Language modelling examples

This folder contains some scripts showing examples of *language model pre-training* with the 🤗 Transformers library. For straightforward use-cases you may be able to use these scripts without modification, although we have also included comments in the code to indicate areas that you may need to adapt to your own projects. The two scripts have almost identical arguments, but they differ in the type of LM they train - a causal language model (like GPT) or a masked language model (like BERT). Masked language models generally train more quickly and perform better when fine-tuned on new tasks with a task-specific output head, like text classification. However, their ability to generate text is weaker than causal language models.

Pre-training versus fine-tuning

These scripts can be used to both *pre-train* a language model completely from scratch, as well as to *fine-tune* a language model on text from your domain of interest. To start with an existing pre-trained language model you can use the `--model_name_or_path` argument, or to train from scratch you can use the `--model_type` argument to indicate the class of model architecture to initialize.

Multi-GPU and TPU usage

By default, these scripts use a `MirroredStrategy` and will use multiple GPUs effectively if they are available. TPUs can also be used by passing the name of the TPU resource with the `--tpu` argument.

run_mlm.py

This script trains a masked language model.

Example command

```
python run_mlm.py \
--model_name_or_path distilbert-base-cased \
--output_dir output \
--dataset_name wikitext \
--dataset_config_name wikitext-103-raw-v1
```

When using a custom dataset, the validation file can be separately passed as an input argument. Otherwise some split (customizable) of training data is used as validation.

```
python run_mlm.py \
--model_name_or_path distilbert-base-cased \
--output_dir output \
--train_file train_file_path
```

run_clm.py

This script trains a causal language model.

Example command

```
python run_clm.py \
--model_name_or_path distilgpt2 \
```

```
--output_dir output \
--dataset_name wikitext \
--dataset_config_name wikitext-103-raw-v1
```

When using a custom dataset, the validation file can be separately passed as an input argument. Otherwise some split (customizable) of training data is used as validation.

```
python run_clm.py \
--model_name_or_path distilgpt2 \
--output_dir output \
--train_file train_file_path
```