DeeBERT: Early Exiting for *BERT

This is the code base for the paper DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference, modified from its original code base.

The original code base also has information for downloading sample models that we have trained in advance.

Usage

There are three scripts in the folder which can be run directly.

In each script, there are several things to modify before running:

- `PATH_TO_DATA`: path to the GLUE dataset.
- `--output_dir`: path for saving fine-tuned models. Default: `./saved_models`.
- `--plot_data_dir`: path for saving evaluation results. Default: `./results`. Results are printed to stdout and also saved to `npy` files in this directory to facilitate plotting figures and further analyses.
- `MODEL_TYPE`: bert or roberta
- `MODEL_SIZE`: base or large
- `DATASET`: SST-2, MRPC, RTE, QNLI, QQP, or MNLI

train_deebert.sh This is for fine-tuning DeeBERT models.

eval_deebert.sh This is for evaluating each exit layer for fine-tuned DeeBERT models.

entropy_eval.sh This is for evaluating fine-tuned DeeBERT models, given a number of different early exit entropy thresholds.

Citation

Please cite our paper if you find the resource useful:

```
@inproceedings{xin-etal-2020-deebert,
    title = "{D}ee{BERT}: Dynamic Early Exiting for Accelerating {BERT} Inference",
    author = "Xin, Ji  and
      Tang, Raphael  and
      Lee, Jaejun  and
      Yu, Yaoliang  and
      Lin, Jimmy",
    booktitle = "Proceedings of the 58th Annual Meeting of the Association for Computational
    month = jul,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://www.aclweb.org/anthology/2020.acl-main.204",
    pages = "2246--2251",
}
```