MSBD5018 Individual Project
Survey on Transformer-based Pretrained Language Models with Application Trend

Lam Chun Ting Jeff (1222973)
ctjlam@connect.ust.hk

Abstract

This paper will be a short survey on Transformer-based Pretrained Language models, for the course in HKUST MSBD5018 Natural Language Processing. With the rise of GPT and Bert, these models use the techniques of transfer learning and supervised learning in different applications. This paper will discuss the foundation technology that comes with tranformer-based pretrained language models, basic methodology category, and the application trend.

1 Introduction

The most interesting and influential moment in the history of natural language processing(NLP) development should be the start of the SCIgen(Stribling et al.), a an automatic CS-paper generator. It successfully faked the paper with the conference, using the technique of N-gram, proposed by Markov (Markov, 1913). However, this kind of old-fashioned approach cannot deal with complicated entity linkage. For example, apple can mean two things, a technology company, or a fruit.

With the rise in computational power in machines and big data technology, the trend of solving NLP problems tends to be using deep learning, and models are trained with large datasets such as Bert(Devlin et al., 2019), RoBERTa(Liu et al., 2019), and GPT(Radford and Narasimhan, 2018), which aiming to solve some complicated problems like entity linking as mentioned before.

In this paper, the basic foundation that comes with pretrained language models will be discussed first. Secondly, the category of transformer-based pretrained language models, and the downstream method will be demonstrated. Thirdly, some related application trends using transformer-based pretrained language models are listed and predicted. Lastly, there will be my own thoughts on the technology.

2 Motivation and Related Works

The content of the paper is mainly inspired by the survey of Transformer-based Pretrained Models in NLP (Kalyan et al., 2022). Among different topics being chosen for course group project presentation, toxic span detection accounts for the majority of topic choices. Students who chose this topic worked on the project using the concept of Bert(Devlin et al., 2019), RoBERTa(Liu et al., 2019), which are actually pretrained language models in solving the problem. Also, due to the outbreak of COVID-19, the clinical NLP has got more attention. Therefore, in the individual report, I would like to delve into Pre-trained Language Models, explaining the basic techniques and predicted trends.

3 Foundation

When it comes to deep learning, there are three main approaches, namely supervised learning, unsupervised learning, and reinforcement learning. In NLP, although there also exist some unsupervised approaches(Radford et al., 2019), and reinforcement learning methods(Uc-Cetina et al., 2022), these two will not be discussed in this paper. However, both corresponding research areas point out that self-supervised learning is the main-stream, with the fact that the nature of self-supervised learning is similar to unsupervised learning. Self-supervised learning is actually closely related to pretrained language model, which has laid the foundation of the whole learning process.

With a trained large scale language models, transfer learning will be used to adapt the model for use in different application, such as conversational AI(Gao et al., 2018), and clinical area (Wu et al., 2020). However, there is also other research pointing out that the models are not aligned with user intentions (Ouyang et al., 2022). Further discussion will be made in section 6.

3.1 Self-supervised Learning

Supervised learning is a method that when given an input, and labels, we could find a method that can output a good prediction, regardless the input is seen or unseen. Supervised learning performs well when there is a lot of data, but the trade-off is it requires huge cost in labeling the data. Self-supervised learning is different from that. Labels can be learnt during the objective-oriented computation, thus saving the labeling cost.

3.2 Transfer Learning

With the fact training from scratch is computationally heavy, and scarcity of data in public, transfer learning is a technique that does not requires training from scratch. Instead, it uses existing models to fine-tune in specific application. In the research done by OpenAI(Sharir et al., 2020), GPT2 training cost around $1.6 million. It is impossible for a normal developer to develop a new language model for certain applications using such a high cost. Therefore, the aim to reduce cost is the reason why transfer learning adopts existing models when it comes to application.

4 Pretrained Language Models

The most significant evolution in pretrained language models is transformer based. Transformer is introduced by Vaswani et al. (Vaswani et al., 2017), which overcomes weakness of traditional deep learning methods by CNN and RNN. The earlier models are GPT (Radford and Narasimhan, 2018) and Bert (Devlin et al., 2019), developed based on transformer. Transformer based pretrained language models have been classified into 4 categories by Kalyan et al, which are the pretraining corpus, model architecture, type of self-supervised learning, and extensions. (Kalyan et al., 2022).

For pretraining corpus, the pretrained models can be trained with different sources, and can be categorized into 4 domains, which are the general corpus, social media corpus, language-based corpus and domain specific. With different corpus, the model can be trained and extended into different application for several specific task. The table 1 shows and explains each corpus.

Corpus	Explanation
General	Less noisy data and in a more formal language
Social Media	More noisy data and in informal language
Language-based	monolingual or multilingual
Domain Specific	Aims in specific field which is not cover in general corpus

Table 1: Pretrained Language Models Corpus Summary

For architecture, it can be generalized into 3 types, which are encoder, decoder, and both.

For the types of self-supervised learning (SSL), it can be developed with the 4 techniques, which are generative SSL, contrastive SSL, adversarial SSL, and hybrid SSL. Refer to table 2 for explanation.

SSL Tech	Explanation
Generative	Helped with token prediction
Contrastive	Helped with token comparison
Adversarial	Helped with corrupted token prediction
Hybrid	Merge different techniques for different purpose

Table 2: Pretrained Language Models SSL Summary

For extensions, there are numbers of different category and list in table 3.

Extensions	Explanation
Compact	Aims to reduce the model size and enhances the efficiency
Character-based	Aims to overcome the rare and misspelled words
Green	Focus on environmentally-friendly method on reduce the cost when training
Sentence-based	Aims to generate better quality in sentence embedding
Tokenize-Free	Aims to overcome the drawbacks brought from tokenizer, tries without token
Large Scale	Focuses on large scale models, by increasing number parameters, training size, and etc.
Knowledge Enriched	Pretrained on large volume of data
Long-Sequence	Aims to reduce the complexity in case of long input sequence
Efficient	Aims to achieve more reliable models with smaller training data and model complexity

Table 3: Pretrained Language Models Extension Summary

5 Downstream Adaption

As mentioned in section 3.2, pretrained models is heavily trained. While there is lack of data, the model can be downstreamed to different application. There are three approaches in pretrained language modeling, namely feature based, fine-tuning based, and prompt-based tuning. Figure 1 shows all the current approaches in downstream adaption methods.

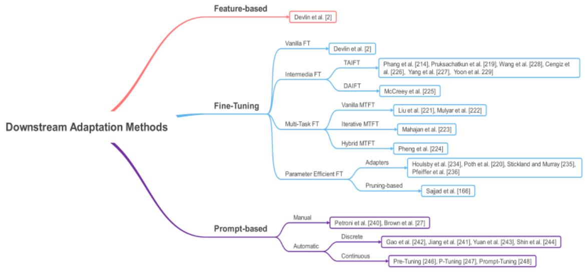

Downstream Adaptation Methods
Feature-based
Devlin et al. [2]
Fine-Tuning
Vanilla FT
Devlin et al. [2]
Intermedia FT
TAIFT
Phang et al. [214], Pruksachatkun et al. [219], Wang et al. [228], Cengiz et al. [226], Yang et al. [227], Yoon et al. 229]
DAIFT
McCreey et al. [225]
Multi-Task FT
Vanilla MTFT
Liu et al. [221], Mulyar et al. [222]
Iterative MTFT
Mahajan et al. [223]
Hybrid MTFT
Pheng et al. [224]
Parameter Efficient FT
Adapters
Houlsby et al. [234], Poth et al. [220], Stickland and Murray [235], Pfeiffer et al. [236]
Pruning-based
Sajjad et al. [166]
Prompt-based
Manual
Petroni et al. [240], Brown et al. [27]
Automatic
Discrete
Gao et al. [242], Jiang et al. [241], Yuan et al. [243], Shin et al. [244]
Continuous
Pre-Tuning [246], P-Tuning [247], Prompt-Tuning [248]

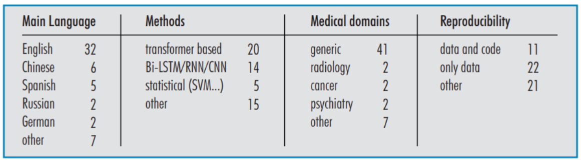
Figure 1: Downstream Adaptation Methods(Kalyan et al., 2022)

6 Application Trend

Apart from the area of natural language processing, transformer based pretrained learning models techniques has also been applied to computer vision (Dosovitskiy et al., 2021) and has led to a great success. Extending such techniques to different fields will definitely be the trend, such as Finance (Yang et al., 2020), Legal (Chalkidis et al., 2020), and Clinical (Wu et al., 2020)(Laparra et al., 2021). Due to the outbreak in COVID-19, the demand of medical system has sharply increased. In this section, the first part will be focused on a research on clinical usage, while some general concern and research in current pretrained learning models will be discussed in the second part.

6.1 Medical Application Trend

Before 2020, the main-stream of utilizing natural language processing on clinical fields is using RNN, and CNN. Application of such technology coincided with the exploration of potential in computer vision, when the potential of transformer based pretrained learning models had not been discovered. In the report written by Wu et al in 2020 (Wu et al., 2020), Deep Learning Method was foreseen to become the trend in later years. To compare, a report was published in 2021, which includes figure 2 below. Unlike the report stated in 2020, the proportion of transformer-based language model has become the mainstream among others methods. Papers published describing active learning layer on top of Bert (Shelmanov et al., 2021) are further reviewed. The seriousness of pandemic in 2021 might account for the relatively fewer papers being published in 2021. It is observed that a clinical NLP workshop, (Naumann) which has been held annually, was cancelled in 2021 but resumed in 2022. Once after the final review on 3rd June in the workshop, the latest trend will be noticed. It is believed that transformer based pretrained models will continue to be the mainstream in this year's research.



Main Language		Methods		Medical domains		Reproducibility	
English	32	transformer based	20	generic	41	data and code	11
Chinese	6	Bi-LSTM/RNN/CNN	14	radiology	2	only data	22
Spanish	5	statistical (SVM...)	5	cancer	2	other	21
Russian	2	other	15	psychiatry	2		
German	2			other	7		
other	7						

Figure 2: NLP in medical area(Laparra et al., 2021)

6.2 General Concerns and Solution Trends

Although pretrained language models perform pretty well in different aspects, concerns are raised regarding this methodology. As stated in the paper "On the Opportunities and Risks of Foundation Models" (Bommasani et al., 2021), concerns such as failures in certain models, AI safety, and so on, appeared. Apart from concerns on the social aspect, there are some foundation issues that when it comes to supervised learning, when giving inputs and predicting outputs using a large amount of data. As everything has become a black box, it is difficult to figure out how the AI makes specific decisions. The model size, scale, and anything related to the model parameters are under active research. (Bender et al.). Apart from researching and reviewing on the concerns of using such methodology, research on solution has also been done. As the garbage-in-garbage-out problem is applied with the model, there is research on the handling of data feedback, which is not desired manually. With merging the technology with reinforcement learning, OpenAI published a paper (Ouyang et al., 2022) on how to match results with user-intention using GPT-3. Therefore, apart from pointing out the problems in language models, the next development trend is solving problems brought by AI.

7 Discussion

In the field of NLP, the major task is text translation, text clasification, chatbots, machine translation, and so on. Different models are developed to accommodate to the different purposes of tasks. Pretrained language models are claimed to perform quite well, and therefore being downstreamed to different tasks. Whether the problems are solved using the right method are uncertain, as many other pretrained language models exist for different specific purposes.

In a anomaly detection problem, there is a paper stated that the majority of researchers are on the wrong track(Wu and Keogh). To solve this problem, deep learning is not required. Instead, the paper stated that the basic method solves the problem better than the deep learning method. It was based on a question that do a ground truth is really the truth. Although transformer based models are the state-of-the-art solution, it is difficult to explain the decision in the models. Moreover, training loss is actually based on the data itself. When some unseen data is not revealed, it is hard to predict the cases and make the models better generalize the problem. The most critical problem is, purpose-specified model is able to solve the specific tasks, but it cannot generalize all the queries.

Stepping one step backward, the reliable query methodology is actually relational database. Structured Query Language has successfully helped people in querying, and getting result. However the limitation is relational database only handles 1- dimensional data. There is actually a field called spatial, textual and multimedia databases (Zeitler). However, when deep learning technology like CNN, RNN is introduced, the field stopped growing. For certain tasks, we might find the answer in that field instead of NLP field, such as generating knowledge graph. Rather then solving problems in a see-then-solve manner, going backwards for universal solution may also be a solution.

8 Conclusion

In this paper, the basics and foundations of transformer-based pretrained language models are covered, with approaches and techniques used being categorized. The application trend of this technology in the clinical field, corresponding concerns, and solution have also been discussed. The paper ends with my personal opinion on the technology and the trend. I hope this paper can bring a better overview of this technology, and whether it is applicable in various fields.

References

Emily M Bender, Angelina McMillan-Major, Timnit Gebru, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big ...

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and et al. 2021. On the opportunities and risks of foundation models.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval*.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: A survey of transformer-based biomedical pretrained language models. *Journal of Biomedical Informatics*, 126:103982.

Egoitz Laparra, Aurelie Mascio, Sumithra Velupillai, and Timothy Miller. 2021. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. *Yearbook of Medical Informatics*, 30(01):239–244.

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 pretraining approach. Cite arxiv:1907.11692.

A. A. Markov. 1913. Essai d'une recherche statistique sur le texte du roman "Eugene Onegin" illustrant la liaison des epreuve en chain ('Example of a statistical investigation of the text of "Eugene Onegin" illustrating the dependence between samples in chain'). *Izvistia Imperatorskoi Akademii Nauk (Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg)*, 7:153–162. English translation by Morris Halle, 1956.

Tristan Naumann. [link].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and et al. 2022. Training language models to follow instructions with human feedback.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners - openai.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training nlp models: A concise overview.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, Alexander Panchenko, and et al. 2021. Active learning for sequence tagging with deep pre-trained models and bayesian uncertainty estimates.

Jeremy Stribling, Max Krohn, and Dan Aguayo. Scigen - an automatic cs paper generator.

Victor Uc-Cetina, Nicolas Navarro-Guerrero, Anabel Martin-Gonzalez, Cornelius Weber, and Stefan Wermter. 2022. Survey on reinforcement learning for language processing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Renjie Wu and Eamonn Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, and et al. 2020. Deep learning in clinical natural language processing: A methodical review.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications.

Erik Zeitler. Spatial, textual and multimedia databases - uppsala university.