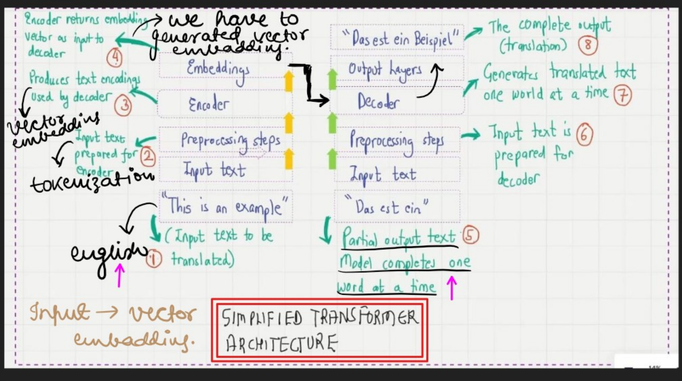
Lecture-4. Basic Intro To Transformer

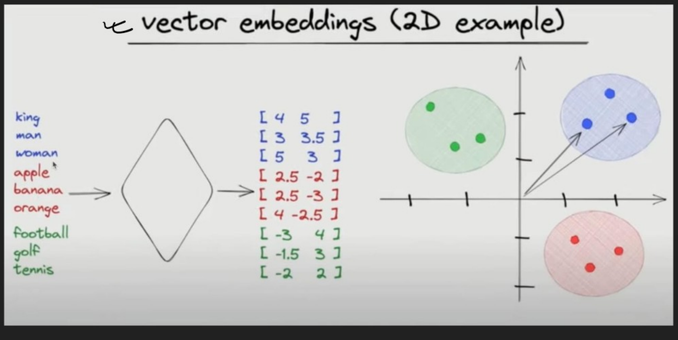
1. Most modern LLMs rely on the the transfer architecture → deep neural network architecture introduced in 2017 paper.
   → Attention is all you need.

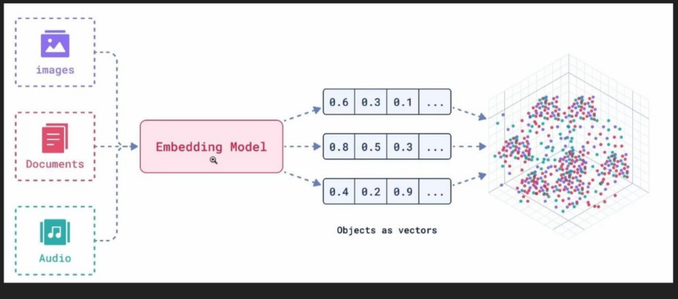
Original Transformer = Developed for machine translation.

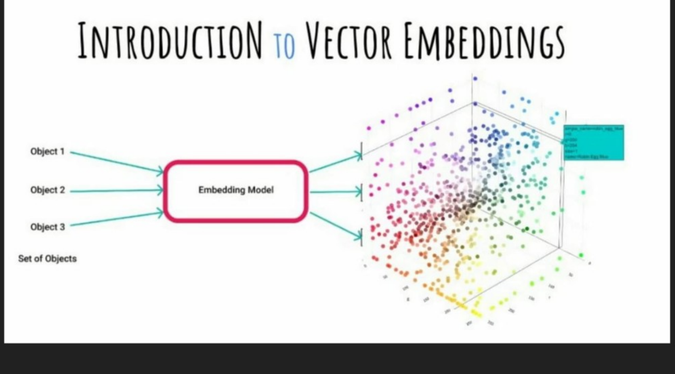
Translating English text to German and French

2.









3. Transformer architecture consist of:

Encoder → encoder input text into vector

Decoder → Generate output text from encoded vectors.

A note on self-attention mechanism:

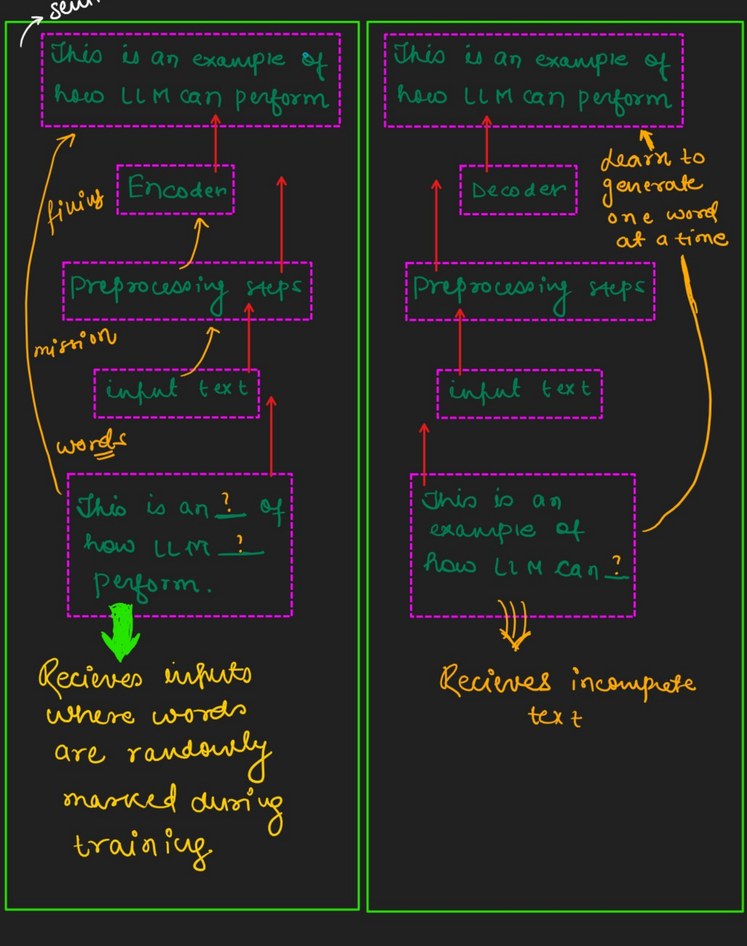
- Key part of transformer: Allow model to weigh importance of different words/tokens relative to each other.
- Enables models to capture long range dependencies.
- we will look at this in detail later.

4. Later variations of transformer architecture

BERT → "Bidirectional encoder representation from transformers". → predicts hidden words in a given sentence

GPT models → "Generative pretrained Transformers". → Generates new words

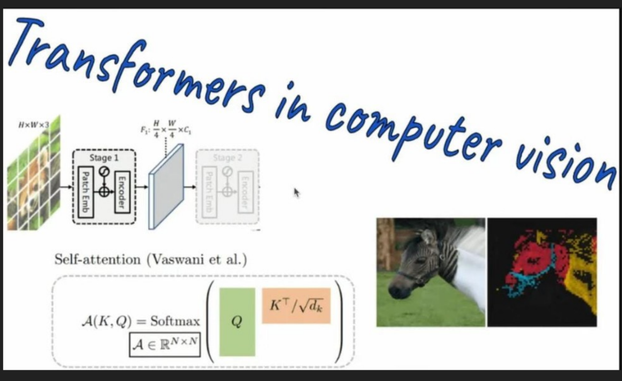
Differences



5. Transformers VS LLMs

→ Not all transformers are LLMs.

→ Transformers can also be used for computer vision.



Music

→ Not all LLMs are transformers.

→ LLMs can be based on recurrent or convolutional architecture as well.