作者：aaronxic
链接：https://www.zhihu.com/question/608820310/answer/3110733292
来源：知乎

最近写了个 LLM 的相关基础知识，分为上篇和下篇，共 3w 字左右。

上篇主要讲 LLM 的基座模型，包括了常见的 3 种 transformer 架构，encoder-only，encoder-decoder 和 decoder-only，包括了

- encoder-only：BERT
- encoder-decoder：T5, GLM-130B, UL2
- decoder-only：GPT 系列, LLaMA, OPT, PaLM, LaMDA, Chinchilla, BLOOM

同时串讲介绍了若干技术

- Norm 位置 3 种: Post-Norm， Pre-Norm 和 Sandwich-Norm
- Norm 方法 3 种: LayerNorm, DeepNorm 和 RMSNorm
- 激活函数 3 种: GeLU, GeGLU 和 SwiGLU
- PE 方法 6 种: Fixed Absolute, Learned Absolute, Fixed Relative, Learned Relative, RoPE, ALiBi

[Transformer 101 系列] 初探 LLM 基座模型 481 赞同 • 34 评论文章

下篇主要讲编程辅助应用和 ChatBot 是怎么炼成的，包括

- 编程辅助应用：Codex 和 AlphaCode
- ChatBot：InstructGPT, Bard, Claud, MOSS, ChatGLM2, LLaMA 系

[Transformer 101 系列] ChatBot 是怎么炼成的?56 赞同 • 3 评论文章

上述两篇是《Transformer 101》系列的第二篇和第三篇，该系列计划从以下五个方面对 transformer 进行介绍

- 算法 1: NLP 中的 transformer 网络结构
- 算法 2: CV 中的 transformer 网络结构
- 算法 3: 多模态下的 transformer 网络结构
- 训练: transformer 的分布式训练
- 部署: transformer 的 tvm 量化与推理

《Transformer 101 系列文章》

指标篇

[Transformer 101 系列] Perplexity 指标究竟是什么?77 赞同 • 7 评论文章

LLM 篇

[Transformer 101 系列] 初探 LLM 基座模型 481 赞同 • 34 评论文章
[Transformer 101 系列] ChatBot 是怎么炼成的?56 赞同 • 3 评论文章

多模态篇

[Transformer 101 系列] 多模态的大一统之路 206 赞同 • 6 评论文章

AIGC 篇

[Transformer 101 系列] AIGC 组成原理(上)60 赞同 • 4 评论文章
[Transformer 101 系列] AIGC 组成原理(下)45 赞同 • 2 评论文章

LLM 训练篇

[Transformer 101 系列] LLM 分布式训练面面观 142 赞同 • 12 评论文章

LLM 量化篇

[Transformer 101 系列] LLM 模型量化世界观(上)122 赞同 • 7 评论文章
[Transformer 101 系列] LLM 模型量化世界观(下)32 赞同 • 1 评论文章

LLM 部署加速

[Transformer 101 系列] 深入 LLM 投机采样(上)43 赞同 • 1 评论文章
[Transformer 101 系列] 深入 LLM 投机采样(下)33 赞同 • 0 评论文章

2024 智源大会速览

[2024 智源大会速览] 大语言模型和 AGI 篇 63 赞同 • 0 评论文章
[2024 智源大会速览] 多模态基础和 AI 工程篇 86 赞同 • 2 评论文章
[2024 智源大会速览] 视频生成篇 17 赞同 • 0 评论文章