A Comparative Study of DistilBERT and ALBERT

Ahmad Saleh
Email: ahmadsaleh0309@gmail.com

Abstract—**Transformer-based language models such as BERT have achieved state-of-the-art performance on many Natural Language Processing tasks. However, these models are computationally expensive and difficult to deploy in real-world environments. To address these challenges, lightweight variants such as DistilBERT and ALBERT were proposed. This paper presents a comparative analysis of these two models, including their architectures, training methodologies, computational efficiency, and practical applications, supported by references from the original research works.**

Index Terms—**Natural Language Processing, BERT, DistilBERT, ALBERT, Transformers**

I. Introduction

The introduction of BERT (Bidirectional Encoder Representations from Transformers) marked a major advancement in NLP by enabling deep bidirectional contextual understanding [1]. BERT-based models achieved state-of-the-art results on a wide range of benchmarks including GLUE, SQuAD, and SWAG [1].

Despite its success, BERT suffers from major limitations:

- Very large model size
- High memory consumption
- Slow inference speed

For example, the original BERT-large model contains approximately 340 million parameters, making it difficult to deploy on resource-constrained devices [1]. To address these issues, researchers proposed more efficient variants such as DistilBERT [2] and ALBERT [3].

This paper analyzes these two models and compares their design philosophies and practical trade-offs.

II. Related Work

Model efficiency has become a major research direction in NLP. Several strategies have been explored to reduce the size and computational requirements of transformer models.

Knowledge distillation has been widely used to compress large models into smaller ones while retaining most of their performance [2]. Other distillation-based approaches include TinyBERT and MobileBERT, which apply similar principles for specific tasks and devices.

Architectural modifications have also been proposed as an alternative to compression. ALBERT introduces parameter-sharing and embedding factorization to reduce redundancy in BERT models [3]. Quantization and pruning techniques have further been explored to optimize transformer inference [2].

DistilBERT and ALBERT represent two distinct approaches: one based on model compression and the other based on architectural redesign.

III. Background: BERT Overview

BERT is a deep bidirectional transformer encoder trained on large corpora using self-supervised learning [1]. It is pre-trained using two main objectives:

- Masked Language Modeling (MLM)
- Next Sentence Prediction (NSP)

The base version of BERT consists of:

- 12 transformer layers
- 768 hidden units
- 110 million parameters

While highly effective, these characteristics make BERT computationally expensive, motivating the development of more efficient alternatives.

IV. DistilBERT

A. Model Motivation

DistilBERT was introduced as a smaller and faster version of BERT through the technique of knowledge distillation [2]. The goal was to reduce model size and inference time while preserving most of BERT's language understanding capability.

B. Architecture

According to the original paper, DistilBERT:

- Reduces the number of layers from 12 to 6
- Keeps the same hidden dimension as BERT-base
- Removes token-type embeddings

These changes result in approximately 40% fewer parameters and significantly faster inference [2].

C. Training Method

DistilBERT is trained using a combination of:

- Distillation loss from the teacher (BERT)
- Masked language modeling loss
- Cosine embedding loss

This multi-task training objective allows the student model to imitate the behavior of the teacher model effectively [2].

D. Advantages and Limitations

Advantages

- Faster inference than BERT
- Smaller memory footprint
- Suitable for real-time applications

Limitations

- Slight reduction in accuracy compared to full BERT
- Less capacity for highly complex reasoning tasks

V. ALBERT

A. Model Motivation

ALBERT (A Lite BERT) was designed to reduce the parameter count of BERT without sacrificing performance [3]. Instead of compressing a trained model, ALBERT redesigns the architecture itself.

B. Key Innovations

ALBERT introduces two main architectural modifications:

1) Factorized Embedding Parameterization: In standard BERT, the embedding matrix is very large because vocabulary size and hidden size are tied together. ALBERT decomposes this matrix into two smaller matrices, significantly reducing the number of parameters [3].

2) Cross-Layer Parameter Sharing: ALBERT shares the same parameters across all transformer layers. This drastically reduces redundancy while maintaining model depth [3].

C. Training Objective

ALBERT replaces Next Sentence Prediction with Sentence Order Prediction (SOP), which was shown to be more effective for modeling inter-sentence coherence [3].

D. Advantages and Limitations

Advantages

- Much fewer parameters than BERT
- High parameter efficiency
- Comparable or better performance on benchmarks

Limitations

- Training can be slower due to parameter sharing
- Inference speed is not as fast as DistilBERT

VI. Comparison

TABLE I: Comparison of DistilBERT and ALBERT based on original publications

Feature	DistilBERT [2]	ALBERT [3]
Main Goal	Faster inference	Parameter reduction
Technique	Knowledge distillation	Parameter sharing
Model Size	Reduced	Highly reduced
Speed	Very fast	Moderate
Accuracy	Slightly lower than BERT	Similar to BERT
Best Use	Real-time systems	Large-scale training

VII. Applications

Both models are widely used in practical NLP systems. Common applications include:

- Text classification
- Sentiment analysis
- Question answering
- Chatbots
- Information retrieval

DistilBERT is commonly preferred in latency-sensitive applications such as mobile or web services, while ALBERT is more suitable for large-scale training environments where memory efficiency is important.

VIII. Conclusion

This paper presented a comparative study of DistilBERT and ALBERT, two influential lightweight alternatives to BERT. DistilBERT focuses on reducing model size and increasing speed through knowledge distillation, whereas ALBERT achieves efficiency through architectural redesign and parameter sharing.

Both models significantly contribute to making transformer-based NLP systems more practical for real-world deployment. The selection between them depends on specific application requirements such as inference speed, memory constraints, and accuracy needs.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, 2019.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[3] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," *International Conference on Learning Representations (ICLR)*, 2020.