Llama 2

We are unlocking the power of large language models. Our latest version of Llama is now accessible to individuals, creators, researchers and businesses of all sizes so that they can experiment, innovate and scale their ideas responsibly.

This release includes model weights and starting code for pretrained and fine-tuned Llama language models — ranging from 7B to 70B parameters.

This repository is intended as a minimal example to load Llama 2 models and run inference. For more detailed examples leveraging Hugging Face, see llama-recipes.

Updates post-launch

See UPDATES.md.

Download

⚠ **7/18: We're aware of people encountering a number of download issues today. Anyone still encountering issues should remove all local files, re-clone the repository, and request a new download link. It's critical to do all of these in case you have local corrupt files.**

In order to download the model weights and tokenizer, please visit the Meta AI website and accept our License.

Once your request is approved, you will receive a signed URL over email. Then run the download.sh script, passing the URL provided when prompted to start the download.

Pre-requisites: Make sure you have `wget` and `md5sum` installed. Then to run the script: `./download.sh`.

Keep in mind that the links expire after 24 hours and a certain amount of downloads. If you start seeing errors such as `403: Forbidden`, you can always re-request a link.

Access on Hugging Face

We are also providing downloads on Hugging Face. You must first request a download from the Meta AI website using the same email address as your Hugging Face account. After doing so, you can request access to any of the models on Hugging Face and within 1-2 days your account will be granted access to all versions.

Setup

In a conda env with PyTorch / CUDA available, clone the repo and run in the top-level directory:

```
pip install -e .
```

Inference

Different models require different model-parallel (MP) values:

Model	MP
7B	1
13B	2
70B	8

All models support sequence length up to 4096 tokens, but we pre-allocate the cache according to `max_seq_len` and `max_batch_size` values. So set those according to your hardware.

Pretrained Models

These models are not finetuned for chat or Q&A. They should be prompted so that the expected answer is the natural continuation of the prompt.

See `example_text_completion.py` for some examples. To illustrate, see the command below to run it with the llama-2-7b model (`nproc_per_node` needs to be set to the `MP` value):

```
torchrun --nproc_per_node 1 example_text_completion.py \
    --ckpt_dir llama-2-7b/ \
    --tokenizer_path tokenizer.model \
    --max_seq_len 128 --max_batch_size 4
```

Fine-tuned Chat Models

The fine-tuned models were trained for dialogue applications. To get the expected features and performance for them, a specific formatting defined in `chat_completion` needs to be followed, including the `INST` and `<<SYS>>` tags, `BOS` and `EOS` tokens, and the whitespaces and breaklines in between (we recommend calling `strip()` on inputs to avoid double-spaces).

You can also deploy additional classifiers for filtering out inputs and outputs that are deemed unsafe. See the llama-recipes repo for an example of how to add a safety checker to the inputs and outputs of your inference code.

Examples using llama-2-7b-chat:

```
torchrun --nproc_per_node 1 example_chat_completion.py \
    --ckpt_dir llama-2-7b-chat/ \
    --tokenizer_path tokenizer.model \
    --max_seq_len 512 --max_batch_size 6
```

Llama 2 is a new technology that carries potential risks with use. Testing conducted to date has not — and could not — cover all scenarios. In order to help developers address these risks, we have created the Responsible Use Guide. More details can be found in our research paper as well.

Issues

Please report any software “bug,” or other problems with the models through one of the following means:

- Reporting issues with the model: github.com/facebookresearch/llama
- Reporting risky content generated by the model: developers.facebook.com/llama_output_feedback
- Reporting bugs and security concerns: facebook.com/whitehat/info

Model Card

See MODEL_CARD.md.

License

Our model and weights are licensed for both researchers and commercial entities, upholding the principles of openness. Our mission is to empower individuals, and industry through this opportunity, while fostering an environment of discovery and ethical AI advancements.

See the LICENSE file, as well as our accompanying Acceptable Use Policy

References

1. Research Paper
2. Llama 2 technical overview
3. Open Innovation AI Research Community

Original LLaMA

The repo for the original llama release is in the `llama_v1` branch.