petals-team / StableBeluga2

Downloads last month: 132,919

Model Description

Stable Beluga 2 is a Llama2 70B model finetuned on an Orca style Dataset. This repository contains the model from the stabilityai/StableBeluga2 repository with the following changes:

- **Storing weights in bfloat16 instead of float32.** This leads to 2x smaller files and a small quality loss, which is not significant compared to the loss caused by NF4 quantization used in Petals by default.
- **Storing weights in small shards.** Each transformer block is stored in its own shard (1.71 GB each). The input and output embeddings and adjacent layernorms are in a separate shard (1.05 GB) too. This way, Petals clients and servers don't have to download any excess data besides the layers they actually use.
- **Using Safetensors instead of Pickle.** This allows faster loading with smaller RAM requirements.

Model Details

Developed by: Stability AI
Model type: Stable Beluga 2 is an auto-regressive language model fine-tuned on Llama2 70B.
Language(s): English
Library: HuggingFace Transformers
License: Fine-tuned checkpoints (Stable Beluga 2) is licensed under the STABLE BELUGA NON-COMMERCIAL COMMUNITY LICENSE AGREEMENT
Contact: For questions and comments about the model, please email lm@stability.ai

Training Dataset

Stable Beluga 2 is trained on our internal Orca-style dataset. Training data is a synthetic dataset that was created to enhance the small model's reasoning abilities. The dataset comprises a diverse collection of tasks aimed at training AI models across various domains, focusing on cautious reasoning and alignment with ethical guidelines. It includes approximately 602,000 zero-shot queries grouped into 23 categories and 126 sub-categories, each sharing a common instruction format to promote consistency. The dataset also features 55,000 few-shot samples to encourage the model's ability to learn from context, around 160,000 math problems sourced from a variety of existing datasets, and 2,000 synthetically generated conversations between doctors and patients designed to test the model's specialized skills.

Intended Uses

The pretrained-only model can be used for prompting for evaluation of downstream tasks as well as text generation. In addition, the model can be fine-tuned on a downstream task. For all other checkpoints, please have a look at the model hub.

Responsible AI Considerations

Beluga is a new technology that carries risks with use:

- Testing conducted to date has been in English, and has not covered, nor could it cover all scenarios.
- Beluga's potential outputs cannot be predicted in advance (as with all LLMs), and the model may in some instances produce inaccurate, biased or other objectionable responses to user prompts.

Therefore, before deploying any applications of Beluga, developers should perform safety testing and tuning tailored to their specific applications of the model.