

# Generative Pretrained Transformer

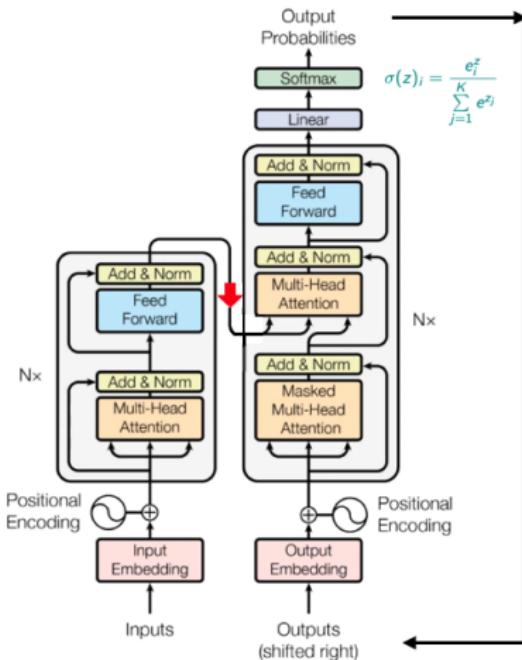
## GPT & LLM

Jaideep Ganguly

August 31, 2024

# Agenda

- ➊ Encode, Positional encoding
- ➋ Key, Query and Value → Attention Filter
- ➌ Multi Head Attention
- ➍ Information preservation, normalisation
- ➎ Decode, Masked Attention
- ➏ Potential Innovations
- ➐ Conclusion



Decoders are autoregressive models;  
They are trained to predict the next token  
after reading the preceding ones

# Generalised Pretrained Transformer (GPT) - 2017

---

## Attention Is All You Need

---

**Ashish Vaswani\***

Google Brain

avaswani@google.com

**Noam Shazeer\***

Google Brain

noam@google.com

**Niki Parmar\***

Google Research

nikip@google.com

**Jakob Uszkoreit\***

Google Research

usz@google.com

**Llion Jones\***

Google Research

llion@google.com

**Aidan N. Gomez\* †**

University of Toronto

aidan@cs.toronto.edu

**Lukasz Kaiser\***

Google Brain

lukaszkaiser@google.com

**Illia Polosukhin\* ‡**

illia.polosukhin@gmail.com

"Neural Machine Translation by Jointly Learning to Align and Translate"  
by Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 2014.  
Introduced concept of attention mechanism and laid the foundation for  
subsequent developments in NLP and DL, including the transformer  
architecture introduced in "Attention Is All You Need."

# Attention

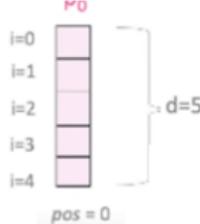
He went to the bank and learned of his empty account, after which he went to a river bank and cried.

- ① Attention mechanism has an infinite reference window
- ② In contrast, Recurring Neural Network (RNN) has a short reference window, Long Short Term Memory (LSTM) has a longer window. RNN does not work, LSTM has limited capability.

# One Hot Encoding, Word Embedding, Positional Encoding

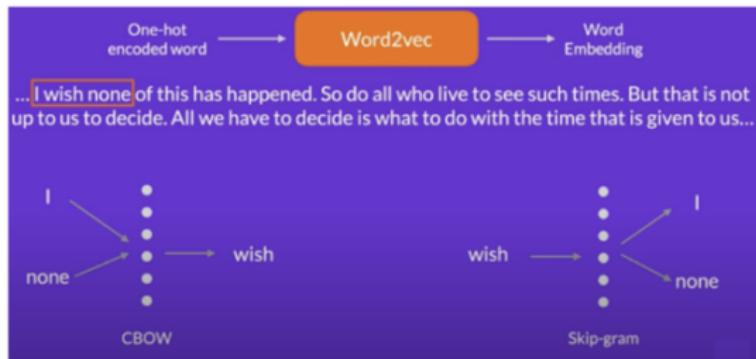
Linear & logistic regression models can deal with numerical data only, not text. In one-hot encoding, the string encoded variable is replaced with new variables of boolean type, e.g., a feature such as color can be encoded as:

	red	green	blue
1	1	0	0
0	0	1	0
0	0	0	1

$$PE_{(pos,2i)} = \sin\left(\frac{pos \cdot 2i}{10000}\right)$$

odd pos sin, even pos cos



$$\begin{bmatrix} 0.37 \\ 0.99 \\ 0.01 \\ 0.08 \end{bmatrix}$$

AJ's dog is a cutie → Position 2  
AJ looks like a dog → Position 5

$$\begin{bmatrix} 0.37 \\ 0.99 \\ 0.01 \\ 0.08 \end{bmatrix}$$

Embedding of "Dog"

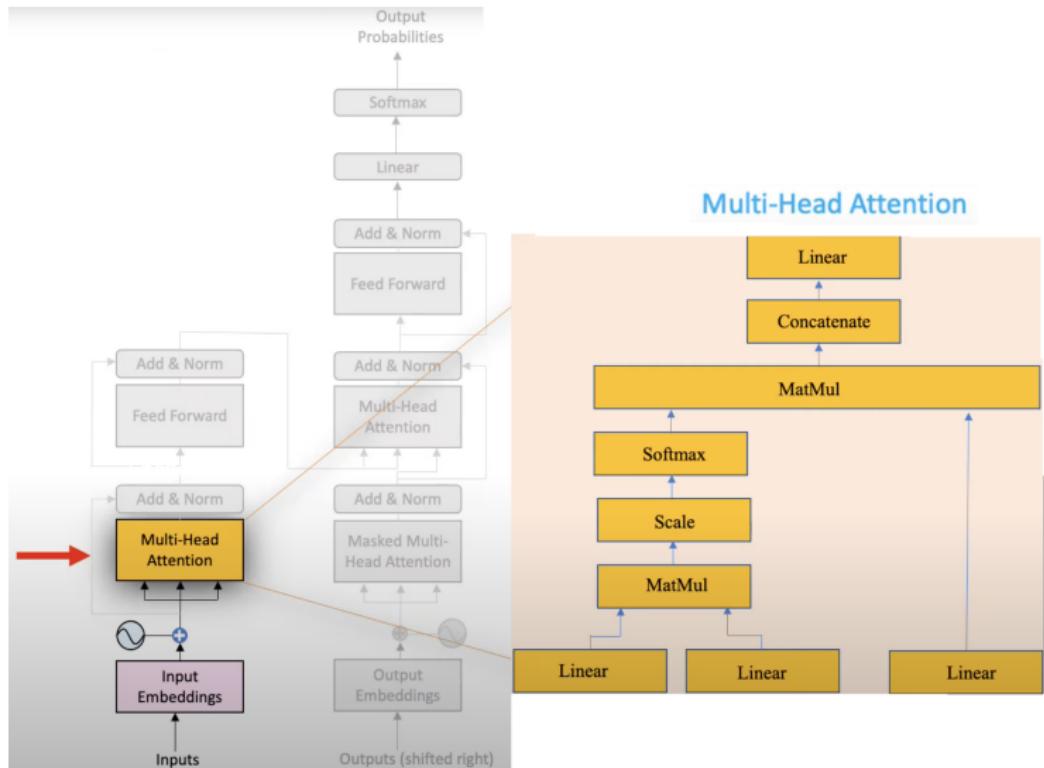
$$\boxed{\text{Positional Encoding}}$$

Vector Encoding of position in sentence

$$\begin{bmatrix} 0.42 \\ 0.84 \\ 0.12 \\ 0.81 \end{bmatrix}$$

Embedding of Dog (with context info)

# Multi Head Attention



# Attention: Query, Key, Value

Concept of:

- ① Query (Q)
- ② Key (K)
- ③ Value (V)

All constructed from the embedding.



I came from your other question [Self-attention original work?](#) The key/value/query formulation of attention is from the paper [Attention Is All You Need](#).

31



How should one understand the queries, keys, and values



+50

The key/value/query concepts come from retrieval systems. For example, when you type a query to search for some video on Youtube, the search engine will map your **query** against a set of **keys** (video title, description etc.) associated with candidate videos in the database, then present you the best matched videos (**values**).

Mimic the retrieval of a value  $v_i$  for a query  $q$  based on a key  $k_i$  in DB.

$$\text{attention}(q, k, v) = \sum_i \text{similarity}(q, k_i) \times v_i \text{ (weighted)}$$

# Attention: Q, K, V and weights $Q_w$ , $K_w$ , $V_w$

Weights are initialised randomly using a random distribution. Example:



Because every input has a dimension of 4, each set of the weights must have a shape of 4x3.  
Weights are initialised randomly, it is done once before training.

Weights for key    Weights for query    Weights for value

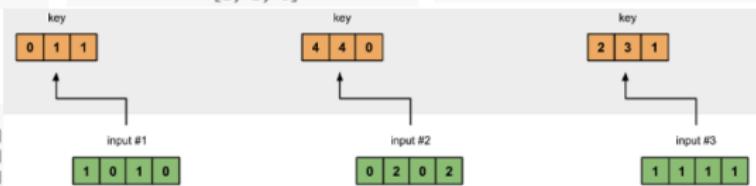
$\begin{bmatrix} 0, 0, 1 \\ 1, 1, 0 \\ 0, 1, 0 \\ 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 1, 0, 1 \\ 1, 0, 0 \\ 0, 0, 1 \\ 0, 1, 1 \end{bmatrix}$	$\begin{bmatrix} 0, 2, 0 \\ 0, 3, 0 \\ 1, 0, 3 \\ 1, 1, 0 \end{bmatrix}$
--	--	--

Key representation  
for input 1:

$$\begin{aligned} & [0, 0, 1] & [0, 0, 1] & [0, 0, 1] \\ & [1, 0, 1, 0] \times [1, 1, 0] = [0, 1, 1] & [0, 2, 0, 2] \times [1, 1, 0] = [4, 4, 0] & [1, 1, 1, 1] \times [1, 1, 0] = [2, 3, 1] \\ & [0, 1, 0] & [0, 1, 0] & [0, 1, 0] \\ & [1, 1, 0] & [1, 1, 0] & [1, 1, 0] \end{aligned}$$

Key representation:  
(Vectorise)

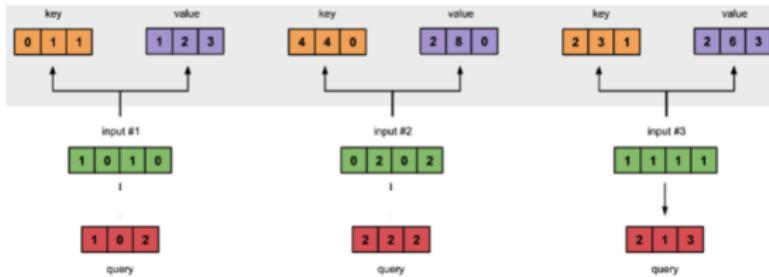
$$\begin{aligned} & [0, 0, 1] & [0, 1, 1] \\ & [1, 0, 1, 0] \quad [1, 1, 0] \quad [0, 1, 1] \\ & [0, 2, 0, 2] \times [0, 1, 0] = [4, 4, 0] \\ & [1, 1, 1, 1] \quad [1, 1, 0] \quad [2, 3, 1] \end{aligned}$$



# Attention: Q, K, V and weights $Q_w$ , $K_w$ , $V_w$

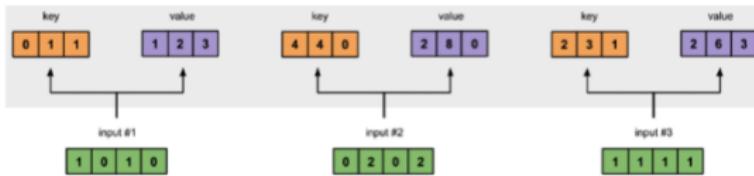
## Query representation:

$$\begin{array}{ccc} [1, 0, 1, 0] & [1, 0, 0] & [1, 0, 2] \\ [0, 2, 0, 2] \times [0, 0, 1] = [2, 2, 2] & & \\ [1, 1, 1, 1] & [0, 1, 1] & [2, 1, 3] \end{array}$$



## Value representation:

$$\begin{array}{ccc} [0, 2, 0] & [0, 3, 0] & [1, 2, 3] \\ [1, 0, 1, 0] \times [1, 0, 3] = [2, 8, 0] & & \\ [0, 2, 0, 2] & [1, 1, 0] & [2, 6, 3] \end{array}$$

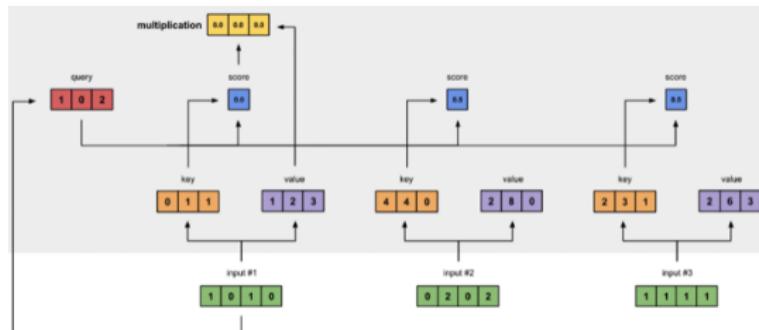


# Attention: Q, K, V and weights $Q_w$ , $K_w$ , $V_w$

Attention scores - dot product between Input 1's query (red) with all keys (orange), including itself

$[0, 4, 2]$   
 $[1, 0, 2] \times [1, 4, 3] = [2, 4, 4]$   
 $[1, 0, 1]$

$\text{softmax}([2, 4, 4]) = [0.0, 0.5, 0.5]$



Multiply Softmax attention scores for each input (blue) by its corresponding value (purple).

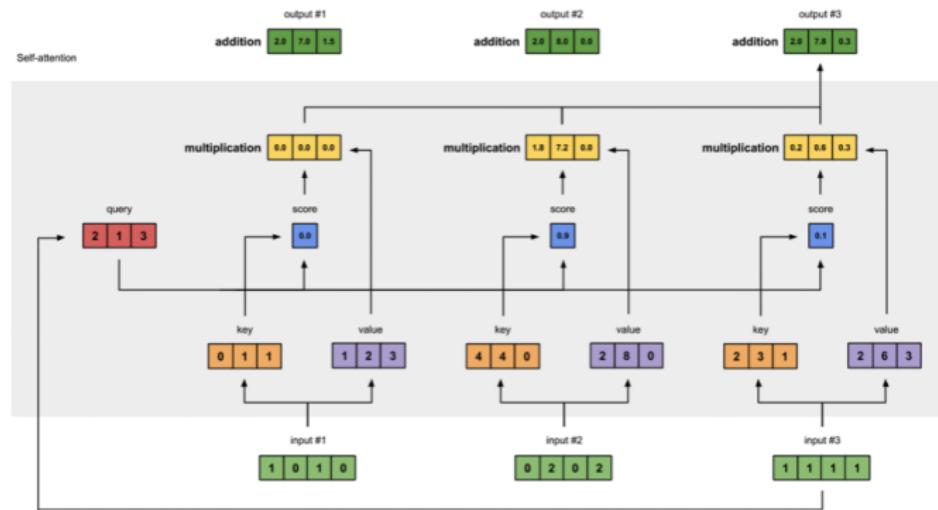
1:  $0.0 * [1, 2, 3] = [0.0, 0.0, 0.0]$   
2:  $0.5 * [2, 8, 0] = [1.0, 4.0, 0.0]$   
3:  $0.5 * [2, 6, 3] = [1.0, 3.0, 1.5]$

# Attention: Q, K, V and weights $Q_w$ , $K_w$ , $V_w$

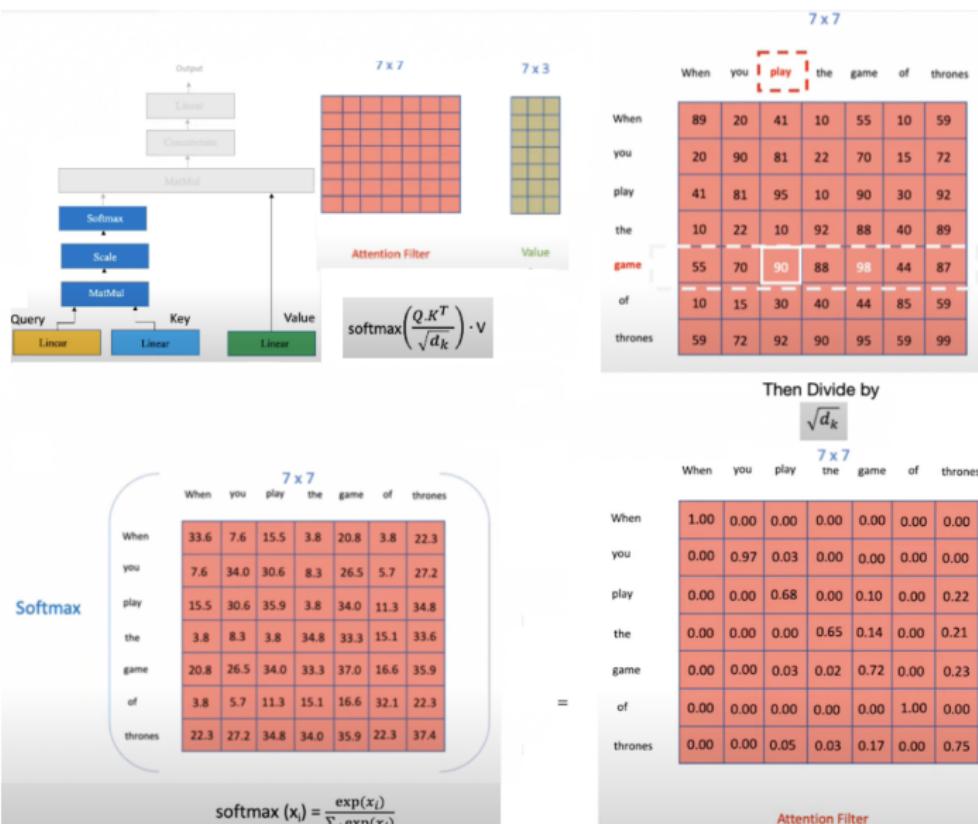
Sum weighted values to get Output 1

$$\begin{array}{l} [0.0, 0.0, 0.0] \\ + [1.0, 4.0, 0.0] \\ + [1.0, 3.0, 1.5] \\ \hline = [2.0, 7.0, 1.5] \end{array}$$

Repeat for  
Input 2 &  
Input 3

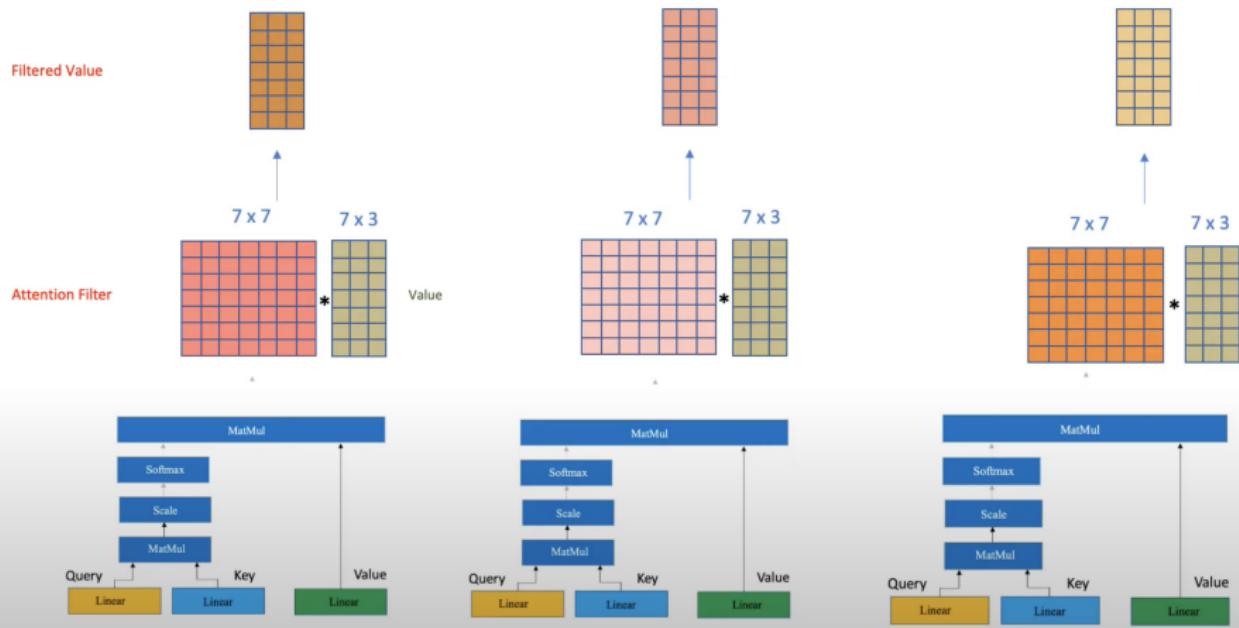


# Attention Filter - filter out unnecessary information (noise)

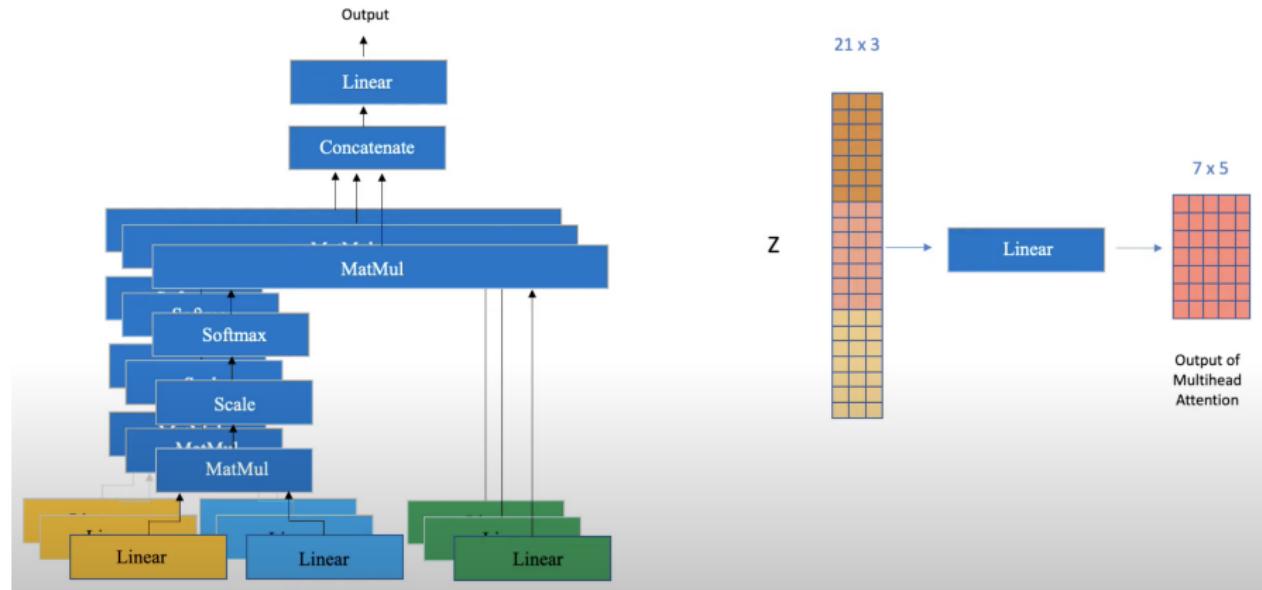


# Multi Head Attention

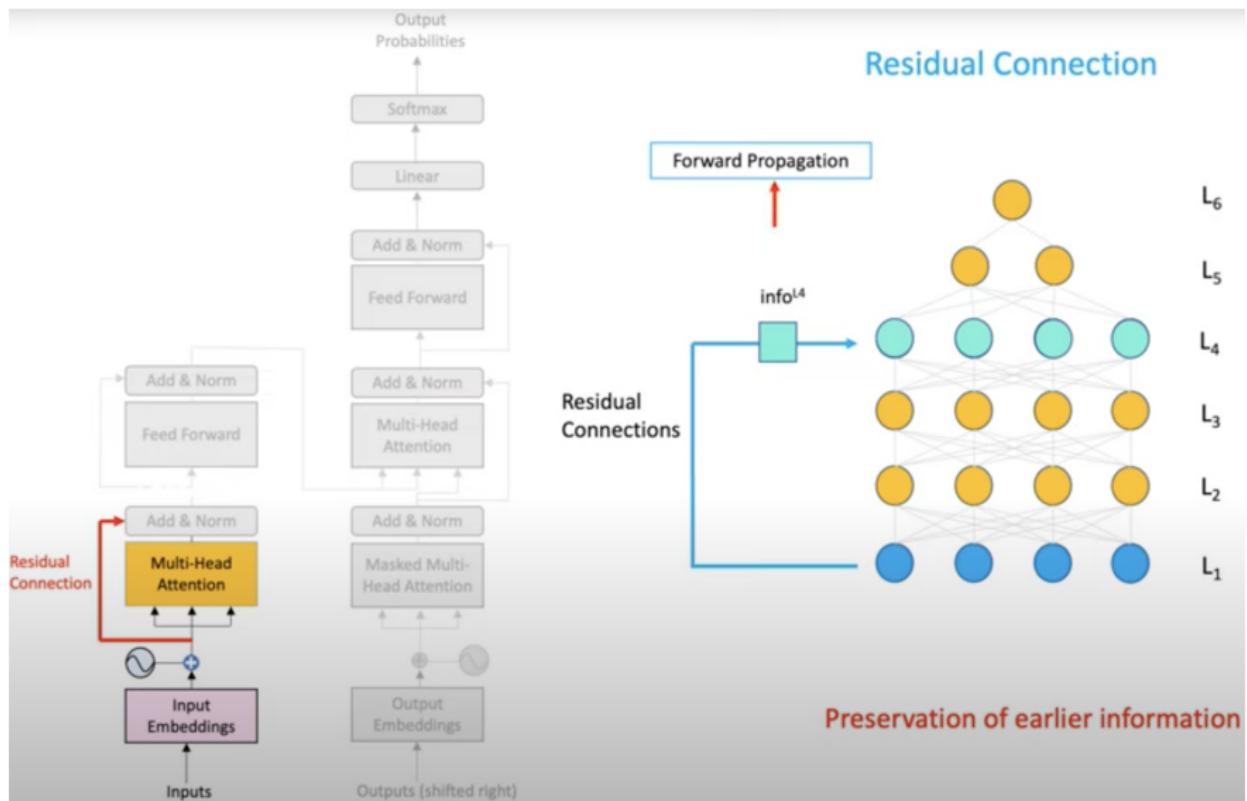
## Multi-Head Attention



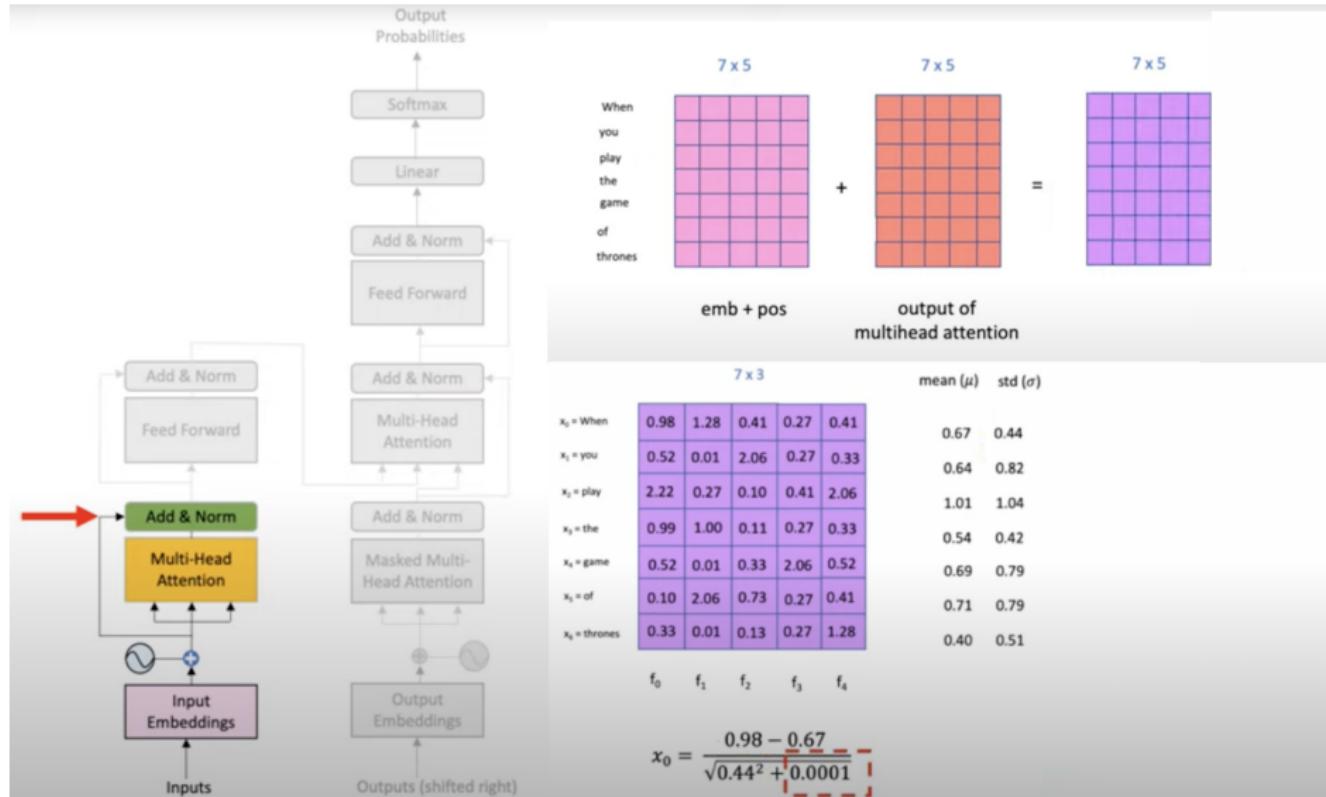
# Multi Head Attention - Concatenation



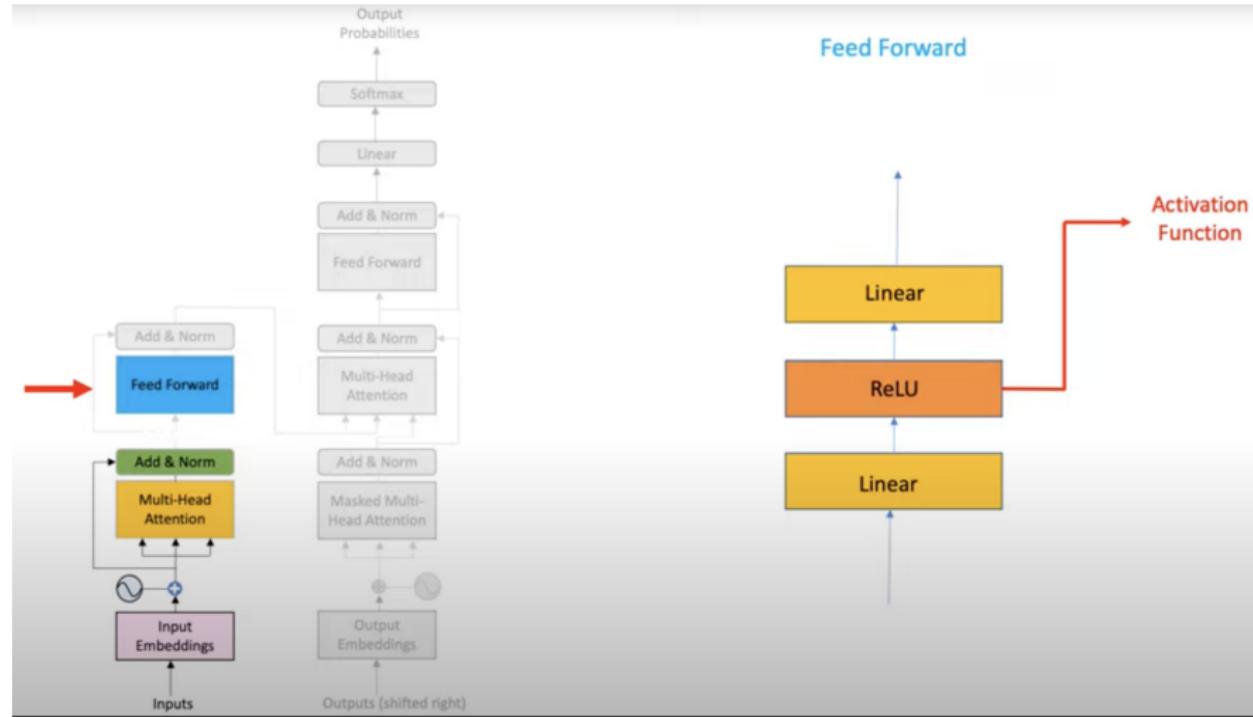
# Information Preservation



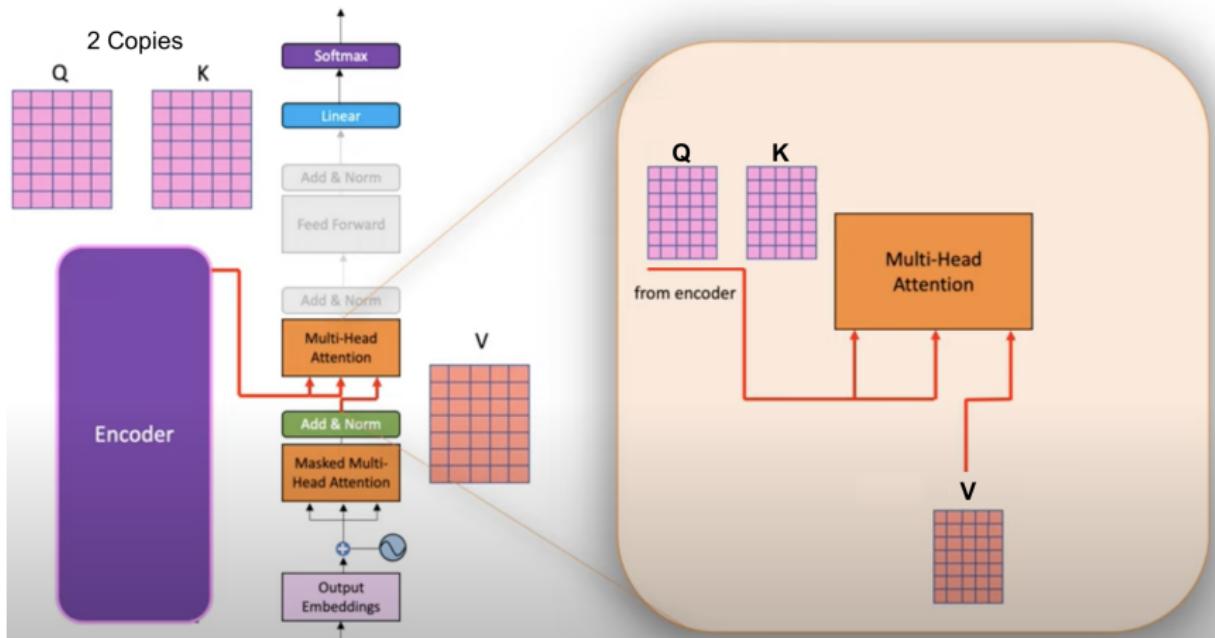
# Add (Preserve Information) & Normalisation



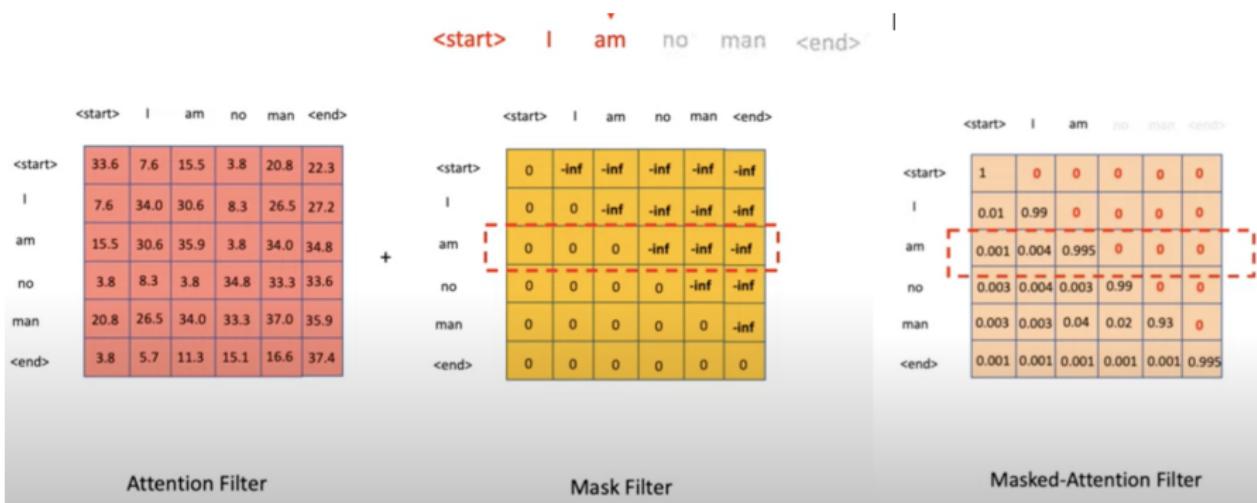
# Feed Forward



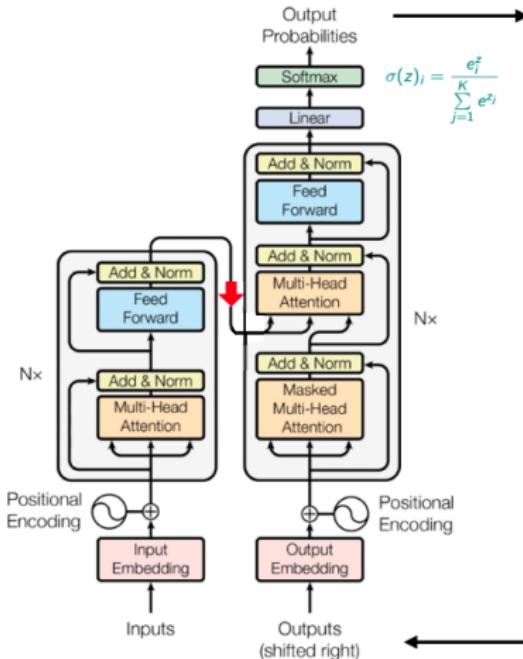
# Decoder Layer



# Masking



# Decoder Output



Decoders are autoregressive models;  
They are trained to predict the next token  
after reading the preceding ones

## Potential Innovations

- ➊ Softmax calculation for output probabilities are independent and is simplistic. In many cases, joint probabilities for bi-gram and tri-gram occurrences are more appropriate.
- ➋ The next token generated in the decode can be a "composite" token in response to the computed attention from the input.
- ➌ Reduce "Cost to Serve (CTS)" through reduced parameters and better efficiency through quantisation.

# Conclusion

- ① GPT does not replicate human writing or speaking processes. Although they imitate human writing, any apparent cleverness primarily arises from our inclination to attribute human characteristics to non-human entities (anthropomorphization).
- ② LLMs are essentially establishing statistical connections among vectors representing words and more extended grammatical structures. Each word within a sentence is linked to the subsequent word in the sequence with an associated probability.
- ③ This diverges significantly from human cognitive processes, where we employ word meanings to construct intricate and precise structures of "meanings" and definitions. For LLMs, definitions are confined to statistical interrelations among intricate vectors encompassing words, sentences, and more extensive grammatical constructs. LLMs cannot innovate, not just yet.
- ④ That said, LLMs are impressive and have massive use cases.

# Questions?

Thank You!