Mengzi: Towards Lightweight yet Ingenious Pre-trained Models for Chinese

Zhuosheng Zhang\textsuperscript{1}, Hanqing Zhang\textsuperscript{2*}, Keming Chen\textsuperscript{3*}, Yuhang Guo\textsuperscript{4*}, Jingyun Hua\textsuperscript{5}, Yulong Wang\textsuperscript{5}, Ming Zhou\textsuperscript{5}

\textsuperscript{1}Shanghai Jiao Tong University, Shanghai, China \textsuperscript{2}Beijing Institute of Technology, Beijing, China \textsuperscript{3}Beijing Jiaotong University, Beijing, China \textsuperscript{4}Peking University, Beijing, China \textsuperscript{5}Langboat Technology, Beijing, China

zhangzs@sjtu.edu.cn, zhanghanqing@bit.edu.cn
20120341@bjtu.edu.cn, yuhangguo@pku.edu.cn
{huajingyun,wangyulong,zhouming}@chuangxin.com

Abstract

Although pre-trained models (PLMs) have achieved remarkable improvements in a wide range of NLP tasks, they are expensive in terms of time and resources. This calls for the study of training more efficient models with less computation but still ensures impressive performance. Instead of pursuing a larger scale, we are committed to developing lightweight yet more powerful models trained with equal or less computation and friendly to rapid deployment. This technical report releases our pre-trained model called Mengzi, which stands for a family of discriminative, generative, domain-specific, and multimodal pre-trained model variants, capable of a wide range of language and vision tasks. Compared with public Chinese PLMs, Mengzi is simple but more powerful. Our lightweight model has achieved new state-of-the-art results on the widely-used CLUE benchmark with our optimized pre-training and fine-tuning techniques. Without modifying the model architecture, our model can be easily employed as an alternative to existing PLMs. Our sources are available at \url{https://github.com/Langboat/Mengzi}.

1 Introduction

Using force to suppress others leads to superficial compromise. Genuine power only comes from practicality. (以力服人者，非心服也，力不赡也。权，然后知轻重；度，然后知长短。)  

\textsuperscript{Mencius} (372 BC - 289 BC)

Pre-trained models (PLMs) have greatly improved performance in a broad spectrum of natural language processing (NLP) tasks and stimulated the development to more practical scenarios [Radford et al., 2018; Peters et al., 2018; Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Lan et al., 2020; Clark et al., 2020]. Various trends have emerged recently: 1) bigger model and more data; 2) more efficient architecture and pre-training methodology; 3) domain- and task-aware pre-training 4) unification of vision and language modeling. With the promising advances above, a variety of pre-trained models have been developed for real-world applications. Despite their convenience of use, PLMs currently consume and require expensive resources and time, which hinders the wide range of practical applications. Therefore, modest-sized but powerful models, i.e., with only 100 million parameters, are much more preferred in light of resource cost and development circle, which desperately calls for the study of efficient methods. From the technical view, the major problems concerning lightweight language models lie within two aspects: effective training objectives that capture knowledge fast and efficient strategies that train language models quickly.

For model effectiveness, although PLMs have shown effectiveness in capturing syntax and semantic knowledge after pre-training [Hewitt and Manning, 2019; Ettinger, 2020], recent studies show that
the current models still suffer from under-fitting issues, and it remains challenging to train a powerful model with less computation (Rogers et al., 2020). Designing effective criteria for language modeling is one of the major topics in training pre-trained models, which decides how the model captures the knowledge from large-scale unlabeled data. Recent studies have investigated denoising strategies (Raffel et al., 2020; Lewis et al., 2020), model architecture (Yang et al., 2019), and auxiliary objectives (Lan et al., 2020; Joshi et al., 2020) to enhance the model capacity in pre-training. However, the cutting-edge researches mainly focus on English; there are a few studies in other languages like Chinese (Wei et al., 2019; Cui et al., 2020; Zhang et al., 2021b; Zeng et al., 2021). Besides, the application requirements in specific domains, e.g., financial analysis and multimodal tasks, further urge the development of effective Chinese pre-trained models.

To the end of efficiency, recent studies have investigated knowledge distillation (Sanh et al., 2019; Jiao et al., 2020; Wang et al., 2020) and model compression techniques (Gordon et al., 2020; Shen et al., 2020; Xu et al., 2020a). However, they are not optimal for real-world applications. Knowledge distillation methods train a light model with the guidance of a large-scale teacher model, which requires two stages of training, and training a teacher model still consumes massive computing resources. Similarly, model compression aims to train a simplified and optimized model from the original one without significantly diminished accuracy. The widely-used techniques include parameter sharing (Lan et al., 2020), module replacement (Xu et al., 2020a), pruning (Gordon et al., 2020), and quantization (Shen et al., 2020). Such a line of methods still needs abundant training. Also, these methods suffer from dramatic changes in the model architecture, so that it would be hard for easy real-world practice as it is incompatible with commonly deployed frameworks like the Transformers toolkit (Wolf et al., 2020).

In this work, instead of pursuing larger model size as the major goal of recent studies, we aim to provide more powerful but much resource-friendly models with a better performance compared with others on the same scale, which are of potential to rapid application to real scenarios and large-scale deployment. Therefore, we seek carefully optimized enhancement on the pre-training objectives, inspired by linguistic analysis and training acceleration, and are also free from a model architecture modification. As a result, we develop Mengzi, which is a family of discriminative, generative, domain-specific, and multimodal pre-trained model variants capable of a wide range of language and vision tasks. To keep consistent with public models and ensure easy application, we build our backbone model on top of the RoBERTa (Liu et al., 2019) following the same model settings. The main contributions of this work are three-fold:

1) We investigate various pre-training strategies to train lightweight language models, showing that well-designed objectives can further significantly improve the model capacity without the need to enlarge the model size.

2) We release Mengzi, including the discriminative, generative, financial, and multimodal model variants, capable of a wide range of language and vision tasks. The text encoders in these models only contain 103 million parameters, which we hope to facilitate the related studies for both academia and industry.

3) Extensive evaluates on widely-used benchmarks demonstrate that Mengzi achieves strong performance on a range of language understanding and generation tasks.

2 Backbone Encoder

Figure 1 shows the family of released Mengzi models and their connections: Mengzi-BERT-base, Mengzi-BERT-base-fin, Mengzi-T5-base, and Mengzi-Oscar-base are derivatives of Mengzi-BERT-base.
application scenario, they range from text-only language models to multimodal variants, from general-purpose training to domain-specific adaptation. The details will be demonstrated in Section 5. From a technical point of view, the last three ones can be regarded as the derivatives of Mengzi-BERT-base because their text encoders follow the same structure as Mengzi-BERT-base and are initialized by the pre-trained parameters of Mengzi-BERT-base. Therefore, in the following experimental parts, for simplicity, we only focus on the fundamental text-only encoder side and report our optimization techniques that are of general effectiveness.2

2.1 Setup

Data, algorithms, and computation are the key to powerful pre-trained language models. In the following part, we will present the details for training Mengzi in view of the three aspects.

Data Processing The pre-training corpus is derived from Chinese Wikipedia, Chinese News, and Common Crawl, with a 300GB data size in total. We clean the data by using exploratory data analysis techniques, i.e., removing HTML tags, URLs, e-mails, emoji, etc. Since there are simplified and traditional Chinese tokens in the original corpus, we convert traditional tokens into the simplified form using OpenCC. Duplicate articles are also removed.

Architecture RoBERTa (Liu et al., 2019) is leveraged as the initial backbone model for Mengzi pre-training. Our Mengzi architecture is based on the base size, where the model consists of 12 transformer layers, with the hidden size of 768, 12 attention heads, and 103M model parameters in total. We keep the model specification the same as the public one to ensure compatibility in real-world deployment and application. Following [Liu et al., 2019], we employ masked language modeling (MLM) as the major pre-training task.

Pre-training Details Our vocabulary contains 21,128 tokens. We limit the length of sentences in each batch to up to 512 tokens, and the batch size is 128. During pre-training, 15% words are randomly masked in each sequence for MLM prediction. We use a mixed-batch training procedure with LAMB optimizer (You et al., 2020), which involves two stages: the first 9/10 of the total epochs use a sequence length of 128, and the last 1/10 of the total epochs use a sequence length of 512. The batch sizes for the two stages are 16384 and 32768, respectively. We employ PostgreSQL to globally sample the training examples to avoid the imbalance of sample weight in the two-stage training. The overall pre-training process takes 1,000,000 steps. We use 32 NVIDIA Tesla 24GB 3090 Ti GPUs, with FP16 and deepspeed for training acceleration.

3 Experiments

3.1 Tasks

For downstream tasks for model evaluation, we use the Chinese Language Understanding Evaluation (CLUE) benchmark [Xu et al., 2020b], which consists of six different natural language understanding tasks: Ant Financial Question Matching (AFQMC), TouTiao Text Classification for News Titles (TNEWS), IFLYTEK (CO, 2019), Chinese-translated Multi-Genre Natural Language Inference (CMNLI), Chinese Winograd Schema Challenge (WSC), and Chinese Scientific Literature (CSL) and three machine reading comprehension (MRC) tasks: Chinese Machine Reading Comprehension (CMRC) 2018 [Cui et al., 2019], Chinese IDiom cloze test (CHID) [Zheng et al., 2019], and Chinese multiple-Choice machine reading Comprehension (C3) [Sun et al., 2019].

3.2 Setup

We build the downstream models for the natural language understanding tasks by adding a linear classifier on top of the “[CLS]” token to predict label probabilities. For the span-based question answering task, CMRC, we packed the question and passage tokens together with special tokens.

2We denote Mengzi-BERT-base as Mengzi for short in subsequent parts unless otherwise specified.

3https://github.com/BYVoid/OpenCC
4https://github.com/microsoft/DeepSpeed
Table 1: Results on the CLUE development datasets. The RoBERTa Dev results is from Cui et al. (2020). The test results except ours are from the CLUE leaderboard. Since there is a lack of accurate numbers of parameters in some public models, we use ~ to indicate the approximate scale. The standard evaluation metric is accuracy. For CMRC18, the reported score is calculated by the average of EM and F1 scores.

3.3 Implementation Details

For the fine-tuning experiments, we use Adam as our optimizer with an initial learning rate in {8e-6, 1e-5, 2e-5, 3e-5} with a warm-up rate of 0.1 and L2 weight decay of 0.01. The batch size is selected in {16, 24, 32}. The maximum number of epochs is set in [2, 5] depending on tasks. Texts are tokenized with a maximum length of 384 for MRC and 256 for other tasks.

3.4 Overall Results

Table 1 shows the performance of Mengzi on CLUE compared with public models. Compared with the RoBERTa baseline, we observe that Mengzi achieves consistent improvements on all the subtasks, showing that Mengzi is an effective alternative. For the public ranking on the test set, our large model has surpassed existing models for over three months. Mengzi not only far exceeds the performance of public models under the same model scale but also outperforms the largest Chinese model with 200 billion parameters, Pangu (Zeng et al., 2021).

Taking the Mengzi model as the backbone, we are interested in whether extra plug-in techniques, like auxiliary training objectives, would further improve the model capacity. In view of industrial deployment, assume that once the PLM is deployed, we would not spare manual labor to update the environment or model framework. The simplest way is to update the existing model weights with a new one. Therefore, we keep the basic criteria that those techniques should be independent of the model architecture, beneficial for pre-training, and dispensable during inference. To this end, we investigate the pre-training and fine-tuning techniques to enhance the capacity of Mengzi further.

5 The large model follows the large setting in Liu et al. (2019) and uses the same pre-training process as our base model.
4 Analysis

4.1 Pre-training Techniques

Linguistic-motivated Objectives Linguistic information has been shown effective for language modeling (Xu et al., 2021; Zhang et al., 2020). Inspired by LIMIT-BERT (Zhou et al., 2020), we employ part-of-speech (POS) and named entity (NE) sequence labeling tasks in conjunction with the original MLM and NSP objective during pre-training. POS and NE tags in the raw texts are annotated by spaCy.

Sequence Relationship Objectives To better model the pairwise information between sentences, we add the sentence order prediction (SOP) task (Lan et al., 2020) in model pre-training.

Dynamic Gradient Correction The widely-used MLM would cause the disturbance of original sentence structure, leading to the loss of semantics and improve the difficulty of prediction, inevitably resulting in insufficient and inefficient training. To alleviate the issue, we propose a series of dynamic gradient correction techniques to improve the model capacity, as well as the robustness.

4.2 Fine-tuning Strategies

Fine-tuning strategies are essential for downstream task performance. We report the results of the general and representative techniques that we have investigated, including knowledge distillation, transfer learning, choice smoothing, adversarial training, and data augmentation to further enhance the fine-tuning performance. Since those strategies mainly aim for competing on the leaderboard, the analysis is based on large models.

Knowledge Distillation We train a teacher model and employ the teacher model to guide the training of the student model. In detail, we calculate the Kullback–Leibler (KL) divergence of the contextualized hidden states from the teacher and student models, respectively, for the same input sequence. The divergence measures the similarity degree between the representations from the teacher and student models, which is minimized during fine-tuning, along with the original downstream task objective.

| Models                      | Accuracy (%) |
|-----------------------------|--------------|
| Baseline                    | 81.4         |
| + Knowledge Distillation    | 82.6         |
| + Data Augmentation         | 85.3         |

Table 2: Ablation results on the CMRC2018 dev set (average accuracy of F1 and EM scores).

Transfer Learning We leverage the parameters from the trained model on the CMNLI dataset to initialize the model training for related datasets like C3. For AFQMC, we use the model trained on LCQMC (Liu et al., 2018) and XNLI (the Chinese part) (Conneau et al., 2018) and initialize the model training on AFQMC. For CMNLI, we first use the OCNLI (Hu et al., 2020), CMNLI, SNLI (Bowman et al., 2015), MNLI (translated) (Nangia et al., 2017), and XNLI (Chinese part) (Conneau et al., 2018) for training an initial model, and then use it for initializing CMNLI model training.

| Models                      | Accuracy (%) |
|-----------------------------|--------------|
| Baseline                    | 75.2         |
| + Knowledge Distillation    | 77.1         |
| + Transfer Learning         | 77.3         |
| + Choice Smoothing          | 77.8         |

Table 3: Ablation results on the C3 dev set.

Choice Smoothing For multi-choice or classification tasks, combining different kinds of training objectives would lead to better performance (Zhang et al., 2021c). For each input example, we apply the cross-entropy and the binary cross-entropy as the loss functions and combine the loss from both sides to help the model learn features from different granularity.

https://github.com/explosion/spaCy

More details will be provided in our latter version.
**Adversarial Training**  To help the model generalize to unseen data, we apply a smoothness-inducing adversarial regularization technique following [Jiang et al. (2020)](https://www.journals.uchicago.edu/doi/10.1086/712644) to encourage the output of the model not to change much when injecting a small perturbation to the input.

**Data Augmentation**  Data augmentation has been widely used for training powerful models, especially for low-resource situations. For tasks like CHID, for each idiom in the given dictionary, we collect the related sentences from our corpora for pre-training and use them as extra training sources. For CMRC2018, we add the training data from DRC and SQuAD (translated) for augmentation. In addition, we use the original version in CLUEWSC2020 as supplemental training set for training WSC models.

| Models                  | Accuracy (%) |
|-------------------------|--------------|
| Baseline                | 85.6         |
| + Choice Smoothing      | 85.8         |
| + Adversarial Training  | 86.7         |
| + Data Augmentation     | 88.4         |

Table 4: Ablation results on the CHID dev set.

Tables 2-4 show the ablation results of the representative fine-tuning strategies, from which we have the following observations:

1) For MRC tasks like CMRC2018 and C^3, knowledge distillation can boost the benchmark performance of the student model with the guidance of teacher predictions.

2) Transfer learning boosts the model performance, which is consistent with practice for English GLUE benchmark ([Wang et al. (2019)](https://arxiv.org/abs/1810.04805)) by using Multi-Genre Natural Language Inference (MNLI) ([Williams et al. (2018)](https://arxiv.org/abs/1710.08560)) for initialization ([Liu et al. (2019)](https://arxiv.org/abs/1907.06012) [Lan et al. (2020)](https://arxiv.org/abs/2001.05787)) that would be beneficial for training models on small-scale datasets. However, we find that such a way of transfer learning is also helpful for large datasets like C^3.

3) Choice smoothing is effective for multi-choice tasks, which may provide fine-grained information from multi-label classification and binary classification, where multi-label classification captures the relationship between the label-wise predictions and binary classification models the prediction confidence for each label.

4) Adversarial training shows obvious improvements on CHID, which might be due to the benefit that using small perturbations in the embeddings might help improve the model’s robustness.

5) Data augmentation is an effective approach to enhance the model capacity. We observe substantial improvements in CMRC and CHID. However, finding a suitable augmentation technique remains a challenge.

## 5 Model Release

We release a family of pre-trained models covering discriminative, generative, multimodal, and financial application areas on the backbone of our ingenious encoders. The details of the release Mengzi models are presented in Table 5. Mengzi-BERT-base initializes the text encoders of Mengzi-BERT-base-fin, Mengzi-T5-base, and Mengzi-Oscar-base.

**Mengzi-BERT-base**  is a discriminative language model compatible with BERT as described in Section 2, which can be used for most NLP tasks like natural language understanding and machine reading comprehension.

**Mengzi-T5-base**  is a generative language model with a decoder module specialized for natural language generation tasks. The overall architecture follows T5 ([Raffel et al. (2020)](https://arxiv.org/abs/1910.10683)).

**Mengzi-BERT-base-fin**  is a domain-specific language model designed for financial scenarios, by continuing training Mengzi-BERT-base using our collected 20G financial corpus composed of financial news, announcements, and financial research reports.

---

*https://github.com/dbiir/UER-py/wiki/CLUE-Classification*
Table 5: Details of the release Mengzi models.

| Model                  | Size  | Features                                                                 | Tasks                                                                 | Corpus                      |
|------------------------|-------|---------------------------------------------------------------------------|----------------------------------------------------------------------|-----------------------------|
| Mengzi-BERT-base       | 103M  | Compatible with BERT as a stronger alternative, powered with linguistic-driven enhancements. | Text classification, entity recognition, relation extraction, reading comprehension, etc. | 300G internet corpus       |
| Mengzi-T5-base         | 220M  | More controllable text generation capacity, better performance than BERT structure and GPT structure. | Article generation, news generation, financial research report generation, etc. | 300G internet corpus       |
| Mengzi-BERT-base-fin   | 103M  | Specific for financial tasks by training Mengzi-BERT-base with financial corpus. | Financial news classification, sentiment analysis of financial research reports, etc. | 20G financial news, announcements, research reports |
| Mengzi-Oscar-base      | 103M  | Applicable to multimodal tasks, on top of Mengzi-BERT-base trained on millions of text-image pairs | Image caption, text retrieval, image retrieval, etc. | 300G internet corpus + 3.7M Chinese text-image pairs |

Table 6: Experimental results in the financial domain. The RoBERTa-wwm-ext baseline is from Cui et al. (2020). The evaluation results of information retrieval are reported by R@10/20. For entity recognition and relation extraction, the metric is F1. For entity linking, we use accuracy.

| Models                  | Information Retrieval | Entity Recognition | Relation Extraction | Entity Linking |
|-------------------------|-----------------------|--------------------|---------------------|---------------|
| RoBERTa-wwm-ext         | 90.20/92.90           | 88.11              | 77.44               | 93.40         |
| Mengzi-BERT-base        | 90.40/92.40           | 88.51              | 77.51               | 93.80         |
| Mengzi-BERT-base-fin    | 91.00/93.50           | 88.53              | 77.57               | 94.10         |

Mengzi-Oscar-base is a multimodal model effective for vision-language tasks, like image caption. The overall architecture follows Oscar (Li et al., 2020; Zhang et al., 2021a), which is a vision-language pre-training method to learn generic image-text representations for vision-language understanding and generation tasks. The language encoder is initialized by our Mengzi-BERT-base.

5.1 Exemplars and Assessment

Marketing Copywriting Figure 2 compares the quality of the generated marketing copywriting texts based on our Mengzi-T5-base model and GPT. Given the input title and keywords, the models are required to generate a corresponding descriptive passage. According to the generated examples, we observe that texts generated by our Mengzi-T5-base model contain much more details and keep fluency at the same time, indicating that generating texts using our model would benefit from satisfactory diversity fluency and coherence.

Financial Tasks We evaluate our Mengzi-BERT-base and Mengzi-BERT-base-fin in financial tasks, such as information retrieval, entity recognition, relation extraction, and entity linking. We extract the entities (e.g., events) from LUGE for the entity recognition task. For evaluation on the other tasks, we use our self-collected datasets. Results in Table 6 show that our methods are capable of the tasks specific for the financial domain, especially our Mengzi-BERT-base-fin yields the best performance.

https://aistudio.baidu.com/aistudio/competition/detail/46/0/task-definition.
Figure 2: Generated marketing copywriting examples from Mengzi-T5-base and GPT.

Image Caption | We compare the image caption performance of Mengzi-Oscar-base with the widely-used Automatic Alt Text technique used in Microsoft 365. Figure 3 shows the case studies based on randomly selected examples from the AIC-ICC Val set (Wu et al., 2017). We observe that our model generates more fluent and informative captions compared with the baseline.

---

10 https://support.microsoft.com/en-us/topic/everything-you-need-to-know-to-write-effective-alt-text-df98f884-ca3d-456c-807b-1a1fa82f5dc2.

5.2 How To Use

Our released Mengzi models are available at [https://github.com/Langboat/Mengzi](https://github.com/Langboat/Mengzi). Our models are also easily accessible by using the HuggingFace Transformers toolkit. For example, Mengzi-BERT-base is available by calling through the following scripts:

```python
from transformers import BertTokenizer, BertModel
tokenizer = BertTokenizer.from_pretrained("Langboat/mengzi-bert-base")
model = BertModel.from_pretrained("Langboat/mengzi-bert-base")
```

6 Conclusion

This technical report presents our exploration of training lightweight language model called Mengzi, which shows remarkable performance improvements compared with the same-sized or even larger-scale models. A series of pre-training and fine-tuning strategies have been verified to be effective for improving model benchmark results. Experimental results show that Mengzi achieves state-of-the-art performance with carefully designed training strategies. Without the modification of the model architecture, Mengzi is easy to be deployed as a powerful alternative to existing PLMs.

References

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In 8th International Conference on Learning Representations (ICLR 2020). OpenReview.net.

LTD IFLYTEK CO. 2019. Iflytek: a multiple categories chinese text classifier. competition official website.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475–2485.

[1] [https://huggingface.co/mengzi](https://huggingface.co/mengzi)
Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics, 8:34–48.

Mitchell Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the effects of weight pruning on transfer learning. In Proceedings of the 5th Workshop on Representation Learning for NLP, pages 143–155, Online. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. Ocnli: Original chinese natural language inference. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 3512–3526.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2177–2190, Online. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4163–4174, Online. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8:64–77.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In 8th International Conference on Learning Representations (ICLR 2020). OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision, pages 121–137. Springer.
Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. Lcqmc: A large-scale chinese question matching corpus. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1952–1962.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pre-training approach. arXiv preprint arXiv:1907.11692.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, pages 1–10.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21:1–67.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 8:842–866.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing in NeurIPS 2019.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2020. Q-bert: Hessian based ultra low precision quantization of bert. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8815–8821.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. Probing prior knowledge needed in challenging chinese machine reading comprehension. arXiv preprint arXiv:1904.09679.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. arXiv preprint arXiv:2002.10957.

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. Nezha: Neural contextualized representation for chinese language understanding. arXiv preprint arXiv:1909.00204.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art
natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. 2017. Ai challenger: A large-scale dataset for going deeper in image understanding. arXiv preprint arXiv:1711.06475.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. 2020a. BERT-of-the-seus: Compressing BERT by progressive module replacing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7859–7869, Online. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020b. CLUE: A Chinese language understanding evaluation benchmark. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4762–4772, Barcelona, Spain (Online).

Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. 2021. Syntax-enhanced pre-trained model. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5412–5422, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, pages 5754–5764.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training BERT in 76 minutes. In 8th International Conference on Learning Representations (ICLR 2020). OpenReview.net.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. arXiv preprint arXiv:2104.12369.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, et al. 2021a. Vinvl: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5579–5588.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, et al. 2021b. CPM: A large-scale generative chinese pre-trained language model. AI Open, 2:93–99.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020. SG-Net: Syntax guided transformer for language representation. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021c. Retrospective reader for machine reading comprehension. In The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021).

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 778–787, Florence, Italy. Association for Computational Linguistics.

Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. 2020. LIMIT-BERT: Linguistics informed multi-task BERT. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4450–4461, Online. Association for Computational Linguistics.