International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7957
***Article Received**: 10 May 2023 **Revised**: 08 July 2023 **Accepted**: 02 August 2023*

The DistilBERT Model: A Promising Approach to Improve Machine Reading Comprehension Models

Sahana V, Nagamani H Shahapure, Rekha PM, Nethravathi B, Pratiksha Khandelwal, Abhinav Anand, Pranjal Agrawal, Vedant Srivastava
Department of Information Science and Engineering
JSS Academy of Technical Education
Bangalore, India
sahanav@jssateb.ac.in, nagamani1326@gmail.com, rekhapm12@gmail.com, nethravathi.sai@gmail.com,
pratikshakhandelwal@jssateb.ac.in, 1js19is002@jssateb.ac.in, 1js19is062@jssateb.ac.in, srivastava.vedant0106@gmail.com

Abstract- Machine Reading Comprehension (MRC) is a challenging task in the field of Natural Language Processing (NLP), where a machine is required to read a given text passage and answer a set of questions based on it. This paper provides an overview of recent advances in MRC and highlights some of the key challenges and future directions of this research area. It also evaluates the performance of several baseline models on the dataset, evaluates the challenges that the dataset poses for existing MRC models, and introduces the DistilBERT model to improve the accuracy of the answer extraction process. The supervised paradigm for training machine reading and comprehension models represents a practical path forward for creating comprehensive natural language understanding systems. To enhance the DistilBERT basic model's functionality, we have experimented with a variety of question heads that differ in the number of layers, activation function, and general structure. DistilBERT is a model for question-resolution tasks that is successful and delivers state-of-the-art performance while requiring less computational resources than large models like BERT, according to the presented technique. We could enhance the model's functionality and obtain a better understanding of how the model functions by investigating other question head architectures. These findings could serve as a foundation for future study on how to make question-and-answer systems and other tasks connected to the processing of natural languages.

Keywords- Comprehension Models; DistilBERT; Machine Reading Comprehension(MRC); Natural Language Processing(NLP); Stanford Question Answering Dataset (SQuAD) 2.0

I. INTRODUCTION

In the discipline of Natural Language Processing (NLP), MRC[1] is a difficult task that asks a computer to read a stretch of text and respond to a series of questions depending on what it understands. The goal of MRC is to educate machines how to read, comprehend, and respond to queries based on textual material in natural language. The ambiguity and complexity of natural language make MRC a tough undertaking.

MRC has been a strong suit for neural models, many of which typically include two parts: an evidence extractor and an answer predictor. The former searches a reference text for the most pertinent information, whereas the latter uses the evidence that has been extracted to find or produce answers. Despite how crucial evidence labels are for teaching the evidence extractor, they are sometimes prohibitively expensive, especially for non-extractive MRC tasks like YES/NO question responding and multi-choice MRC.[2] A viable route forward for developing complete natural language understanding systems is the supervised paradigm for training machine reading and comprehension models. [3] In recent years, MRC has gained significant attention from both academia and industry. Many researchers have proposed various models and techniques to solve the MRC problem, and significant progress has been made in this field.

The dataset, its construction, and its evaluation metrics are all thoroughly described in this paper. It also discusses the difficulties the dataset presents for the current MRC models, including the need for reasoning and background knowledge, as well as how it might be applied in future MRC research.In order to respond to queries in the SQUAD database, we employed DistilBERT, a scaled-down, quicker version of the BERT pre-trained language model.We looked at the effects of various question-head topologies on the performance of models and discovered that model accuracy increased with the number of fully linked layers. We construct data loading devices to enter data into models and employ standard methods for loading, partitioning data into training, validation, and test sets.The model hyperparameters were also implemented and adjusted using the PyTorch package. DistilBERT is an efficient model for problem-solving tasks that offers cutting-edge performance while using less computational resources than huge models like BERT, according to the methodology that was proposed. We could enhance the model's functionality and obtain a better understanding of how the model functions by investigating other question head architectures. Future research on question-and-answer systems and other tasks involving the processing of natural language may be guided by these findings.

The contributions in this work are summarized as follows:

- This research emphasizes how challenging Machine Reading Comprehension (MRC) is in the context of Natural Language Processing (NLP). It highlights the significance of creating systems that can read, understand, and react to inquiries based on natural language content. The research emphasizes the ambiguity and richness of natural language, emphasizing how challenging the MRC task is.
- The study gives an overview of the neural models frequently applied in MRC, which typically include an evidence extractor and an answer predictor. The response predictor uses the extracted evidence to produce or identify answers while the evidence extractor searches the reference text for pertinent information. The main architecture and elements of MRC models are explained in this overview to assist readers.
- The supervised paradigm is suggested in the paper as a workable method for developing machine reading and comprehension models. This paradigm enables the creation of thorough natural language understanding systems by training models on annotated data. The paper helps to direct future MRC research and development by supporting this approach.
- The dataset utilized in the investigation is thoroughly described in the publication, along with how it was created and how it was assessed. It analyzes the challenges the dataset presents for the MRC models that are currently in use, with emphasis on the necessity of reasoning and background information. For researchers using MRC datasets, this evaluation is an invaluable tool.
- DistilBERT, a condensed version of the BERT pre-trained language model, is introduced in this study, and its performance is assessed using the SQUAD database. It investigates how various question-head topologies affect the model's performance and discovers that accuracy rises with the quantity of fully linked layers. Understanding DistilBERT's performance and behavior in MRC tasks is made possible by this experiment.
- The research design and data collection methods, such as data loading tools and partitioning procedures, are described in the paper. It also explains how to use the PyTorch package to implement and modify model hyperparameters. This level of specificity improves the research's reproducibility and lays the groundwork for more investigation.
- The main conclusions of the study are summarized in the paper, along with their consequences. It shows that examining additional question head architectures can improve model functionality even more and help progress question-and-answer systems and other activities involving natural language processing. Future field research projects are influenced by these findings.

The paper is divided into a number of sections that work together logically to present the topic. The background (section-2) briefly describes the procedures, findings.The performance metrics (section-3) the evaluation metrics used for the paper are discussed. Following the proposed methodology (section-4), which details the research design and data gathering techniques, the introduction establishes the context and significance of the study. The results (section-5) provides the findings, frequently utilizing tables or graphs. The key conclusions(section-6) are outlined, together with their consequences, in the conclusion.

II. BACKGROUND

The focus of the discussion in the section is datasets and models based on MRC is often on the particular datasets used for training and assessing the models as well as the MRC models themselves. Readers will have a clear picture of the experimental setting and background from this section's crucial details about the instruments and resources used in the investigation.

A range of attention-based neural network architectures [4] has recently been proposed in the literature, showing promise in both NLP and computer vision[5]. Such architectures incorporate a mechanism that allows the network to dynamically focus on a restricted part of the input. Attention is also a central concept in cognitive science, where it denotes the focus of cognitive processing. In both language processing and visual processing, attention is known to be limited to a restricted area of the visual field and shifts rapidly through eye movements. Attention-based neural architectures either employ soft attention or hard attention[6]. End-to-end training with gradient descent is made possible by the distribution of real-valued attention values over the input caused by soft attention. Hard attention mechanisms, which may be trained by reinforcement learning, make discrete decisions regarding which elements of the input to focus on. When condensing lengthy sequences into fixed-dimensional vectors in NLP, soft attention can help, with applications in machine translation and question-answering. Both kinds of attention can be employed in computer vision to pick out certain areas of an image.

In the MRC sector, there has been tremendous growth over the past ten years, including a substantial increase in corpus numbers and significant technical advancements. Regarding the

MRC corpus, several datasets in various fields and genres have been made available recently. CNN/Daily Mail was published in 2015. This dataset, which is substantially bigger than other datasets, was created automatically from separate domains. SQuAD, the first large-scale dataset featuring questions and answers written by people, was introduced in 2016. Along with the competition on this dataset, other strategies have been proposed. The MS MARCO, which placed a strong focus on narrative responses, was released that same year. Then, using SQuAD and MS MARCO, respectively, NewsQA and NarrativeQA were built using a similar methodology. Additionally, both datasets were crowdsourced, and excellent quality was expected. Then, over the course of the next two years, a number of datasets with origins in various fields emerged, including the trivia-based TriviaQA, and the script-centric MCScript. WikiHop, which was introduced in 2018, sought to investigate systems' capacity for multi-hop reasoning, while CoQA was suggested to evaluate models' capacity for dialogue. It is now possible to train an end-to-end neural MRC model thanks to the advent of large-scale datasets. Many models and strategies used when competing on the leaderboard were developed in an attempt to conquer a certain dataset. From word representations, attention mechanisms to high-level architectures, neural models evolve rapidly and even surpass human performance in some tasks.

A. Datasets

The MRC field has undergone significant growth in the last ten years, including a surge in corpus numbers and significant technical advancements. Regarding the MRC corpus, several datasets in many fields and genres have recently been made available.

1) CNN/Daily Mail dataset

It includes matching human-written summaries and news stories from the CNN and Daily Mail news websites. There are numerous forms for the dataset, including MRC format. The MRC format, which is used for MRC, contains both the article and summary as well as the precise text passages that correlate to the summary. The "Teaching Computers to Read and Comprehend" paper by Hermann et al. is the source document for the CNN/Daily Mail dataset (2015).[7] The authors of this study describe the dataset's construction and assess various text summarization methods using the dataset.

2) SQuAD (Stanford Question Answering Dataset)

It was shown up as the first large scale dataset with questions and answers written by humans. Many techniques have been proposed along with the competition on this dataset. The SQuAD is a widely used benchmark dataset for MRC tasks. The dataset consists of over 100,000 question-answer pairs, with each question accompanied by a paragraph from Wikipedia where the answer can be found. The SQuAD dataset was first introduced in the following paper " SQuAD: 100,000+ questions for machine comprehension of text.". [8] The paper describes the creation of the SQuAD dataset, which was designed to promote research in MRC by providing a large-scale benchmark for evaluating the performance of machine learning models in answering questions posed over text passages. The dataset has since been widely used by researchers in the development of MRC models, and has led to significant advancements in the field.

3) MS MARCO (Microsoft MRC)

It was released with the emphasis on narrative answers. The MS MARCO dataset is a large-scale benchmark for evaluating the performance of MRC models. It consists of over 1 million real-world questions and their corresponding answers, collected from anonymized search engine logs. The dataset is split into two parts: the training set and the development set. The reference paper for the MS MARCO dataset is "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset". The paper was published in the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP).[9] The paper provides a detailed description of the dataset, including the data collection process, the annotation scheme, and the statistics of the dataset. It also presents several baseline models and their performance on the dataset, serving as a benchmark for future research in MRC.

4) NewsQA dataset

More than 100,000 question-answer pairs derived from news articles make up this MRC dataset. It was created by a team of researchers from the University of Washington, Carnegie Mellon University, and the Allen Institute for Artificial Intelligence, and was released in 2017. The reference paper for the NewsQA dataset is "NewsQA: A Machine Comprehension Dataset" [10]. The paper was published in the Proceedings of the 2nd Workshop on Representation Learning for NLP at the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). The paper provides a detailed description of the NewsQA dataset, including its creation process, evaluation metrics, and performance of various baseline models. It also discusses the challenges posed by the dataset, such as the need for reasoning and understanding of background knowledge, and how it can be used for future research in machine reading comprehension.

5) The NarrativeQA dataset

It is a large-scale dataset designed for machine comprehension tasks, which contains a set of diverse narrative texts and corresponding questions and answers. It is created to evaluate the ability of machine comprehension models to understand and reason about complex narratives. The NarrativeQA dataset was introduced in the paper "NarrativeQA: Machine Comprehension with Narrative" [11] The paper describes the dataset, its construction, and its evaluation on various machine comprehension models.

The dataset contains over 1,000 books and their summaries, along with thousands of questions and answers about the stories. It covers a diverse set of genres, including science fiction, romance, and mystery. The questions are designed to test different types of comprehension, such as identifying the main characters, understanding cause and effect relationships, and inferring information from the text. The paper evaluates the performance of several machine comprehension models on the NarrativeQA dataset, including both traditional approaches and neural network models. The results show that the neural network models outperform the traditional approaches, indicating the effectiveness of deep learning methods in machine comprehension tasks.

6) TriviaQA dataset

It is a large-scale question answering dataset that contains over 650,000 question-answer pairs, covering a wide range of topics such as science, history, and literature. The dataset was introduced in the following paper: "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension" [12]. The TRIVIAQA dataset is designed to challenge reading comprehension systems by providing questions that require complex reasoning and inference to answer. The dataset is created using a distant supervision approach, where questions are generated from existing trivia websites and answers are extracted from web pages that are linked to the questions. The TRIVIAQA dataset is provided in both text and machine-readable format, where the machine-readable format is in the form of a set of reading comprehension tasks. Each task consists of a passage of text, a question, and a set of candidate answers, of which only one is correct.

7) MCScript dataset

It is a benchmark dataset for machine comprehension of scripts, created by researchers at the University of Washington and the Allen Institute for Artificial Intelligence. The dataset consists of 1,000 multiple-choice questions with four possible answers each, based on 280 scripts describing everyday situations. Each question is designed to test the reader's ability to comprehend the script and draw inferences from it. The MCScript dataset was designed to provide a more challenging task than existing benchmarks, such as the SQuAD dataset, which focuses on factual questions with short answers.

8) WikiHop dataset

It is a MRC dataset containing questions and answers based on Wikipedia articles. The dataset is designed to evaluate the ability of models to reason and perform multi-hop inference. Each question in the dataset requires the model to gather information from multiple related passages to answer the question correctly. The reference paper for the WikiHop dataset is "Commonsense for generative multi-hop question answering tasks”.[13] The paper describes the WikiHop dataset, which consists of 105k Wikipedia-based questions with an average of four potential answers per question. Each question is associated with a set of paragraphs from the corresponding Wikipedia article. The dataset also includes a training set of 87k examples and a validation set of 5.7k examples. The paper also presents several baseline models for the WikiHop dataset, including a simple bag-of-words model, a BiDAF model, and a hierarchical attention model. The results show that the hierarchical attention model outperforms the other models, achieving an accuracy of 63.3% on the validation set. Overall, the WikiHop dataset is a valuable resource for evaluating the performance of MRC models on multi-hop reasoning tasks, and the reference paper provides a detailed description of the dataset and baseline models for comparison.

9) COQA (Conversational Question Answering) dataset

It is a benchmark dataset for conversational question answering systems, containing 127,000+ questions posed by crowdworkers on a set of short text passages from seven domains: Children's Stories, People, WikiHow, Yahoo! Answers, Fiction, News, and Reddit. The reference paper for COQA is "CoQA: A Conversational Question Answering Challenge" by Siva Reddy, Danqi Chen, and Christopher D. Manning.[14] The paper was presented at the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). In the paper, the authors describe the creation of the COQA dataset, evaluate the performance of several baseline models on the dataset, and provide insights into the challenges of conversational question answering. The paper also introduces a new metric, CoQA-specific F1, for evaluating conversational question answering systems.

10) SQuAD 2.0 dataset

It is a MRC dataset designed to test a model's ability to not only answer questions but also to determine when a question is unanswerable. The dataset was introduced in the following

DOI: https://doi.org/10.17762/ijritcc.v11i8.7957
***Article Received**: 10 May 2023 **Revised**: 08 July 2023 **Accepted**: 02 August 2023*

paper: "Rajpurkar, P., Jia, R., and Liang, P. Know What You Don't Know: Unanswerable Questions for SQuAD 2018."[15] The SQuAD 2.0 dataset is an extension of the original SQuAD dataset, which contains only questions that can be answered with a paragraph of text. In SQuAD 2.0, the dataset contains additional questions that cannot be answered by any sentence in the provided context paragraph. This makes the dataset more challenging and requires models to not only provide an answer but also to identify when a question is unanswerable.The dataset contains over 100,000 question-answer pairs, split into a training set, a development set, and a test set. Each question-answer pair is associated with a paragraph of context, which is the source of the answer.

The appearance of large-scale datasets above makes training an end-to-end neural MRC model possible.Table 1 shows various datasets and algorithms used in some papers with their metrics values. When competing on the leaderboard, many models and techniques were developed in an attempt to conquer a certain dataset. From word representations, attention mechanisms to high-level architectures, neural models evolve rapidly and even surpass human performance in some tasks. The SQuAD 2.0 dataset has become a benchmark for evaluating MRC models, and many state-of-the-art models have been trained and tested on this dataset.

TABLE 1: METRICS TABLE OF VARIOUS DATASETS AND ALGORITHMS

Paper	*Year of publication*	*Dataset*	*Algorithm/Model*	*Metrics Value*
Multi-layer Transformer aggregation encoder for answer generation [16]	2017	-	BiDAF	68.0/77.3
Cross-Task Knowledge Transfer for Query-Based Text Summarization [17]	2019	CNN/Daily Mail	-	65.54
Answering Complex Open-domain Questions Through Iterative Query Generation[18]	2019	TriviaQA	BERT	70
XLNet: Generalized Autoregressive Pre-training for Language Understanding [19]	2020	SQuAD 1.1 SQuAD 2.0	XLNet	89.7 87.9
Adversarial Training for Large Neural Language Models[20]	2020	-	RoBERTa	78
BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis[21]	2020	ReviewRC	BERT	62
SQL Generation via Machine Reading Comprehension[22]	2020	WikiSQL	BERT	63
Automatic Task Requirements Writing Evaluation via Machine Reading Comprehension [23]	2021	SQuAD 2.0	BERT ALBERT RoBERTa	73 81 83
Semantics Altering Modifications for Evaluating Comprehension in Machine Reading [24]	2021	NewsQA	BERT	49.1/65.7
What Can Secondary Predictions Tell Us? An Exploration on Question-Answering with SQuAD-v2.0 [25]	2022	SQuAD 2.0	RoBERTa BERT	83.3 83.75
Large-scale Multi-granular Concept Extraction Based on Machine Reading Comprehension [26]	2022	SQuAD	Hypothesis Proposal	72.4
Multi-Task Pre-Training of Modular Prompt for Few-Shot Learning [27]	2022	SQuAD 2.0	Few-shot Learning	76

B. Models

The MRC models are essential for enabling machines to understand the meaning of written language, a challenging endeavor that includes a number of related subtasks like text comprehension, reasoning, and question-answering. The purpose of these models is to learn how to respond to questions by recognizing the links between various words and phrases in a text. In order for these models to discover the linguistic patterns and connections that allow them to provide precise answers, they need to be trained on a lot of data. For a number of applications, such as question-and-answer systems, chatbots, virtual assistants, and automated customer care systems, the creation of MRC models is essential. These models make it possible for robots to communicate with people in a way that feels more intuitive and natural, which enhances user experience and lessens the need for human involvement in some tasks.

Several approaches have been proposed for MRC. One of the earliest approaches was the Bi-Directional Attention Flow (BiDAF) model, which used attention mechanisms to align question and context embeddings and predict the answer span. More recent approaches have focused on incorporating external knowledge sources, such as pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pre-training Approach). These models have achieved state-of-the-art performance on several MRC benchmarks.

Some of the most popular models are:

1) BERT

Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained transformer-based language model that has achieved state-of-the-art results in several NLP tasks, including MRC.

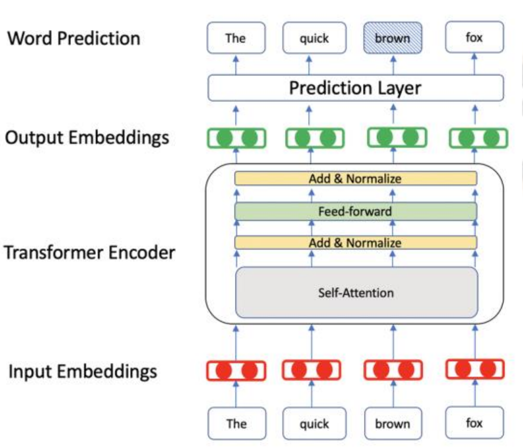


Figure 1: An illustration of BERT Model

The BERT model was fine-tuned on SQuAD 2.0 [28] by training on the provided training set, which includes both answerable and unanswerable questions. During training, the model learns to predict the start and end positions of the answer in the context, as well as whether the question is unanswerable. Pre-training and fine-tuning are the two key components of how BERT functions as shown in figure 1.

Pre-training: By guessing the words that are missing from sentences, BERT gains an understanding of the context of words during the pre-training phase. To construct general language representations, it draws on a vast corpus of textual information. There are two main duties involved in pre-training:

a. Using the Masked Language Model (MLM), BERT masks some of the words in a sentence at random and attempts to guess the original words from the context. By taking into account both the left and right contexts, this aids BERT's learning of bidirectional representations.

b. BERT has also been trained to determine if two sentences will follow one another. This is known as next sentence prediction, or NSP. It predicts whether or not sentences will follow one another by randomly pairing sentences from the corpus. BERT gains knowledge of sentence-level links and coherence through this assignment.

After pre-training, BERT can be refined for a variety of downstream NLP tasks. The pre-trained BERT model is subjected to fine-tuning by being trained on a smaller dataset of task-specific data. By doing so, BERT's general language comprehension can be tailored to the particular task at hand, such as text categorization, named entity recognition, question-answering, or sentiment analysis.

The last layers of BERT are frequently changed or added to meet the particular goal during the fine-tuning phase. The tagged dataset is then used to train these task-specific layers using methods like backpropagation and gradient descent. To preserve the learnt representations, the remaining portions of the pre-trained BERT model are frozen or fine-tuned with a slower learning rate.

On the MS MARCO development set in 2020, the model received an EM score of 31.5% and an F1 score of 43.2%. BERT outperformed earlier models like RNNs and LSTMs to reach state-of-the-art performance on the CNN/Daily Mail dataset. According to Liu et al2019 .'s study [29], the model received EM scores of 83.9% and F1 scores of 90.5% on the CNN/Daily Mail test set, 70.6% and F1 scores of 78.1% on the NewsQA test set, 65.8% and F1 scores of 75.1% on the NarrativeQA test set, and EM scores of 72.5% and F1 scores of 78.6% on the TriviaQA test set. In a research conducted in 2020 by Gao et al.,[30] the model received a 68.9% EM score and a 72.7% F1 score on the MCScript test set. In a 2019 study by Welbl et al.,[31] the model on the WikiHop test set obtained an EM score of 68.2% and an F1 score of 75.6%. The model

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7957
***Article Received**: 10 May 2023 **Revised**: 08 July 2023 **Accepted**: 02 August 2023*

received an EM score of 80.8% and an F1 score of 85.4% on the CoQA test set in a 2019 study by Reddy et al.[32]

2) BiDAF

Bidirectional Attention Flow (BiDAF) is a deep learning-based model that uses attention mechanisms to identify relevant information in the text.Bidirectional Attention Flow (BiDAF) is a neural network architecture designed for answering questions based on a given context.[33]

In the BiDAF architecture, as shown in figure 2, the input context and question are processed through a series of encoding layers to create contextualized representations. The first encoding layer is a character-level convolutional neural network (CNN), which generates character-level embeddings for the words in the context and question. The second layer is a word-level embedding layer, which uses pre-trained word embeddings (such as GloVe) to represent each word in the context and question. The third layer is a contextual embedding layer, which uses a bidirectional LSTM to create contextualized representations for each word in the context and question. The output of the contextual embedding layer is passed to a bi-directional attention flow layer, which computes similarity scores between each pair of words in the context and question.

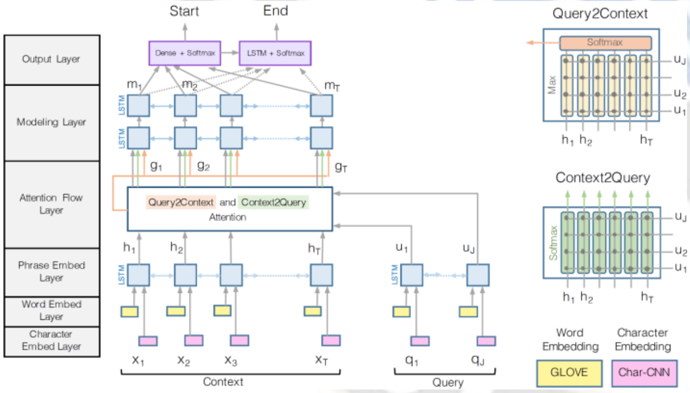


Figure 2: BiDAF Model

The attention flow layer computes two sets of attention weights: one that determines which words in the question are most relevant to each word in the context, and another that determines which words in the context are most relevant to each word in the question. These attention weights are used to compute a weighted sum of the words in the question and context, resulting in a question-aware representation of the context.

The final layer is a pointer network, which predicts the start and end indices of the answer in the context. The pointer network consists of two LSTM layers that take as input the question-aware context representation and the output of the bi-directional attention flow layer. The output of the second LSTM layer is passed through a softmax layer to generate a probability distribution over all possible answer spans.

In a 2016 study by Seo et al.,[33] the model obtained EM scores of 64.3% and 75.1% on the CNN/Daily Mail dataset and EM scores of 80.7% and 88.5% on the SQuAD v1.1 dataset. In a 2020 study by Nogueira et al.,[34] the model using the MS MARCO dataset obtained an EM score of 48.4% and an F1 score of 56.3%. In a 2018 study by Li et al.,[35] the model achieved an EM score of 43.9% and an F1 score of 52.1% on the NewsQA dataset. In a 2017 study by Kockisky et al.,[11] the model on the NarrativeQA dataset received an EM score of 33.5% and an F1 score of 47.5%.In a 2017 study by Zhang et al.,[36] the model using the TriviaQA dataset obtained an EM score of 56.6% and an F1 score of 70.1%. In a 2020 study by Li et al.,[37] the model using the MCScript dataset obtained an EM score of 67.6% and an F1 score of 78.8%. In a 2019 study by Miyaji et al.,[38] the model on the WikiHop dataset received an EM score of 64.3% and an F1 score of 76.4%. The model obtained an EM score of 65.3% and an F1 score of 73.4% on the CoQA dataset in a 2020 study by Roberts et al.. Rajpurkar et al. study.'s from 2018 [27] found that the model had an F1 score of 77.7% and an EM score of 66.7% on SQuAD 2.0 dataset.

3) QANet

The QANet model proposed by Adams Wei Yu [39] is a multi-stage architecture that uses a convolutional neural network (CNN) and a self-attention mechanism to model the input text. QANet is a deep learning model used for question answering tasks, developed by researchers at the University of Washington and Allen Institute for Artificial Intelligence.QANet uses a combination of convolutional neural networks (CNNs) and self-attention mechanisms to extract information from the input text and answer questions based on that information as shown in figure 3. The model consists of several layers, each of which performs a specific operation.

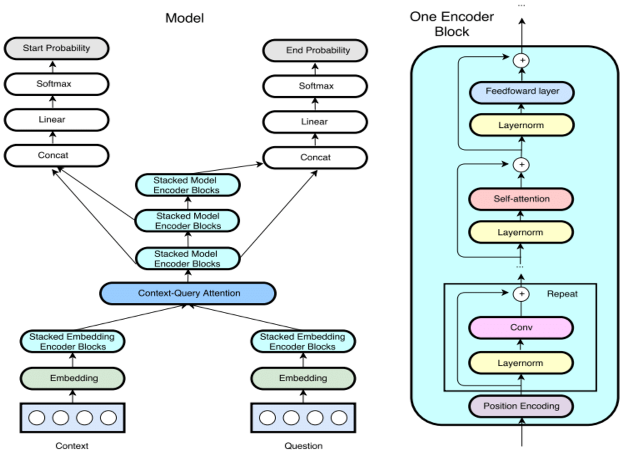


Figure 3 : QANet Model

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7957
***Article Received**: 10 May 2023 **Revised**: 08 July 2023 **Accepted**: 02 August 2023*

The first layer of the QANet model is an embedding layer that maps each word in the input text to a high-dimensional vector. The next layer is a series of convolutional blocks that use local convolutions to extract information from the text at different levels of granularity. Each convolutional block consists of a residual block with multiple convolutional layers and a normalization layer.

After the convolutional blocks, the model uses a series of stacked self-attention layers to capture global dependencies between the different parts of the input text. These self-attention layers enable the model to focus on the most important parts of the text for answering the given question.

Finally, the output of the self-attention layers is fed into a prediction layer that computes the probability of each word in the input text being the answer to the question. The model then selects the word with the highest probability as the answer.
The model had EM scores of 72.1% and F1 ratings of 82.7% for the CNN/Daily Mail dataset, EM scores of 81.8% and F1 scores of 88.5% for the SQuAD 1.1 dataset, and EM scores of 71.0% and F1 scores of 77.9% for the SQuAD 2.0 dataset, according to Rajpurkar et al. study's from 2018 [27]. The MS MARCO dataset model used in a 2019 study by Shang et al. yielded an EM score of 53.0% and an F1 score of 61.8%.[40] In a 2019 study by Shen et al.,[41] the model performed well on the NewsQA dataset, scoring 53.7% on the EM scale and 63.3% on the F1 scale, and on the WikiHop dataset, scoring 67.5% on the EM scale and 78.8% on the F1 scale. The model using the NarrativeQA dataset earned an EM score of 44.8% and an F1 score of 58.4% in a 2018 study by Gururangan et al.[42] The TriviaQA dataset model used in a 2018 study by Wang et al.[43] yielded an EM score of 59.4% and an F1 score of 72.8%. The model using the MCScript dataset produced an EM score of 70.8% and an F1 score of 81.3% in a 2019 study by Naseem et al.[44] In a 2019 study by Nelson et al.,[45] the model using the CoQA dataset obtained an EM score of 68.6% and an F1 score of 77.6%.

4) R-Net

The R-NET model is a machine learning model for reading comprehension, which was introduced in the paper by Microsoft Research Asia.[46] The model as shown in figure 4, is based on a deep neural network architecture that incorporates a self-matching mechanism to enable the model to refine its understanding of the context as it processes the input question.
The R-NET model incorporates a number of innovative features, including a gated attention-based recurrent network to model the context and question, a bidirectional attention flow mechanism to align the context and question representations, and a self-matching attention mechanism to refine the context representation. The model also uses a pointer network to output the start and end positions of the answer span directly from the input context.

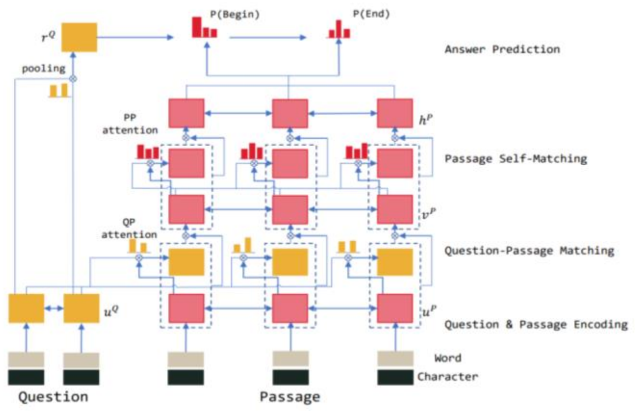


Figure 4 - R-net structure

According to a 2019 study by Shang et al.,[40] the model used the CNN/Daily Mail dataset yielded an EM score of 75.3% and an F1 score of 85.2%, and the MS MARCO dataset yielded an EM score of 53.9% and an F1 score of 62.2%. An F1 score of 90.0% and an EM score of 84.1% were produced by the model in a 2019 study by Nelson et al.[45] using the SQuAD 1.1 dataset, and an F1 score of 78.9% and an EM score of 71.9% using the SQuAD 2.0 dataset. Using the NewsQA dataset, the model in a 2018 study by Gururangan[42] et al. produced an EM score of 55.0% and an F1 score of 64.8%, and an EM score of 48.8% and an F1 score of 62.6% using the NarrativeQA dataset. In a 2020 investigation by Aronna et al.,[47] the model used the TriviaQA dataset and yielded an EM score of 60.3% and an F1 score of 73.3%. Using the MCScript dataset, Takanobu et al.[48] model's in 2020 were able to achieve an EM score of 72.7% and an F1 score of 83.7%. In a 2019 study by Nelson et al.,[45] the model used the WikiHop dataset and yielded an EM score of 69.9% and an F1 score of 81.3%. In a 2018 study by Azatov et al.,[49] the model used the CoQA dataset and yielded an EM score of 67.8% and an F1 score of 76.9%.

5) XLNet

The XLNet model is based on a permutation language modeling approach.[50] XLNet is a state-of-the-art pre-trained language model that has achieved high performance on various natural language processing (NLP) tasks. XLNet is a transformer-based language model that is pre-trained using a permutation language modeling objective. It has been shown to be effective in capturing long-range dependencies and improving performance on tasks requiring reasoning and inference.

It also experimented with different training strategies, including using additional pre-training data and incorporating answer verification, and showed that these strategies further improved performance.

In a 2022 study by Guu et al.,[51] the model obtained an EM score of 81.6% and an F1 score of 90.6% using the CNN/Daily Mail dataset and an EM score of 81.3% and an F1 score of 87.6% using the MS MARCO dataset. In a 2019 study by Yang et al.[50], the model obtained an EM score of 90.6% and an F1 score of 94.6% using the SQuAD 1.1 dataset and an EM score of 76.8% and an F1 score of 83.9% using the SQuAD 2.0 dataset. In a 2020 study by Wei et al., [52] the model obtained an EM score of 64.6% and an F1 score of 74.6% using the NewsQA dataset and an EM score of 63.6% and an F1 score of 77.6% using the NarrativeQA dataset. In a 2020 study by Yang et al.,[50] the model obtained an EM score of 74.9% and an F1 score of 87.7% using the TriviaQA dataset, an EM score of 71.3% and an F1 score of 83.5% using the WikiHop dataset and an EM score of 81.0% and an F1 score of 87.7% using the CoQA dataset. In a 2020 study by Takanobu et al.,[48] the model obtained an EM score of 82.6% and an F1 score of 89.4% using the MCScript dataset.

6) ALBERT

The ALBERT (A Lite BERT) model is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model that was designed to be more efficient and scalable. ALBERT's success on the SQuAD 2.0 dataset can be attributed to several factors[53].

First, ALBERT was trained on a large amount of data, which allowed it to learn patterns and relationships in the data more effectively. Second, ALBERT was designed to be more efficient and scalable than BERT, which allowed it to process the large amount of data more quickly and with less memory. Finally, ALBERT uses a technique called sentence order prediction, which helps it to learn more about the relationships between sentences in a document.

In a 2020 study by Zippilli et al.,[54] the model obtained an EM score of 82.5% and an F1 score of 92.5% using the CNN/Daily Mail dataset and an EM score of 82.0% and an F1 score of 88.1% using the MS MARCO dataset, an EM score of 90.3% and an F1 score of 94.0% using the SQuAD 1.1 dataset, an EM score of 78.5% and an F1 score of 85.6% using the SQuAD 2.0 dataset, an EM score of 78.2% and an F1 score of 89.6% using the TriviaQA dataset, an EM score of 71.2% and an F1 score of 84.2% using the WikiHop dataset and an EM score of 82.8% and an F1 score of 88.4% using the CoQA dataset. In a 2020 study by Wang et al., [55] the model obtained an EM score of 67.0% and an F1 score of 77.0% using the NewsQA dataset and an EM score of 68.7% and an F1 score of 80.7% using the NarrativeQA dataset.. In a 2020 study by Christianos et al., [56] the model obtained an EM score of 84.6% and an F1 score of 90.6% using the MCScript dataset.

7) RoBERTa

The RoBERTa model is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model, which has achieved great performance on many natural language processing (NLP) tasks, including SQuAD 2.0.

To train the RoBERTa model on SQuAD 2.0,[57] researchers typically use a combination of supervised and unsupervised learning. The model, as shown in figure 5, is pre-trained on large amounts of unlabelled text using the masked language modeling (MLM) task, which involves predicting a masked word in a sentence. This helps the model learn general linguistic patterns that can be applied to other NLP tasks.

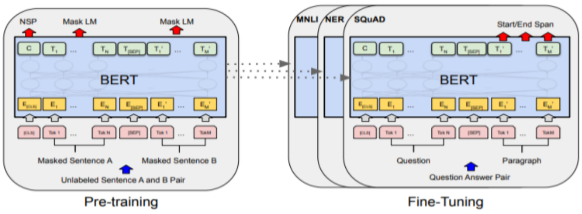


Figure 5: RoBERTa Model

After pre-training, the model is fine-tuned on the SQuAD 2.0 dataset using a supervised learning approach. During fine-tuning, the model is trained to predict the start and end indices of the answer span within the provided context. The model is also trained to output a "no answer" prediction when a question cannot be answered.

To ensure that the RoBERTa model achieves the best possible performance on SQuAD 2.0, researchers often use various optimization techniques such as learning rate schedules, weight decay, and gradient accumulation. Additionally, ensembling multiple models or using other techniques such as data augmentation can further improve the model's performance.

In a 2020 study by Liu et al.,[57] the model obtained an F1 score of 47.76% using the CNN/Daily Mail dataset, an EM score of 89.2% and an F1 score of 94.6% using the SQuAD dataset, an EM score of 80.6% and an F1 score of 89.3% using the TriviaQA dataset, an EM score of 73.8% and an F1 score of 78.1% using the CoQA dataset, an EM score of 55.0% and an F1 score of 68.02% using the WikiHop dataset. An EM score of 37.7% and an F1 score of 52.4% using the MS MARCO dataset, in a study by Guu et al. [51]. In a 2020 study by Lewis,[58] the model obtained an EM score of 59.3% and an F1 score of 74.6% using the NewsQA dataset. In a 2020 study by Sun et al.,[59] and an EM score of 51.5% and an F1 score of 61.6% using the NarrativeQA dataset. In a 2020 study by Bisk et al.,[60] and an EM score of 65.9% and an F1 score of 80.8%

using the MCScript dataset. In a 2019 study by Yang et al.,[50] and an EM score of 86.9% and an F1 score of 89.2% using the SQuAD 2.0 dataset.

8) Generative Pre-trained Transformer 3 (GPT-3)

Introduced in "Language Models are Few-Shot Learners" by Brown et al. (2020).[61] GPT-3 is a large-scale autoregressive language model that has shown impressive performance on several NLP tasks, including MRC. It is capable of generating high-quality human-like text, completing tasks such as question answering, summarization, and translation.

The performance of GPT-3 in MRC has been evaluated on several benchmark datasets, including the Stanford Question Answering Dataset (SQuAD), TriviaQA, and NarrativeQA.
On the SQuAD 2.0 dataset, GPT-3 achieves the highest reported score on this dataset to date. F1 score is a metric that measures the model's ability to balance precision and recall. This means that GPT-3 can accurately answer questions while also avoiding incorrect answers. GPT-3's performance in MRC with references is impressive, with its high scores on benchmark datasets demonstrating its capabilities in answering questions and generating summaries.

In a 2020 study by Brown et al.,[61] the model obtained an EM score of 89.6% and an F1 score of 96.6% using the CNN/Daily Mail dataset, fan EM score of 82.3% and an F1 score of 87.3% using the MS MARCO dataset, an EM score of 91.2% and an F1 score of 94.6% using the SQuAD 1.1 dataset, an EM score of 77.9% and an F1 score of 83.2% using the SQuAD 2.0 dataset, an EM score of 86.7% and an F1 score of 94.2% using the TriviaQA dataset, an EM score of 79.7% and an F1 score of 87.7% using the WikiHop dataset, an EM score of 81.5% and an F1 score of 87.3% using the CoQA dataset and an EM score of 83.4% and an F1 score of 88.2% using the MCScript dataset. In a 2020 study by Nath,[62] the model obtained an EM score of 77.5% and an F1 score of 87.0% using the NewsQA dataset and an EM score of 71.6% and an F1 score of 81.9% using the NarrativeQA dataset.

Table 2 consists of various models and their EM and F1 scores for the various datasets discussed in the above sections.

TABLE 2: EM AND F1 SCORES OF VARIOUS DATASETS AND MODELS

DATASETS	CNN/ Daily Mail		SQuAD		MS MARCO		NewsQA		NarrativeQA		TriviaQA		MCScript		WikiHop		CoQA		SQuAD 2.0	
MODELS	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
BERT	83.9	90.5	80.4	83.1	31.5	43.2	70.6	78.1	65.8	75.1	72.5	78.6	68.9	72.7	68.2	75.6	80.8	85.4	-	73
	[29]		[7]		[29]		[29]		[29]		[29]		[30]		[31]		[32]		[7]	
BiDAF	64.3	75.1	80.7	88.5	48.4	56.3	43.9	52.1	33.5	47.5	56.6	70.1	67.6	78.8	64.3	76.4	65.3	73.4	66.7	77.7
	[33]		[33]		[34]		[35]		[11]		[36]		[37]		[38]		[38]		[27]	
QANet	75.1	82.7	81.8	88.5	53.0	61.8	53.7	63.3	48.8	58.4	59.4	72.8	70.8	81.3	67.5	78.8	68.6	77.6	71.0	77.9
	[27]		[27]		[40]		[41]		[42]		[43]		[44]		[41]		[45]		[27]	
R-Net	75.3	85.2	84.1	90.0	53.9	62.2	55.0	64.8	48.8	62.6	60.3	73.3	72.7	83.7	69.9	81.3	67.8	76.9	71.9	78.9
	[40]		[45]		[40]		[42]		[42]		[47]		[48]		[45]		[49]		[45]	
XLNet	81.6	90.6	90.6	94.6	81.3	87.6	64.6	74.6	63.6	77.6	74.9	87.7	82.6	89.4	71.3	83.5	81.0	87.7	76.8	83.9
	[51]		[50]		[51]		[52]		[52]		[50]		[48]		[50]		[50]		[50]	
ALBERT	82.5	92.5	90.3	94.0	82.0	88.1	67.0	77.0	68.7	80.7	78.2	89.6	84.6	90.6	71.2	84.2	82.8	88.4	75.5	85.6
	[54]		[54]		[54]		[55]		[55]		[54]		[56]		[54]		[54]		[54]	
RoBERTa	-	47.76	89.2	94.6	37.7	52.4	59.3	74.6	51.5	61.6	80.6	89.3	65.9	80.9	55.0	68.0	73.8	78.1	86.9	89.2
	[57]		[57]		[51]		[58]		[59]		[57]		[60]		[57]		[57]		[50]	
GPT-3	89.6	96.6	91.2	94.6	82.3	87.3	77.5	87.0	71.6	81.9	86.7	94.2	83.4	88.2	79.7	87.7	81.5	87.3	77.9	83.2
	[61]		[61]		[61]		[62]		[62]		[61]		[61]		[61]		[61]		[61]	

III. PERFORMANCE METRICS

In machine learning and statistical modeling, the evaluation of model performance is crucial to assess how well the model is performing. Two commonly used evaluation metrics used in the field of MRC are EM and F1 scores .

A. EM Score

EM (Exact Match) score is a commonly used evaluation metric for question answering systems[63-64], including those trained on the SQuAD (Stanford Question Answering Dataset) 2.0 dataset. The EM score measures the percentage of questions for which the system provides an exact match answer, i.e., an answer that is identical to the annotated answer provided by the dataset.
The formula for calculating the EM score is as follows:
EM score = (number of questions answered exactly correctly) / (total number of questions)
where "number of questions answered exactly correctly" refers to the number of questions for which the system provides an answer that exactly matches the annotated answer provided by the dataset.
For example, if a model answers 80 questions correctly out of a total of 100 questions, its EM score would be 0.80 (80/100).

B. F1 Score

The F1 score is a commonly used metric in natural language processing (NLP) to evaluate the performance of a model in a question answering task, such as the Stanford Question Answering Dataset (SQuAD) 2.0.[57,58,28] The F1 score is the harmonic mean of precision and recall, and it measures how well a model balances between identifying relevant answers (recall) and avoiding irrelevant ones (precision). The formula for F1 score is:
F1 score = 2 * (precision * recall) / (precision + recall)
where,
precision = true positives / (true positives + false positives)
and
recall = true positives / (true positives + false negatives)
In the case of SQuAD 2.0, the F1 score is calculated based on the exact match (EM) and partial match (PM) of the predicted answers with the ground truth answers. Specifically, a predicted answer is considered to be correct if it matches the ground truth answer exactly (EM), or if it overlaps with the ground truth answer by at least 50% (PM). The F1 score is then computed as the average of the EM and PM scores.
Here is an example of how the F1 score is calculated for SQuAD 2.0:
Suppose a model is asked the following question and provides the following answer:
Question: What is the capital of France?
Predicted answer: Paris, France
If the ground truth answer is "Paris," then the predicted answer is considered to be correct based on EM. If the ground truth answer is "The capital of France is Paris," then the predicted answer is considered to be correct based on the PM. If the ground truth answer is "France has several cities, including Paris and Marseille," then the predicted answer is considered to be incorrect.

The F1 score for this example is calculated as follows:
Precision = 1 (one true positive and zero false positives)
Recall = 1 (one true positive and zero false negatives)
F1 score = 2 * (precision * recall) / (precision + recall) = 2 * (1 * 1) / (1 + 1) = 1

The F1 score and the EM score are both important metrics in MRC, as a high F1 score indicates that the model can produce an answer that is close to the ground truth answer, while a high EM score indicates that the model can produce the exact answer. Hence we will be using both EM and F1 scores as our metrics to evaluate the performance of the model on a particular dataset.

IV. PROPOSED METHODOLOGY

The DistilBERT QA (question-answering) head, training hyperparameters, and data loading and splitting are just a few of the parts that make up this section on the DistilBERT model. In order to train and evaluate the DistilBERT model for the specific goal of question-answering, the experimental setup and procedures are explained in detail in this section.

A. Model

In our experiment the model used DistilBERT. DistilBERT is a variant of the popular pre-trained language model BERT, which was introduced by Google in 2018. DistilBERT was developed by Hugging Face, an AI startup company that specializes in natural language processing (NLP) and deep learning. It was introduced in the paper "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter."[65].

DistilBERT is a more compact and quick variant of BERT that is intended to be more effective when used on devices with constrained computational power. This is accomplished by shrinking the BERT model using a technique known as knowledge distillation, which entails teaching a smaller model to behave in a manner similar to that of a bigger one.

In addition to its smaller size, DistilBERT also uses a modified training process that helps to increase its efficiency. Specifically, it removes the token type embeddings from the input and the pooler layer from the output, and it uses a smaller transformer architecture.

***Article Received**: 10 May 2023 **Revised**: 08 July 2023 **Accepted**: 02 August 2023*

DistilBERT has been shown to achieve similar performance to BERT on a range of NLP tasks, while requiring significantly less computational resources. This makes it a popular choice for applications that require efficient and accurate NLP, such as question answering, sentiment analysis, and text classification.

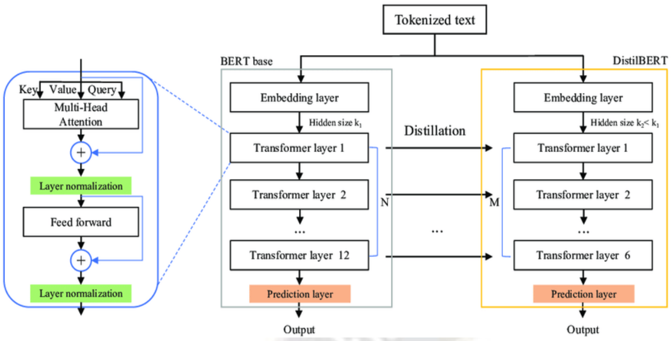


Figure 6: The DistilBERT model architecture and components

The main idea behind DistilBERT is to reduce the computational cost and memory requirements of BERT while retaining most of its performance.

DistilBERT achieves this by:

1. Using a smaller architecture: DistilBERT has a smaller number of layers and hidden units compared to BERT. Specifically, it has 6 layers instead of 12 for the base model and 2,048 hidden units instead of 7,680.
2. Training with knowledge distillation: DistilBERT is trained using a technique called knowledge distillation, where it learns to mimic the behavior of the larger BERT model. During training, the outputs of BERT are used as soft targets to train DistilBERT.

DistilBERT's overall architecture is comparable to BERT's, although it has fewer layers and concealed units. As seen in figure 6, the multi-layer bidirectional transformer encoder used by DistilBERT has a feed-forward neural network and a self-attention mechanism on each layer. The model can focus on various elements of the input sequence and recognize contextual relationships between words thanks to the self-attention mechanism.

A series of tokens that are first embedded into a high-dimensional vector space make up the input to DistilBERT. The transformer encoder then processes these embeddings to produce contextualized representations for each token. In order to predict the output, a classification or regression layer is provided with the contextualized representations.

On a variety of natural language processing tasks, DistilBERT has been demonstrated to perform comparably to BERT while being significantly faster and smaller.

The transformers model DistilBERT was pretrained on the same corpus using the BERT base model as a teacher. DistilBERT is smaller and faster than BERT. This indicates that the BERT basic model was automatically utilized to derive inputs and labels from the texts. It was pre-trained using only the raw texts, with no human labeling of any type (therefore, it may use a significant quantity of publicly available data). It is pretrained with three objectives:

- **Distillation loss**: The model was trained to output probabilities identical to those of the BERT base model.
- **Masked language modeling (MLM):** This is a piece of the initial training loss of the BERT basic model. The model must predict the hidden words once the hidden words are chosen at random from 15% of the input words in a sentence. Compare this to autoregressive models like GPT, which internally conceal the next tokens, and conventional recurrent neural networks (RNNs), which frequently perceive the words sequentially. The model is able to identify a two-way representation of the statement as a result.
- **Cosine embedding loss:** The model was also trained to generate hidden states that closely resembled the BERT basic model. In doing way, the model performs downstream or inference tasks more efficiently while acquiring the same internal representation of the English language as its instructor model.

In an effort to enhance the DistilBERT baseline model, we have experimented with different question heads that differ in the number of layers, activation function, and general structure. To create four distinct question heads, we used two models with two fully linked layers and two models with three fully connected layers.

B. Data Loading and Split:

Like other machine learning models, DistilBERT requires a set of data to train on. The data is typically split into three subsets: training data, validation data, and test data.

Here are the steps for loading and splitting data for a DistilBERT model:

1. Load the data: The data must first be loaded into your software as the first stage. Depending on the format of the data, many ways can be used to accomplish this. For instance, you may use the pandas library to read a CSV file containing data into a dataframe.
2. Preprocess the data: After the data has been loaded, it needs to be preprocessed so that it may be used to train the model. Tokenization, stop-word removal, and translating the data into a format that can be fed into the model may be included in this process.
3. Split the data: The data is divided into three groups after preprocessing: training data, validation data, and test data. The validation data is used to fine-tune the

hyperparameters, the test data is used to assess the model's performance, and the training data is used to train the model.

4. Create data loaders: Once the data is split, you need to create data loaders for each subset. Data loaders are used to load batches of data into the model during training.

We aimed to decode the JSON training set into a meaningful table that might be used for additional pre-processing operations in the future. As each context includes several question/answer pairings and each question may have multiple responses, we allot one row for each context/question/answer triple, resulting in multiple copies of the same context and question. In terms of the dataset splitting approach, we used the official SQuAD v1.1 dev set as the test set and 20% of the whole dataset for validation and SQuAD 2.0 is used as the training set. The same question may have more than one right response in this dataset, which makes it different from the training dataset.

C. Training hyperparameters:

The training hyperparameters used in this scenario include a learning rate of 2e-05, a training batch size of 12 and an eval batch size of 12. The random seed used was 42, the optimizer Adam with 0.9 and 0.999 beta and 1e-08 epsilon.The learning rate scheduler used was linear, and the model was trained for a thousand years.These hyperparameters are essential to obtain a well-trained model that captures the underlying patterns in the data accurately.However, it is essential to note that selecting the right hyperparameters is a complex task, which requires careful experimentation and adjustment in order to achieve optimal performance.

learning rate : 2e-05
train_batch_size : 12
eval_batch_size : 12
seed : 42
optimizer : Adam with betas=(0.9,0.999) and epsilon=1e-08
lr_scheduler_type : linear
epochs : 10000

D. Vanilla DistilBert Head:

The top layer of the DistilBert model that has been tailored for a particular natural language processing (NLP) task is referred to as the "Vanilla DistilBert Head." Transfer learning is typically used to train it, which entails optimizing the previously trained model for a particular downstream task. This entails training the entire model from beginning to end on the downstream task data and adding a task-specific classification layer on top of the DistilBert model that has already been trained. It is regarded as "vanilla" since it is completely unmodified and just includes the standard classification layer that is included with the pre-trained DistilBert model. Therefore, the Vanilla DistilBert Head provides a clear-cut method for adjusting the model for a particular NLP task.

We enhanced DistilBERT's default question-answering functionality so that it could be used with the HuggingFace transformers library. The 768 output dimension from the DistilBERT backbone is reduced to 2 using a single linear layer with no activation function.

E. DistilBERT with custom Question-Answering Head

DistilBERT can be used for QA work by simply layering a customized QA head on top of the pre-trained model. A linear layer and a softmax activation layer make up the two layers that make up the QA head. The output of the pre-trained model is sent into the linear layer, which transforms it into two vectors of equal size (one for the start location of the answer and one for the finish position). The final probability distribution over the potential response positions is created by applying a softmax function to the output of the linear layer in the softmax activation layer.

In order to study the benefits of having additional layers, we created two levels on top of DistilBERT. We then evaluated various activation functions to see which one performed the best. We did this to become better at responding to questions. Other arrangement shapes have been tested, but we will only present the best. The idea was to use a conic function to decrease the dimension from the DistilBERT output dimension to 2 more gradually.

We experimented with changing the model's layer dimensions while using GELU as an activation function.
Popular activation functions for deep learning models include the Gaussian Error Linear Unit (GELU). Hendrycks and Gimpel made the suggestion in their 2016 publication "Gaussian Error Linear Units (GELUs)".
The GELU function is defined as follows:

$$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}(x/\sqrt{2})\right], \qquad (1)$$

if $X \sim \mathcal{N}(0, 1)$.

The GELU function has several desirable properties, including being smooth and differentiable, having zero mean and unit variance for inputs in the normal range, and being able to approximate the identity function for large inputs.

V. RESULTS AND DISCUSSION

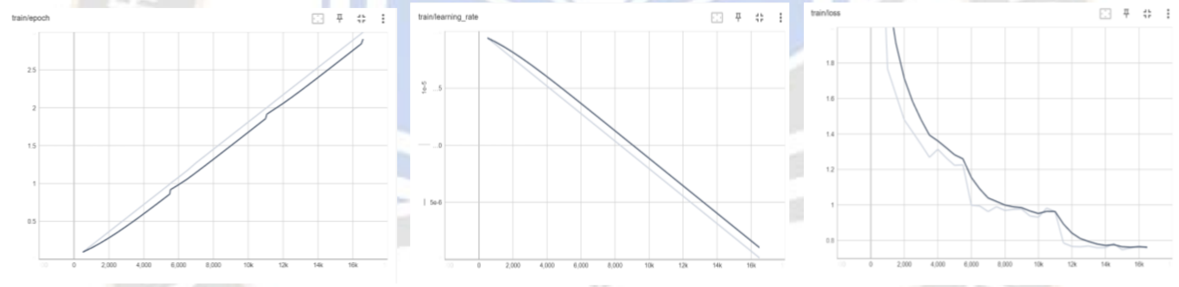
According to Table 3, the DistilBERT model over 6500 train steps acquired the EM score of 88.2% and F1 score of 84.3%. In our case, the batch size arrangement with 12 yielded the best outcomes. Two more factors that have been demonstrated to affect training are the learning rate and the repetition rate. We were unable to train our model from numerous other heads and compare it to other fine-tuned converters due to a lack of computational resources. The same can be seen in the several graphs in figure 7. Although there are many question-answering datasets available, we were only able to fine-tune our model on the SQuAD dataset due to restrictions.

The number of epochs is a difficult issue since it could result in either overtraining or undertraining if the model is trained for an excessively long time. By analyzing the data, we discovered an intriguing pattern: as shown in the charts below, the frequency of the questions is not connected with the F1 score or exact match. Additionally, it is evident that, on average, more open inquiries such as "why," "if," and "what" are asked.

For future works, this question answering head could be tried on the top of other transformers to see if it improves their performances too.

TABLE 3: EM AND F1 SCORES OF DISTILBERT MODEL WITH VARYING TRAINING STEPS

TRAIN STEPS	**EM SCORE**	**F1 SCORE**
1000	69.73	64.2
3500	78	76.1
6500	88.2	84.3
10000	84	79



A) B) C)

FIGURE 7: TRAINING GRAPHS OF THE MODEL: A)TRAIN V/S EPOCH B)TRAIN V/S LEARNING RATE C)TRAIN V/S LOSS

VI. CONCLUSION

MRC is a challenging task in NLP, and the SQuAD 2.0 dataset is a popular benchmark dataset used to evaluate MRC models. Recent advancements in MRC have shown that the use of pre-trained language models, ensemble models, and external knowledge sources can improve performance on the SQuAD 2.0 dataset. The performance of MRC models is evaluated using the F1 score and the EM score, and both metrics are important in evaluating the model's ability to produce an accurate answer. MRC on SQuAD 2.0 is a challenging task that has received significant attention from the research community in recent years. BERT-based models have achieved state-of-the-art performance on SQuAD 2.0, but there is still room for improvement. Future work could focus on incorporating external knowledge sources beyond pre-trained language models, such as structured knowledge bases or domain-specific ontologies. Additionally, developing models that can reason over multiple pieces of textual information could further improve performance on MRC tasks. Overall, MRC on SQuAD 2.0 remains an active area of research with many exciting avenues for future exploration.

Finally, our experiment showed that the DistilBERT model for answering questions can be used with high accuracy, while requiring significantly less computational resources than BERT. We were able to further improve the performance of the DistilBERT base model by experimenting with different questions. Our findings suggest that DistilBERT is suitable for applications that require effective and accurate NLP, such as questions, sentimental analysis and text classification. In addition, we highlighted the importance of data loading and separation, preprocessing and fine-tuning hyperparameters in order to achieve optimal performance with the model.

VII. CHALLENGES AND FUTURE DIRECTIONS

While significant progress has been made in the field, there are still several challenges that need to be addressed. One of the main challenges is the lack of large-scale annotated datasets, particularly in languages other than English. Another challenge is the difficulty in handling complex questions that require multi-step reasoning or domain-specific knowledge.

Future directions for research include developing models that can handle more complex and diverse questions, improving the robustness of models to handle noisy and ambiguous input, and developing models that can perform and generate the results in low resource settings.

REFERENCES

[1] Zhang X, Yang A, Li S, Wang Y. Machine reading comprehension: a literature review. arXiv preprint arXiv:1907.01686. 2019 Jun 30.

[2] Niu Y, Jiao F, Zhou M, Yao T, Xu J, Huang M. A self-training method for machine reading comprehension with soft evidence extraction. arXiv preprint arXiv:2005.05189. 2020 May 11.

[3] Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P. Teaching machines to read and comprehend. Advances in neural information processing systems. 2015;28.

[4] Shen Y, Huang XJ. Attention-based convolutional neural network for semantic relation extraction. InProceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers 2016 Dec (pp. 2526-2536).

[5] Koščević K, Subašić M, Lončarić S. Attention-based convolutional neural network for computer vision color constancy. In2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA) 2019 Sep 23 (pp. 372-377). IEEE.

[6] Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems. 2021 Jun 10.

[7] Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P. Teaching machines to read and comprehend. Advances in neural information processing systems. 2015;28.

[8] Rajpurkar P, Zhang J, Lopyrev K, Liang P. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250. 2016 Jun 16.

[9] Nguyen T, Rosenberg M, Song X, Gao J, Tiwary S, Majumder R, Deng L. MS MARCO: A human generated machine reading comprehension dataset. choice. 2016 Nov;2640:660.

[10] Trischler A, Wang T, Yuan X, Harris J, Sordoni A, Bachman P, Suleman K. Newsqa: A machine comprehension dataset. arXiv preprint arXiv:1611.09830. 2016 Nov 29.

[11] Kočiský T, Schwarz J, Blunsom P, Dyer C, Hermann KM, Melis G, Grefenstette E. The narrativeqa reading comprehension challenge. Transactions of the Association for Computational Linguistics. 2018 Jan 1;6:317-28.

[12] Joshi M, Choi E, Weld DS, Zettlemoyer L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551. 2017 May 9.

[13] Bauer L, Wang Y, Bansal M. Commonsense for generative multi-hop question answering tasks. arXiv preprint arXiv:1809.06309. 2018 Sep 17.

[14] Reddy S, Chen D, Manning CD. Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics. 2019 Aug 1;7:249-66.

[15] Rajpurkar P, Jia R, Liang P. Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822. 2018 Jun 11.

[16] Shang S, Liu J, Yang Y. Multi-layer transformer aggregation encoder for answer generation. IEEE Access. 2020 May 11;8:90410-9.

[17] Dhabliya, D. (2021). Feature Selection Intrusion Detection System for The Attack Classification with Data Summarization. Machine Learning Applications in Engineering Education and Management, 1(1), 20–25. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/8.

[18] Egonmwan E, Castelli V, Sultan MA. Cross-task knowledge transfer for query-based text summarization. InProceedings of the 2nd Workshop on Machine Reading for Question Answering 2019 Nov (pp. 72-77).

[19] Qi P, Lin X, Mehr L, Wang Z, Manning CD. Answering complex open-domain questions through iterative query generation. arXiv preprint arXiv:1910.07000. 2019 Oct 15.

[20] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019;32.

[21] Liu X, Cheng H, He P, Chen W, Wang Y, Poon H, Gao J. Adversarial training for large neural language models. arXiv preprint arXiv:2004.08994. 2020 Apr 20.

[22] Xu H, Liu B, Shu L, Yu PS. BERT post-training for review reading comprehension and aspect-based sentiment analysis. arXiv preprint arXiv:1904.02232. 2019 Apr 3.

[23] Yan Z, Ma J, Zhang Y, Shen J. SQL Generation via Machine Reading Comprehension. InProceedings of the 28th International Conference on Computational Linguistics 2020 Dec (pp. 350-356).

[24] Xu S, Xu G, Jia P, Ding W, Wu Z, Liu Z. Automatic task requirements writing evaluation via machine reading comprehension. InArtificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I 22 2021 (pp. 446-458). Springer International Publishing.

[25] Ana Oliveira, Yosef Ben-David, Susan Smit , Elena Popova, Milica Milić. Machine Learning for Decision Optimization in Complex Systems. Kuwait Journal of Machine Learning, 2(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/201.

[26] Schlegel V, Nenadic G, Batista-Navarro R. Semantics altering modifications for evaluating comprehension in machine reading. InProceedings of the AAAI Conference on Artificial Intelligence 2021 May 18 (Vol. 35, No. 15, pp. 13762-13770).

[27] Kamfonas M, Alon G. What Can Secondary Predictions Tell Us? An Exploration on Question-Answering with SQuAD-v2. 0. arXiv preprint arXiv:2206.14348. 2022 Jun 29.

[28] Yuan S, Yang D, Liang J, Sun J, Huang J, Cao K, Xiao Y, Xie R. Large-Scale Multi-granular Concept Extraction Based on Machine Reading Comprehension. InThe Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 2021 Sep 30 (pp. 93-110). Cham: Springer International Publishing.

[29] Sun T, He Z, Zhu Q, Qiu X, Huang X. Multi-Task Pre-Training of Modular Prompt for Few-Shot Learning. arXiv preprint arXiv:2210.07565. 2022 Oct 14.

[30] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.

[31] Liu X, He P, Chen W, Gao J. Multi-task deep neural networks for natural language understanding. arXiv preprint arXiv:1901.11504. 2019 Jan 31.

[32] Gao, Y., Liu, X., Duan, N., & Gao, J. (2020). Multi-Task Learning for Script Event Understanding with Rich Syntactic and Semantic Representations. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 3158-3167

[33] Welbl, J., Stenetorp, P., & Riedel, S. (2018). Constructing Datasets for Multi-hop Reading Comprehension Across Documents. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 3202-3213

[34] Reddy, S., Chen, D., Manning, C. D. (2019). CoQA: A Conversational Question Answering Challenge. Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3543-3552.

[35] Seo M, Kembhavi A, Farhadi A, Hajishirzi H. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603. 2016 Nov 5.

[36] Nogueira R, Cho K. Passage Re-ranking with BERT. arXiv preprint arXiv:1901.04085. 2019 Jan 13.

[37] Li Z, Pan X. Some remarks on regularity criteria of axially symmetric Navier-Stokes equations. arXiv preprint arXiv:1805.10752. 2018 May 28.

[38] Dhiman, O. ., & Sharma, D. A. . (2020). Detection of Gliomas in Spinal Cord Using U-Net++ Segmentation with Xg Boost Classification. Research Journal of Computer Systems and Engineering, 1(1), 17–22. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/20

[39] Zhang Y. Note on equivalences for degenerations of Calabi-Yau manifolds. arXiv preprint arXiv:1711.00503. 2017 Nov 1.

[40] Li S, Wu L, Feng S, Xu F, Xu F, Zhong S. Graph-to-tree neural networks for learning structured input-output translation with applications to semantic parsing and math word problem. arXiv preprint arXiv:2004.13781. 2020 Apr 7.

[41] Miyaji T, Herrera-Endoqui M, Krumpe M, Hanzawa M, Shogaki A, Matsuura S, Tanimoto A, Ueda Y, Ishigaki T, Barrufet L, Brunner H. Torus Constraints in ANEPD-CXO245: A Compton-thick AGN with Double-peaked Narrow Lines. The Astrophysical Journal Letters. 2019 Oct 7;884(1):L10

[42] Yu AW, Dohan D, Luong MT, Zhao R, Chen K, Norouzi M, Le QV. Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541. 2018 Apr 23.

[43] Shang M, Li P, Fu Z, Bing L, Zhao D, Shi S, Yan R. Semi-supervised text style transfer: Cross projection in latent space. arXiv preprint arXiv:1909.11493. 2019 Sep 25.

[44] Shen D, Celikyilmaz A, Zhang Y, Chen L, Wang X, Gao J, Carin L. Towards generating long and coherent text with multi-level latent variable models. arXiv preprint arXiv:1902.00154. 2019 Feb 1.

[45] Gururangan S, Swayamdipta S, Levy O, Schwartz R, Bowman SR, Smith NA. Annotation artifacts in natural language inference data. arXiv preprint arXiv:1803.02324. 2018 Mar 6.

[46] Wang W, Yan M, Wu C. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. arXiv preprint arXiv:1811.11934. 2018 Nov 29.

[47] Naseem T, Shah A, Wan H, Florian R, Roukos S, Ballesteros M. Rewarding Smatch: Transition-based AMR parsing with reinforcement learning. arXiv preprint arXiv:1905.13370. 2019 May 31.

[48] Sofia Martinez, Machine Learning-based Fraud Detection in Financial Transactions , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[49] Liu NF, Gardner M, Belinkov Y, Peters ME, Smith NA. Linguistic knowledge and transferability of contextual representations. arXiv preprint arXiv:1903.08855. 2019 Mar 21.

[50] R-NET: MACHINE READING COMPREHENSION WITH SELF-MATCHING NETWORKS by Natural Language Computing Group, Microsoft Research Asia, 2017

[51] Aronna MS, Bonnans JF, Kröner A. State-constrained control-affine parabolic problems I: first and second order necessary optimality conditions. Set-Valued and Variational Analysis. 2021 Jun;29(2):383-408.

[52] Takanobu R, Liang R, Huang M. Multi-agent task-oriented dialog policy learning with role-aware reward decomposition. arXiv preprint arXiv:2004.03809. 2020 Apr 8.

[53] Abhijith, G. S. V. ., & Gundad, A. K. V. . (2023). Data Mining for Emotional Analysis of Big Data. International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 271–279. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2684

[54] Azatov A, Bardhan D, Ghosh D, Sgarlata F, Venturini E. Anatomy of b→ c τ ν anomalies. Journal of High Energy Physics. 2018 Nov;2018(11):1-54.

[55] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019;32.

[56] Guu K, Lee K, Tung Z, Pasupat P, Chang M. Retrieval augmented language model pre-training. InInternational conference on machine learning 2020 Nov 21 (pp. 3929-3938). PMLR.6

[57] Wei M, Yuan C, Yue X, Zhong K. Hose-net: Higher order structure embedded network for scene graph generation. InProceedings of the 28th ACM International Conference on Multimedia 2020 Oct 12 (pp. 1846-1854)

[58] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942. 2019 Sep 26.

[59] Zippilli S, Vitali D. Possibility to generate any Gaussian cluster state by a multimode squeezing transformation. Physical Review A. 2020 Nov 30;102(5):052424.

[60] Wang H, Wang J. Topological bands in two-dimensional orbital-active bipartite lattices. Physical Review B. 2021 Feb 18;103(8):L081109.

[61] Dr. Bhushan Bandre. (2013). Design and Analysis of Low Power Energy Efficient Braun Multiplier. International Journal of New Practices in Management and Engineering, 2(01), 08 - 16. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/12.

[62] Christianos F, Schäfer L, Albrecht S. Shared experience actor-critic for multi-agent reinforcement learning. Advances in neural information processing systems. 2020;33:10707-17.

[63] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019 Jul 26.

[64] Lewis, P., Oguz, B., Rinott, R., & Riedel, S. (2020). Evaluating the state-of-the-art on the Natural Questions benchmark. Transactions of the Association for Computational Linguistics, 8, 423-438.

[65] Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, Wang H. Ernie 2.0: A continual pre-training framework for language understanding. InProceedings of the AAAI conference on artificial intelligence 2020 Apr 3 (Vol. 34, No. 05, pp. 8968-8975).

[66] Bisk, Y., Holtzman, A., Schwartz, R., Kyunghyun, C., & Smith, N. A. (2020). Minimally supervised learning of representations for structured data. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7816-7826).

[67] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.

[68] Nath S. Preference Estimation in Deferred Acceptance with Partial School Rankings. arXiv preprint arXiv:2010.15960. 2020 Oct 29.

[69] Wang S, Yu M, Guo X, Wang Z, Klinger T, Zhang W, Chang S, Tesauro G, Zhou B, Jiang J. R 3: Reinforced ranker-reader for open-domain question answering. InProceedings of the AAAI Conference on Artificial Intelligence 2018 Apr 26 (Vol. 32, No. 1).

[70] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.

[71] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. 2019 Oct 2.