Getting Started with Llama 3 on AMD Radeon and Instinct GPUs

Garrett Byrd (Fluid Numerics)
Dr. Joe Schoonover (Fluid Numerics)

AMD
together we advance_

What is an LLM?

What is an LLM?

An LLM is a **Large Language Model**, a natural language processing model that utilizes neural networks and machine learning (most notably, **transformers**) to execute language processing tasks. This is often in the form of some generative output, such as text, images, audio, and video.

An LLM is a machine learning model that probabilistically predicts a sequence of words based on training data.

It's a robot you can have a conversation with.

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

What is a token?

In the context of LLMs, a **token** is a set of characters.*

Conceptually, we can think of tokens as words, although in practice a model's set of tokens includes portions of words, punctuation, and tokenizer-specific symbols (e.g., the **end of sequence** token).

Examples:

- 'unun'
- 'heiten'
- 'andid'
- 'Execute'
- '_Message'
- '.getClass'
- '_lifetime'
- 'ĠWebsite'
- '.'
- 'film'
- 'ĠDeposit'
- '<|eot_id|>'

* In practice, depending on the modality of the model, a token could also be a portion of an image, audio, video, etc.

What is an embedding?

An **embedding** is the representation of a token as an array. The weights/parameters of a transformer model are derived from embeddings of its tokens. The matrix made up of all token embeddings is the **embedding matrix**.

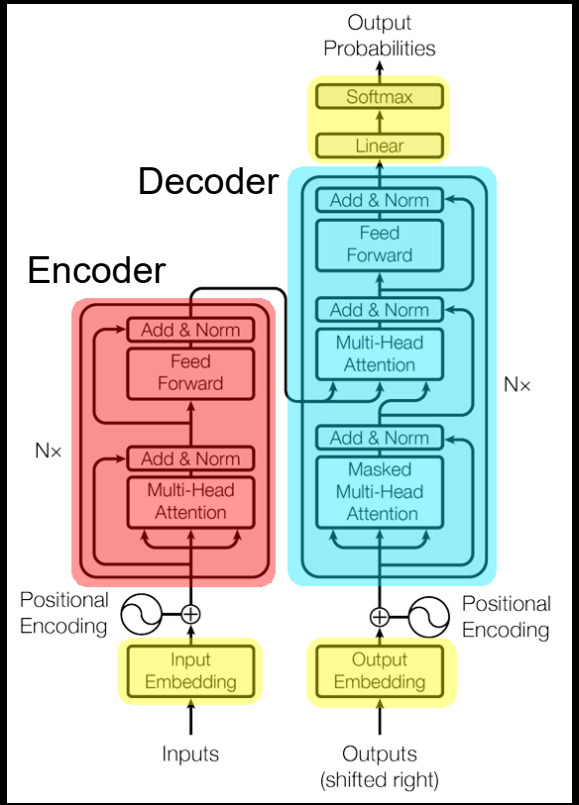
$$\text{“instinct”} \rightarrow \begin{bmatrix} 0.13 \\ -0.42 \\ \vdots \\ 13.37 \\ 8.34 \end{bmatrix}$$

What is a Transformer?

A **transformer** breaks down an input (usually text) into tokens, and then uses trained parameters to probabilistically generate an output. **Encoding** considers the position of each token in the input.

Example: The bassist for The Beatles is ???

Token	Probability
John	5%
Paul	85%
George	5%
Ringo	4%
Yoko	1%



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

What do I need? (hardware/OS)

A ROCm-compatible AMD GPU

AMD Instinct | AMD Radeon PRO | AMD Radeon

GPU	Architecture	LLVM target	Support
AMD Instinct MI300X	CDNA3	gfx942	✅ [1]
AMD Instinct MI300A	CDNA3	gfx942	✅ [2]
AMD Instinct MI250X	CDNA2	gfx90a	✅
AMD Instinct MI250	CDNA2	gfx90a	✅
AMD Instinct MI210	CDNA2	gfx90a	✅
AMD Instinct MI100	CDNA	gfx908	✅
AMD Instinct MI50	GCN5.1	gfx906	⚠️
AMD Instinct MI25	GCN5.0	gfx900	❌

✅: **Supported** - AMD enables these GPUs in our software distributions for the corresponding ROCm product.

⚠️: **Deprecated** - Support will be removed in a future release.

❌: **Unsupported** - This configuration is not enabled in our software distributions.

https://rocm.docs.amd.com/projects/install-on-linux/en/latest/reference/system-requirements.html (Last accessed June 3, 2024)

A ROCm-compatible Linux Distribution

Operating system	Kernel	Support
RHEL 9.3	5.14.0-362	✅
RHEL 9.2	5.14.0-362	✅
RHEL 8.9	4.18-513	✅
RHEL 8.8	4.18-513	✅
CentOS 7.9	3.10	✅
SLES 15 SP5	5.14.21-150500	✅
SLES 15 SP4	5.14.21-150500	✅
Ubuntu 22.04.4	5.15 [GA], 6.5 [HWE]	✅
Ubuntu 22.04.3	5.15 [GA], 6.2 [HWE]	✅
Ubuntu 20.04.6	5.15 [HWE]	✅
Ubuntu 20.04.5	5.15 [HWE]	✅

https://rocm.docs.amd.com/projects/install-on-linux/en/latest/reference/system-requirements.html (Last accessed June 3, 2024)

What do I need to install?

What do I need to install?

- ROCm (version 6.0+)
- conda (Anaconda3 or Miniconda3)
- PyTorch
- transformers (by Hugging Face)
- A few other things (packages like pip, jupyter, accelerate)

Installing ROCm

```
# install ROCm
# https://rocm.docs.amd.com/projects/install-on-linux/en/latest/tutorial/quick-start.html
sudo apt install "linux-headers-$(uname -r)" "linux-modules-extra-$(uname -r)"
sudo usermod -a -G render,video $LOGNAME # Adding current user to Video, Render groups. See prerequisites.
wget https://repo.radeon.com/amdgpu-install/6.1.1/ubuntu/jammy/amdgpu-install_6.1.60101-1_all.deb
sudo apt install ./amdgpu-install_6.1.60101-1_all.deb
sudo apt update
sudo apt install amdgpu-dkms
sudo apt install rocm
echo "Please reboot system for all settings to take effect."
```


https://rocm.docs.amd.com/projects/install-on-linux/en/latest/tutorial/quick-start.html (Last accessed June 3, 2024)

Installing conda

```
# install conda (miniconda)
# https://docs.anaconda.com/free/miniconda/
mkdir -p ~/miniconda3
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh -O ~/miniconda3/miniconda.sh
bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3
rm -rf ~/miniconda3/miniconda.sh
```


https://docs.anaconda.com/free/miniconda/#quick-command-line-install (Last accessed June 3, 2024)

Set up conda environment and git-lfs

```
# set up conda environment
conda create -n llama-3-env
conda activate llama-3-env
conda install pip
pip install jupyterlab

# install pytorch
# https://pytorch.org/get-started/locally/
pip install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/rocm6.0

# install transformers
pip install transformers accelerate

# install git-lfs
# https://git-lfs.com/
wget https://github.com/git-lfs/git-lfs/releases/download/v3.5.1/git-lfs-linux-amd64-v3.5.1.tar.gz
tar -xzf git-lfs-linux-amd64-v3.5.1.tar.gz
cd git-lfs-3.5.1/
sudo ./install.sh
git lfs install
```


https://git-lfs.com/ (Last accessed June 7, 2024)

Set up conda environment

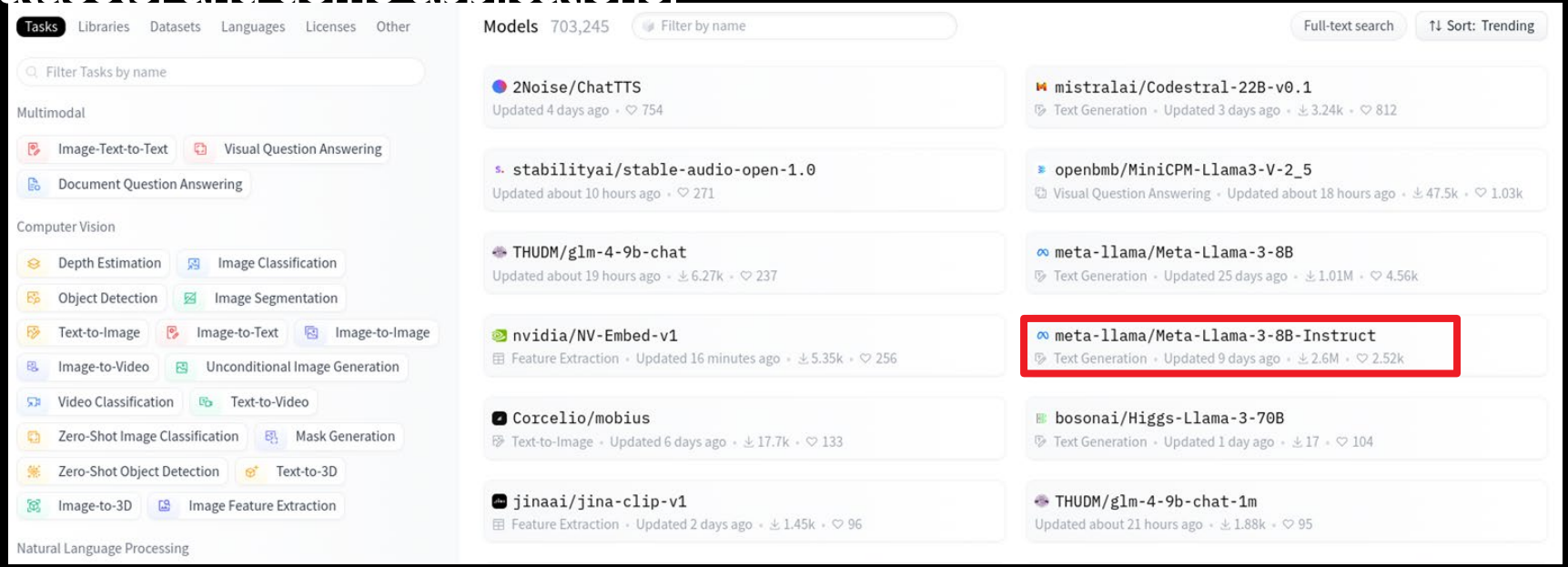
PyTorch Build	**Stable (2.3.0)**		Preview (Nightly)	
Your OS	**Linux**	Mac	Windows	
Package	Conda	**Pip**	LibTorch	Source
Language	**Python**		C++ / Java	
Compute Platform	CUDA 11.8	CUDA 12.1	~~CUDA 12.4~~	**ROCm 6.0**	CPU
Run this Command:	`pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/rocm6.0`			

https://pytorch.org/get-started/locally/ (Last accessed June 3, 2024)

Where do I get a model?
What model do I want?

Hugging Face Hub

The Hugging Face Hub is a platform that provides open source models, datasets, and demo applications.



https://huggingface.co/models (Last accessed June 6, 2024)

Hugging Face 🤗 Transformers Library

- Python API for PyTorch, TensorFlow, and JAX
- Provides REST endpoints
- Can be used for:
  - Natural Language Processing
  - Computer Vision
  - Audio
  - Multimodal Inputs

What model do I choose?

Llama 3

- Open source model developed by Meta Platforms, Inc.
- Pretrained with 15 trillion tokens
- 8 billion and 70 billion parameter versions
- Context length of 8K tokens
- High scores on various LLM benchmarks (e.g., MMLU)
- The Llama family has 5 million+ downloads on Hugging Face

https://llama.meta.com/llama3/ (last accessed June 6 2024); https://ai.meta.com/blog/meta-llama-3/ (last accessed June 6 2024)

Coding Example

Follow along:

github.com/FluidNumerics/amd-ml-examples

ROCm Blogs

- **Accelerating Large Language Models with Flash Attention on AMD GPUs**
  https://rocm.blogs.amd.com/artificial-intelligence/flash-attention/README.html
- **Unveiling performance insights with PyTorch Profiler on an AMD GPU**
  https://rocm.blogs.amd.com/artificial-intelligence/torch_profiler/README.html
- **PyTorch C++ Extension on AMD GPU**
  https://rocm.blogs.amd.com/artificial-intelligence/cpp-extn/readme.html
- **ResNet for image classification using AMD GPUs**
  https://rocm.blogs.amd.com/artificial-intelligence/resnet/README.html
- **Using LoRA for efficient fine-tuning: Fundamental principles**
  https://rocm.blogs.amd.com/artificial-intelligence/lora-fundamentals/README.html
- **Fine-tune Llama 2 with LoRA: Customizing a large language model for question-answering**
  https://rocm.blogs.amd.com/artificial-intelligence/llama2-lora/README.html

https://rocm.blogs.amd.com/ (last accessed June 3 2024)

Find Fluid Numerics Online

www.fluidnumerics.com


r/fluidnumerics


github.com/fluidnumerics


youtube.com/@FluidNumerics


linkedin.com/company/fluidnumerics

Q & A