

大语言模型使用与微调

主讲人: 於方仁

大纲

1. 大语言模型简介
2. 大语言模型的使用生态
3. fine-tuning
4. prompt-tuning

大语言模型简介

- 大语言模型（Large Language Models, LLM）是基于深度学习技术的自然语言处理模型。
 - 语言模型即用以处理自然语言文本的模型。
 - “大”体现在它的模型目标是普遍的，也可以不严谨的认为它的模型相对来说很大。
- 大语言模型分为Base model与Fine-tuned model两种。
- Base Model :
 1. GPT（Generative Pre-trained Transformer）系列：由 OpenAI 提出的一种基于 Transformer 模型的单向生成器，可以根据给定的上下文生成连贯的自然语言文本，广泛应用于自然语言生成领域。
 2. BERT（Bidirectional Encoder Representations from Transformers）系列：由 Google 提出的一种基于 Transformer 模型的双向编码器，可以学习到文本中词语之间的上下文关系和语义信息，广泛应用于自然语言理解领域。
 3. 其他, 例如XLNet, RoBERTa, ELECTRA
- Fine-tuned Model:
 - 数不胜数

大语言模型使用生态

1. 直接使用

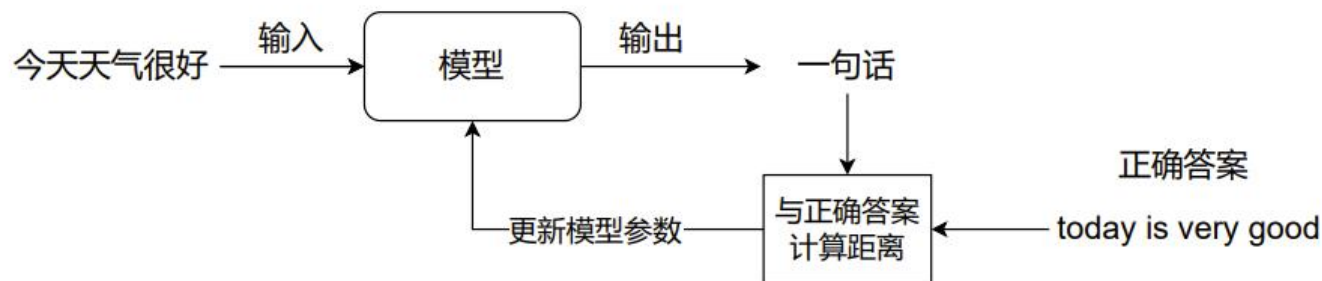
2. 微调

3. 相关网站

- [Introducing ChatGPT \(openai.com\)](#), [API Reference – OpenAI API](#)
- [Models – Hugging Face](#)

微调 (FINE-TUNING)

- 模型训练



用不同的句子对，不断重复这个过程使得模型的预测不断趋近正确答案。

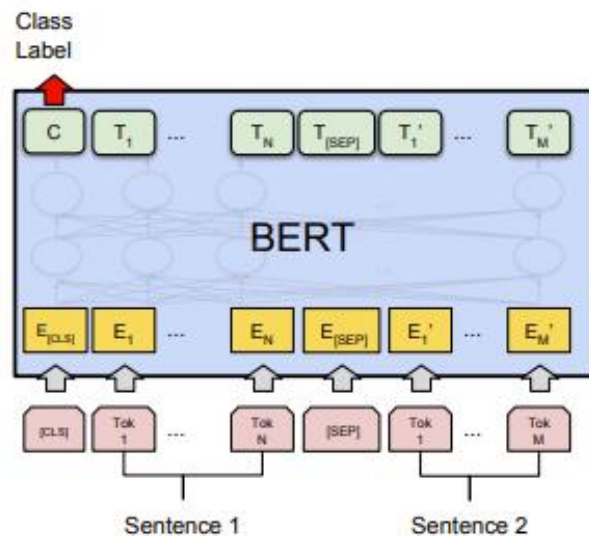
- 预训练与微调

用容易收集的数据及容易训练的任务先训练一个模型。该过程就被称为预训练。

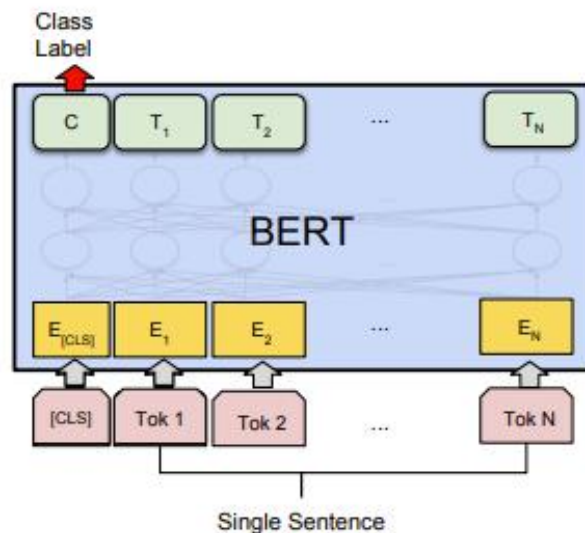


再用该模型用作别的任务继续训练。该过程则被称为微调。

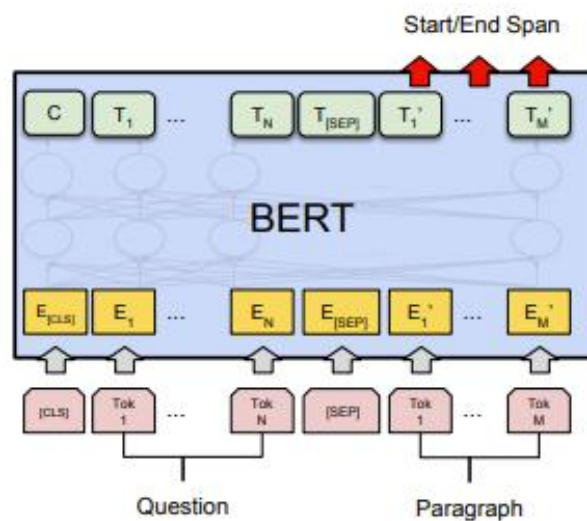
基于BERT的微调



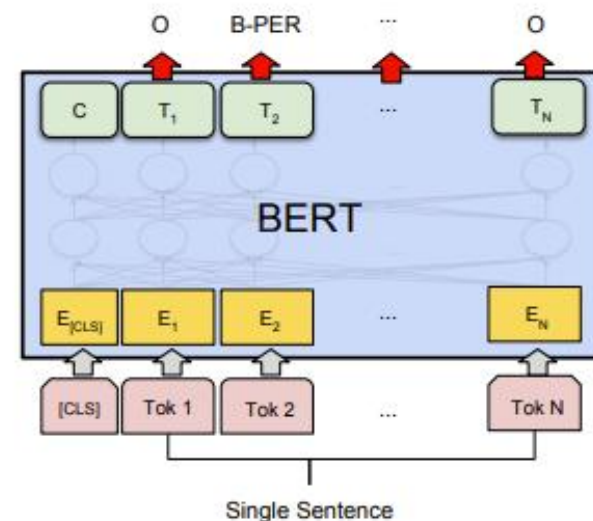
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA



(c) Question Answering Tasks:
SQuAD v1.1



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

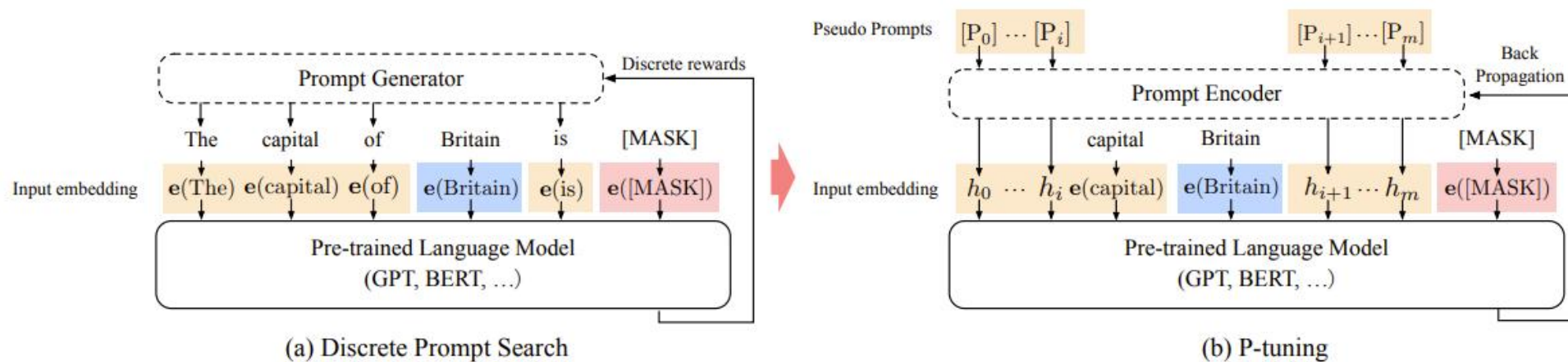
微调的瓶颈

1. LLM越来越大, 微调成本越来越大。
2. LLM越来越好, 微调的意义相对减弱。
3. 微调的难度上限太大。

PROMPT-TURNING

1. 提示词微调(prompt-turning), 针对提示词的微调训练, 在2021年正式提出。
2. Prompt, 可以认为是请求LLM时发送的文本, 或文本模板。
3. Prompt learning, 即如何有效使用Prompt。
4. Prompt-Turning, 指通过微调训练得到Prompt。

PROMPT-TURNING

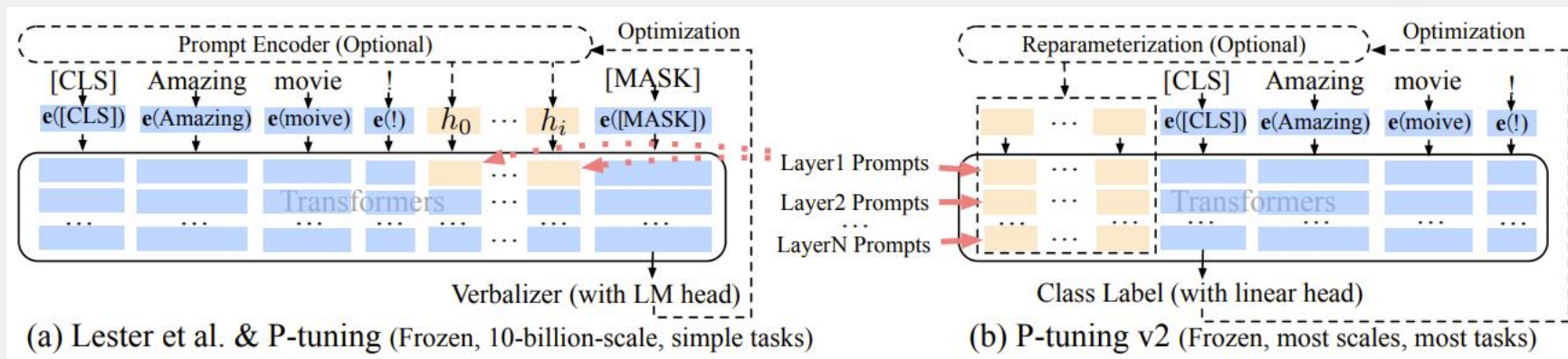


1. 将离散提示词用可迭代训练的连续向量替代。
2. 给定训练数据，训练代表提示词的向量。

PROMPT-TURNING 的优劣势

- 优势：
 1. 训练成本小，速度快
- 劣势：
 1. 效果不及fine-tuning。
 2. 太依赖LLM本身的质量，或者说LLM在目标任务的适合度。

PROMPT-TURNING V1



1. 与V1版本的区别:

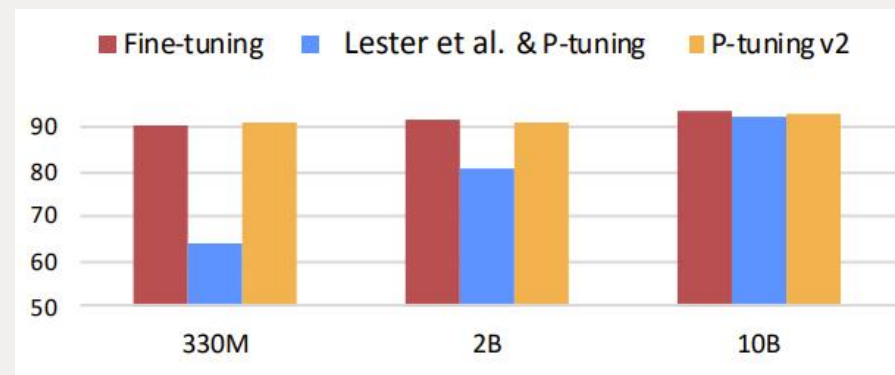
- V1的prompt只是当作模型最外层的输入。V2的prompt输入进模型每一层的网络中。

2. 优势:

- 效果直逼fine-tuning。
- 相对fine-tuning而言，速度与成本不是一个量级。

劣势:

- 相对V1而言，速度略慢。



结束