BERT, which stands for **Bidirectional Encoder Representations from Transformers**, is a new language representation model. Unlike previous models such as ELMo and OpenAI GPT, BERT is designed to pre-train **deep bidirectional representations** from unlabeled text. It achieves this by jointly conditioning on both left and right context in all layers, which is a key difference from standard unidirectional language models that limit the context available during pre-training. The limitation of unidirectional models is particularly harmful for token-level tasks like question answering where both left and right context are crucial.

The pre-training process for BERT uses two unsupervised tasks:

1. **Masked Language Model (MLM)**: Inspired by the Cloze task, this involves randomly masking a percentage (15% in experiments) of input tokens and training the model to predict the original vocabulary ID of the masked word based on its context. To mitigate the mismatch between pre-training (where [MASK] tokens appear) and fine-tuning (where they do not), a mixed strategy is used for the 15% chosen tokens: 80% are replaced with [MASK], 10% with a random token, and 10% remain unchanged. This objective enables the representation to fuse left and right context, allowing for the pre-training of a deep bidirectional Transformer.
2. **Next Sentence Prediction (NSP)**: This task involves presenting the model with two sentences, A and B, and training it to predict whether B is the actual next sentence that follows A (labeled IsNext) or a random sentence from the corpus (labeled NotNext). This helps the model understand sentence relationships, which is beneficial for downstream tasks like Question Answering and Natural Language Inference.

For input, BERT can represent both single sentences and pairs of sentences unambiguously in a single token sequence. A special classification token ([CLS]) is always the first token, and its final hidden state is used as the aggregate sequence representation for classification tasks. Sentence pairs are separated by a special token ([SEP]), and a learned embedding indicates which sentence each token belongs to. The input representation for a token is the sum of its token, segment, and position embeddings.

Following pre-training, the BERT model can be **fine-tuned** with just one additional output layer for various tasks. Fine-tuning involves initializing the model with pre-trained parameters and training all parameters end-to-end using labeled data from the downstream task. The architecture is unified across tasks with minimal differences from the pre-trained model. Fine-tuning is relatively inexpensive compared to pre-training. BERT uses bidirectional self-attention in its Transformer architecture, contrasting with OpenAI GPT's left-context-only self-attention.

BERT has demonstrated empirical power by achieving new state-of-the-art results on eleven natural language processing tasks. It significantly improved scores on benchmarks like GLUE

(General Language Understanding Evaluation) and SQuAD (Stanford Question Answering Dataset). For instance, BERTBASE and BERTLARGE outperformed previous systems on GLUE tasks by a substantial margin. On SQuAD v1.1, BERTLARGE set a new state of the art, even as a single model compared to prior ensemble systems. For SQuAD v2.0, BERT achieved a significant F1 improvement over the previous best system.

Ablation studies confirmed the importance of the pre-training tasks and model architecture. Removing the NSP task hurt performance on tasks like QNLI, MNLI, and SQuAD. Comparing the MLM model to a left-to-right (LTR) model (similar to OpenAI GPT), the LTR model performed worse on all tasks, with notable drops on MRPC and SQuAD, highlighting the benefit of bidirectional representations. Even adding a BiLSTM to the LTR model during fine-tuning did not match the performance of pre-trained bidirectional BERT models. The studies also showed that **larger model sizes** (like BERTLARGE) lead to significant accuracy improvements, even on small datasets, suggesting that fine-tuning allows downstream tasks to benefit from more expressive pre-trained representations. BERT is also shown to be effective for both the fine-tuning approach and a feature-based approach where contextual embeddings are extracted from the pre-trained model.