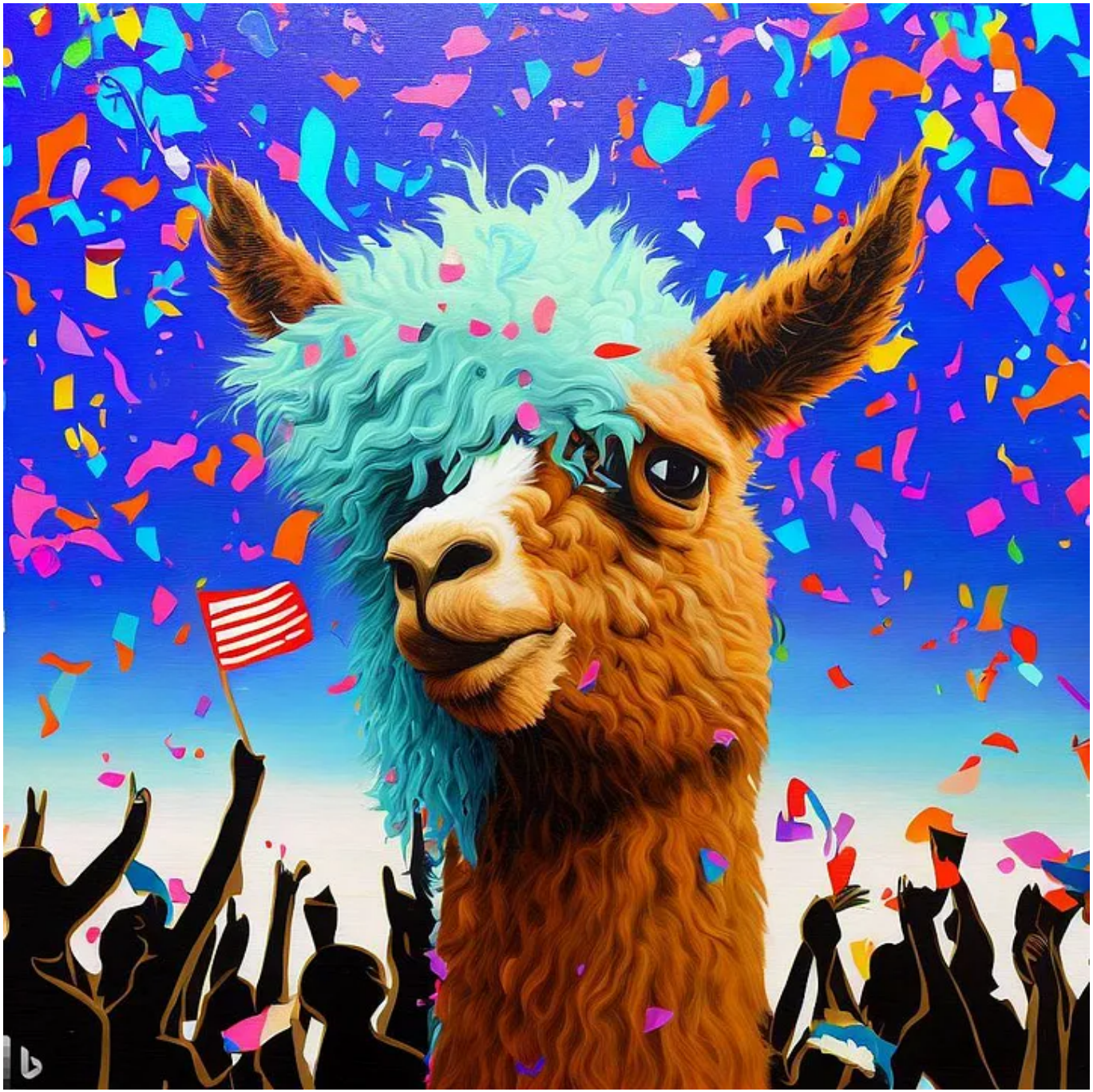
Democratizing Large Language Models (LLMs) Using Petals

FHIRFLY · Follow

3 min read · Aug 30, 2023



In recent years, large language models (LLMs) have transformed the way we interact with technology, enabling capabilities such as chatbots, language translation, and text generation. However, the sheer size and computational requirements of these models have posed challenges for wider access and adoption. Enter Petals, a groundbreaking initiative that aims to democratize LLMs by making them accessible and usable by a broader audience, including individual developers and researchers. In this article, we will explore how Petals empowers users to harness the power of LLMs like Llama 2 and BLOOM-176B for their tasks.

The Power of LLMs and the Access Challenge

LLMs, such as GPT-3.5 and BERT, have demonstrated remarkable capabilities in understanding and generating human-like text. These models are typically trained on massive amounts of data and comprise billions of parameters. However, utilizing these models effectively requires significant computational resources, which often limits access to large organizations with access to high-performance hardware.

Petals aims to bridge this gap by creating a collaborative network where individuals can contribute their GPU resources to host and run parts of these large models. By distributing the computational load across a network of

devices, Petals enables even those without access to cutting-edge hardware to leverage the power of LLMs.

Introducing Petals and Its Models

Petals provides a user-friendly interface for connecting to its distributed network of hosted model layers. This is made possible through the `AutoDistributedModelForCausalLM` class provided by the Petals library. By simply loading a small part of a model, users can join forces with others to collectively run inference or fine-tuning tasks.

```
from transformers import AutoTokenizer
from petals import AutoDistributedModelForCausalLM
```

```
model_name = "petals-team/StableBeluga2"

tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoDistributedModelForCausalLM.from_pretrained(model_name)

inputs = tokenizer("A cat sat", return_tensors="pt")["input_ids"]
outputs = model.generate(inputs, max_new_tokens=5)
print(tokenizer.decode(outputs[0]))  # A cat sat on a mat...
```

Making GPUs Accessible with Petals

Petals operates on a shared GPU model, where users contribute their GPU resources to the network. The process of hosting a part of a model on your GPU is straightforward and varies based on your operating system. Whether you're using Linux, Windows with Windows Subsystem for Linux (WSL), Docker, or even macOS with Apple M1/M2 GPU, Petals provides installation and setup instructions for each environment.

For instance, on Linux using Anaconda, you can host a part of the Stable Beluga 2 model with the following commands:

```
conda install pytorch pytorch-cuda=11.7 -c pytorch -c nvidia
pip install git+https://github.com/bigscience-workshop/petals
python -m petals.cli.run_server petals-team/StableBeluga2
```

Collaborative Inference and Beyond

Petals not only facilitates single-batch inference at impressive speeds but also allows for parallel inference, which significantly enhances the throughput. This collaborative approach ensures that even resource-intensive tasks can be completed efficiently. Moreover, Petals goes beyond traditional LLM APIs, allowing users to employ fine-tuning, sampling

methods, custom paths through the model, and even inspect hidden states, all while enjoying the comfort of an API with the flexibility of PyTorch.

Democratizing LLMs: A Community Effort

Petals operates as a community-driven system, relying on individuals sharing their GPU resources. By contributing your GPU power, you become part of a collaborative effort to democratize LLMs and make their capabilities accessible to a wider audience. As a token of gratitude, once you've loaded and hosted a certain number of blocks, your name or link can be displayed on the swarm monitor.

Conclusion

Petals revolutionizes the accessibility of large language models by creating a network where individuals can collectively host and run parts of these models. By democratizing access to powerful LLMs like Llama 2 and BLOOM-176B, Petals empowers developers, researchers, and enthusiasts to unlock the potential of natural language understanding and generation. This innovative approach brings us one step closer to a future where AI technologies are accessible to all, regardless of hardware limitations. So, why not join the Petals community and be a part of this exciting journey?

Disclaimer: The information provided in this article is based on the README.md from the project's GitHub repository as of the knowledge cutoff date in September 2021. For the most up-to-date information and instructions, please refer to the project's official documentation and repository.

Llm Llama 2 AI Healthcare Democracy

Written by FHIRFLY

Follow

74 Followers

SECURE. PRIVATE. AVAILABLE. CONFIDENTIAL. INTEGRAL. INTEROPERABLE. OUT OF THE DARKNESS COMES LIGHT.