基础课程 3：大语言模型及其应用

一．语言模型：

1. 对语言进行建模：

- 建模语言中句子

$$P(X) = P(x_1, x_2, \ldots, x_n)$$

X1: 从前 有 座 山 。
X2: I am happy .
X3: I are happy .

- 判断不同种语言
  - P_ENG(X1) v.s. P_CHS(X1) ?
  - P_ENG(X2) v.s. P_CHS(X2) ?
- 判断同种语言的不同句子
  - P_ENG(X2) v.s. P_ENG (X3) ?

- 句子数量随着词汇量呈几何级数增长，因此难以直接统计估计

2. 根据条件概率公式：

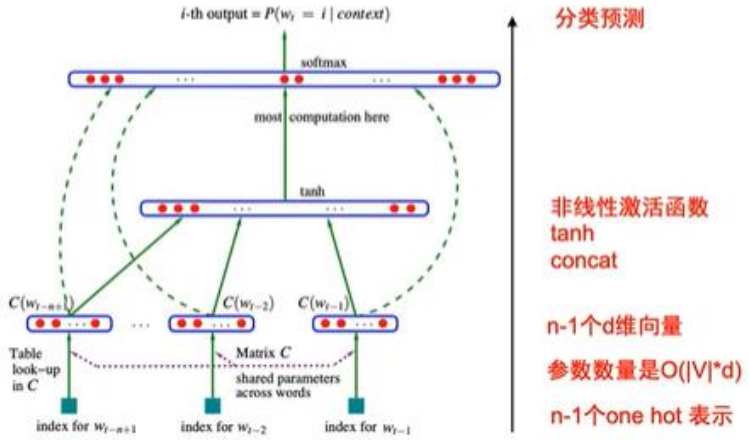
$$\begin{aligned} P(X) &= P(x_1, x_2, \ldots, x_n) \\ &= P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * \ldots \\ &\quad * P(x_n|x_1, x_2, \ldots, x_{n-1}) \\ &= \prod_{i=1}^{n} P(x_i|x_1, x_2, \ldots, x_{i-1}) \end{aligned}$$

预测下一个词
next token prediction

- P(从前 有 座 山) = P(从前) * P(有|从前)*
  P(座|从前, 有)*
  P(山|从前, 有, 座)

3. 神经网络语言模型（NPLM）

- 使用神经网络建模n-gram概率



A Neural Probabilistic Language Model. Bengio et al. NIPS 2001, ICML 2003

4. 神经网络语言模型的问题：

- **大量的参数需要大量的数据进行训练**
  - overfitting风险
- **大规模的模型需要较大的训练开销**
  - 计算资源
  - 训练时间

解决方法：预训练 pretrain

- 预训练embedding： word2vec
- 预训练contextualized representation: ELMo
- 预训练更多模型参数（多层Attention等）

5. 与训练和微调

预训练和微调

- **预训练的大规模参数可以为下游任务提供支持**
  - GPT、BERT等（12层Transformer block）
- **在下游任务的标记数据上进行进一步训练**
  - finetuning微调（防止灾难性遗忘）
  - 保持原有模型能力，学习完成新的任务

- **不同的任务可以从共同的参数基础开始训练**

6. From PLM to LLM

- **如果仅仅使用共同的参数，不进行进一步训练呢?**

- **部分任务有效，不及有监督学习；随着模型增大，能力在增强**

- **继续扩大模型规模**
- **300B token数据**

用上下文学习突破有监督学习的范式
(in context learning)

prompt和example/demonstration都起到了作用

即使不再改变参数，也能获得
逼近甚至超过监督学习的性能！

为什么模型能够具有这样的能力?

- **模型可能在大规模数据学习中进行了隐式的任务学习**
  - 无监督学习过程中可能包含大量的上下文学习样例
  - Meta learning with in context learning

显式的任务instruction学习可以取得更好的任务泛化效果！

7. CoT 推理能力

8. 听从人类指令的能力，完成人类的指令，结果倾向（Align）

让模型学会人类对回答的倾向/偏好！

二. 进一步提升大模型的能力

- **更长的上下文（long-context）**
- **混合专家模型（Mixture-of-Experts）**
- **高效参数训练（e.g. LoRA）**
- **高效推理（量化、剪枝、蒸馏）**
- **多模态模型（与语音、图片融合）**
- **……**

1. 使用 LLM

- **使用开源模型**
  - Llama、Qwen、GLM 、DeepSeek...
  - 规模相对小；需要算力进行部署；可以训练和更新
- **使用web/API**
  - GPT、Claude、Gemini、 ...
  - 规模更大能力更强；付费使用；以推理为主，不更新模型参数

- **大模型能力已经很强，但与实际任务要求仍有距离**
  - 改变参数：继续预训练、指令学习、偏好优化
  - 改变输入：检索增强、智能体

2. 改变模型参数

学习方式和代价

- Full-finetune：更新全部参数
- LoRA：仅更新新增的低秩矩阵

多语言翻译指令微调

- 训练模型执行 Translate [l1] xxx into [l2]
- 增强翻译（执行翻译指令）能力，可以泛化到未微调的语言

针对特定任务进行微调

- 微调Baicuan等模型，执行反馈、类比、合成等功能

改变模型偏好

- 将语言使用当做一种“倾向”

3. LLM as an Agent

- 什么是智能体?
  - 感知和动作 Perception & Action
  - 交互 Natural Language Interaction
  - 知识 Knowledge
  - 记忆 Memory
  - 推理和规划 Reasoning and Planning
  - 使用工具 Tool-using
  - .....

4. 检索增强（RAG）

检索增强

- 训练一个可以使用辅助信息的模型

使用检索结果作为额外输入

- **训练语言模型使用外部信息**

二．总结

总结和讨论

- **理解大语言模型**
  - 什么是大语言模型?
  - 什么是语言模型的基础能力?
  - 大语言模型为什么能获得更多能力?
- **使用大语言模型 v.s.通用智能AGI**
  - 目前仍需要通过特定手段进行增强!
- **建议练习：让LLM完成给定任务（如，讲故事，修改语法错误等）**
  - 考虑prompting、finetuning、RLHF及它们发挥的不同作用