AI Models on Consumer Hardware

Rambling opinions of Richard Cottrill

Building Blocks (today)

GPU
>8GB VRAM,
Nvidia

CPU
Some more,
please

RAM
lots, >16GB

Storage
do you have a
lazy terabyte?

Linux
pref. Ubuntu

CUDA 11.7
drivers

Python
64bit 3.10

Mad skillz

- Compiling things
- Python (basic)
- Command-line
- Willingness to break things

Components in the ecosystem

1

Tensors

- Arrays of floating point numbers

2

Transformers

- *GPU* focussed processing

3

*BLAS libraries

4

CUDA / Nvidia

5

Python libraries (pytorch, transformers, ...)

Players in the ecosystem

- Open-source nerds
- langchain (middleware)
- Hugging face (think github but builds more stuff)
- Facebook
- Google
- Microsoft
- Nvidia
- Intel (barely)

Models: LLAMA and Stable Diffusion

- LLAMA is an LLM (Large Language Model), open-source from Facebook.
- llama.cpp is a project to run llama-derived models on commodity hardware (starting with Apple silicon).
- Runs on CPU and GPU (for Llama)
- Quantize (compress) model/weights to smaller values (8, 4, 1 bit)

- Stable Diffusion is a text-to-image model, open-source from LMU Munich
- Runs in 6-8GB VRAM

What's in an AI?

Base Model	Weights	*Memory*	*Prompts*
• *LLM: <300m – 170T parameters* • *SD: 890m parameters*	• *pre-tuning* • *fine tuning* • *LORA*	• *Vector index/database*	• *General techniques* • *model-specific*

What can I do with this AI stuff?

- Build an AI to "know" a domain
- Llama

  DIY ChatGPT https://gpt4all.io/index.html

  Impersonate a character Fine-tune LLaMA to speak like Homer Simpson - Replicate – Replicate
- BLOOMZ

  Q&A against a knowledge base *https://medium.com/@dvianna/fine-tuning-bloomz-for-legal-question-answering-f964b5c0657f*
- Stable Diffusion

  Stylized QR codes https://stable-diffusion-art.com/qr-code/

Baby steps

- https://gpt4all.io/index.html
- https://github.com/AUTOMATIC1111/stable-diffusion-webui
- https://github.com/oobabooga/text-generation-webui
- https://github.com/ggerganov/llama.cpp