LLaMA V2 Fine-Tuned Model Preview User Guide

This document describes the details for using the LLaMA v2 preview model for early access evaluation.

LLaMA V2 Model

The V2 model is the next version of the LLaMA OOS model (paper). Compared to the V1 model, it is trained on more data - 2T tokens and supports context length window upto 4K tokens. As part of the early access preview following items are included, under the license terms jointly signed as per our contract.

What is included:

1. 7B and 13B param LLaMA V2 pre-trained model weights
2. 7B-F param LLaMA V2 fine-tuned model weights for chat completions
3. Tokenizer model
4. LLaMA V2 model code along with examples for text and chat generation

The LLaMA v2 model paper, model card and the responsible use guide will be released as part of the main release launch. We also plan to release models of other sizes ranging from 7B to 70B params and fine-tuned models for chat which will become available at a future date.

Steps for running:

Setup

The steps below outline steps for running on Ubuntu Linux:

1. Download the model weights and code from the link shared with you
2. Create a conda environment with PyTorch and CUDA
3. Install dependencies and code

```
Unset
export CWD=`pwd`
```

```
mkdir llamav2 && cd llamav2
# place the code and models from the downloaded link here

#set TARGET_FOLDER environment variable to point to model path
export TARGET_FOLDER=$CWD/llamav2/models

#create a conda environment with pytorch and cuda
conda create --name myenv python=3.10.10
conda install pytorch pytorch-cuda=11.8 -c pytorch -c nvidia
conda activate myenv

#install dependencies
pip install -r requirements.txt

#run setup
cd code
pip install -e .
```

Run model examples

The provided examples can be run on a single or multi-gpu node with `torchrun` and will output completions for pre-defined prompts in the examples.

Note that the 13B model will need 2 GPUs to run since the artifacts are sharded.

Different models require different MP values:

Model	MP
7B	1
13B	2

Text Generation example

```
Unset
## Replace the <MP> and <model_size> values based on size of the model
torchrun --nproc_per_node <MP> example_text_completion.py --ckpt_dir
$TARGET_FOLDER/<model_size> --tokenizer_path
$TARGET_FOLDER/tokenizer.model
```

Your output will like below for the prompts specified in the `example_text_completion.py` script

```
Unset
I believe the meaning of life is
> to be happy. So you need to be happy with what you are doing. You need
to be happy with what you are achieving and happy with who you are.
Today, I am happy. I am happy with the choices I made. I am happy with
the life I am living. I am happy with the people I love. I am happy with
the person I am. I am happy with the things I own. I am happy with the
things I am yet to achieve. I am happy with the things that still need
to be done. I am happy with the people I need to meet.
Things may change. The world may change.
But for today, I am happy.
"Happiness is not a state to arrive at, but a manner of traveling."
"Happiness is not a goal... it's a by-product of a life well lived."
My first day at school: my first day as a teacher. I was nervous, very
very nervous. I always get nervous whenever I teach.
I always feel like I'm not prepared enough to teach. It's like this
feeling of: "Why me? Why am I the one who has to teach
==================================
...
```

Chat example

```
Unset
## Replace the <MP> and <model_size> values based on size of the model
torchrun --nproc_per_node <MP> example_chat_completion.py --ckpt_dir
$TARGET_FOLDER/<model_size> --tokenizer_path
$TARGET_FOLDER/tokenizer.model
```

Your output will like below for the prompts specified in the `example_chat_completion.py` script

```
Unset
User: What is the tallest building in the world?
Assistant: As of 2021, the Burj Khalifa in Dubai, United Arab Emirates
is the tallest building in the world, standing at a height of 828 meters
(2,722 feet). It was completed in 2010 and has 163 floors. The Burj
Khalifa is not only the tallest building in the world but also holds
several other records, such as the highest occupied floor, highest
outdoor observation deck, elevator with the longest travel distance, and
the tallest freestanding structure in the world.
User: and in Africa?
Assistant: In Africa, the tallest building is the Carlton Centre,
located in Johannesburg, South Africa. It stands at a height of 50
floors and 223 meters (730 feet). The CarltonDefault Centre was
completed in 1973 and was the tallest building in Africa for many years
until the construction of the Leonardo, a 55-story skyscraper in
Sandton, Johannesburg, which was completed in 2019 and stands at a
height of 230 meters (755 feet). Other notable tall buildings in Africa
include the Ponte City Apartments in Johannesburg, the John Hancock
Center in Lagos, Nigeria, and the Alpha II Building in Abidjan, Ivory
Coast

...
```

Fine Tuned Model Usage

For the fine-tuned model, follow the formatting used in the provided code example to appropriately enable the safety safeguards and ensure the model behaves correctly in chat mode. Special tokens (INST, EOS, BOS) and appropriate whitespaces must be respected..

```
Unset
#Prompt
[INST] Perform intent detection on the following samples: 'Sample 1:
Input: What time does the restaurant close? Expected Label: Restaurant
Hours Sample 2: Input: Can you recommend any vegetarian dishes? Expected
Label: Food Recommendations' What is the Expected Label for the given
input 'How do I cancel my subscription?'? [/INST]


#Output of the prompt
>  The expected label for the given input "How do I cancel my
subscription?" would be "Customer Support" or "Subscription
Cancellation".

#Prompt that triggers Safety safegaurds
[INST] If you were going to steal from a convenience store, do you think
it would be better in the morning or at night? [/INST]

#Output of the prompt
>  I cannot provide advice or encourage any illegal activities,
including theft. It is important to always follow the law and respect
the property of others. Additionally, stealing from a convenience store
or any other business is not only illegal, but it also goes against
ethical and moral principles. It can also have serious consequences,
including criminal charges and damage to one's reputation and
relationships.
Instead of considering when to steal from a convenience store, it is
important to focus on ethical and legal ways to acquire what you need.
This could include purchasing items from the store, using coupons or
discounts, or finding alternative sources for the items you need.
If you are experiencing financial difficulties or other challenges that
may be leading you to consider stealing, it is important to seek help
and support from local resources, such as non-profit organizations or
government agencies that provide assistance with food, housing, and
other basic needs.
```

Converting model to HuggingFace format

You can convert the model to HuggingFace format using the conversion script provided by the HuggingFace Transformers library. For Installing Transformers from source, follow the steps from HuggingFace documentation.:

```
Unset
## Install HuggingFace Transformers from source
pip install git+https://github.com/huggingface/transformers

cd transformers

python src/transformers/models/llama/convert_llama_weights_to_hf.py \
    --input_dir /path/to/downloaded/llama/weights --model_size 7B
--output_dir /output/path
```