prediction from model but not obvious.

Analysis: From context showed in Figure 4b we know, time complexity and the size of input will are the characters of models. However, if people don’t have prerequisite, they may determine the wrong answer as the model. The reason behind it we think is, the distance between ground true and the information of question is long and both of them has very long length. Most attention mechanism are based on word not thought groups. Thus, for a answer that has long length, it’s hard for make it associate with the previous concept.

4.3 Paraphrase problems

Analysis: Figure 5b. Both Catholic orthodoxy and Christian is related to religion. However, the model can’t paraphrase Christian piety to exponents of the Catholic orthodoxy. The reason behind this wrong prediction we think is their different proportion in training set of pretrained model. The word ”Christian” occurs frequently in text, most of which are related to religion while the



(a) 4.2 Attention layer visualization(1)

(b) 4.2 Attention layer visualization(2)

Figure 4: Syntactic complications and ambiguities

word ”Catholic orthodoxy” appears less. So if the question is corresponding to religion, the attention will focus on Christian.

Analysis: Figure 5a. The phrase ”final form” in question is hard for machine to paraphrase. As we known, when people refer to ”final”, there must be a series and this word means at the end of the series. The reason why the prediction is ”the algo saxon language” is the model can distinguish which one in text the ”language” in question refers to while they can't determine whether it's the final form or not.

Analysis: Figure 5c. The length of this question is too long to locate the position of answer. By analyzing the context for the question, we find the relationship between two related answer ”analysis of algorithm” and ”computational complex theory” is complected. The text proposed them at almost same time and use several sentence to explain the difference between them. But for a model which only have attentions based on words, it's hard to understand the logic behind them and thus can't paraphrase them in the right way.

4.4 External knowledge

This error type contributes to only 3% of the observed errors. However, it remains unsolved across all models to some extent.

We list one example in this categories for further discussion:
Example: 56de179dcffd8e1900b4b5db
Context: “The Normans had a profound effect on Irish culture and history after their invasion at Bannow Bay in 1169. . . The Normans settled mostly in an area in the east of Ireland. . . ”
Question: “What country did the Normans invade in 1169?”
BERT Prediction: “irish”
Albert Prediction: “irish”
Answer: Ireland

As described in Section 3.1, we visualize the attention matrices to see how BERT and Albert compose representations internally.

For simplicity, we only display 100 tokens on the x-axis since other tokens do not contain useful information in this example.

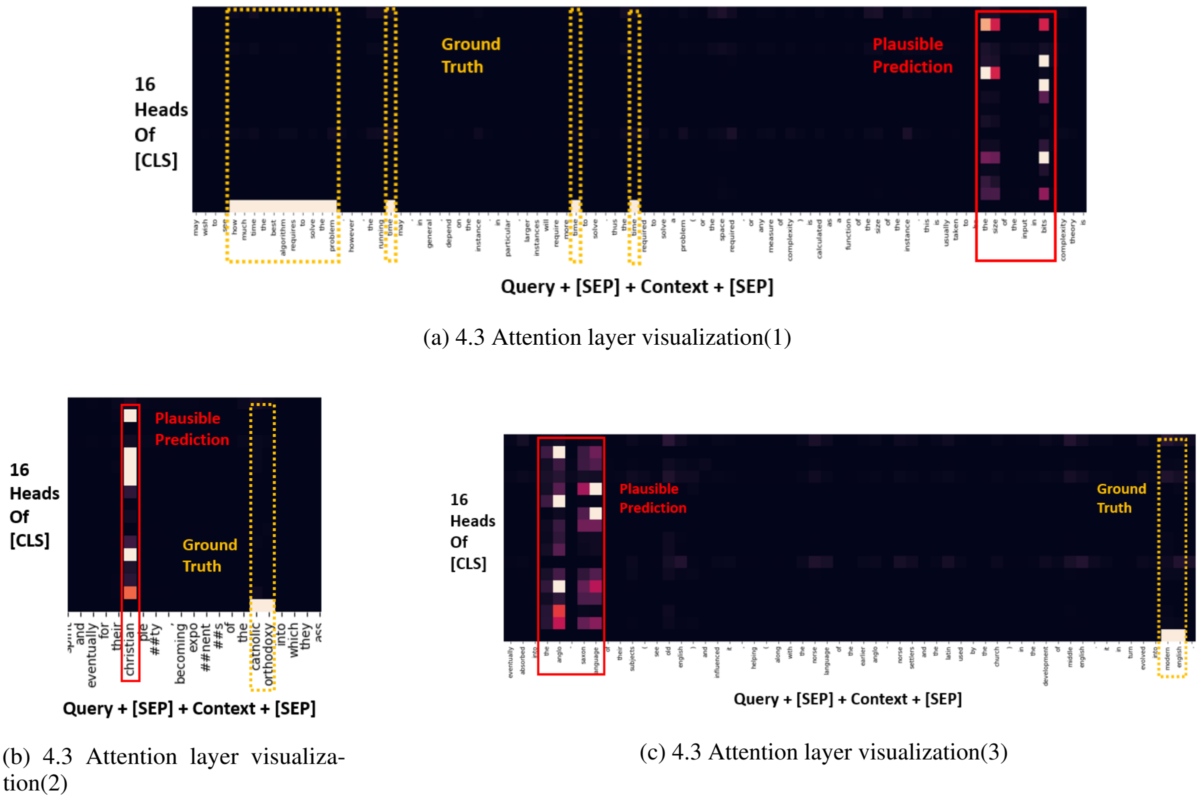
As Figure 6b and 6a shown, both BERT and Albert wrongly address attention on ”Irish” instead of ”Ireland”. Intuitively, we know that ”Irish” is indeed the key word to answer this question. But in order to obtain exact answer, we need external knowledge as Irish culture implicitly refers to the country Ireland.

Even though Albert-xxlargev1 model archives F1 and EM scores up to 89.3% and 85.7%, we observe little improvement on external knowledge errors. We would leave this error type in-depth investigation to future work, which we think it might break limitations of state-of-the-art QA models.

4.5 Same Entity Type Confusion

This kind of error includes those cases that question is about an entity type of specific type of works. And the model returns a entity that is the same type as the ground truth entity but different meaning. This kind of error is common in the wrong predictions. We will extract 3 types of examples to illustrate this type of errors, they are (ground truth in the square bracket):

digit
Q: How many Victorians are non-religious?



(a) 4.3 Attention layer visualization(1)

(b) 4.3 Attention layer visualization(2)

(c) 4.3 Attention layer visualization(3)

Figure 5: Paraphrase problems

A: 168637. [20]

geo name
Q: Where was Victoria first set to be located in Australia?
A: sullivan bay. [new south wales]
Q: What present-day area was this settlement near?
A: fort caroline. [parris island]

time
Q: When was this proclamation issued?
A: 1598. [1629]

Note, in the visualized attention matrix, more weights are also put on the wrong answer tokens.

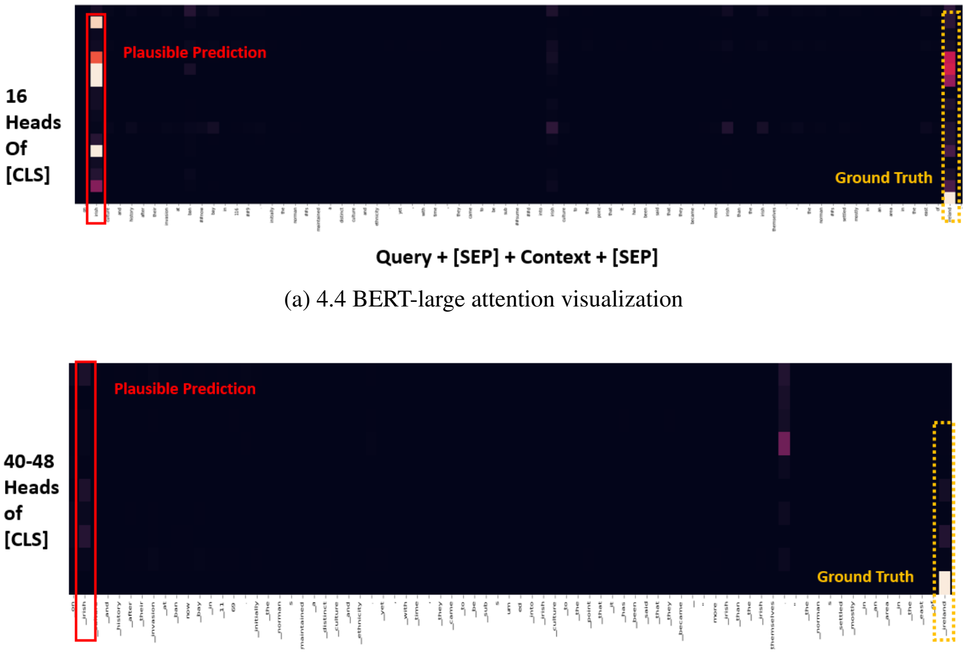
4.6 Incorrect preprocessing

We extract 3 typical examples to analyze the reasons of the wrong prediction mainly based on the attention weights of CLS to all.

Take the first example as following:
Context: Southern California contains a Mediterranean climate, with infrequent rain and many sunny days. Summers are hot and dry, while winters are a bit warm or mild and wet. Serious rain can occur unusually. In the summers, temperature ranges are 90-60's while as winters are 70-50's, usually all of Southern California have Mediterranean climate. But snow is very rare in the Southwest of the state, it occurs on the Southeast of the state.
Question: What is the low end of the temperature range in summer?
Ground truth: 60s
BERT Prediction: 9060s

Analysis: We can find that 90 in not in the ground truth, but the predicted result includes it. When we look into the attention layer Fig 7a, we can find that ”90”, ”60”, ”’” in sentence ”90-60’s” have higher weights, similarly, ”70”, ”50”, ”’” in sentence ”70-50’s” have high weights, so we can know that the attention captures ”range” in the query, however, the ”low end” is ignored, so the answer focus on the words after ”range” but fail to get the meaning of ”low end”.

Another typical example:
Context: BSkyB initially charged additional subscription fees for using a Sky+ PVR with their service; waiving the charge for subscribers whose package included two or more premium channels. This changed as from 1 July 2007, and



(a) 4.4 BERT-large attention visualization

(b) 4.4 Albert-xxlargev1 attention visualization

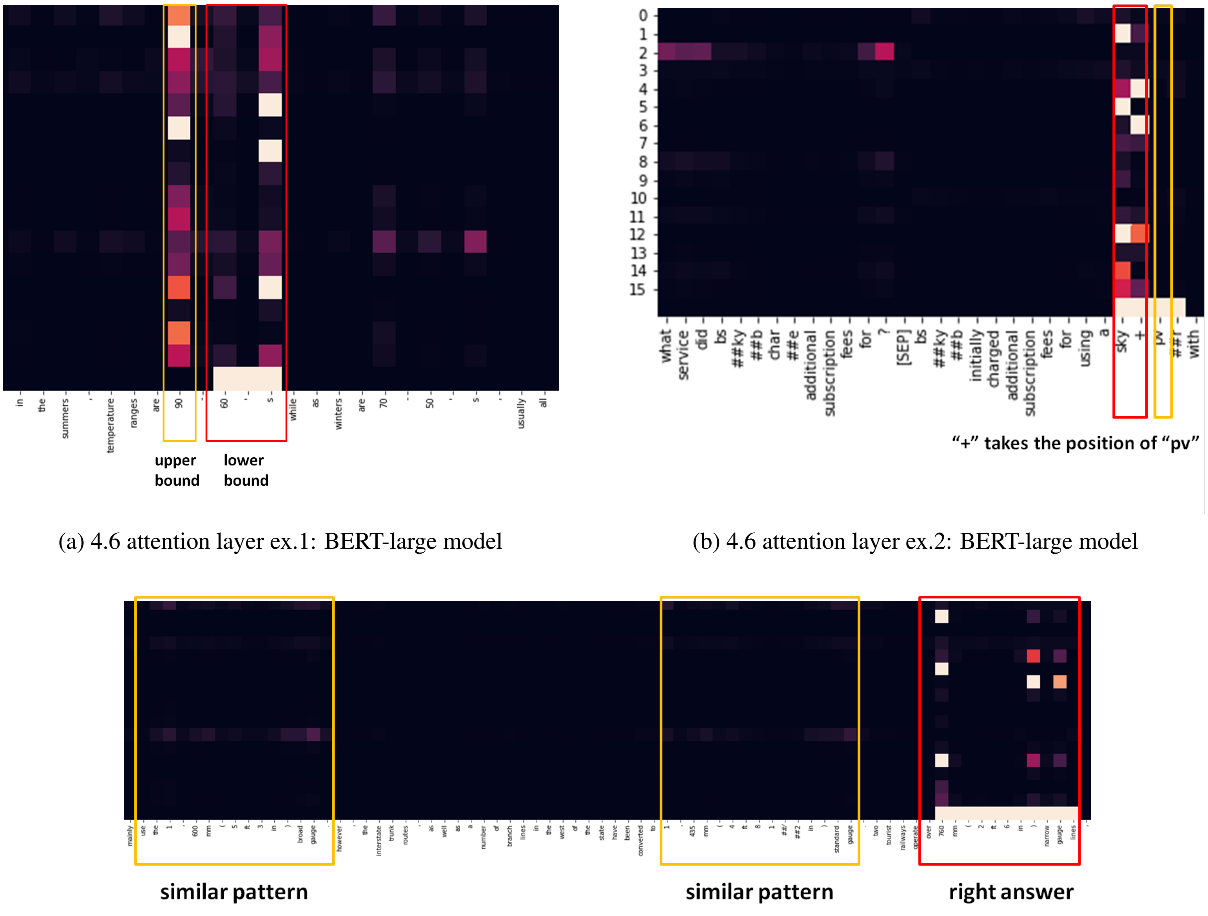
Figure 6: External knowledge

now customers that have Sky+ and subscribe to any BSkyB subscription package get Sky+ included at no extra charge. Customers that do not subscribe to BSkyB's channels can still pay a monthly fee to enable Sky+ functions. In January 2010 BSkyB discontinued the Sky+ Box, limited the standard Sky Box to Multiroom upgrade only and started to issue the Sky+HD Box as standard, thus giving all new subscribers the functions of Sky+. In February 2011 BSkyB discontinued the non-HD variant of its Multiroom box, offering a smaller version of the SkyHD box without Sky+ functionality. In September 2007, Sky launched a new TV advertising campaign targeting Sky+ at women. As of 31 March 2008, Sky had 3,393,000 Sky+ users.
Question: What service did BSkyB chare additional subscription fees for?
Ground truth: sky pvr
BERT Prediction: sky

Analysis: In Fig 7b we can find that the weights of word "sky" and its following symbol "+" is highest among all the words, comparing with the ground truth "sky pvr", we get to know that the predicted result spot on the right answer, however, it fails to understand the meaning of symbol "+". "+" denotes that the word on both sides of it should be equally adopted, that is to say, "+" is similar to "and", the attention layer only captures that the answer has two words which occupy two positions but fail to know the use of "+" here.

The last example is:
Context: There are also several smaller freight operators and numerous tourist railways operating over lines which were once parts of a state-owned system. Victorian lines mainly use the 1,600 mm (5 ft 3 in) broad gauge. However, the interstate trunk routes, as well as a number of branch lines in the west of the state have been converted to 1,435 mm (4 ft 8 12 in) standard gauge. Two tourist railways operate over 760 mm (2 ft 6 in) narrow gauge lines, which are the remnants of five formerly government-owned lines which were built in mountainous areas.
Question: What gauge of rail lines do two tourist lines use?
Ground truth: [760 mm—760 mm 2 ft 6 in narrow gauge lines—760 mm 2 ft 6 in narrow gauge lines]
BERT Prediction: 760 mm 2 ft 6 in

Analysis: In Fig 7c we can see from the weights of [CLS] to all that when the pattern like "number + mm + ft + number + in + adj + gauge" such as "1,600 mm (5 ft 3 in) broad gauge", "1,435 mm (4 ft 8 12 in) standard gauge", "760 mm (2 ft 6 in)



(a) 4.6 attention layer ex.1: BERT-large model

(b) 4.6 attention layer ex.2: BERT-large model

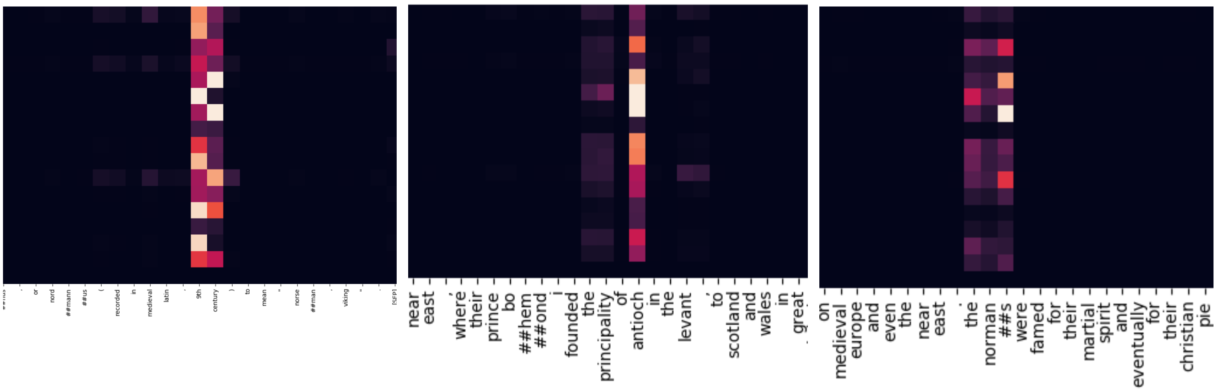
(c) 4.6 attention layer ex.3: BERT-large model

Figure 7: Incorrect preprocessing: (a) is the figure of example 1 of the incorrect preprocessing; (b) is the figure of example 2 of the incorrect preprocessing; (c) is the figure of example 3 of the incorrect preprocessing

narrow gauge”appears, the weights of such pattern are higher than the others. This shows that the attention layer has already captured the answer-like pattern and successfully spot on the right location of answer, however, the reason why it is different from ground truth is that some of the words in the answer are lost. In all, it predict the right answer in some way but not accurate enough.

4.7 No Answer

In the prediction result of our reproduced BERT, 41.9% of wrong prediction in the ”has answer” group are due to the **no answer prediction**, which can tremendously improve the algorithm performance if effective investigation is operated. Because in SQuAD 2.0, we have two groups of dataset: has answer and no answer, while in the no answer group, plausible answers are provided. However, the official metric doesn’t take advantage of this information while evaluating the model performance. The model are not only required to locate the answer position in the corpus, but also need to tell whether the question has answer or not through a binary classifier. And the binary classifier, are to be blamed for the 41.9% error predictions in the has answer group. We investigate the attention layer of the no answer group, and find out that even for no answer inputs, the attention still puts very high weight on the tokens that are labeled as the plausible answers. We analyze 100 no answer samples by giving visualization, and 91 out of them are as expected. we shows three of them, with plausible answers ”Antioch”, ”Normans” and ”9th century” correspondingly. As we can see in Figure 8, the lightest column are basically from those tokens. In conclusion, the reason that model giving no answer prediction for has answer data is not because of the attentions, instead, we will investigate on the binary classifier in the future, which might be improved to solve this problem



(a) 4.7 BERT-large attention matrices Shown: BERT-large model(1)

(b) 4.7 BERT-large attention matrices Shown: BERT-large model(2)

(c) 4.7 BERT-large attention matrices Shown: BERT-large model(3)

Figure 8: Three No ANSWER examples

5 Future work

5.1 Binary classifier analysis for no answer

For most of sample we observed with ”no answer” prediction while the ground true has answer, we find that the attention of these questions is very similar to ground true. However, we have no idea why the model at last judge the answer as ”no answer”. In the future, we will further analyze the classifier from the downstream model, to see what cause the model misclassify the class ”has answer” into ”no answer”. A probable solution is we use a single binary classifier for no answer question. Since the large proportion of ”no answer” prediction, after get the output of attention, we can add recurrent neural network or LSTM(Hochreiter and Schmidhuber, 1997) to make binary prediction. If the model predict ”has answer” for the question, the data will move to next step which predicts the start and end position of answer.

5.2 Attention based on thought group

Nowadays most attention mechanism is based on words or characters. However, for sentence that longer than 100 words, attentions seems hard to focus on the correct things if the question need ”Logic” to answer. A possible method to solve this problem is Attention based on thought group. Each thought group contains lots of words and will filter some words that are trivial. This also can accelerate the training speed. However, the problem is obvious. How can we divide the words into several different thought group. We will focus on this after a while.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Soumya Wadhwa, Khyathi Raghavi Chandu, and Eric Nyberg. 2018. Comparative analysis of neural qa models on squad. *arXiv preprint arXiv:1806.06972*.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Fastqa: A simple and efficient neural architecture for question answering. *arXiv preprint arXiv:1703.04816*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

A Appendices

Model configurations

	BERT	ALBERT	Roberta
setting	large-uncased	xxlarge-v1	large
parameters	340M	223M	355
#layers	24	12	24
#hiddens	1024	4096	1024
#heads	16	64	16
f1	81.8	89.3	87.9
exact	79.0	85.7	84.5

B Supplemental Material

SQuAD 2.0/1.1 Dataset can be download via https://rajpurkar.github.io/SQuAD-explorer
Implementation code is publicly available from https://github.com/Xiaohui9607/FlyWithNLP
Reproduced model result are publicly available via https://tufts.box.com/s/addbv9i1nbwqes7r2