



DISTILBERT & ALBERT



DISTILBERT & ALBERT: EFFICIENT TRANSFORMERS
NOURAN SHABAN EL-SHORA

1. Introduction to BERT and the Need for Efficient Models

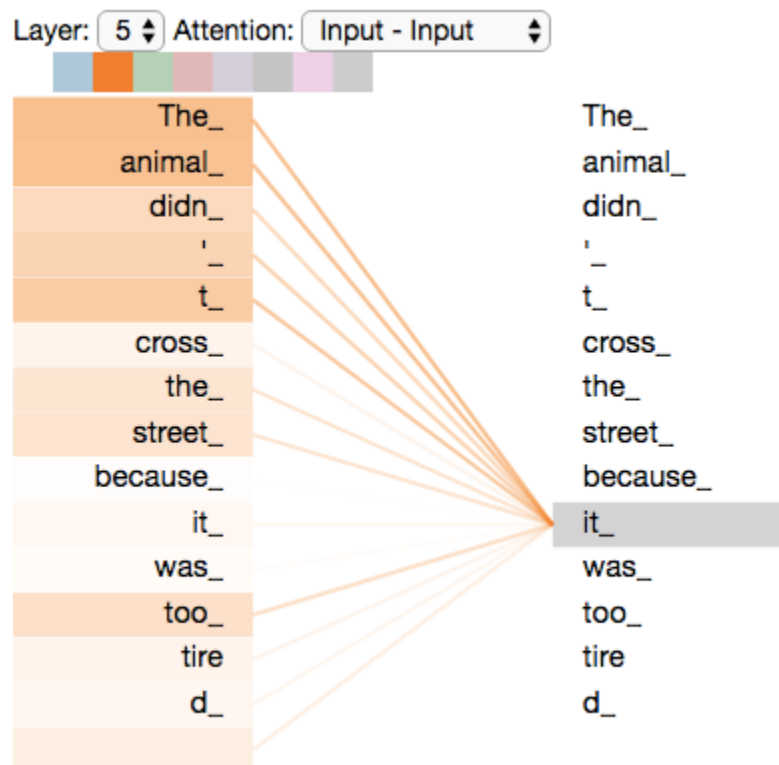
Transformers have changed the way we do Natural Language Processing (NLP). They allow models to understand context in sentences using a mechanism called **self-attention**.

Among transformer-based models, **BERT (Bidirectional Encoder Representations from Transformers)** is one of the most famous. It achieved incredible results in various NLP tasks, from sentiment analysis to question answering.

However, BERT is **huge and resource-hungry**. Its base model has 110 million parameters, while the large version has 340 million. Running BERT requires powerful GPUs, and deploying it in real-world applications like mobile apps or edge devices is difficult.

This led researchers to create **lighter, faster models** that keep most of BERT's accuracy while reducing size and inference time. Two of the most popular ones are **DistilBERT** and **ALBERT**.

- Transformer self-attention diagram



2. DistilBERT

What is DistilBERT?

DistilBERT is a **smaller, faster version of BERT** introduced in 2019. The main idea is to **compress BERT** using a technique called **knowledge distillation**. In knowledge distillation, a smaller “student” model learns to mimic a bigger “teacher” model.

How it works:

- DistilBERT keeps **the same hidden size and attention heads** as BERT but uses **only 6 layers instead of 12**.
- It is trained to **reproduce the output distributions of BERT**, so it learns what the original BERT knows.
- Despite having fewer layers, it retains **97% of BERT’s language understanding ability**.

Benefits of DistilBERT:

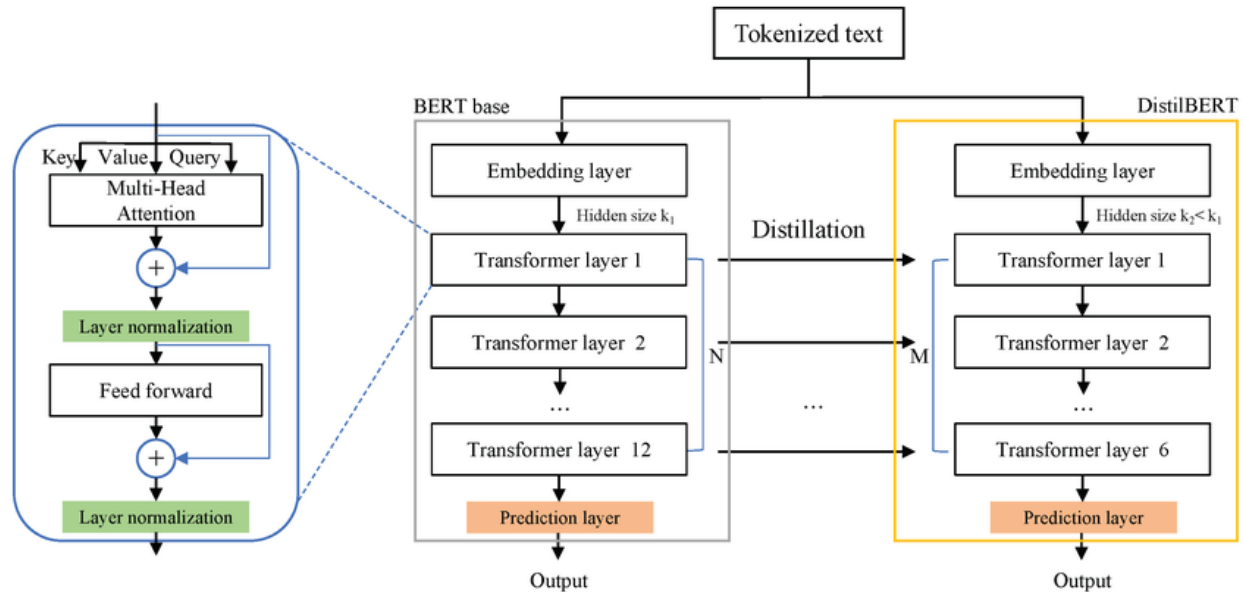
- 40% smaller than BERT
- 60% faster during inference
- Requires less memory, making it ideal for mobile devices

Use Cases:

- Real-time text classification
- Mobile NLP applications
- Chatbots or recommendation systems

Model	Layers	Parameters	Inference Speed	Accuracy (GLUE)
BERT-base 12	12	110M	1x	100%
DistilBERT 6	6	66M	1.6x	97%

- **DistilBERT architecture diagram**



3. ALBERT

What is ALBERT?

ALBERT, or "A Lite BERT," was introduced in 2019 to **reduce the number of parameters in BERT** while keeping high accuracy. The goal was to **improve memory efficiency without losing performance**.

Key innovations:

1. **Factorized Embedding Parameterization** – instead of having huge embedding matrices, ALBERT splits the embedding size from hidden layers. This reduces parameters significantly.
2. **Cross-Layer Parameter Sharing** – weights are shared across all layers. Instead of each layer having its own set of weights, the same weights are reused.

Benefits of ALBERT:

- Much smaller memory footprint
- Maintains high accuracy on benchmarks like GLUE and SQuAD
- Faster training due to fewer parameters

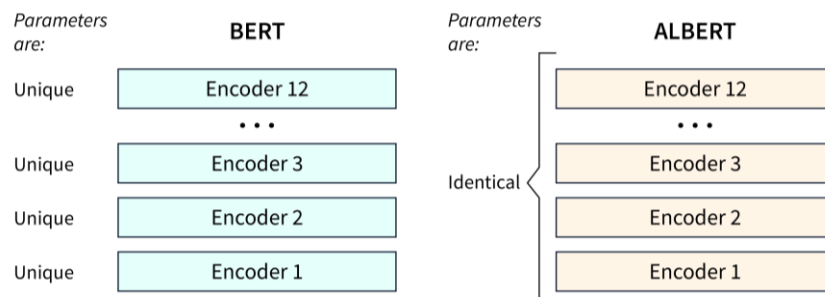
Limitations:

- Slightly slower inference than DistilBERT because it uses all layers at runtime
- More complex architecture

Use Cases:

- Large-scale NLP tasks where memory is limited
- Training on datasets that require long sequences

ALBERT vs BERT parameter sharing diagram



Conclusion:

- DistilBERT is **ideal for speed and light memory usage**, while ALBERT is **ideal for memory efficiency without losing accuracy**.
- Both models are crucial in making transformer models more accessible and usable in real-world applications.