Multi-head self-attention analysis: using BERT and ALBERT

Insop Song (first name at)

Motivation

- Large pre-trained models have had great success in NLP
- BERT (Bidirectional Encoder Representations from Transformers)[1] & ALBERT (A Lite BERT)[2] are two of the pre-trained models
- BERT and ALBERT are based on Transformer
- **MHA (multi-head self-attentions)** is the key component of Transformer model
- Is it not clear *what multi-headed attentions do in the model?*
- Our goal: *understand MHA in BERT & ALBERT*

Multi-head self-attentions

- Self-attention output is is weighted sum of value
- MHA(Multi-head attention) is concatenation of individual head
- MHA effectively ensembles attention heads

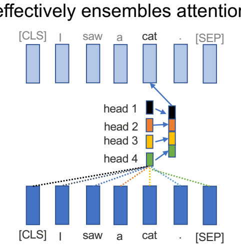

Fig0a. Multi-head attention

Transformer

- A transformer is an architecture that processes sequence of tokens without recurrent unit

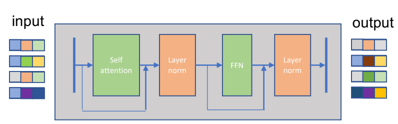

Fig0b. Transformer block

BERT and ALBERT

- BERT [1]: Consist of stack of transformer block
- Unsupervised training using MLM (masked LM)
- Can be used for various downstream task with fine tuning
- ALBERT [2]: A lite version of BERT, it reduced parameters by factorized embedded parameters and cross-layer parameter sharing

	layers	heads	Param.	Hidden	embedding	Param sharing
BERT base	12	12	108M	768	768	False
ALBERT base	12	12	12M	768	128	True

Table 1. The configuration of BERT and ALBERT models

Multi-head attention

- Use the attention weight, a_{ij} for analysis
- Shows few behaviors: attends broadly, attends to next or previous token. Or attends SEP or period

$$\alpha_{ij} = \frac{exp(q_i^T k_j)}{\sum_{l=1}^{n} exp(q_i^T k_l)}$$

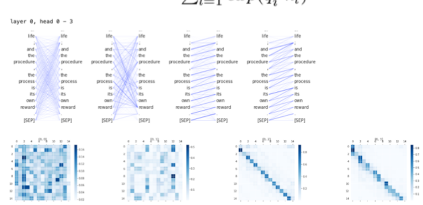

Fig1. BERT attention map, Layer 0 head 0-4

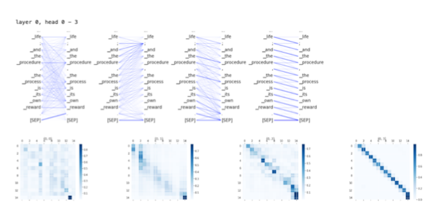

Fig2. ALBERT attention map, Layer 0 head 0-4

Confidence map

- To find out which head contributes more
- Average of max attention weight
- ALBERT shows smooth confidence, *likely due to* shared parameters

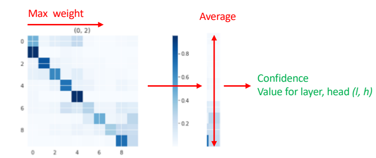

Fig6. Confidence map from multi-head attention

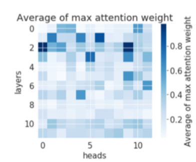

Fig3a. BERT confidence w/o SEP

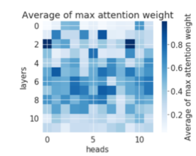

Fig3b. BERT confidence w/ SEP

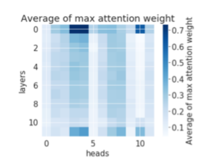

Fig4a. ALBERT confidence w/o SEP

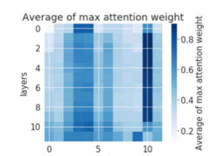

Fig4b. ALBERT confidence w/ SEP

Ablating one head

- Run GLUE MNLI test with single head pruned
- ALBERT shows head with importance collocated

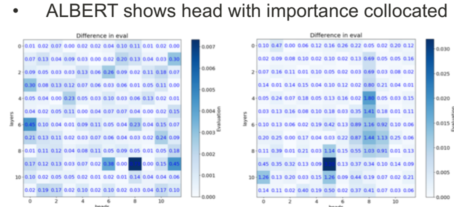

Fig5. BERT (left) ALBERT (right) evaluation score changes when one head is pruned

Less number of head can be *better* ?

- Prune heads based on ablation results, i.e. prune heads that caused eval score changed
- BERT: performance improves with 20 heads pruned
- ALBERT: more sensitive to head pruned, also *likely due to* shared parameters

# of pruned heads	Base -line	3	10	20	30
BERT base for MNLI	84.2	84.37	84.33	84.37	83.9
ALBERT base, MNLI	84.7	84.8	83.8	82.7	79.9

Conclusion

- First (*as far as we know*) comparison analysis for BERT & ALBERT using attention weight, confidence map, ablation experiment
- Successfully show the different multi head attention behaviors between BERT & ALBERT
- Provide confidence map for attention analysis
- Based on ablation result, showed head pruning behaviors in BERT & ALBERT

Future work

- Apply the analysis to other type of tests: SQuAD and NMT
- Apply systematic search for pruning, such as Beam search

References and code

1. Jacob Devlin, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, github.com/google-research/bert.git
2. Zhenzhong Lan, et al., ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, github.com/google-research/google-research
3. Kevin Clark, et al., What Does BERT Look At? An Analysis of BERT's Attention, github.com/clarkkev/attention-analysis, **MAIN REFERENCE**
4. Elena Voita, et al, Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned
5. Paul Michel, et al, Are Sixteen Heads Really Better than One?

- Special appreciation to *Reid* for the project advice