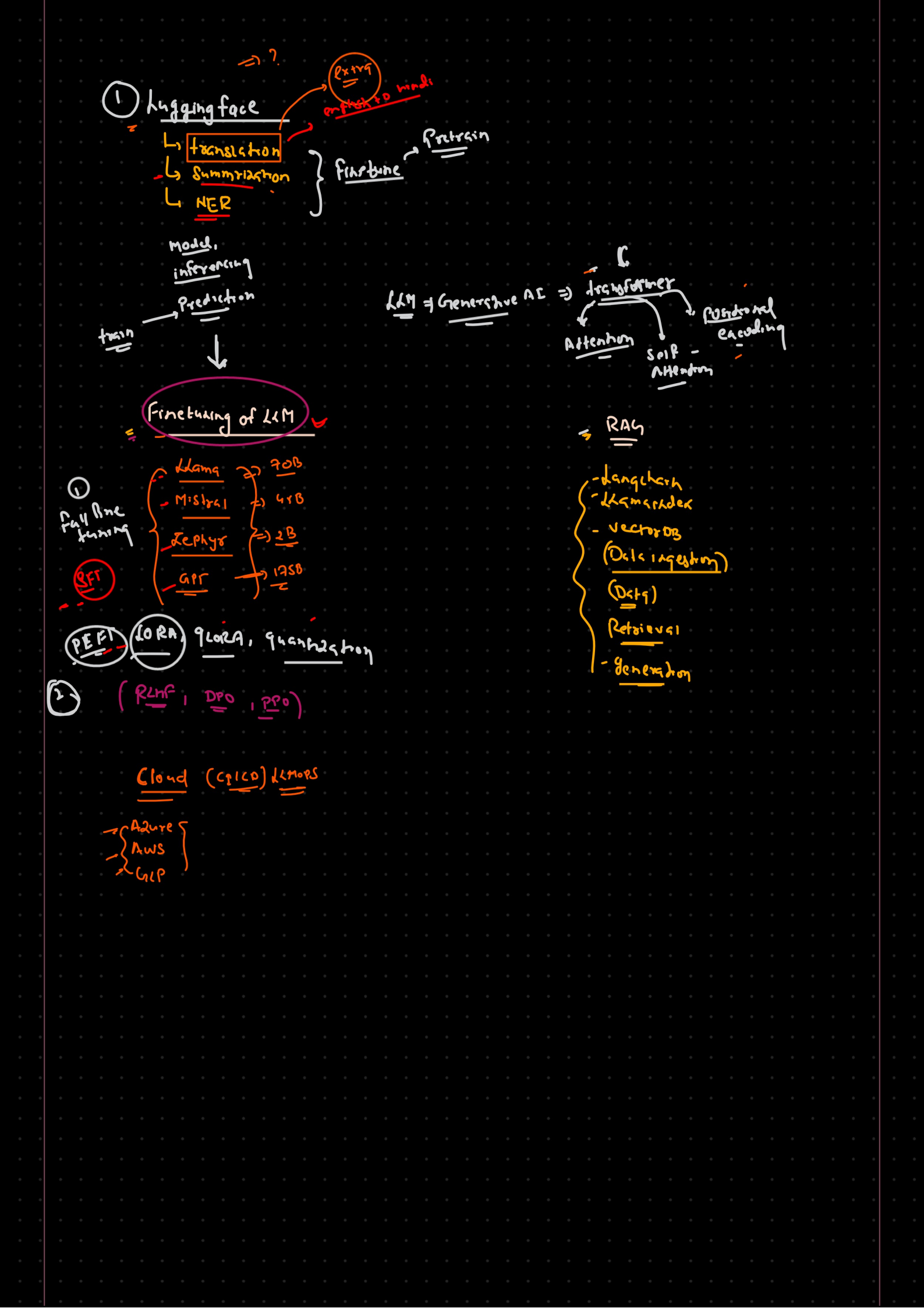
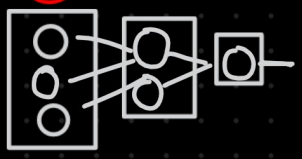

1 huggingface
translation
Summarization
NER
finetune
Pretrain
Extra
model, inferencing
Prediction
train
Finetuning of LLM
LLM = Generative AI
transformer
Attention
Self - Attention
Positional encoding
RAG
-langchain
-llamaindex
- vectorDB
(Data ingestion)
(Data)
Retrieval
- generation
Llama
Mistral
Zephyr
GPT
70B
4TB
2B
175B
full fine tuning
SFT
PEFT
LORA
QLORA, quantization
2
(RLHF, DPO, PPO)
Cloud (CI/CD) LLMOPS
Azure
AWS
GCP

Transformer ⇒ Self attention, Contextual embedding

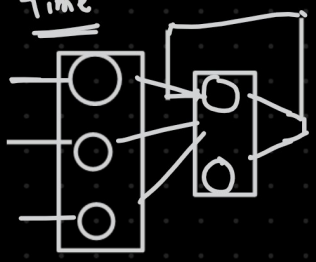
CNN — Image processing

① NN



② CNN

③ RNN — Recurrent → Repetation time

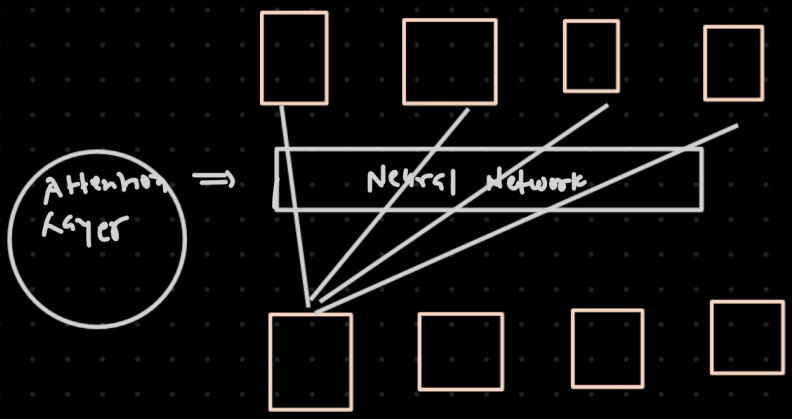


⇒ LSTM, GRU

mapping → input → { Many to one, One to many, many to many } → output

⇒ Encoder - Decoder (many to many) Asynchronize (Diff length)
(LSTM & GRU)

⇒ Encoder Decoder with attention



(capture the info.) → Semantic meaning

Data → Model → mathematical eq.

{ Summarization, generation, classification }

{ Text, Image, tabular }

NLP — text → Model

Number

Embedding & Encoding ⇒ Numeric form

{ OHE, BOW, N-grams, TF-IDF } ⇒ my data → vector ⇒ Numbers

OHE ⇒ → Sunny Sevita

[[1 0] [0 1]]

BOW ⇒ Count the frequency

N-gram ⇒ Count the frequency (pair) N = 2

TFIDF ⇒ TF * IDF ⇒ BM25 ← latest ← Google

TF ↑ term frequency

IDF ↑ Counting

GLOVE ⇒ Matrix factorization

Word2vec ⇒ embedding ⇒ weight of layer (H → o/p)

NN → F1 F2 F3 F4 F5 → 5-D

King → [0.2, 0.9, 0.1, 0.5, 1]

Queen → [0.3, 0.5, 0.1, 0.9, 1]

(words)

→ Semantic meaning of data

Problem with word2vec

1. An Apple a day keeps doctor away.
2. Apple is healthy
3. Apple is better than orange
4. Apple makes great phone.

0.3

0.7

Word2vec

Apple = [Tech, fruit]

[0.2 0.8]
[0.1 0.9]
[0.3 0.7]
[0.9 0.1]

[0.3 0.7]

Average meaning of Apple

[0.99, 0.01]

[0.01, 0.99]

Context

Dynamic

Contextual embedding

Self attention

Apple launched a new phone while I was eating an apple

[0.3, 0.7] [0.3, 0.7]

Dynamic embedding / contextual embedding

like human

(DE & CE)

Embedding

Self attention

Sentence → embedding → Static & Avg → Dynamic

Word2vec

Data → embedding → Self Attention → Contextual embedding → NN

Attention all your need

one head K, Q, V, linear transformation, softmax

multi head → best to best from Data

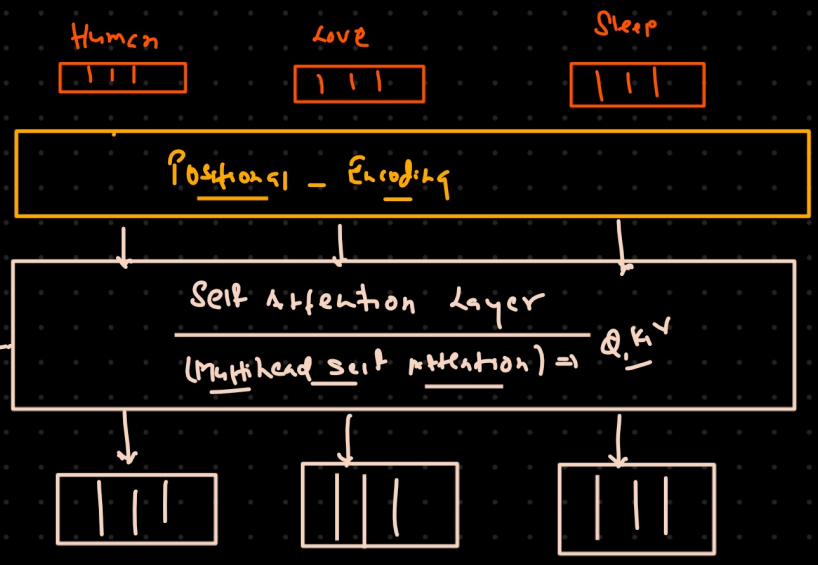
transformer

Money bank → Diff → 0.6 | 0.2 | 0.1 | 0.7

River bank → Diff → 0.6 | 0.2 | 0.1 | 0.7

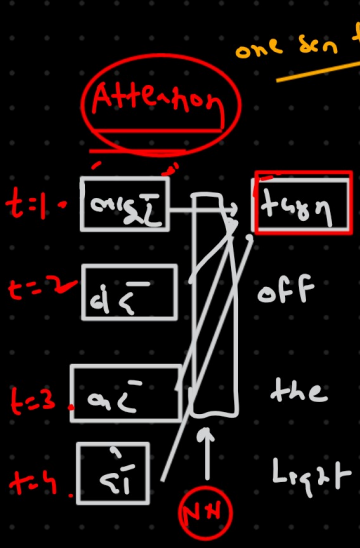
Context

eg:- Human Love Sleep



Query x Key = Value

$\triangleq$ $\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$



Example

(Q,K,V)

Self Attention

① Money bank grows

② River bank flows

Bank → ① ②

Seat | bank ⇒ 0.3x Money + 0.3x bank + 0.1x grows

(ω) (ω) (ω)

Seat bank → 0.5 x River + 0.3 x Bank + 0.2 x flows

Self attention

e(Money) → 0.7 money + 0.2 bank + 0.1 grows

e(bank) → 0.25 money + 0.7 bank + 0.05 grows

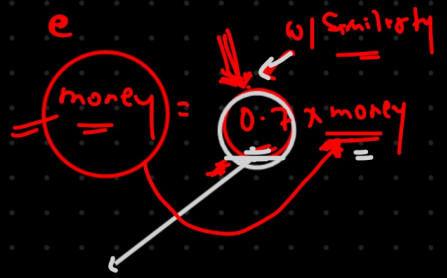
e(grows) → 0.1 money + 0.2 bank + 0.7 grows

e(River) → 0.8 River + 0.15 bank + 0.05 flows

e(bank) → 0.2 River + 0.7 bank + 0.1 flows

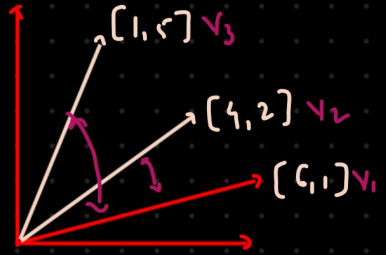
e(flows) → 0.4 River + 0.01 bank + 0.59 flows

Seat → SelfAttn → embedding



$e_{bank} \cdot e_{money}$ Similarity ⇒ Dot Product

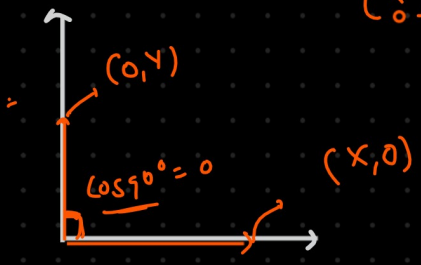
initially



$V_1 \cdot V_2 = [6,1]\begin{bmatrix}4\\2\end{bmatrix} = 6\times4 + 2\times1 = 26$

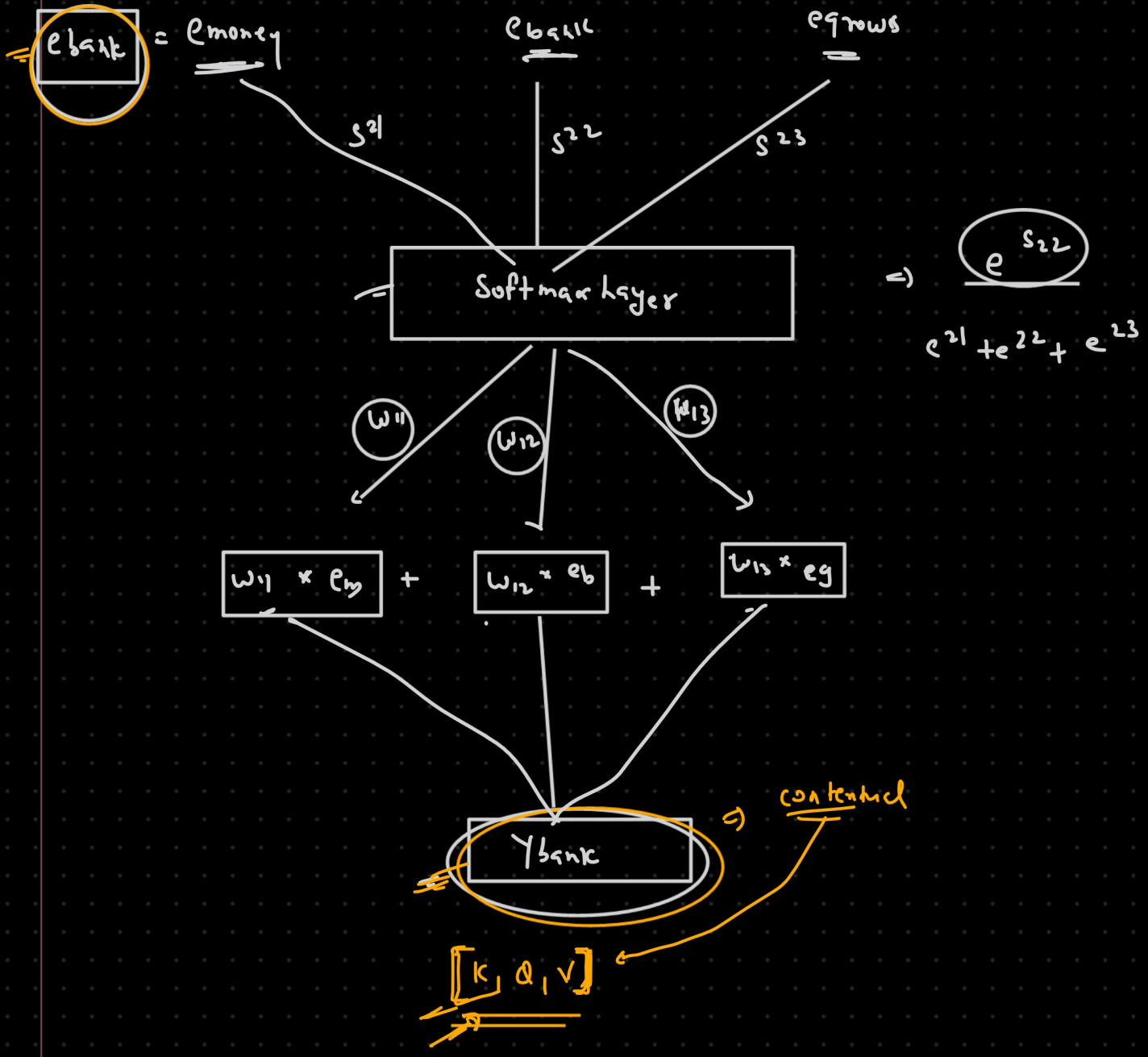
$V_1 \cdot V_3 = [6,1]\begin{bmatrix}1\\5\end{bmatrix} = 6\times1 + 1\times5 = 11$

$[x\ 0]\begin{bmatrix}0\\y\end{bmatrix} = x*0 + y*0 = 0$



$$e_{bank} = [e_{bank} \cdot e_{money}] * money + [e_{bank} \cdot e_{bank}] * bank + [e_{bank} \cdot e_{grows}] * grows$$

Similarity score (money) | Similarity score (bank) | Similarity score (grows)



Point — (1). Parallel operation

(2) there are no Parameter → trainable → NN

{ How are you → આપ કેવા છો! }

{ Piece of Cake → કેક ના ટુકડા ✓
→ ઘણું જ સહેલું ✓

ambiguity

task specific → Reading → Parameter (NN)

K q V

D = {a:1, b:2}

key

query → (2) → D[a] → (1)

value (3)

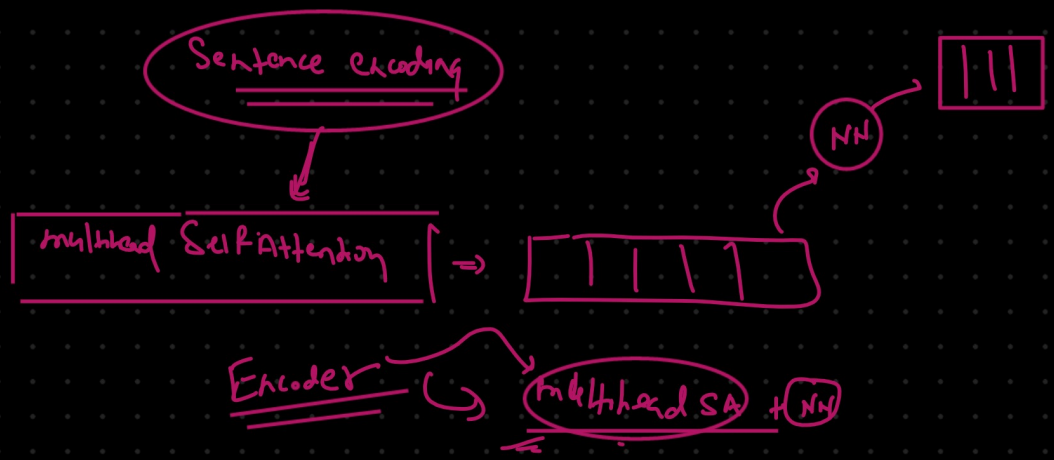
$K \times q = V$

Linear transformation, multihead attention, $\text{softmax}\left(\frac{q \times k}{\sqrt{d}}\right) \cdot V$

1.1:30

Positional encoding

Python → Code


Sentence encoding
multihead SelfAttention
NN
Encoder
multihead SA + NN