BERT and its Applications

Agenda

- Key Questions
- What is BERT?
- What are the variants of BERT?
- How is BERT trained?
- What are the downstream applications of BERT?

Key Questions

What is BERT?

What are the variants of BERT?

How is BERT trained?

What are the downstream applications of BERT?

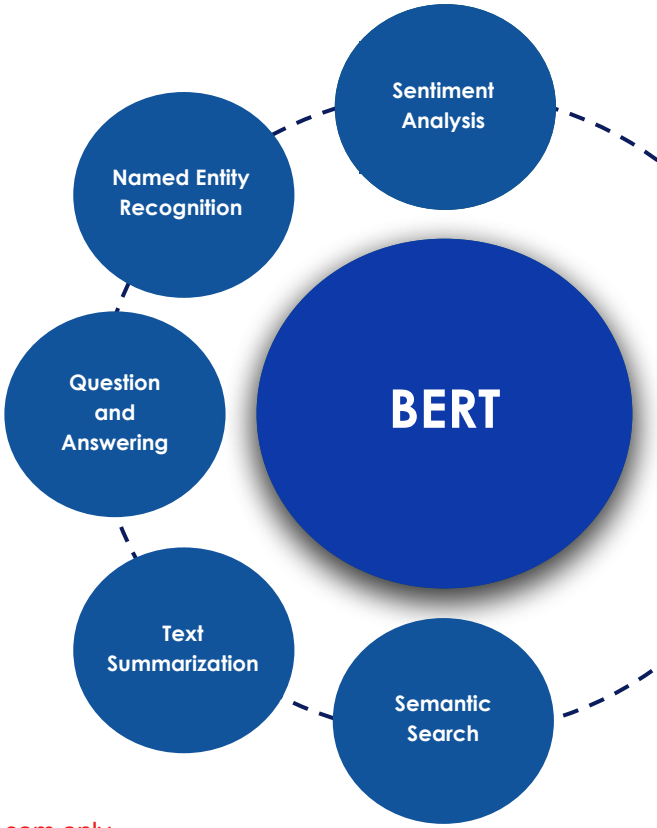
What is BERT?

BERT stands for **B**idirectional **E**ncoder **R**epresentations **T**ransformer

It was introduced in a **research paper** by **Jacob Devlin** and his **colleagues** at Google Research in **2018**.

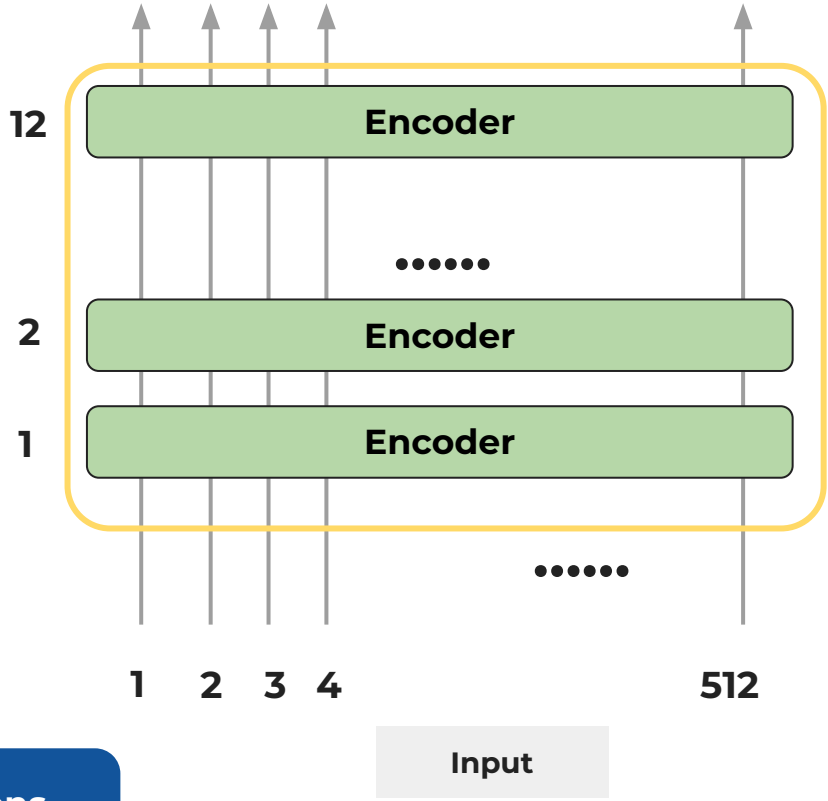
BERT is an **encoder only transformer** model that leverages a **stack** of **encoders** to get an deeper **understanding** of **language context**

BERT **revolutionized** Natural Language Processing (NLP) by exhibiting **state-of-the-art performance** in a variety of NLP tasks



What are the variants of BERT?

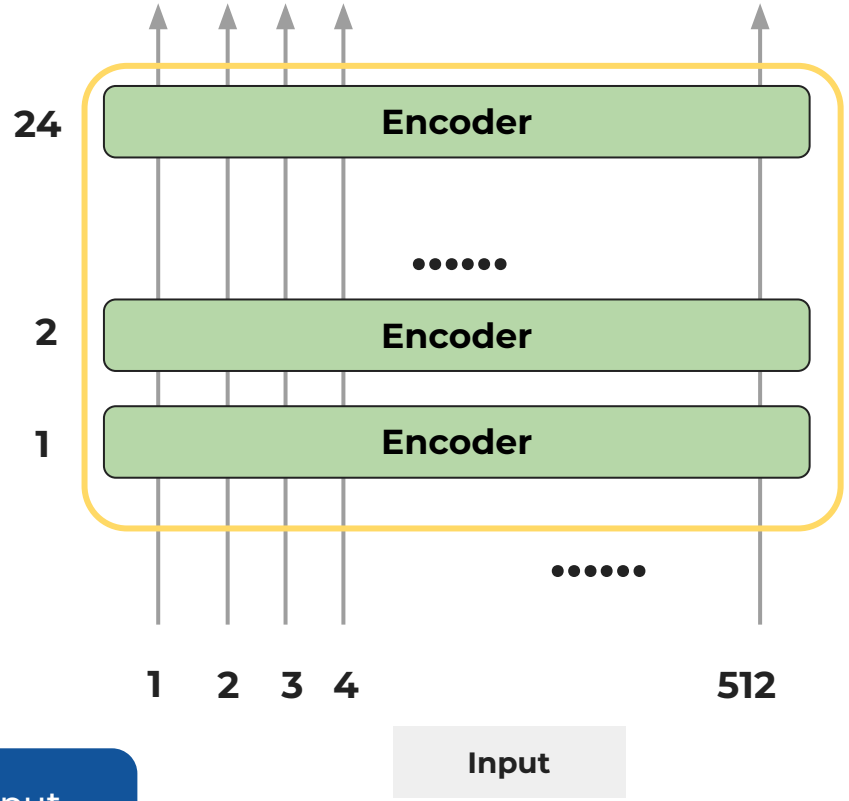
	$BERT_{BASE}$
# parameters	110M
# encoders	12
# attention heads	12
word embeddings dimension	768



$BERT_{BASE}$ has a **limitation** on the **number of input tokens** - it can take a maximum of **512** input tokens

What are the variants of BERT?

	$BERT_{LARGE}$
# parameters	340M
# encoders	24
# attention heads	16
word embeddings dimension	1024



$BERT_{LARGE}$ has a similar limitation on the number of input tokens as $BERT_{BASE}$

How is BERT trained?

BERT's training essentially consists of two stages - a **Pre-training** Stage and a **Fine-tuning** Stage

Pre-training Stage

The model builds a foundational understanding of language

Involves exposing the model to a vast amount of text data, where it learns about word relationships and gains contextual knowledge

+

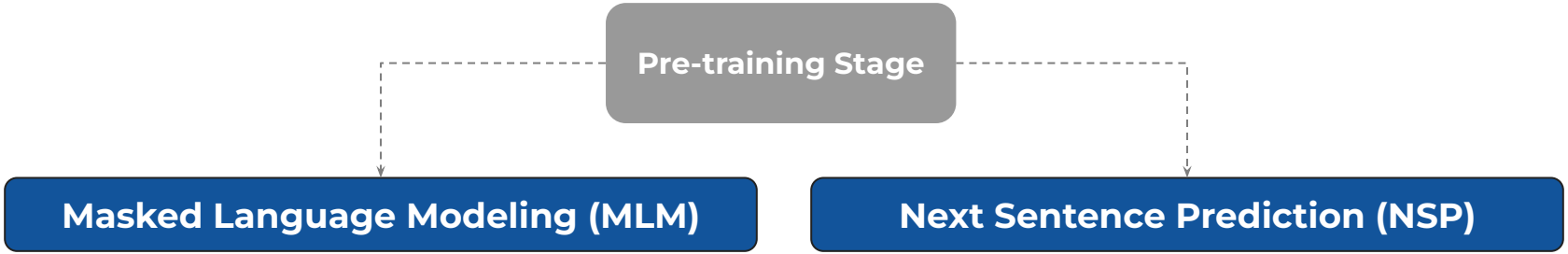
Fine-tuning Stage

The model adapts its foundational understanding of language to perform better on downstream tasks

Involves adjustment of model parameters (weights) to specialize in a particular task.

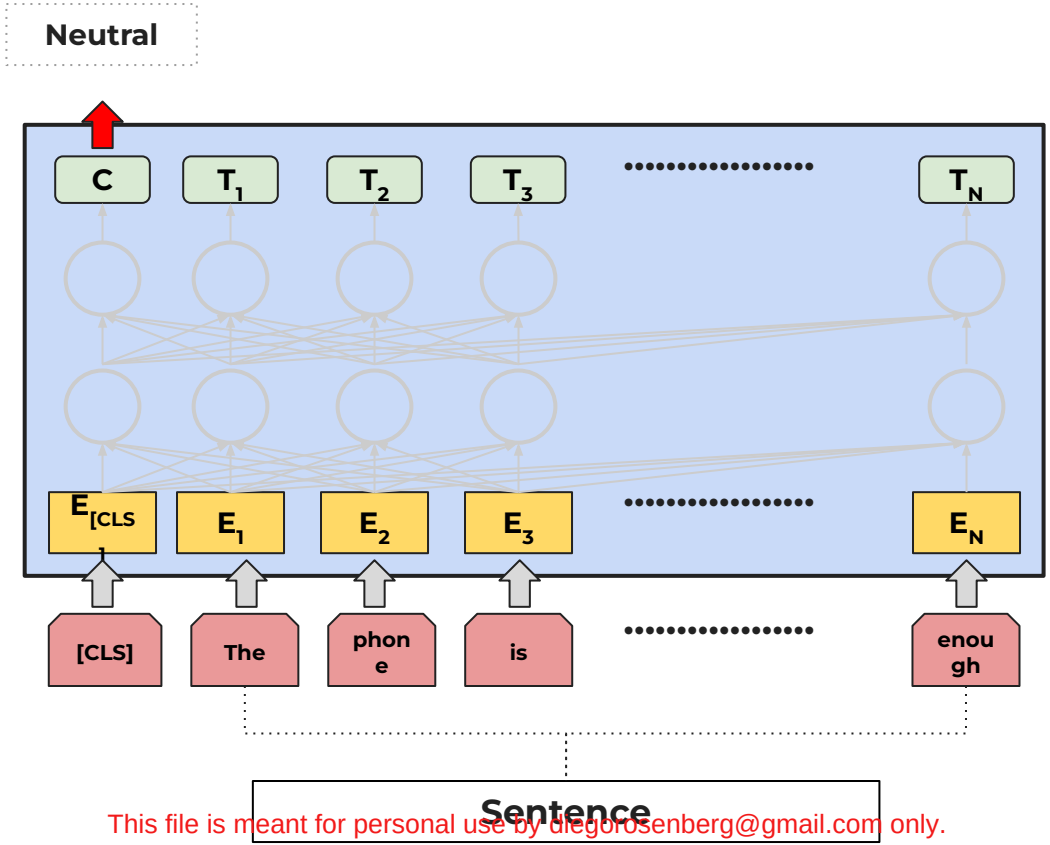
BERT - Pre-training Stage

BERT's **pre-training** stage is done in **two parts**



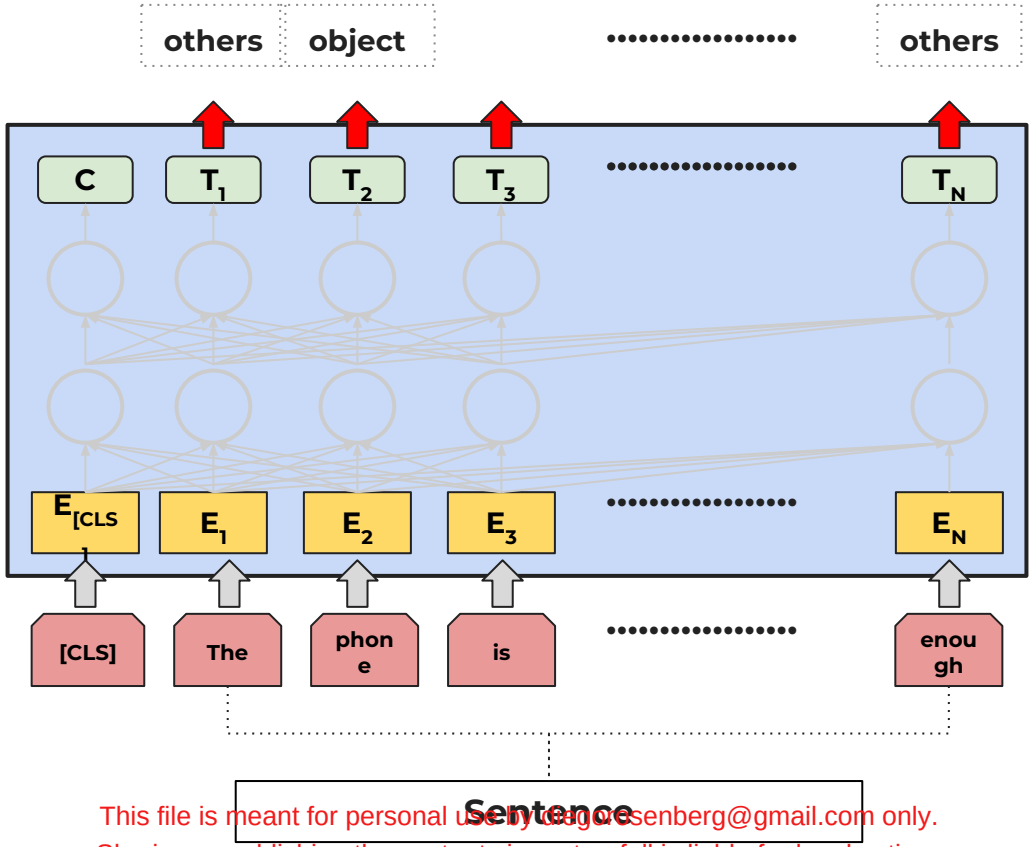
What are the downstream applications of BERT?

Sentiment Classification



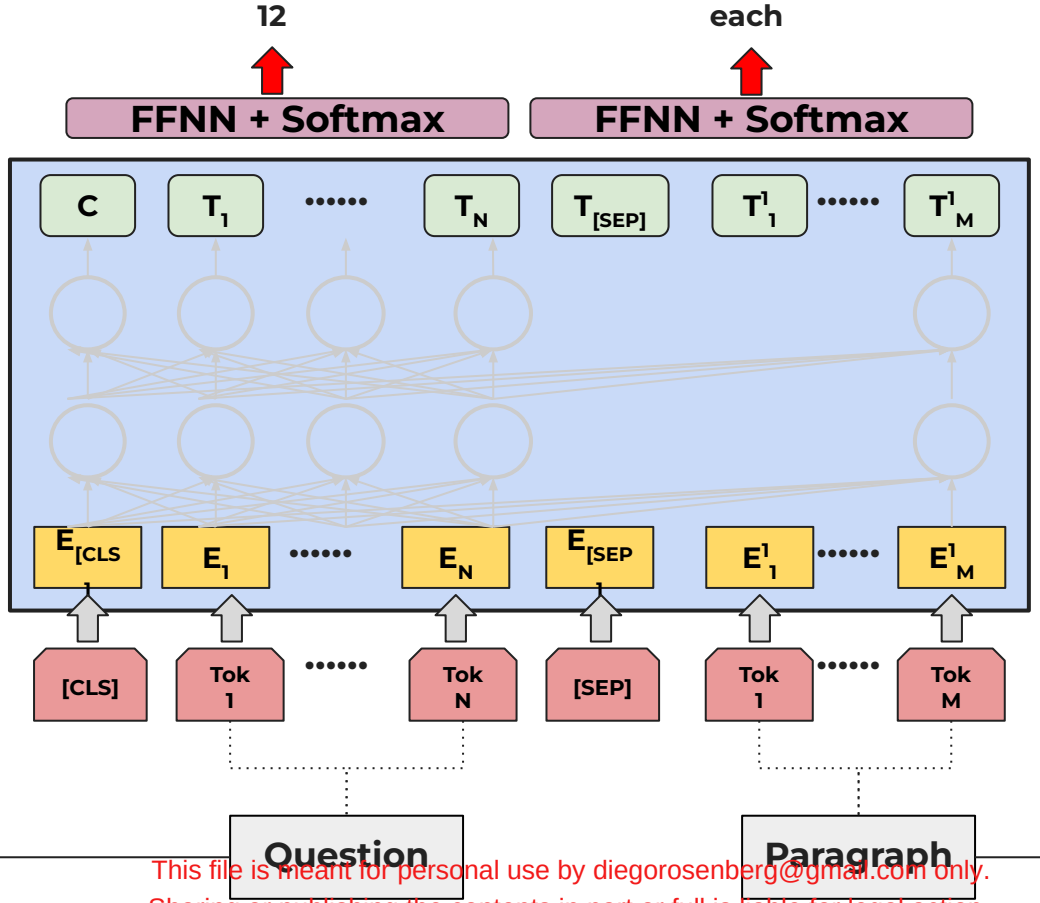
What are the downstream applications of BERT?

Named Entity Recognition



What are the downstream applications of BERT?

Question Answering



How many attention heads does each transformer encoder in BERT have?

BERT consists of 12 transformer encoders with **12 attention heads in each**, totaling approximately 110 million parameters. This design allows BERT to deeply understand language nuances and relationships, making it adept at various language tasks.

Great Learning

Happy Learning !