BMInf

Release v1

OpenBMB

Apr 06, 2022

GETTING STARTED

BMInf (Big Model Inference) is a low-resource inference package for large-scale pretrained language models (PLMs).

CHAPTER ONE

INTRODUCTION

BMInf (Big Model Inference) is a low-resource inference package for large-scale pretrained language models (PLMs). It has following features:

- **Hardware Friendly.** BMInf supports running models with more than 10 billion parameters on a single NVIDIA GTX 1060 GPU in its minimum requirements. Running with better GPUs leads to better performance. In cases where the GPU memory supports the large model inference (such as V100 or A100), BMInf still has a significant performance improvement over the existing PyTorch implementation.
- **Open.** The parameters of models are open. Users can access large models locally with their own machines without applying or accessing an online API.
- **Comprehensive Ability.** BMInf supports generative model CPM1 [*1*], general language model CPM2.1 [*2*], and dialogue model EVA [*3*]. The abilities of these models cover text completion, text generation, and dialogue generation.
- **Upgraded Model.** Based on CPM2 [*2*], the newly upgraded model CPM2.1 is currently supported. Based on continual learning, the text generation ability of CPM2.1 is greatly improved compared to CPM2.
- **Convenient Deployment.** Using BMInf, it will be fast and convenient to develop interesting downstream applications.

1.1 Supported Models

BMInf currently supports these models:

- **CPM2.1.** CPM2.1 is an upgraded version of CPM2 [*1*], which is a general Chinese pre-trained language model with 11 billion parameters. Based on CPM2, CPM2.1 introduces a generative pre-training task and was trained via the continual learning paradigm. In experiments, CPM2.1 has a better generation ability than CPM2.
- **CPM1.** CPM1 [*2*] is a generative Chinese pre-trained language model with 2.6 billion parameters. The architecture of CPM1 is similar to GPT [*4*] and it can be used in various NLP tasks such as conversation, essay generation, cloze test, and language understanding.
- **EVA.** EVA [*3*] is a Chinese pre-trained dialogue model with 2.8 billion parameters. EVA performs well on many dialogue tasks, especially in the multi-turn interaction of human-bot conversations.

Besides these models, we are now working on adding more PLMs especially large-scale PLMs. We welcome every contributor to add their models to this project by proposing an issue.

1.2 Performances

Here we report the speeds of CPM2 encoder and decoder we have tested on different platforms. You can also run `benchmark/cpm2/encoder.py` and `benchmark/cpm2/decoder.py` to test the speed on your machine!

1.3 Contributing

Here is the QRCode to our WeChat user community and we welcome others to contribute codes following our contributing guidelines.

1.4 License

The package is released under the Apache 2.0 License.

1.5 References

1. CPM-2: Large-scale Cost-efficient Pre-trained Language Models. Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, Yanzheng Cai, Guoyang Zeng, Zhixing Tan, Zhiyuan Liu, Minlie Huang, Wentao Han, Yang Liu, Xiaoyan Zhu, Maosong Sun.
2. CPM: A Large-scale Generative Chinese Pre-trained Language Model. Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, Maosong Sun.
3. EVA: An Open-Domain Chinese Dialogue System with Large-Scale Generative Pre-Training. Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, Jie Tang.

4. Language Models are Unsupervised Multitask Learners. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever.

CHAPTER

TWO

INSTALLATION

2.1 From pip (Recommended)

```
pip install bminf
```

2.2 From Source

```
git clone https://github.com/OpenBMB/BMInf.git
cd BMInf
python setup.py install
```

2.3 From Docker

```
docker run -it --gpus 1 -v $HOME/.cache/bigmodels:/root/.cache/bigmodels --rm openbmb/
↪bminf python3 examples/fill_blank.py
```

After installation, you can run an example in the `examples` folder to find if it is installed correctly.

```
python examples/fill_blank.py
```

2.4 Hardware Requirement

Here we list the minimum and recommended configurations for running BMInf.

GPUs with compute capability 6.1 or higher are supported by BMInf. Refer to the table to check whether your GPU is supported.

2.5 Software Requirement

BMInf requires CUDA version >= 10.1 and all the dependencies can be automaticlly installed by the installation process.

- **python** >= 3.6
- **requests**
- **tqdm**
- **jieba**
- **numpy**
- **cpm_kernels** >= 1.0.9

If you want to use the backpropagation function with PyTorch, make sure `torch` is installed on your device.

CHAPTER THREE

ABOUT US

BMInf is developed and maintained by OpenBMB (Open Lab for Big Model Base). OpenBMB is founded and supported by Beijing Academy of Artificial Intelligence (BAAI) and Tsinghua University.

The goal of OpenBMB is to build the model base and toolkits for large-scale pre-trained language models. We aim to accelerate the process of training, tuning, and inference for big models (with more than 10 billion parameters) and lower the barriers to use them. Based on this, we further aim to build the open-source community to promote the open-source ecosystems of pre-trained language models, build the AI infrastructure, and define the application paradigm in the intelligent era.

3.1 Our Team

3.1.1 Manager

3.1.2 Member

3.2 Contact Us

Join our WeChat community to contact us:

CHAPTER

FOUR

DEMO INTRODUCTION

Demo Code

BMInf-Demos includes application examples designed according to models in BMInf. These examples are:

- **Fill Blank.** It is a use case based on CPM2.1. It supports arbitrary input of a paragraph and can generate corresponding content in the blank according to the context.
- **Generate Story.** It is an example based on CPM1. You only need to write the beginning of a paragraph, and it can create a coherent essay for you.
- **Dialogue.** It is an example based on the EVA model. Here, you can talk freely with the machine.

4.1 Demonstration

- Fill Blank
- Generate Story
- Dialogue

4.2 Install

- Run the follow command after installing `nvidia-docker2`:

```
$ docker run -it --gpus 1 -v $HOME/.cache/bigmodels:/root/.cache/bigmodels -p 0.0.0.
↪0:8000:8000 --rm openbmb/bminf-demos
```

- Open your browser and visit http://localhost:8000/ to access the demo.

CHAPTER

FIVE

MODELS

5.1 CPM-1

5.2 CPM-2

5.3 EVA