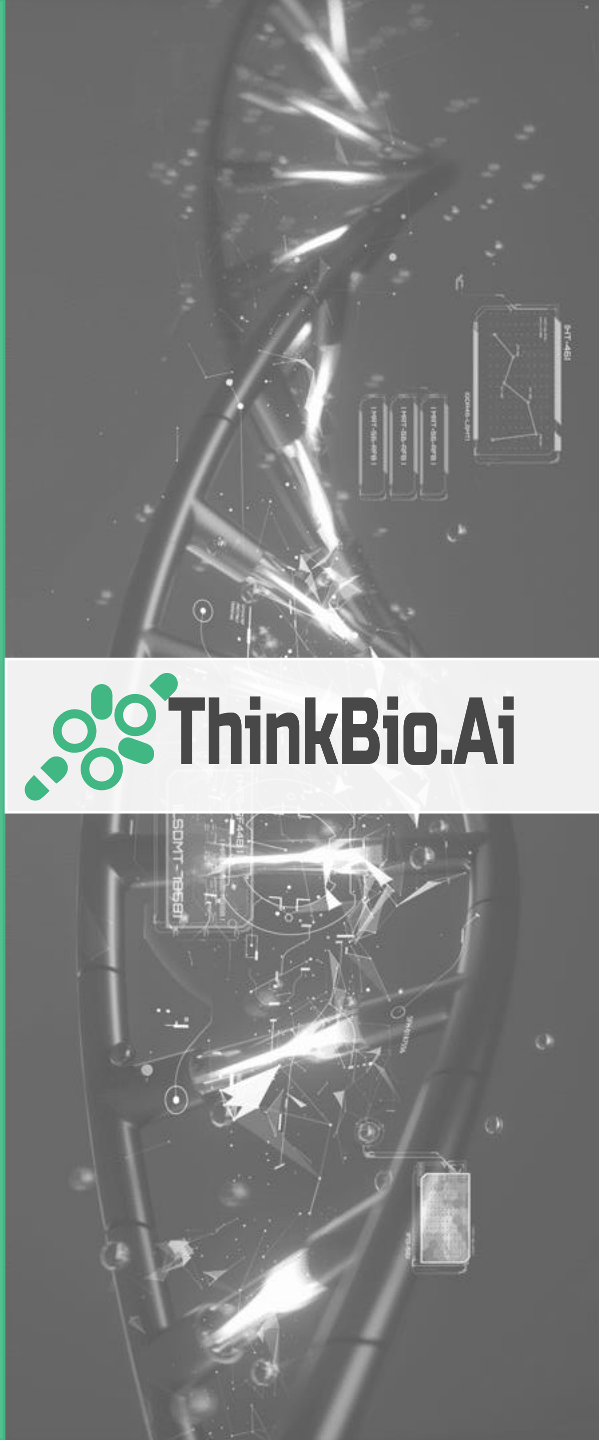
Discussion session
November – 1 - 2024

TRANSFORMER INTUITION : A prerequisite for LLMs and Biomedical Foundational Models

Vasudev R
Empid – 10073
Software Developer Trainee In Machine Learning
Machine Learning Department
Feathersoft Info IT Solutions Kochi



Agenda For Discussion

01 Sequence to Sequence Learning

02 Why Attention Mechanism

03 Mathematical Intuition Of Attention Mechanism

04 Transformer Architecture

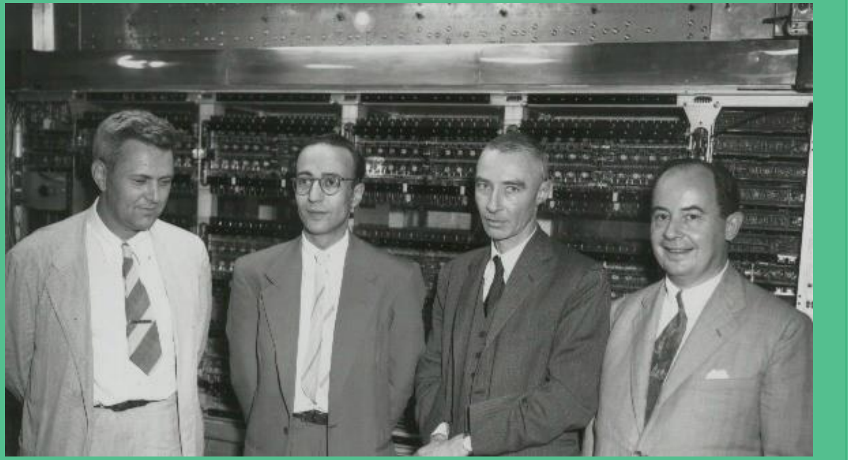
History of Seq2Seq Models

What is Seq2seq ?

- Input : a sequence (like sentence)
- Output : a sequence

Applications include

- Machine Translation
- Text Summarization
- Conversational Models

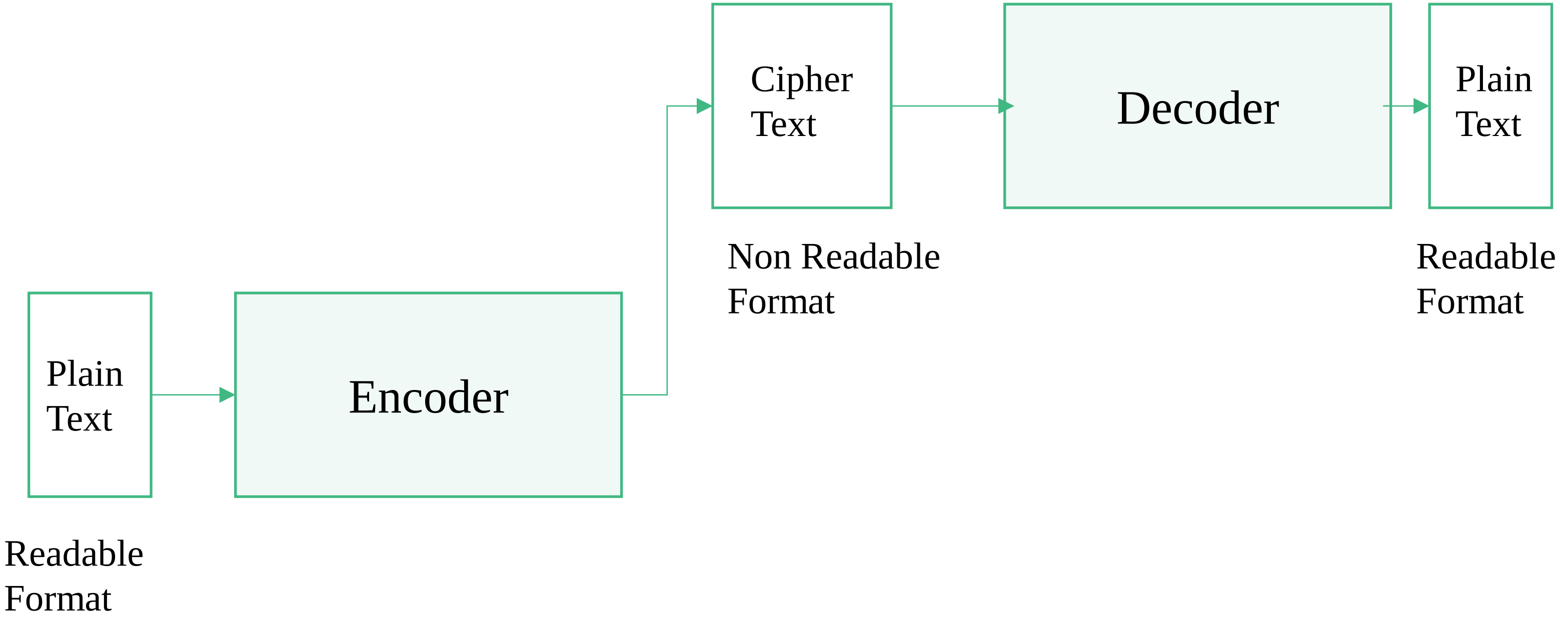


Warren Weaver, Letter to Norbert Wiener, March 4, 1947

" *One naturally wonders if the* ***problem of translation*** *could conceivably be treated as a* ***problem in cryptography.*** "

Cryptography Technique

High level Architecture – Encoder Decoder Model



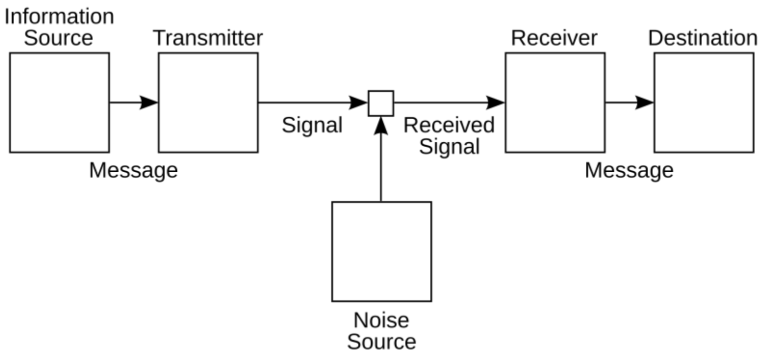
Two process is involved – Encryption and Decryption

Key Features

- Confidentiality
- Integrity
- Authentication

Sequential to Sequential Models

Signal Transduction Process



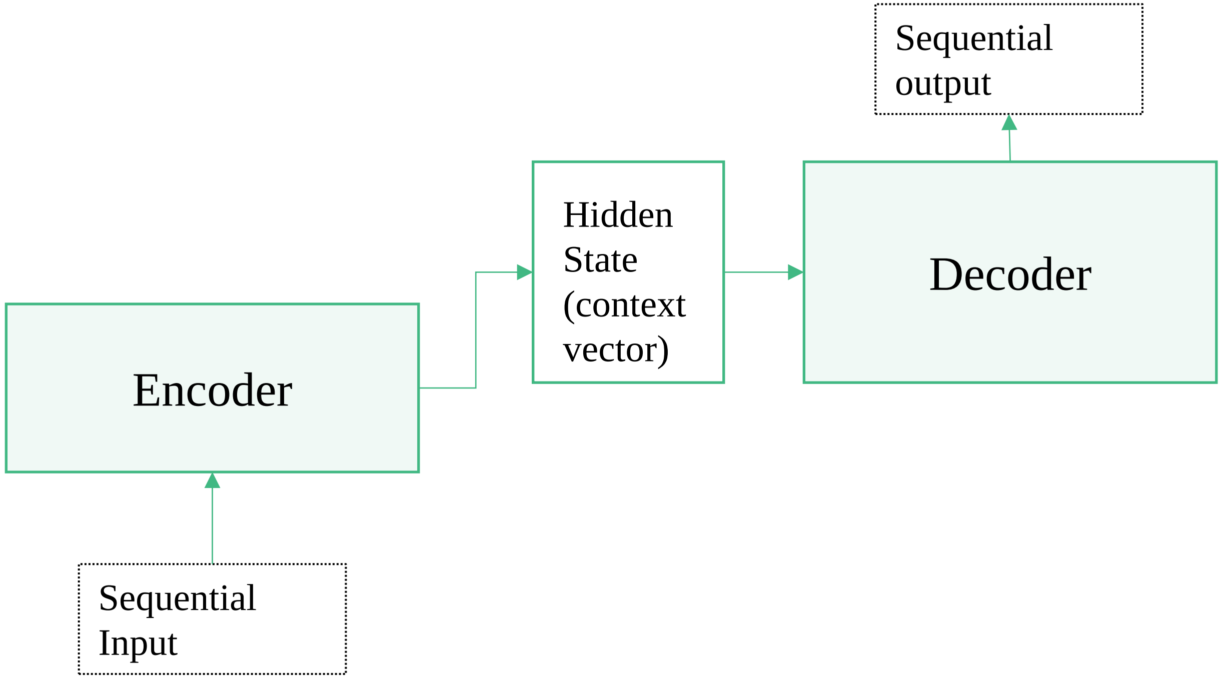
Sequential to Sequential Model in Natural Lang Processing

- Input : Sequential data (sequence of words or sentences)
  Output: Sequential data

- Before Seq2seq
  1. Statistical Methods
  2. Phrase Based Methods

  Unable to handle long term dependencies

- Seq2Seq Model
  Use RNN based Networks for input processing and as well as Output generation


Sequential output

Hidden State (context vector)

Decoder

Encoder

Sequential Input

- Encoder
  - Process the input sequence
  - convert into a fixed size hidden representation

- Decoder
  - use hidden state representation
  - produce target sequence

- Context vector – semantic information and other important information
- Advanced version - **Transformer**

Agenda For Discussion

01 Sequence to Sequence Learning

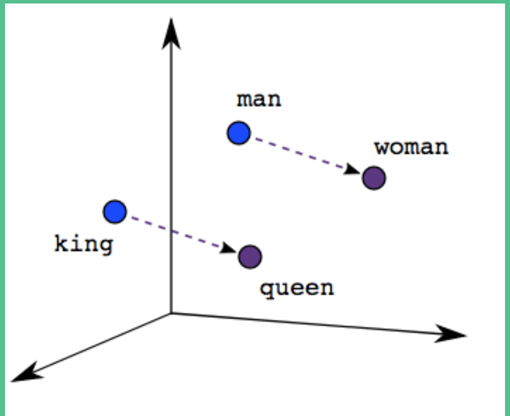
02 Why Attention Mechanism

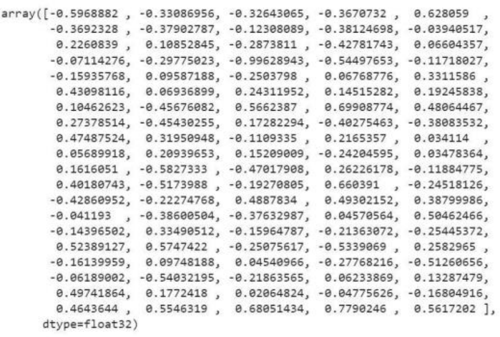
03 Mathematical Intuition Of Attention Mechanism

04 Transformer Architecture

Pre-requisite : Embeddings

- Bridge between humans and computers
  Text to numbers





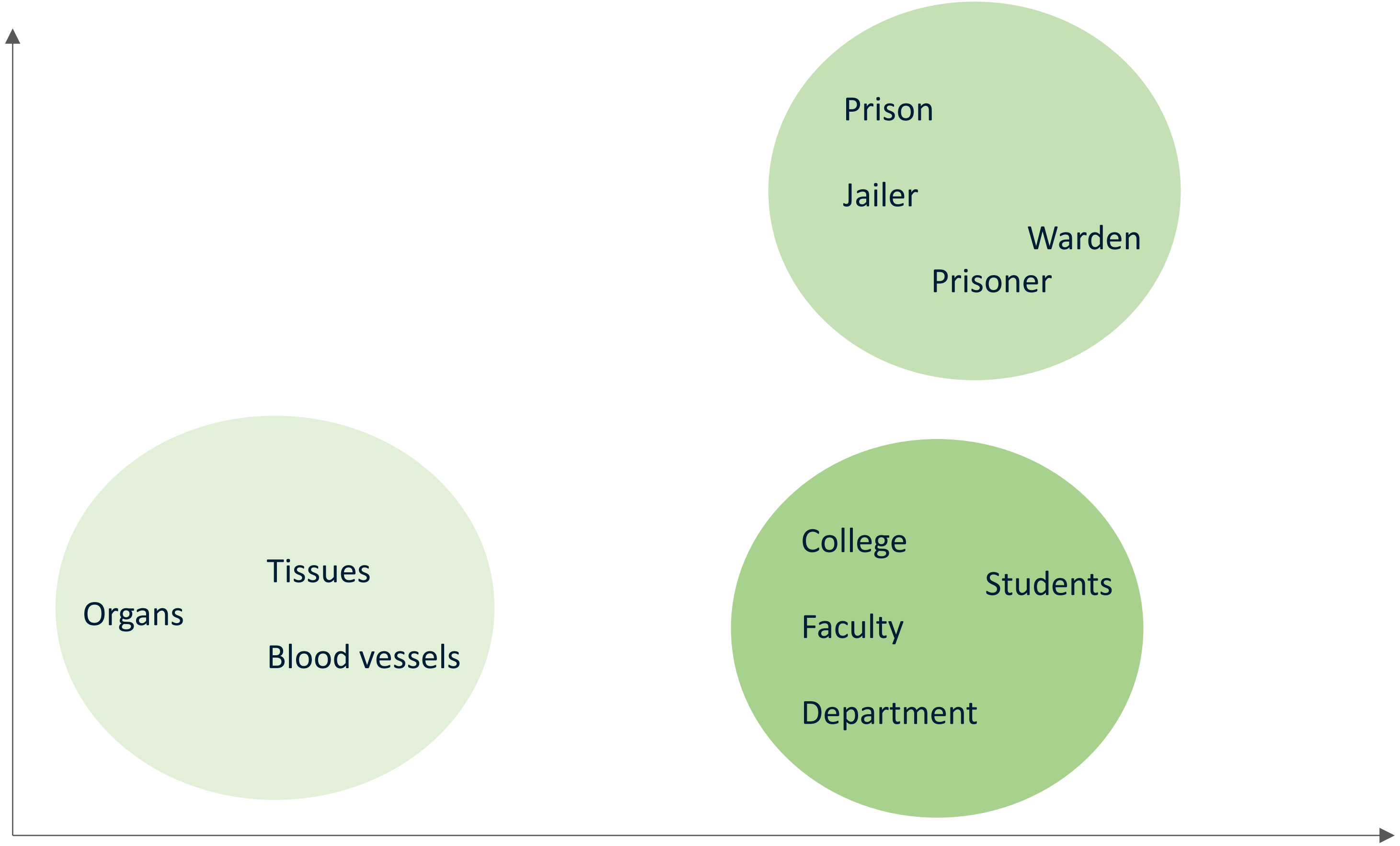
```
array([-0.5968882 , -0.33086956, -0.32643065, -0.3670732 ,  0.628059  ,
       -0.3692328 , -0.37902787, -0.12308089, -0.38124698, -0.03940517,
        0.2260839 ,  0.10852845, -0.2873811 , -0.42781743,  0.06604357,
       -0.07114276, -0.29775023, -0.99628943, -0.54497653, -0.11718027,
       -0.15935768,  0.09587188, -0.2503798 ,  0.06768776,  0.3311586 ,
        0.43098116,  0.06936899,  0.24311952,  0.14515282,  0.19245838,
        0.10462623, -0.45676082,  0.5662387 ,  0.69908774,  0.48064467,
        0.27378514, -0.45430255,  0.17282294, -0.40275463, -0.38083532,
        0.47487524,  0.31950948, -0.1109335 ,  0.2165357 ,  0.034114  ,
        0.05689918,  0.20939653,  0.15209009, -0.24204595,  0.03478364,
        0.1616051 , -0.5827333 , -0.47017908,  0.26226178, -0.11884775,
        0.40180743, -0.5173988 , -0.19270805,  0.660391  , -0.24518126,
       -0.42860952, -0.22274768,  0.4887834 ,  0.49302152,  0.38799986,
       -0.041193  , -0.38600504, -0.37632987,  0.04570564,  0.50462466,
       -0.14396502,  0.33490512, -0.15964787, -0.21363072, -0.25445372,
        0.52389127,  0.5747422 , -0.25075617, -0.5339069 ,  0.2582965 ,
       -0.16139959,  0.09748188,  0.04540966, -0.27768216, -0.51260656,
       -0.06189002, -0.54032195, -0.21863565,  0.06233869,  0.13287479,
        0.49741864,  0.1772418 ,  0.02064824, -0.04775626, -0.16804916,
        0.4643644 ,  0.5546319 ,  0.68051434,  0.7790246 ,  0.5617202 ],
      dtype=float32)
```

- Better the embedding better the model prediction will be

Why Attention Mechanism ?

Limitations with the Word Embeddings

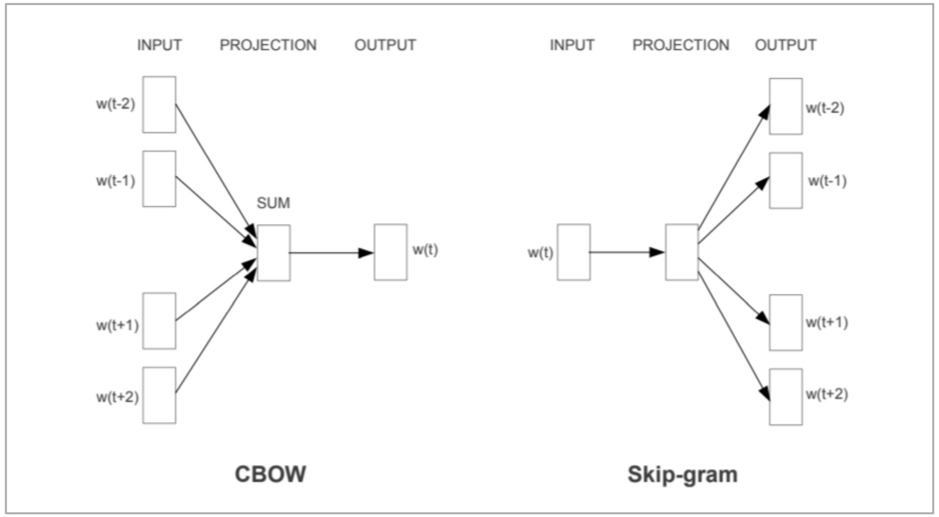
Where will you keep the word Cell ?

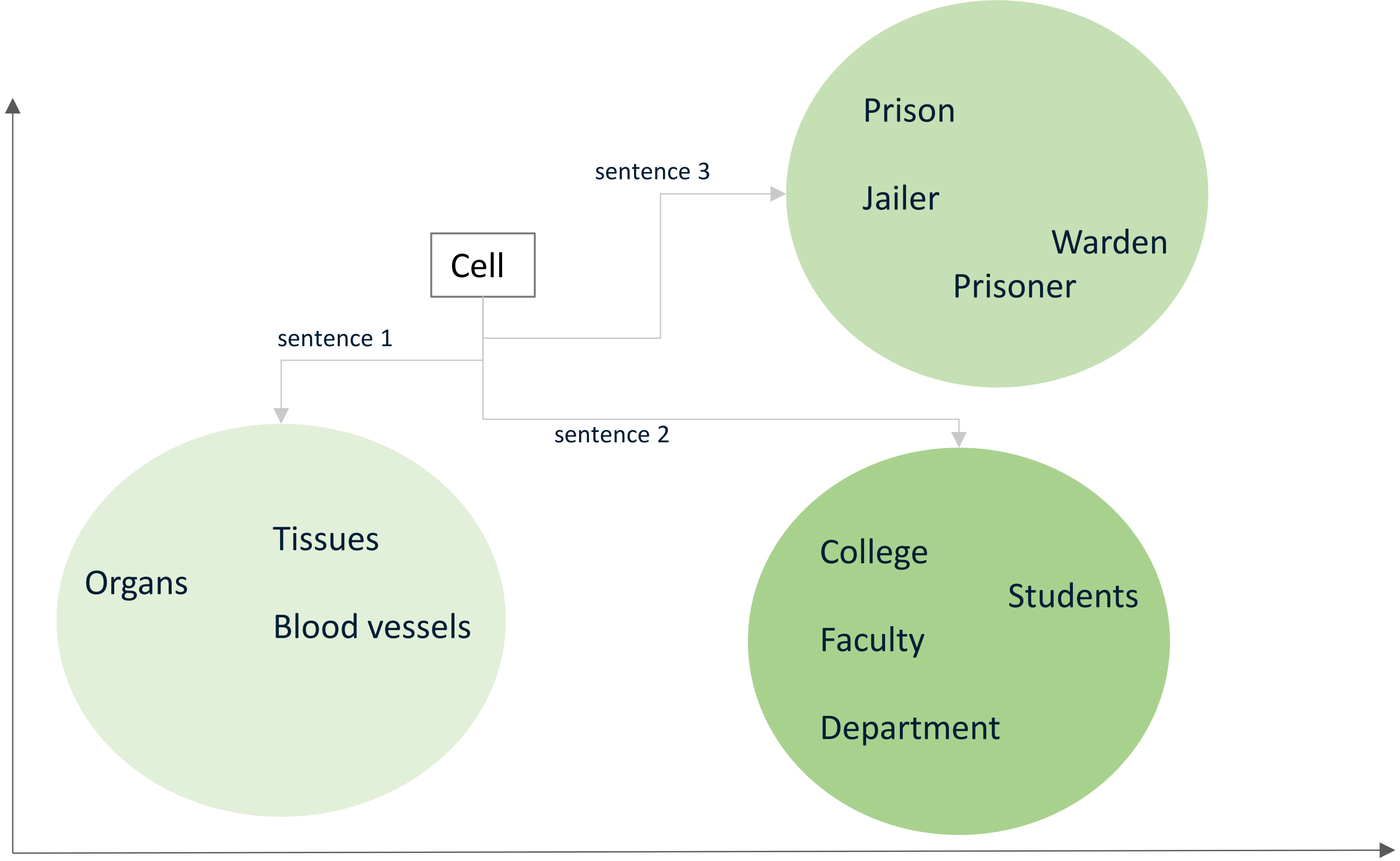


What Attention does ?

Uses the context of the whole sentence to know what we are talking about by adjusting the embeddings values

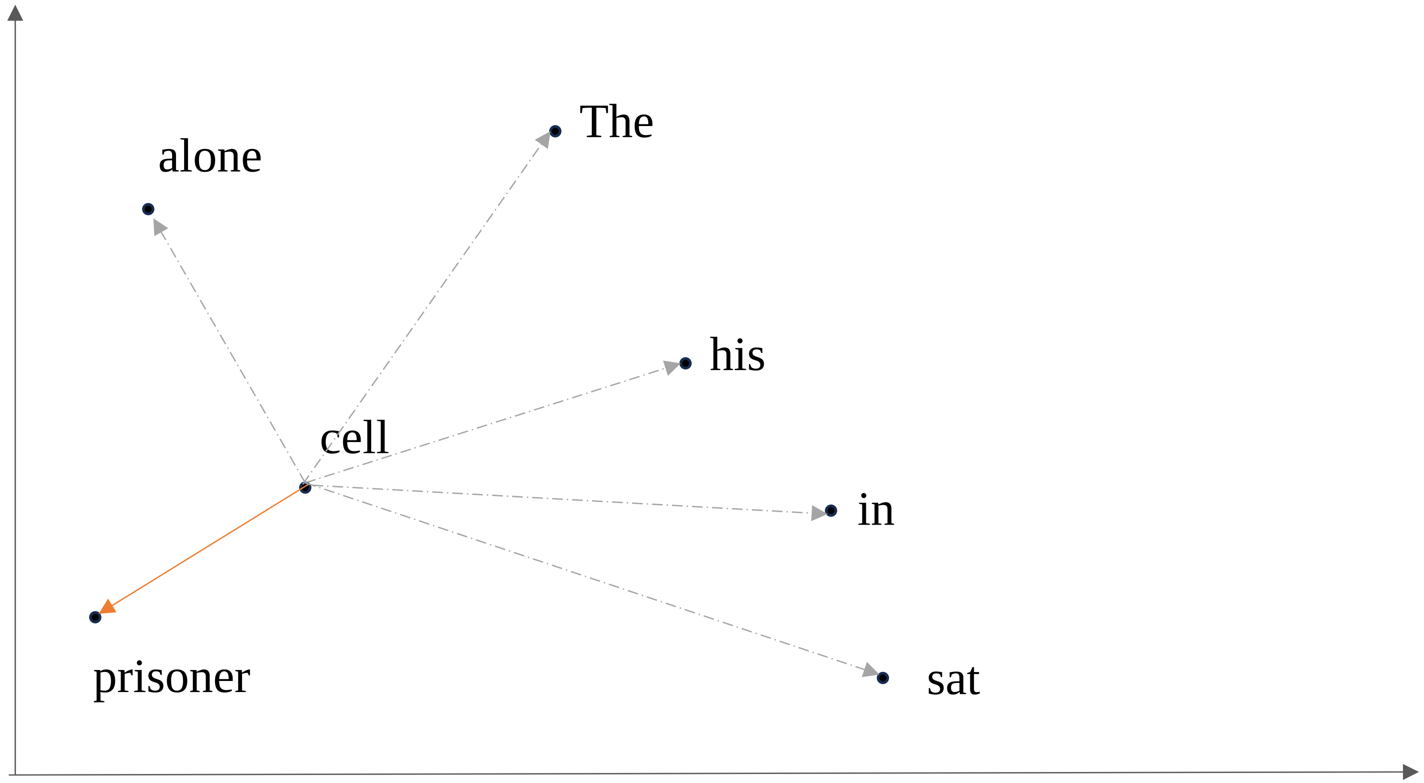
- "The cell houses essential components like the nucleus, mitochondria, and cytoplasm...“
- Each department formed a cell to drive innovation and quick solutions
- The prisoner sat alone in his small cell.

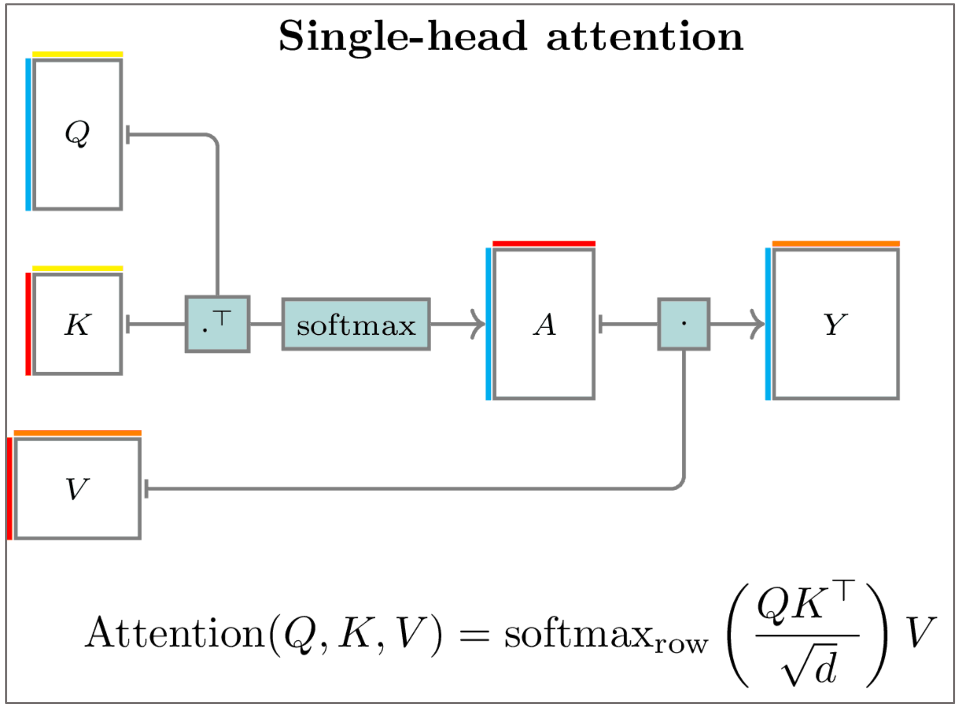




What about other words ?

The prisoner sat alone in his small cell.

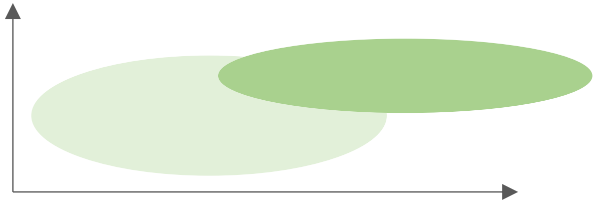


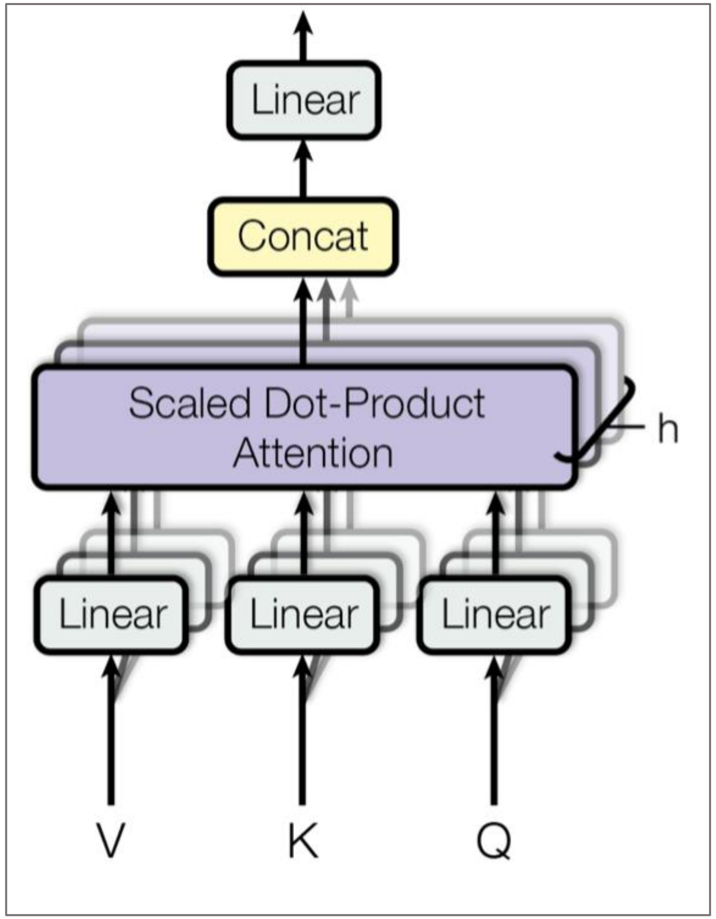


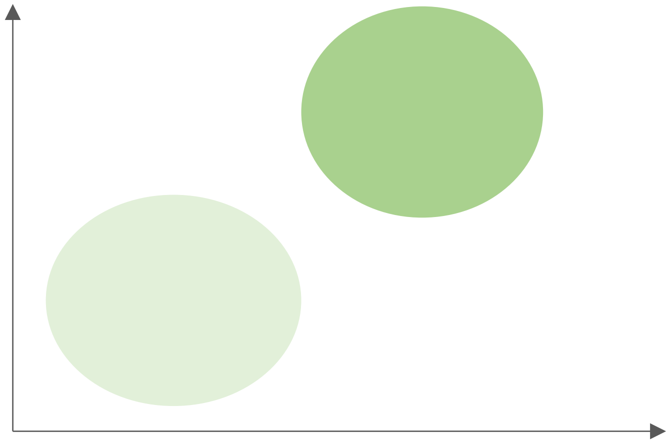
$$\text{Attention}(Q,K,V) = \text{softmax}_{\text{row}}\left(\frac{QK^\top}{\sqrt{d}}\right)V$$

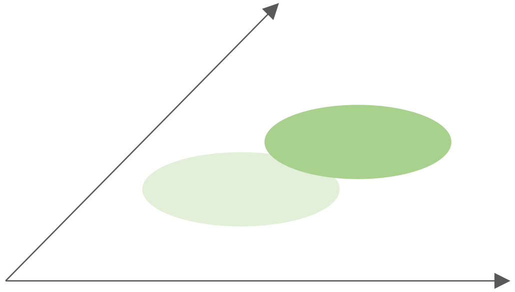
Is Single Head Attention Enough ?

- No its not, You need multiple attention aka multihead Attention
- More specifically multiple embeddings are created
- Among the multiple embeddings we select the embeddings with the best separation of cluster









- Select the best clusters from a set of clusters for clustering
- One embedding is created – Apply linear transformations on it
    - Shear
    - Stretch
    - Rotate
    - Combination of all

Agenda For Discussion

01 Sequence to Sequence Learning

02 Why Attention Mechanism

03 Mathematical Intuition Of Attention Mechanism

04 Transformer Architecture

Mathematical Intuition Of Attention Mechanism

Two Process is involved

1. Similarity capturing between every words
2. Normalization and Exponential

Similarity measures in Euclidian space

- Dot product
- Cosine Similarity
- Scaled Dot Product

$$\vec{A}^{T} = \begin{bmatrix} A_1 & A_2 & A_3 \end{bmatrix} \qquad \vec{B} = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix}$$

$$\begin{bmatrix} A_1 & A_2 & A_3 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = A_1B_1 + A_2B_2 + A_3B_3 = \vec{A}.\vec{B}$$

Eg. The prisoner sat in his cell.

	The	Prisoner	Sat	In	His	Cell	Tissue	College
The	1	0.2	0	0	0	0	0	0
Prisoner	0.2	1	0	0	0	0.75	0	0
Sat	0	0	1	0	0	0	0	0
In	0	0	0	1	0	0	0	0
The	0	0	0	0	1	0	0	0
Cell	0	0.75	0	0	0	1	0	0
Tissue	0	0	0	0	0	0.8	1	0
College	0	0	0	0	0	0.8	0	1

The =

Prisoner = 1 * Prisoner + 0.75 * Cell

Sat =

in =

his =

Cell = 0.75 * Prisoner + 1 * Cell

Normalization and exponentiation

Eg. The prisoner sat in his cell.

	The	Prisoner	Sat	In	His	Cell	Tissue	College
The	1	0.2	0	0	0	0	0	0
Prisoner	0.2	1	0	0	0	0.75	0	0
Sat	0	0	1	0	0	0	0	0
In	0	0	0	1	0	0	0	0
The	0	0	0	0	1	0	0	0
Cell	0	0.75	0	0	0	1	0	0
Tissue	0	0	0	0	0	0.8	1	0
College	0	0	0	0	0	0.8	0	1

The =

Prisoner = 1 * Prisoner + 0.75 * Cell

Sat =

in =

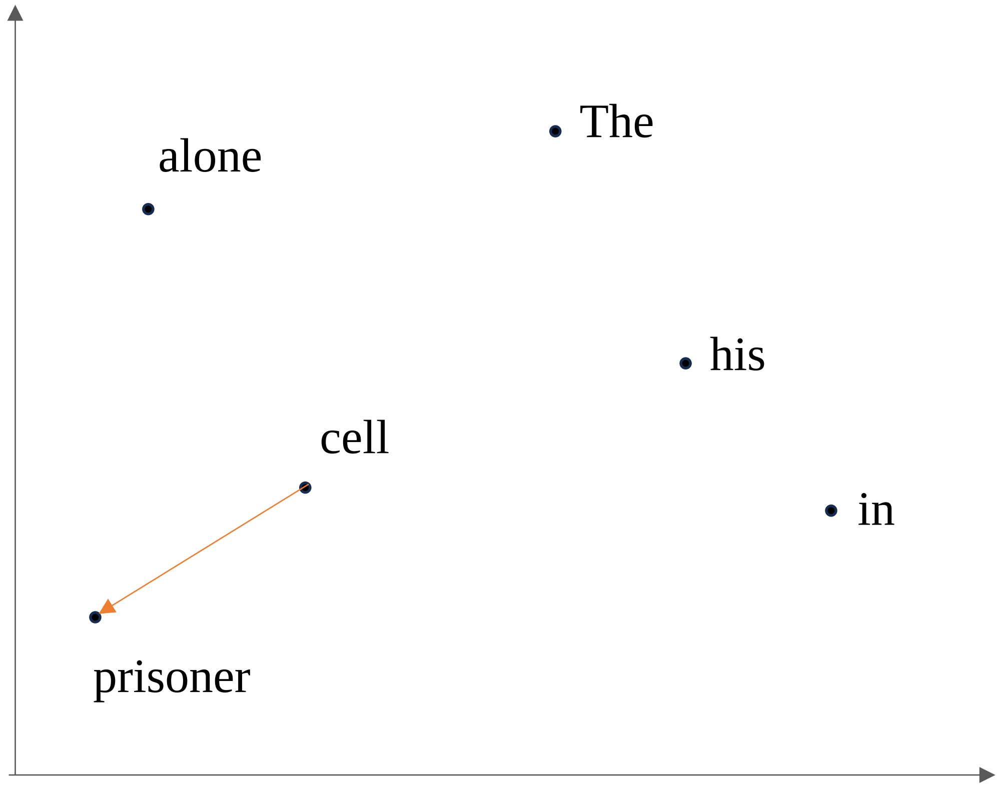
his =

Cell = 0.75 * Prisoner + 1 * Cell

$$\text{Prisoner} = \frac{1 * \text{Prisoner} + 0.75 * \text{Cell}}{1+ 0.75} = 0.58 \text{ Prisoner} + 0.42 \text{ Cell}$$

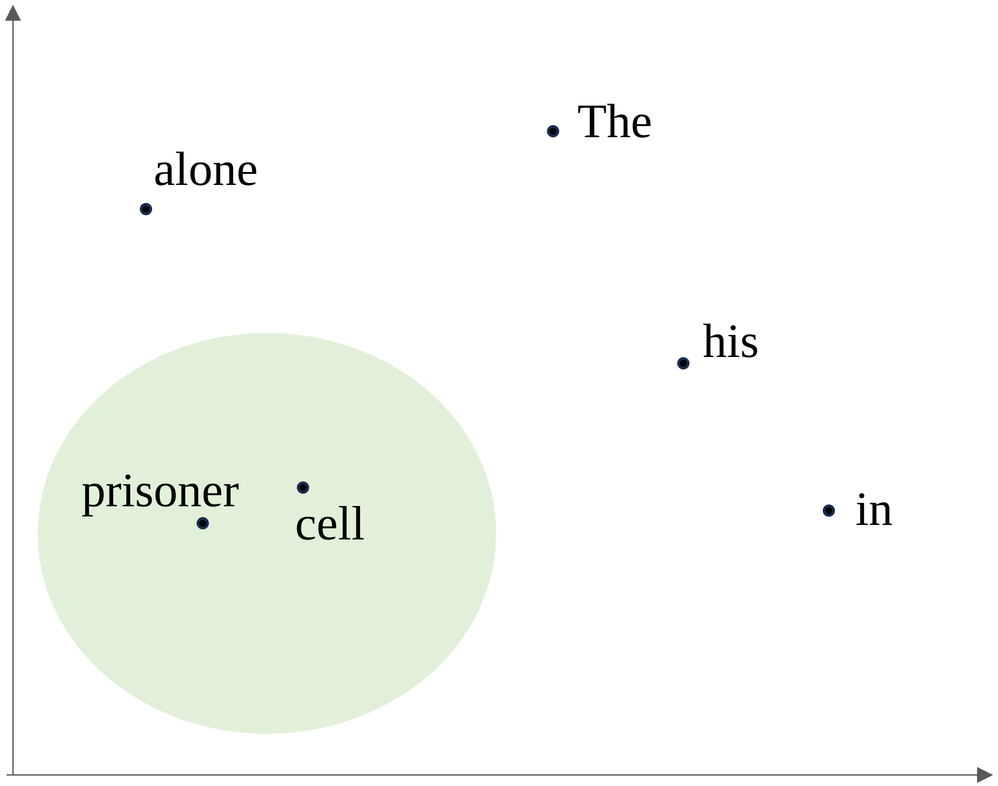
$$\text{Prisoner} = \frac{(e \wedge 1) * \text{Prisoner} + (e \wedge 0.75) * \text{Cell}}{e \wedge (1+ 0.75)} = 0.58 \text{ Prisoner} + 0.42 \text{ Cell}$$

Eg. The prisoner sat in his cell.



Cell = 0.58 Prisoner + 0.42 Cell

Prisoner = 0.42 Prisoner + 0.58 Cell



Agenda For Discussion

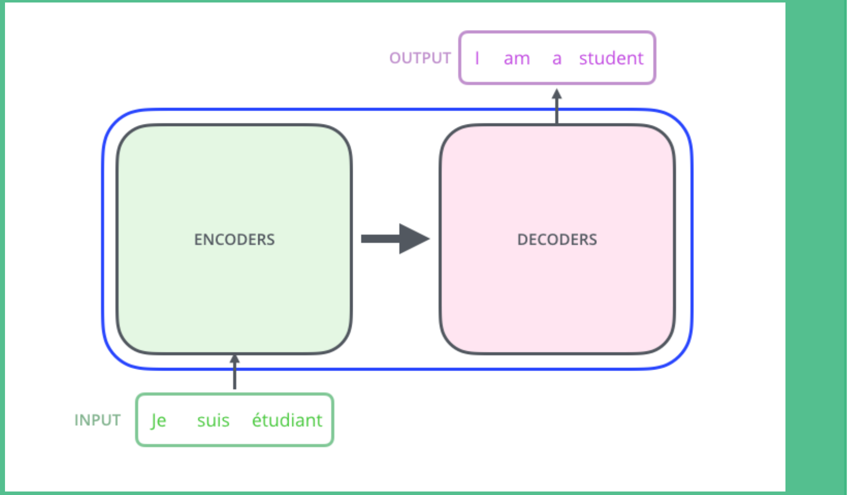
01 Sequence to Sequence Learning

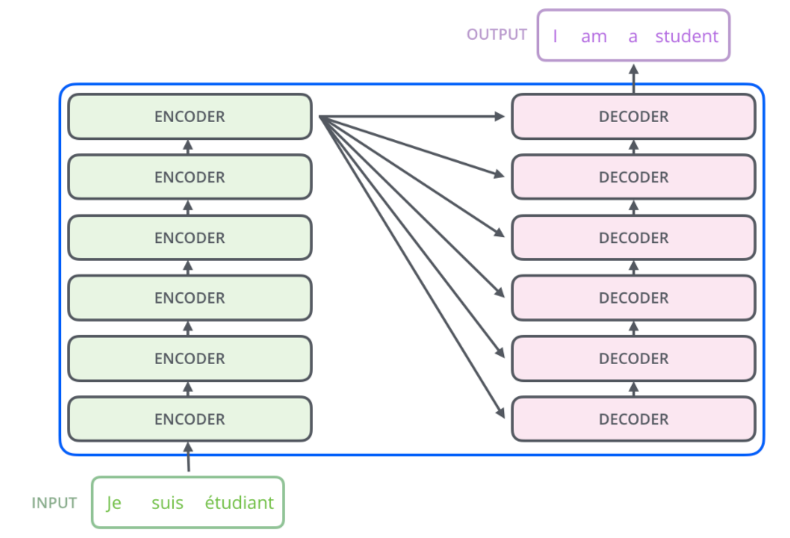
02 Why Attention Mechanism

03 Mathematical Intuition Of Attention Mechanism

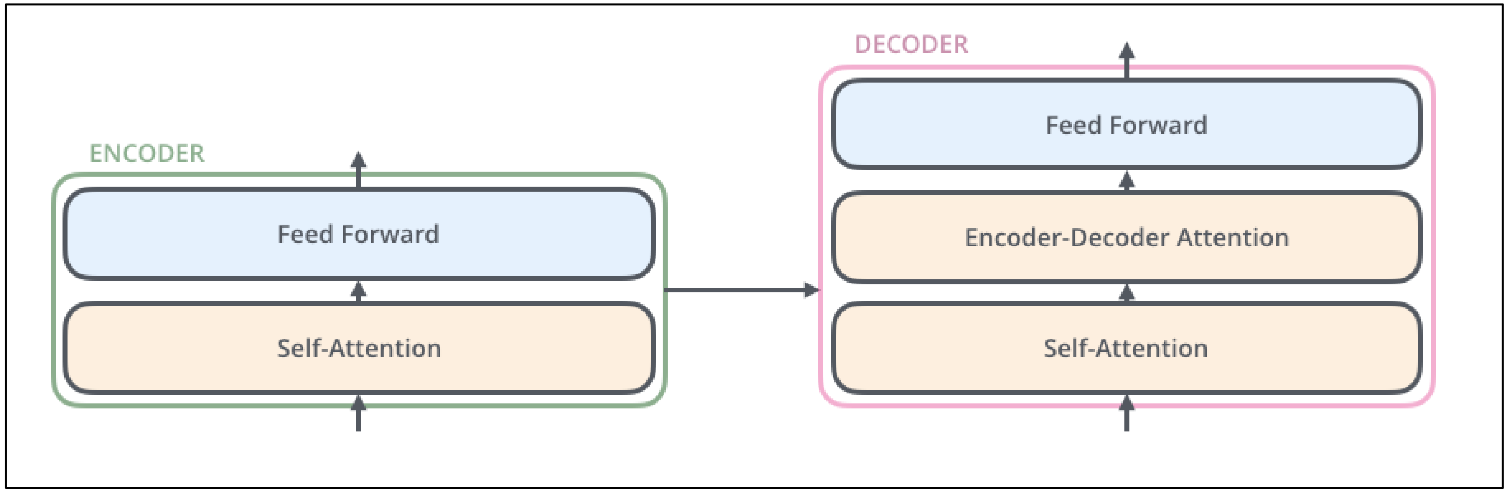
04 Transformer Architecture

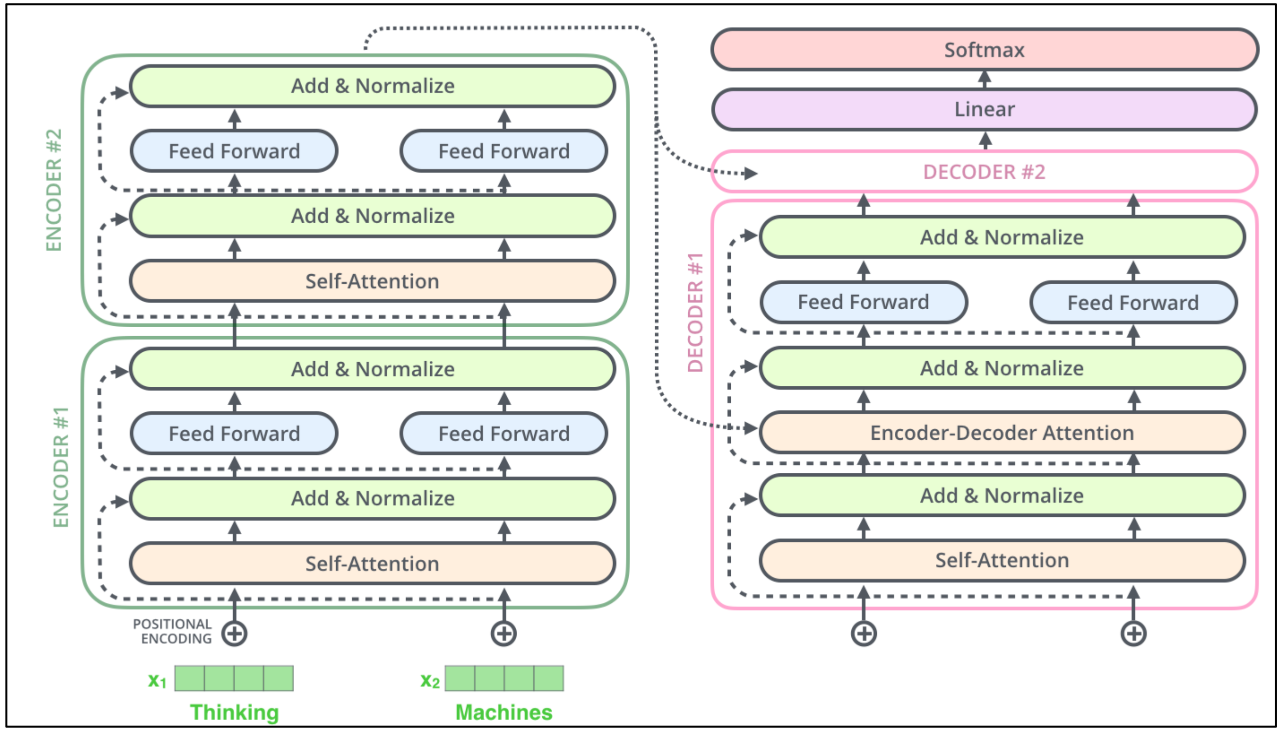
High level Architecture





Whats inside Encoder and Decoder ?





The Transformer Timeline

1986

Word Embeddings

Hinton proposed the idea of “learning distributed representation of words”

- Representing semantics of a word by mapping it into a higher dimension space.
- Such that words that are together have similar meaning.

2013

Word2vec

- This was a breakthrough in NLP
- Embeddings generated were called Neural Embeddings
- These embedding were of lower dimensions also.

2017

Transformer (Attention)

- Update the embedding values
- Updated values will be able to capture wrt to context of the sentence.

Thank You For your Valuable Time.