Paper presentation narrator

The bidirectional encoder transformer (BERT) has shown striking improvements in various NLP tasks, and a series of variants have been proposed to further improve the performance of pre-trained language models. In this paper, the goal is to revisit the Chinese pre-trained language model to examine its effectiveness in non-English languages and release a Chinese pre-trained language models to the community. The paper also proposes a simple but effective model, called MacBERT, which improves on RoBERTa in several aspects, especially the use of MLM as a corrected (Mac) masked strategy. It conducted extensive experiments on eight Chinese NLP tasks to revisit existing pre-trained language models as well as the proposed MacBERT. Experiment results show that MacBERT achieves state-of-the-art performance on many NLP tasks, and also simplify the details with some findings to help future research.

Starting from BERT, the community has made great and rapid progress in optimizing pre-trained language models, such as ERNIE (Sun et al., 2019a), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019) , SpanBERT (Joshi et al., 2019), ALBERT (Lan et al. 2019), ELECTRA (Clark et al., 2020), etc. However, pre-training language models for TRANSFORMER-based (Vaswani et al., 2017) is not as simple as we used to train word embeddings or other traditional neural networks. Usually, to make feature-powerful BERT large models with a 24-layer transformer with 330 million parameters converge requires high-memory computing devices, such as TPUs, which are very expensive. On the other hand, although various pre-trained language models have been published, most of them are based on English and few powerful pre-trained language models have been built on other languages.

**P3 [Problems being solved]**

* Chinese and English are most used language in the world, but non-English natural language processing (NLP) faces several challenges, such as the lack of large-scale annotated datasets, domain adaptation, and the complexity of Chinese syntax and semantics.
* Traditional NLP methods often require hand-crafted features and domain-specific knowledge, which are time-consuming and expensive to develop.
* Pre-trained models, which learn generic representations of language from large amounts of unlabeled text, have shown great success in English NLP, but their effectiveness in Chinese NLP is still limited by several factors.
* This paper aims to address these issues and investigate the effectiveness of pre-trained models for various Chinese NLP tasks.

This paper aims to build a pre-trained language models for Chinese and release them to the public to facilitate the research community. We will revisit existing popular pre-trained language models and adapt them to Chinese to see if these models generalize well to languages other than English. It proposes a new pre-trained language model called MacBERT, which replaces the original MLM task and reduces the differences between the pre-training and fine-tuning phases. Extensive experiments were conducted on eight popular Chinese NLP datasets, ranging from sentence-level to document-level, such as machine reading comprehension, text classification, etc. The results show that the proposed MacBERT achieves significant approches in most tasks compared to other training language models, as well as examines the components with detailed research.

P5 [Related work - BERT]

* BERT (Bidirectional Encoder Representation from Transformer) has proven successful in natural language processing research. BERT pre-trains deep bidirectional representations by co-adjusting the left and right contexts of all Transformer layers. BERT consists of two main pre-training tasks: mask language model (MLM ) and next sentence prediction (NSP).
* ERNIE (Enhanced Representation through kNowledge IntEgration) is designed to optimize the masking process of BERT, which includes entity-level masking and phrase-level masking.

P6

* XLNet, which was based on Transformer-XL, is designed to alleviate the discrepancy of the pre-training and fine-tuning stage.
* RoBERTa (Robustly Optimized BERT Pretraining Approach) aims to adopt original BERT architecture but make much more precise modifications to show the powerfulness of BERT.
* ALBERT (A Lite BERT) primarily tackles the problems of higher memory assumption and slow training speed of BERT.
* ELECTRA (Efficiently Learning an Encoder that Classifiers Token Replacements Accurately) employs a new generator-discriminator framework that is similar to GAN.

P7 [Contributions]

* Extensive empirical studies are carried out to revisit the performance of **Chinese pre-trained models** on various tasks with careful analyses.
* The paper proposes a new pre-trained model called **MacBERT** that **mitigate the gap between the pre-training and fine-tuning stage** by **masking the word** with its **similar word**, which has proven to be effective on down-stream tasks.
* To further accelerate future research on Chinese NLP, it also creates and releases the Chinese pre-trained model series to the community.