AI4ICPS IIT Kharagpur

Introduction to Transformer Model

Jiaul Paik

Department of AI
IIT Kharagpur

jiaul@ai.iitkgp.ac.in

© 2026 IITKGP AI4ICPS I Hub Foundation. All rights reserved.

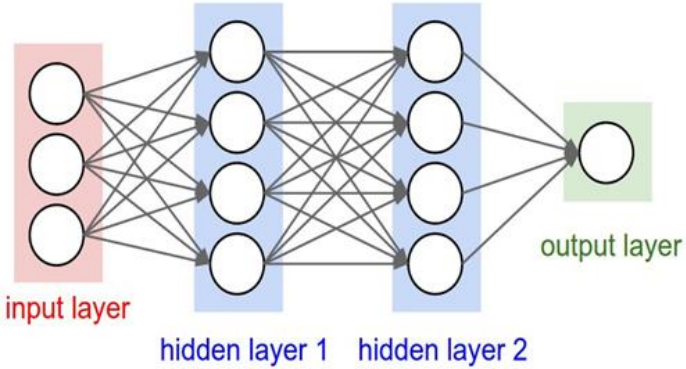
Recap of Necessary Concepts

- Projection
- Feedforward Neural Net and its matrix form
- Softmax
- Layer Normalization

Projection and Learning

AI4ICPS IIT Kharagpur

Feedforward Neural Net and its Matrix Form



Softmax

AI4ICPS IIT Kharagpur

Layer Normalization

For a given data input and layer, it computes the mean and variance over all the neurons in the layer

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}, \quad y_i = \gamma_i \hat{x}_i + \beta_i$$

where:

$$\mu = \frac{1}{D}\sum_{i=1}^{D} x_i, \quad \sigma^2 = \frac{1}{D}\sum_{i=1}^{D}(x_i - \mu)^2$$

AI4ICPS IIT Kharagpur

What is Transformer?

A Neural network that processes sequential data, understands the context and performs a set of tasks such as

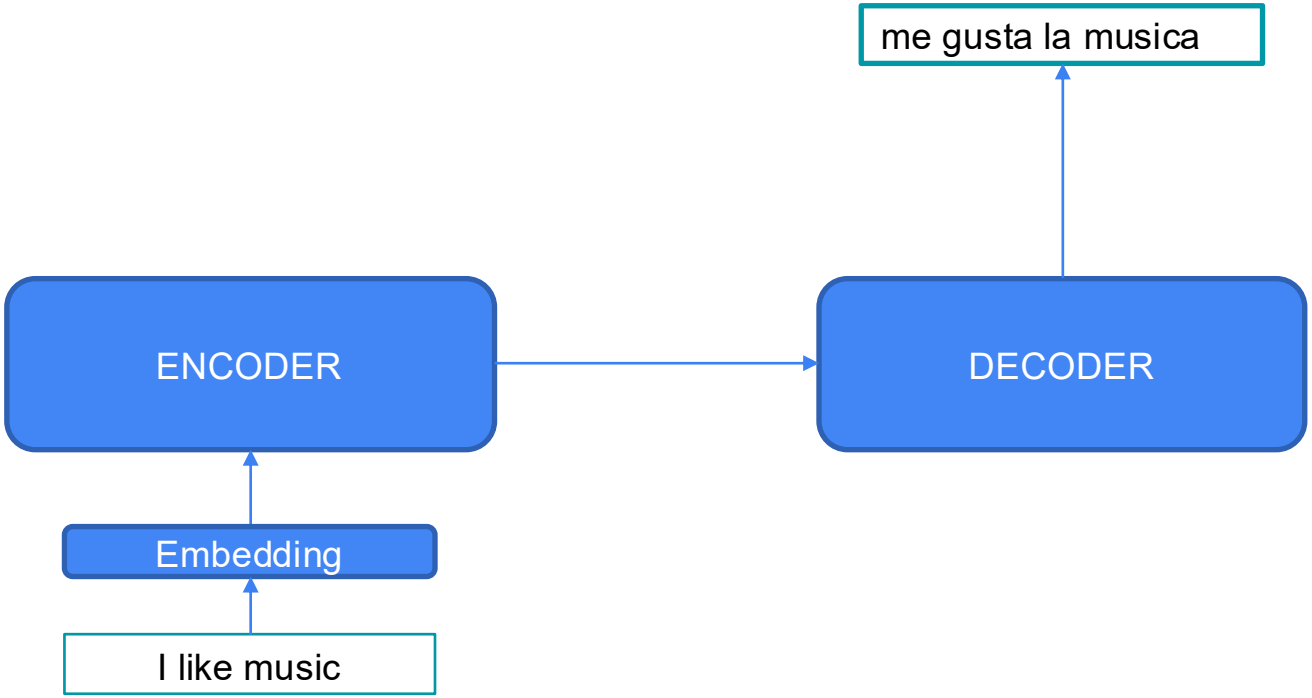
- Classification
- Translation
- Language Understanding

AI4ICPS IIT Kharagpur

What is Transformer?

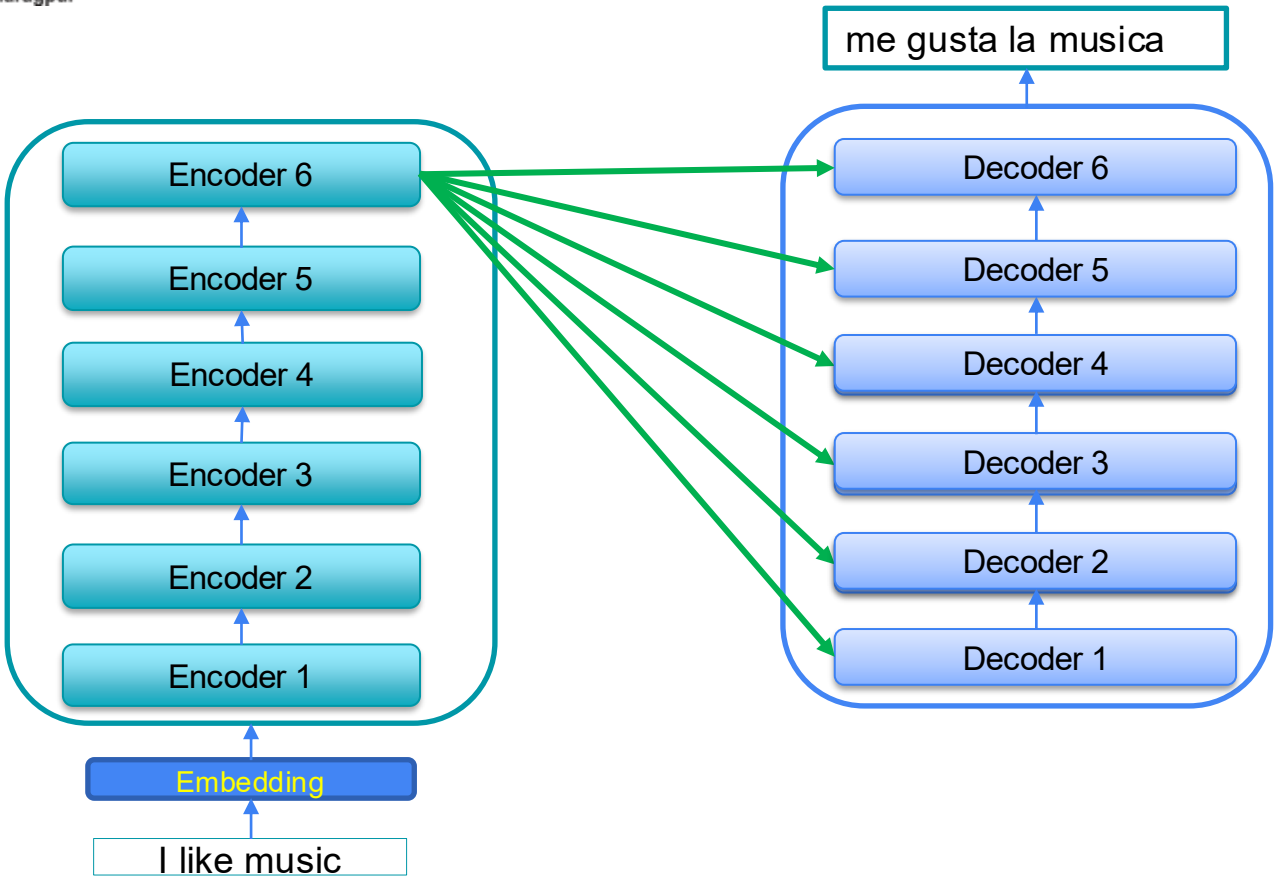
AI4ICPS IIT Kharagpur

Transformer Architecture



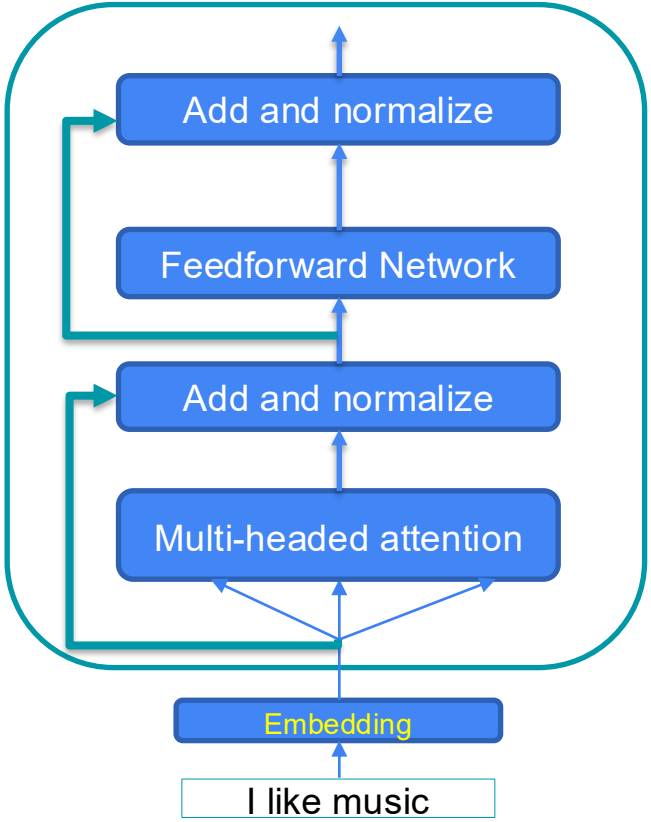
AI4ICPS IIT Kharagpur

Transformer Architecture



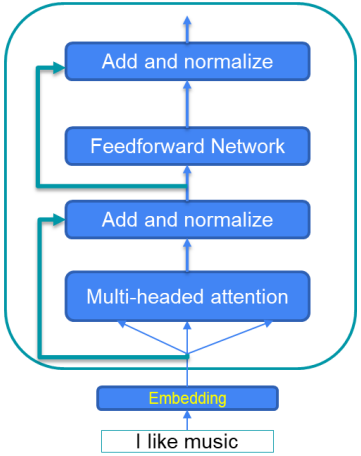
AI4ICPS IIT Kharagpur

Encoder Block: Inside Look

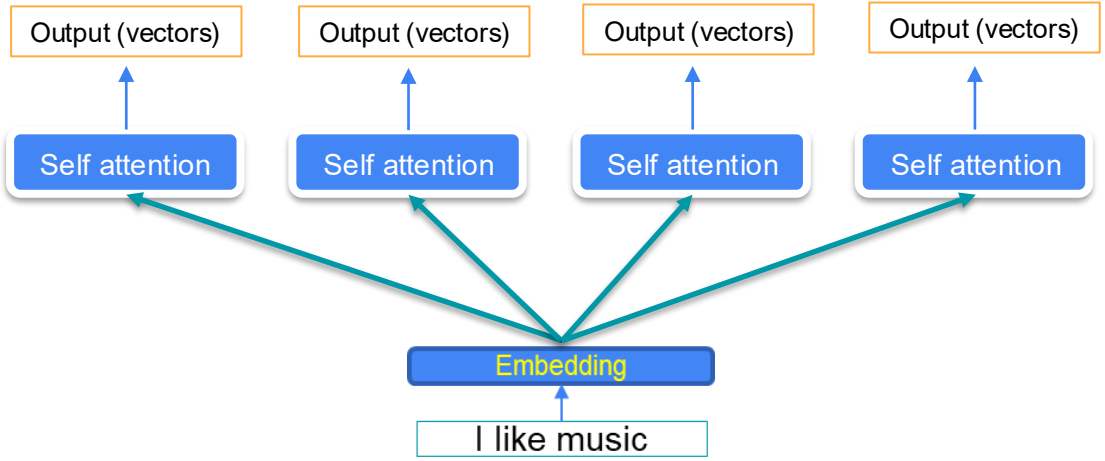


AI4ICPS IIT Kharagpur

Encoder Block: Multi-headed Attention



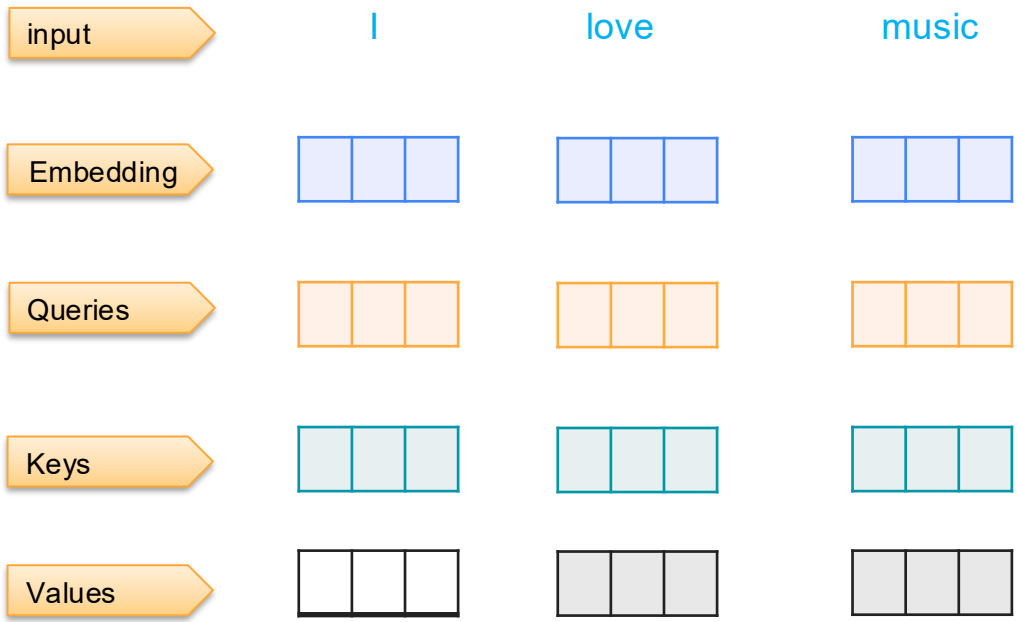
Application of self attention mechanism on the input multiple times to generate multiple representation



AI4ICPS IIT Kharagpur

Understanding Self-attention

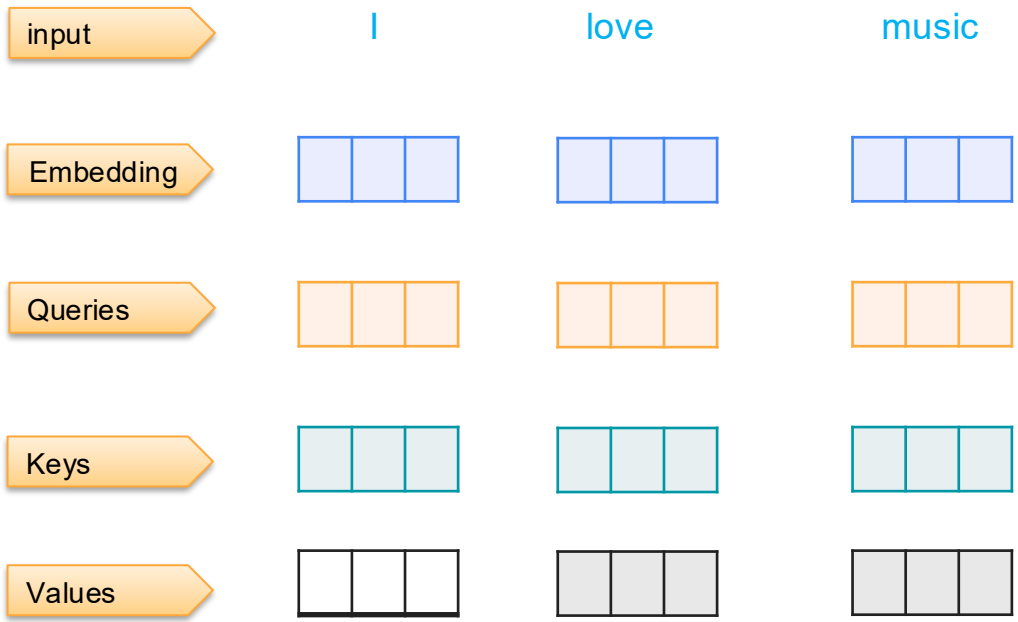
Goal: to understand the relationship between different units (e.g, words) in the input sequence.



AI4ICPS IIT Kharagpur

Generating Queries, Keys and Values

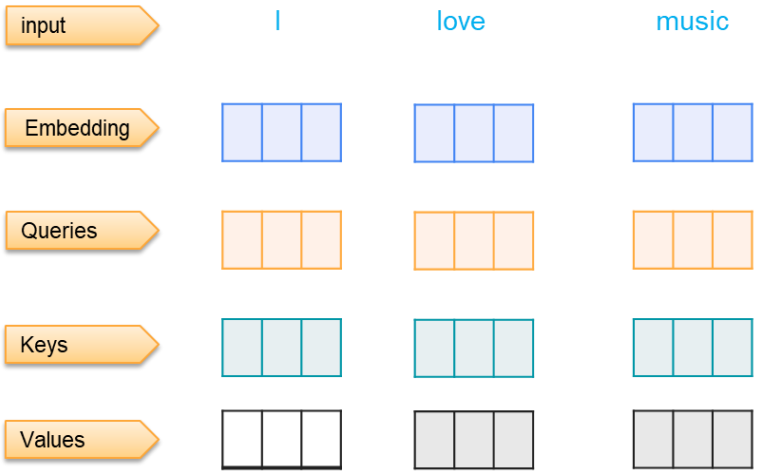
Goal: to understand the relationship between different units (e.g, words) in the input sequence.



AI4ICPS IIT Kharagpur

Self-attention: Scaled Dot Product

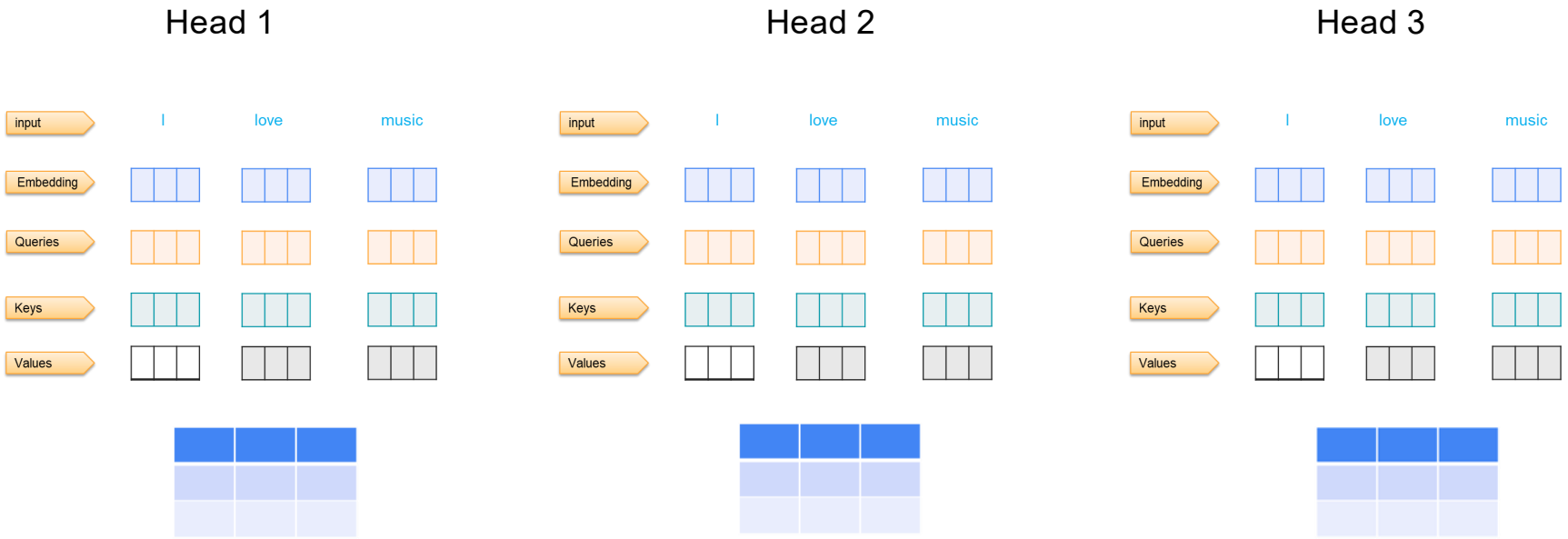
Goal: to understand the relationship between different units (e.g, words) in the input sequence.



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

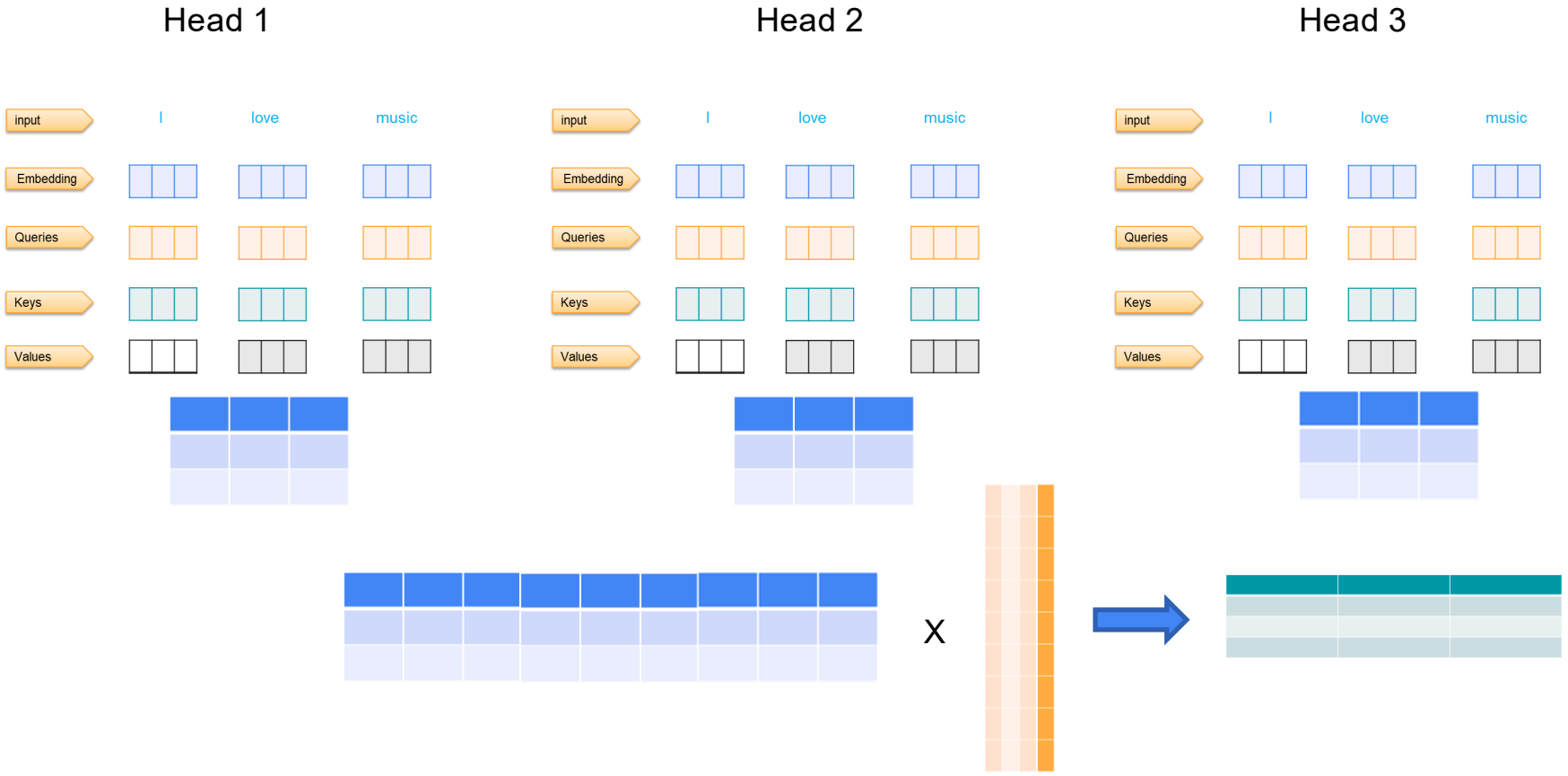
AI4ICPS IIT Kharagpur

Understanding Multi-headed attention



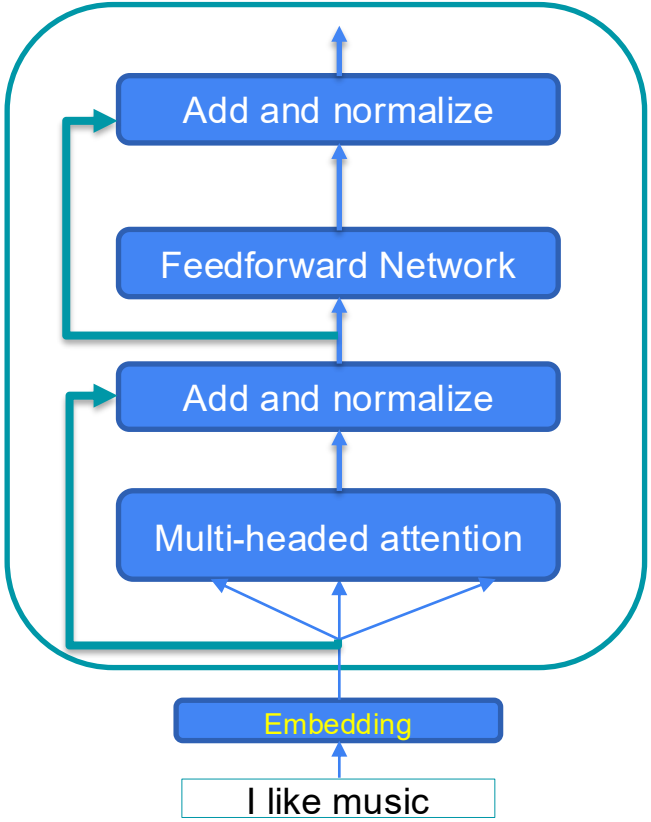
AI4ICPS IIT Kharagpur

Understanding Multi-headed attention



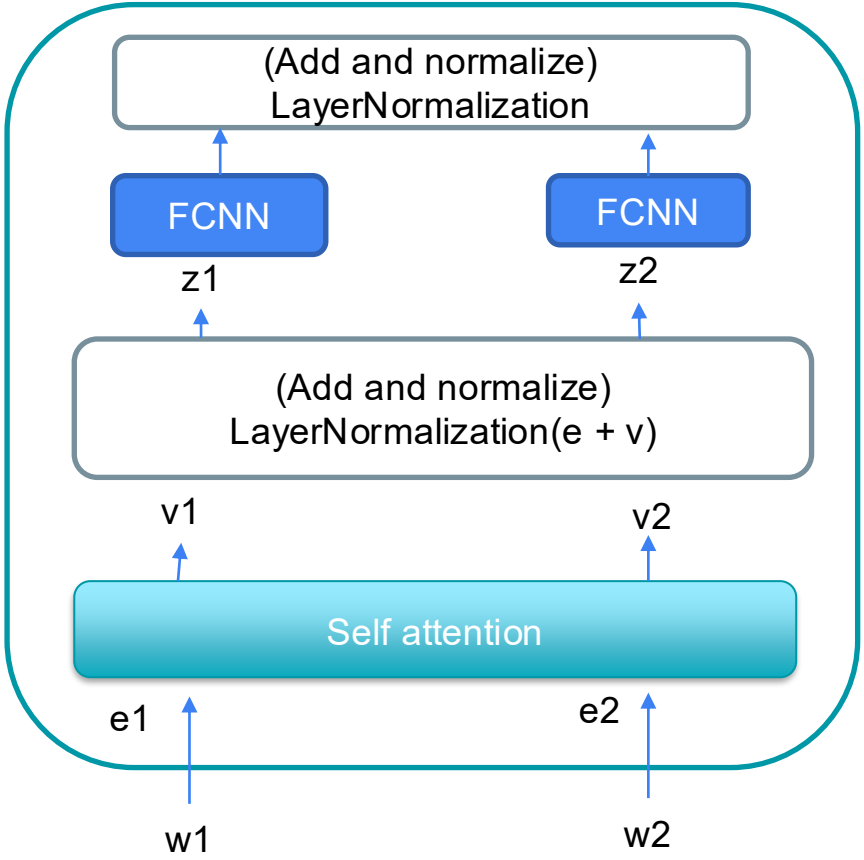
AI4ICPS IIT Kharagpur

Encoder Block



AI4ICPS IIT Kharagpur

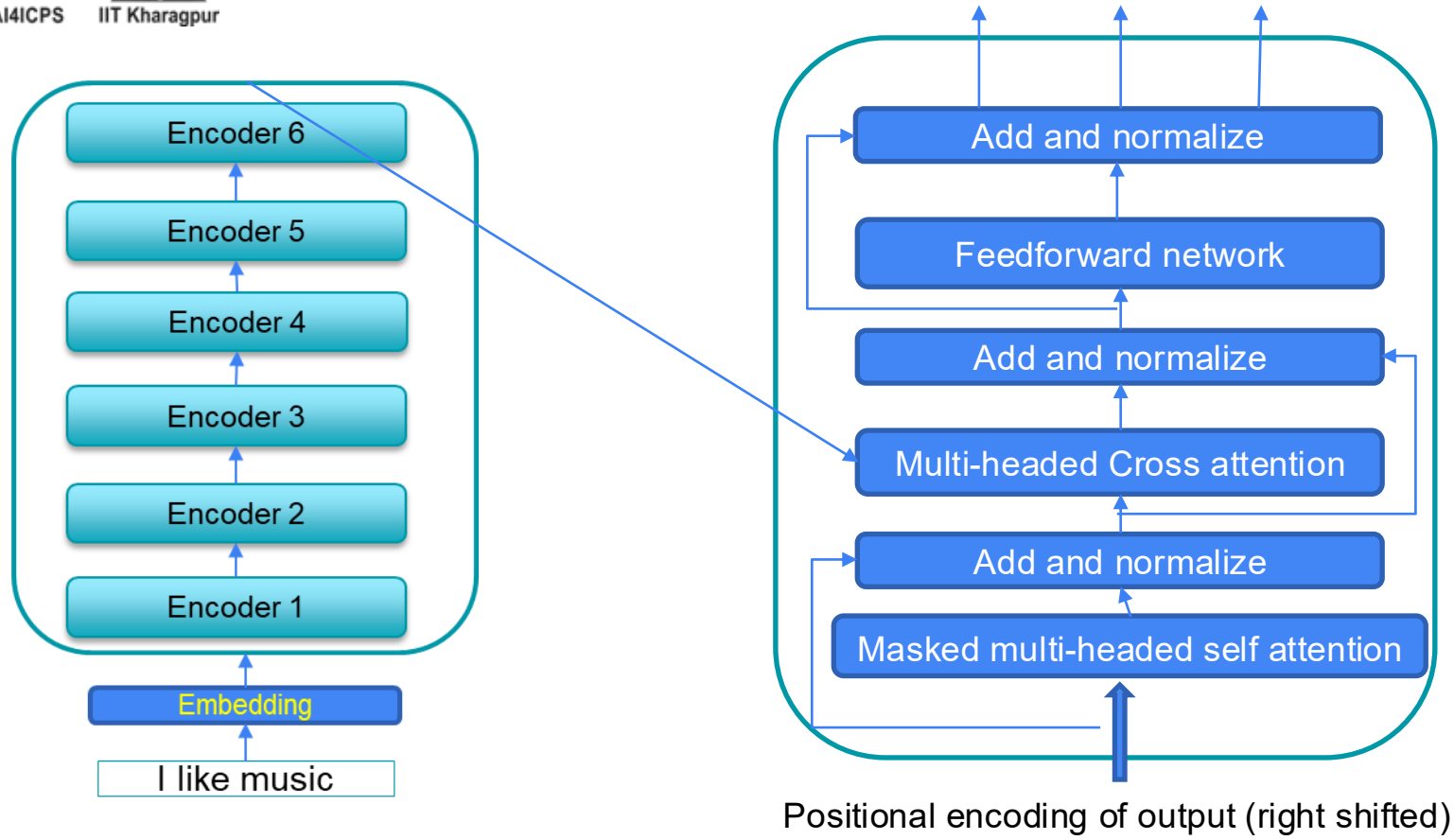
Add and Normalize



DECODER

AI4ICPS IIT Kharagpur

Decoder Block: Inside Look

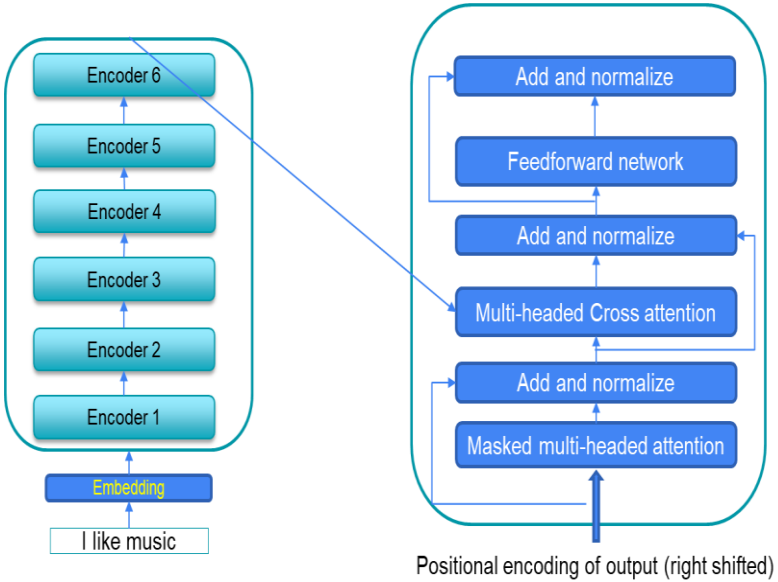


Masked Self-attention

- Similar to self-attention, but
  - Stops positions from attending to subsequent positions
  - Words do not get influenced by future tokens (can't see the future)
  - For example, when we are generating "good" in "Life is good", the masked self attention mechanism will only consider "Life" and "is" to generate "good".

AI4ICPS IIT Kharagpur

Encoder-decoder attention



- Interplay between encoder and decoder
- Queries come from the previous decoder layer
- Keys and values come from the output of the encoder

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


AI4ICPS

IIT Kharagpur

Positional Encoding

Motivation

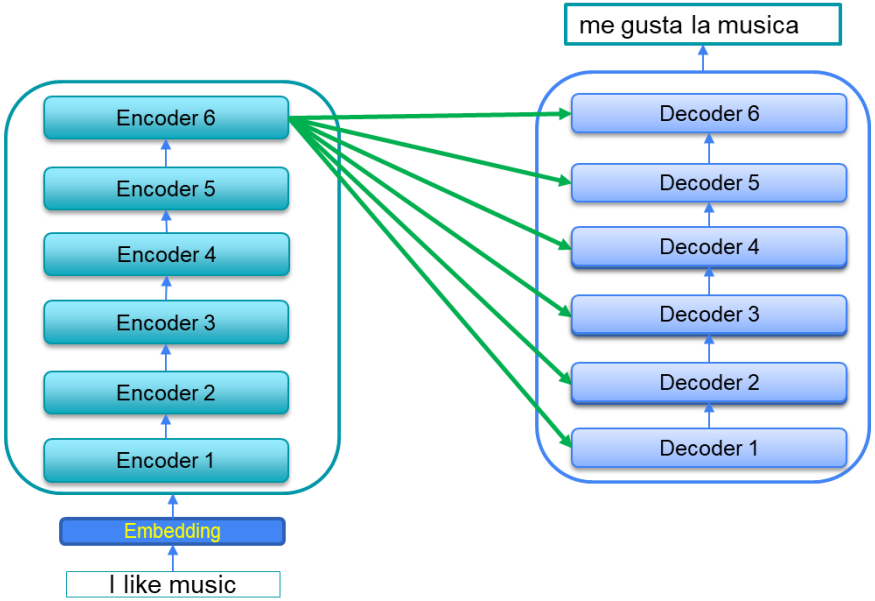
Recurrent Networks explicitly takes into account sequential information

But transformers do not

AI4ICPS IIT Kharagpur

Motivation

Words flow simultaneously through the encoder/decoder stacks



Properties

1. Unique positional encoding for each time step
2. Reasonable notion of relative distance
3. Independent of sentence size
4. Deterministic

AI4ICPS IIT Kharagpur

Sinusoidal/Cosinusoidal Positional Embedding

- Position will be a vector (instead of scalar)

$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k . t), & \text{if } i = 2k \\ \cos(\omega_k . t), & \text{if } i = 2k+1 \end{cases}$$

$$\omega_k = \frac{1}{10000^{2k/d}}$$

$$\overrightarrow{p_t} = \begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \vdots \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}_{d \times 1}$$

AI4ICPS IIT Kharagpur

Positional Encoding: Example

$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k . t), & \text{if } i = 2k \\ \cos(\omega_k . t), & \text{if } i = 2k+1 \end{cases}$$

$$\omega_k = \frac{1}{10000^{2k/d}}$$

Tokens
i
love
music

Index
0
1
2

i=0	i=0	i=1	i=1
Sin(0)	Cos(0)	Sin(0)	Cos(0)

Thank you!