# Model parallel language model training example

The following example showcases how to train/fine-tune GPTNeo model with model parallelism using the JAX/Flax backend and the `pjit` transformation.

> Note: The example is experimental and might have bugs. Also currently it only supports single V3-8.

The `partition.py` file defines the `PyTree` of `ParitionSpec` for the GPTNeo model which describes how the model will be sharded. The actual sharding is auto-matically handled by `pjit`. The weights are sharded accross all local devices. To adapt the script for other models, we need to also change the `ParitionSpec` accordingly.

TODO: Add more explantion.

Before training, let's prepare our model first. To be able to shard the model, the sharded dimention needs to be a multiple of devices it'll be sharded on. But GPTNeo's vocab size is 50257, so we need to resize the embeddings accordingly.

```
from transformers import FlaxGPTNeoForCausalLM, GPTNeoConfig
model = FlaxGPTNeoForCausalLM.from_pretrained("EleutherAI/gpt-neo-1.3B")

emb = jnp.zeros((50264, model.config.hidden_size))
# update the first 50257 weights using pre-trained weights
emb = emb.at[:50257, :].set(model.params["transformer"]["wte"]["embedding"])
params = model.params
params["transformer"]["wte"]["embedding"] = emb

# initialize a random model with the right vocab_size
config = GPTNeoConfig.from_pretrained("EleutherAI/gpt-neo-1.3B", vocab_size=50264)
model = FlaxGPTNeoForCausalLM(config)

# assign the pre-trained weights and save the model.
model.params = params
model.save_pretrained("gpt-neo-1.3B")
```

### Train Model

```
python run_clm_mp.py \
    --model_name_or_path gpt-neo-1.3B  \
    --tokenizer_name gpt2 \
    --dataset_name wikitext --dataset_config_name wikitext-2-raw-v1 \
    --do_train  --do_eval \
    --block_size 1024 \
    --num_train_epochs 5 \
    --learning_rate 4e-6 \
    --per_device_train_batch_size 3 --per_device_eval_batch_size 3 \
    --overwrite_output_dir --output_dir ~/tmp/flax-clm \
```

```
--cache_dir ~/datasets_cache/wikitext --dtype bfloat16 \
--logging_steps 96 --eval_steps 96
```