Transformer → Interview

↓

(3 - 2 Min)

→ Transformation Architecture →

1) Long term Depencies
2) Context Embedding
3) Paralism
4) Muli Modality

Self Attention ⇒

→ ANN → RNN → LSTM → GRU →

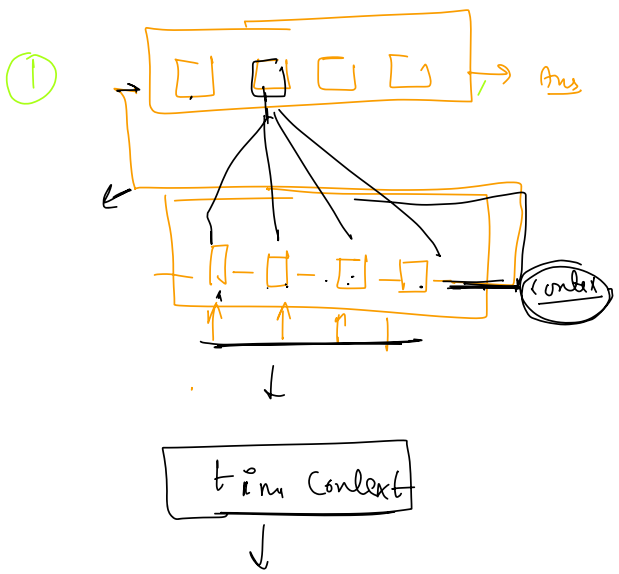
Encoder – Decoder

↓

Attention Mechism

↓

Self Attention



long term context

↓

Attention Mechism

Book

work →

time Context

↓

① Long term · Dependetie

Encoder – Decoder → (Long) – Attention Mechism

↓

Long ✗

↓

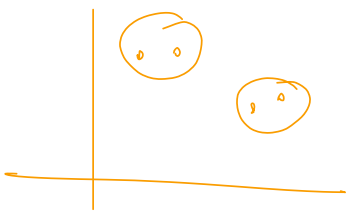
X Parallelism

Self Attention

① Parallelism

Word Embedding

Text → Embedding



Amazon

Company → 70%

River — 30%

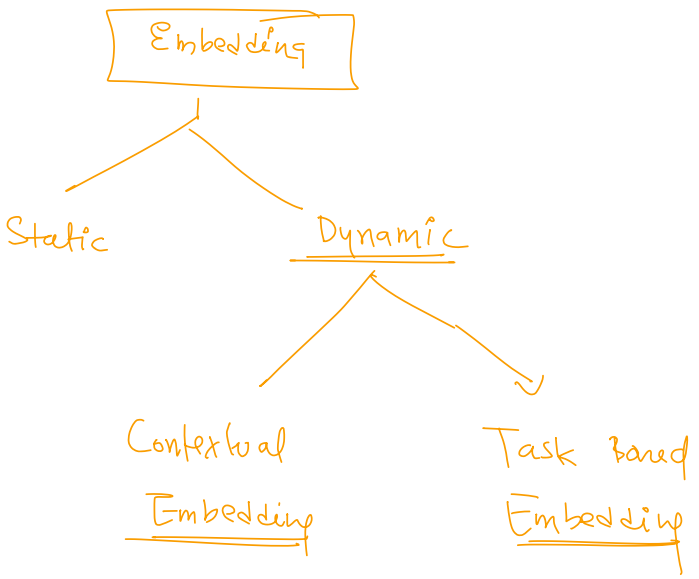
Amazon → Company (X)

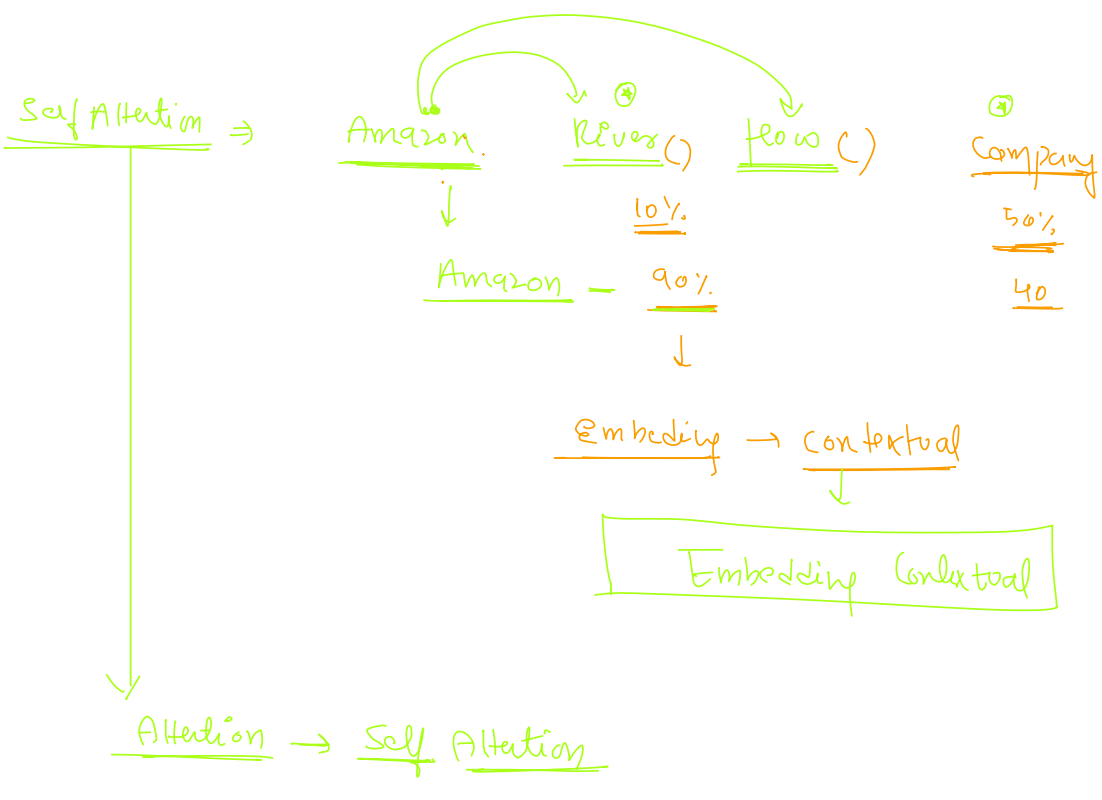
70 > 30

Amazon → company.

Self Attention

← 1) parallism
2) Average Embedding



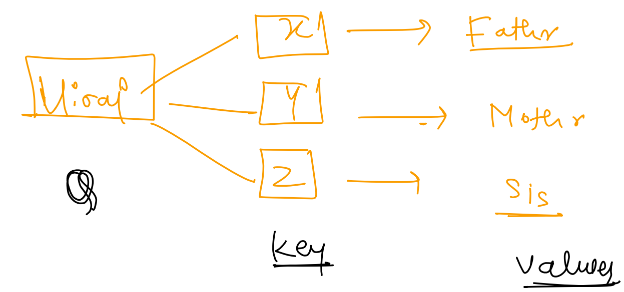


Cosine similarity = $\cos\theta = \frac{A \cdot B}{|A||B|}$

Amazon ⇒ 0.3 · (River) + 0.7 · Amazon + 0.2 · flows

① similarity

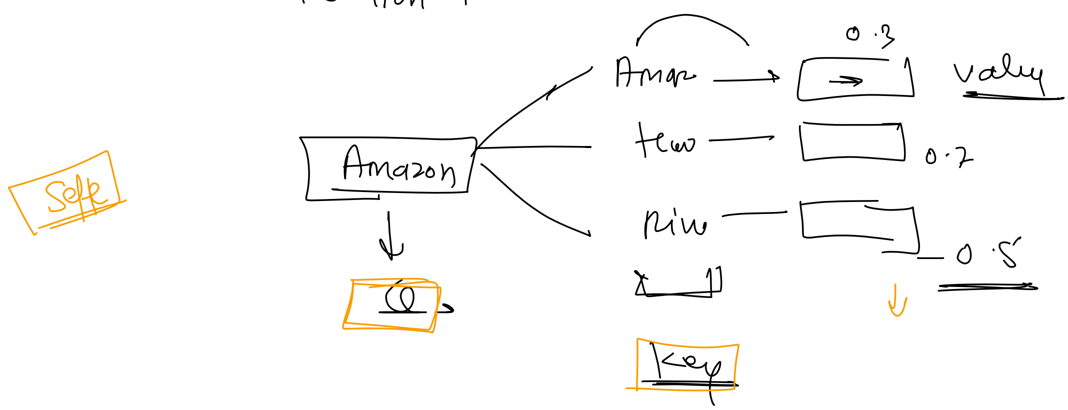
Ex :



कितना →

किसके → 0.3

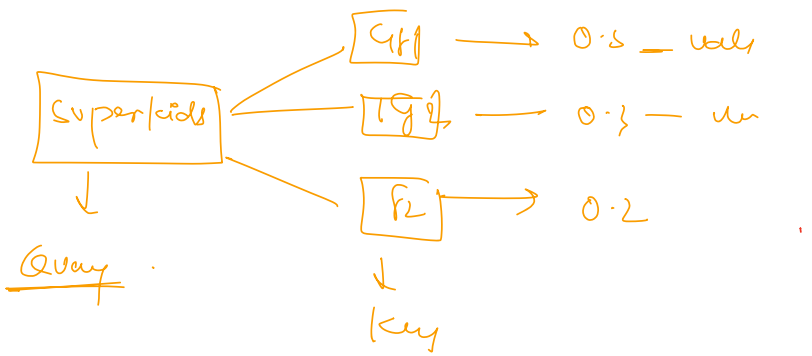
Relation →



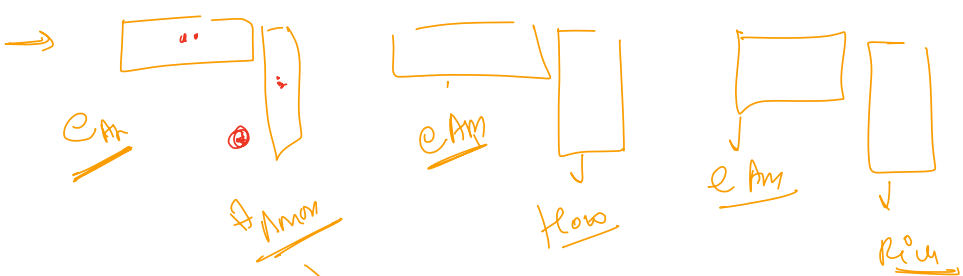
MCQ → Q =

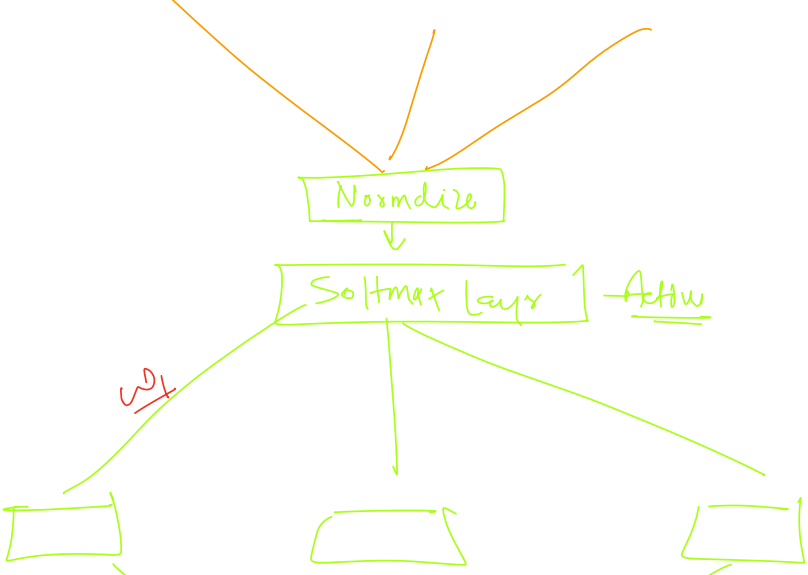
MCQ op → — — — — .

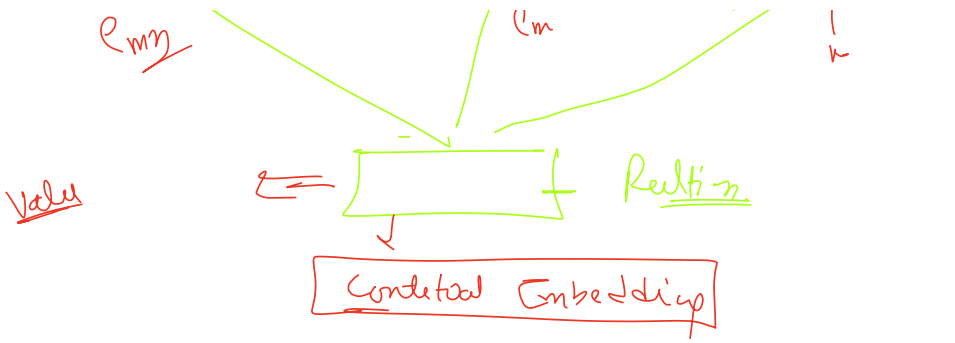
Imp Vals



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







Amazon → Amazon + Am (River) + Am (flow)

Contextual Embedding

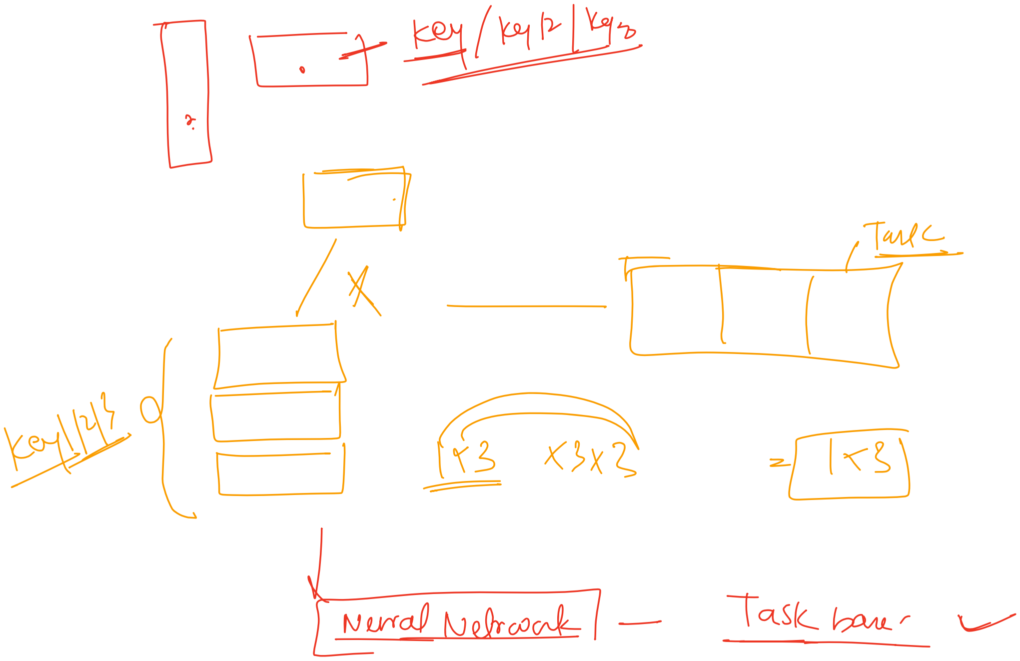
④ Idiom → Kick the Bucket →

① He painted House that is located near the park
② He did painting while being near park

Contextual Embedding → Fail

Task Based Sentence

Task Based Embedding



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

Max/min

dimension of vector

Q →

$\sqrt{d_k}$ = Stability / scaled

Q → why self Attenti called = Scaled Dot Product =

$\sqrt{d_k}$

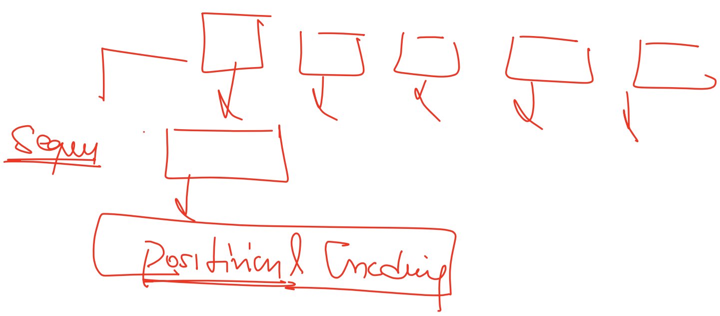
Normal

Softmax $\left(\frac{Q \cdot K^T}{\sqrt{d_k}} \cdot V \right)$

training

Self - parallelism ⟹ (1) ← (2) parall

Ans



1) Normalisation
2) Batch Normati.
(3) Layer Normalisation

Gallt

Vanila

2) Exp
3) Convex
4) Stability