DistilBERT & AlBERT

BERT

Bidirectional Encoder Representations from Transformers introduced by Google researchers in 2018.
The core strength of BERT lies in its bidirectional approach to language understanding. Unlike previous models such as GPT that process text unidirectionally (left-to-right or right-to-left), so BERT understands the word form the text before and after it enabling deeper contextual understanding.

Architecture:

BERT is a multi-layer bidirectional Transformer encoder architecture

- **BERT-Base:** 12 Transformer layers, 768 hidden units, 12 attention heads, ~110 million parameters
- **BERT-Large:** 24 Transformer layers, 1024 hidden units, 16 attention heads, ~340 million parameters

Pre-training:

pre-trained PERT to understand the language using two strategies

1. **Masked Language Modeling (MLM):**
    a. 15% of input tokens are replaced with [Mask] token , and the model learns to predict these masked tokens based on non masked tokens in the sequence. This creates a true bidirectional learning signal, which is impossible with traditional left-to-right language modeling.
2. **Next Sentence Prediction (NSP):**
    a. **The model is trained to predict whether sentence B follows sentence A in the original corpus or is a random sentence. This task helps the model understand inter-sentence relationships.**
    b. **its a binary classification the output is 0 or 1**
    c. **[cls] token appears in the beginning of the sentence**
    d. **[sep] token appears in the end of each sequence**

Fine-tuning:

After pre-training on massive unlabeled corpora, BERT can be fine-tuned for a specific task such as sentiment analysis, question answering, named entity recognition, and text classification.
We only add a small layer to the core model and most hyperparameters stay the same as in bert training.

DistilBERT

Distilled version of BERT introduced in 2019 by Hugging Face., DistilBERT addresses BERT's computational limitations through knowledge distillation a technique where a smaller (student) model learns from a larger (teacher) model BERT

Distillation:

Distillation is a technique where a large, powerful model (*teacher*) trains a smaller, faster model (*student*) to behave similarly.
compressing knowledge from a big model into a smaller one without retraining from scratch

well-trained model such as BERT (*teacher*) that achieves high accuracy.
Generate soft labels instead of producing only hard labels (0/1), the teacher outputs probability distributions, showing its confidence across all classes
The student model (DistilBERT) is trained to mimic the teacher's outputs by matching these soft probabilities.
We can get a lighter and faster model that preserves most of the teacher's performance.

Architecture:

- DistilBERT uses fewer layers than BERT (6 layers instead of 12 in BERT-base).
- The hidden dimensions are the same as BERT 768 hidden units.
- It uses 12 attention heads Same as BERT-base
- The Next Sentence Prediction (NSP) objective is removed, which simplifies the model.

Performance:

- ~40% fewer parameters than BERT
- ~60% faster inference
- Retains ~97% of BERT's performance
- Requires minimal computational resources compared to pre-training BERT from scratch

ALBERT(A Lite BERT)

ALBERT introduced in 2019 by Google Research.This model uses parameter reduction techniques to make the model smaller and faster.
ALBERT shares parameters across layers and factorizes the embedding layer, reducing memory and computation requirements and the model can still capture rich language representations

Architecture:

- **ALBERT-base:** 768 hidden units, 12 layers (shared), 128 embedding size [11M] parameters vs BERT-base's [110M]
- **ALBERT-large:** 1024 hidden units, 24 layers (shared), 128 embedding size [18M] parameters vs BERT-large's [340M]
- **ALBERT-xxlarge:** 4096 hidden units, 12 layers (shared), 128 embedding size [223M] parameters

Factorized Embedding Parameterization

Large vocabulary embeddings are factorized into smaller matrices, decreasing the model size
ALBERT decomposes the embedding matrix **v * h** into two smaller matrices: **v * e and e * h**, where **e << h** ,e is the embedding size,h is the hidden size and v is the vocabulary size.

Cross-layer Parameter Sharing

Instead of learning unique parameters for each Transformer layer, ALBERT shares parameters across layers.The model learns one set of parameters for the Transformer block (multi-head attention + feed-forward network) and reuses them across all layers.

Sentence-Order Prediction (SOP)

ALBERT replaces BERT's Next Sentence Prediction (NSP) with Sentence Order Prediction (SOP), where the NSP task **is sentence B the actual next sentence after A?**
SOP task **Are two consecutive sentences in correct or reversed order?**
SOP is more challenging and better captures inter-sentence coherence.