General AI/ML

Unit 1: Intro to AI, ML and DL

1.2.4

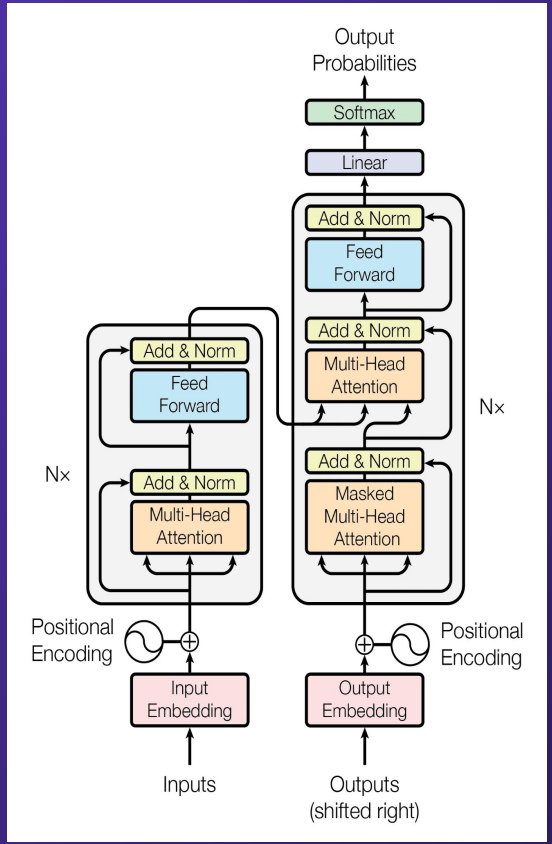
Introduction to Deep Learning

Introduction to Transformer Architecture

TIL-AI
TODAY I LEARNED AI

Introduction to Transformer Architecture

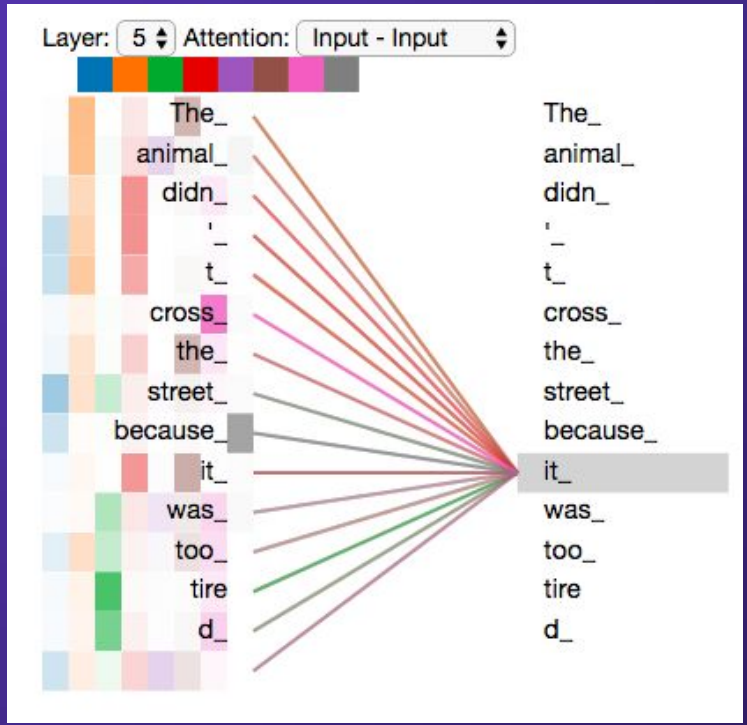
- A Novel Approach to Neural Machine Translation
- Introduced in the 2017 paper "Attention is All You Need"
- Replaced recurrent architectures (RNNs, LSTMs) in many NLP tasks
- Now a foundational architecture for various AI applications
- Key innovation: Self-attention mechanism for understanding relationships within sequences



TIL-AI
TODAY I LEARNED-AI

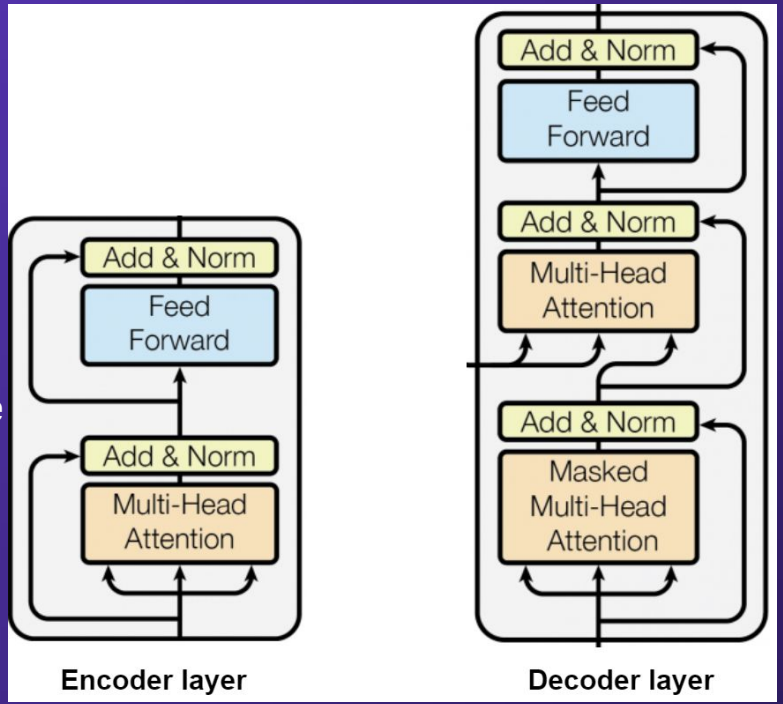
The Core of the Transformer: Attention

- The Attention Mechanism: A New Way to Focus
- Considers the relationships between tokens in an input sequence
- Calculates attention scores for each token, focusing on relevant parts
- Enables the model to capture long-range dependencies



Encoder & Decoder

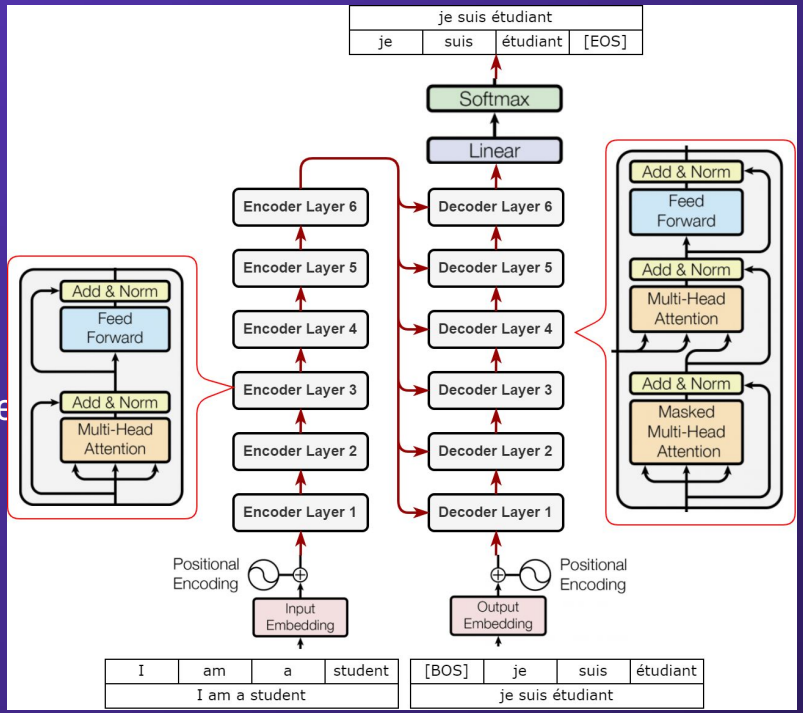
- The Encoder: Processing the Input Sequence
- Stacks of identical encoder layers
- Each layer uses self-attention to understand relationships within the input
- The Decoder: Generating the Output Sequence
- Uses a masked self-attention mechanism to prevent information leakage
- Incorporates attention from the encoder's output for context



TIL-AI
TODAY I LEARNED-AI

Encoder & Decoder

- The Encoder: Processing the Input Sequence
- Stacks of identical encoder layers
- Each layer uses self-attention to understand relationships within the input
- The Decoder: Generating the Output Sequence
- Uses a masked self-attention mechanism to prevent information leakage
- Incorporates attention from the encoder's output for context



Advantages of Transformer Architecture

- Highly Parallel Processing: Enables efficient training on large datasets
- Long-Range Dependency Modeling: Captures complex relationships within sequences
- Superior Performance: Achieved state-of-the-art results on various NLP tasks
- Versatile Applications: Used in machine translation, text summarization, question answering, and more

The Transformer's Dominance in NLP

- State-of-the-art performance in machine translation
- Advanced text summarization and generation
- Enhanced question-answering systems
- More sophisticated language understanding across various tasks

Transformers Beyond NLP

- Computer Vision: Efficient image classification and object detection
- Code Generation: Creating code from natural language descriptions
- Time-Series Analysis: Forecasting and anomaly detection
- Multimodal Applications: Combining language, images, and other modalities