Notes on Large Language Models

Andreev, Anton

Le-Point-Technique, June/2023

abstract: In this technical brief, we review some common questionning about Large Language Models (LLM), such as: What are the types of LLM? What is fine-tuning? or What is prompting?

keywords: LLM, Prompting, Fine-tuning

Common applications of LLM

Common applications of LLM can be separated in two domains: dialog and natural language processsing (NLP) tasks.

In terms of dialog, applications regroup: - task-oriented dialog - conversational question answering - open domain dialog

In terms of NLP tasks, applications regroup: - sentence classification - named entity recognition - summarization - masked language modeling - many more ...

Popular LLM are ChatGPT (OpenAI, San Francisco, the US), LLaMa (Meta, Menlo Park, the US) or again Hugging Chat (Hugging Face, New York City, the US). You can find a list of LLMs here.

What are the types of language modelling?

- *Masked language modeling* (MLM) predicts a masked token in a sequence, and the model can attend to tokens bidirectionally. This means the model has full access to the tokens on the left and right. Masked language modeling is great for tasks that require a good contextual understanding of an entire sequence. MLM is used by bi-directional models like BERT, in which a certain percentage of words in the training set are masked, and the task of the model is to predict these missing words. Note that in this task, the model can see the words preceding as well as succeeding the missing word and that's why its called bi-directional. For pre-training BERT, another task called Next Sentence Prediction (NSP) was also used, but researchers have found its utility to be marginal and MLM being good enough for all practical purposes.
- *Causal language models* (CLM) are frequently used for text generation. You can use these models for creative applications like choosing your own text adventure or an intelligent coding assistant like Copilot or CodeParrot. Causal language modeling predicts the next token in a sequence of tokens, and the model can only attend to tokens on the left. This means the model cannot see future tokens. GPT-2 is an example of a causal language model.

- An example of auto-regressive model is GPT. These models are uni-directional (one directional) and they are trained to predict the next word without seeing the succeeding ones. This is because these auto-regressive models are specifically designed for better language generation, which makes it necessary for the model to be pre-trained in a uni-directional manner.

Comparison of the number of parameters in both types of models: - BERT with a few hundred million parameters (MLM). - GPT-like models which have several billion parameters (AR).

What is PLM?

"PLM" stands for pre-trained language models, as compared to from-scratch model. We will review here some technics to use these pre-trained models.

What is fine-tuning?

It is also commonly referred to as *transfer learning*. There are two types of fine-tuning: *task oriented* and *domain adaptation*:

- For many NLP applications involving Transformer models, you can simply take a pretrained model from the Hugging Face Hub and fine-tune it directly on your data for the task at hand. Provided that the corpus used for pretraining is not too different from the corpus used for fine-tuning, transfer learning will usually produce good results.
- However, there are a few cases where you will want to first fine-tune the language models on your data, before training a task-specific head. For example, if your dataset contains legal contracts or scientific articles, a vanilla Transformer model like BERT will typically treat the domain-specific words in your corpus as rare tokens, and the resulting performance may be less than satisfactory. By fine-tuning the language model on in-domain data you can boost the performance. This process of fine-tuning a pretrained language model on in-domain data is usually called domain adaptation.

What is prompting?

Prompt is where you can describe to the model what you want it to do. It like giving somebody an instruction. For example, we can give the instruction "Write a sentence using the word ocean" and it can respond with: "The ocean is vast and beautiful."

Prompting can be separated 3 categories (although not necessarily): - instruction: "Write a sentence", "Write a resume on the subject of . . . " - examples for few-shot learning - the model will take cues from the writing style and try to use the

same style of writing - context stuffing - these are specific details such as name, materials or places to be used in the output generation

At its base LLM do a word prediction trying to complete a phrase. This means that even the instructions we give can be taken as a text to be completed instead of producing the desired result. In such case you need to specify the instruction more clearly.

An example using Cohere Playground:

```
Name: Peter
Age: 38
Occupation: doctor
Description: Answer the question "What is your name?"
```

And the output is: `My name is Peter.`

Note how the first three lines are used to provide a context and the last line is the instruction.

What is in-context learning (few-shot learning)?

In-context or few-shot learning (FSL) through prompt design is the process of fine-tuning a model with a hundred or thousand input texts, the model just takes a few task-specific examples (<10 usually) as input, and then quickly figures out how to perform well on that task. Note that in this process, there is no update of the model weight that happens! No backpropagation and no gradient descent! Researchers have suggested that GPT-like models do some kind of Bayesian inference using the prompts. In simple words, using just a few examples, the model is able to figure out the context of the task, and then it makes the prediction using this information.

Few-Shot Learning (FSL) is a Machine Learning framework that enables a pre-trained model to generalize over new categories of data (that the pre-trained model has not seen during training) using only a few labeled samples per class.

Few-Shot learning (LSL) is where we prompt the model with a few examples so that it picks up on the pattern and style we are going for.

What is RLHF training?

At a high-level, reinforcement learning with human feedback (RLHF) is a technique for training large language models that has been critical to OpenAI's ChatGPT and InstructGPT models, Anthropic's Claude, and more. RLHF enables training LLMs to understand instructions and generate helpful responses.

What is federated learning?

Under federated learning, multiple people remotely share their data to collaboratively train a single deep learning model, improving on it iteratively, like a team

presentation or report. Each party downloads the model from a datacenter in the cloud, usually a pre-trained foundation model. They train it on their private data, then summarize and encrypt the model new configuration. The model updates are sent back to the cloud, decrypted, averaged, and integrated into the centralized model. Iteration after iteration, the collaborative training continues until the model is fully trained.

Classification of LLMs

LLM can be regrouped in two categories:

- Grounded - is the linking of concepts to context and within NLP context is often a knowledge base, images or conversation
- Instruct GPT - is a powerful tool that allows users to fine-tune the language generation capabilities of the GPT (Generative Pre-trained Transformer) model. Developed by OpenAI, Instruct GPT allows users to train the model on specific tasks and generate text that is tailored to their specific needs.
  - Example:

```
Correct spelling and grammar from the following text.
I do not wan to go
```

References

A. Andreev, 'List of Large Language Models and APIs'. https://github.com/tonc ho11/ML_examples/wiki/List-of-Large-Language-Models-and-APIs.

'What are foundation models? | IBM Research Blog'. https://research.ibm.com /blog/what-are-foundation-models.