Question Answering examples

Based on the script `run_qa.py`.

Note: This script only works with models that have a fast tokenizer (backed by the 🤗 Tokenizers library) as it uses special features of those tokenizers. You can check if your favorite model has a fast tokenizer in this table, if it doesn't you can still use the old version of the script.

The following example fine-tunes BERT on SQuAD:

```
python run_qa.py \
  --model_name_or_path bert-base-uncased \
  --dataset_name squad \
  --do_train   \
  --do_eval   \
  --max_seq_length 384 \
  --doc_stride 128 \
  --learning_rate 3e-5 \
  --num_train_epochs 2 \
  --per_device_train_batch_size 12 \
  --output_dir ./bert-qa-squad \
  --eval_steps 1000 \
  --push_to_hub
```

Using the command above, the script will train for 2 epochs and run eval after each epoch. Metrics and hyperparameters are stored in Tensorflow event files in `--output_dir`. You can see the results by running `tensorboard` in that directory:

```
$ tensorboard --logdir .
```

or directly on the hub under *Training metrics*.

Training with the previously defined hyper-parameters yields the following results:

```
f1 = 88.62
exact_match = 81.34
```

sample Metrics - tfhub.dev

Here is an example training on 4 TITAN RTX GPUs and Bert Whole Word Masking uncased model to reach a F1 > 93 on SQuAD1.1:

```
export CUDA_VISIBLE_DEVICES=0,1,2,3
python run_qa.py   \
--model_name_or_path bert-large-uncased-whole-word-masking   \
--dataset_name squad   \
--do_train   \
--do_eval   \
--per_device_train_batch_size 6   \
--learning_rate 3e-5   \
```

```
--num_train_epochs 2   \
--max_seq_length 384   \
--doc_stride 128   \
--output_dir ./wwm_uncased_finetuned_squad/ \
--eval_steps 1000 \
--push_to_hub
```

Training with the previously defined hyper-parameters yields the following results:

```
f1 = 93.31
exact_match = 87.04
```

Usage notes

Note that when contexts are long they may be split into multiple training cases, not all of which may contain the answer span.

As-is, the example script will train on SQuAD or any other question-answering dataset formatted the same way, and can handle user inputs as well.

Memory usage and data loading

One thing to note is that all data is loaded into memory in this script. Most question answering datasets are small enough that this is not an issue, but if you have a very large dataset you will need to modify the script to handle data streaming.