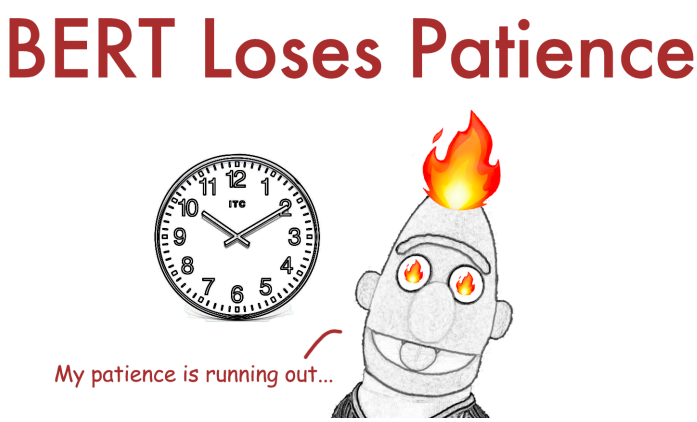
Patience-based Early Exit

Patience-based Early Exit (PABEE) is a plug-and-play inference method for pretrained language models. We have already implemented it on BERT and ALBERT. Basically, you can make your LM faster and more robust with PABEE. It can even improve the performance of ALBERT on GLUE. The only sacrifice is that the batch size can only be 1. Learn more in the paper ["BERT Loses Patience: Fast and Robust Inference with Early Exit"](#) and the official GitHub repo.



Training

You can fine-tune a pretrained language model (you can choose from BERT and ALBERT) and train the internal classifiers by:

```
export GLUE_DIR=/path/to/glue_data
export TASK_NAME=MRPC

python ./run_glue_with_pabee.py \
  --model_type albert \
  --model_name_or_path bert-base-uncased/albert-base-v2 \
  --task_name $TASK_NAME \
  --do_train \
  --do_eval \
  --do_lower_case \
  --data_dir "$GLUE_DIR/$TASK_NAME" \
  --max_seq_length 128 \
  --per_gpu_train_batch_size 32 \
  --per_gpu_eval_batch_size 32 \
  --learning_rate 2e-5 \
  --save_steps 50 \
  --logging_steps 50 \
  --num_train_epochs 5 \
  --output_dir /path/to/save/ \
  --evaluate_during_training
```

Inference

You can inference with different patience settings by:

```
export GLUE_DIR=/path/to/glue_data
export TASK_NAME=MRPC

python ./run_glue_with_pabee.py \
  --model_type albert \
  --model_name_or_path /path/to/save/ \
  --task_name $TASK_NAME \
  --do_eval \
  --do_lower_case \
  --data_dir "$GLUE_DIR/$TASK_NAME" \
  --max_seq_length 128 \
  --per_gpu_eval_batch_size 1 \
  --learning_rate 2e-5 \
  --logging_steps 50 \
  --num_train_epochs 15 \
  --output_dir /path/to/save/ \
  --eval_all_checkpoints \
  --patience 3,4,5,6,7,8
```

where `patience` can be a list of patience settings, separated by a comma. It will help determine which patience works best.

When evaluating on a regression task (STS-B), you may add `--regression_threshold 0.1` to define the regression threshold.

Results

On the GLUE dev set:

Model	#Param	Speed	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B
ALBERT-base	12M		58.9	84.6	89.5	91.7	89.6	78.6	92.8	89.5
+PABEE	12M	1.57x	61.2	85.1	90.0	91.8	89.6	80.1	93.0	90.1

Model	#Param	Speed-up	MNLI	SST-2	STS-B
BERT-base	108M		84.5	92.1	88.9
+PABEE	108M	1.62x	83.6	92.0	88.7
ALBERT-large	18M		86.4	94.9	90.4
+PABEE	18M	2.42x	86.8	95.2	90.6

Citation

If you find this resource useful, please consider citing the following paper:

```
@misc{zhou2020bert,
    title={BERT Loses Patience: Fast and Robust Inference with Early Exit},
    author={Wangchunshu Zhou and Canwen Xu and Tao Ge and Julian McAuley and Ke Xu
and Furu Wei},
    year={2020},
    eprint={2006.04152},
    archivePrefix={arXiv},
    primaryClass={cs.CL}
}
```