A Study of Pre-trained Language Models in Natural Language Processing

Jiajia Duan
Henan International Joint Laboratory of Theories and Key Technologies on Intelligence Networks
Henan University
Kaifeng, 475000, China
robots7@163.com

Hui Zhao
Educational Information Technology Laboratory
Henan University
Kaifeng, 475000, China
zhh@henu.edu.cn

Qian Zhou*
Art Department
Yellow River Water Conservancy Vocational and Technical College
Kaifeng, 475000, China
Corresponding author:
65706698@qq.com

Meikang Qiu
Department of Computer Science
Harrisburg University of S&T
Harrisburg, PA 17101, USA
mqiu@harrisburgu.edu

Meiqin Liu
College of Electrical Engineering
Zhejiang University
Hangzhou 310027, China
liumeiqin@zju.edu.cn

***Abstract*—Pre-trained Language Model (PLM) is a very popular topic in natural language processing (NLP). It is the rapid development of pre-trained language models (PLMs) that has led to the achievements of natural language today. In this article, we give a review of important PLMs. First, we generally introduce the development history and achievements of PLMs. Second, we present several extraordinary PLMs, including BERT, the variants of BERT, Multimodal PLMs, PLMs combined with Knowledge Graph and PLMs applied to natural language generation. In the end, we summarize and look into the future of PLMs. We expect this article will provide a practical guide for learners to understanding, using and developing PLMs with the abundant literature existing for various NLP tasks. (*Abstract*)**

Keywords—Pre-trained, Embedding, BERT, Cross-modal, KG, Natural Language Generation

I. INTRODUCTION

With the dawn of the Internet Age, especially in mobile system area [1], embedded system area [2] and modern multiprocessor technology [3], civilization has undergone profound, rapid-fire changes that we are experiencing more than ever today [4]. NLP research, in particular, has not evolved at the same pace as other technologies in the past decades of years. NLP is an important field in computer science and artificial intelligence, which aims to study various theories and methods that can realize effective communication between human and computer with natural language. Using natural language to communicate with computers has very important practical application significance as well as revolutionary theoretical significance. More recently, the application of pre-trained technology has been proved bring amount of NLP tasks to the state-of-the-art level.

Language Model (LM) is the basis of various NLP tasks. In its historical development, LM has gone through expert grammar rules models (up to the 1980s), statistical language models (up to 2000), and neural network language models (up to now). Early NLP systems are based primarily on manually written rules, which is time-consuming and laborious and cannot cover a variety of linguistic phenomena. In the 1980s, statistical LMs [5] were proposed, among which the N-gram Model is commonly used, but without obtaining the above long-term dependence and enough generalization ability. In 2003, Bengio [6] first employed neural network to solve language model, namely Neural Network Language Model (NNLM), which unifies the characteristic form of NLP that is embedding, while there are also some problems that require too much computation and cannot solve the long-term dependence problem. On the basis of NNLM, Recurrent Neural Network Language Model (RNNLM) [7] was proposed. It breaks the limitation of context window and uses the state of the hidden layer to summarize all historical context information, while it is difficult to capture longer distance dependent information.

NLP has become more popular as word embedding which is embedding a word into a multidimensional space. After training on a relatively large corpus, it will be used to capture the specific relationship between words. The well-known word vector training models include Word2vec [8] and GloVe [9]. Word embedding can be used to initialize the first layer of the downstream model and plus other functional layers to build the entire model. However, the early word embedding method has its limitations that cannot retain the information of each word context.

Pre-trained technology is first proposed in Computer Vision (CV) field. In the field of NLP, PLMs were first proposed by Dai, A. Etc. [10] in 2015, but only recently have they been shown to be effective for a large number of different types of tasks. Context2vec, a neural model that learns a generic embedding function for variable length contexts of target words, was proposed in 2016 [11]. Context2vec is a link between traditional word embedding and pre-trained language model. The Google Brain team [12] at the 2017 ICLR conference presented a general unsupervised learning method to improve the accuracy of sequence to sequence (seq2seq) models, proving that the pre-training process can directly improve the generalization ability of seq2seq model, and the importance and universality of pre-training are put forward again. Prajit Ramachandran et al. [13] proposed the then famous ELMo in 2018, showing the importance of deep bidirectional language models.

978-1-7281-6547-9/20/$31.00 ©2020 IEEE
DOI 10.1109/SmartCloud49737.2020.00030

ULMFiT, proposed Jeremy Howard et al. [15], is a key technology for fine-tuning language model.

Vaswani et al. [14] proposed the Transformer model to handle multiple NLP tasks with a single model in 2017. Based on the Transformer architecture, amount of PLMs began to appear at the end of 2018, refreshing abundant NLP tasks and forming new milestones. Transformer has been introduced into the training of language models, starting with GPT [16], BERT [17], which is very famous, then GPT2 [18], and even ALBERT [19], which in particular BERT, driving the progress of NLP as a whole.

PLMs have minimum training cost, better training parameters and higher training performance. This is especially important for some of more scarce tasks of training data. When the neural network parameters are very large, they may not be fully trained by the task of their own training data. Therefore, via the pre-training method, the model can be studied based on a better initial state, which can achieve higher performance.

This article contributes to BERT, BERT variants and some important PLMs with a brief. First, we review the birth and crucial points of BERT (see Chapter 2). Second, we introduce and summarize some improved model of BERT (see Chapter 3). Third, we present the BERT in multimodal (see Chapter 4). Then, we introduce the combined with KG of PLMs (see chapter 5). Lastly, we present MSAA (see Section 6.1) and UNILM (see Section 6.2) which are applied for Natural Language Generation.

II. BERT

The BERT model ushers in a new era of NLP. It is the king of PLMs in efficiency, usability and universality, outperforming large amount of PLMs. The appearance of BERT completely changes the relationship between pre-trained word vectors and downstream specific NLP tasks. This section presents the birth, design ideas and crucial points of BERT.

A. The birth of BERT

The biggest difference between BERT [17] and EMLo [13] is that the language model of EMLo [13] is LSTM, while BERT is Transformer [14]. The main disadvantages of the LSTM sequence model are as follows: First, it is a one-way language model. Even BiLSTM model, it only makes a simple addition at loss, namely it makes reasoning in order without considering the data in the other direction. Second, it is a sequence model with poor parallel computing power.

After ELMo [13], GPT [16] proposed by Alec Radford et al. employs Transformer [14] to encode, but its principle is similar to EMLo [13], which uses the preceding word to predict the next word. It is unidirectional and loses the information below. Therefore, BERT appeared, Transformer encoded and context-sensitive.

B. The training steps of BERT

BERT standing for Bidirectional Encoder Representations from Transformers is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers [17]. BERT has two main pre-training tasks, namely Masked LM and Next Sentence Prediction, which are its biggest innovation.

The first is Masked LM. BERT adopts a deep bidirectional model which standard conditional language models cannot be trained by. Therefore, BERT employs Masked LM to solve that issue, namely masking percentage of the input tokens at random, and then predicting those masked tokens. Further, the training data generator chooses 15% of the token positions at random for prediction [17].

The second is Next Sentence Prediction. Since tasks such as Question Answering (QA) [35] and Natural Language Inference (NLI) [36] are involved, a binarized next sentence prediction task is added to make the model understand the relationship between two sentences. When inputs to the training are sentences A and B, 50% of the time B is the actual next sentence that follows A, and 50% of the time it is a random sentence from the corpus [20]. After input these two sentences, the model predicts whether B is the next sentence of A. Pre-training can achieve 97-98% accuracy.

C. Strengths and weaknesses of BERT

The greatest strength of BERT is that it advances the state of the art for eleven NLP tasks with pre-training and fine-tuning [17]. Moreover, Transformer encoder with Self-attention Mechanism is adopted, and hence the upper and lower layers of the model are all directly connected with each other. Compared with RNN and LSTM, it is more efficient and can capture longer distance dependence.

The weaknesses of BERT are mainly related to the mask. First, the [mask] will not appear in the actual prediction, and excessive use of the [mask] in training will affect the performance of the model. Second, only 15% of each batch of tokens is predicted, so BERT converges more slowly than the left-right model which predicts each token.

III. THE IMPORTANT IMPROVED MODELS OF BERT

The birth of BERT brings NLP into a new world. In this chapter, we briefly introduce several important improved models of BERT, respectively introducing the principle, advantages and disadvantages of these models.

A. XLNet

XLNet, a generalized autoregressive pretraining method [20], This method combines autoregressive and autoencoder methods and presents a permutation and combination language model target, opening up an idea of how autoregressive language models can be introduced into the context.

Compared with BERT, XLNet still follows a two-stage process. The first stage is the language model pre-training stage. The second stage is the fine-tuning stage of task data. However, XLNet has improved the first phase by adopting Permutation Language Model.

The advantages of XLNet are as follows: (1) Learning bidirectional context by maximizing the probability expectation of factoring order; (2) The autoregression formula is added to overcome BERT's limitations. (3) Integrating the idea of the latest autoregressive model Transformer XL into the pre-training. However, XLNet also

has some drawbacks in the pre-training phase. It only makes use of the information in the co-occurance between tokens. when the two tokens have similar context, the final coding would have a very high similarity.

B. RoBERTa

RoBERTa, a robustly optimized BERT pretraining approach [21], is established on the basis of BERT language masking strategy, and modifies the key hyperparameters in BERT, including deleting the pre-training target for the next sentence, and training with greater Bacth size and learning rate.

RoBERTa makes more refined tuning of BERT. It chooses larger training model, batch, and data, and deletes the target predicted in the next sentence. It trains longer sequences, as well as dynamically changes the masking mode applied to the training data. However, Its training time is too long.

C. ALBERT

ALBERT, a lite BERT for self-supervised learning of language representation [19], proposes two parameter reduction techniques and introduces a self-supervised loss function, which improved BERT in three aspects, reduced the overall number of parameters, accelerated the training speed and increased the model effect.

ALBERT has the following advantages:

1) Decompose the embedded matrix to remove the relationship between word embedding and the size of the hidden layer, so as to expand the size of the hidden layer without increasing the model parameters;
2) Cross-layer parameter sharing, although slightly reducing the performance to a certain extent, can greatly reduce the model parameters, with a high yield;
3) Give up NSP and introduce SOP to learn the coherence between sentences more effectively.

However, ALBERT Base and Large produce effects that are not as good as BERT in the same inference time.

IV. BERT WITH CROSS-MODAL

Deep learning has implied a breakthrough in multimodal image-text [22]. Now, BERT has a stronger learning ability, which has been gradually applied to the multimodal field since 2019. The cross-modal solution based on pre-training can be divided into two branches, video-linguistic BERT and visual-linguistic BERT. The main difficulty is how to integrate non-text information into the framework of BERT.

A. Visual-linguistic BERT

The latest multimodal models are basically Transformer backbone, which is usually extended over a PLM to achieve better modal processing capability. Here, we introduce five major Visual-Linguistic BERT models, that are VisualBERT [23], ViLBERT [24], LXMERT [25], VL-BERT [26] and Unicoder-VL [27]. Their comparison is shown in Table 1.

VisualBERT, a simple and performant baseline for vision and lingual proposed by Li et al., consists of a stack of Transformer layers that implicitly align elements in an input text and regions in a related input image with self-attention [23]. In addition, two visual- lingual relational learning targets on image description data are used in the pre-training of VisualBERT. It can also establish connections between language elements and regions in an image without any explicit supervision and be sensitive to syntactic relationships and tracing.

ViLBERT (short for Vision-and-Language BERT), a model for learning task-agnostic joint representations of image content and natural language [24], is an extension of the BERT architecture to a multimodal model that supports two streams of input, preprocessing visual and text input respectively, and interacting in the Transformer layer of united attention. This model provides a new mind for learning the relationship between vision and lingual, which is no longer limited to the learning in the training process of a specific task, but takes the vision-language relationship as a pre-trainable and transferable model ability.

LXMERT, learning cross-modality encoder representations from Transformers [25], is used to learn the relationship between lingual and vision. The model is constructed based on Transformers encoder and novel cross-modal encoder. Then the model is pretrained with various pretraining tasks on a large data set of image and sentence pairs.

TABLE I. THE COMPARISON VISUAL-LINGUISTIC BERT MODELS

Method	Architecture	Pre-train Tasks	Downstream tasks
VisualBERT	single cross-modal Transformer	(1) sentence-image alignment (2) masked language modeling	(1) visual question answering (2) visual commonsense reasoning (3) natural language visual reasoning (4) grounding phases
ViLBERT	one single-modal Transformer(language) +one cross-modal Transformer(with restricted attention pattern)	(1) sentence-image alignment (2) masked language modeling (3) masked visual-feature classification	(1) visual question answering (2) visual commonsense reasoning (3) grounding referring expression (4) image retrieval (5) zero-shot image retrieval
LXMERT	two single-modal Transformer(vision & language respectively) + one cross-modal Transformer	(1) sentence-image alignment (2) masked language modeing (3) masked visual-feature classifucation (4) masked visul-feature regression (5) visual question answering	(1) visula question answering (2) natural language viasual reasoning
VL-BERT	single cross-modal Transformer	(1) masked language modeling (2) masked visual-feature classification	(1) visual question answering (2) visual commonsense reasoning (3) grounding referring expressions
Unicoder-VL	single cross-modal Transformer	(1) sentence-image alignment (2) masked language modeling (3) masked visual-feature classification	(1) image-text retrieval (2) zero-shot image-text retrieval

VL-BERT, pre-training of generic visual-linguistic representations [26], can be used as a new pre-trainable generic representation for visual-linguistic tasks. VL-BERT adopts the simple yet powerful Transformer model as the backbone, and extends it to take both visual and linguistic embedded features as input. VL-BERT proves to achieved the first place of single model on the leaderboard of the VCR benchmark.

Unicoder-VL, a universal encoder for vision and lingual by cross-modal pre-training [27], draws on the design ideas of cross-lingual and pre-trained models such as XLM and Unicoder, and both visual and linguistic contents will be introduced into a multi-tier Transformer as a cross-modal pre-training stage.

B. Video-linguistic BERT

The first model is VideoBERT [28] proposed by Google. VideoBERT is a typical model that combines BERT and video to learn cross-mode. It discretizes the feature vectors extracted from the video through clustering, and then adds visual token on the basis of text token to learn visual and text information together.

The VideoBERT training process is briefly reviewed. The first is to process the video text data. For the text, the text in the video is extracted with Automatic Speech Recognition tools, and the sentences are broken with LSTM-based language model, and then processed with the original BERT. Vector quantization is used to obtain video characteristics for video data. 1024-dimensional feature vectors obtained by pre-training are classified by hierarchical clustering, and then each segment is taken as the lower half of BERT inputs. Next, the two self-monitoring tasks MLM and NSP of original Bert are adjusted. MLM can be extended directly into Visual Token, while NSP becomes predictive in VideoBERT of whether text sequence and Visual Sequence are identical, namely whether both are extracted from the same video. VideoBERT can be applied to three downstream tasks: text-to-video, video-to-text and unimodal fashion.

VideoBERT is an interesting study. Based on BERT, video information is integrated into the model through pre-training. In the future, more information can be integrated through more different tasks to build a general AI model.

The second model is CBT [29]. Since VideoBERT organizes the visual features of continuous real value into a finite number of class centers through clustering, it may lose many details contained in the video. CBT no longer uses clustering to discretize real value continuous visual features, but directly uses real value vector Visual features to realize the multimodal transformation of BERT through fine-tuning of model algorithm.

V. PRETRAINED MODEL COMBINED WITH KNOWLEDGE GRAPH

When reading a specific field text, the average can only understand the words according to the context, while the expert can infer from the relevant domain knowledge. At present, BERT, GPT, XLNet [20] and other pre-trained models are all obtained from the pre-trained of corpus in the open field, which can read the general text but lack certain background knowledge of the text in the professional field as the average. One way to solve this issue is to use the professional corpus pre-trained model, but the process is very time-consuming and consuming computing resources, and usually difficult for ordinary researchers to realize. However, Knowledge Graph (KG) is a good solution. In this chapter, we give a brief introduction to three PLMs combined with KG, namely ERNIE(Tsinghua) [30], ERNIE(Baidu) [31] and K-BERT [32].

A. ERNIE (Tsinghua)

The enhanced language representation with informative entities (ERNIE) is obtained by utilizing both large-scale textual corpora and KGs to train a language that can make full use of lexical, syntactic and knowledge information simultaneously, so as to help it solve more complex and abstract NLP issues [30].

ERNIE is divided into two steps which are extracting knowledge information and training language model. First, the entity in the sentence is identified and matched with the entity in the KG. Then an independent TransE algorithm is utilized to obtain the entity vector, and the entity vector is embedded into BERT.

The biggest advantage of ERNIE (Tsinghua) is the addition of KGs. However, it also depends on the accuracy of NER extraction, and thus the model complexity is too high.

B. ERNIE (Baidu)

The enhanced representation through knowledge integration (ERNIE) implicitly learns something called entity relations and entity attributes by changing masking strategies rather than directly feeding some external knowledge information [31].

The ERNIE model is compared with the BERT model. In terms of ERNIE model architecture, there is basically no change. From the perspective of input data, the input data of ERNIE model is also token generated by plain text, except that text data needs to be segmented at different granularity through lexical analysis tools first. For the training task, the ERNIE model adds an extra task to predict the mask entity, predicting each token corresponding to the entity directly based on the tokens in the context.

The ERNIE model also has the following disadvantages: It is based on the co-occurrence of words and sentences to train the model, and ignores other valuable information such as lexical, syntactic and semantic information.

C. K-BERT

K-BERT, a knowledge-enabled language representation model with knowledge graphs (KGs), introduces soft-position and visual matrix to limit the impact of knowledge to overcome knowledge noise [32]. By introducing the triplet information of KG into the pre-training of BERT, K-BERT enables the model to acquire the semantic knowledge of special fields, so as to improve its performance in knowledge-driven tasks.

Compared with BERT, K-BERT has made some improvements. The first is the generation of the sentence tree. The input sentences pass through a knowledge layer and retrieve the KG to form a sentence tree with rich background knowledge. The second is the sentence tree sequence information which is expressed by soft-position encoding each token. The third is the encoding of sentence tree information. In order to introduce the structural information in the sentence tree into BERT, K-BERT modifies the self-attention in the Transformer-encoder.

The main innovation of K-BERT model lies in the soft-position and visible Matrix. In addition, the K-BERT model is consistent with Google BERT, which makes K-BERT compatible with BERT well and saves a lot of computing resources. By combining with KG, K-BERT clearly outperforms BERT in both domain-specific and open-domain tasks.

VI. Models applied to Natural Language Generation

BERT achieves great success in the field of natural language understanding, but does not perform well in the field of natural language generation, which is decided by the language model adopted by BERT in tis training. BERT only learns the ability of contextual representation of words, namely the ability to understand language, but not organize language. MASS [33] and UNILM [34] models proposed by Microsoft have tried the natural language generation aspect and obtain good results.

A. MASS

Masked Sequence to Sequence pre-training (MASS) for encoder-decoder based language generation adopts Encoder-Decoder framework to learn text generation, and both Encoder and Decoder use Transformer as feature extractors, which outperforms BERT and GPT in the sequence-to-sequence natural language generation task [33].

MASS has made some significant improvements. First, it introduces Seq2Seq to train. Second, it removes a single piece of tokens in a mask instead of a discrete mask to help the model's ability to generate language. Lastly, its encoder is a sequence dropped from a mask and decoder is a token that corresponds to a mask.

MASS has the following advantages:

1) Other words on the decoder side are blocked to encourage the decoder to extract information from the encoder side to assist in the prediction of continuous fragments, which can promote the joint training of encoder-attention-decoder structure;
2) In order to provide more useful information to the decoder, the encoder is forced to extract the semantics of unmasked words to improve the encoder ability to understand the text of source sequence;
3) To improve the language modeling capability of the decoder by enabling the decoder to predict continuous sequence fragments.

B. UNILM

UNILM, a new unified pre-trained language model, can be fine-tuned for both natural language understanding and generation tasks [34]. The basic unit of UNILM is still multi-layer Transformer, but the Transformer network is pre-trained on multiple language models, Unidirectional LM, Bidirectional LM and Seq2Seq LM.

The advantages of the UNILM are as follows:

1) UNILM unifies pre-training process so that only one Transformer language model can be used. This Transformer model shares parameters on different LMS, which eliminates the need for training and configuration on multiple LMS separately;
2) Parameter sharing among multiple LM makes the learned text representation have stronger generalization ability. Joint optimizations on different language model goals allow context to be used in different ways, and also slow down overfitting on a single LM;
3) In addition to being applicable to natural language understanding tasks, UNILM can also be used as a link-to-sequence LM to handle natural language generation tasks, such as abstract generation and problem generation.

In the future, we can continue to improve model performance with larger epochs, larger models, and more data, as well as extend unified pretraining models, such as support for cross-language tasks.

VII. Conclusion

PLMs have been successful in abundant NLP tasks recently, especially BERT and its improved models. This article reviews the recent achievements of PLMs. We briefly introduce BERT, some important improved models of BERT, across-models, models combined with KG and models used to natural language generation, as well as analyze the advantages and disadvantages of each model. Although PLMs have achieved advanced results, it is not the ultimate solution of NLP. The processing forms of natural language are semantics, reasoning, pragmatic and other stages. Most of the proposed PLMs can handle semantic problems well, but they are not good at inferencing tasks. In the future, how to use more pre-training data, how to compress and accelerate models, how to deal with long documents and how to deal with anti-attack, etc. all need to be studied and solved with more efforts, so that the field of natural language processing can reach a higher level.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No.61728303) and the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China (No. ICT20025).

References

[1] K. Gai, M. Qiu, H. Zhao, “Privacy-preserving data encryption strategy for big data in mobile cloud computing”, IEEE Transactions on Big Data, 2017.

[2] Z. Shao, C. Xue, Q. Zhuge, M. Qiu, B. Xiao, EHM. Sha, “Security protection and checking for embedded system integration against buffer overflow attacks via hardware/software”, IEEE Transactions on Computers 55 (4), 443-453, 2006.

[3] M. Qiu, Z. Jia, C. Xue, Z. Shao, EHM. Sha, “Voltage assignment with guaranteed probability satisfying timing constraint for real-time multiproceesor DSP”, The Journal of VLSI Signal Processing Systems for Signal, Image, and Video, 2007.

[4] Cambria. E and White. B, “A review of natural language processing research”, Computational Intelligence Magazine IEEE 9(2), 48-57, 2014.

[5] Jing. K, Xu. J and He. B, “A survey on neural network language models”, 2019.

[6] Bengio. Y, Ducharme. R, Vincent. P and Jauvin. C, “A neural probabilistic language model”, Journal of machine learning research 3, Feb, 2003, 1137–1155.

[7] Mikolov. T, Karafiát. M, Burget. L, Cernock´. y. J, Khudanpur. S, “Recurrent neural network based language model”, INTERSPEECH, pp. 1045–1048, 2010.

[8] Mikolov. T, Chen. K and Corrado. G, “Efficient Estimation of Word Representations in Vector Space”, Computerence, 2013.

[9] Pennington. J, Socher. R and Manning. C, “Glove: Global Vectors for Word Representation”, Conference on Empirical Methods in Natural Language Processing, 2014.

[10] Dai. A. M and Le. Q. V, “Semi-supervised sequence learning”, In Advances in Neural Information Processing Systems, 2015, pp. 3079–3087.

[11] Melamud. O, Goldberger. J and Dagan. I, "Context2vec: Learning Generic Context Embedding with Bidirectional LSTM", Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, 2016.

[12] Prajit. Ramachandran, Peter. J. Liu and Quoc. V. Le, "Unsupervised Pretraining for Sequence to Sequence Learning", 2016, arXiv preprint arXiv:1611.02683.

[13] Peters. M. E, Neumann. M, Iyyer. M, Gardner. M, Clark. C and Lee. K, "Deep contextualized word representations", 2018, arXiv preprint arXiv:1802.05365.

[14] Vaswani. A, Shazeer. N, Parmar. N, Uszkoreit. J, Jones. L, Gomez. A. N, Kaiser. Ł and Polosukhin. I, "Attention is all you need", In Advances in neural information processing systems, 5998–6008, 2017.

[15] Jeremy. Howar and Sebastian. Ruder, "Universal language model fine-tuning for text classification", 2018, arXiv preprint arXiv:1801.06146.

[16] Radford. A, Narasimhan. K, Salimans. T and Sutskever. I, "Improving language understanding by generative pre-training", Technical report, OpenAI, 2018.

[17] Devlin. J, Chang. M. W, Lee. K, and Toutanova, K, "BERT: Pre-training of deep bidirectional transformers for language understanding", 2018, arXiv preprint arXiv:1810.04805.

[18] Alec. Radford, Jeffrey. Wu, Rewon. Child, David. Luan, Dario. Amodei and Ily. Sutskever, "Language models are unsupervised multitask learners", Technical report, OpenAI, 2019.

[19] Zhenzhong. Lan, Mingda. Chen, Sebastian. Goodman, Kevin. Gimpel, Piyush. Sharma and Radu. Soricut, "Albert: A lite BERT for self-supervised learning of language representations", 2019, arXiv preprint arXiv:1909.11942.

[20] Zhilin. Yang, Zihang. Dai, Yiming. Yang, Jaime. G. Carbonell, Ruslan. Salakhutdinov and Quoc. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding", In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems2019, NeurIPS 2019, 8-14 December 2019, Vancouver , BC, Canada, 2019, pp. 5754–5764.

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

[22] Q. Guo, Y. Li, Y. Song, D. Wang and W. Chen, "Intelligent Fault Diagnosis Method Based on Full 1-D Convolutional Generative Adversarial Network," IEEE Transactions on Industrial Informatics, vol. 16, no. 3, Mar. 2020, pp. 2044-2053.

[23] Liunian. Harold. Li, Mark. Yatskar, Da. Yin, Cho-Jui. Hsieh and Kai-Wei. Chang, "VisualBERT: A simple and performant baseline for vision and language", 2019, arXiv preprint arXiv:1908.03557.

[24] Jiasen. Lu, Dhruv. Batra, Devi. Parikh and Stefan. Lee, "ViLBERT: Pre-training task agnostic visiolinguistic representations for vision-and-language tasks", NeurIPS, 2019.

[25] Tan. H and Bansal. M, "Lxmert: learning cross-modality encoder representations from transformers", 2019.

[26] Weijie. Su, Xizhou. Zhu, Yue. Cao, Bin. Li, Lewei. Lu, Furu. Wei and Jifeng. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations", 2019, arXiv preprint arXiv:1908.08530.

[27] Gen. Li, Nan. Duan, Yejian. Fang, Daxin. Jiang and Ming. Zhou, "Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training", 2019, arXiv preprint arXiv:1908.06066.

[28] Chen. Sun, Austin. Myers, Carl. Vondrick, Kevin. Murphy and Cordelia. Schmid, "VideoBERT: A joint model for video and language representation learning", ICCV, 2019.

[29] Sun. C, Baradel. F, Murphy and K, Schmid. C, "Learning video representations using contrastive bidirectional transformer", arXiv. 2019.

[30] Zhang. Z, Han. X, Liu. Z, Jiang, X, Sun. M and Liu. Q, "ERNIE: Enhanced language representation with informative entities", 2019, arXiv preprint arXiv:1905.07129.

[31] Sun. Y, Wang. S, Li. Y, Feng. S, Chen. X, Zhang. H, Tian. X, Zhu. D, Tian. H and Wu. H, "ERNIE: Enhanced representation through knowledge integration", 2019, arXiv preprint arXiv:1904.09223.

[32] Liu. W, Zhou. P, Zhao. Z, Wang. Z, Ju. Q and Deng. H, "K-BERT: enabling language representation with knowledge graph", 2019.

[33] Kaitao. Song, Xu. Tan, Tao. Qin, Jianfeng. Lu and Tieyan. Liu, "MASS: Masked sequence to sequence pre-training for language generation", In International Conference on Machine Learning, 2019.

[34] Li. Dong, Nan. Yang, Wenhui. Wang, Furu. Wei, Xiaodong. Liu, Yu. Wang, Jianfeng. Gao, Ming. Zhou and Hsiao-Wuen. Hon, "Unified language model pre-training for natural language understanding and generation", 2019, arXiv preprint arXiv:1905.03197.

[35] Pranav. Rajpurkar, Jian. Zhang, Konstantin. Lopyrev and Percy. Liang, "Squad: 100,000+ questions for machine comprehension of text", In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 383–2392.

[36] Alexis. Conneau, Douwe Kiela, Holger. Schwenk, Loıc. Barrault and Antoine. Bordes, "Supervised learning of universal sentence representations from natural language inference data", In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 670–680, Copenhagen, Denmark. Association for Computational Linguistics, 2017.