Research Report: DistilBERT and ALBERT

Yousef Mahmoud Ali

September 16, 2025

Abstract

This report explains two models from the BERT family: DistilBERT and ALBERT. Both are designed to make BERT more efficient, but they do so in different ways. DistilBERT focuses on reducing the size of the model while keeping most of its performance, while ALBERT focuses on parameter sharing and reducing memory usage. This paper explains their goals, methods, advantages, limitations, and applications in simple terms.

1 Introduction

BERT (Bidirectional Encoder Representations from Transformers) is one of the most important models in natural language processing (NLP). It showed how powerful transformers can be for tasks like classification, translation, and question answering. But BERT is very large and expensive to train and use. Because of this, smaller and more efficient versions were developed, such as DistilBERT and ALBERT.

2 DistilBERT

2.1 Idea Behind DistilBERT

DistilBERT is a smaller version of BERT that keeps around 95% of its performance while using 40% fewer parameters. It was created by a method called **knowledge distillation**, which means transferring knowledge from a big model (the teacher, BERT) into a smaller model (the student, DistilBERT).

2.2 How it Works

The student model learns not only from the original data but also from the outputs of the teacher model. This helps the smaller model mimic the teacher's behavior without needing all its size. DistilBERT has 6 layers instead of 12 in BERT, but it is trained carefully to stay close in accuracy.

2.3 Advantages

- Faster and smaller, which makes it good for real-time applications.
- Uses less memory, so it can run on normal hardware like laptops or mobile devices.
- Keeps most of BERT's accuracy.

2.4 Limitations

- Slightly less accurate than full BERT.
- Still requires large datasets and teacher models to train properly.

2.5 Applications

- Chatbots and virtual assistants where speed is important.
- Mobile applications with limited resources.
- Real-time translation and summarization tools.

3 ALBERT

3.1 Idea Behind ALBERT

ALBERT stands for **A Lite BERT**. It is not about making BERT shorter like DistilBERT, but about making it more memory efficient. It reduces the number of parameters while keeping the same number of layers.

3.2 How it Works

ALBERT uses two main tricks:

- **Parameter Sharing:** Instead of having separate parameters for each layer, ALBERT reuses them across layers.
- **Factorized Embedding Parameterization:** It breaks the embedding layer into two smaller matrices, which reduces the number of parameters drastically.

3.3 Advantages

- Much smaller in memory size compared to BERT.
- Faster training because fewer parameters need to be updated.
- Despite being smaller, ALBERT achieves results close to or even better than BERT in some benchmarks.

3.4 Limitations

- More difficult to train because of parameter sharing.
- Sometimes less flexible compared to standard BERT.

3.5 Applications

- Large-scale NLP tasks where memory is limited.
- Research environments with limited GPU resources.
- Deployments where model size matters, like cloud-based services.

4 Comparison Between DistilBERT and ALBERT

5 Conclusion

DistilBERT and ALBERT are two different ways to make BERT models more efficient. DistilBERT reduces the number of layers and uses knowledge distillation, while ALBERT focuses on memory efficiency with parameter sharing. Both approaches make large language models easier

Feature	**DistilBERT**	**ALBERT**
Main Goal	Smaller, faster version	Reduce memory size
Method	Knowledge distillation	Parameter sharing + factorization
Layers	6 (half of BERT)	Same as BERT but shared
Performance	95% of BERT	Close to or better than BERT

Table 1: Simple comparison of DistilBERT and ALBERT

to use in real-world applications without always needing powerful hardware. Together, they show how research is moving towards making NLP models smaller, faster, and more accessible to everyone.