Computing for Medicine

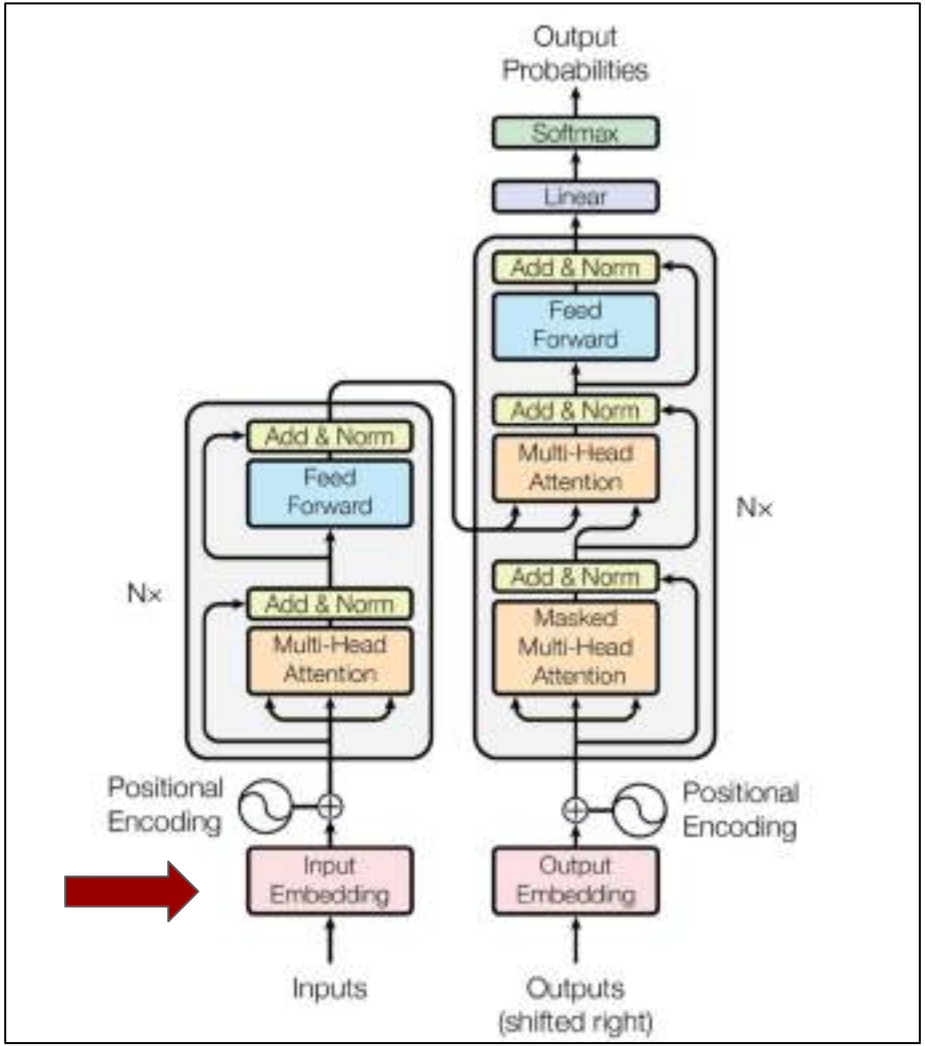
Google Classroom Code: dnd5qkt5

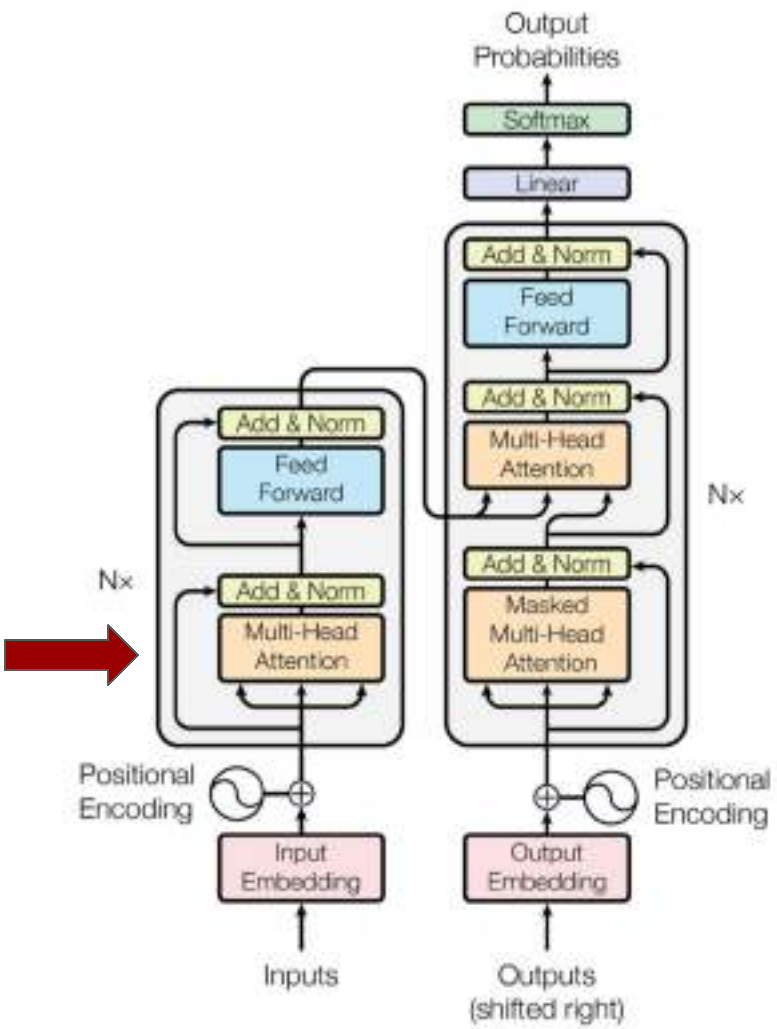
Monsoon 2025

Lecture 8

Transformer (Contd)

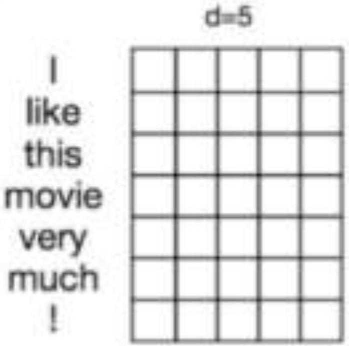
Recap Transformer

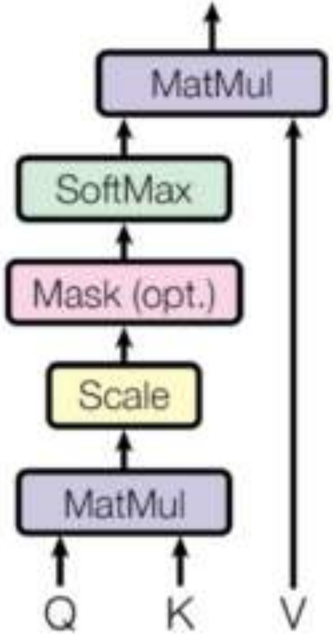


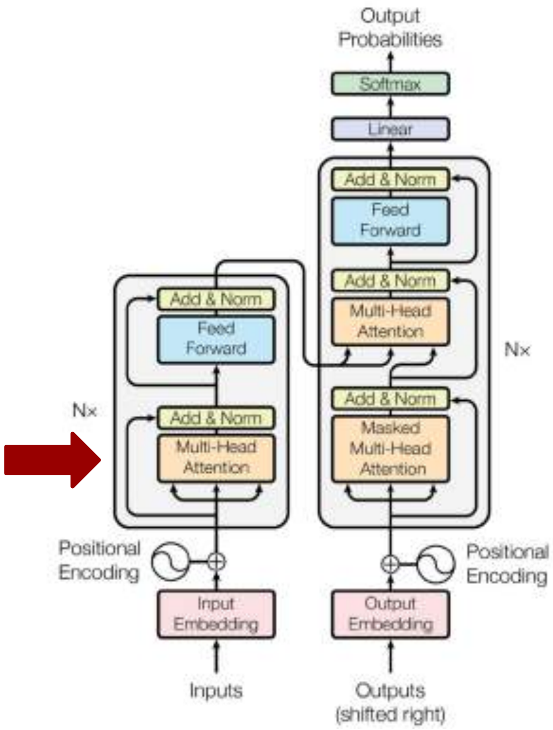

Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

Scaled Dot Product (Similarity with Context)

$$Attention(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

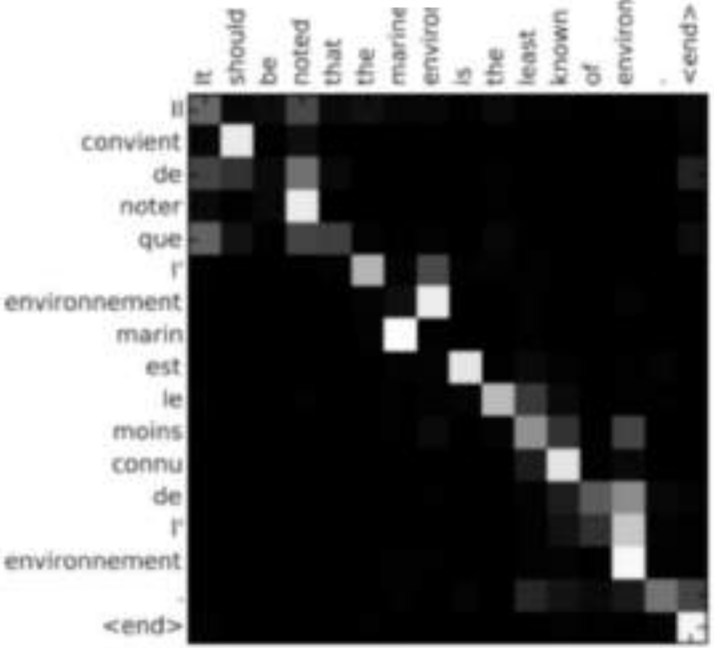




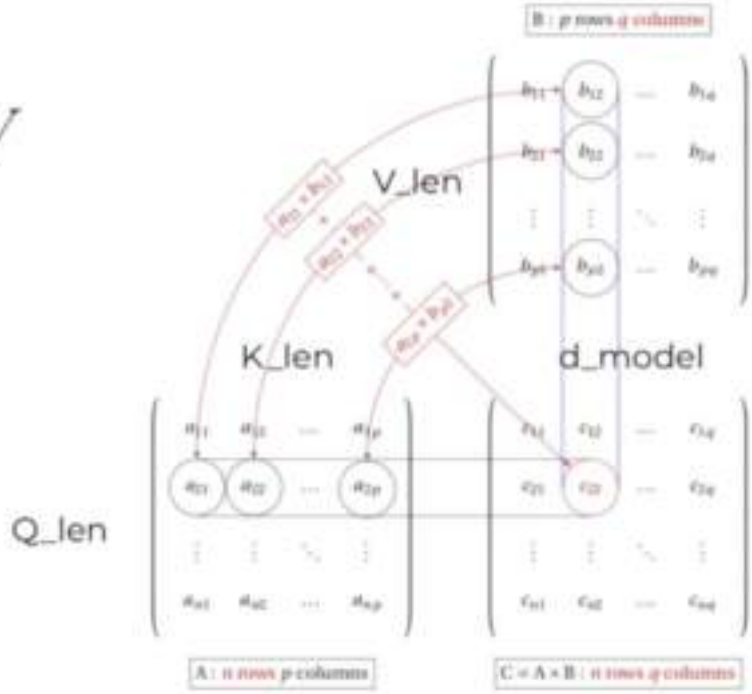


Mechanics of Attention

$$Attention(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



V



Breaking it down with an Example

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Q: queries $(n \times d_k)$
- K: keys $(m \times d_k)$
- V: values $(m \times d_v)$

Step 1: Generate a Similarity Score

$$S = QK^T \quad \Rightarrow \quad (n \times m)$$

E.g. n judges scoring m keys

Breaking it down with an Example

Step 2: Scale and Softmax

$$A = \text{softmax}\left(\frac{S}{\sqrt{d_k}}\right) \quad \Rightarrow \quad (n \times m)$$

Softmax turns each row into a probability distribution

Step 3: Retrieve the values (output embeddings)

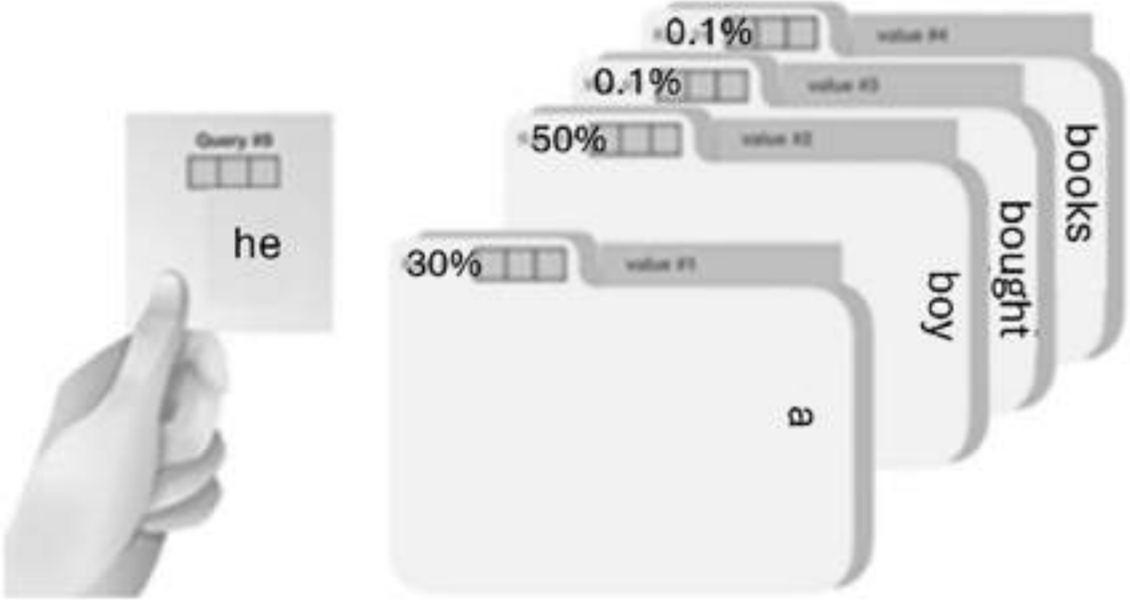
$$O = AV \quad \Rightarrow \quad (n \times d_v)$$

Analogies to understand Query, Key, Value

- **Keys (K)** → "Where should I look?" (addressing mechanism)
- **Queries (Q)** → "What am I asking about?" (the search intent)
- **Values (V)** → "What information do I retrieve once I've found the right spots?" (the actual content)

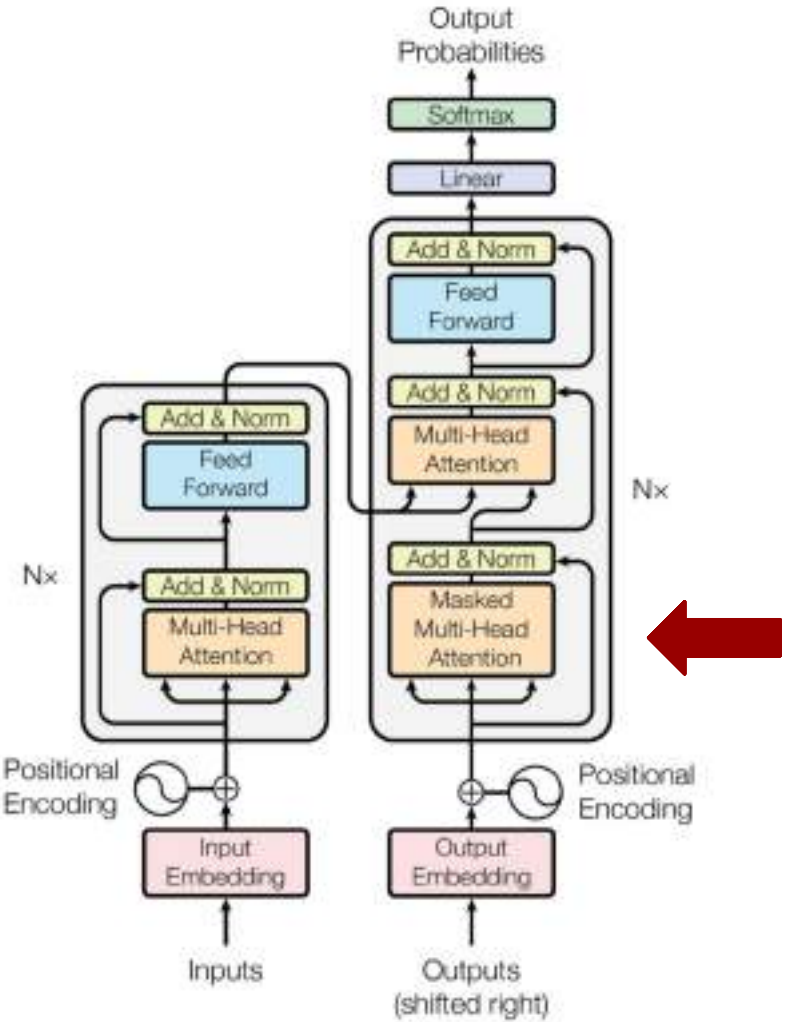
- **Key** = book's index card (used to locate it).
- **Query** = your search question.
- **Value** = the actual book content you take home once you've matched the index card.

Library Analogy: Query, Key, Value

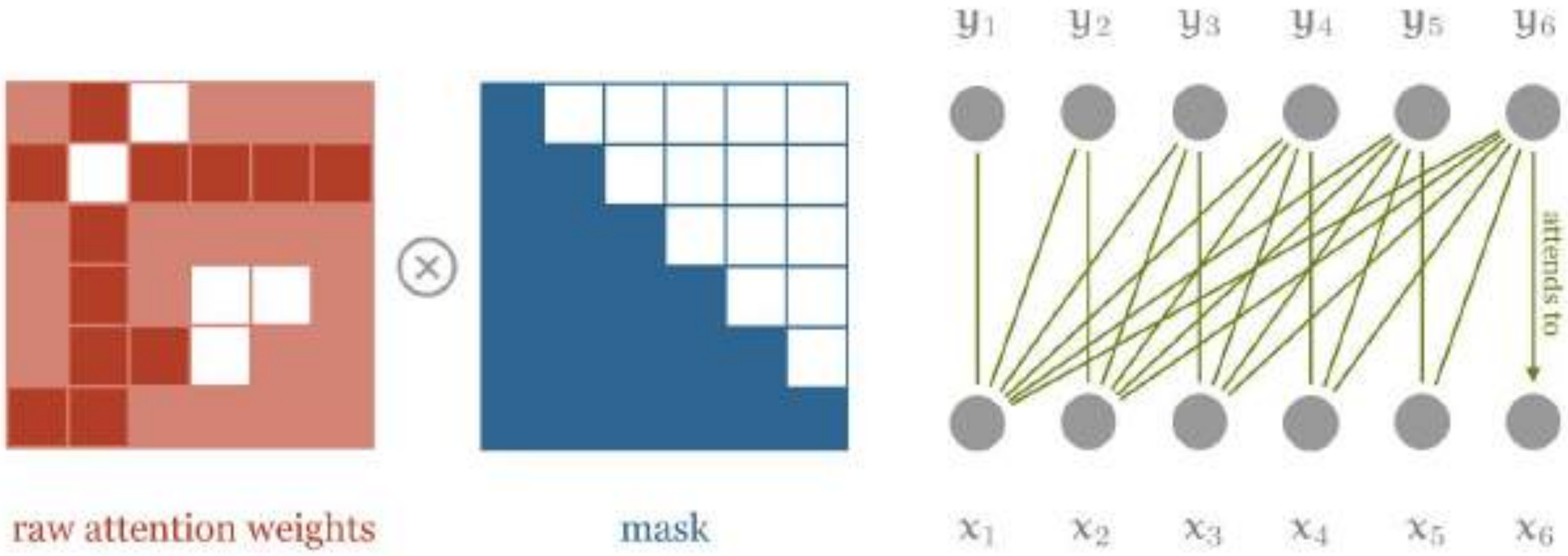


Masked Attention

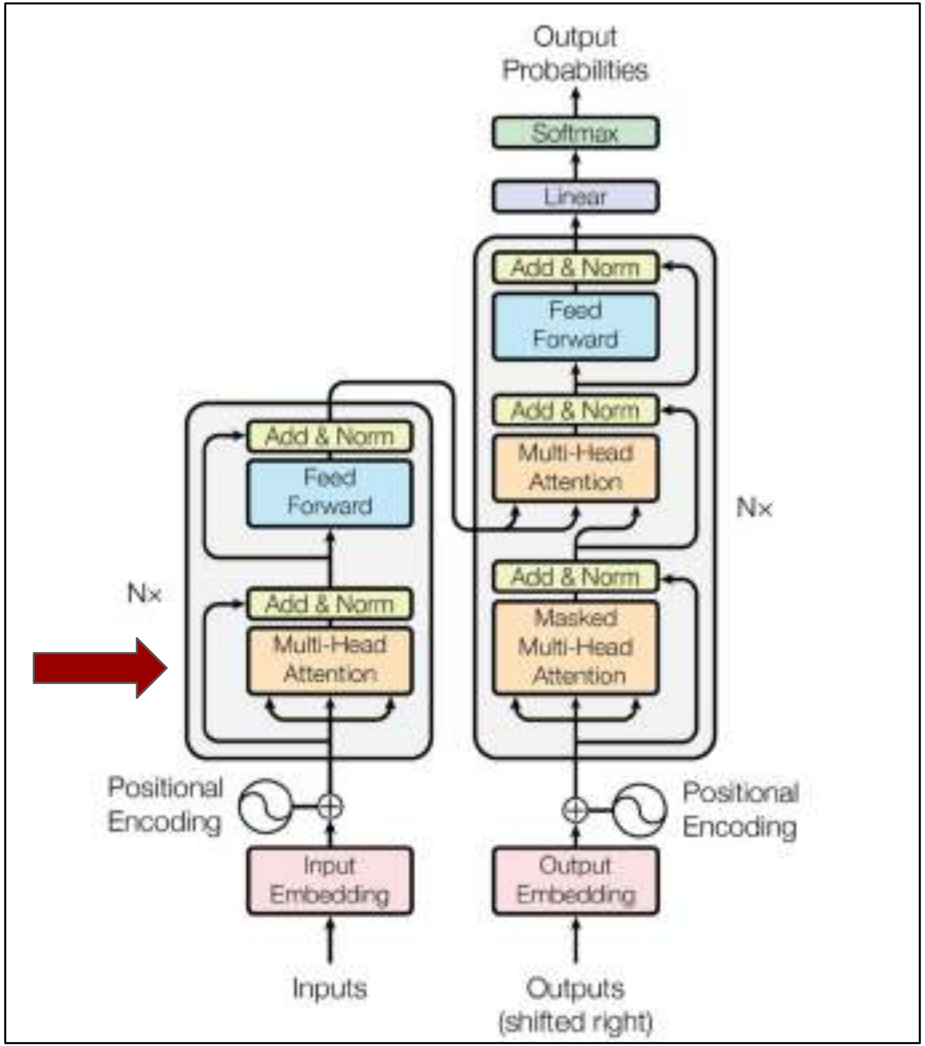
$$QK^T \quad * \quad \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$



Look ahead mask

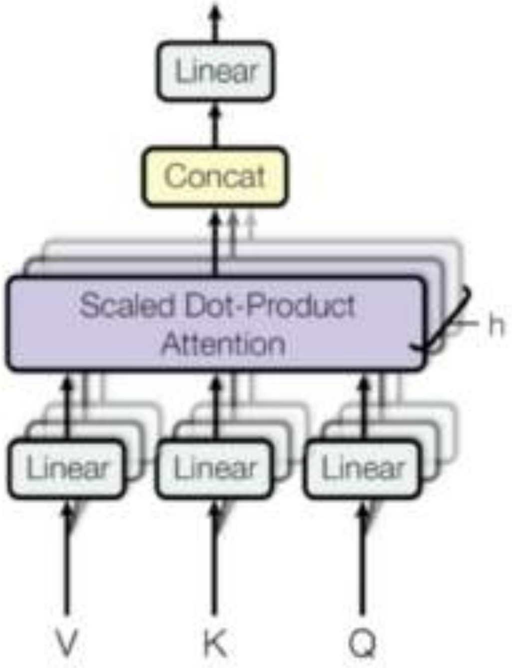


http://peterbloem.nl/blog/transformers

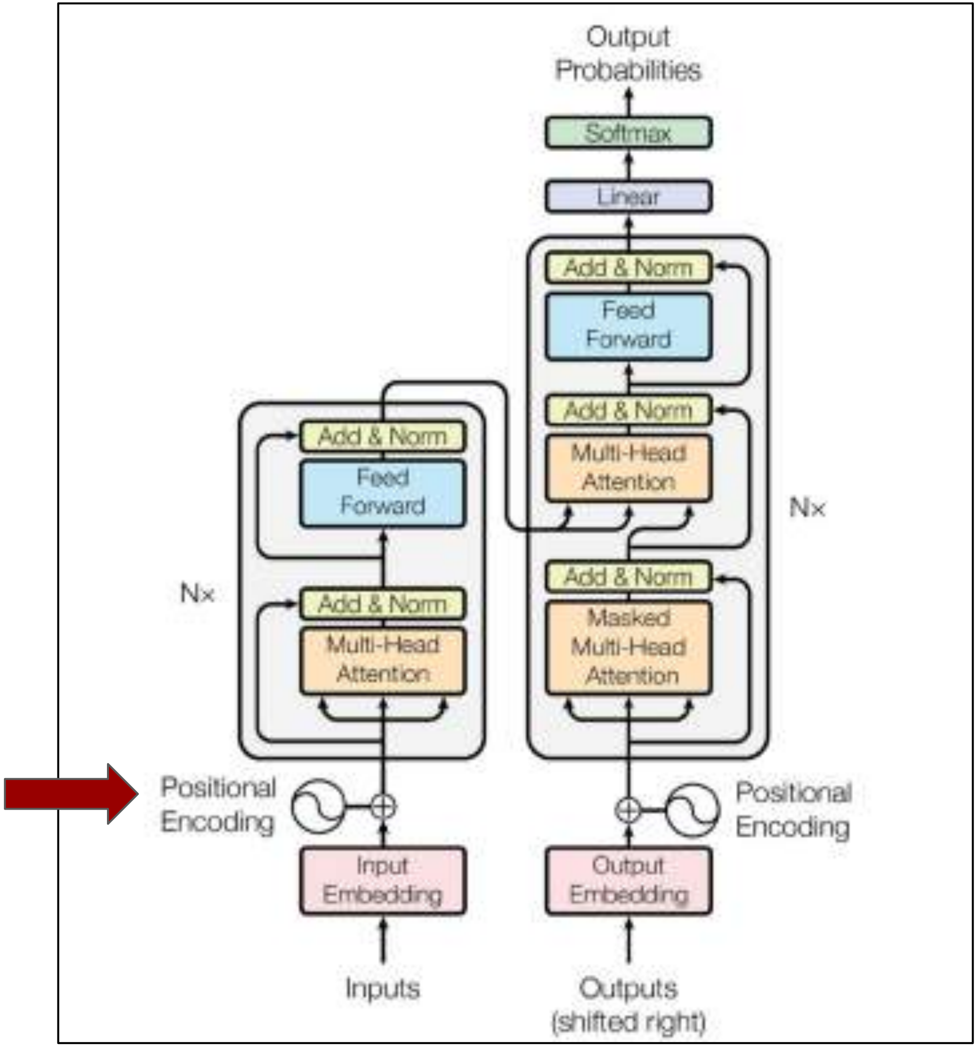


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Nx
Add & Norm
Multi-Head Attention
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Multi-head attention



Mathematically: one big linear function, and then a splitting allows each subspace to compose with the full original vector. Splitting and then applying a linear function restricts the possibilities.



Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

Positional Encodings

No convolution, no recurrence: Let's add a positional encoding!

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/dmodel})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/dmodel})$$

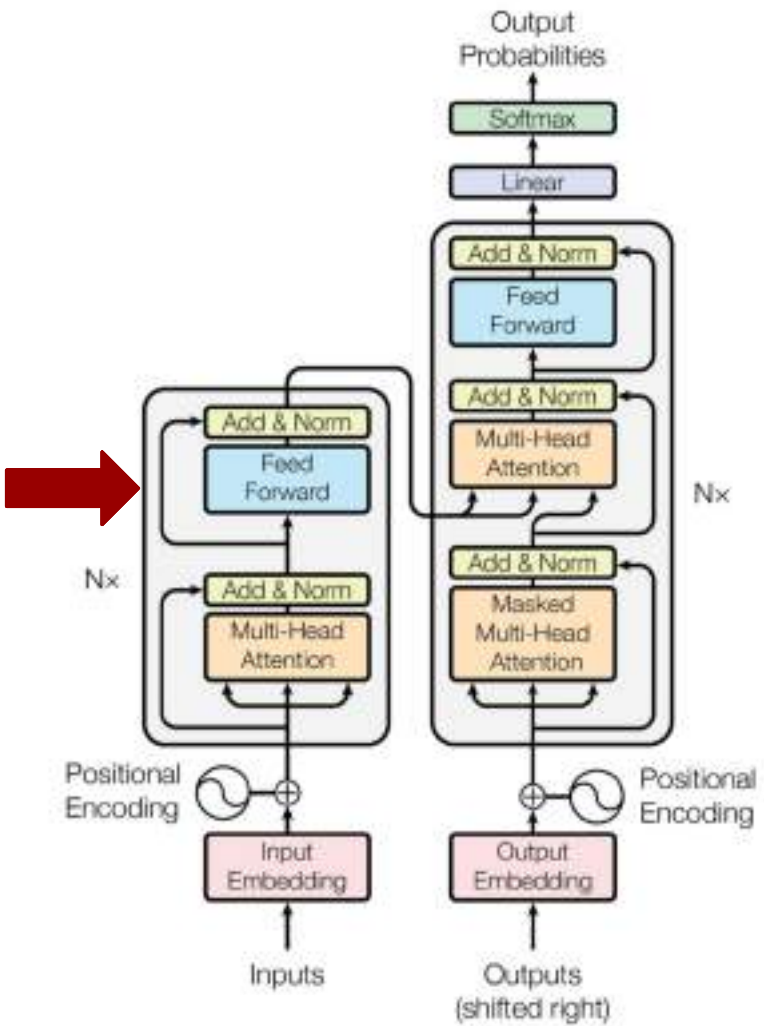
$$e(t) = \boldsymbol{E}_{t,:} := \begin{bmatrix} \sin\left(\frac{t}{f_1}\right) \\ \cos\left(\frac{t}{f_1}\right) \\ \sin\left(\frac{t}{f_2}\right) \\ \cos\left(\frac{t}{f_2}\right) \\ \vdots \\ \sin\left(\frac{t}{f_{\frac{d_{\text{model}}}{2}}}\right) \\ \cos\left(\frac{t}{f_{\frac{d_{\text{model}}}{2}}}\right) \end{bmatrix},$$

https://timodenk.com/blog/linear-relationships-in-the-transformers-positional-encoding/

Feed Forward Layers

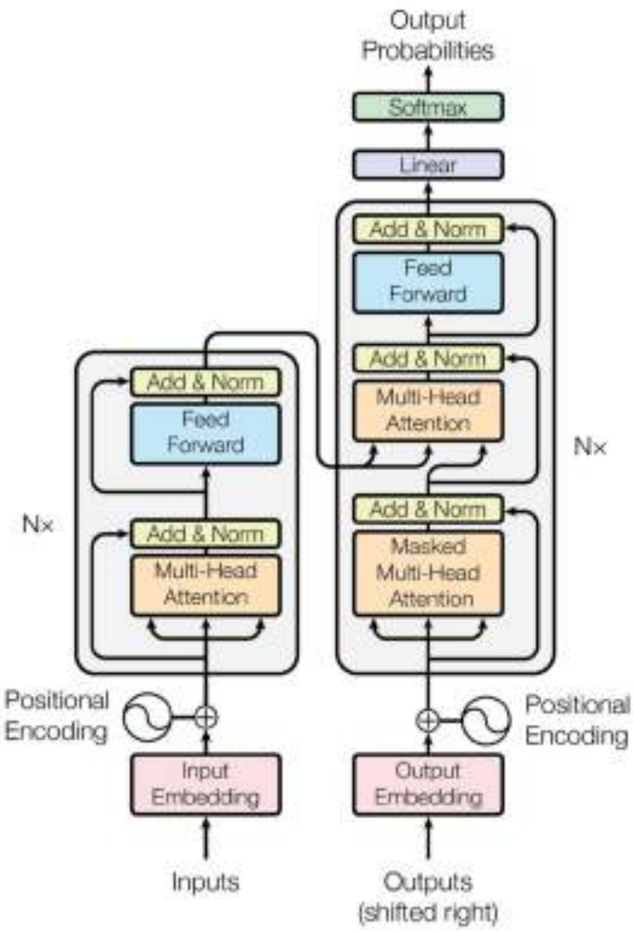
$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

- Fully Connected Dense Layers for learning
- Each FCN is different for each Nx
- Nx = 8 in original paper



Dropout

"We apply dropout to the output of each sub-layer before it is added to the sub-layer and normalized. In addition, we apply dropout to the sums of embeddings and positional encodings in both encoder and decoder stacks"



http://nlp.seas.harvard.edu/2018/04/03/attention.html

Thanks for attending the class!