A Comparative Study of DistilBERT and ALBERT

Abstract

Language models based on transformers, like DistilBERT and ALBERT, are now used for many tasks related to Natural Language Processing (NLP). Unfortunately, because of their size and computational cost, they cannot be used in practice without modifications. As a result, researchers have developed more efficient transformer models like DistilBERT and ALBERT to help address these issues and make it possible to use language models on typical hardware. In this paper, we will compare DistilBERT and ALBERT from multiple perspectives, including design, training, performance, and the potential applications of each model to the current landscape of NLP-based problems. Through the comparison of the two models, we will highlight how each balances efficiency against accuracy and we will provide examples of how both can be utilized for different situations within NLP.

1. Introduction

Introduced in late 2017, the Transformer architecture transformed Natural Language Processing (NLP) by allowing for parallelization of sequence models using self-attention to compute relationships between all tokens in a sequence of text simultaneously. The development of models like BERT has greatly improved the accuracy and performance of various NLP tasks such as text classification, question answering, and named entity recognition. However, the downside of these powerful models is that they are very resource intensive, consuming very large amounts of memory (footprint) or taking a significant amount of time per inference (time to generate an output based on an input) to process.

As NLP systems are tra nsitioning from research or academic settings to being used in production, efficiency of the models has become an important consideration. This has prompted the development of lightweight transformer models intended to maintain the same level of performance as the original transformer models while using significantly fewer computational resources. Some of the most well-known examples of lightweight transformer models include DistilBERT and ALBERT.

This paper examines the key motivations driving the development of these two lightweight transformer models, their design principles, and the strengths and weaknesses of each model, comparing and contrasting them.

2. Background: BERT and the Efficiency Problem

BERT created bidirectional contextual representations with multiple transformer encoder layers stacked together. Although successful, the base version of BERT has about 110 million parameters, which limits its deployment on mobile devices and in low-latency applications.

The 4 major issues with large transformer models include:

- Excessive memory use
- Limited inference speed
- More energy consumption
- Challenges in deploying at scale

DistilBERT and ALBERT tackle these issues in different ways (knowledge distillation and parameter sharing).

3. DistilBERT

3.1 Introduction

DistilBERT is a smaller version of BERT that offers nearly all the same performance results by creating a significantly smaller model throughout the knowledge distillation process. The knowledge distillation process consists of creating a smaller, so-called “student” model to learn how to predict the same outputs as a larger “teacher” model.

3.2 Architecture

The DistilBERT model consists of fewer layers but has the same functionality as those found in BERT. The main features of this model are:

- Fewer encoder layers
- No token embeddings
- Maintains the functionality of bidirectional self-attention

This approach allows DistilBERT to provide high-quality contextual representations with fewer parameters than BERT.

3.3 Training Process

The model was trained using two distinct losses: (1) on the outputs of the teacher model using distillation loss, (2) on the internal representations of the teacher model using cosine embedding loss and (3) on the word and phrase placements of the masked language modeled data. This multi-objective training approach allows DistilBERT to learn both the output and internal representation of the teacher successfully.

3.4 Performance and Efficiency

Numerous empirical studies have demonstrated that DistilBERT provides approximately 95-97% of the performance available to BERT while using many fewer parameters and generating in much less time. Because of this, DistilBERT is well suited for use with real-time systems such as chatbots and content moderation systems.

4. ALBERT

4.1 Introduction

ALBERT (orchest for A Lite BERT) utilizes an entirely different method of achieving efficient performance than its predecessors do. Instead of relying solely on reducing depth, they focus on using an architecture that minimizes redundancy in a parameterization sense.

4.2 Factorized Embedding Parameterization

In typical transformers, the dimensions that define the word embedding(s) are equal to the dimensions that define the hidden layer(s). By decoupling these two sets of dimensionality, ALBERT can have smaller embedding sizes but still represent the same type of expressive hidden representations thereby greatly reducing the number of overall parameters.

4.3 Cross-Layer Parameter Sharing

In the case of ALBERT, all parameters that would normally be possessed per-layer within transformer architecture are shared amongst all layers as opposed to being unique to each layer, leading to a significantly smaller model size without sacrificing depth.

4.4 Sentence Order Prediction

Rather than use BERT’s Next Sentence Prediction task, ALBERT replaces it with a new task, called Sentence Order Prediction which has been found to produce better inter-sentence coherence and therefore increased downstream performance.

4.5 Performance Characteristics

Although ALBERT has far-fewer parameters than BERT, results show that performance in certain benchmarks is either comparable or better than previous models with respect to their own respective benchmark comparisons. Model size reduction through cross-layer sharing may increase complexity during training and may slow rate of convergence at times.

5. Comparative Analysis

5.1 Model Size & Performance

DistilBERT decreases its size by reducing the number of layers used to create the model.

ALBERT reduces its overall size by sharing parameters and creating a smaller embedding matrix.

While ALBERT can have a larger reduction on its total number of parameters than DistilBERT, DistilBERT is able to have faster inference times due to it being easier to understand and quicker to execute.

5.2 Training Complexity

DistilBERT can only train based on the fact that it has already trained from an existing large model (teacher model).

ALBERT, however, does not require any distillation of the model but has added complexity due to architectural differences.

5.3 Trade-offs of Performance

DistilBERT provides excellent performance with extremely few architectural changes.

ALBERT has a greater success rate than BERT on some tasks even though it is generally smaller in size.

5.4 Use-Case Suitability

Use Case	Preferred Model
Real-time inference	DistilBERT
Memory-constrained systems	ALBERT
Research experimentation	ALBERT
Fast deployment	DistilBERT

6. Applications

- Both models can be used for:
-
- Text Classification
-
- Question Answering
-
- Sentiment Analysis
-
- Toxic Content Detection
-
- Search/Recommendation Systems

DistilBERT tends to be more popular in production settings where low latency is necessary, whereas ALBERT can serve well in larger research environments that rely on memory efficiency.

7. Limitations and Future Directions

Though both models are efficient; they have limitations:

- Smaller than full-scale transformers
- Some loss of performance when arbitrary complexity is introduced into a problem
- Ongoing need for large pre-training corpus

Future study could include the use of sparse and quantized techniques with these models; another area of focus could be the potential merging of distillation and share-and-use parameters.

8. Conclusion

Both ALBERT and DistillBERT are impressive examples of improving transformer efficiency from different perspectives. While DistillBERT improves speed and uses knowledge distilled from a previous version of BERT, ALBERT focuses on reducing the number of parameters used in its overall architecture through innovative design techniques. Therefore, both approaches are helping to make transformer-based NLP more efficiently produced and more available for implementation in real-world applications.

References

1. Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018
2. Sanh et al., *DistilBERT, a distilled version of BERT*, 2019

3. Lan et al., *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*, 2019

4. Vaswani et al., *Attention Is All You Need*, 2017