LLMs for Question Answering

Shah Avish Vipulkumar
Department of Electrical Engineering
IISc Bangalore
avishv@iisc.ac.in

1 Introduction

In recent years, the rapid advancements in Natural Language Processing (NLP) have transformed the way machines understand and respond to human language. One of the most impactful applications of NLP is Question Answering (QA), where the goal is to enable machines to accurately comprehend a given context and provide precise answers to user queries. QA systems are foundational for a variety of use cases, such as virtual assistants, search engines, and automated customer support.Traditional QA systems relied on rule-based or statistical methods, which often lacked flexibility and struggled with the complexities of natural language. However, the advent of Large Language Models (LLMs), such as BERT and its lighter variant DistilBERT, has revolutionized the QA landscape. These transformer-based models, pre-trained on vast amounts of text data, have demonstrated remarkable capabilities in understanding language semantics, context, and relationships, enabling state-of-the-art performance across numerous NLP benchmarks.

2 Problem Statement

Despite the impressive performance of modern LLMs, the task of building an effective and efficient Question Answering system remains challenging due to several reasons such as context understanding and relevance, efficiency in resource-constrained settings and domain specific generalization.In this paper, I have tried using a vanilla DistilBERT tuned for question answering and have customized it to see if I get better results.

3 System Description

3.1 BERT

The key innovation of the BERT model is applying bidirectional training of Transformers to language modelling. The entire input sequence though is read a once, so it would be more correct to de ne it non-directional. BERT is basically an Encoder stack of transformer architecture. BERTBASE has 12 layers in the Encoder stack while BERTLARGE has 24 layers in the Encoder stack. These are more than the Transformer architecture described in the original paper (6 encoder layers). BERT architectures (BASE and LARGE) also have larger feedforward-networks (768 and 1024 hidden units respectively), and more attention heads (12 and 16 respectively) than the Transformer architecture suggested in the original paper. Bert model is pre-trained on 2 tasks: Masked LM, where a percentage of the input tokens are masked and then predicted, and Next Sentence Prediction, in order to learn the relationship between two sentences, which is not directly modelled by language modelling

3.2 DistilBERT

DistilBERT is a small, fast, cheap and light Transformer model trained by distilling BERT base. It has 40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

BERTs performances as measured on the GLUE language understanding benchmark. Knowledge Distillation (KB) could be seen as a transfer learning technique, even though it has a di erent aim. Knowledge Distillation is a form of compression from a huge high precision model to a smaller one, without losing too much in generalization. It is a reinforcement learning technique that wants to reduce the dimension of huge models (like BERT) and the training time by transferring knowledge between two models. In particular, the Teacher-Student paradigm is composed of a small network, the Student, and a bigger one, the Teacher, which should be trained on the complete dataset. The training phase usually needs a considerable amount of time for models. The nal aim is to teach the Student how to simulate the Teachers behaviour, but with a model smaller in size and computation.

4 Models used

4.1 Vanilla DistilBERT

I have finetuned the vanilla version of DistilBERT for question answering. It uses a single linear layer, with no activation function, to reduce the 768 output dimension that comes from the DistilBERT backbone to 2.

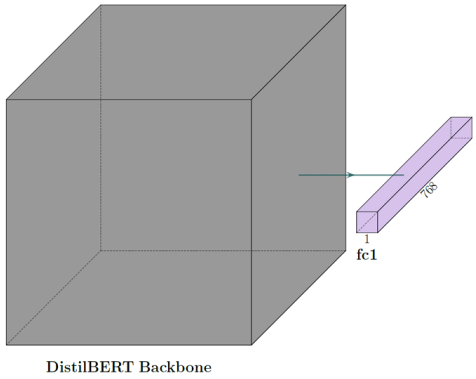

Figure 1: Vanilla DistilBERT for question answering

4.2 Custom DistilBERT

I have added 2 more layers on the top of DistilBERT and tried using different activation functions to understand which one works better for the task of question answering

4.2.1 2 layers with tanh activation function

In this version with tanh activation function, I have gone for a reduction of dimension, first from 768 to 384 and then from 384 to 2.

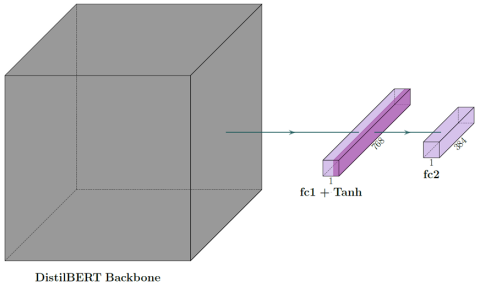

Figure 2: DistilBERT with tanh activation function

4.2.2 2 layers with GELU activation function

I have used GELU(Gaussian Error Linear Unit) activation function. In this model, the same input dimension is kept in both layers but skip connections are introduced.The output of DistilBERT is fed to the first layer and then added to its output before feeding everything to the second layer.

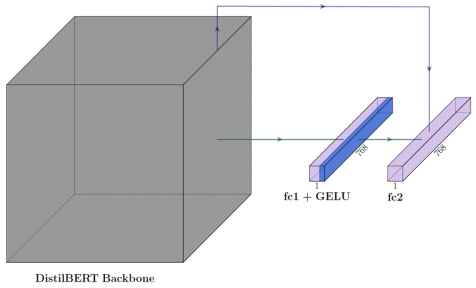

Figure 3: DistilBERT with GELU activation function

5 Results

I have used 2 different configurations and have presented the results of the model with GELU activation function only. For question answering, I have used F1 score and exact match(EM) as metrics.

Configuration	Learning Rate	Batch Size	Num of Epochs
configuration1	2.00E−04	32	4
configuration2	5.00E−04	16	4

Table 1: Hyperparameter configurations for training.

Configurations	VanillaQADistilBERT		QADistilBERT	
	F1(%)	EM(%)	F1(%)	EM(%)
configuration1	77.38	61.29	78.11	62.02
configuration2	77.30	60.85	77.92	61.37

Table 2: Comparison of models across different configurations.

VanillaDistilBERT		DistilBERT	
F1 (%)	EM (%)	F1 (%)	EM (%)
84.38	75.53	84.48	75.82

Table 3: Comparision on test set(SQUAD 1.1 dev set)

The link to the code and report is `https://github.com/avishshah02/DLNLP_term_paper`

6 Analysis

One of the important observations is that more open questions like 'why','if' and 'what' gets worse results than direct questions like 'when', 'who' and 'which'. There are 2 types of errors which were observed:'Lack of precise answers' and 'Misunderstanding of the question's type'

Error: Lack of precise answers

Context: The most widely used symbol is the flag of Greece, which features nine equal horizontal stripes of blue alternating with white representing the nine syllables of the Greek national motto Eleftheria i thanatos (freedom or death), which was the motto of the Greek War of Independence.

The blue square in the upper hoist-side corner bears a white cross, which represents Greek Orthodoxy. The Greek flag is widely used by the Greek Cypriots, although Cyprus has officially adopted a neutral flag to ease ethnic tensions with the Turkish Cypriot minority see ag of Cyprus).

Question: Have the people of Greece done anything to make the matter more palatable for the people of Turkey ?

Correct answer: Cyprus has officially adopted a neutral flag

Prediction: The Greek flag is widely used by the Greek Cypriots, although Cyprus has officially adopted a neutral flag to ease ethnic tensions with the Turkish Cypriot minority see flag of Cyprus).

Error: Misunderstanding of the questions type

Context: At over 5 million, Puerto Ricans are easily the 2nd largest Hispanic group. Of all major Hispanic groups, Puerto Ricans are the least likely to be pro cient in Spanish, but millions of Puerto Rican Americans living in the U.S. mainland nonetheless are fuent in Spanish. Puerto Ricans are natural-born U.S. citizens, and many Puerto Ricans have migrated to New York City, Orlando, Philadelphia, and other areas of the Eastern United States, increasing the Spanish-speaking populations and in some areas being the majority of the Hispanophone population, especially in Central Florida. In Hawaii, where Puerto Rican farm laborers and Mexican ranchers have settled since the late 19th century, 7.0 per cent of the islands people are either Hispanic or Hispanophone or both.

Question: Where are the biggest population of Puerto Ricans on the mainland?

Correct answer: many Puerto Ricans have migrated to New York City,Orlando, Philadelphia, and other areas of the Eastern United States

Prediction:over 5 million

Bibliography

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton.*Layer Normalization*.2016.arXiv(`https://arxiv.org/abs/1607.06450`)

[2] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov.*Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*.2019.arXiv(`https://arxiv.org/abs/1901.02860`)

[3] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. "Universal transformers. In 7th International Conference on Learning Representations".

[4] "DistilBERT for question answering"

[5] Dan Hendrycks and Kevin Gimpel.*Gaussian Error Linear Units (GELUs)*.2020.arXiv(`https://arxiv.org/abs/1606.08415`)

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter".2020.arXiv(`https://arxiv.org/abs/1910.01108`)