Author: @vasudevgupta7

Intro

In this project, we fine-tuned **BigBird** on **natural-questions** dataset for **question-answering** task on long documents. **BigBird**, is a **sparse-attention based transformer** which extends Transformer based models, such as BERT to much **longer sequences**.

Read more about BigBird at https://huggingface.co/blog/big-bird

Fine-tuning

Setup

You need to install jax yourself by following the official docs (refer this). Other requirements for this project can be installed by running following command:

```
pip3 install -qr requirements.txt
```

Download & prepare dataset

The Natural Questions corpus contains questions from real users, and it requires QA systems to read and comprehend an entire Wikipedia article that may or may not contain the answer to the question. This corpus takes ~100 GB on disk. We have used HuggingFace datasets to download & process the dataset.

```
# just run following CMD
python3 prepare_natural_questions.py

# this will download the whole dataset from HuggingFace Hub & will make it ready for training
# this script takes ~3 hours to process the dataset
```

Launch Training

We have trained on Cloud's TPU v3-8. Each epoch took around 4.5 hours and the model got converged in just 2 epochs. You can see complete training args in this script.

```
# just run following CMD
python3 train.py

# In case, you want to try hparams tuning, you can run wandb sweep
wandb sweep --project=bigbird sweep_flax.yaml
wandb agent <agent-id-obtained-by-above-CMD>
```

Evaluation

Our evaluation script is different from the original script and we are evaluating sequences with length up to 4096 for simplicity. We managed to get the **EM score of ~55.2** using our evaluation script.

```
# download validation-dataset first
mkdir natural-questions-validation
wget https://huggingface.co/datasets/vasudevgupta/natural-questions-
validation/resolve/main/natural_questions-validation.arrow -P natural-questions-
validation
wget https://huggingface.co/datasets/vasudevgupta/natural-questions-
validation/resolve/main/dataset_info.json -P natural-questions-validation
wget https://huggingface.co/datasets/vasudevgupta/natural-questions-
validation/resolve/main/state.json -P natural-questions-validation

# simply run following command
python3 evaluate.py
```

You can find our checkpoint on HuggingFace Hub (see this). In case you are interested in PyTorch BigBird fine-tuning, you can refer to this repositary.