Result Summary of finetuning the ALBERT

**The structure of ALBERT for GLUE task is like following:**

E

ALBERT Transformer

L

L

ALBERT (Pretrained)

Output Layer

E: Embedding Layer

L: Linear Layer

The ALBERT Transformer contains one attention block, three linear layers, an activation layer, and a norm layer, which is much simpler than BERT. In the model, they only implement one ALBERT Transformer. When running, the ALBERT Transformer will be used recurrently, like RNN.

**I test the GLUE task in serval different settings:**

1. Finetuning all parameters,

2. Freezing the whole ALBERT and only finetuning the output layers,

3. Only freezing the embedding layer and finetuning other parts,

4. Freezing embedding layer and transformer and finetuning two linear layers.

Except the layers to finetune, all other experiment settings are same for these four setting.

**The results show that:**

1. Comparing to finetuning all parameters, only finetuning the output layer will cause over 30% performance drop on SST-2, STS-B, MNLI (I will give more detail about these tasks later). Only in WNLI the performance drop is less than 5%. In other tasks, the performance drop is about 20-15%. Except for WNLI, the overall performance is too bad. Note that the ALBERT model in the repo has some problem that for CoLA task, it always output 0% accuracy. This issue has been reported serval times but not fixed yet. Thus, all results include CoLA task.

2. The performance of freezing the embedding layer and finetuning other parts is on par with (or even exceed) the performance of finetuning all parameters in all tasks.

3. Freezing embedding layer and transformer, and finetuning two linear layers make and about average 10% performance drop on all tasks. It is much better than only finetuning the output layer, but still unacceptable. However, in WNLI, the performance is even worse than only finetuning the output layer.

**Parts of GELU task description:**

SST-2: It is a binary single-sentence classification task consisting of sentences extracted from movie reviews with human annotations of their sentiment, which includes fine-grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences.

STS-B: It is a collection of sentence pairs drawn from news headlines and other sources. They were annotated with a score from 1 to 5 denoting how similar the two sentences are in terms of semantic meaning. The benchmark comprises 8628 sentence pairs.

MNLI: It is a large-scale, crowdsourced entailment classification task. Given a pair of sentences, the goal is to predict whether the second sentence is an entailment, contradiction, or neutral with respect to the first one. It contains 433k sentence pairs.

In the above three tasks, finetuning output layer has the biggest performance drop. As we can see, the performance drop does not relate to the dataset size.

WNLI: It is a small natural language inference dataset, which collected 150 Winograd schemas. A Winograd schema is a pair of sentences that differ in only one or two words and that contain an ambiguity that is resolved in opposite ways in the two sentences and requires the use of world knowledge and reasoning for its resolution.

In the above task, finetuning output layer has the smallest performance drop.