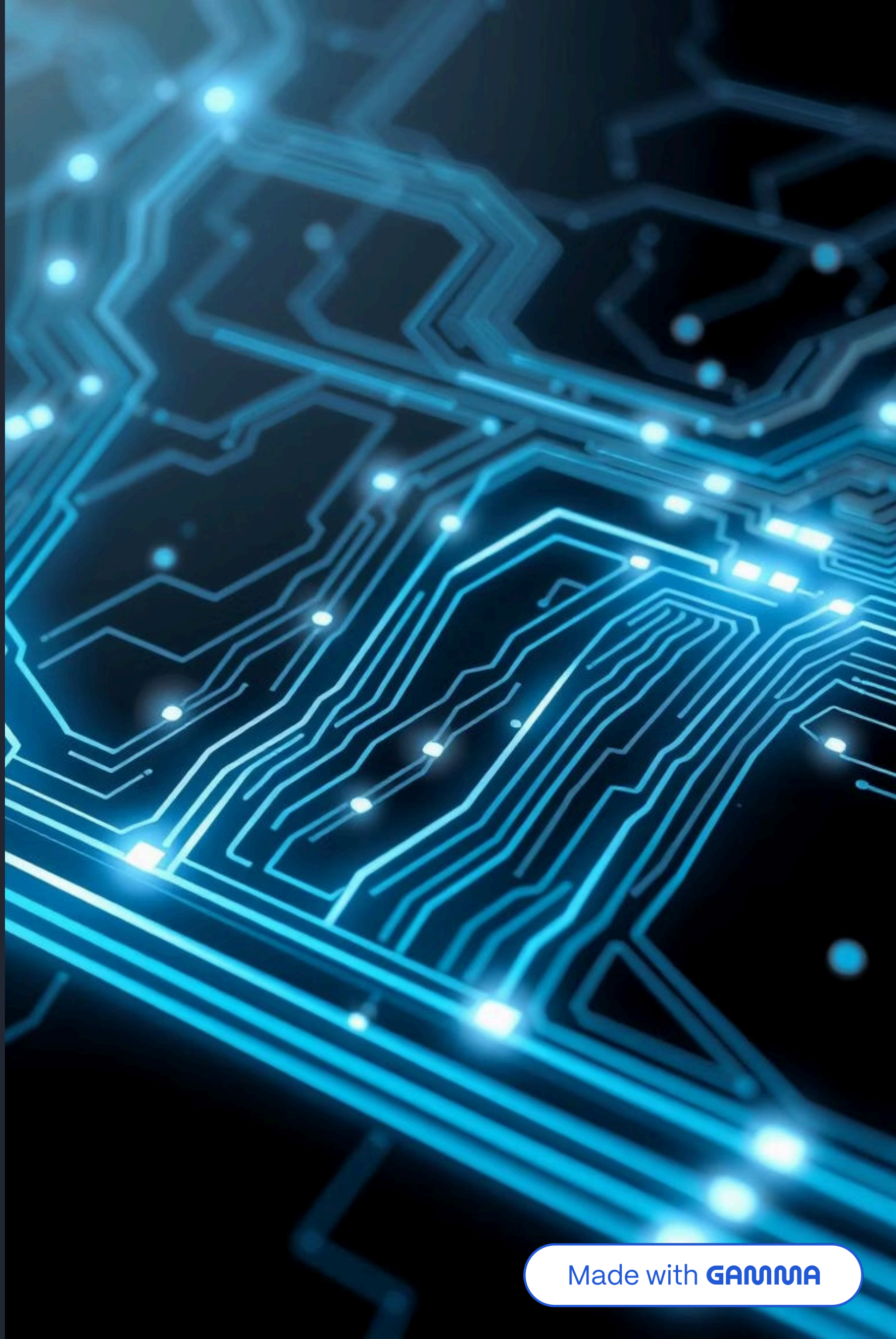
Attention & Transformers

Revolutionizing Natural Language Processing (Applied Department)

by: Ayoub Benyamina



Opening Questions

Can a machine understand human emotions?

Can a model detect if a movie review is positive or negative?

How do Transformers use attention mechanisms to interpret context?

Project Goal

Build and fine-tune a Transformer-based model (BERT-base) to classify movie reviews as **Positive** or **Negative**, using the IMDB dataset and a custom PyTorch training loop.

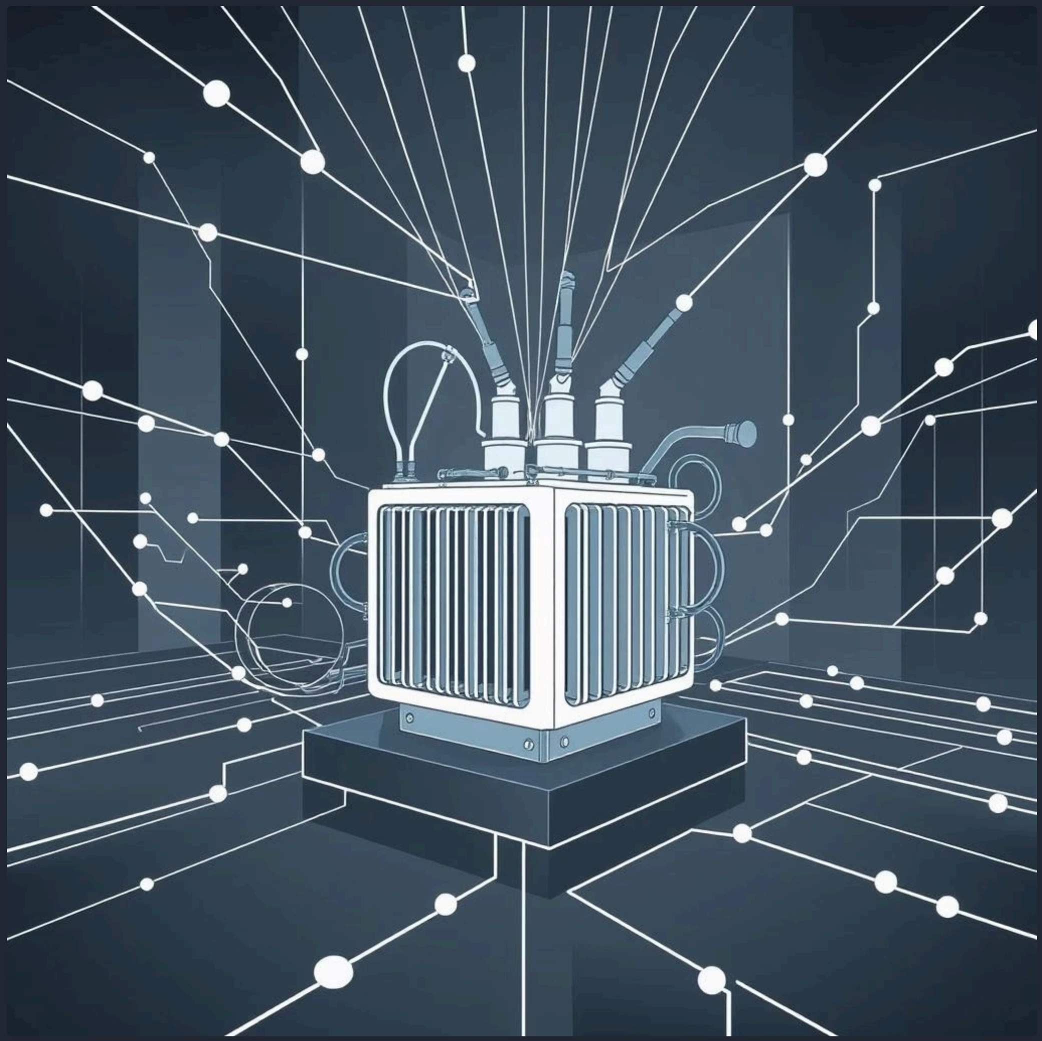
What Are Transformers & Attention?

What is a Transformer?

A Transformer is a neural network architecture designed for NLP and sequence tasks. It processes words **in parallel**, not sequentially, enabling:

- Faster training
- Better long-range understanding
- Deep contextual representation

Transformers are the foundation of modern models like BERT, GPT, T5, and RoBERTa.



What is Attention?

Attention is a mechanism that lets the model **focus on the most relevant words** in a sentence.

Example: “The movie was good but the ending was terrible.” The word *terrible* carries more emotional weight → attention gives it a higher score.

What is Self-Attention?

Self-Attention allows the model to compare **every word with every other word**, finding relationships and dependencies.

It helps the model understand:

- Meaning
- Tone
- Emotional context

Relation Between Transformers and Attention

- Transformers **are built entirely on self-attention** (no RNNs or CNNs).
- BERT uses **multi-head self-attention** to look at the text from different perspectives.
- This makes Transformers powerful for sentiment analysis and emotion understanding.

Why Transformers?

1

Contextual Understanding

Utilize Self-Attention mechanisms to deeply understand the context and relationships between words in a sequence.

2

Long-Range Dependencies

Outperform traditional RNNs/LSTMs in capturing dependencies across long text sequences, crucial for complex sentences.

3

State-of-the-Art Performance

Consistently achieve top results in various Natural Language Processing tasks, setting new benchmarks.

4

Bidirectional Processing

BERT's unique bidirectional approach reads sentences both forward and backward, leading to a richer understanding.

5

Text Classification Excellence

Highly effective for tasks like sentiment analysis, making them an ideal choice for classifying movie reviews.

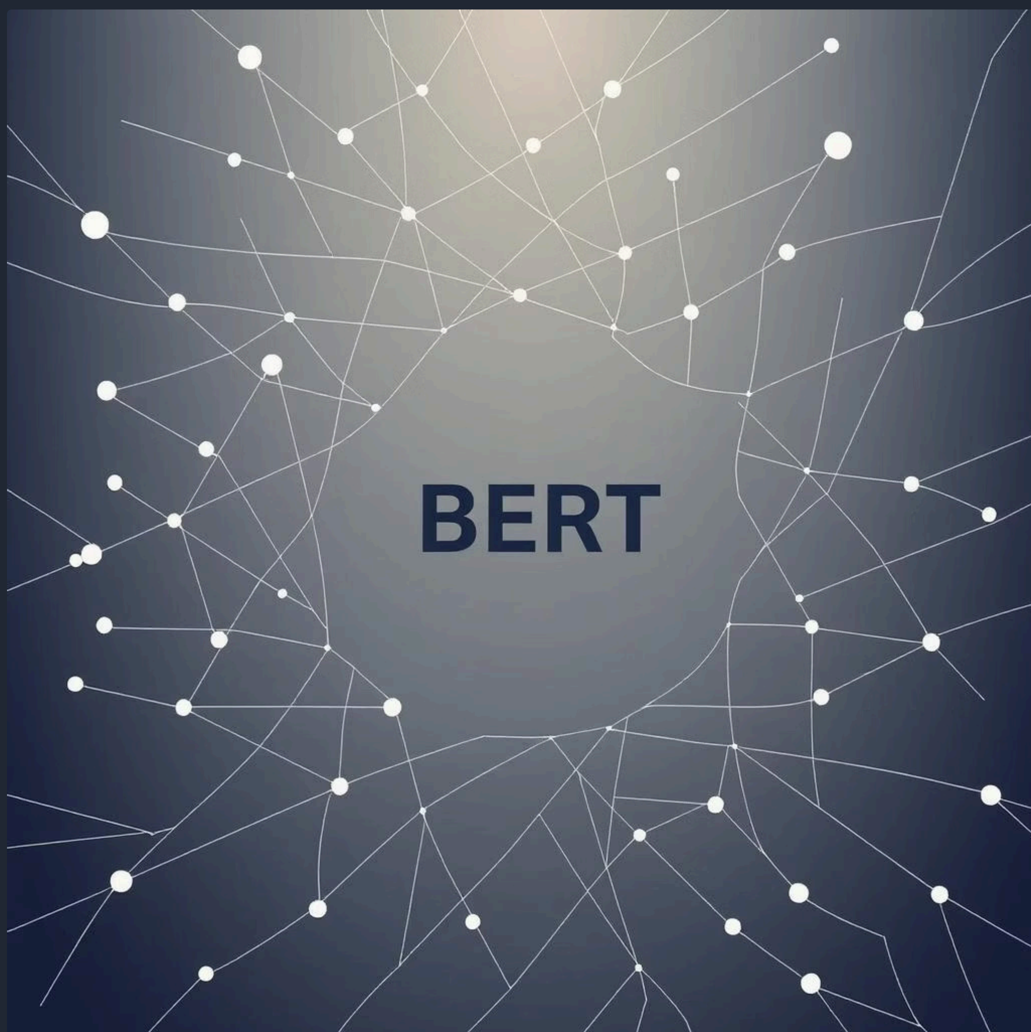
Dataset: IMDB Movie Reviews

- **50,000 labeled movie reviews**: A comprehensive collection of text data.
- **Balanced Dataset**: Consists of 25,000 positive and 25,000 negative reviews, ensuring unbiased learning.
- **Clean and Widely Used**: Renowned for its quality and serves as a common benchmark in NLP research.
- **Strategically Split**: Divided into Training, Validation, and Test sets for robust model development and evaluation.
- **Ideal for Sentiment Learning**: Its characteristics make it perfect for training models to understand and classify emotional tone in text.

The IMDB dataset provides a solid foundation for developing and testing our sentiment analysis model, allowing us to accurately gauge its performance.

Model & Fine-Tuning

Model: BERT-base



- **110 Million Parameters**: A robust pre-trained model capable of learning complex linguistic patterns.
- **Fine-tuned for Sequence Classification**: Adapted specifically to predict whether a review is Positive or Negative.
- **Leveraging Pre-trained Knowledge**: Benefits from extensive prior training on a massive text corpus, accelerating learning for our specific task.

Training Steps

1

Tokenization

Converting raw text into a format BERT understands using its specialized tokenizer.

2

Custom PyTorch Loop

Implementing a flexible and efficient training pipeline.

3

Forward Pass & Loss

Processing data through the model and quantifying prediction errors.

4

Backpropagation & Optimization

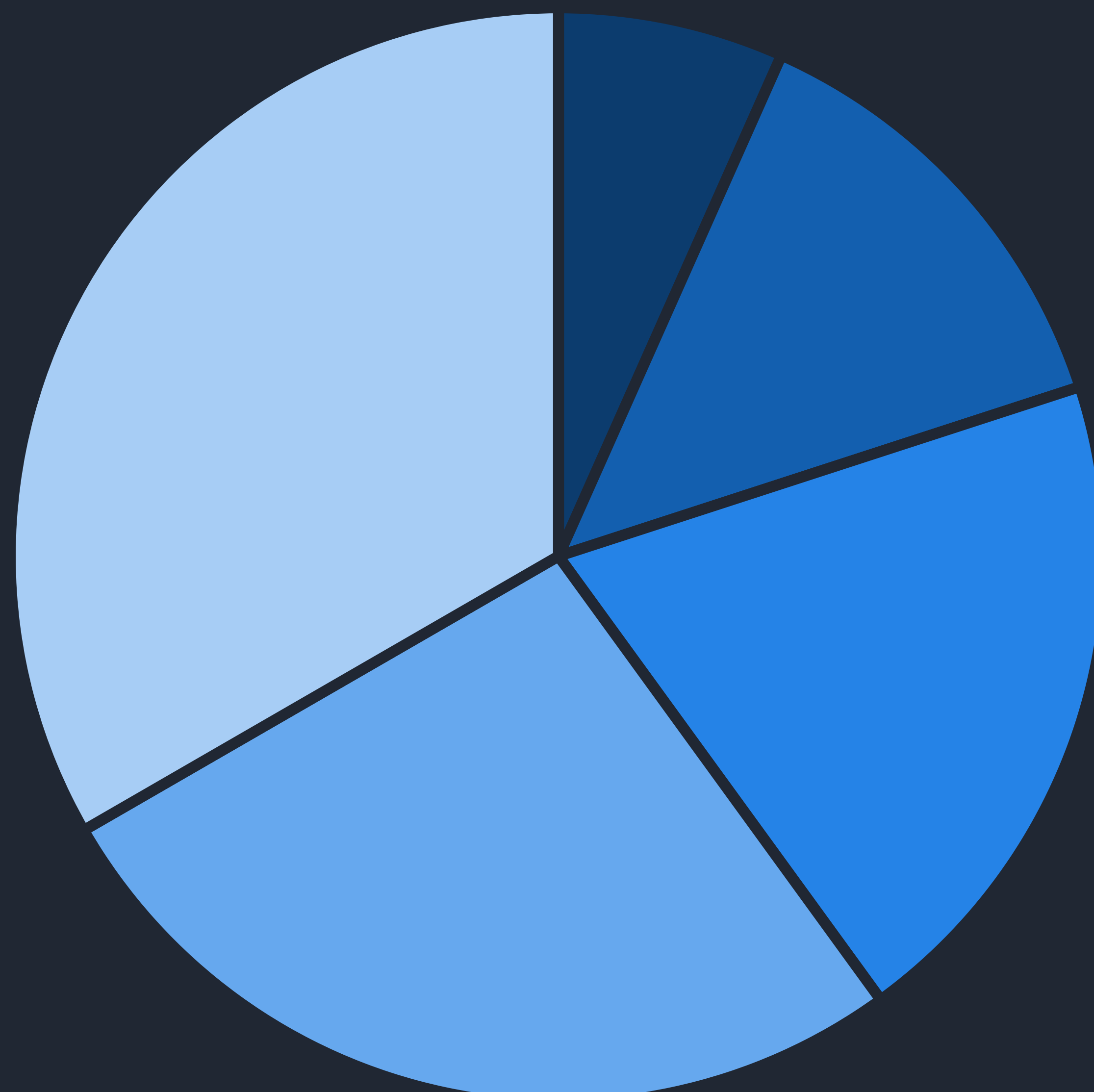
Adjusting model weights to minimize loss and improve accuracy.

5

Validation

Regularly assessing model performance on unseen data after each epoch to prevent overfitting.

Training Results: A Stable Learning Curve



The line chart illustrates the performance of the model during training, showing:

- **Gradual Decrease in Training Loss**: Indicating effective learning and convergence over epochs.
- **Consistent Increase in Validation Accuracy**: Demonstrating the model's ability to generalize well to new, unseen data.
- **Robust Generalization**: The model successfully learns transferable patterns, performing well beyond its training set.
- **Stable Learning Curves**: Confirming a healthy training process without significant fluctuations, ensuring reliable performance.

Model Testing: Real-World Application

Example 1

Input: “This movie was absolutely amazing!”

Output: Positive

Example 2

Input: “The film was boring and too long.”

Output: Negative

These examples demonstrate the model's capability to correctly recognize and classify the emotional tone of movie reviews.

Challenges Faced During Development

GPU Limitations on Colab

Navigating the constraints of available computational resources for intensive deep learning tasks.

Handling Long Sequences

Managing BERT's maximum input length for extensive movie reviews, requiring careful truncation or segmentation.

Preprocessing Large Texts

Efficiently cleaning and preparing vast amounts of text data for model input.

Avoiding Overfitting

Implementing strategies during fine-tuning to ensure the model generalizes well rather than just memorizing training data.

Conclusion & Future Work

- ## Transformers Revolutionized NLP

  Through the innovative use of self-attention mechanisms, setting new standards for language understanding.

- ## BERT's Strong Performance

  Delivers robust and accurate results on sentiment analysis tasks, validating its effectiveness.

- ## Fine-Tuning for Real-World Applications

  Enables the adaptation of powerful pre-trained models to specific, practical challenges.

- ## Modern NLP's Power

  This project vividly demonstrates the advanced capabilities of contemporary Natural Language Processing.

Future Work

To further enhance this project, we plan to:

- **Test Alternative Models**: Explore DistilBERT, RoBERTa, or other advanced Transformer architectures.
- **Utilize Larger Datasets**: Expand the scope with more extensive and diverse movie review datasets for broader applicability.
- **Explore Advanced Fine-tuning Techniques**: Implement more sophisticated training methodologies to push performance boundaries.