Gen AI

→ generate new data based on training data.

Unstructured data → Understand pattern & learn. → generate content - Audio, Video, Image, text

Input (Cat or dog.) → discriminitive model → Cat / Dog.

Sample data → generative model → [dotted box] → new data.

- LLM → Large language model (foundational models) use deep learning algorithm to understand natural language.

→ text to text, text to image, text to video. summarizer, translator, code generate.

- pipeline -

1) Data acquisation
2) Data preparation -
3) Feature engineering
4) Modelling
5) Evaluation
6) Deployement
7) Monitoring.

(1) Data acquisation -

→ Available data (csv, txt, pdf, xlsx, docs)
→ Other data (Internet, DB, API, Scrapping)
→ No data (Create own data)
+ LLM to generate data.

Note: less data → Data augmentation.

I am Anu, my name is Anu
+ replace with synonyms.
+ Biagram flip (2टा शब्द अदल बदल)
+ Back translate
+ Add additional noise.

(2) Data Preprocessing —

(i) Cleanup : HTML removal, emoji, spelling correction
(ii) Basic preprocessing
(iii) Advance preprocessing

Basic Preprocessing :
→ tokenization → word
→ sentence

Optional preprocessing :
→ Stop word removal → lametization.
→ steaming → punctuation removal.
→ lower case → language detection

Steaming.

play played playing

Play
root representation.

lametization.

→ root representation readable.

Advance preprocessing —
→ parts of speech tagging
→ Pansing (Parse tree)
→ Corefesrence resolution.
Anamika is a good girl, she is talented also.

(3) Feature Engineering —

→ text Vectorization
→ TF IDF
→ Bug of word.
→ wordzvec
→ one HOT
→ Transformess model.

(4) Modelling —

→ Choose models
open source model
LLm
paid

(5) Evaluation –

(i) Intrinsive → metrices → Gen AI eng.

(ii) Extrinsive – after doing deployement → production

(6) Deployement

monitoring + Retraining

◉ Data Representation –

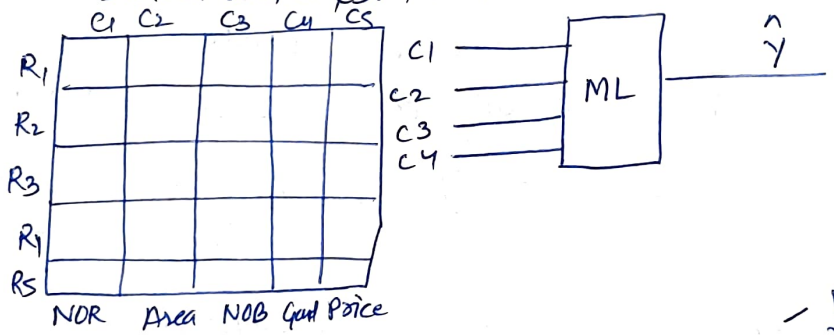
(1) feauture extraction from text / Image –

Vector representation from text / Image

↓

Numbers.

#. ML: tabular data (csv, xlsx)



#. CV: Image, Videos. – Pixels. (0-255) – Black. white

#. Audio: frequency.

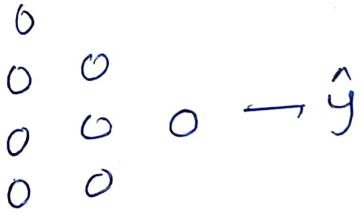
#. text: unstructured data.

→ dimensionily issue

(i) My name is Anu. ×

(ii) I am Anamika. ×

(4) (5) (6)



techniques:

(1) One – Hot encoding

(2) Bag of word (BOW)

One-Hot Encoding

P_1	I am Anamika.
P_2	Anamika is studying.
P_3	Anamika is learning AI.
P_4	She is a smart girl.

(i) Corpus : Entire data.

$P_1 + P_2 + P_3 + P_4$.

→ find unique words.

↓

n = 10 unique words.

I, am, Anamika, is, Studying, learning, AI, She, a, Smart, girl.

I	am	is	Anamika	Studying	lea	AI	She	a	Smart	girl.
1	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0

(word find kro 1 likho, aur 0 dgao)

↓

make a matrix.

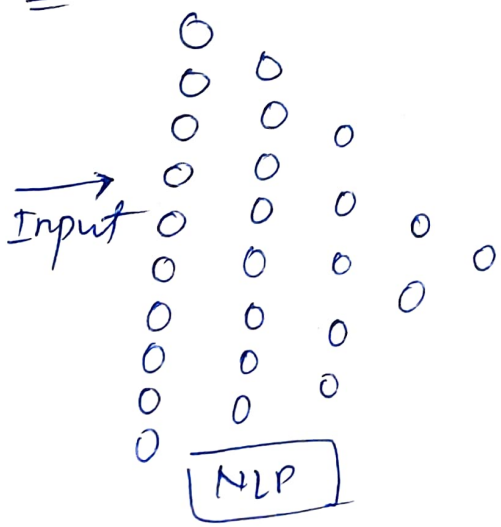
P_1 = for each paragraph make 2-D matrices.

= [[1 0 0 0 0 ..] > finding the unique words in the sentence.
[- - -]

P_2 = . Perform the same operation.

(ii) make a neural network of unique words (n)

=



* Drawbacks

→ Spacity (lots of zeroes)

→ No fixed size.

→ out of vocabulary

→ not capturing semantic relationship.

Bag of Words.

(i) Corpus — whole data (unique)

n = 6

P_1	I am Anu.
P_2	I am a girl.
P_3	I am a Student.

I	am	Anu	a	girl	Student

→ It will see how many times each word present in any sentence and make matrix.

I	am	Anu	a	girl	student
1	1	1	0	0	0
1	1	0	1	1	0
1	1	0	1	0	1

→ pandas
→ sklearn
countvectorizer.

→ Semantic picture not capture.

Transformer Tree

transformers.

Encoder | decoder

BERT
XLM

GPT

Open Source
- mistral
- Llama
- Gemini
- Falcon
- Claude
- StableLm

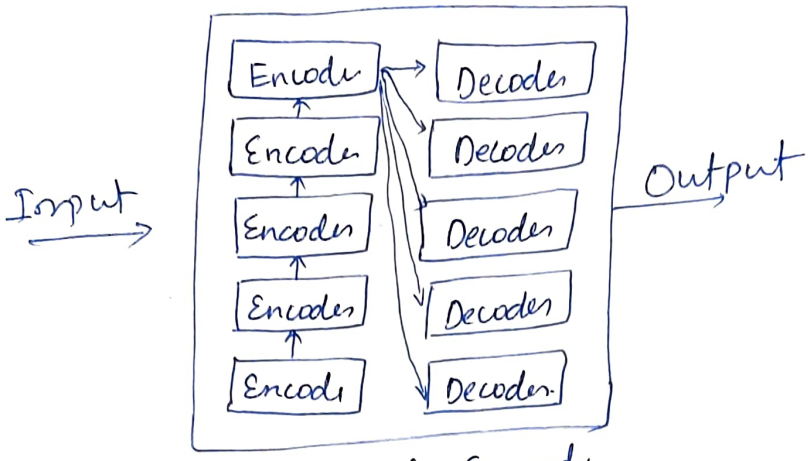
• Prompt Designing.

Zeroshot Learning — using a single command.
Few shot Learning — command + Example.

transformer Architecture.

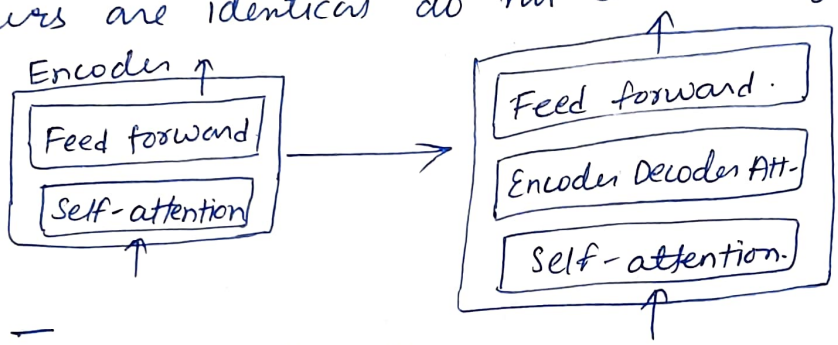
Input → transformer → Output.

Input → Encoder → Decoder → Output



Stack of Encoder & Decoder.

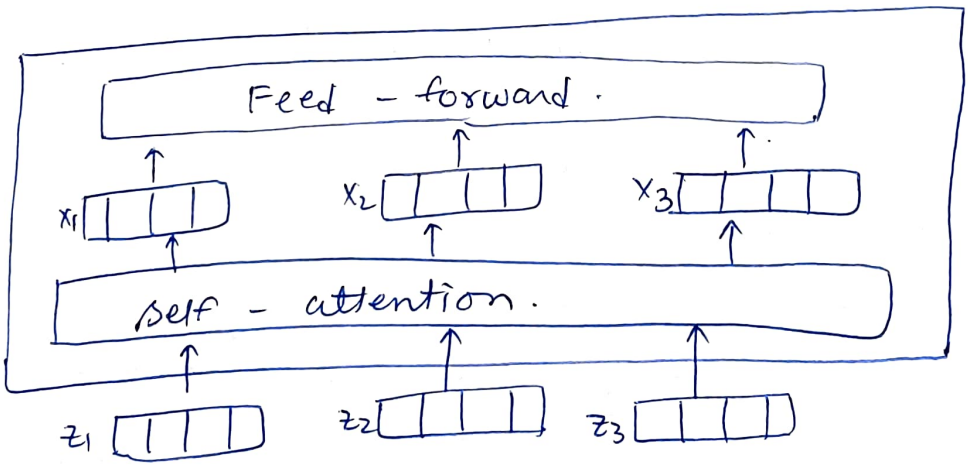
→ Encoders are identical do not share weights -



Working -

I am Anamika.

Word 2 Vec. z_1 z_2 z_3



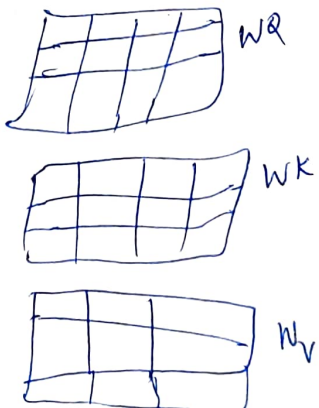
- **Self - attention** - (Single head attention)

Input Embedding: Input x_1, thinking x_2, machines

Queries: q_1, q_2

Keys: K_1, K_2

Value: V_1, V_2

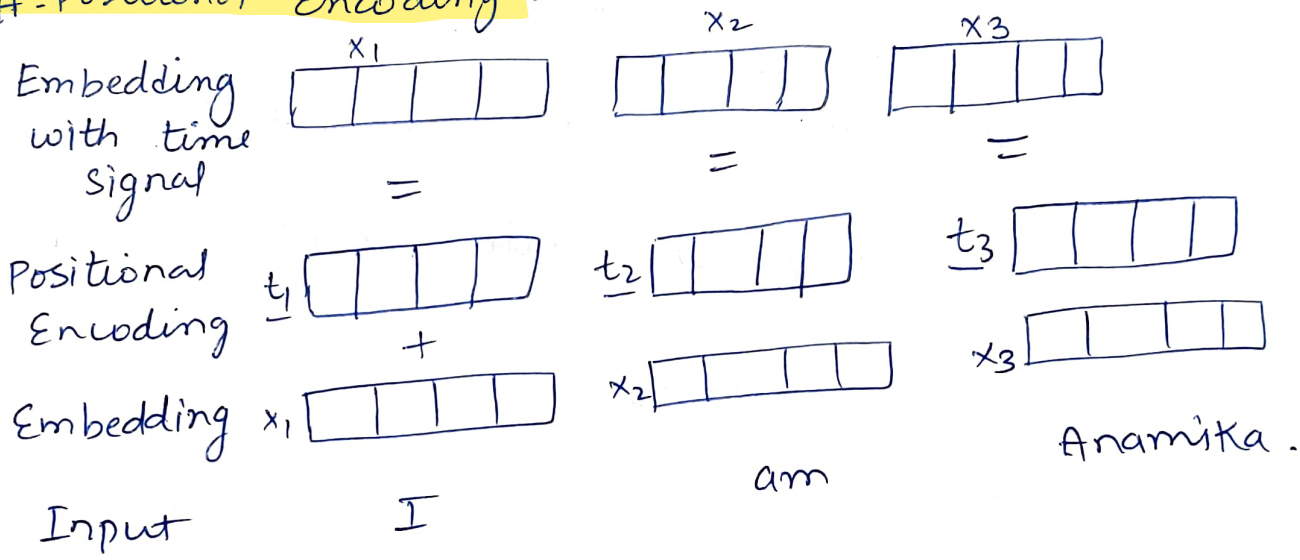


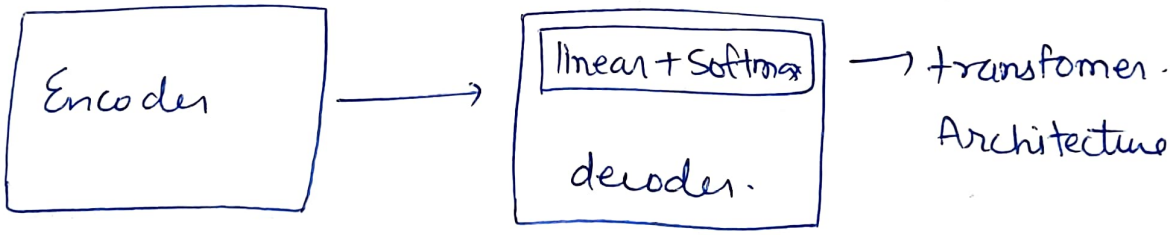
Score $q_1 \cdot K_1 =$ $q_1 \cdot K_2 =$

Softmax → activation function

$$\text{Softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) V = Z$$

#. Positional Encoding -





Linguistics – study of language.

- Clear instruction
- Adopt a persona
- Specify the format
- Avoid leading the answer
- Limit the scope

Vector DataBase

Unstructured data.

High Dimensional data (pdf, Image, word)

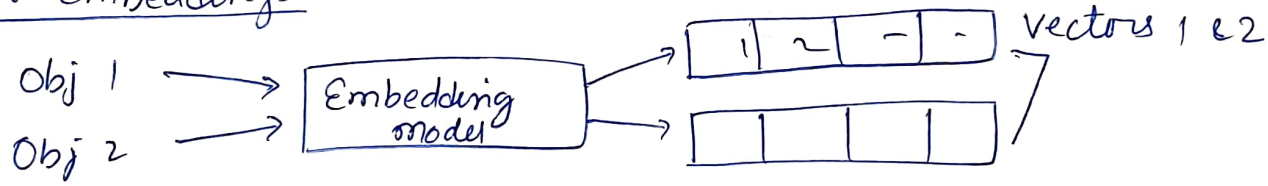
↓

High dimensional ~~data~~ vector 

↓

Vector database

- Vector embeddings –



Ex- Unstructure data as text –

Embedding model.

Feature	King	Queen	man	Women	Monkey
gender	1	0	1	0	1
Wealth	1	1	0.5	0.3	0
Power	1	1	0.5	0.3	0
Weight	1	0.5	0.5	0.3	0.2

male → 1
female → 0
yes → 1
NO → 0

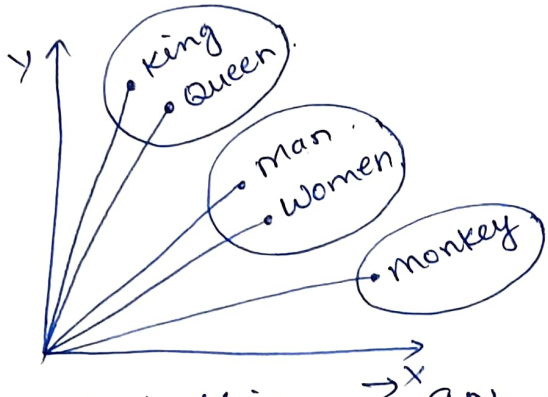
King => [1, 1, 1, 1]

Queen => [0, 1, 1, 0.5]

man => [1, 0.5, 0.5, 0.5]

Women => [0, 0.3, 0.3, 0.3]

Monkey => [1, 0, 0, 0.2]



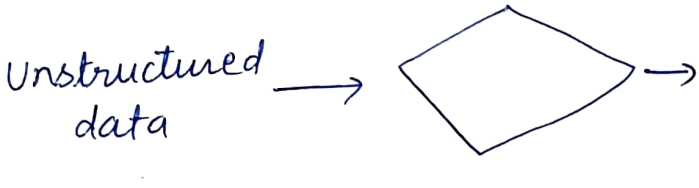
Prince → King, Queen.

→ Understand the essance of other words.

→ Open AI Embedding → gpt model
→ Hugging face Embedding → open source.
→ llama 2 Embedding → facebook
→ Google palm embedding → Google

Vector DB
- → Pinecone
- Neo4j
- Weaviate
- Chroma DB
- FAISS

- **Vector Indexes –**



→ It will check the new data is falling under which cluster and then search only in those indexes.

→ - Long term memory for LLMs.
- Use for Semantic Search based on meaning of conte
- Similarity Search.
- Recommendation engine as well.