Jakob Hördt

Mapping the LLM Ecosystem

Software, Models, Frameworks, APIs

Models

- Blobs of numbers
- Open weight / open source models you can run yourself.
- Notable open weight base models
  - Llama 3 by Meta
  - Mixtral by Mistral AI
  - Qwen by Alibaba
- Vision Language Models (VLMs), built atop Transformer Architecture
- LoRA adapters make serving and training fine-tuned models cheap
  https://arxiv.org/abs/2106.09685

Deep Learning Frameworks

- Linear Algebra libraries
- automatic differentiation (autograd) and back-propagation
- Make architecture definition easy
- Notable frameworks
- Training and Serving
  - PyTorch, still most popular `https://pytorch.org/`
  - TensorFlow `https://www.tensorflow.org/`
  - Jax, newer, higher performing `https://jax.readthedocs.io/en/latest/`
  - Keras, higher level, can uses the above three `https://keras.io/`

Inference Engines

- Deep learning frameworks not optimal for serving
- llama.cpp https://github.com/ggerganov/llama.cpp
  - original goal: run llama on Macbook using single C++ file
  - Now supports wide variety of hardware
  - GGUF model format which eg. PyTorch models can be converted to.
  - https://llamafile.ai/ project by Mozilla
  - great for running locally
- vLLM https://vllm.ai/
  - Python library
  - high performance using novel paged attention
  - used in https://chat-ai.academiccloud.de/
  - OpenAI compatible API

OpenAI API

- The “narrow waist” of LLMs
- HTTP API by OpenAI
- Defacto standard for accessing LLMs remotely
- `https://platform.openai.com/docs/api-reference/introduction`

Applications

- https://www.librechat.ai/
- https://ollama.com/
- Most popular CLI https://llm.datasette.io/en/stable/
- code completion https://docs.continue.dev/intro

Retrieval Augmented Generation (RAG) / Embeddings

- RAG
  - https://docs.privategpt.dev/
  - https://github.com/h2oai/h2ogpt
- Embeddings Databases
  - Chroma
  - PGVector uses Postgres

Frameworks

- Huggingface defacto standard platform for Models
- Gradio
  - Python/Javascript framework for AI user interfaces
  - Can use Gradio as an HTTP API as well
  - by huggingface
- Langchain
  - Similar Application framework for Python/JS

Notable People in LLMs (highly subjective)

- Grant Sanderson aka 3Blue1Brown
  - Math YouTuber with fantastic transformer explanation
  - https://www.youtube.com/watch?v=wjZofJX0v4M
- Robert Miles
  - Leading AI Safety researcher
  - AI Safety YouTube: https://www.youtube.com/@RobertMilesAI
- Simon Willison
  - Software Engineer focusing on LLMs, great blog
  - Great introduction to LLMs: https://simonwillison.net/2023/Aug/3/weird-world-of-llms/

Conclusion

Not covered

- Hardware
- Science
- Architectures

Further reading

- https://simonwillison.net/2023/Aug/3/weird-world-of-llms/
- https://github.com/Hannibal046/Awesome-LLM
- https://letsbuild.ai/