BERT: Pre-training of Deep Bidirectional Transformers

Traditional language models (e.g., GPT, ELMo) process text either left-to-right *or* right-to-left, limiting their ability to capture full bidirectional context. **BERT** solves this by introducing a **deep bidirectional Transformer** pretrained on two novel tasks, enabling state-of-the-art performance across NLP benchmarks with minimal task-specific modifications.

Key Innovations

1. **Masked Language Modeling (MLM)**
    - Randomly masks 15% of input tokens and predicts them using *full bidirectional context*.
    - Solves the "see-itself" problem in naive bidirectional models (e.g., "the [MASK] sat on the mat" → "cat").

2. **Next Sentence Prediction (NSP)**
    - Pretrains the model to predict if two sentences are consecutive (e.g., "[CLS] A [SEP] B [SEP]" → IsNext).
    - Improves performance on tasks requiring sentence-pair understanding (e.g., Q&A, inference).

3. **Transformer Architecture**
    - Uses the **Transformer encoder** (from "Attention Is All You Need") but *bidirectionally*.
    - Two model sizes:
        - **BERT-Base**: 12 layers, 768 hidden dim, 12 heads (110M params).
        - **BERT-Large**: 24 layers, 1024 hidden dim, 16 heads (340M params).

Model Structure

Input Representation

- Token embeddings (WordPiece), segment embeddings (sentence A/B), and positional embeddings.
- Special tokens: [CLS] (classification), [SEP] (sentence separator), [MASK] (masked token).

Pre-training

- Trained on **BooksCorpus (800M words)** + **English Wikipedia (2.5B words)**.
- MLM and NSP are jointly optimized.

Fine-tuning

- Requires only **task-specific output layers** (e.g., softmax for classification).
- Processes pairs of sentences (e.g., Q&A, entailment) by concatenating them with [SEP].

Results GLUE Benchmark: Outperforms prior models by **7.7% average accuracy**.

- **SQuAD 1.1**: Achieves **93.2% F1** (single model), surpassing human performance (91.2%).
- **Named Entity Recognition (NER)**: New SOTA (**96.4% F1** on CoNLL-2003).

Why It Worked:

- **Bidirectional context** captures deeper linguistic patterns than unidirectional models.
- **Pre-training + fine-tuning** reduces need for task-specific architectures.