ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

1 Contribution

The paper has proposed

- two parameter reduction techniques for BERT, and
- a novel inter-sentence coherence loss

2 The Parameter Reduction Techniques

Factorized embedding parameterization

In BERT, the size of the word embedding matrix is $V \times H$ where V is the vocabulary size and H is the hidden dimension. Since V can be in the order of tens or hundreds of thousands, the embedding matrix can easily introduce millions/billions of parameters.

ALBERT has proposed to factorize the embedding matrix M into multiplication of two low rank matrice i.e. $M = AB$ where $A \in R^{V \times E}$ and $A \in R^{E \times H}$ and $E \ll H$. This factorization reduces the number of parameters due to word embedding from $V \times H$ to $V \times E + E \times H$.

Cross Layer Parameter Sharing

ALBERT has proposed to share parameters across all layers. They have also tested several combinations of parameter sharing - sharing only the parameters of feed-forward layer, sharing only the parameters of attention layer, sharing the parameters of both feed forward layer and attention layer.

3 The Inter-sentence Coherence Loss

BERT uses next sentence prediction loss to capture inter-sentence coherence. The authors have claimed that next sentence prediction task is not effective to capture inter-sentence coherence as it falls back to the easier topic prediction task.

The paper has proposed a novel inter-sentence coherence loss called sentence-order prediction loss. In sentence-order prediction task, two sentences of the input segments are swapped with probability 0.5. The task is to predict whether the sentences of the input segment have been swapped.

4 ALBERT Model Configurations

The paper has proposed four ALBERT model configurations. They are shown Figure 1. Due to the design choices discussed above, ALBERT models have

Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

Table 1: The configurations of the main BERT and ALBERT models analyzed in this paper.

Figure 1: ALBERT model configurations

much smaller parameter size compared to corresponding BERT models. ALBERT-large has about 18x fewer parameters compared to BERT-large, 18M versus 334M. An ALBERT-xlarge configuration with H = 2048 has only 60M parameters and an ALBERT-xxlarge configuration with $H = 4096$ has 233M parameters, i.e., around 70% of BERT-large's parameters.

5 Experimental Results

The comparison of ALBERT with BERT in various down-stream tasks is shown in Figure 2.

The effect of vocabulary embedding size and cross-layer parameter sharing strategies on the performance of ALBERT-base are shown in Figure 3 and 4 respectively.

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3	4.7x
	large	334M	92.2/85.5	85.0/82.2	86.6	93.0	73.9	85.2	1.0
ALBERT	base	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1	5.6x
	large	18M	90.6/83.9	82.3/79.4	83.5	91.7	68.5	82.4	1.7x
	xlarge	60M	92.5/86.1	86.1/83.1	86.4	92.4	74.8	85.5	0.6x
	xxlarge	235M	**94.1/88.3**	**88.1/85.1**	**88.0**	**95.2**	**82.3**	**88.7**	0.3x

Table 2: Dev set results for models pretrained over BOOKCORPUS and Wikipedia for 125k steps. Here and everywhere else, the Avg column is computed by averaging the scores of the downstream tasks to its left (the two numbers of F1 and EM for each SQuAD are first averaged).

Figure 2: Performance of ALBERT on various down-stream tasks.

Model	E	Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
ALBERT base not-shared	64	87M	89.9/82.9	80.1/77.8	82.9	91.5	66.7	81.3
	128	89M	89.9/82.8	80.3/77.3	83.7	91.5	67.9	81.7
	256	93M	90.2/83.2	80.3/77.4	84.1	91.9	67.3	81.8
	768	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3
ALBERT base all-shared	64	10M	88.7/81.4	77.5/74.8	80.8	89.4	63.5	79.0
	128	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1
	256	16M	88.8/81.5	79.1/76.3	81.5	90.3	63.4	79.6
	768	31M	88.6/81.5	79.2/76.6	82.0	90.6	63.3	79.8

Table 3: The effect of vocabulary embedding size on the performance of ALBERT-base.

Figure 3: The effect of vocabulary embedding size on the performance of ALBERT-base

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
ALBERT base E=768	all-shared	31M	88.6/81.5	79.2/76.6	82.0	90.6	63.3	79.8
	shared-attention	83M	89.9/82.7	80.0/77.2	84.0	91.4	67.7	81.6
	shared-FFN	57M	89.2/82.1	78.2/75.4	81.5	90.8	62.6	79.5
	not-shared	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3
ALBERT base E=128	all-shared	12M	89.3/82.3	80.0/77.1	82.0	90.3	64.0	80.1
	shared-attention	64M	89.9/82.8	80.7/77.9	83.4	91.9	67.6	81.7
	shared-FFN	38M	88.9/81.6	78.6/75.6	82.3	91.7	64.4	80.2
	not-shared	89M	89.9/82.8	80.3/77.3	83.2	91.5	67.9	81.6

Table 4: The effect of cross-layer parameter-sharing strategies, ALBERT-base configuration.

Figure 4: The effect of cross-layer parameter sharing strategies on the performance of ALBERT-base