Pretrained Language Models for Text Generation

Advanced Techniques in Artificial Intelligence

Unai Sainz de la Maza, Unai Salas and Adrián San Segundo

University of the Basque Country
2021/2022

INTRODUCTION

Text generation is one of the most important tasks in natural language processing (NLP). The main goal is to produce plausible and readable text in human language from input data like a sequence and keywords. There are a lot of applications where text generation is used, such as machine translation, dialogue systems, text summarization, text paraphrasing, and more.

Given the rise of deep learning, several works have been proposed to solve text generation tasks using deep neural networks, e.g., recurrent neural networks (RNN), convolutional neural networks (CNN), graph neural networks (GNN), and attention mechanisms. A major performance bottleneck of these deep neural networks lies in the lack of large datasets. Nowadays, the existing datasets are rather small. As deep neural networks tend to overfit within small datasets, they do not generalize well in practice.

One of the most widely used paradigms today are the pre-trained language models (PLMs). The key idea behind this is to first pretrain the models in large-scale corpus and then fine-tune these models in various downstream tasks (those supervised-learning tasks that utilize a pre-trained model or component). With this technique, we can avoid training a new model from scratch.

TASK DEFINITION AND APPLICATIONS

The core of text generation is to generate a sequence of discrete tokens $Y = \langle y1, \ldots, yj, \ldots, yn \rangle$ where each yj is drawn from a word vocabulary V. In most cases, text generation is conditioned on input data, such as attributes, text and structured data, which is denoted as X. Formally, the text generation task can be described as:

$P(Y|X) = P(y1, \ldots, yj, \ldots, yn|X)$.

There are several types of applications according to input X:

- If X is not provided, this task will be to generate text without any constraint (unconditional generation task).
- If X is a set of discrete attributes (e.g., topic words, sentiment labels), the task becomes topic-to-text generation or attribute-based generation. The input X plays the role of “guiding” the text generation.
- If X is structured data, like a knowledge graph or a table, the task will be

considered a data-to-text generation. It aims to generate descriptive text about structured data.

- If X is multimedia input like an image, the task becomes image caption (generate a description from an image), and if the input is an speech, the task becomes speech recognition (process human speech to text).
- If X is a text sequence, there exist several applications such as machine translation (translate from one language to another automatically), summarization (generate condensed summary of a long document), and dialogue system (converse with humans using natural language).

WHY PRETRAINED MODELS ARE CHOSEN

First of all, let's introduce the concept of pretrained models (PTMs), or more specifically, the pretrained language models (PLMs). These models are pre-trained with a mass of unlabelled text data and can be fine-tuned on downstreams tasks. Is a way to reuse the models trained (previously) for different tasks. They are basically constructed by chopping off the top layer that does actual classification, and fitting another layer on top of those features.

With the development of deep learning, the number of model parameters has increased rapidly. The much larger dataset is needed to fully train model parameters and prevent overfitting. However, building large-scale labeled datasets is a great challenge for most NLP tasks due to the extremely expensive annotation costs, especially for syntax and semantically related tasks.

In contrast, large-scale unlabeled corpora are relatively easy to construct. To leverage the huge unlabeled text data, we can first learn a good representation from them and then use these representations for other tasks.

Pretrained on large-scale corpus, PLMs encode massive linguistic and world knowledge into vast amounts of parameters, which can enhance the understanding of language and improve the generation quality.

The advantages of pre-training can be summarized as follows:

1. Pre-training on the huge text corpus can learn universal language representation and help with the downstream tasks.
2. Pre-training provides a better model initialization, which usually leads to a better generalization performance and speeds up convergence on the target task.

3. Pre-training can be regarded as a kind of regularization to avoid overfitting on small data.

STANDARD ARCHITECTURES

Nowadays, almost all PLMs employ the backbone of Transformer. For text generation tasks are actually two main arquitectures:

- **Encoder-decoder:** is the standard Transformer, composed with two stacks of Transformer blocks. The encoder is fed with an input sequence, while the decoder aims to generate the output sequence based on the encoder-decoder self-attention mechanism. Examples of models representative of this family:
    - BART.
    - mBART.
    - Marian.
    - T5.
- **Decoder-only:** this one employs a single Transformer decoder block, which is typically used for language modeling. They apply unidirectional self-attention masking that each token can only attend to previous tokens. Examples of models that use it:
    - CTRL.
    - GPT.
    - GPT-2.
    - Transformer XL.

Even if the transformer is an effective architecture for pre-training, its computational complexity (quadratic to the input length) limits it. Because of that, the actual GPUs can't handle sequences with more than 512 tokens.

FUTURE WORK

While PLMs have demonstrated their potential for a variety of NLP tasks, they are still challenges to overcome.

We present several directions of work that are being studied:

- **Upper bound of PLMs:** at the present, PLMs have not yet reached its upper bound. A vast majority of the current PLMs can be further improved by more

training steps and larger corpora. Also the current models can be deeper, however this leads us to develop more sophisticated and efficient training techniques.

- **Interpretability of PLMs:** due the deep non-linear architecture, the decision-making process is highly non-transparent. The interpretability and reliability of PLMs remain to be explored further in many aspects, it is fundamental to understand how PLMs work in order to improve and make better use of them.
- **Language-agnostic PLMs:** most of the PLMs for text generation are mainly based on English. This makes it challenging for these models to deal with non-English generation tasks. One of the most promising directions is how to reuse existing English-based PLMs for text generation in non-english languages.
- **Ethical concern:** nowadays, PLMs are pre-trained on large corpus crawled from the web without fine-grained filtering, these causes potential ethical issues. Further research is needed to avoid these ethical issues.

REFERENCES

1. QIU XiPeng1 , SUN TianXiang , XU YiGe , SHAO YunFan , DAI Ning & HUANG XuanJing. Pre-trained models for natural language processing: A survey. School of Computer Science, Fudan University, Shanghai 200433, China
2. Junyi Li , Tianyi Tang, Wayne Xin Zhao and Ji-Rong Wen. Pretrained Language Models for Text Generation: A Survey. Gaoling School of Artificial Intelligence, Renmin University of China.
3. Junyi Li , Tianyi Tang , Gaole He , Jinhao Jiang , Xiaoxuan Hu , Puzhao Xie , Zhipeng Chen , Zhuohao Yu , Wayne Xin Zhao, Ji-Rong Wen. TextBox: A Unified, Modularized, and Extensible Framework for Text Generation. Gaoling School of Artificial Intelligence, Renmin University of China