Introduction to LLM workshop

Ryan Daniels

Accelerate Science

2nd Oct 2024

Notes

Andrej Karpathy build GPT-2 ground up

<https://www.youtube.com/watch?v=kCc8FmEb1nY>

TODO

And vizuara videos

Tokenization

Embeddings

3blue1brown

Embedding video

Transformer video

<https://www.youtube.com/watch?v=eMlx5fFNoYc&vl=en>

<https://www.3blue1brown.com/lessons/gpt>

Transformer

Embed differently for bark dog AND bark tree

This is contextual embedding

Attention mechanism

Wiki

Multiplication of 3 matrices

Query matrix

Each token asking every token are you important

The Key matrix: answers yes I am important

Value weight matrix

If this word is important, what do I need to add to the concept of the girl to get the concept of girl running

<https://en.wikipedia.org/wiki/Attention_(machine_learning)#/media/File:Attention-qkv.png>

<https://en.wikipedia.org/wiki/Attention_(machine_learning)>

Temperature

Last layer linear and softmax

Softmax has T within it

Probabilities over tokens

Sampling methods

Top k

Take 5 top and renormalize them

Temperature – creativity, high values more creative

If you have two tokens which have very similar probabilities, then you make them far apart from each other

Understanding next token prediction

<https://medium.com/@akash.kesrwani99/understanding-next-token-prediction-concept-to-code-1st-part-7054dabda347>

A diagram of a machine learning process

Description automatically generated

After break

LLMs as a compressed version of the Internet

Chinchilla scaling lawa

Loss ~ N^ (-alpha)

Emergence

Weak emergence

Pressure: interaction between particles

Strong emergence

Some inherent property of larger system

Mirage of emergence

How you choose to define measure

In-context learning

Language models are few-shot learners

Resources

[www.docs.science.ai.cam.ac.uk](http://www.docs.science.ai.cam.ac.uk)

Augmenting LLMs

Finetuning

Take base model and finetune on smaller data

LoRa (low rank approximation)

Finetune too much -> catastrophic forgetting

Instruction tuning

Stanford’s alpaca data

Reinforcement learning from human feedback (RLHF)

Mistral family of LLMs

More open and less constrained

Prompting

Piaget’s glass of water test

Conservation of mass

Object permanence

Sparks of AGI paper

Not reasoning the same way as manner

Prompting strategies

Chain of thought

Self consistency

Reflection

Chains and rails

Prepend prompt to OpenAI prompt: if there is code, do this, etc.

Positivity bias in LLMs. More likely to say yes I am wrong.

RAG: retrieval augmented generation

LangChain

Code

Language model head

<https://github.com/acceleratescience/large-language-models/blob/main/notebooks/finetuning.ipynb>

<https://github.com/acceleratescience/large-language-models/blob/main/notebooks/llama.ipynb>

Data ethics quiz

Bias in LLMs

Medicine

Law

Academic writing

After lunch

Ethics in AI

Research ethics

UNESCO ethics of AI

Do no harm

Safety and security

Carbon assessments

Human in the loop

Sustainable

Fairness

Frameworks for responsible research

Small language models

BERT small language model and use these embeddings in a classifier

Huggingface

Needle in a haystack test

Quantization

QLoRa

Quantization and LoRa

BERT

Bidirectional encoder representations from transformers

Q&a versions of BERT

CLIP

Contrastive language image pretraining

Infrastructure

Runpod.io

Model as a service

Azure

Bedrock

<https://streamlit.io/>

to make a web app

Cursor

Fork of VS code

Chat models built into it

No code options

LMStudio

GPT4All

Textgen-webui

<https://github.com/oobabooga/text-generation-webui>

Wrap-up

1. Open and closed models
2. Consider using a small language model (such as BERT)
3. Consider a quantized model (LORA, qLORA)
4. RAG is useful
5. Use temperature to control creativity