IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, January 2024

LEBERT :Lite and Efficiently Optimized BERT PRetraining Approach

Priyanka Yadav and Anjali Sharma
Institute of Distance and Open Learning, Mumbai, Maharashtra, India

Abstract: *The extensive generalization of these models can lead to overfitting, causing the model to perform poorly on unseen data and thereby not realizing its full potential. To address this challenge systematically, we propose a novel approach for lightweight and efficient fine-tuning of BERT (**Bidirectional Encoder Representations from Transformers**) that aims to achieve improved generalization and harness the maximum capabilities of BERT. Our proposed approach incorporates various regularization techniques designed to adaptively manage the model's complexity. We plan to conduct experiments using this approach across various NLP tasks, including GLUE (Wang et al., 2019), RACE (Lai et al., 2017), and SQuAD (Rajpurkar et al., 2016).*

Keywords: BERT

I. INTRODUCTION

All forms of human expression, whether spoken or written, contain vast reservoirs of information. The nuances of our thoughts, our tone, and our carefully chosen words all contribute valuable data that can be extracted and analyzed. In theory, this wealth of information could allow us to not only comprehend but even predict human behavior.

However, a significant challenge arises when an individual generates extensive, and at times, hundreds or even thousands of words in a single statement, each sentence brimming with its unique complexities. When attempting to scale and analyze numerous individuals, variations, or declarations within a given context, the task becomes unmanageable.

Conversations, declarations, and even the succinct messages conveyed through tweets serve as prime examples of unstructured data. This type of information does not easily conform to the traditional row and column structure of relational databases and, in fact, constitutes the predominant share of data encountered in the real world. It is inherently chaotic and arduous to manipulate.

Nevertheless, the field of Natural Language Processing (NLP) has undergone a remarkable transformation in recent times. The focus has shifted away from the conventional mechanical approach of interpreting text or speech based solely on keywords. Instead, contemporary efforts center on comprehending the inherent meaning behind these words, embracing a more cognitive approach."

natural Language Processing (NLP) poses a significant challenge in the field of computer science. This difficulty primarily arises from the intricate nature of human language. The rules governing the use of natural language are far from straightforward, making it a formidable task for computers to grasp.

These language rules can vary widely in complexity. Some are high-level and abstract, such as understanding sarcasm, while others are low-level, such as recognizing plural forms marked by the letter 's.'NLP necessitates a comprehensive understanding of not only individual words but also how these words interconnect to convey intended messages. Although humans naturally acquire language skills, the idiosyncrasies, ambiguities, and abstract aspects of natural languages present formidable challenges for machines.In the field of NLP, algorithms play a pivotal role. They are employed to identify and extract the underlying language rules, enabling the transformation of unstructured language data into a format that computers can comprehend.

When processing text, computers utilize algorithms to extract meaning from each sentence, gathering essential information from the provided content."

At times, computers may struggle to accurately understand the meaning of a sentence, leading to unclear outcomes (Garbade, 2018). With the advancements in Deep Learning (DL), neural network architectures such as Recurrent

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, January 2024

Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, as well as Convolutional Neural Networks (CNN), have shown significant improvements in various Natural Language Processing (NLP) tasks like text classification, language modeling, and machine translation.

However, the performance of DL models in NLP still lags behind the remarkable achievements of deep learning in computer vision. One of the primary reasons for this slower progress is the limited availability of large labeled text datasets. Most labeled text datasets are not sufficiently large to effectively train deep neural networks, given that these networks have a large number of parameters, and training them on small datasets often leads to overfitting.

Another significant factor contributing to NLP's slower advancement compared to computer vision is the lack of transfer learning in NLP. Transfer learning has played a crucial role in the success of DL in computer vision. This success is attributed to the existence of extensive labeled datasets like ImageNet, on which deep CNN-based models were trained and later used as pre-trained models for a wide range of computer vision tasks. This was not the case in NLP until 2018 when Google introduced the transformer model."

Many tasks in Natural Language Processing (NLP), such as text classification, language modeling, and machine translation, involve modeling sequences of data. Traditional machine learning models and conventional neural networks struggle to capture the sequential information inherent in text data. To address this, researchers turned to recurrent neural networks (RNN) and Long Short-Term Memory (LSTM) architectures (Sherstinsky, 2021) as they are capable of modeling sequential data within text.

However, RNNs have their own challenges, notably their inability to parallelize processing because they take one input at a time. In the context of a text sequence, an RNN or LSTM processes one token at a time, resulting in slow training on large datasets. This highlighted the need for transfer learning in NLP.

In 2018, Google introduced the transformer architecture in the paper 'Attention is All You Need' (Vaswani et al., 2017), marking a significant breakthrough in NLP. Following this, numerous transformer-based models emerged for various NLP tasks. Transformer-based models offer multiple advantages. Firstly, they process entire input sequences in a single pass, enabling efficient GPU utilization. Secondly, these models do not require labeled data for pre-training. Instead, they can be trained on a large volume of unlabeled text data and subsequently applied to various NLP tasks, such as text classification, named entity recognition, and text generation.

This is the essence of how transfer learning operates in NLP. BERT (Bidirectional Encoder Representations from Transformers) and GPT-3 are among the most renowned transformer-based models. In this context, we will focus on BERT and explore how a pre-trained BERT model can be effectively employed for text classification with enhanced optimization. BERT is a recent paper published by Google AI Language, known for achieving outstanding results across various NLP tasks, including Question Answering (SQuAD v1.1) and Natural Language Inference (MNLI). BERT's groundbreaking innovation lies in its bidirectional training of the Transformer model, a widely adopted attention-based model, for language modeling."

In contrast to earlier approaches that examined text sequences in either a left-to-right or a combined left-to-right and right-to-left manner, a groundbreaking study (Devlin et al., 2020) introduced a bidirectional language model. The results of this research demonstrate that a bidirectionally trained language model possesses a richer understanding of language context and flow compared to single-direction language models.

The research paper outlines a novel technique called Masked Language Model (MLM), which enables bidirectional training in models where it was previously deemed impossible. BERT (Bidirectional Encoder Representations from Transformers) employs the Transformer architecture, an attention mechanism designed to learn contextual relationships among words or sub-words in a text. In its basic form, Transformer consists of two components: an encoder that processes the text input and a decoder responsible for task predictions. Given that BERT aims to create a language model, it utilizes only the encoder component.

The inner workings of Transformer are elaborated in a Google paper. Unlike directional models that sequentially process text input (either left-to-right or right-to-left), the Transformer encoder simultaneously processes the entire word sequence. Therefore, it is considered bidirectional, although it would be more precise to describe it as non-directional. This characteristic allows the model to understand a word's context based on its complete surroundings, encompassing both left and right sides of the word.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, January 2024

A major challenge associated with using BERT and similar large neural language models is the substantial computational resources required for training, fine-tuning, and making inferences. However, recent research has proposed various techniques to address this issue. Knowledge distillation, quantization, pruning, and other approaches have made BERT models more feasible for production environments. Our approach focuses on regularizing existing models to make them suitable for a wide range of NLP tasks."

II. LITERATURE SURVEY

The concept of transformer models was initially introduced by Vaswani et al. in 2017 for the purpose of neural machine translation. The remarkable performance achieved by these models served as an inspiration for Devlin et al. in 2019 to propose a bidirectional transformer-based language model known as BERT (Bidirectional Encoder Representations from Transformers).

In their work, Devlin et al. (2019) undertook the pre-training of the BERT model using an extensive corpus, all without the need for human annotation, through unsupervised learning tasks. BERT's success sparked subsequent research efforts aimed at refining the pre-training process. These efforts included the introduction of new unsupervised learning tasks (Yang et al., 2020), increasing the model's size (Lan et al., 2019; Raffel et al., 2019), expanding the training data (Liu et al., 2019c; Yang et al., 2019; Raffel et al., 2019), and exploring multi-task learning (Liu et al., 2019).

2.1 Scaling up Representation Learning for Natural Language

Learning representations of natural language has proven to be incredibly advantageous in a wide range of Natural Language Processing (NLP) tasks and has gained widespread acceptance in the field (Mikolov et al., 2013; Le & Mikolov, 2014; Dai & Le, 2015; Peters et al., 2018; Devlin et al., 2019; Radford et al., 2018; 2019). Notably, a significant shift in recent years is the move away from pre-training word embeddings, whether they are static (Mikolov et al., 2013; Pennington et al., 2014) or contextualized (McCann et al., 2017; Peters et al., 2018), towards comprehensive full-model pre-training, followed by fine-tuning for specific tasks (Dai & Le, 2015; Radford et al., 2018; Devlin et al., 2019).

Research in this area consistently demonstrates that larger model sizes lead to improved performance. For example, Devlin et al. (2019) showed that when addressing three carefully selected natural language understanding tasks, using larger hidden dimensions, additional layers, and increased attention heads consistently resulted in better outcomes. However, they set a limit on the hidden dimension at 1024, likely due to constraints related to model size."

The challenge of dealing with large models is compounded by computational constraints, particularly concerning GPU/TPU memory limitations. Given that contemporary models often comprise hundreds of millions or even billions of parameters, memory constraints become a significant concern. To address this issue, Chen et al. (2016) introduced a technique called gradient check pointing, which reduces memory requirements in a sub-linear manner but adds the cost of an additional forward pass. Gomez et al. (2017) proposed a method to reconstruct the activations of each layer from the subsequent layer, eliminating the need to store intermediate activations.

Both of these approaches reduce memory consumption, albeit at the expense of speed. In contrast, Raffel et al. (2019) suggested employing model parallelization to train a large model. In our case, our parameter reduction techniques decrease memory usage while simultaneously enhancing training speed.

2.2 Cross-Layer Parameter Sharing

The concept of sharing parameters throughout layers has been formerly explored with the Transformer structure (Vaswani et al., 2017), however this earlier paintings has centered on schooling for trendy encoder decoder obligations in preference to the pretraining/finetuning setting. Different from our observations, Dehghani et al. (2018) display that networks with cross-layer parameter sharing (Universal Transformer, UT) get higher overall performance on language modelling and subject-verb settlement than the usual transformer. Very recently, Bai et al. (2019) advocate a Deep Equilibrium Model (DQE) for transformer networks and display that DQE can attain an equilibrium factor for which the enter embedding and the output embedding of a sure

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, January 2024

layer live the same. Our observations display that our embeddings are oscillating in preference to converging. Hao et al. (2019) integrate a parameter-sharing transformer with the usual one, which in addition will increase the quantity of parameters of the usual transformer.

2.3 Sentence Ordering Objectives

ALBERT employs a pretraining objective based on predicting the order of consecutive text segments. Various researchers have explored pretraining objectives related to discourse coherence. Discourse coherence and cohesion have been extensively studied, revealing various phenomena that connect adjacent text segments (Hobbs, 1979; Halliday & Hasan, 1976; Grosz et al., 1995). Many effective objectives in practice are relatively straightforward. For example, Skip Thought (Kiros et al., 2015) and FastSent (Hill et al., 2016) sentence embeddings are learned by encoding a sentence to predict words in neighboring sentences. Other objectives for sentence embedding learning include predicting future sentences rather than just neighboring ones (Gan et al., 2017) and predicting specific discourse markers (Jernite et al., 2017; Nie et al., 2019). Our loss function closely resembles the sentence ordering objective proposed by Jernite et al. (2017), where sentence embeddings are learned to determine the order of consecutive sentences. However, it's worth noting that our loss is defined for textual segments rather than individual sentences.

In the context of BERT (Devlin et al., 2019), their loss function is based on predicting whether the second segment in a pair has been swapped with a segment from another document. In your experiments, you found that the task of sentence ordering, where you predict the correct order of consecutive segments of text, is more challenging during pretraining. Additionally, this sentence ordering task appears to be more beneficial for certain downstream tasks when compared to BERT's original loss function.

Simultaneously, Wang et al. (2019) also worked on predicting the order of consecutive text segments. However, they combined this task with the original next sentence prediction task in a three-way classification framework. This suggests that there is ongoing research into improving pretraining tasks for models like BERT to enhance their performance on various natural language processing taskstaking in preference to empirically evaluating .

III. METHODOLOGY

BERT, which stands for Bidirectional Encoder Representations from Transformers, constitutes a substantial neural network architecture with a substantial parameter count, which can range from 100 million to as much as 300 million. Consequently, attempting to train a BERT model from the ground up using a limited dataset could result in overfitting (Joshi, 2020). Therefore, a more prudent approach is to leverage a pre-trained BERT model that has been trained on a vast dataset as a starting point. Subsequently, the model can be further refined or fine-tuned on our comparatively smaller dataset. This process is commonly referred to as model fine-tuning.

Different Fine-Tuning Techniques

Full-model training involves instructing the entire pre-trained model on our dataset and passing its output to a SoftMax layer. In this scenario, errors are propagated throughout the entire architecture, and the pre-existing model weights are adjusted based on the new dataset

Another approach to utilize a pre-trained model is to train it selectively by freezing certain layers while retraining others. In this method, we maintain the weights of the initial layers of the model in a fixed state while exclusively retraining the upper layers. The determination of which layers to freeze and which ones to retrain can be experimentally explored

An alternative strategy is to completely freeze the entire structure of the model, including all of its layers, and then attach additional neural network layers of our own. This new composite model is then trained, with the weights of only the appended layers being updated during the training process

IV. CONCLUSION

BERT turned into capable of enhance the accuracy on many Natural Language Processing and Language Modelling obligations. The foremost leap forward this is furnished via way of means of this paper is permitting using semi-supervised gaining knowledge of for plenty NLP undertaking that permits switch gaining knowledge of in NLP.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, January 2024

REFERENCES

[1] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Maten, Yanqi Zhou, Wei Li, Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.2020

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss Gretchen Krueger, Tom Henighan Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray Benjamin Chess Jack Clark Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. Language Models are Few-Shot Learners.2020

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-schooling of Deep Bidirectional Transformers for Language Understanding.2019

[4] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy & Samuel R. Bowman. GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING. 2019.

[5] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang and Eduard Hovy. RACE: Large-scale Reading Comprehension Dataset from Examinations.2017

[6] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev and Percy Liang. SQuAD: a hundred,000+ Questions for Machine Comprehension of Text.2016

[7] Dr. Michael J. Garbade. A Simple Introduction to Natural Language Processing.2018

[8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A massive-scale hierarchical photo database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[9] Ashish Vaswani, Noam Shazee, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention Is All You Need. 2017

[10] Alex Sherstinsky. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. 2021

[11] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, Tuo Zhao. SMART: Robust and Efficient Fine-Tuning for Pre-educated Natural Language Models via Principled Regularized Optimization. 2021

[12] Prateek Joshi, July 21, 2020. Transfer Learning for NLP: Fine-Tuning BERT for Text Classification.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Che, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach.2019

[14] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS. 2020