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

Alexandra Ioana Neagu
TU Delft
neagu.alexandra0206@gmail.com

Abstract

This paper review critically evaluates the research paper titled ***ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*** (Lan et al., 2019), focusing on its strengths, weaknesses, experimental methodology, and reproducibility. The paper demonstrated innovative approaches, particularly in experimental design and addressing research questions. However, its reproducibility and clarity of presentation could be further improved. Overall, the review offers insights into the paper's contributions and areas for enhancement, contributing to a deeper understanding of its significance in the field of NLP.

1 What is this paper about, and what contributions to NLP does it make?

The paper introduces *ALBERT (A Lite BERT)*, a model that addresses the memory and computational inefficiencies of large-scale pre-trained language models (LMs) like BERT. It introduces two primary parameter reduction techniques, factorized embedding parameterization and cross-layer parameterization. The first technique decomposes the large vocabulary embedding matrix into two smaller matrices, separating the size of the hidden layers from the size of the vocabulary embedding, allowing for a larger hidden size without significantly increasing the parameter size. The latter method prevents the parameter count from growing with the network depth by sharing parameters across different layers, significantly reducing the overall number of parameters.

Additionally, the paper presents a new self-supervised loss for sentence-order prediction (SOP), aimed at improving inter-sentence coherence, which is more effective than BERT's original next sentence prediction (NSP) loss.

In my opinion, the main contributions of the paper are outlined as follows:

1. **Parameter reduction techniques**: The factorized embedding parameterization and cross-layer parameter sharing collectively reduce the number of parameters in ALBERT models compared to traditional BERT models. This leads to more efficient use of memory and faster training times without sacrificing performance.

2. **Self-supervised loss for SOP**: The introduction of SOP addresses the limitations of NSP in BERT, enhancing the model's ability to understand and predict the order of sentences, which is crucial for tasks involving multi-sentence inputs.

3. **Performance improvements**: Despite having fewer parameters, ALBERT models outperform their predecessors on several benchmarks. Specifically, they achieve state-of-the-art results on GLUE, RACE, and SQuAD benchmarks, demonstrating the effectiveness of the proposed techniques.

The primary problem addressed by this paper is the scalability of large pre-trained LMs in terms of memory consumption and training speed. As models grow in size to improve performance on various NLP tasks, they become increasingly challenging to train due to hardware limitations and extended training times. The contributions of this paper are significant as they tackle this critical issue of scalability in NLP model training. The introduction of efficient parameter reduction techniques and a novel self-supervised loss function enhances the practicality and performance of pre-trained models. These advancements not only push the boundaries of what is achievable with smaller models but also provide a more sustainable approach to scaling NLP models, which is essential for their deployment in real-world applications. While the novelty of the individual techniques might be debated, their

combined impact and the resulting performance improvements are noteworthy contributions to the field.

1.1 Paper summary

The related work section discusses the evolution of natural language representation learning, noting the shift from pre-training word embeddings to full-network pre-training followed by task-specific fine-tuning. This transition has led to significant performance improvements in NLP tasks, with larger models generally performing better. However, the increasing size of models presents challenges in terms of memory limitations and computational costs, which prior methods like gradient checkpointing and layer reconstruction attempt to address. The section also explores cross-layer parameter sharing, previously applied in standard encoder-decoder tasks and showing mixed results in terms of performance. Finally, the paper examines sentence ordering objectives, comparing ALBERT's SOP loss with other pre-training objectives like BERT's NSP loss. The authors find that SOP, which focuses on predicting the order of two consecutive text segments, is more challenging and beneficial for downstream tasks than NSP.

Section 3 outlines the design choices for the ALBERT model, comparing it to the original BERT architecture. ALBERT retains the transformer encoder backbone with GELU non-linearities but introduces three key innovations:

1. **Factorized Embedding Parameterization**: ALBERT separates the WordPiece embedding size from the hidden layer size, projecting one-hot vectors first into a lower-dimensional embedding space and then into the hidden space, significantly reducing the number of embedding parameters and allowing for larger hidden sizes without a proportional increase in parameters.

2. **Cross-Layer Parameter Sharing**: To improve parameter efficiency, ALBERT shares parameters across layers, which stabilizes network parameters and reduces memory consumption. This approach contrasts with previous models like Universal Transformer (Dehghani et al., 2018) and Deep Equilibrium Models (Bai et al., 2019), showing smoother transitions in embeddings between layers.

3. **Inter-Sentence Coherence Loss**: ALBERT replaces BERT's NSP loss with a SOP loss, which focuses on inter-sentence coherence rather than topic prediction. This change addresses the limitations of NSP, leading to better performance on downstream tasks that require multi-sentence encoding.

The section concludes by showing that ALBERT models have significantly fewer parameters than BERT models with comparable configurations, highlighting the efficiency and performance improvements achieved by these design choices.

Section 4 presents the experimental setup for the study and its results. Results indicate that ALBERT models, despite having fewer parameters than BERT-large, achieve significant improvements in several tasks, including SQuAD and MNLI. ALBERT models also exhibit higher data throughput during training due to reduced communication and computation requirements. Factorized embedding parameterization experiments reveal that smaller embedding sizes under the shared condition (*ALBERT-style*) perform better, leading to the choice of an embedding size of 128 for future settings. Cross-layer parameter sharing experiments show that sharing parameters across all layers slightly hurts performance, particularly with larger embeddings. However, sharing attention parameters alone has minimal impact, suggesting that more granular sharing strategies could be beneficial.

The study also explores the impact of SOP as an additional loss function. Results show that SOP significantly improves performance on downstream tasks compared to no inter-sentence loss or the NSP used in BERT. Further experiments controlling for training time rather than steps reveal that **ALBERT-xxlarge outperforms BERT-large** even when training for the same duration.

Additional training data and the removal of dropout are tested, with results indicating significant improvements in model performance. Notably, removing dropout boosts MLM accuracy and downstream task performance, challenging the conventional use of dropout in large Transformer-based models. The **ALBERT model achieves SOTA results on several benchmarks**, including *GLUE* and *RACE*, demonstrating its superior performance over previous models like BERT, XLNet, and RoBERTa.

Finally, the study reports on the performance of single-model and ensemble configurations of AL-

BERT on various NLP tasks, showing substantial improvements and setting new benchmarks. The findings underscore the efficiency and effectiveness of ALBERT's design choices in achieving high performance with fewer parameters and faster training times.

The paper concludes by highlighting that while ALBERT-xxlarge has fewer parameters than BERT-large and achieves significantly better results, it is computationally more expensive due to its larger structure. Future steps to improve ALBERT include speeding up training and inference through methods like sparse attention and block attention. Additionally, integrating techniques such as hard example mining and more efficient language modelling training could further enhance representation power. Despite the demonstrated benefits of sentence order prediction as a learning task, the authors suggest that there may be other dimensions not yet captured by current self-supervised training losses that could further enhance the model's language representations.

2 What strengths does this paper have?

This paper demonstrates several strengths. Firstly, it presents a novel approach, ALBERT, which achieves significant improvements in performance while using fewer parameters compared to BERT. This innovation highlights the effectiveness of the proposed design choices in enhancing the efficiency and effectiveness of language representation models. Additionally, the paper provides comprehensive experimental results and comparisons with existing models, showcasing the superiority of ALBERT across various downstream tasks. This thorough evaluation enhances the credibility of the findings and underscores the potential impact of ALBERT in advancing natural language understanding tasks.

Furthermore, in my opinion, the paper is well-written and clearly articulates the motivation behind ALBERT's design choices, the experimental methodology, and the interpretation of results. The discussion section offers valuable insights into future research directions, suggesting avenues for further improving language representation models. Overall, the paper's clarity, novelty, and empirical rigour contribute to its strength as a significant contribution to the field of NLP.

3 What weaknesses does this paper have?

While the paper demonstrates several strengths, it also has some weaknesses. One notable aspect is the limited exploration of certain design choices and their potential implications. For instance, the paper discusses the impact of varying vocabulary embedding sizes and cross-layer parameter-sharing strategies but does not delve deeply into alternative configurations or their comparative performance. This lack of comprehensive exploration may leave some questions unanswered regarding the optimal choices for specific model architectures or tasks.

Additionally, although the paper discusses the potential for future research directions, such as sparse attention mechanisms and alternative training methodologies, it could provide more detailed insights into these areas. A more thorough discussion of the limitations of ALBERT and potential avenues for addressing them could enhance the paper's completeness and relevance to the broader research community.

Furthermore, while the experimental results are extensive, the paper could benefit from a more in-depth analysis of the underlying reasons for ALBERT's performance improvements. Exploring the model's behaviour through qualitative analysis or probing techniques could provide deeper insights into its strengths and weaknesses, thereby enriching the interpretation of the experimental findings.

Overall, while the paper presents a compelling case for the effectiveness of ALBERT, addressing these weaknesses could further strengthen its contribution to the field of NLP.

4 What are the research questions that are answered in the experimental evaluation?

In the experimental evaluation, the paper addresses several research questions:

1. **RQ1: What impact do the different design choices have on the model's performance?** The first research question focuses on assessing the impact of specific design choices, such as parameter efficiency and training methodologies, on the performance of ALBERT compared to BERT. Key results include demonstrating that ALBERT achieves significant improvements in performance over BERT despite having fewer parameters, showcasing the effectiveness of its design choices in en-

hancing model efficiency. Insights include the observation that ALBERT models achieve higher data throughput during training compared to BERT models due to less communication and fewer computations, contributing to their improved performance.

2. **RQ2: Is the use of SOP effective?** Another research question explores the effectiveness of the SOP task in improving downstream task performance compared to traditional NSP tasks. The experimental results reveal that SOP leads to consistent improvements in downstream task performance, indicating that modelling sentence coherence through SOP can result in better language representations. Insights from this evaluation suggest that SOP captures more nuanced dependencies between sentences compared to NSP, thereby enhancing the overall effectiveness of pre-training strategies.

Overall, the experimental evaluation addresses these research questions by providing empirical evidence and insights into the impact of design choices and auxiliary training objectives on the performance of ALBERT models.

5 Can you think of how you would improve the work presented in this paper?

One potential improvement that I suggest to the work presented in this paper could involve exploring the incorporation of domain-specific knowledge or context into the pre-training process of ALBERT. This could be achieved by fine-tuning the model on domain-specific datasets or by incorporating domain-specific embeddings during pre-training.

RQ1: How does fine-tuning ALBERT on domain-specific datasets impact its performance on downstream tasks within that domain?

Justification: Fine-tuning ALBERT on domain-specific datasets could enhance its ability to capture domain-specific nuances and improve its performance on tasks within that domain. This would provide insights into the effectiveness of domain adaptation techniques for ALBERT.

RQ2: What is the impact of incorporating domain-specific embeddings into ALBERT's pre-training process?

Justification: Incorporating domain-specific embeddings during pre-training could provide ALBERT with better contextual understanding of domain-specific terms and concepts. This could lead to improved performance on downstream tasks within that domain and could enhance the model's overall adaptability to different domains.

By addressing these research questions, we could potentially enhance the versatility and performance of ALBERT across a wide range of domain-specific tasks, making it more applicable to real-world applications in various fields.

6 How reproducible is the work?

In my opinion, the reproducibility of the work described in the paper would likely fall into the score **3 out of 4**. Researchers and practitioners could mostly reproduce the results described here, possibly by substituting public data for proprietary data. The experimental setup, including datasets and training procedures, seems well-defined and based on publicly available corpora like BOOKCORPUS and English Wikipedia. Additionally, the authors provide the code and the pre-trained models. However, there might be some challenges in precisely replicating certain aspects of the experiments due to factors like hardware specifications (which were not shared in the paper) or specific implementations of training procedures. Additionally, while the paper provides detailed descriptions of the experimental setup, there may still be some subjective elements or parameters that could pose difficulties in exact reproduction without further clarification or standardization. Overall, with some effort and possibly minor adjustments, researchers and practitioners should be able to largely reproduce the reported results (Raff, 2019).

7 How confident are you in your assessment of this paper?

I would rate my confidence in the assessment of this paper as a **4 out of 5**. I tried to check the important points carefully, but there is always a chance that I might have missed something that could affect my ratings. While I have a good understanding of the general area, I did not deeply dive into the paper's more specific details, such as the mathematical aspects, intricacies of the picked benchmarks, or the decisions chosen for the experimental design. However, I feel reasonably confident in my evaluation overall.

References

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2019. Deep equilibrium models. *Advances in neural information processing systems*, 32.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2018. Universal transformers. *arXiv preprint arXiv:1807.03819*.

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

Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems*, 32.