MobileBERT (MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices)

MobileBERT is a thin version of BERT_LARGE, while equipped with bottleneck structures and a carefully designed balance between self-attentions and feed-forward networks.

To train MobileBERT, we first train a specially designed teacher model, an inverted-bottleneck incorporated BERT_LARGE model. Then, we conduct knowledge transfer from this teacher to MobileBERT. Empirical studies show that MobileBERT is 4.3x smaller and 5.5x faster than BERT_BASE while achieving competitive results on well-known benchmarks. This repository contains TensorFlow 2.x implementation for MobileBERT.

Network Implementations

Following MobileBERT TF1 implementation, we re-implemented MobileBERT encoder and layers using `tf.keras` APIs in NLP modeling library:

- mobile_bert_encoder.py contains `MobileBERTEncoder` implementation.
- mobile_bert_layers.py contains `MobileBertEmbedding`, `MobileBertTransformer` and `MobileBertMaskedLM` implementation.

Pre-trained Models

We converted the originial TF 1.x pretrained English MobileBERT checkpoint to TF 2.x checkpoint, which is compatible with the above implementations. In addition, we also provide new multiple-lingual MobileBERT checkpoint trained using multi-lingual Wiki data. Furthermore, we export the checkpoints to TF-HUB SavedModel. Please find the details in the following table:

Model	Configuration	Number of Parameters	Training Data	Checkpoint & Vocabulary	TF-Hub SavedModel	Metrics
MobileBERT uncased English	uncased_L-24_H-128_B-512_A-4_F-4_OPT	15.3 Million	Wiki + Books	Download	TF-Hub	Squad v1.1 F1 90.0, GLUE 77.7

Model	Configuration	Number of Parameters	Training Data	Checkpoint & Vocabulary	TF-Hub SavedModel	Metrics
MobileBERT-cased Multilingual	multi_cased_L-24_H-128_B-512_A-4_F-4_OPT	36 Million	Wiki	Download	TF-Hub	XNLI (zero-short):64.7

Restoring from Checkpoints

To load the pre-trained MobileBERT checkpoint in your code, please follow the example below:

```python
import tensorflow as tf
from official.nlp.projects.mobilebert import model_utils

bert_config_file = ...
model_checkpoint_path = ...

bert_config = model_utils.BertConfig.from_json_file(bert_config_file)

# `pretrainer` is an instance of `nlp.modeling.models.BertPretrainerV2`.
pretrainer = model_utils.create_mobilebert_pretrainer(bert_config)
checkpoint = tf.train.Checkpoint(**pretrainer.checkpoint_items)
checkpoint.restore(model_checkpoint_path).assert_existing_objects_matched()

# `mobilebert_encoder` is an instance of
# `nlp.modeling.networks.MobileBERTEncoder`.
mobilebert_encoder = pretrainer.encoder_network
```

Use TF-Hub models

For the usage of MobileBert TF-Hub model, please see the TF-Hub site (English model or Multilingual model).