Project Report

GROUP-8, PROJECT-P9

Automated query processing from passages

TEAM MEMBERS:

18CH10019- CHAITANYA SAI SAHU
18ME33017- SUSMIT KUMAR MISHRA
18ME33018- CHEKURI SRINIVASA VARMA

CONTENTS

1. Problem Statement
2. Dataset information
3. Modern Methods description and comparison
4. Implementation of modern methods
5. Results (Modern methods)
6. Traditional Methods description
7. Results (traditional methods)
8. Observations
9. Running the source code
10. Individual contributions
11. References

Problem statement

Given an input passage of 8000 words the task is to predict the answer text span in the passage. Questions and passages will be in natural language. Case-study on the existing state-of-the-art methods in use, for such a question-answering task and traditional methods previously used, their performance etc.should be performed. Pros and cons of the methods should be mentioned.Finally one rule-based traditional method and one modern method that employ machine learning strategies should be choosed and the reasoning behind choosing these methods to be explained and Implementation of these two methods should be done in Python.

Dataset

- We have chosen the first four chapters from Alice in Wonderland for testing our models.
- We compiled a list of ten questions across the passage and its answers in a text file.
- We have chosen ten diverse questions whose answers vary from a single word to multiple sentences.
- Both questions and passages are in natural language.

Case Study

Modern Models on Question Answering:

1. **BERT**
   BERT(Bidirectional Encoder Representation Representation from Transformers) makes use of Transformer, an attention mechanism that learns contextual relations between words in a text. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary. In contrast to directional models BERT reads the entire sentence of words at once (thus it is **non-directional**). As BERT is trained on 2.5 billion words from wikipedia it can be easily pretrained on diverse datasets and generalises well for various NLP tasks.
   BERT relies on the concept of transfer learning to learn unsupervised from a corpus of unlabeled data.

   The main advantages of BERT are it is Bi- directional, it captures complex syntactic meaning from natural language, and it can be easily fine-tuned for general datasets.
   BERT for question answering scores Test F1 **91.8** on the SQuAD 1.1 dataset. It has 24 layers of encoders, 1024 hidden layer size,16 attention heads and 340M parameters.

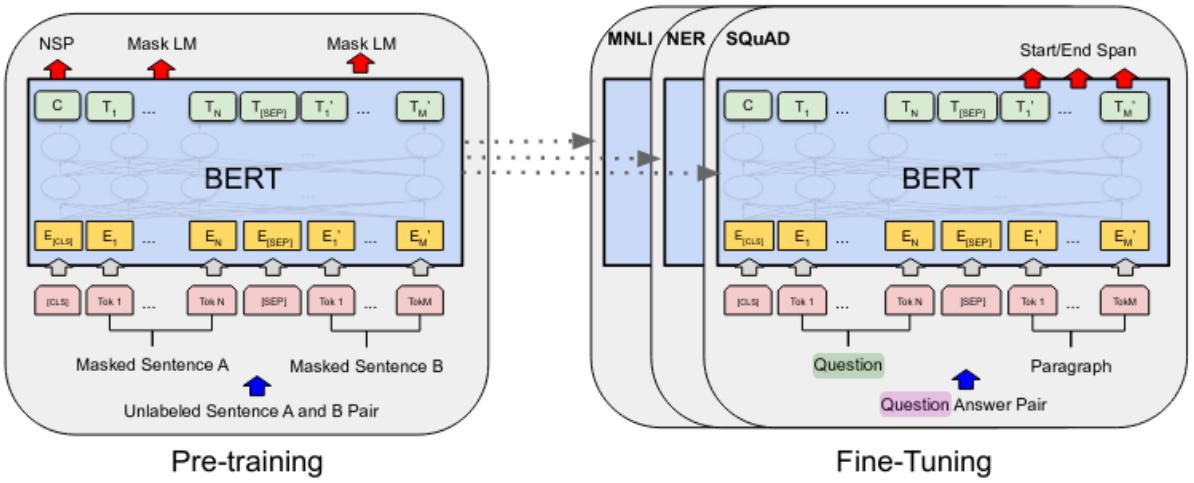


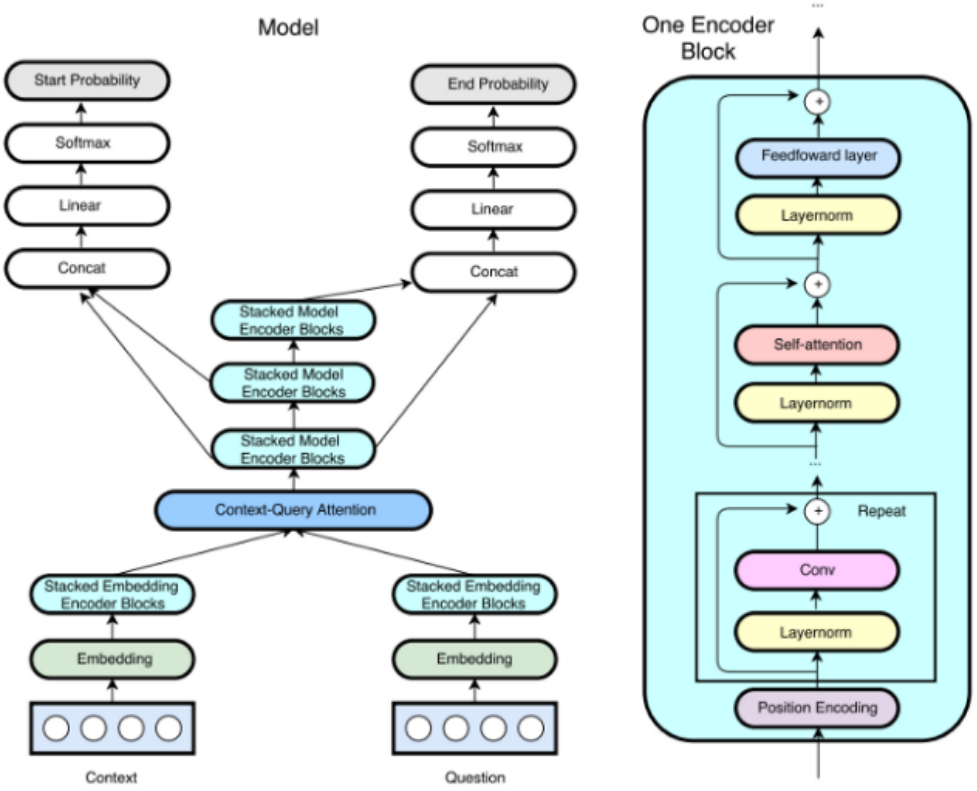
Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

---

2. **QANet**

QANet model is transformer version of using **BIDAF** (**BID**irectional **A**ttention **F**low) method.The structure of this model contains five major components:
1.Embedding layer: Generates and combines the word and character embeddings into a singular representation for every word.
2.Encoder layer: This layer uses one encoder block that has 4 convolutions. The output of this layer is of 128 dimensions (hidden_size) as the input is immediately converted to this dimension by a one-dimensional convolution.
3.Context-Query Attention layer: The output from the encoders comes to this layer. This layer combines context and question and produces a representation for each word in context.
4.Model encoder layer: This layer has 3 stacked encoder blocks that share weights and the number of encoder blocks within each stack is 7. Each block has 2 convolutions.
5.Output layer: Here,probabilities of each position in the context being start and end of the answer span are calculated. These probabilities are turned into a span consisting of a start index and an end index, indicating the location of the answer in the context paragraphs.

The main difference between QANet and other architectures is that for both the embedding and modeling encoders, we only use convolutional and self-attention mechanisms, and not RNNs. As a result, the QANet model is much faster, as it can process the input tokens in parallel.



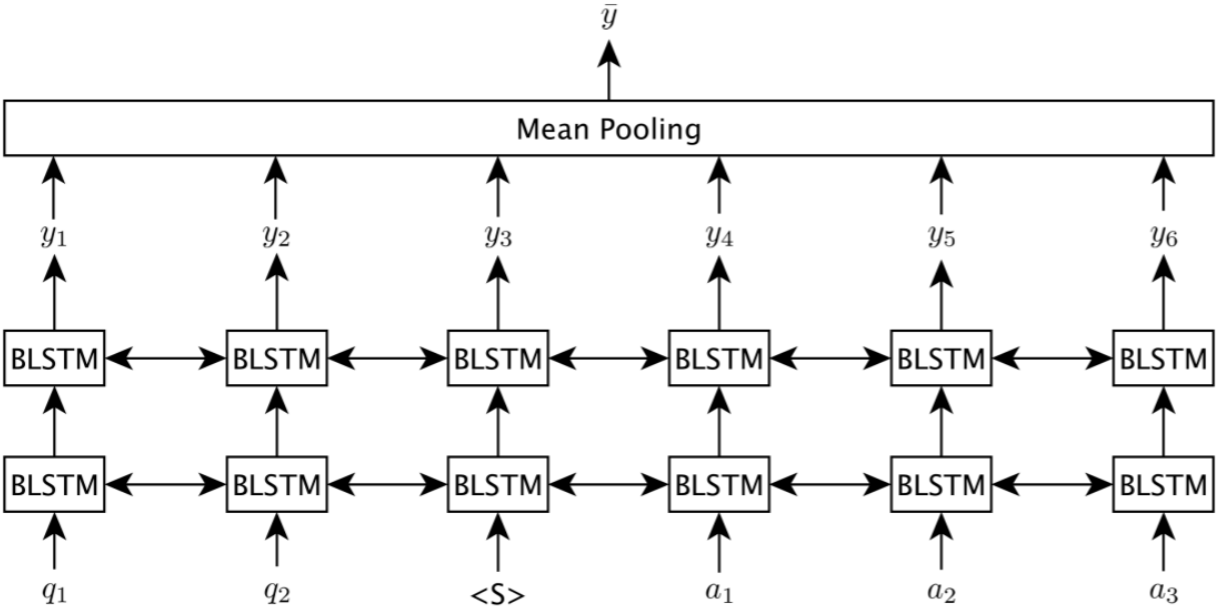
Architecture of the QANet model

There are fewer than 5M parameters present in QANet. On the SQuAD dataset, this model, trained with augmented data, achieves 84.6 F1 score on the test set.

---

3. **BiLSTM**

   Bidirectional LSTMs (BiLSTMs) enable additional training by traversing the input context twice (left-to-right right-to-left). A stacked BiLSTM neural network is resorted to attain the vector representation of the input sentence, which can effectively capture the semantics of the sentence. Wang and Nyberg [https://aclanthology.org/P15-2116.pdf] used a stacked BiLSTM network to sequentially read words from the question and answer sentences, which did not require any syntactic parsing or external knowledge resources such as WordNet. The model architecture consists of 3 high level major steps, 1) question analysis and retrieval of individual subpassages; 2) ranking and selecting of subpassages which contain the answer; 3) extracting and verifying the answer.



Typical Traditional Model on QA sentence relevance based on stacked BLSTMs

The words of input sentences are first converted to vector representations learned from word2vec tool. The question and answer sentences word vectors are sequentially read by BLSTM from both directions. In this way, the contextual information across words in both question and answer sentences is modeled by employing temporal recurrence in BLSTM.
The final output of each time step is the label indicating whether the candidate answer sentence should be selected as the correct answer sentence for the input question. This objective encourages the BLSTMs to learn a weight matrix that outputs a positive label if there is overlapping context information between two LSTM cell memories. During the test phase, we collect mean, sum and max poolings as features.

The major advantage of using a stacked BLSTM does not require any syntactic features or external resources and its ability to bidirectionally remember the contextual information.On SQuAD dataset, the BLSTM model achieves about 70% F1 score on test set.

F1-score comparison for modern question answering methods

MODEL	F1 SCORE (ON SQUAD dataset)
BERT	91.8
QANET	84.6
BiLSTM	70.4

We choose BERT over QANet and BiLSTM because of the following reasons:

- BERT is trained over 2.5 billion words from wikipedia and over 800 books, hence it can easily generalise to our dataset.
- With the help of the attention mechanism and the transformer based encoder decoder architecture it can capture complex relations in natural language.
- BERT is non-directional as it takes the entire sentence simultaneously in contrast to RNN and other sequence based models which process the words sequentially. As a result, these models are hard to parallelize and poor at retaining contextual relationships across long text inputs.
- BERT-large uses 340M parameters, much more than typical RNNs or BiLSTM models. Thus it will not overfit when fine-tuning and also give better performance.
- BLSTMs or stacked BLSTM networks tend to be computationally costly, because of the architecture, upon being a double LSTM. Also, with all the resources and time put into it, the accuracy metrics from BLSTM networks are visibly poor compared to other SOTA methods like BERT.

Dataset for fine tuning BERT

SQuAD Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.
On SQuAD, BERT achieves 93.2% F1 score (a measure of accuracy), surpassing the previous state-of-the-art score of 91.6% and human-level score of 91.2%. We used the huggingface transformers library to import the fine-tuned bertforquestionanswering weights.

Implementation

- We preprocessed the passage by removing punctuation and non alphanumeric characters. Our processed document contained ~300 sentences.
- The question and passage tokens were tokenized using BertTokenizer. We concatenated the query and the passage ids together as input to the model.
- The BERT model returned the vectors, start and end probabilities. We choose the max start probability for starting index and max end probability for ending index of answer.

- Since BERT can accurately answer the query if passage length is a few sentences and it cannot answer properly if we input a complete story with many paragraphs we decided to give multiple segments of sentences (like paragraphs) as input and store the outputs as candidate answers.
- Later we sort these outputs based on their starting and ending probabilities and pick the top one.
- Adam Learning rate for training 1e-4, Bath Size =32, gelu activation(Hendrycks and Gimpel, 2016), for fine tuning learning rate 1e-5, epochs 4. More details about the training and finetuning can be found BERT

Results

Our model with BERT fine-tuned on SQuAD predicted 8 out of 10 queries correctly.

	Query	Correct answer	Predicted answer
1	What is the colour of eyes of the white rabbit?	pink	white
2	What is the label on the jar on the shelf?	ORANGE MARMALADE	orange marmalade
3	What was on the solid glass three legged table?	golden key	a tiny golden key
4	Who was splendidly dressed and carrying a pair of white kid gloves in one hand and a large fan in the other?	white rabbit	the white rabbit
5	Who was afraid of being drowned in own tears?	alice	alice
6	What did Alice pull out of her pocket as prizes?	box of comfits	a box of comfits
7	What was written on the door of bright brass plate?	W RABBIT	w rabbit
8	Whom did Alice give one sharp kick?	BILL	bill
9	What happened when Alice swallowed one of the cakes?	She began shrinking	she began shrinking directly
10	What was the large blue caterpillar doing?	Sitting on the top with its arms folded, quietly smoking a long hookah, and taking not the smallest notice of her or of anything else.	digging for apples

Case Study on Traditional Methods

Traditional method in passage querying :

1. **Cosine similarity using count based approach :**
   - In this method we preprocess the passage in the same way as we did for the BERT model.
   - We used count and frequency of words to represent a sentence into a vector.
   - Similarly we got the vector representation for questions.
   - Using cosine similarity we compared the question and the answer vectors.
   - The sentence which matched most with the question was chosen as the answer.

   However when use used dependency parsing on the selected sentence to get span we were not able to find the correct span in majority of the cases so we decided to report the implementation only upto sentence level identification.

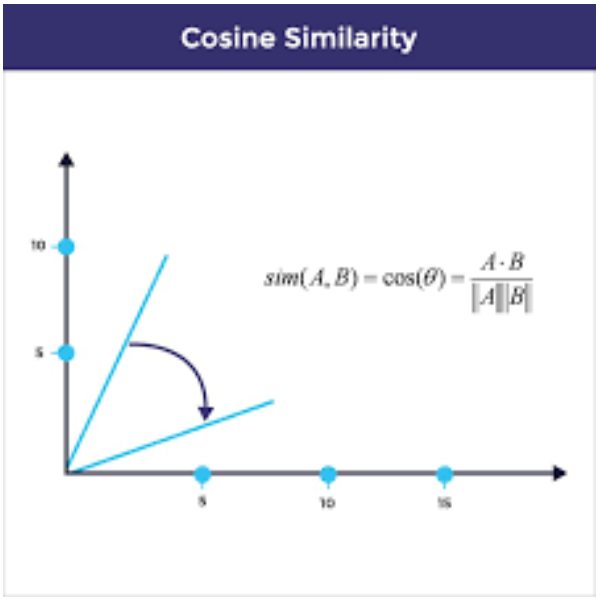


Figure showing the cosine similarity between two vectors in the 2-D space.

---

2. **Euclidean distance based approach :**
   - This method ranks the sentences in the corpus that are similar to the question asked i.e.number of words common between sentence and question regardless of their proximity within the sentence.
   - Firstly stop words are removed and words are stemmed using Porter algorithm from both sentence and question.
   - All sentences are represented as vectors by forming a word frequency dictionary. The vector size in this case is the length of the dictionary.
   - Then we compare each sentence with the question using the euclidean distance between two vectors formula. The answer lies in the sentence which is at least distance from the question.

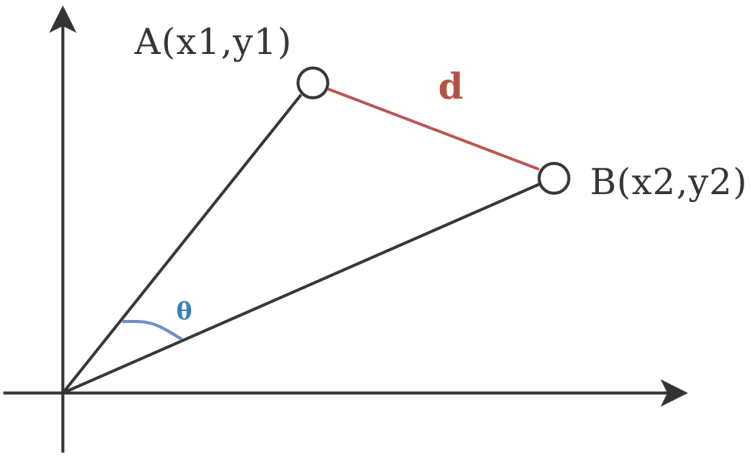


Figure showing the Euclidean distance(d) between two vectors (A,B).

In Euclidean distance approach length of the sentences are not taken into consideration. So in the Euclidean distance method large sentences will be at greater distance generally to the question but they still have a chance to contain the answer.Due to this demerit we are implementing Cosine similarity Based approach.

Results*

	QUERY	ANSWER	Predicted answer
1	What is the colour of eyes of the white rabbit?	pink	down the rabbithole alice was beginning to get very tired of sitting by her sister on the bank and of having nothing to do once or twice she had peeped into the book her sister was reading but it had no pictures or conversations in it and what is the use of a book thought alice without pictures or conversations so she was considering in her own mind as well as she could for the hot day made her feel very sleepy and stupid whether the pleasure of making a daisychain would be worth the trouble of getting up and picking the daisies when suddenly a white rabbit with pink eyes ran close by her
2	What is the label on the jar on the shelf?	ORANGE MARMALADE	the duchess
3	What was on the solid glass three legged table?	golden key	the little door was shut again and the little golden key was lying on the glass table as before and things are worse than ever thought the poor child for i never was so small as this before never
4	Who was splendidly dressed and carrying a pair of white kid gloves in one hand and a large fan in the other?	white rabbit	it was the white rabbit returning splendidly dressed with a pair of white kid gloves in one hand and a large fan in the other he came trotting along in a great hurry muttering to himself as he came o
5	Who was afraid of being drowned in own tears?	alice	i shall be punished for it now i suppose by being drowned in my own tea

6	What did Alice pull out of her pocket as prizes?	box of comfits	prizes alice had no idea what to do and in despair she put her hand in her pocket and pulled out a box of comfits luckily the salt water had not got into it and handed them round as prizes
7	What was written on the door of bright brass plate?	W RABBIT	but i would better take him his fan and gloves that is if i can find them as she said this she came upon a neat little house on the door of which was a bright brass plate with the name w rabbit engraved upon it
8	Whom did Alice give one sharp kick?	BILL	poor alice
9	What happened when Alice swallowed one of the cakes?	She began shrinking	if i eat one of these cakes she thought it is sure to make some change in my size and as it can not possibly make me larger it must make me smaller i suppose so she swallowed one of the cakes and was delighted to find that she began shrinking directly
10	What was the large blue caterpillar doing?	Sitting on the top with its arms folded, quietly smoking a long hookah, and taking not the smallest notice of her or of anything else.	she stretched herself up on tiptoe and peeped over the edge of the mushroom and her eyes immediately met those of a large blue caterpillar that was sitting on the top with its arms folded quietly smoking a long hookah and taking not the smallest notice of her or of anything else.

* These results are based on sentence level identification.

Observations:

1. Running BERT among passages was not working so we had to divide the document into smaller passages.
2. Traditional Methods was not able to correctly find spans.
3. Cosine similarity method was not taking context into consideration so it was not able to associate the pronoun with its corresponding proper noun.
4. BERT was able to understand distant context and this giving better spans and correct answers.
5. However BERT requires more time to execute than traditional methods.
6. Also the BERT model required more resources(such as GPU and memory) compared to traditional methods.

Running the Source Code

Method one (from notebook)
Open and execute the cells

Method two (from github)
In this method questions will be taken from question_list.txt. You can add or modify this file(locally) before running the main.py

```
git clone git@github.com:susmitmishra125/QA_NLP.git
cd QA_NLP
pip install -r requirements.txt
python main.py
```

Individual contributions*:

- **Chaitanya Sai Sahu**
  Case study on BiLSTM, Implementation of BERT based model, Case study on Cosine similarity based method
- **Susmit Kumar Mishra**
  Case study on BERT, Implementation of BERT based model, Implementation of cosine similarity based method.
- **Chekuri Srinivasa Varma:**
  Case study on QANet, Implementation of BERT based model, Case Study on Euclidean distance based method.

*** The report of these sections was also prepared by each of us respectively.**

References:

1. https://arxiv.org/pdf/1804.09541.pdf QANET: COMBINING LOCAL CONVOLUTION WITH GLOBAL SELF-ATTENTION FOR READING COMPREHENSION
2. https://web.stanford.edu/class/cs224n/reports/final_reports/report269.pdf QANet+: Improving QANet for Question Answering
3. https://www.sciencedirect.com/topics/computer-science/cosine-similarity Cosine similarity for text analysis
4. https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089 TF-IDF document ranking approach
5. arXiv:1810.04805v2 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
6. arXiv:1706.03762 Attention Is All You Need Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin
7. Different techniques to represent words as vectors (Word Embeddings)
8. https://aclanthology.org/P15-2116.pdf BiLSTM
9. Jurafsky & Martin Speech And Language Processing 2Ed 2007
