**Build a chatbot based on BERT and the SQuAD 2.0 dataset**

Name: Quoc Khoa Vo

Student ID: 4584844

Supervisor: Gongqi Lin

Date: 19/08/2019

# Literature Review

Nowadays, with the fast-paced development of deep learning, especially in natural language processing (NLP), a lot of organizations are investigating to build systems that can solve their daily tasks automatically. Therefore, this research is to conduct a novelty chatbot for question answering based on the state-of-the-art technologies in NLP.

By using exact match keywords “chatbot question answering” in the IEEE Explore research database, only one result is returned, “An Ergonomics Evaluation to Chatbot Equipped with Knowledge-Rich Mind” (published in 2015). Moreover, searching again with the same keywords, not exact match, this search engine returns 32 articles published from 2017 to 2019. Two of the researches are about FAQ chatbot; two articles focus on using RNN, LSTM which are quite successive models in NLP; some researches attempt to provide chatbot in several domains such as education, e-commerce, elderly care, technical support. The articles list as below:

1. Chatbot for university related FAQs
2. Question answer system for online feedable new born Chatbot
3. Automated Thai-FAQ Chatbot using RNN-LSTM
4. Design of E-commerce chat robot for automatically answering customer question
5. Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots
6. A chatbot using LSTM-based multi-layer embedding for elderly care
7. PESUBot: An Empathetic Goal Oriented Chatbot
8. Nicky: Toward a Virtual Assistant for Test and Measurement Instrument Recommendations
9. Cognitive Natural Language Search Using Calibrated Quantum Mesh
10. Implementation of Chatbot for ITSM Application Using IBM Watson
11. Problem solving chatbot for data structures
12. Chatbot: An automated conversation system for the educational domain
13. KNADIA: Enterprise KNowledge Assisted DIAlogue Systems Using Deep Learning
14. Dinus Intelligent Assistance (DINA) Chatbot for University Admission Services
15. Improvement of Chatbot in Trading System for SMEs by Using Deep Neural Network
16. Question Answering on Structured Data using NLIDB Approach
17. Chatbot Testing Using AI Planning
18. BotHook: An option against Cyberpedophilia
19. Chatbot Application on Cryptocurrency
20. An Efficient Search for Context-Based Chatbots
21. Excitement and Concerns about Machine Learning-Based Chatbots and Talkbots: A Survey
22. Applications of Sequence to Sequence Models for Technical Support Automation
23. AskCO: A Multi-language and Extensible Smart Virtual Assistant
24. Expert system for remote process automation in multiplatform servers, through human machine conversation
25. Rich Short Text Conversation Using Semantic-Key-Controlled Sequence Generation

Searching the keyword “chatbot question answering” in the ScienceDirect database returns 202 articles and 144 articles in the past three years. However, most of them mention about chatbot in the title but not relate to question answering.

Searching again the keyword “chatbot question answering” in the Google Scholar returns 9,840 results and 6,750 results in recent three years. However, almost results do not involve to chatbot question answering much. Moreover, if using the exact match searching, there are only 4 articles from 2017 to 2019 range.

It indicates that chatbot is the hot trend technology, the number of researches mainly focus on its application in many domains. Nevertheless, in question answering domain, researchers do not make use of state-of-the-art solutions in NLP such as Google BERT, SQuAD. Searching with keywords “chatbot bert” or “chatbot bert squad” in IEEE explore database and ScientDirect database return zero result. It shows that chatbot for question answering may be provided in more specific industry domains, or still has room to conduct this research. To implement a solution for this kind of chatbot using BERT and SQuAD, researchers must deeply understand them through reading these articles as below:

1. Deep contextualized word representations
2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
3. XLNet: Generalized Autoregressive Pretraining for Language Understanding

# Related Work

## Deep contextualized word representations

ELMo stands for Embeddings from Language Models. ELMo use language models to obtain embeddings for individual words while taking the entire sentence or paragraph into account.

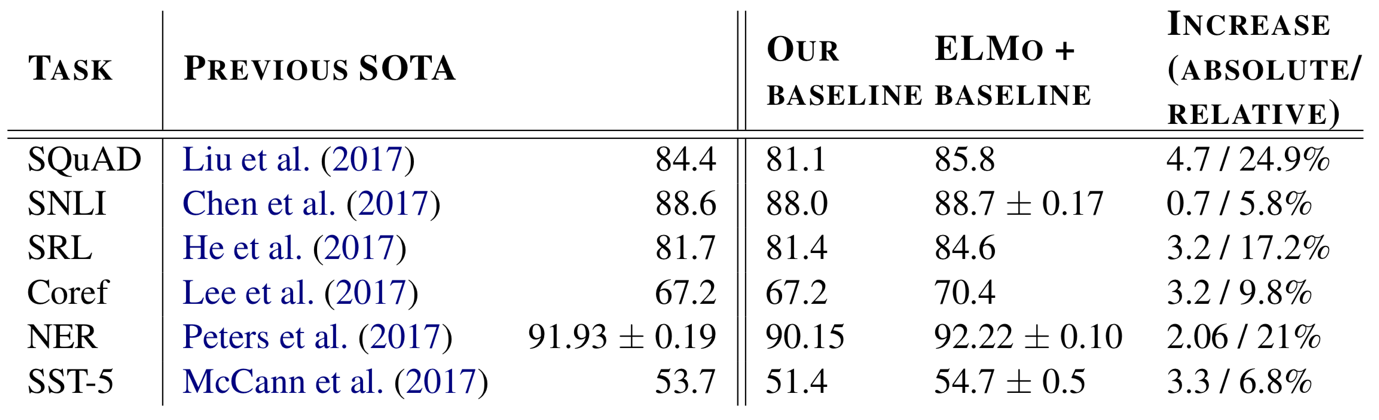
According to Matthew et al. (2018), they introduce new deep contextualized word representations to:

* take context of words into account because of polysemy
* complex characteristics of word use (e.g., syn-tax and semantics)

The authors use a pre-trained, multi-layer, bi-directional, LSTM-based language model and extract the hidden state of each layer for the input sequence of words. Then, they compute a weighted sum of those hidden states to obtain an embedding for each word. ELMo representations are deep in the sense that they are a function of all of the internal layers of the bidirectional language model (biLM).

They show that it is easy to include words embeddings to existing models and enhance the cutting-edge six NLP tasks, including question answering, textual entailment and sentiment analysis. Moreover, they expose that it is vital to implement deep internals of the pre-trained networks which allowing downstream models to mix different types of semi-supervision signals.

Their main achievements are as below:



However, ELMo predicts a single word from both left and right context using LSTMs, but not at the same time which could cause problem in the fine-tuning phase for token-level task where context is extremely significant.

## BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT stands for Bidirectional Encoder Representations from Transformers.

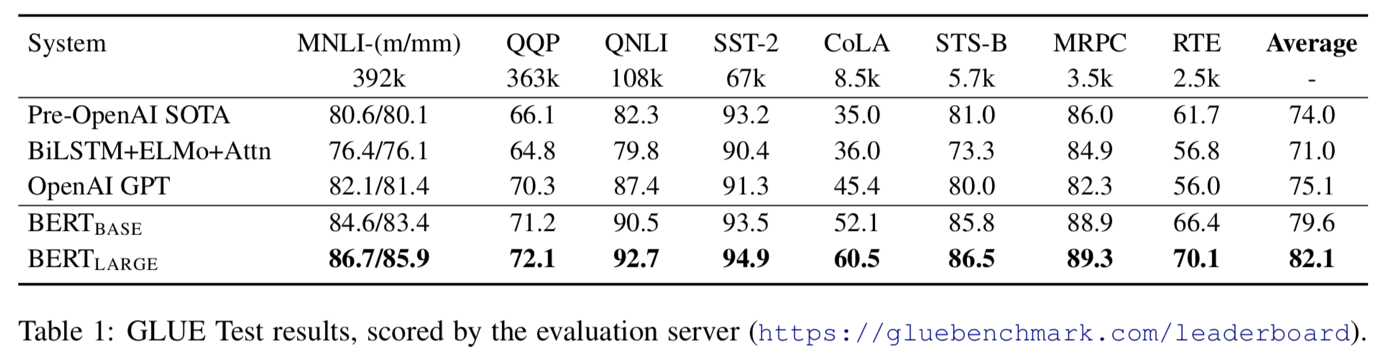
Language modeling is to predict a probability of a word in a sentence within a given context. In NLP, it is an effective task for using unlabeled data to pretrain neural network. Some standard language models attempt to predict a next token in a sequence from left-to-right or from right-to-left, but not at the same time. According to Jacob et al. (2018), these models are named unidirectional, and they may cause the limit of architectural choices that can be used during-pretraining. For example, left-to-right architecture are used by the authors in OpenAI GPT, where transformer’s self-attention layers every token can only show up to previous tokens. This might cause problems for downstream tasks such as question answering, where the relation of words in a sentence, and sentence next to sentence in a context are significant. In these two sentences “I went to the bank to deposit money”, “I sat on the bank of a river”, the word “bank” is considered the same meaning by using standard language models.

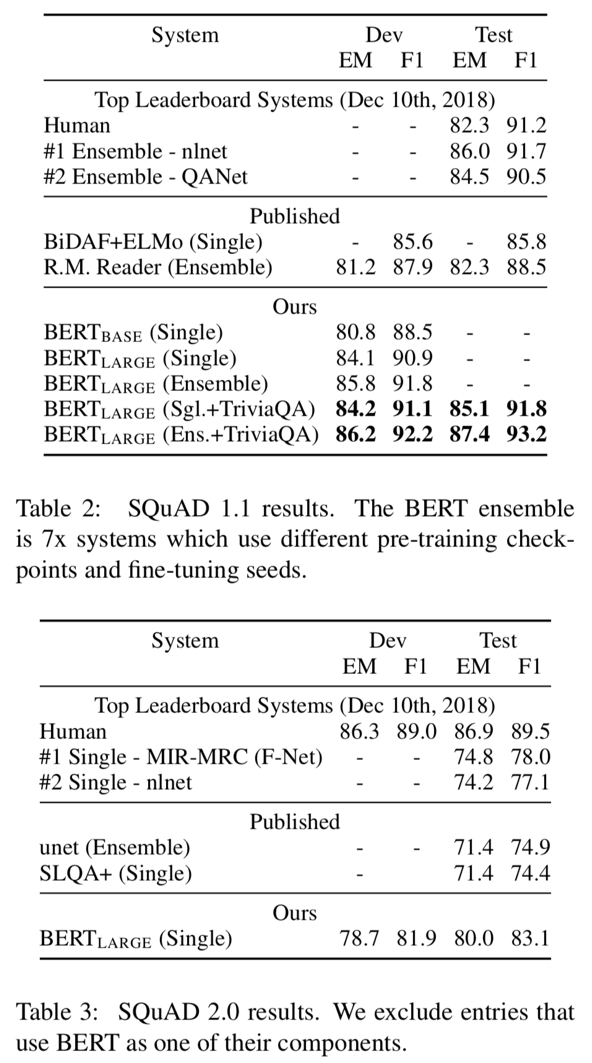
The limitations have motivated the authors to create BERT. The key innovation in their solution is they design BERT to train the language model bidirectional representations deeply from unlabeled dataset, from left-to-right and right-to-left simultaneously in all layers. Thanks to this approach, they contribute to:

* demo the importance of bidirectional pre-training for language representations
* outperform many task-specific architectures
* perform eleven NLP tasks as a cutting-edge technology

As a result, the authors outperform many tasks in NLP as below:

* Push the GLUE score to 80.5% (7.7% improvement)
* Push MultiNLI accuracy to 86.7% (4.6% improvement)
* SQuAD v1.1 question answering Test F1 to 93.2 (1.5 improvement)
* SquAD v2.0 Test F1 to 83.1 (5.1 improvement)





BERT not only is outstanding in pre-train model task but also it is simple to apply the pre-trained BERT representations for fine-tuning without modifying task-specific architecture. To create models to perform wide range of tasks such as question answering and language inference, the pre-trained BERT representations can be fine-tuned with just one additional output layer.

Although BERT gains many outstanding results in many NLP tasks, it also exposes some problems that can be improved in the next version. BERT uses [MASK] for pre-training, but this [MASK] can cause the pretrain-finetune discrepancy because does not appear in the fine-tune phase:

* So, would it really learn to produce meaningful representations for non-masked tokens?
* what if there are no [MASK] tokens in the input sentence.
* BERT supposes that unmasked tokens and predicted (masked) tokens are independent of each other

## XLNet: Generalized Autoregressive Pretraining for Language Understanding

XLNet is a generalized autoregressive pretraining method that:

* allow learning bidirectional contexts using permutation language modeling
* improve BERT’s limitation.

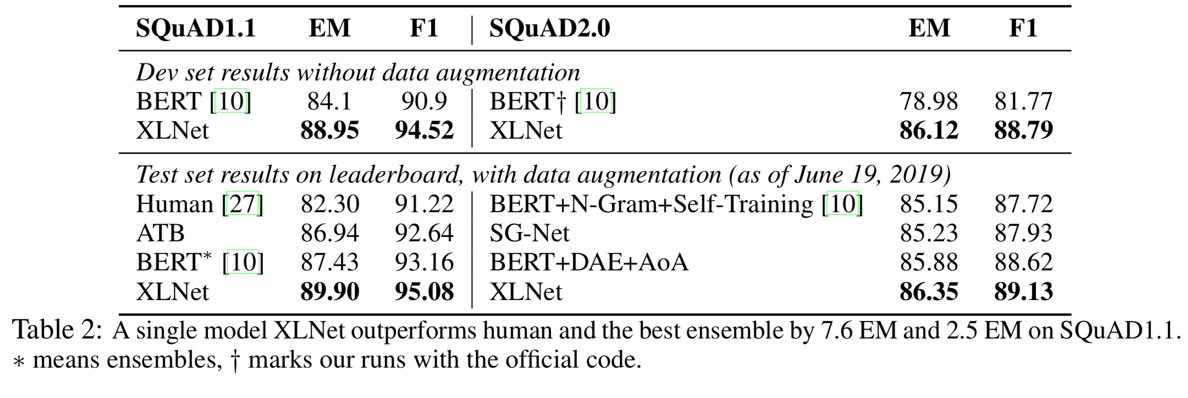
Because BERT exposes potential problems as below:

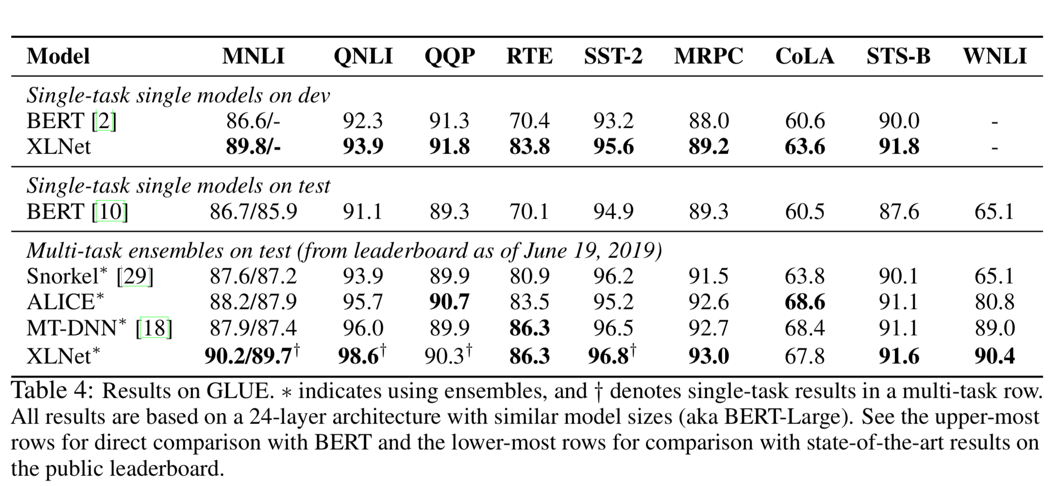
* In pre-training phase, BERT uses the [MASK] token to train the model for predicting the next token in the sequence, but the [MASK] token does not appear during fine-tuning phase. What happens for tokens not replaced with [MASK]? Non-masked tokens might be copied to the output. Then, for non-masked tokens, does BERT learn to create meaningful representations? What if there are no [MASK] tokens in the input sentence?
* BERT supposes that unmasked tokens and predicted (masked) tokens are independent of each other

Zhilin et al. (2019) want to overcome these limitations of BERT to bridge a gap between language modeling and pretraining because of the lack of the capability of bidirectional context modeling.

The authors create an innovation language modeling called “permutation language modeling” to achieve the goal. This language modeling produces bidirectional contexts and maintains benefits of autoregressive models. Traditional language models take the previous n tokens and predict the next one in a given context; however, permutation language models are trained to predict tokens in random order.

Their main achievement is that XLNet consistently outperforms BERT on 20 tasks, often by a large margin.





# Research Problem

The research will build a chatbot for question answering based on BERT/XLNet and SQuAD dataset. The chatbot can learn and answer new knowledge from knowledge base, product manual, or documentation in different field such as education, healthcare, travelling, online commerce.

## References

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer†, 2018, *Deep contextualized word representations*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2018, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le, 2019, *XLNet: Generalized Autoregressive Pretraining for Language Understanding*

**Appendix A.1. search for keywords “chatbot question answering”, exact match**

In the IEEE Explore research database, only one result is returned, “An Ergonomics Evaluation to Chatbot Equipped with Knowledge-Rich Mind” (published in 2015).

**Appendix A.2. search again for keywords “chatbot question answering”, not exact match**

In IEEE Explore returns 32 articles published from 2017 to 2019.

Review:

* Two of the researches are about FAQ chatbot
* Two articles focus on using RNN, LSTM which are quite successive models in NLP
* Some researches attempt to provide chatbot in several domains such as education, e-commerce, elderly care, technical support

The first 25 articles are listed as below:

1. Chatbot for university related FAQs

2. Question answer system for online feedable new born Chatbot

3. Automated Thai-FAQ Chatbot using RNN-LSTM

4. Design of E-commerce chat robot for automatically answering customer question

5. Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots

6. A chatbot using LSTM-based multi-layer embedding for elderly care

7. PESUBot: An Empathetic Goal Oriented Chatbot

8. Nicky: Toward a Virtual Assistant for Test and Measurement Instrument Recommendations

9. Cognitive Natural Language Search Using Calibrated Quantum Mesh

10. Implementation of Chatbot for ITSM Application Using IBM Watson

11. Problem solving chatbot for data structures

12. Chatbot: An automated conversation system for the educational domain

13. KNADIA: Enterprise KNowledge Assisted DIAlogue Systems Using Deep Learning

14. Dinus Intelligent Assistance (DINA) Chatbot for University Admission Services

15. Improvement of Chatbot in Trading System for SMEs by Using Deep Neural Network

16. Question Answering on Structured Data using NLIDB Approach

17. Chatbot Testing Using AI Planning

18. BotHook: An option against Cyberpedophilia

19. Chatbot Application on Cryptocurrency

20. An Efficient Search for Context-Based Chatbots

21. Excitement and Concerns about Machine Learning-Based Chatbots and Talkbots: A Survey

22. Applications of Sequence to Sequence Models for Technical Support Automation

23. AskCO: A Multi-language and Extensible Smart Virtual Assistant

24. Expert system for remote process automation in multiplatform servers, through human machine conversation

25. Rich Short Text Conversation Using Semantic-Key-Controlled Sequence Generation

**Appendix A.3. search for keyword “chatbot question answering” in the ScienceDirect database**

Return 202 articles and 144 articles in the past three years. However, most of them mention about chatbot in the title but not relate to question answering.

**Appendix A.4. search for keyword “chatbot question answering in the Google Scholar”**

Return 9,840 results and 6,750 results in recent three years. However, almost results do not involve to chatbot question answering much.

If using the exact match searching, there are only 4 articles from 2017 to 2019 range.

**Appendix A.4. search for keywords “chatbot bert” or “chatbot bert squad” in IEEE explore database and ScientDirect database**

Return 0 result.

**Summary:** It indicates that chatbot is the hot trend technology, the number of researches mainly focus on its application in many domains. Nevertheless, in question answering domain, researchers do not make use of state-of-the-art solutions in NLP such as Google BERT, SQuAD. It shows that chatbot for question answering may be provided in more specific industry domains, or still has room to conduct this research.

**Appendix B.1. Deep contextualized word representations**

1. Why did the authors conduct the research (existing problems)?
   1. take context of words into account because of polysemy
   2. complex characteristics of word use (e.g., syn-tax and semantics)
2. What did the authors do (key contributions)?

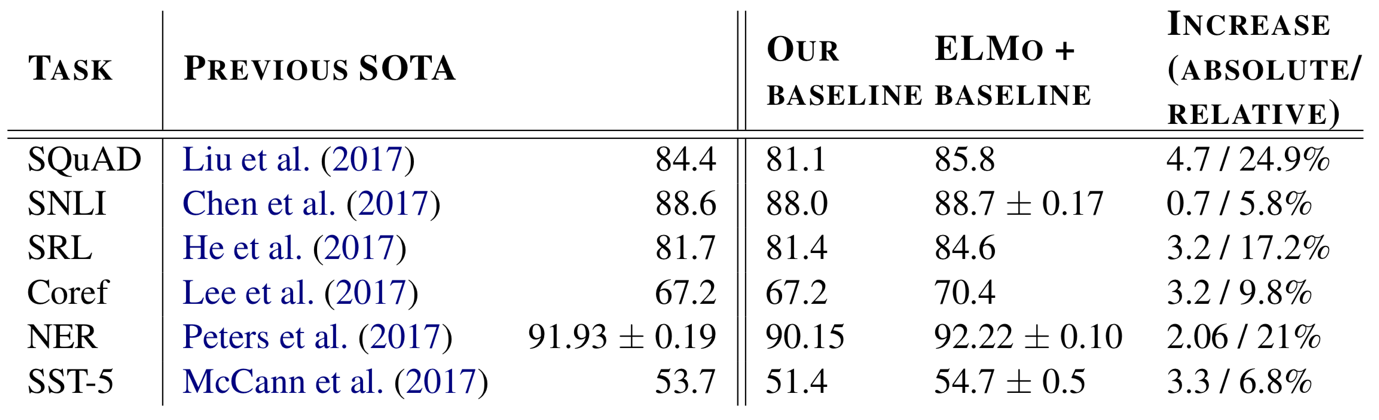
* it is easy to include words embeddings to existing models and enhance the cutting-edge six NLP tasks
* it is vital to implement deep internals of the pre-trained networks which allowing downstream models to mix different types of semi-supervision signals.

1. What are the key differences in the method/approach (innovation)?

Using a pre-trained, multi-layer, bi-directional, LSTM-based language model, extract the hidden state of each layer for the input sequence of words. Then, they compute a weighted sum of those hidden states to obtain an embedding for each word.

ELMo representations are deep in the sense that they are a function of all of the internal layers of the bidirectional language model (biLM).

1. What are main achievements, significance?



1. What can be further improved?

ELMo predicts a single word from both left and right context using LSTMs, but not at the same time which could cause problem in the fine-tuning phase for token-level task where context is extremely significant.

**Appendix B.2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

1. Why did the authors conduct the research (existing problems)?

Some standard language models attempt to predict a next token in a sequence from left-to-right or from right-to-left, but not at the same time. This might cause problems for downstream tasks such as question answering, where the relation of words in a sentence, and sentence next to sentence in a context are significant.

1. What did the authors do (key contributions)?

* demo the importance of bidirectional pre-training for language representations
* outperform many task-specific architectures
* perform eleven NLP tasks as a cutting-edge technology

1. What are the key differences in the method/approach (innovation)?

The key innovation in their solution is they design BERT to train the language model bidirectional representations deeply from unlabeled dataset, from left-to-right and right-to-left simultaneously in all layers.

1. What are main achievements, significance?

* Push the GLUE score to 80.5% (7.7% improvement)
* Push MultiNLI accuracy to 86.7% (4.6% improvement)
* SQuAD v1.1 question answering Test F1 to 93.2 (1.5 improvement)
* SquAD v2.0 Test F1 to 83.1 (5.1 improvement)

1. What can be further improved?

BERT uses [MASK] for pre-training, but this [MASK] can cause the pretrain-finetune discrepancy because does not appear in the fine-tune phase:

* So, would it really learn to produce meaningful representations for non-masked tokens?
* what if there are no [MASK] tokens in the input sentence.
* BERT supposes that unmasked tokens and predicted (masked) tokens are independent of each other

**Appendix B.3. XLNet: Generalized Autoregressive Pretraining for Language Understanding**

1. Why did the authors conduct the research (existing problems)?

Relying on corrupting the input with masks, BERT neglects dependency between the masked positions and suffers from a pretrain-finetune discrepancy

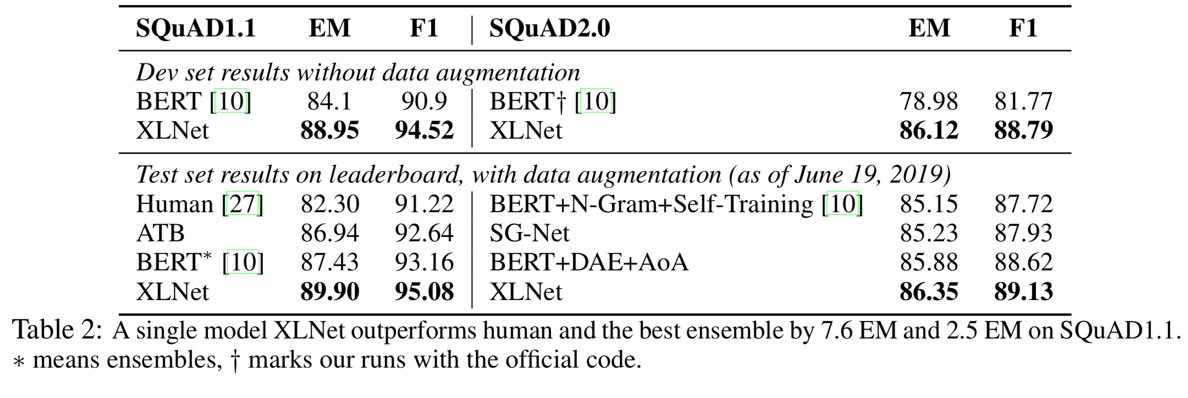
1. What did the authors do (key contributions)?

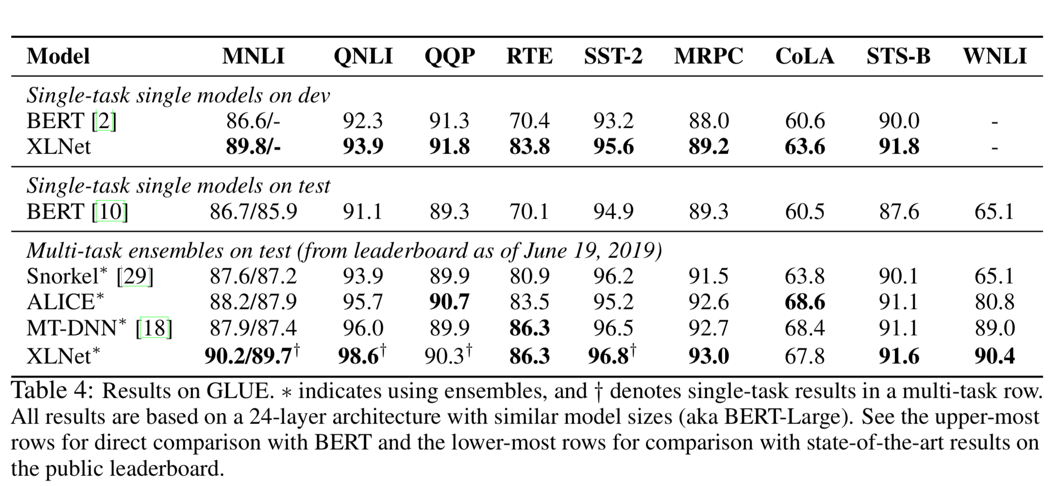
Bridge a gap between language modeling and pretraining because of the lack of the capability of bidirectional context modeling

1. What are the key differences in the method/approach (innovation)?

The authors create an innovation language modeling called “permutation language modeling” to achieve the goal. This language modeling produces bidirectional contexts and maintains benefits of autoregressive models. Traditional language models take the previous n tokens and predict the next one in a given context; however, permutation language models are trained to predict tokens in random order.

1. What are main achievements, significance?





1. What can be further improved?

N/A