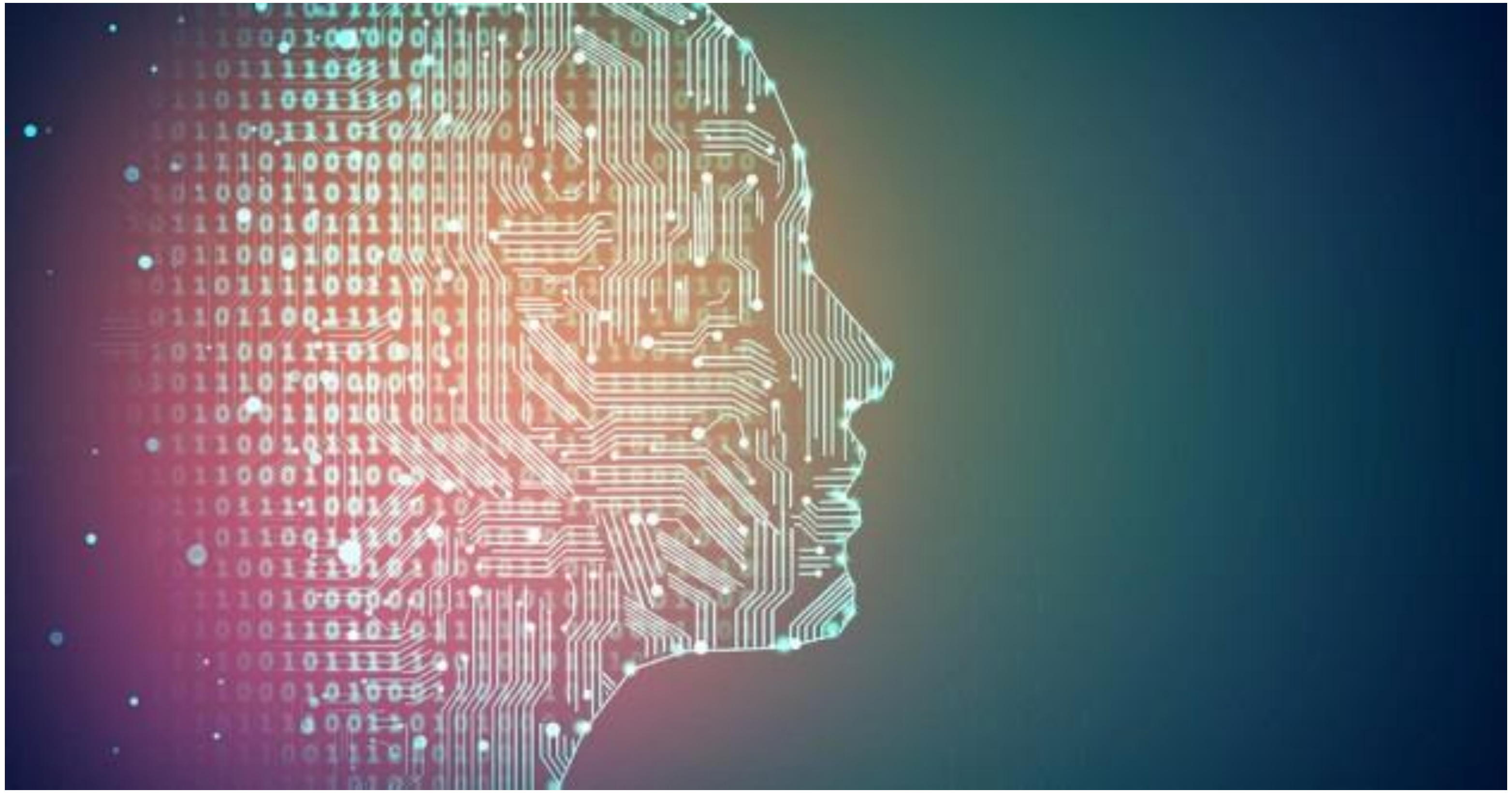
Generative AI: Transformers: Architecture PART II: Encoder

Thuan L Nguyen, PhD

2: Generative AI: LLM: Transformers: Architecture



AI Deep learning (Source: mindovermachines.com)

3: Generative AI: LLM: Transformers: Architecture

1. Generative AI: Transformers: Architecture: Overview
2. Generative AI: Transformers: Architecture: Encoder-Decoder Structure
3. Generative AI: Transformers: Architecture: Revolutionary Core Features
4. Generative AI: Transformers: Architecture: Input Embedding & Positional Encoding
5. Generative AI: Transformers: Architecture: Attention & Multi-Header Attention
6. Generative AI: Transformers: Architecture: Encoder & Encoding Stack of Layers
7. Generative AI: Transformers: Architecture: Decoder & Decoding Stack of Layers

4: Generative AI: LLM: Transformers: Architecture

Artificial Intelligence: Generative AI

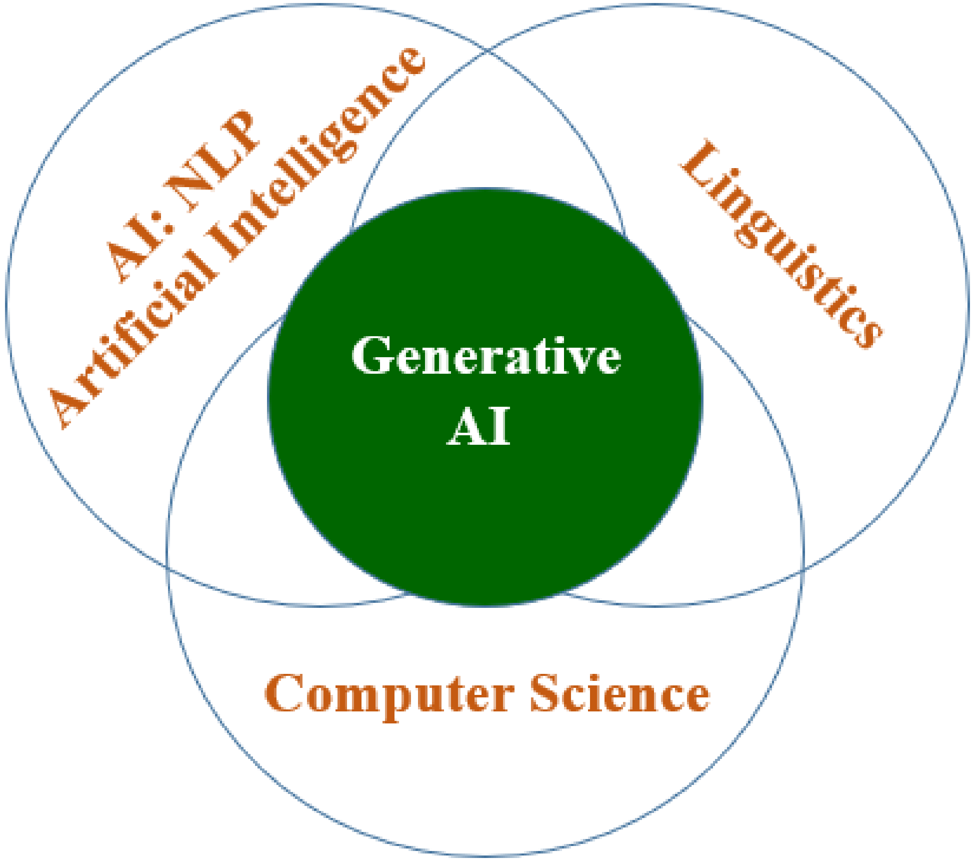
What is It?

Generative AI: A category of artificial intelligence focused on using AI deep learning models to generate new contents, including text, images, audio, video, and more. The contents are novel but look realistic and may be indistinguishable from human-created ones.

5: Generative AI: LLM: Transformers: Architecture

Artificial Intelligence: Generative AI: LLM

Foundational Sciences & Technologies



Generative AI is based on the NLP technologies such as Natural Language Understanding (NLU) and Conversational AI (AI Dialogues) - Those among the most challenging tasks AI needs to solve.

6: Generative AI: LLM: Transformers: Architecture

Artificial Intelligence: Generative AI: LLMs

Large Language Models

Large Language Models (LLMs) are revolutionary AI **Deep Learning neural networks** that excel in natural language understanding (NLU) and content generation.

- “LARGE" in LLMs refers to the vast scale of data and parameters used to train them, allowing LLMs to develop a comprehensive understanding of language.
- Being particularly transformer-based models trained on massive text datasets using deep learning techniques,
- Able to learn complex language patterns, capture nuances like grammar and tone, and generate coherent and contextually relevant text

7: Generative AI: LLM: Transformers: Architecture

Artificial Intelligence: Generative AI: LLMs

Large Language Models: Transformers

***Photo Source**: **Prompt** by Thuan L Nguyen – Generated by **Generative AI Text-To-Images**: Google-DeepMind ImageFX*

8: Generative AI: LLM: Transformers: Architecture

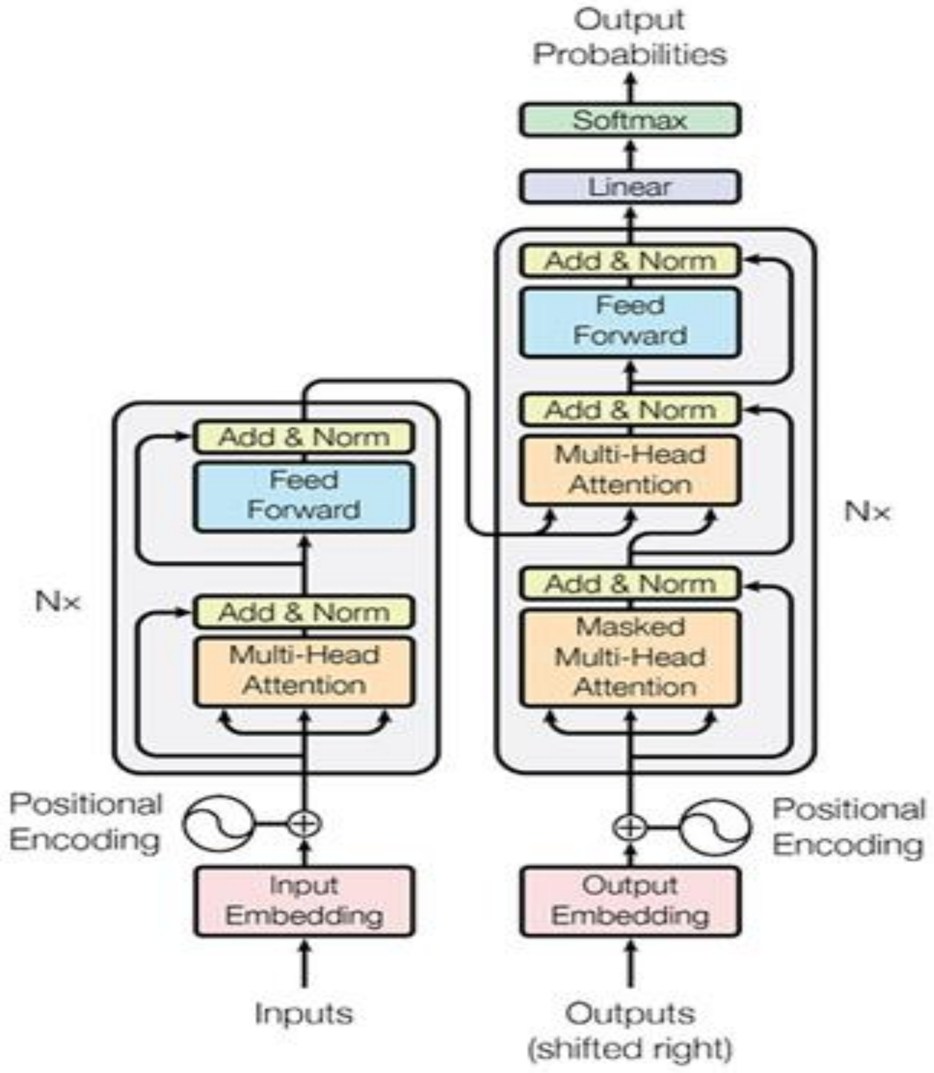
“**Attention Is All You Need!**”

__Photo Source__: __Prompt__ by Thuan L Nguyen – Generated by __Generative AI Text-To-Images__: Google-DeepMind ImageFX

9: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: Architecture

Encoder-Decoder



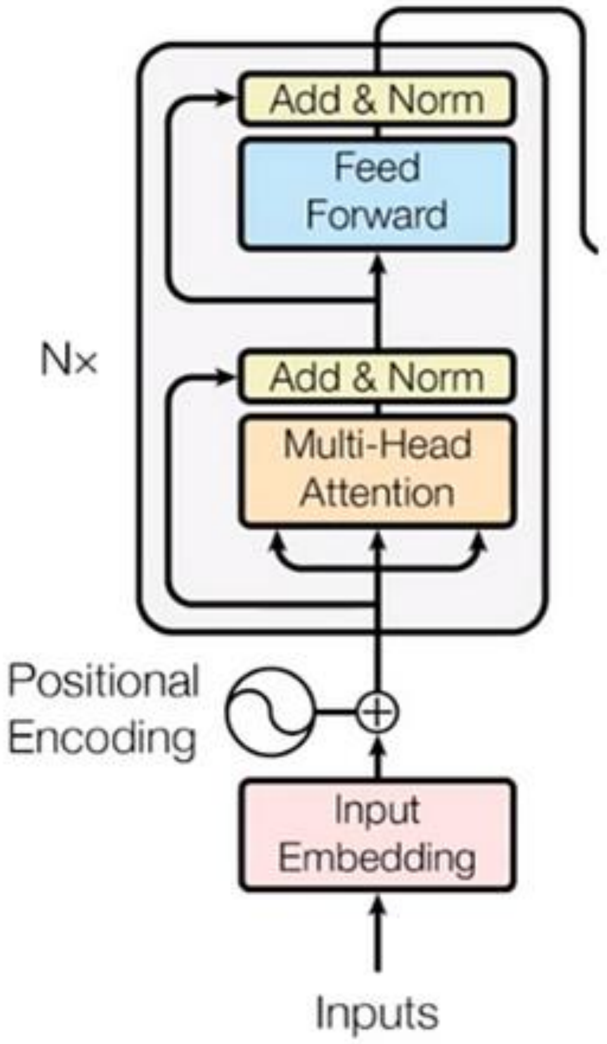
The Transformer - model architecture.

source: arXiv:1706.03762

10: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: Architecture

Encoder



source: arXiv:1706.03762

11: Generative AI: LLM: Transformers: Architecture

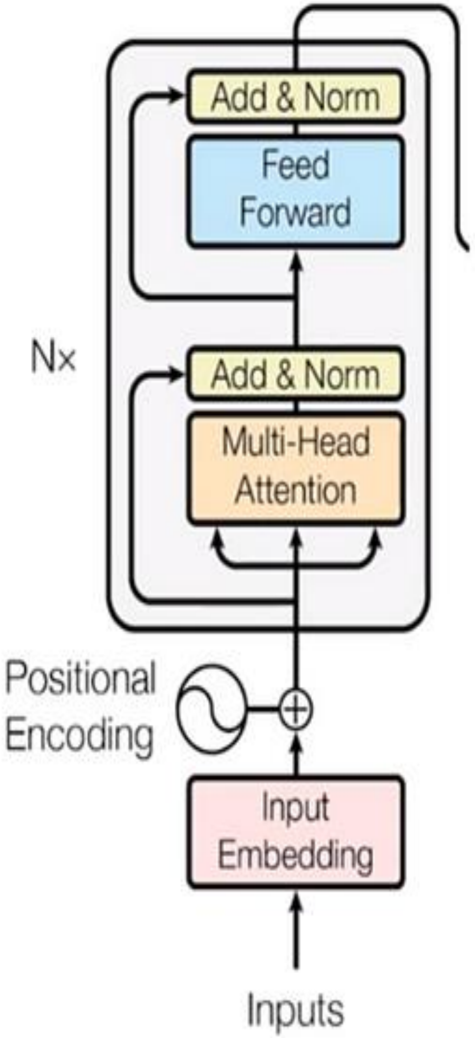
LLM: Transformers: Architecture

Encoder

Main Layers of the Transformer: Encoder

- **Input Embeddings**: The input words are converted into word embeddings (dense vector representations).
- **Multi-Head Attention (Self-Attention)**: Words "attend" to each other within the input sequence.
- **Layer Normalization**: Help stabilization during training.
- **Feed-Forward Network**: Further refine the representations of words
- **Residual Connections**: Shortcut connections add the input to the output of each sub-layer, making training more manageable.

A Transformer has **stacked encoder layers** (usually 6 or more), with each layer **refining** the representations of words or contents from the previous.



source: arXiv:1706.03762

12: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

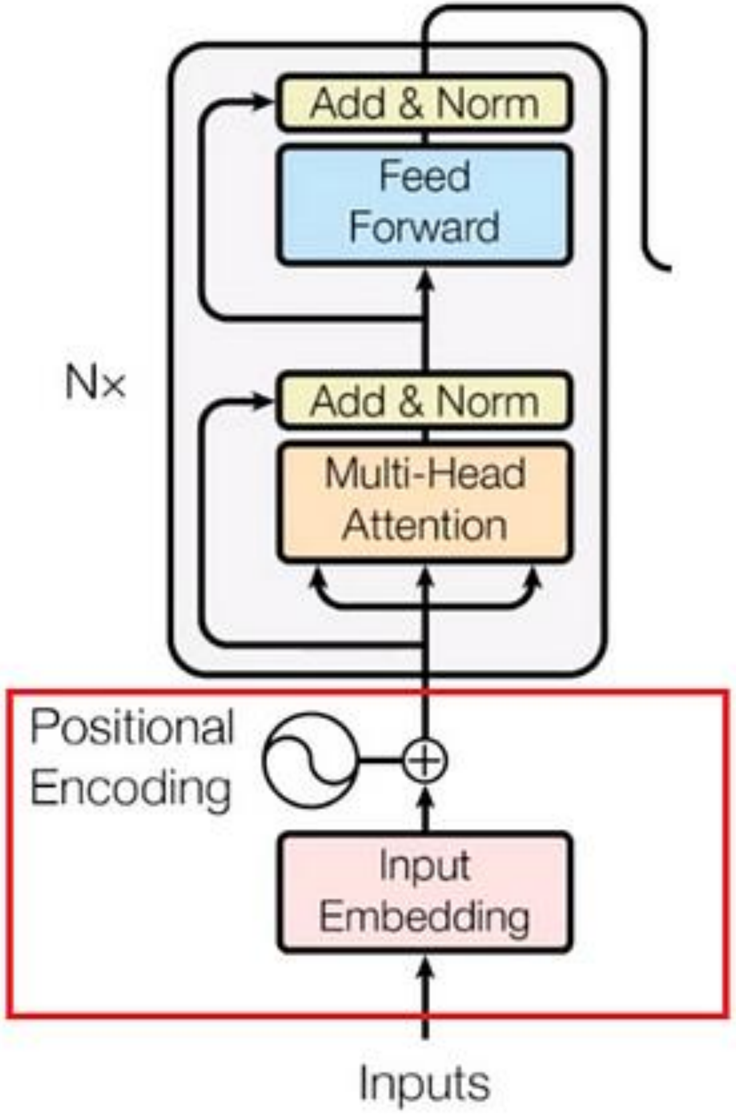
Encoder

Computers does not understand words. Computers can only works on numbers, vectors or matrices.

Words → Vectors: **Embedding Space**

Embedding Space:

- A dictionary where words of similar meanings are grouped together and are present close to each other.
- Every word, according to its meaning, is mapped and assigned with a particular value, i.e., a vector.



source: arXiv:1706.03762

13: Generative AI: LLM: Transformers: Architecture

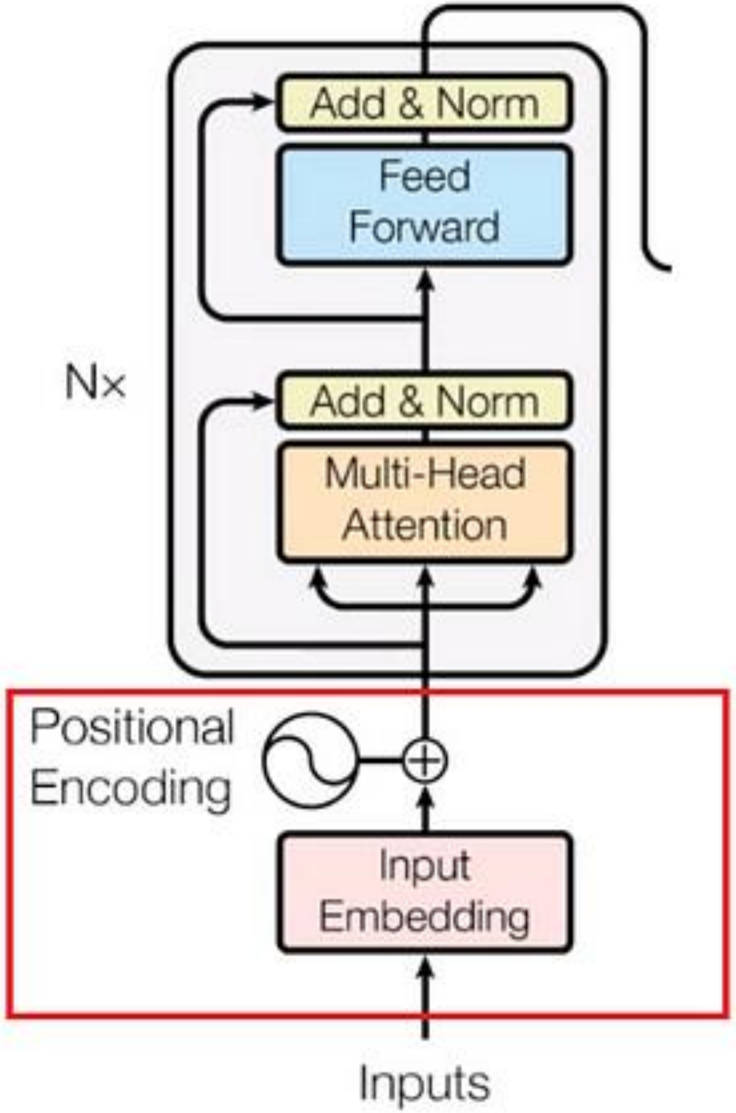
LLM: Transformers: How Does it Works?

Encoder

Every word in different sentences has different meanings.

Positional Encoding → **Positional Encoders**

Positional Encoder: A **vector** that gives the **context** of each word according to the **position** of the word in a sentence.



source: arXiv:1706.03762

14: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Encoder

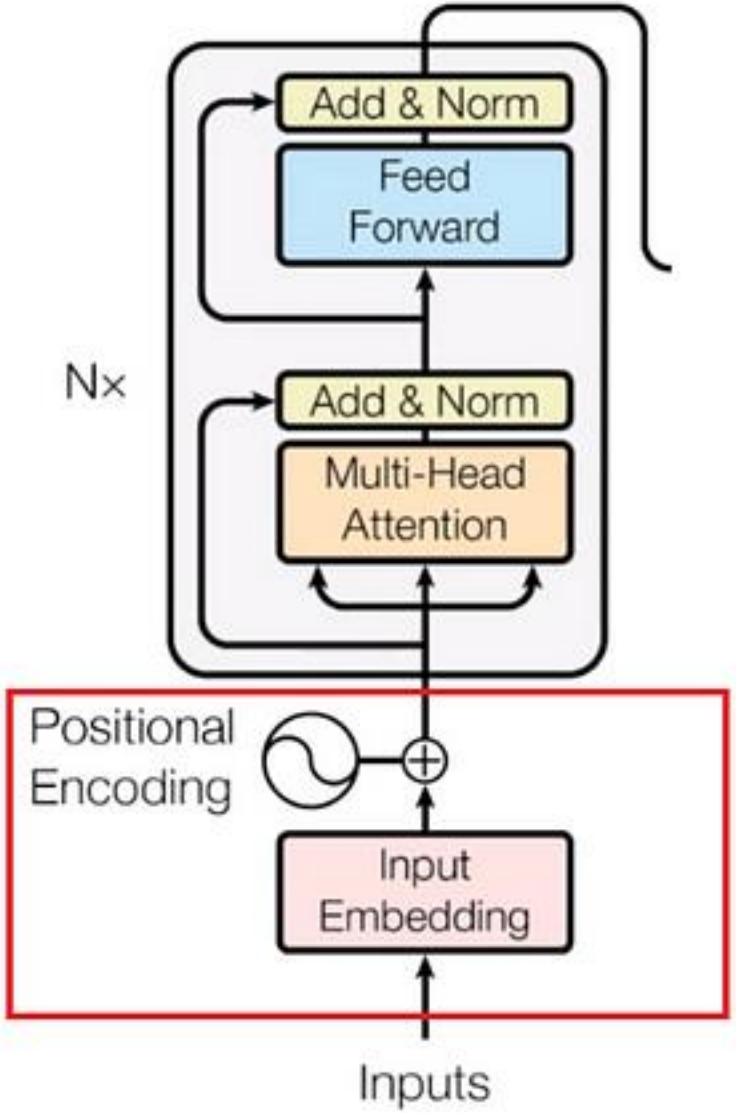
Each Word:

→ Word Embedding Vector (Input Embedding)

→ Positional Embedding Vector (Positional Embedding)

→**Final Vector**: **Context**

- Ready to be fed into the encoder block



source: arXiv:1706.03762

15: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Encoder

Multi-Head Attention (Self-Attention):

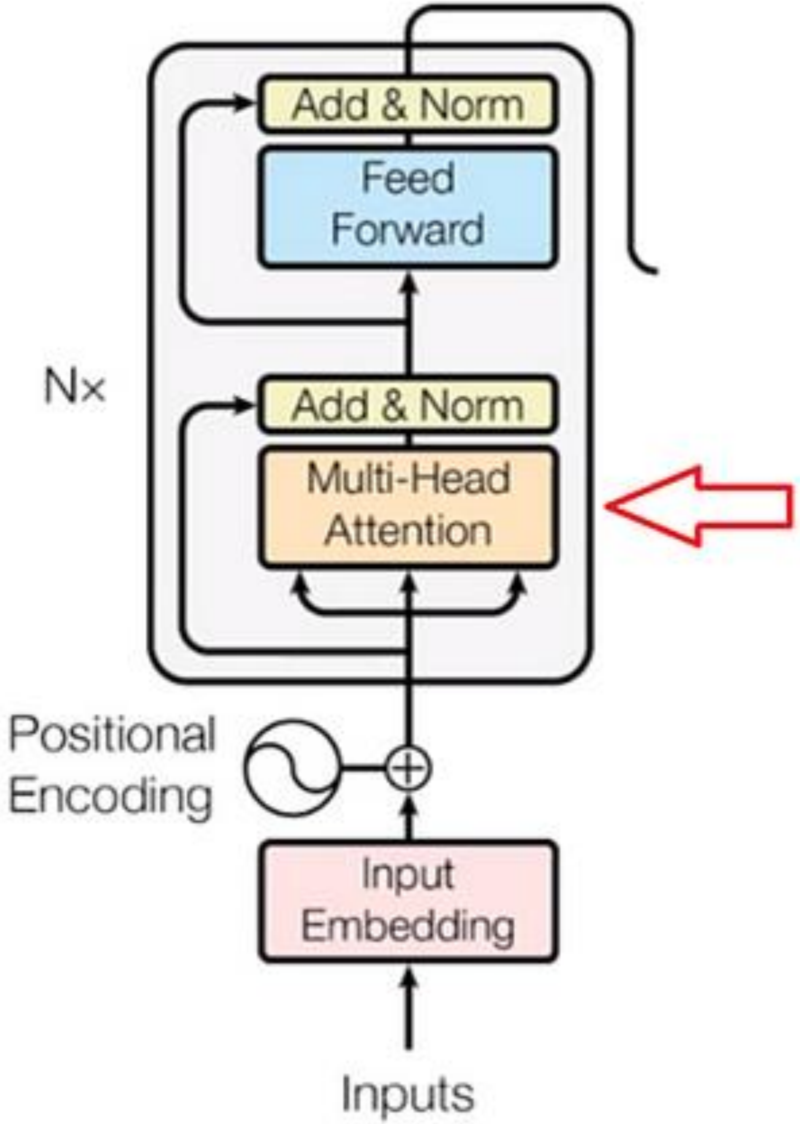
TO DO: Find out how relevant a particular word is w.r.t to other words in that sentence.

OUTPUT: **Attention vectors**

- Each vector represents the **relevance** of a word w.r.t other words.

Every word → An **attention vector**

- Capture the **contextual relationship** between the word and other words in that sentence.



source: arXiv:1706.03762

16: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Encoder

Multi-Head Attention (Self-Attention):

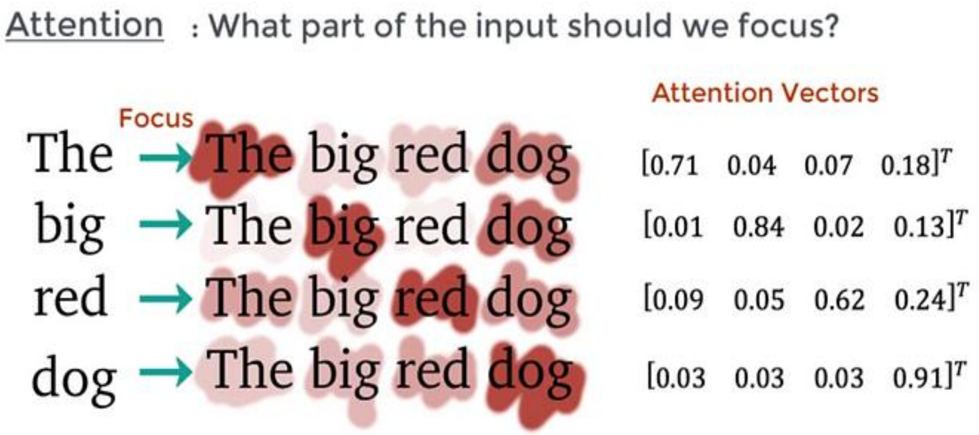
TO DO: Find out how relevant a particular word is w.r.t to other words in that sentence.

OUTPUT: **Attention vectors**

- Each vector represents the **relevance** of a word w.r.t other words.

Every word → An **attention vector**

- Capture the **contextual relationship** between the word and other words in that sentence.



Attention : What part of the input should we focus?

$[0.71 \quad 0.04 \quad 0.07 \quad 0.18]^T$

$[0.01 \quad 0.84 \quad 0.02 \quad 0.13]^T$

$[0.09 \quad 0.05 \quad 0.62 \quad 0.24]^T$

$[0.03 \quad 0.03 \quad 0.03 \quad 0.91]^T$

Source: https://www.youtube.com/watch?v=rPFkX5fJdRY

17: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Encoder

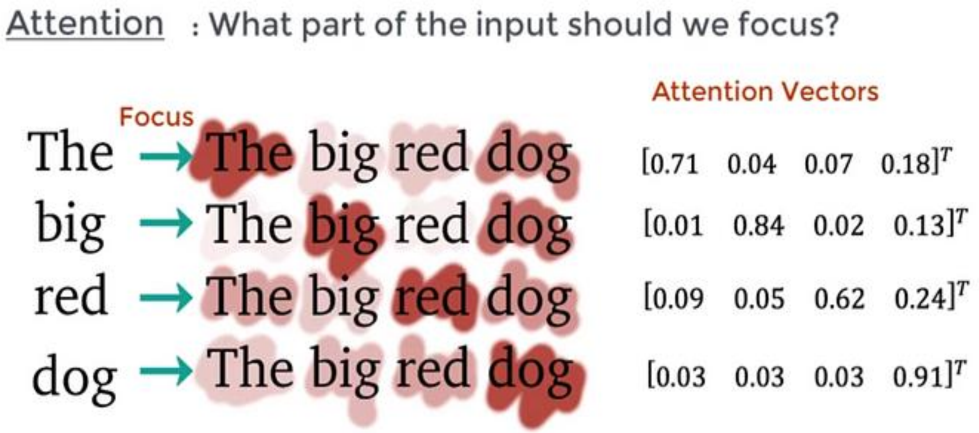
Multi-Head Attention (Self-Attention):

For **every word**:

- The attention on itself is much higher than that of other words in the sentence.

- **Multiple** attention vectors are calculated per word → Multi-Head Attention.

- **Final attention** vector: A weighted average of all attention vectors is calculated.



Attention : What part of the input should we focus?

$[0.71 \quad 0.04 \quad 0.07 \quad 0.18]^T$

$[0.01 \quad 0.84 \quad 0.02 \quad 0.13]^T$

$[0.09 \quad 0.05 \quad 0.62 \quad 0.24]^T$

$[0.03 \quad 0.03 \quad 0.03 \quad 0.91]^T$

Source: https://www.youtube.com/watch?v=rPFkX5fJdRY

18: Generative AI: LLM: Transformers: Architecture

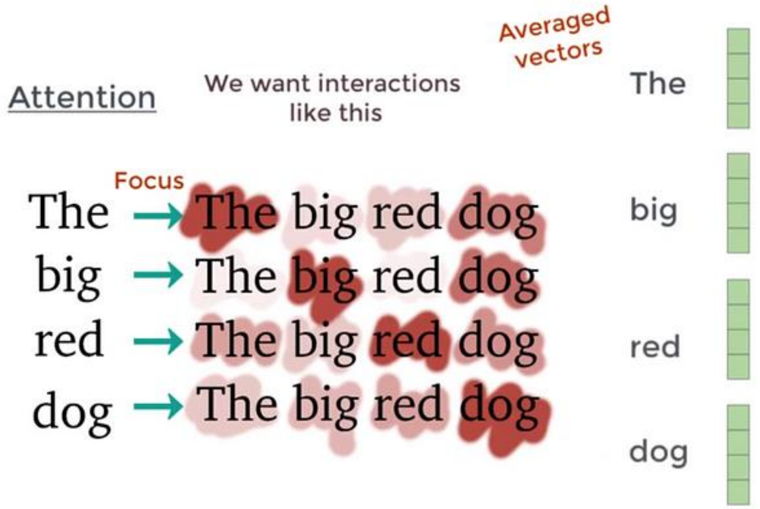
LLM: Transformers: How Does it Works?

Encoder

Multi-Head Attention (Self-Attention):

For **every word**:

- The attention on itself is much higher than that of other words in the sentence.
- **Multiple** attention vectors are calculated per word → Multi-Head Attention.
- **Final attention** vector: A weighted average of all attention vectors is calculated.

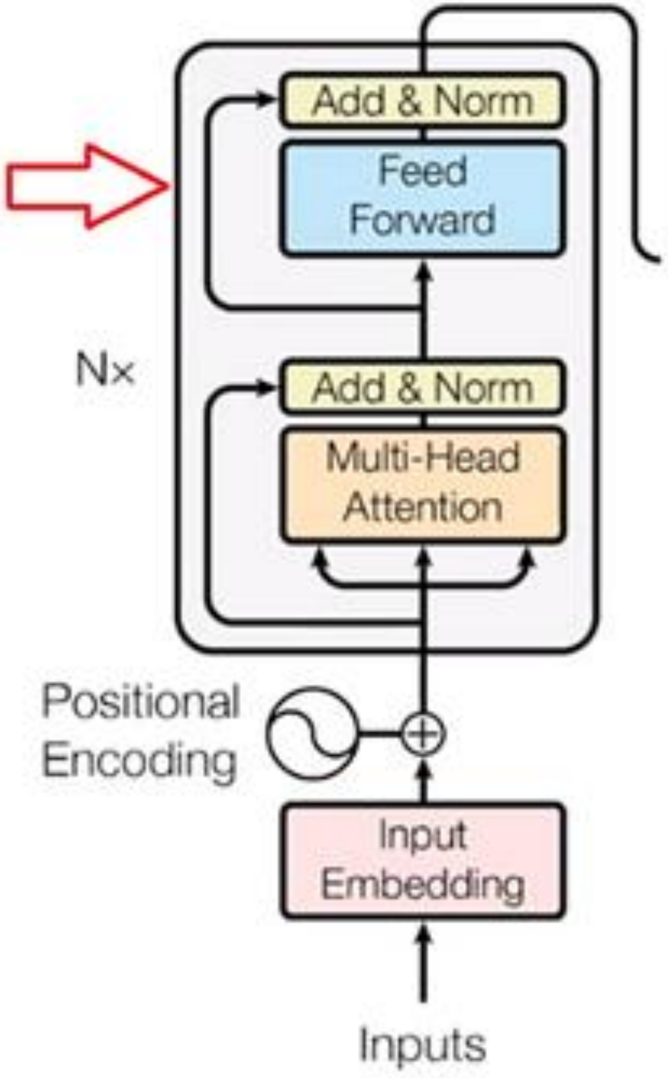


Source: https://www.youtube.com/watch?v=rPFkX5fJdRY

19: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Encoder



source: arXiv:1706.03762

Feed-Forward Neural Network

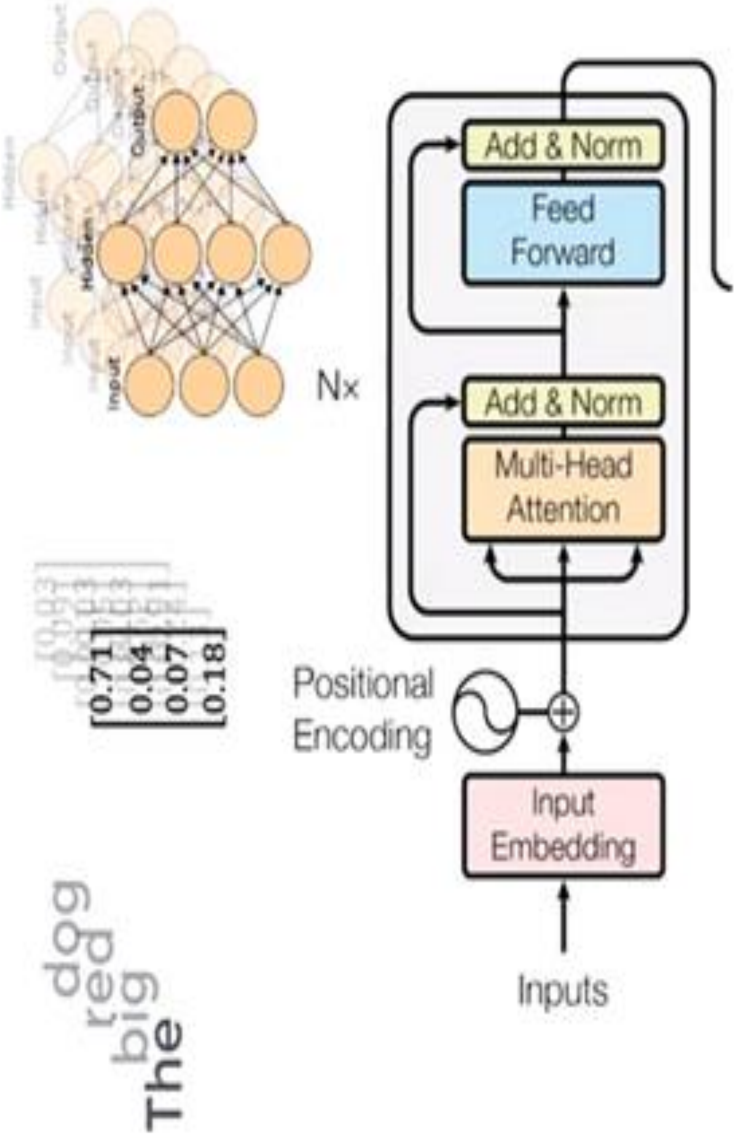
TO DO: Transform each attention vector into a form acceptable to the next encoder-decoder level.

OUTPUT: A data structure (Flattened Vector) acceptable to the next encoder-decoder level.

20: Generative AI: LLM: Transformers: Architecture

LLM: Transformers: How Does it Works?

Encoder



Feed-Forward Neural Network

TO DO: Transform each attention vector into a form acceptable to the next encoder-decoder level.

OUTPUT: A data structure (Flattened Vector) acceptable to the next encoder-decoder level.

source: arXiv:1706.03762

21: Generative AI: LLM: Transformers: Architecture

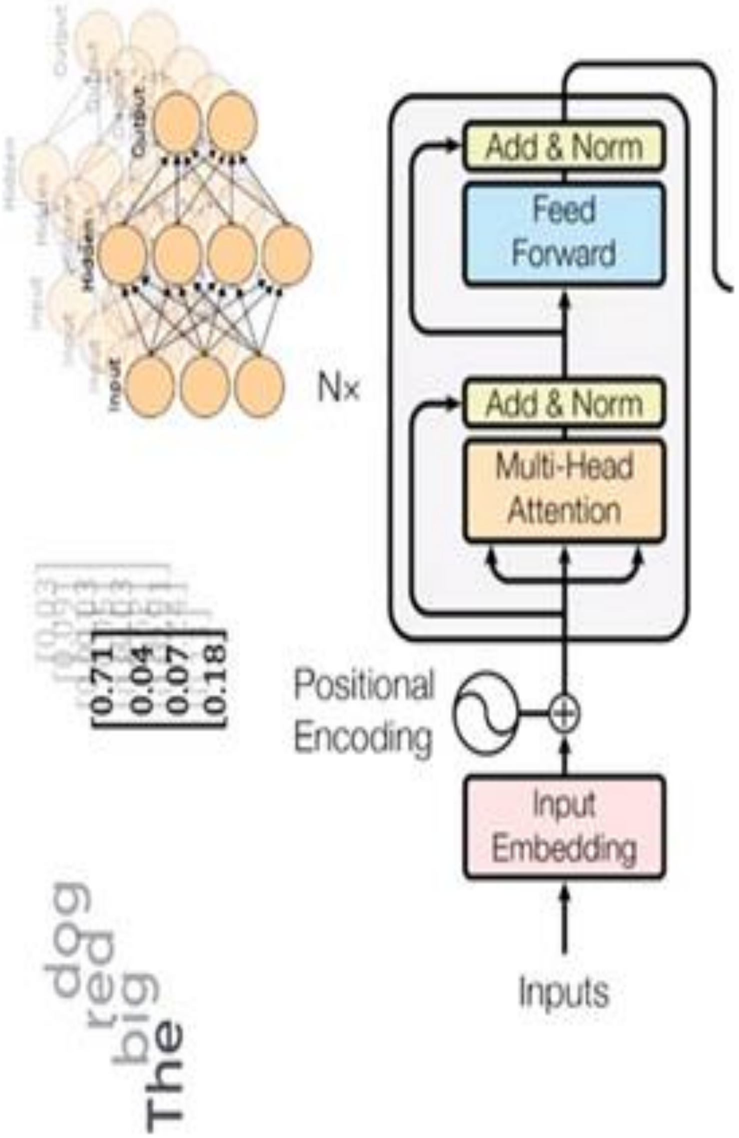
LLM: Transformers: How Does it Works?

Encoder

Feed-Forward Neural Network

Feed Forward Neural Network accepts **only one** attention vector at a time.

- Each attention vector is **independent** of each other.
- **Parallel computing** can be applied here, *and that makes all the difference.*



source: *arXiv:1706.03762*

22: Generative AI: LLM: Transformers: Architecture

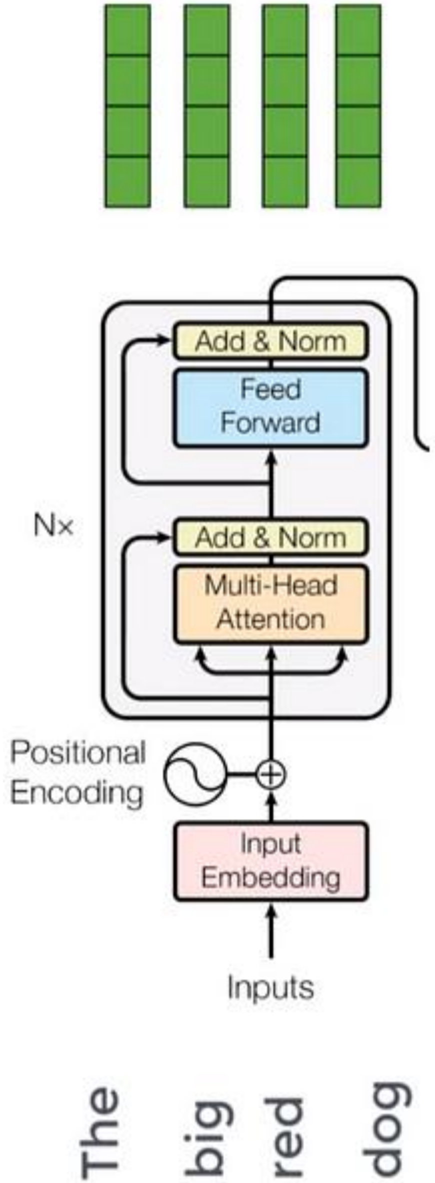
LLM: Transformers: How Does it Works?

Encoder

Encoder: Feed-Forward Neural Network

Feed Forward Neural Network accepts **only one** attention vector at a time. However, each attention vector is **independent** of each other.

- **Parallel computing** can be applied.
- Possible to pass **all** the words at the same time into the **encoder** block, and get the set of Encoded Vectors for every word **simultaneously**.



source: arXiv:1706.03762