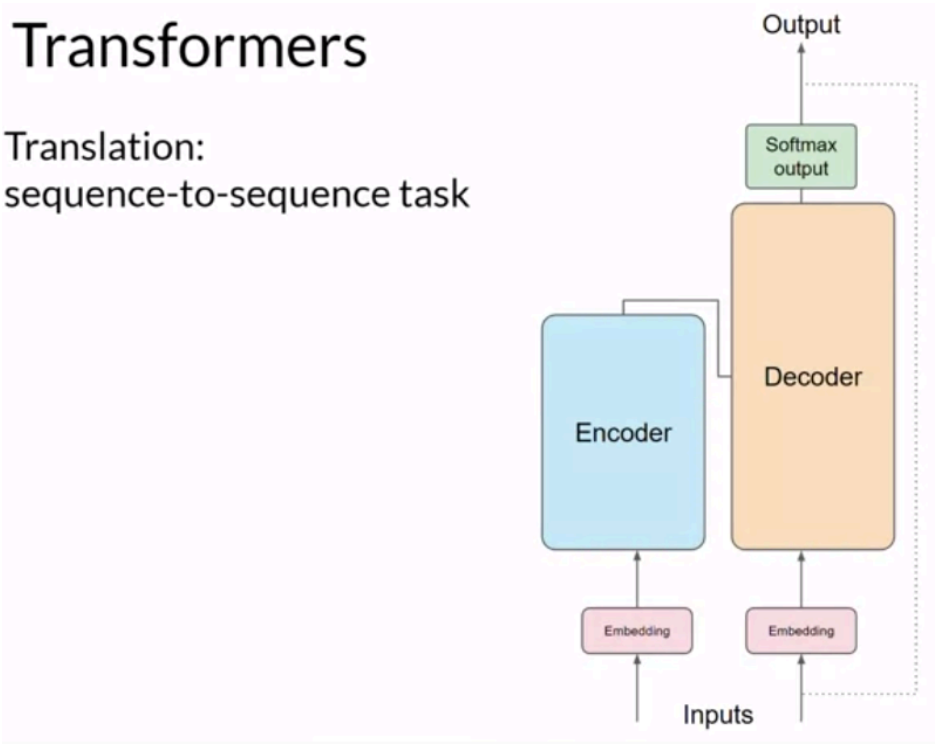
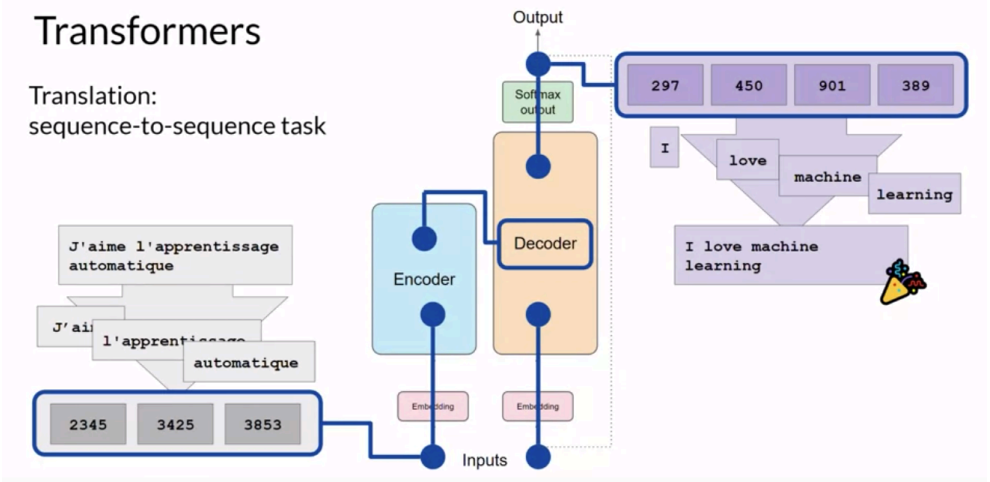
Generating text with transformers

- Example task: translating French phrases into English using a transformer model.

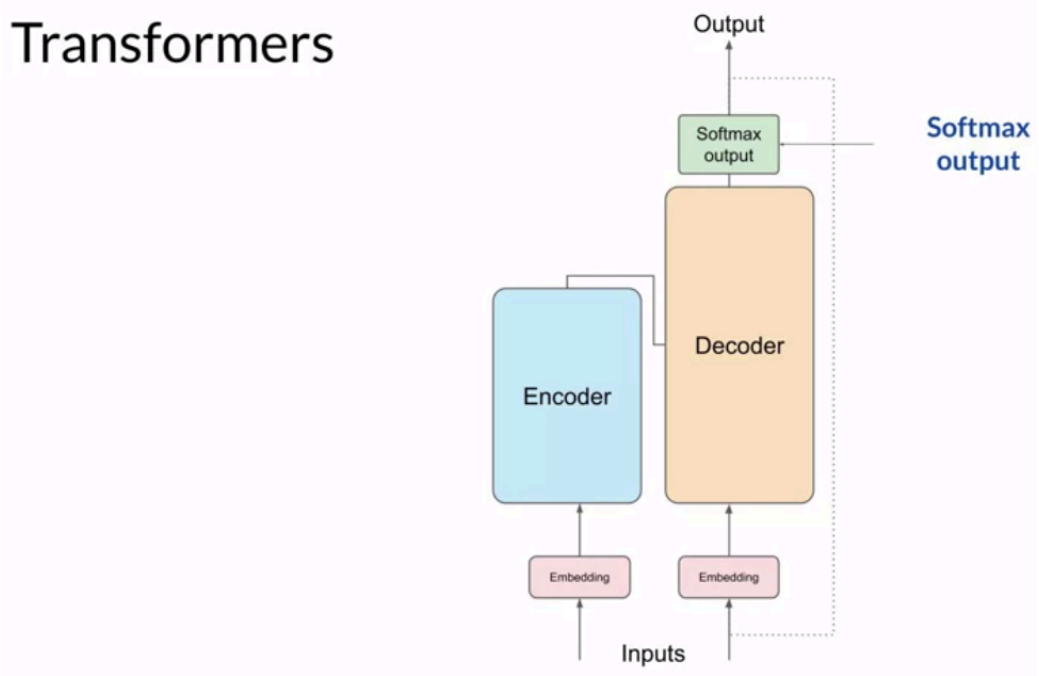


- Tokenization: Input words tokenized using the same tokenizer used in training.



- Tokens added to the encoder, passed through embedding and multi-headed attention layers.
- Outputs from attention layers fed through feed-forward network.

- Encoder output represents deep structure and meaning of input sequence.
- Encoder output inserted into decoder to influence its self-attention.
- Decoder initiates prediction loop with start-of-sequence token.
- Contextual understanding from the encoder guides decoders in predicting the next token.
- Output of decoder's self-attention layers passed through feed-forward network and softmax layer.



- Loop continues until the end-of-sequence token is predicted.
- Final sequence detokenized into words to obtain output.
- Variations in Transformer Architecture:
- Transformer consists of encoder and decoder components.
- Encoder encodes input sequences, while decoder generates new tokens based on the encoder's understanding.
- Encoder-only models used for tasks like sentiment analysis, with input and output sequences of equal length.
- Encoder-decoder models effective for sequence-to-sequence tasks with varying input and output lengths.
- Decoder-only models widely used, capable of generalizing to most tasks.
- Types of Transformer Models:
    - Encoder-only models like BERT.
    - Encoder-decoder models like BART and T5.
    - Decoder-only models such as GPT, BLOOM, Jurassic, LLaMA, etc.

Transformers