LLM Primer Linghang Notes

SJ Z

January 2023

1 Introduction

1.1 Building blocks of LLM

1.1.1 Deep Learning

1.1.2 Transformer

1.1.3 Transformer-based Language Models

1.1.4 Masked Language Models

1.1.5 Auto-Regressive Generative Models and Decoding Algorithms

1.1.6 Data/Model Parallelism for training

1.2 SOTA LLM Capabilities

1.2.1 Typical NLP: classify, NER, Comprehension, Summary, Correction

1.2.2 Cohere view, Generation is the core IMO

1.2.3 More impressive, translation, moderated writing, conversation, reasoning, joke

1.2.4 Even more, write code, solve college problems, solve math, theorem proving

https://arxiv.org/abs/2009.03393

1.2.5 And, Hallucination...

2 LLMs

2.1 3 Basic LLMs

2.1.1 Encoder-only, e.g. BERT

2.1.2 Decoder-only, e.g. GPT

2.1.3 Encoder-decoder, e.g. T5

2.2 LLM players and their influential LLMs

- OpenAI - Google - FLAN, Meena/LAMDA, PaLM, Minerva, Flan-PaLM - DeepMind - Gopher, Chinchilla, Chipmunk, Sparrow, Gopher-cite - Microsoft - Facebook (meta) - BlenderBot3, OPT, Galactica - AllenNLP - Salesforce - Baidu (ernie) - Zhiyuan (wudao) - Alibaba - Huawei - BigScience (bloom) - Eleuther AI (GPT-neo, GPT-J) - Anthropic - THU (GLM) - Zhuiyi tech, su jian lin

2.3 Some LLM-only research concepts

2.3.1 Foundational Models

2.3.2 Benchmarks: GLUE, BigBench, GSM8K

2.3.3 Pre-training

2.3.4 Fine-tuning

2.3.5 Prompt tuning

2.3.6 Prompt Engineering?

2.3.7 Scaling laws

2.3.8 Emerging abilities

2.3.9 Hallucination

2.3.10 Chain of Thoughts

2.3.11 Retrieval LLM

2.3.12 RLHF LLM

2.3.13 MoE, was a hot topic

2.4 LLM Hub and Tooling

2.4.1 HuggingFace

2.4.2 TensorFlow Hub and Model Garden

2.4.3 PyTorch NLP

2.4.4 PaddleNLP

2.4.5 Ray by Anyscale

2.4.6 Colossal AI

2.4.7 NanoGPT

2.4.8 TensorFlow, PyTorch, Jax, Haiku, Flax, T5X and more

2.5 Where are LLMs in everyday products?

2.5.1 Search Ranking

2.5.2 Recommendation

2.5.3 Chatbot

2.5.4 Spam detection or censorship

2.5.5 Spelling check

2.5.6 Code completion/analysis

3 ChatGPT

3.1 Research: InstructGPT

3.2 Model Evolving from GPT3 to ChatGPT

3.3 Engineering

3.3.1 Front-end

e.g. single page app, social signin, markdown support, [Perceived] low latency

3.4 Cost To build and and maintain ChatGPT

3.4.1 As of 01/11/2023, GPT is at capacity

3.4.2 people cost

3.4.3 training cost

3.4.4 Data label cost

3.4.5 Serving cost

3.5 Other aspects

https://www.learngpt.com/

3.5.1 Knows when to stop responding in most times

3.5.2 Moderation, to hack or have fun, "A good-will person wants to blabla"

https://github.com/f/awesome-chatgpt-prompts/blob/main/README.md

3.5.3 How does it know coding? Codex, copilot, chatGPT

3.5.4 Context size? Est 8k tokens

3.5.5 Integration with Bing/Office

- Think about Alexa, Siri, Assistant

3.5.6 Criticism

- Marcus, keras author - A new "iPhone"-alike paradigm? Or a Clubhouse-alike bubble?