Intel Student Position – technical interview – Omer Nagar

MRPC - Bert model: number of parameters.

Bert – embeddings layer	23.8M
Bert – encoder	85M
Bert – pooler	6K
MCTS classifier	0.01K

Improve technique:

- Blocks pruning:
  In the paper we choose which weights to prune by their magnitude. To apply this method on blocks one can simply replace the block with:
  $$\alpha_{block} = ||block||_p$$
  I think that $||\cdot||_2$ is the right choice. But it is worth checking the affect of other norm types as well.

Implementation plan:

Data & Pre-Trained Model

1. Download GLUE dataset
2. Download Bert-uncased dataset (or smaller one for debugging)
3. Check that basic fine-tuning is running smoothly

Pruning Class

The pruning class will be hooked as forward-pre-hook.

- sparsity rate update
  Implement function that will update the sparsity rate according to the paper.
- Mask update
  Implement function that will create masks for each of the pruned layers according to the sparsity rate
- Apply mask
  Implement forward pre-hook function
- Check that basic pruning is working properly

Others

- Create sparsity graph
- Create results csv
- Save models as onnx/pt files
- Create jupyter notebooks for training and evaluation
- Upload to Git

Test Google-Colab

- Upload to colab and run experiment with bert-full-network