2024/05/16 ゼミ

关于大语言模型发展现状的调研 + 研究目标

WU MENGZE 吳夢澤

曹研究室

埼玉工業大学大学院工学研究科

大语言模型とは?

大语言模型是一种具有 实现 **通用语言产生 和 其他自然语言处理** 任务（例如分类）的能力的计算模型。 它可以通过**获取输入文本**并反复**预测接下来**的token或单词来用于文本生成，是一种**生成式AI**。

大语言模型是一种人工神经网络。 截至目前，规模最大、生成能力最强的大语言模型是基于**Transformer**的体系结构、使用**解码器**架构的模型（例如**ChatGPT**）。其次是**编码-解码器**架构的模型（例如清华大学开发的**ChatGLM**）。**编码器**架构的模型则在如今缺乏竞争力，尽管它是Transformer体系模型的先驱（例如谷歌开发的**BERT**）

Transformer是由谷歌开发的深度学习体系结构，并基于2017年论文“Attention is all you need”的多头注意力机制。文本被转换为称为token的数值表示，每个token都通过查询 单词嵌入表 转换为向量。然后，在上下文窗口的范围内将每个token通过平行的多头注意力机制在上下文窗口范围内与上下文进行相关性处理，从而可以**放大关键token的信号，并减少不重要的token的信号。**

编码器（Encoder）的特点是它可以**考虑到输入序列中所有元素的上下文关系**，将输入序列转换为一组连续的表示，而这些表示考虑了输入序列中每个元素的上下文。

解码器（Decoder）的特点是它在处理当前元素时，**只考虑到该元素及其之前的元素**，不考虑其后的元素，这种机制被称为掩码自注意力。它基于前面已经存在的元素，生成下一个元素。

关于大语言模型发展现状的调研
发展历史与主要模型（国外、2023年之前）

发布时间	模型名称	作者	架构	许可证
2018-06	GPT-1	OpenAI	解码器	MIT
2018-11	BERT	Google	编码器	Apache-2.0
2019-02	GPT-2	OpenAI	解码器	MIT
2019-10	T5	Google	编码-解码器	Apache-2.0
	BART	Meta	编码-解码器	Apache-2.0
2020-05	GPT-3	OpenAI	解码器	专有
2021-06	GPT-3.5	OpenAI	解码器	专有
2021-12	Gopher	DeepMind	解码器	专有
2022-01	LaMDA	Google	解码器	专有
2022-04	PaLM	Google	解码器	专有
2022-07	BLOOM	BigScience	解码器	RAIL
2022-11	ChatGPT	OpenAI	解码器	专有

关于大语言模型发展现状的调研
发展历史与主要模型（国外、2023年）

发布时间	模型名称	作者	架构	许可证
2023-02	LLaMA	Meta	解码器	不得用于盈利
2023-03	GPT-4	OpenAI	解码器	专有
	Bard	Google	解码器	专有
2023-04	**Stable Diffusion XL**	**Stability AI**	**解码器**	**MIT、RAIL**
2023-05	PaLM-2	Google	解码器	专有
2023-07	LLaMA2	Meta	解码器	不得用于盈利
2023-09	Mistral-7B	Mistral AI	解码器	Apache-2.0
2023-11	Grok-0	xAI	解码器	专有
2023-12	Gemini	Google	解码器	专有
	Mixtral-8x7B	Mistral AI	解码器	Apache-2.0
	Phi-2	Microsoft	解码器	MIT

关于大语言模型发展现状的调研
发展历史与主要模型（国外、截止至2024年5月初）

发布时间	模型名称	作者	架构	许可证
2024-01	LLaMA-Pro-8B	Meta	解码器	不得用于盈利
2024-02	Sora	OpenAI	解码器	专有
	Gemini 1.5 Pro	Google	解码器	专有
2024-03	Claude 3	Anthropic	解码器	专有
	Grok-1	xAI	解码器	Apache-2.0
2024-04	LLaMA3	Meta	解码器	不得用于盈利
	Phi-3-mini	Microsoft	解码器	MIT
	Med-Gemini	**Google**	**解码器**	**专有**
2024-05	KAN	MIT等	新架构	MIT
	Meditron	**EPFL LLM Team**	**解码器**	**不得用于盈利**
	MemGPT	伯克利大学	解码器	Apache-2.0

关于大语言模型发展现状的调研
发展历史与主要模型（国内，截止至2024年5月初）

发布时间	模型名称	作者	架构	许可证
2021-12	ERNIE 3.0 Titan	百度	解码器	专有
2022-08	GLM-130B	清华大学	编码-解码器	Apache-2.0
2023-02	MOSS	复旦大学	解码器	Apache-2.0、AGPL-3.0
2023-03	ChatGLM	清华大学	编码-解码器	Apache-2.0
	ERNIE Bot	百度	解码器	专有
2023-05	VisualGLM-6B	清华大学	编码-解码器	Apache-2.0
2023-10	ERNIE 4.0	百度	解码器	专有
2024-03	aiXcoder-7B	北京大学	解码器	Apache-2.0
2024-04	Vidu	清华大学	编码-解码器?	专有
2024-05	DeepSeek-V2	深度求索	解码器?	MIT等

KAN vs MLP?

MLP——多层感知器、全连接前馈神经网络
KAN——Kolmogorov-Arnold Network
算法上的重要区别：MLP的激活函数（神经网络里的传递函数）通常是固定的，而KAN具有“可学习”的激活函数
KAN以200参数复现DeepMind的30万参数MLP发现数学定理的研究；在函数拟合、偏微分方程求解、处理凝聚态物理方面的任务比MLP效果好
可规避灾难性遗忘问题
训练速度慢（1、可学习的激活函数评估成本比固定激活函数成本更高；2、研究者暂未在效率优化上做足够多的尝试）
暂无成熟的方案适配Transformer架构（主要问题：1、如何用KAN取代Transformer中的MLP层，实现“Kansformer”；2、如何在GPU上高效实现KAN层）

医疗类大模型现状

今年四月底五月初出现了两个很强大的医疗大模型：Med-Gemini、Meditron

Meditron为基于Llama 3的开源模型；Med-Gemini是谷歌的专有模型

Meditron的参数规模较小，约80亿；Med-Gemini基于Gemini，预估参数规模1750

Med-Gemini、Meditron都是以医学问答与医学影像分析为主的大模型，但Med-Gemini还具备基因组风险的分析能力

目前似乎还不具备能通过分析EEG数据精准诊断神经退行性疾病，并给出诊疗建议的大语言模型？（EEGNet的诞生比谷歌的《Attention Is All You Need》论文发表还早一年多，应当不属于上述提到的大语言模型的范畴）

研究目标

近期目标1：使用近期流行的医疗大模型与我们研究室基于DenseNet开发的新冠胸片诊断系统分别对大量的胸片数据样本做分析，**对比两者 通过医学影像分析 诊断新冠的准确度。**

近期目标2：通过对现有的开源大语言模型（例如Stable Diffusion XL）进行基于指令的微调训练，来让模型能**按照指定需求**生成图像和视频（例如结合我们研究室正在做的课题，**生成一些包含与阿尔茨海默症、脑波等意象有关的图片或视频)**

远期目标：基于现有的开源大模型（例如Meditron），针对早期预测神经退行性疾病（例如阿尔茨海默症）、精神科疾病（例如抑郁症）、癫痫等的需求进行训练，**开发能依靠EEG信号数据高精度早期预测上述疾病的模型，并开发相应的图形化界面，制作成方便用户使用的跨平台app**

潜在难点与风险？

大商业公司和相关研究经验积累较多的学术机构可能率先作为“功能补充”实现该功能
——解决思路：尽可能在研究的突破口上围绕我们研究室的强势内容，BCI、神经科学研究、脑死判定、睡眠分析等等…

大语言模型的训练需要算力与时间，算力需要金钱成本，时间则包含机会成本…
根据近期研究，目前很多参数规模较小的大模型存在过拟合问题，可能意味着存在数据污染，例如Phi-3、Mistral 8x22B，这可能意味着：基于这些模型进行训练可能结果不佳，而且这些模型可能很快会被升级取代
——解决思路：
1、选择在医疗方面已经进行相对深入的微调的大模型进行训练；
2、试错本身也有学术意义，尝试在小模型中挑刺也许也是值得研究的内容

潜在的帮助开发的平台?

Ryght https://www.ryght.ai/

Hugging Face 🤗 针对 医疗保健和生命科学领域 开发的企业级生成式AI平台
“快速提升技能”——提高研究效率
“在多变的环境中随时了解最新情况”——避免重复发明车轮
“在众多方案中找到最经济的机器学习方案”——缩短开发时间，提高研究效率
“开发专注于安全性、隐私性、灵活性的高性能解决方案”——可能有利于研究与开发的创新性
企业级。。。（寻找平替？GPT-4O会是个好的助手吗？）

谢谢