NLP/ASR

Unit 5: Advanced Topics in ASR/NLP

5.1.2

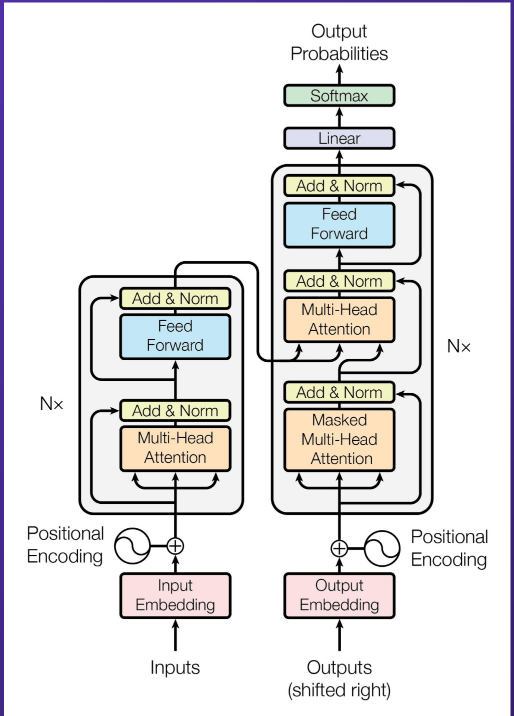
Attention Mechanisms and Transformers

Transformer Architecture & NLP

TIL-AI
TODAY I LEARNED-AI

Introduction to Transformer

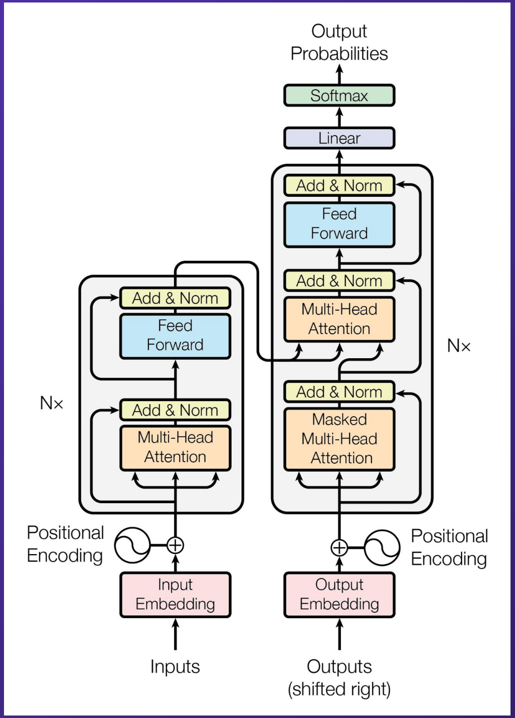
- Introduced in the 2017 paper "Attention is All You Need"
- Key innovation: the self-attention mechanism
- Self-attention allows the model to weigh the importance of different words in a sentence simultaneously
- No need for sequential processing, enabling massive parallelization



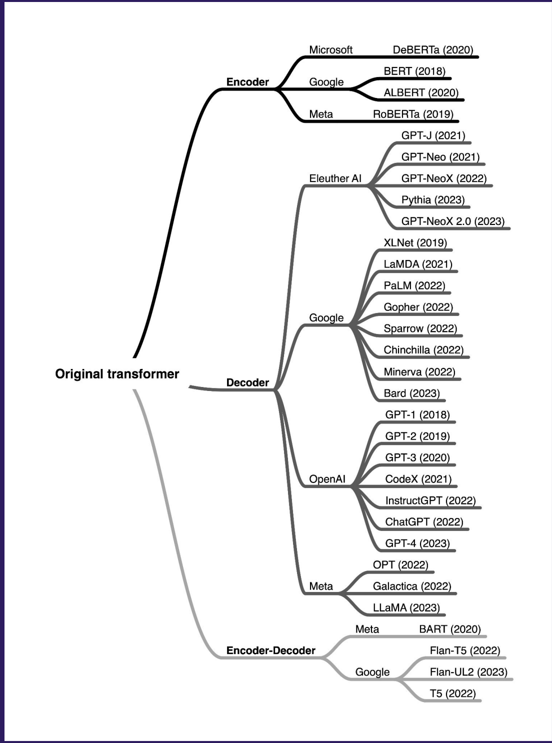
TIL-AI
TODAY I LEARNED-AI

Introduction to Transformer

- Introduced in the 2017 paper "Attention is All You Need"
- Key innovation: the self-attention mechanism
- Self-attention allows the model to weigh the importance of different words in a sentence simultaneously
- No need for sequential processing, enabling massive parallelization



Transformer Architecture Types



TIL-AI
TODAY I LEARNED-AI

Transformer Architecture Types

Encoder-Only

- Processes input sequences only, no output generation
- Focuses on understanding and encoding the input text into a fixed length vector representation or sequence of vectors
- Used for: Text Classification, NER, Sentiment Analysis
- Examples: BERT, RoBERTa, ALBERT

Transformer Architecture Types

Encoder-Decoder

- Processes both input and output sequences
- Encoder - creates a deep understanding of the input text
- Decoder - uses the encoder's understanding to generate the output text
- Used for - Translation, Summarization, Image Captioning
- Examples: T5, BART

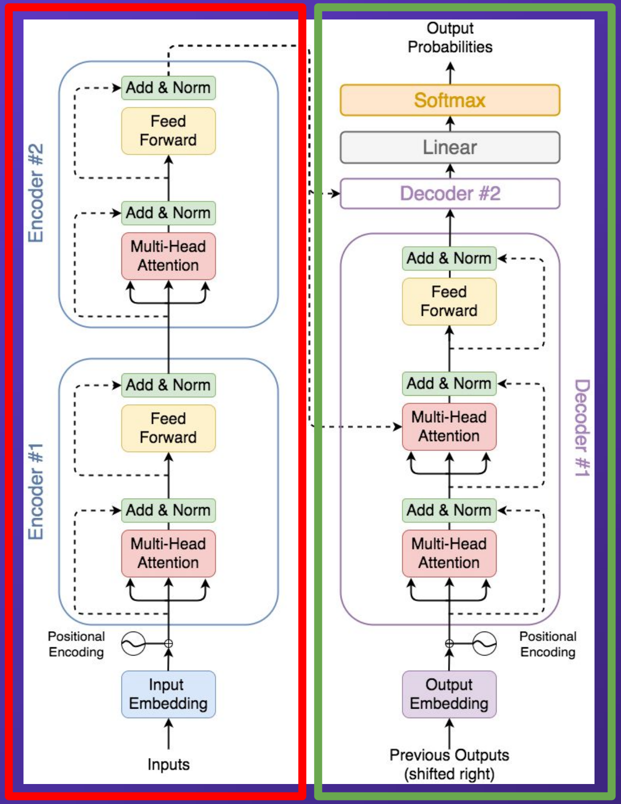
Transformer Architecture Types

Decoder-only

- Processes no input sequence, focuses solely on output generation
- Generates text based on a starting prompt or limited input
- Decoder blocks produce outputs one item at a time
- Used for: Text generation
- Examples: GPT-3, GPT-4, PaLM

Encoder-Decoder Transformer

Encoder

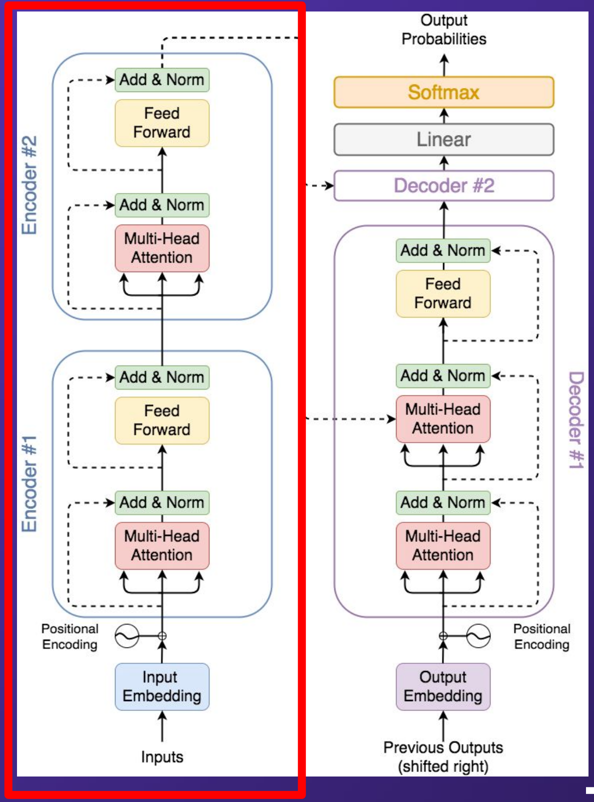


Decoder

TIL-AI
TODAY I LEARNED-AI

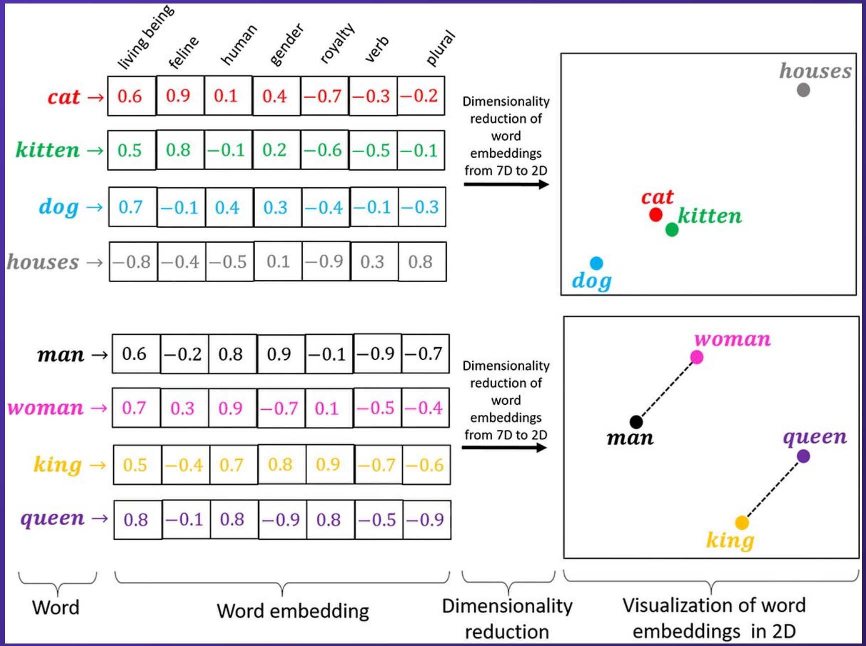
Encoder Blocks

- The encoder consists of:
  - Input Embeddings
  - Positional Encoding
  - Attention blocks that include:
    - Multi-head self-attention layer
    - Feed-forward network for further processing
    - Add & normalize layers for stabilizing training



Input Embeddings

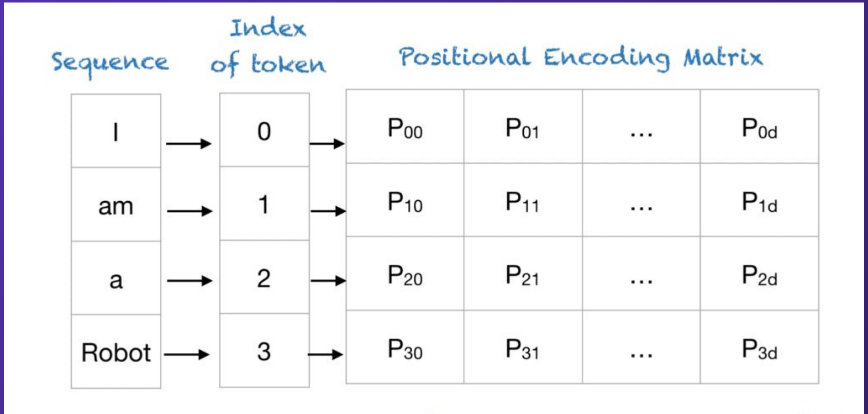
- Before text can be processed, we need numerical representations
- Input words are converted into dense vectors called embeddings
- Embeddings capture semantic relationships between words (words with similar meanings have similar embeddings)



TIL-AI
TODAY I LEARNED-AI

Positional Encoding

- Unlike RNNs, transformers don't process words sequentially
- Positional encoding adds information about the order of words within the sentence
- This allows the model to understand the structure and flow of the text

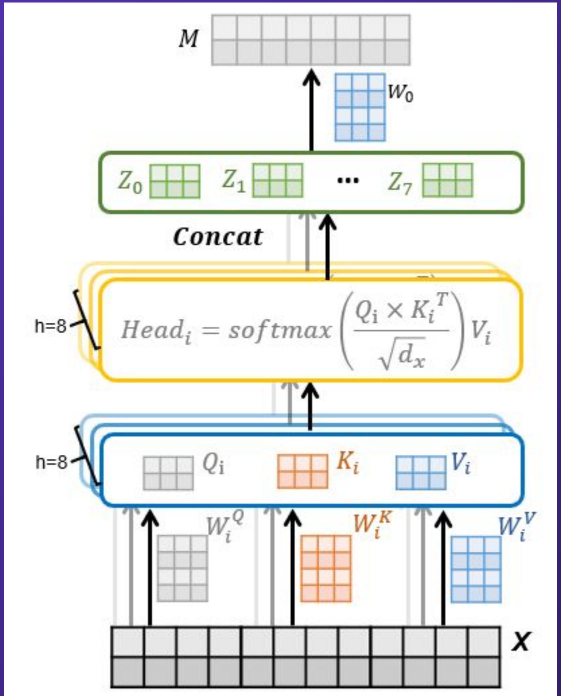


Positional Encoding Matrix for the sequence 'I am a robot'

TIL-AI
TODAY I LEARNED-AI

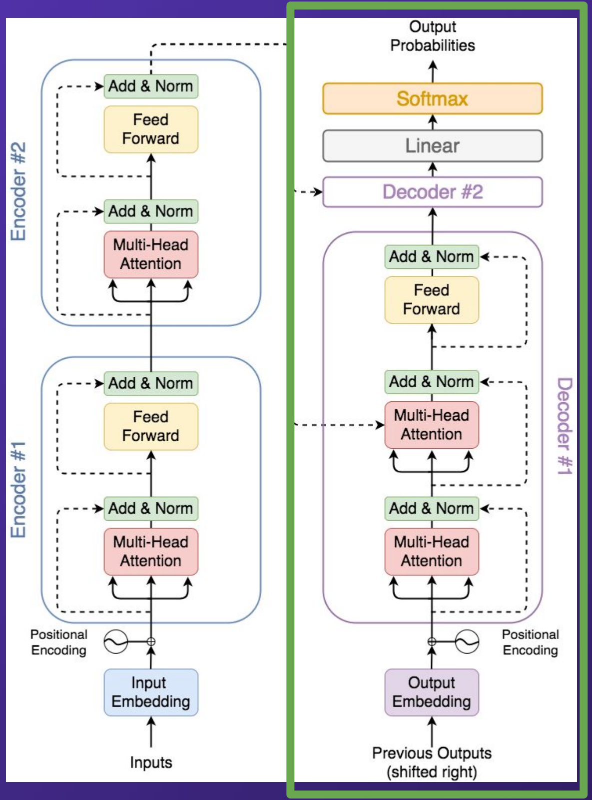
Multi-Head Self-Attention

- The heart of the transformer
- Each word in the sentence attends to all other words, determining relationships and importance
- Multi-head means multiple sets of attention calculations run in parallel, each focusing on different aspects of the input



Decoder Blocks

- The decoder has a similar structure to the encoder but with an extra element
- Masked multi-head attention: Ensures the model only looks at past words when generating text



TIL-AI
TODAY I LEARNED-AI

Output Generation

- The final decoder block produces an output representation
- A linear layer maps this representation to a vocabulary of possible words
- A softmax function is applied to generate probabilities for each word, selecting the most likely next word in the sequence

Transformers in NLP

- Machine translation: Transformers have achieved state-of-the-art results in machine translation tasks
- Text generation: Transformers are used in text generation tasks such as chatbots, language models, and text summarization
- Question answering: Transformers are used in question answering tasks such as SQuAD and TriviaQA

TIL-AI
TODAY I LEARNED-AI

Significance in NLP

- Parallelization: Transformers can be parallelized, making them faster to train and evaluate on text data
- Scalability: Transformers can handle long input sequences, making them more suitable for long-range dependencies in text
- Flexibility: Transformers can be used for a wide range of NLP tasks, making them a versatile architecture

TIL-AI
TODAY I LEARNED-AI

Challenges and Limitations

- Despite their advantages, transformers require substantial computational resources
- They can also be prone to biases present in the training data

Conclusion

- Attention mechanisms are pivotal in modern NLP
- They offer a more nuanced understanding of language and data processing
- Continuous advancements are making these models more versatile and powerful