BERT Architecture

Definition of BERT

BERT (Bidirectional Encoder Representations from Transformers) is an **encoder-only Transformer architecture** designed for **natural language understanding** tasks rather than text generation.

Formally, BERT processes the entire input sequence simultaneously and learns contextual representations using both left and right context.

Architecture Type

- Encoder-only Transformer architecture
- Employs bidirectional self-attention
- Does not generate text autoregressively

High-Level Architectural Flow

Input sentence

Subword tokenization

Token Embedding + Positional Encoding

Bidirectional Multi−Head Self−Attention

Feed Forward Neural Network

Stacked Encoder Blocks (N layers)

Context–aware token embeddings

Core Components of BERT

1. Token Embeddings

- Convert input tokens into continuous vector representations
- Utilize subword tokenization (WordPiece)
- Effectively handle rare and unknown words

2. Positional Encoding

- Incorporates word order information into embeddings
- Necessary because attention mechanisms lack inherent positional awareness

3. Bidirectional Self-Attention

- Allows each token to attend to both preceding and succeeding tokens
- No masking is applied during attention computation

Illustrative Example:

The bank approved the loan

To interpret the word **“bank”**, BERT attends to:

approved + loan

This enables correct semantic interpretation as a financial institution.

4. Feed Forward Neural Network

- Applies position-wise non-linear transformations
- Enhances the expressive capacity of token representations

5. Encoder Block (Repeated)

Each encoder layer in BERT consists of:

- Multi-head self-attention
- Feed forward neural network
- Residual connections
- Layer normalization

Training Objectives of BERT

1. Masked Language Modeling (MLM)

- Random tokens are replaced with a special `[MASK]` token
- The model learns to predict the masked tokens using surrounding context

Example:

```
The bank [MASK] the loan
```

Predicted token:

```
approved
```

2. Next Sentence Prediction (NSP)

- Determines whether one sentence logically follows another
- Facilitates learning sentence-level relationships

Rationale for Bidirectional Processing

Understanding natural language requires access to both:

- Prior context
- Subsequent context

Unlike autoregressive models, BERT processes the complete sentence simultaneously, enabling bidirectional contextual understanding.

Applications of BERT

- Sentence and document embeddings
- Semantic search
- Text classification
- Question answering
- Named entity recognition

Limitations of BERT

- Not designed for text generation
- Does not support autoregressive prediction
- Unsuitable for chatbot systems without architectural modification

Concise Definition for Exams or Interviews

> *BERT is an encoder-only Transformer that learns bidirectional contextual representations for language understanding tasks.*

Memory Aid

> **BERT reads the entire sequence before performing interpretation.**

Comparison with GPT

Feature	**GPT**	**BERT**
Architecture	Decoder-only	Encoder-only
Attention	Masked	Bidirectional
Text generation	Yes	No
Primary use	Generation	Understanding