# Summary of BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

The main goal of the paper is to **improve the performance of NLP models** by introducing a **new way of pre-training a language model** that understands language **from both directions (left and right context)** at the same time.

- **BERT** stands for **Bidirectional Encoder Representations from Transformers**.
- It is a **pre-trained language model** based on the **Transformer architecture**.
- Unlike previous models that read text **left-to-right or right-to-left**, BERT reads the **entire sentence at once (bidirectionally)** to get a better understanding of context.
- It is **pre-trained** on a large amount of text using two tasks:

- **Masked Language Modeling (MLM):** Random words in a sentence are masked, and the model tries to guess them.

- **Next Sentence Prediction (NSP):** The model learns if one sentence logically follows another.

- After pre-training, BERT can be **fine-tuned** on specific tasks like sentiment analysis, question answering, and named entity recognition — achieving **state-of-the-art results**.

**Why BERT is Important:**

- It set a new standard in NLP by allowing models to learn **deep understanding of language** with **bidirectional context**.
- It made it easy to apply the same model to many different NLP tasks with just a small amount of fine-tuning.

## Example Sentence:

**"The man wore a mask to the party."**

Imagine we want the model to understand the word **"mask"** in this sentence.

**Traditional (left-to-right) models:**

They only see:
**"The man wore a"**, so they might guess “hat” or “suit”.

**BERT (bidirectional):**

It sees the **full sentence** — “The man wore a ____ to the party.”
So it understands that **"mask"** fits best because it's something you might wear to a party.

### BERT’s Pre-training Tasks:

**1. Masked Language Modeling (MLM):**

BERT is trained to guess missing words.

**Input to model:**
**"The man wore a [MASK] to the party."**
Model tries to predict: **"mask"**

It uses **both left ("The man wore a") and right ("to the party") context** to predict.

---

**2. Next Sentence Prediction (NSP):**

BERT also learns if one sentence follows another logically.

**Sentence A:** "The man wore a mask to the party."
**Sentence B:** "Everyone thought his costume was amazing."

The model learns that **B follows A**, because they are logically connected.

If we give:

**Sentence A:** "The man wore a mask to the party."
**Sentence B:** "Bananas are yellow."

The model learns this is **not** a logical next sentence.

---

**After pre-training:**

You can fine-tune BERT on real tasks like:

- Sentiment analysis: Is a review positive or negative?
- Question answering: Answering questions based on a passage.
- Named entity recognition: Finding names, places, etc., in text.