Language Models are Unsupervised Multitask Learners

In this paper, Radford et al. (0) proposed an improved language model (LM) for natural language processing (NLP) tasks named GPT-2. While it is a direct successor of GPT (1) with the design largely unchanged, it demonstrated several improvements: a huge scale up to more than ten times of parameters and dataset (2), doubled context size, an increase in vocabulary and batch size, and a minor modification of adding normalisation layers. Radford et al. (0) benchmarked their modified LMs on various typical NLP tasks and demonstrated the potential of improving the performance of scaling up during the pre-training stage.

A typical machine learning task will involve data processing, feature selection, model selection, training and validation. Originally, features are selected by domain experts. This process essentially provided some prior knowledge and reduced data required to train the model. The choice of features is crucial for the performance, yet it becomes unrealistic for human with deeper models and larger datasets. Thus, End-to-End neural network model was proposed. In such models, only a few hyper-parameters require manual tuning. The model takes text as input directly and generates their own "features" with the first few layers in the neural networks during the training. While enabling the model to tackle complicated tasks, the tremendous amount of parameter introduced makes the model hard to converge, i.e., hard to train in practice, especially when the dataset is limited. One solution is to use a two-stage training strategy by adopting a pre-trained model as feature extractors.

In the prior work, Radford et al. (1) proposed GPT which combined transformers (3) (feature extractors) and unsupervised pre-training (two-stage training) (4). While it achieved state-of-the-art (SOTA) results in many tasks (1), it can be improved in several ways. This work chose to explore on its scalability. To improve the generalisability, Radford et al. (0) introduced a much larger model, GPT-2, which contains 15 times more parameters (up to 1.5B). Besides, Radford et al. (0) introduced WebText, a 40 GB corpus with 8 million documents that covers a large range of texts referenced by Reddit (0). A careful data cleaning procedure was performed. Interestingly, the little impurity of the dataset (it contains about 10MB of French text) enables GPT-2 to generate French in translation task to some extent.

This paper demonstrated the potential of unsupervised learning models. As Radford et al. (0) argued, the unsupervised model shares the same training target to the supervised versions, thus it can still achieve the desired outcome if it converges in practice. This belief was supported by the experiments conducted. Without further fine-tuning (in a zero-shot setting), GPT-2 performed well on various tasks in different domains. This further suggested that GPT-2 learnt a wide range of syntactical and semantical knowledge during the unsupervised pre-training process from WebText, which shows its potential as a general-purpose feature extractor in NLP tasks.

An interesting observation is the hints used to guide GPT-2 for desired output like `TL;DR:` (for summarisation) which suggests some common pattern observed in the corpus that may not be generalisable (0).

Comparing GPT-2 to another popular model Bert (5), it uses unidirectional transformers rather than bidirectional one. While Bert exceeds the performance of the original GPT (5) as a pre-trained extractor due to the adoption of bidirectional pre-training, GPT-2 instead attempted to improve via scaling up the model and dataset. Focusing more on the performance of the pre-trained extractor, this paper demonstrated the generalisation power of the vanilla extractor rather than engaging in the second stage ("fine-tune stage") based on a specific task.

This paper paved a promising path to improving LMs, which is to use a deeper pre-train model on a larger unsupervised dataset. It is possible to scale up significantly as the transformer model is easy to compute in parallel (6). Thus, improving performance is transformed from a research topic into an engineering problem to some extent.

Besides, this paper shows a powerful generative LM which can be used to generate a wide range of text on different tasks directly with few attempts (2). From the examples, it seems that the syntactic issues are solved completely. Thus, it may lead to a good text generation tool if semantical and contextual information can be inculcated into the model to strengthen the logical flow of the generated text.

Lastly, this paper evaluated the overlapping between training and testing sets in some commonly used datasets. This may account for some unexpected high performances. The analysis in section 3.7 also demonstrated the necessity of analysing the dataset in some circumstances.

As Radford et al. (0) indicated, the ceiling of this model with fine-tuning is still unclear. It is not too optimistic to expect this approach of using transformers and two-stage training with a large model and dataset can excel in a great range of tasks in NLP.

Reference

(0) Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI Blog*. 2019 Feb. Available From: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [Accessed 06th Nov 2019]

(1) Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018 Jun. Available From: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf [Accessed 24 Nov 2019]

(2) Radford A, Wu J, Amodei D, Amodei D, Clark J, Brundage M, Sutskever I. *Better Language Models and Their Implications.* Available From: https://openai.com/blog/better-language-models/ [Accessed 24 Nov 2019]

(3) Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: *Advances in neural information processing systems 2017*. p.5998-6008.

(4) Radford A. *Improving Language Understanding with Unsupervised Learning.* Available From: https://openai.com/blog/language-unsupervised/ [Accessed 24 Nov 2019]

(5) Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* [preprint] 2018 Oct 11. Available From: https://arxiv.org/abs/1810.04805 [Accessed 24 Nov 2019]

(6) Alammar J. *The Illustrated Transformer*. Available From: https://jalammar.github.io/illustrated-transformer/ [Accessed 24 Nov 2019]

Acknowledgement

Some other secondary literatures also helped me during composing this summary, including:

(7) Zhang J. *效果惊人的GPT 2.0模型：它告诉了我们什么*. Available From: https://zhuanlan.zhihu.com/p/56865533 [Accessed 24 Nov 2019]

(8) Zhang J. *从Word Embedding到Bert模型—自然语言处理中的预训练技术发展史*. Available From: https://zhuanlan.zhihu.com/p/49271699 [Accessed 24 Nov 2019]