Critique Report on the Paper: "ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS [3]"

Sumaiya Tabassum Nimi

Background Information and Purpose of the Research

Pre-trained language models like BERT [1] have been found to exhibit excellent performance on benchmark downstream tasks in NLP. However, the general observation associated with these pre-trained language models is that larger models tend to perform better. But larger models also come along with higher memory requirement (often exceeding the limit of commonly available hardware configurations) and reduced training speed in distributed settings. In order to resolve these issues, in the paper [3], two parameter reduction strategies were proposed that led to significant compression of the popular BERT model, whereas also leading to superior performance at the same time.

Methods Proposed

Exactly three technical contributions, as discussed below, were made in the paper [3] in the design of proposed model called ALBERT, for obtaining superior performance on benchmark tasks compared to the popular BERT model, while also reducing the parameter space at the same time.

- The size of $WordPiece$ embedding, E, was decoupled from the size of the hidden layer, H. The vocabulary embedding matrix was factorized into two smaller matrices, through first projecting one-hot vectors of size V (Vocabulary Size) into a compressed embedding space of size E, from which it was ultimately projected into the hidden space of size H. This proposed factorization resulted in reduced embedding parameter space of size $\mathcal{O}(V \times E + E \times H)$ from its original space of magnitude $\mathcal{O}(V \times H)$, leading to significant reduction when $E << H$. This decoupling allowed for the growth of the magnitude of the hidden layer, without significantly affecting the size of the embedding matrix.

- The second parameter reduction strategy proposed involved sharing of all parameters across the layers of the BERT model. With this strategy deployed, the model could be made deeper without increasing the number of trainable parameters.

- The third contribution made in the paper was the design of a novel self-supervised loss function called SOP that enabled the model to understand coherence between consecutive sentences. This loss function was proposed as a more effective alternative to the NSP loss function used while training the original BERT model, and hence led to better performance.

Strength of the Work

- The greatest strength of the work was achievement of two supposedly conflicting objectives simultaneously. The proposed parameter reduction strategies compressed the original BERT model significantly, while these strategies combined with the proposed novel loss function resulted in both significantly superior performance and faster training. So much so that

the proposed model ALBERT had the potential to replace the traditional BERT model in all target application scenarios.

- Thorough ablation studies done and reported, showing the impact of all possible design decisions.
- The proposed model was demonstrated to perform better than the BERT model on a large number of benchmark tasks, solidifying the claim of superiority of the proposed model. The performance gain on the task called RACE test [2] was particularly significant.

Limitations of the Work

- It can be seen from Table-2 that the three variants of the proposed ALBERT model; ALBERT-large, ALBERT-xlarge, and ALBERT-xxlarge; resulted in less speedup on data iteration (assuming BERT-large as the base) compared to the BERT-base model. This was counter-intuitive, considering the reduced number of parameters in the ALBERT models. Reasons behind these discrepancies should have been discussed in the paper.
- From Table-3, we observed proportional degradation in performance as E was increased from 128, under the condition "all-shared". This was a counter-intuitive trend, that was not observed under the condition "not-shared". This counter-intuitive trend should have been analyzed further, so that future deployment of the model could have benefited from such analyses regarding setting the values of the hyperparameters for the best results for target applications.

Questions Unanswered

- In Table-4, under the condition "$E = 128$", why was the performance on SQuAD tasks similar for cases "all-shared", "shared-attention" and "not-shared"? Shouldn't the "not-shared" case achieve better results, like it did under the condition "$E = 768$"?
- In Section-3.1, it was said, "We observe that the transitions from layer to layer are much smoother for ALBERT than for BERT.". The meaning of "Smoother Transition" was not clear. Although this might mean that both L2 and cosine distances between layer-wise inputs and outputs did not vary from layer to layer, as was seen from Figure-1, it was not clear how this property was significant.Is this property desirable in an ideally trained deep learning model, for best performance? If it was, relevant works establishing the claim should have been cited. Overall, the authors should have explained the significance of the observations from Figure-1 more clearly.

Suggested Future Studies

- Both original BERT and proposed ALBERT models adopted Transformer-based model architectures. It would be interesting to explore efficacy/limits of even lighter architecture settings while designing pre-trained language models.
- Triplet loss like loss function could be incorporated for separating coherent sentences from the incoherent ones. This family of loss functions achieved significant performance gain in computer vision tasks like face detection, it would be interesting to explore the potential of these contrasting loss functions in NLP domain.
- As a sideline task, future works could explore optimizing the inference time of the ALBERT-large, ALBERT-xlarge and ALBERT-xxlarge models, since it was observed from Table-2 of the paper that all these three versions of ALBERT were slower compared to the BERT-base model.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

[3] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.