The What

Self attention computer x

NLP → words → numbers → vectorization

OHE

1) mat cat mat
2) cat rat rat

	mat	cat	rat
mat	1	0	0
cat	0	1	0
rat	0	0	1

num → [1 0 0] [0 1 0] [1 0 0]

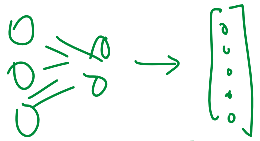
BoW

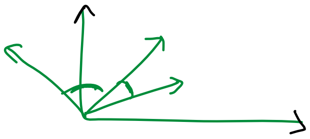
	mat	rat	cat
S1	[2	0	1]
S2	[0	2	1]

Tf Idf → word embedding → semantic meaning

number

training wiki → → 5 dim





cricket [0.2 0.9 0.9 0.90 1] n-dim vector 256, 512 64

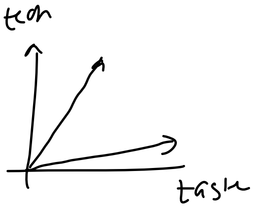
king → [0.9 0.1 1 0 0.9] → similar

queen → [0.9 0.2 0.4 1 0]

similar

royalty

meaning → vector



Apple → vector

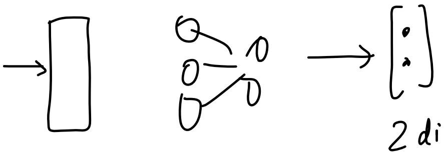
The problem of "Average Meaning"

→ dataset

1) An apple a day keeps the doctor away
2) Apple is healthy
3) Apple is better than orange
4) Apple makes great phones

10000
9000 fruit
1000 Phone

9000 tech
1000 fru



[x y]

taste technology

2 dim

[0.8 0.2] → [0.9 0.3]

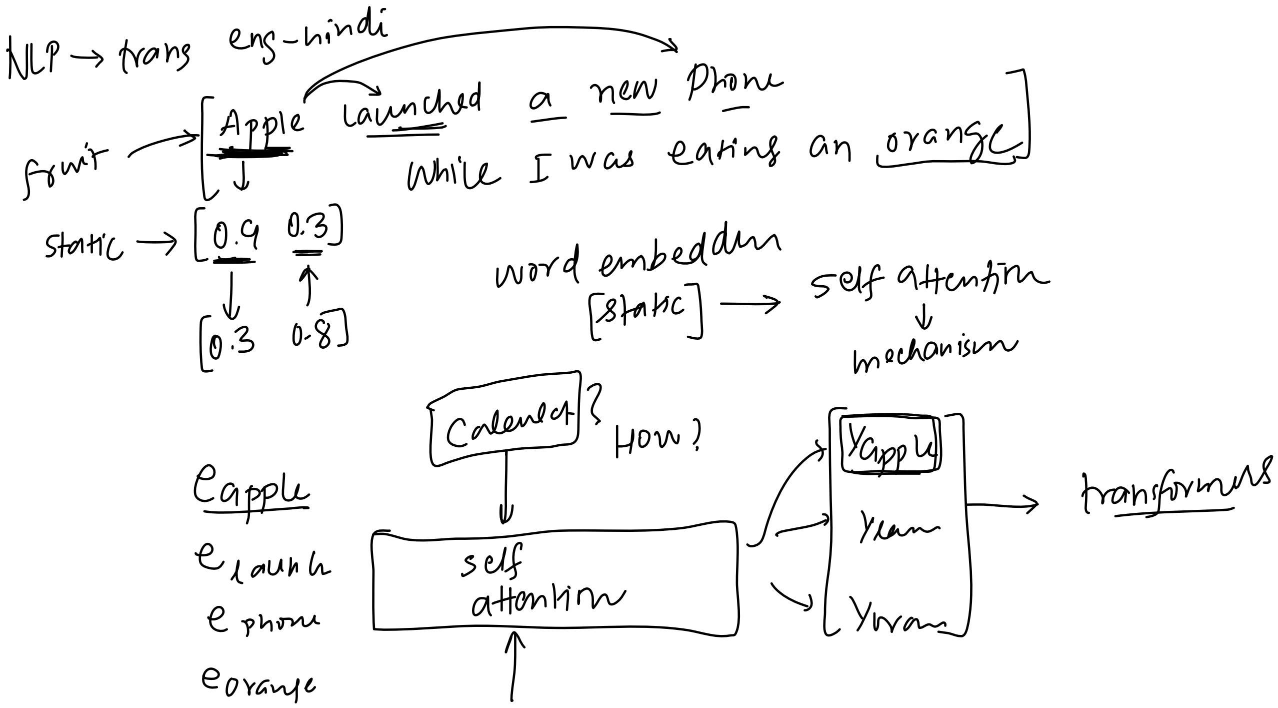
word embeddings

create → use → static

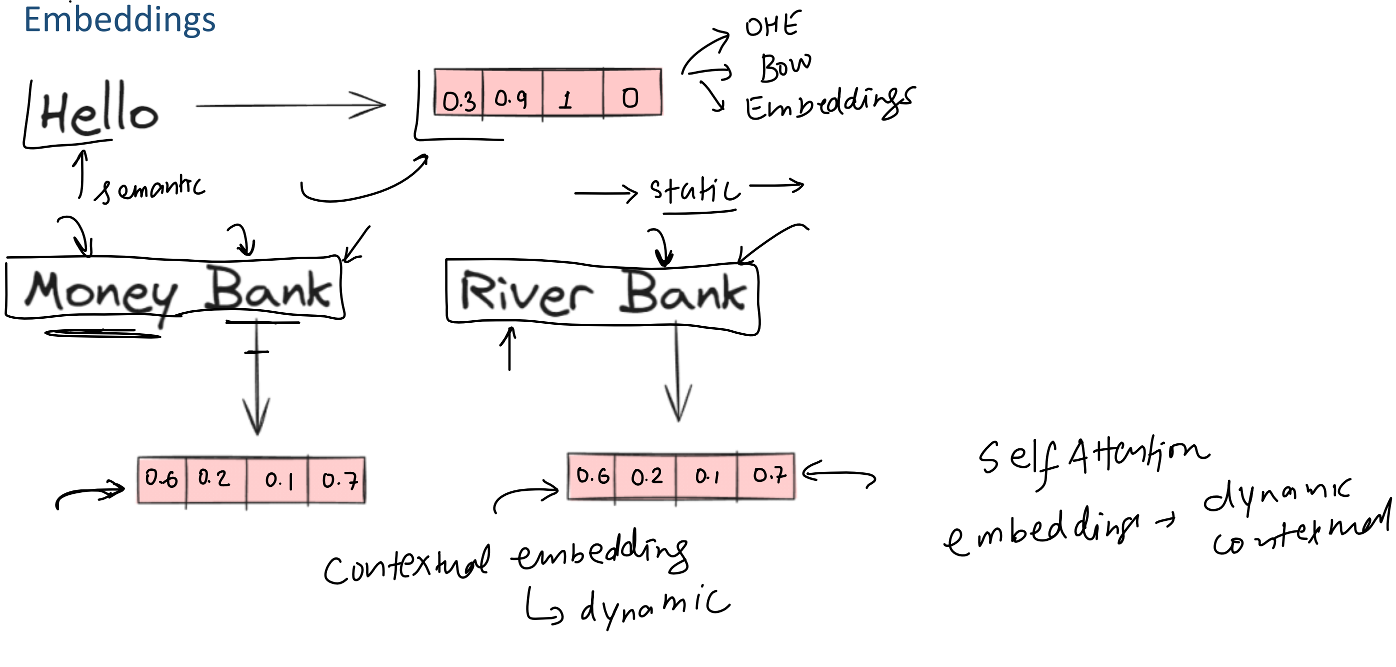
smart

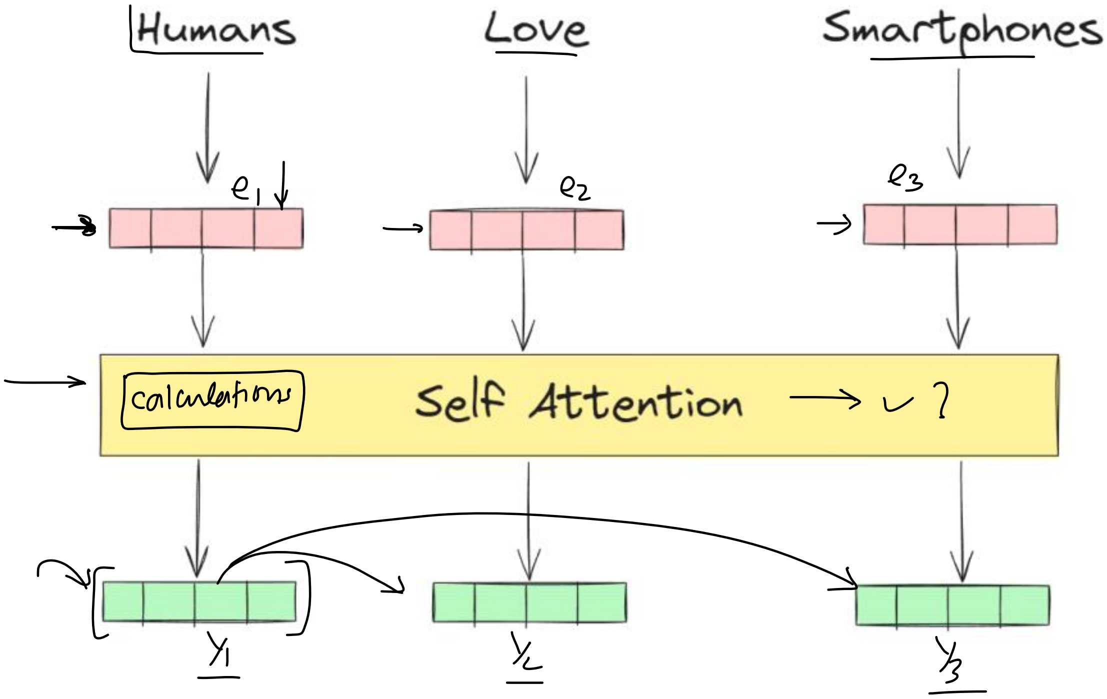
static embeddin → [contextual embeddin]

LLP → trans eng-hindi


NLP → trans eng-hindi
fruit
Apple launched a new Phone
While I was eating an orange
static → [0.9 0.3]
[0.3 0.8]
word embeddin
[static] → self attention
mechanism
Calender?
How?
e_{apple}
e_{launch}
e_{phone}
e_{orange}
self attention
y_{apple}
y_{lau}
y_{oran}
transformers

Embeddings


Hello
0.3
0.9
1
0
OHE
Bow
Embeddings
semantic
Money Bank
River Bank
static
0.6
0.2
0.1
0.7
0.6
0.2
0.1
0.7
Contextual embedding
dynamic
Self Attention
embeddings → dynamic
contextual


Humans
Love
Smartphones
e_1
e_2
e_3
calculations
Self Attention
→ ✓ ?
y_1
y_2
y_3

First Principle Approach

04 February 2024 17:55

$\boxed{\text{money bank grows}}$

bank → bank

bank → $\boxed{0.3\,\text{money} + 0.7\,\text{bank} + 0.1\,\text{grows}}$

$\boxed{\text{river bank flows}}$

bank = $\boxed{0.5\,\text{river} + 0.4\,\text{bank} + 0.1\,\text{flows}}$

S1 word embe

$$\text{money} = 0.7\,\text{money} + 0.2\,\text{bank} + 0.1\,\text{grows}$$
$$\boxed{\text{bank} = 0.25\,\text{money} + 0.7\,\text{bank} + 0.05\,\text{grows}}$$
$$\text{grows} = 0.1\,\text{money} + 0.2\,\text{bank} + 0.7\,\text{grows}$$

S2

$$\text{river} = 0.8\,\text{river} + 0.15\,\text{bank} + 0.05\,\text{flows}$$
$$\boxed{\text{bank} = 0.2\,\text{river} + 0.78\,\text{bank} + 0.02\,\text{flows}}$$
$$\text{flows} = 0.4\,\text{river} + 0.01\,\text{bank} + 0.59\,\text{flows}$$

n dim

$\boxed{1 \mid 1}$ n dim →

$$\left[\begin{aligned} e_{money}^{(new)} &= \boxed{0.7}\,e_{money} + \boxed{0.2}\,e_{bank} + \boxed{0.1}\,e_{grows} \\ e_{bank}^{(new)} &= \boxed{0.25}\,e_{money} + \boxed{0.7}\,e_{bank} + \boxed{0.05}\,e_{grows} \\ e_{grows}^{(new)} &= \boxed{0.1}\,e_{money} + \boxed{0.2}\,e_{bank} + \boxed{0.7}\,e_{grows} \end{aligned}\right]$$

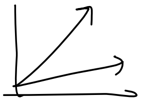
similarity

e_{money} e_{money}

e_{money} e_{bank}

e_{money} e_{grow}

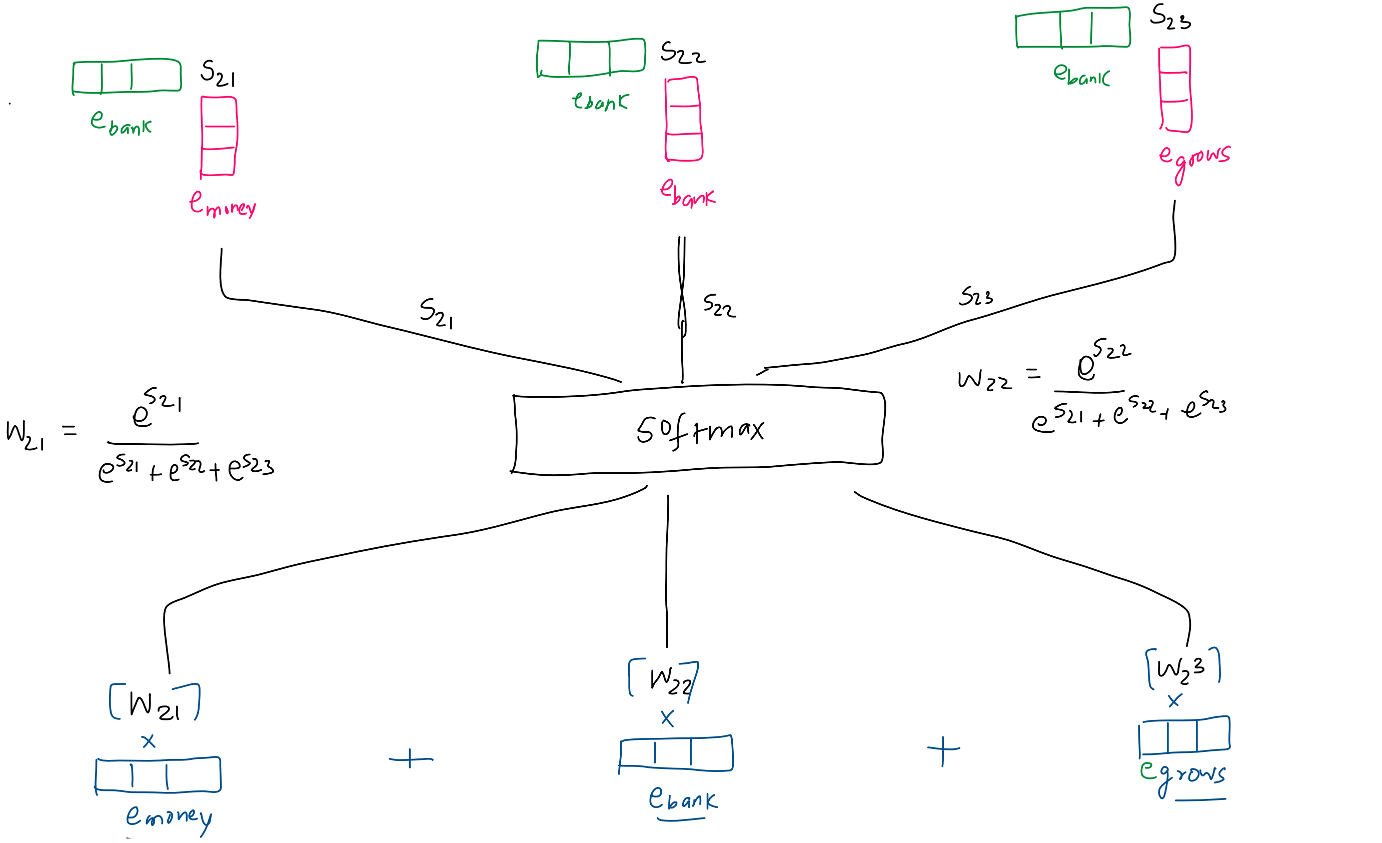
e_{grows} e_{money}



dot products

(1) normalized

$$e_{bank}^{(new)} = [e_{bank} \cdot e_{money}^T]\,e_{money} + [e_{bank} \cdot e_{bank}^T]\,e_{bank} + [e_{bank} \cdot e_{grows}^T]\,e_{grows}$$



$$W_{21} = \frac{e^{S_{21}}}{e^{S_{21}} + e^{S_{22}} + e^{S_{23}}}$$

$$W_{22} = \frac{e^{S_{22}}}{e^{S_{21}} + e^{S_{22}} + e^{S_{23}}}$$

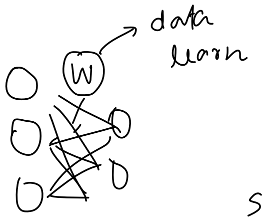
$\hat{e}_{money}$ + e_{bank} + e_{grows}

$e_{bank}^{(new)} = \boxed{y_{bank}}$ → contextual word embeddings

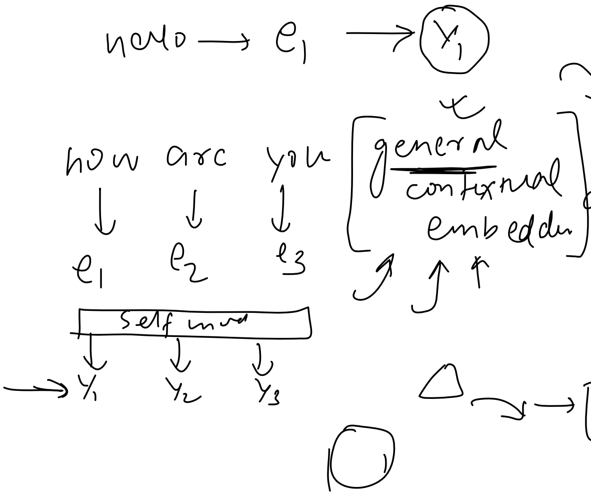
Points to consider

→ This operation is a parallel operation

→ There are no parameters involved



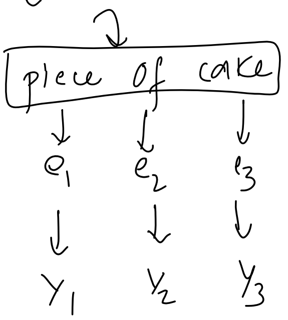
self model



machine translation

eng	hindi
how are you	कैसे हो।
I am good	मैं बढ़िया हूँ
piece of cake →	बहुत आसान काम
—	[illegible]

X [task specific contextual embeddings]



→ [केक का टुकड़ा।]

break a leg → टांग तोड़ दो

Progress

06 February 2024 00:42

Words → machine

↓

embedding → semantic

→ Context

river bank money bank

→ contextual embeddings

simple self model → parameters

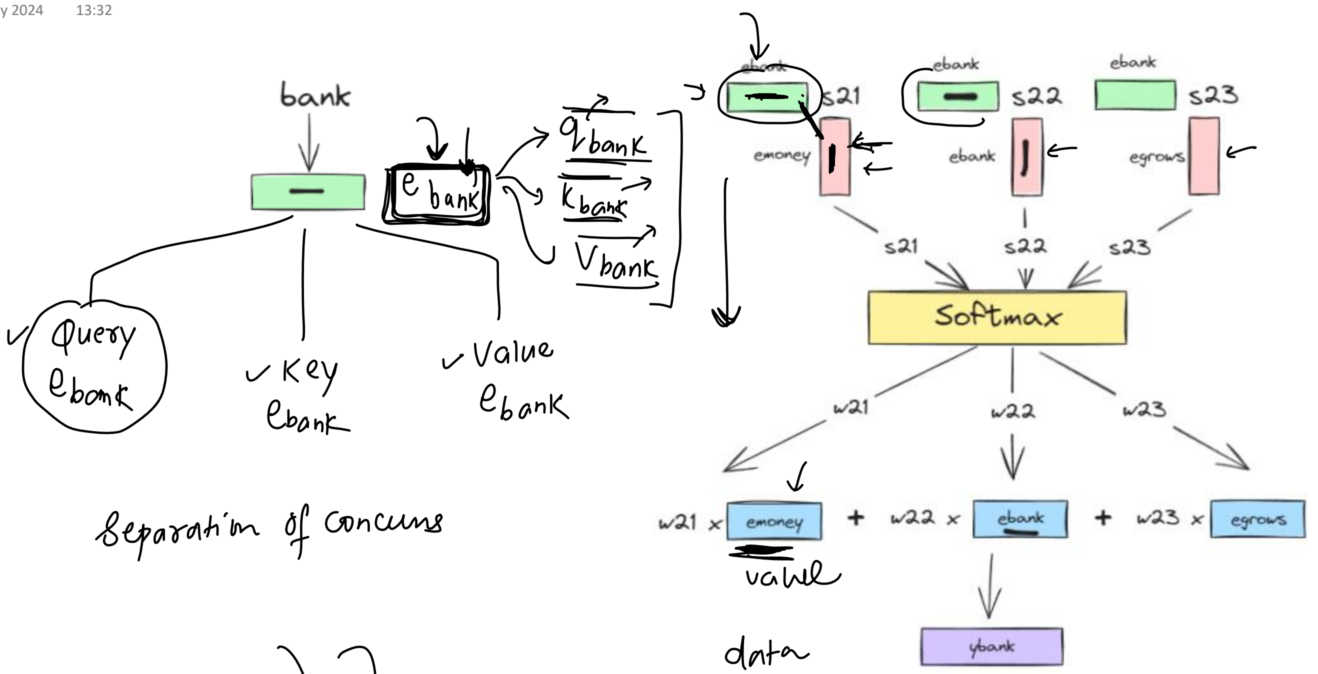
general

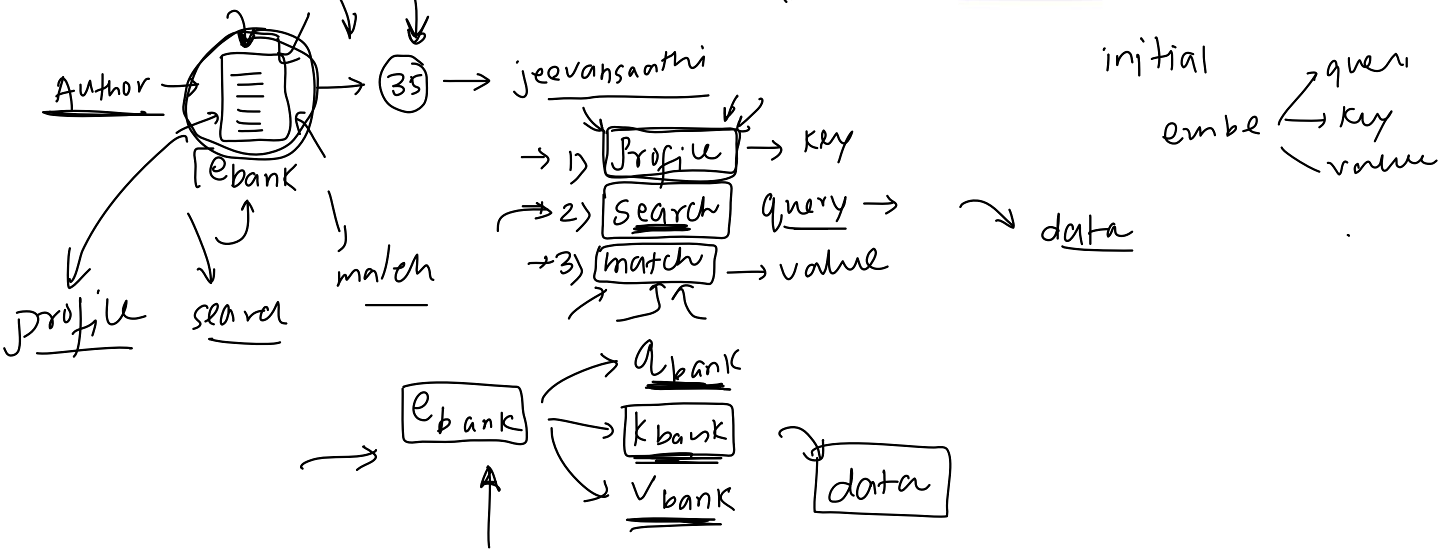
task-specific

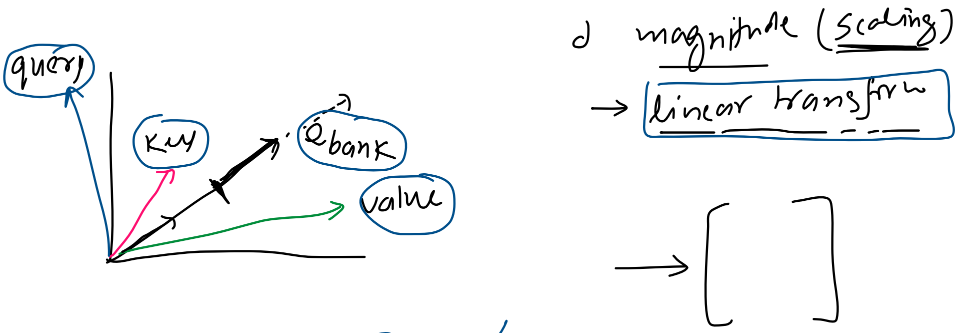
learnable parameter → general

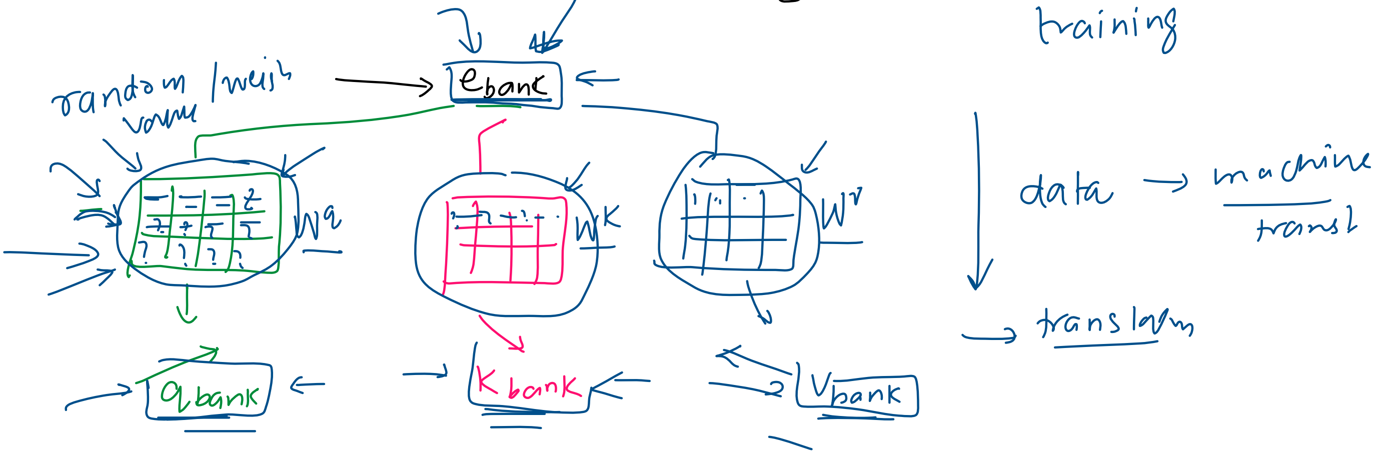
Query, Key & Value Vectors

06 February 2024 13:32









3x3 → dot → (9) values

Revision

28 February 2024 16:25

money bank grows
embeddings

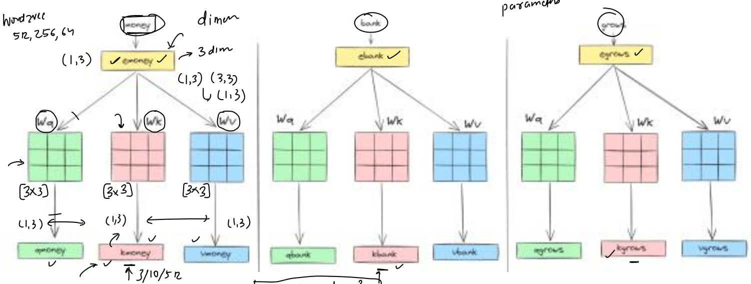
Wq Wk Wv
parameters

word2vec
512, 256, 64

dimen

(1,3) → 3 dim

(1,3) (3,3)
↳ (1,3)



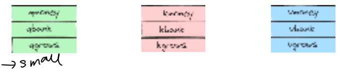
[3x3] [3x3] [3x3]

(1,3) (1,3) (1,3)

↑ 3/10/512

d_k → dim of the K vector
key

$\boxed{d_k = d_q = d_v = 3}$



(3) (3) → $\boxed{(9)\ \text{dot}}$ → small
→ high

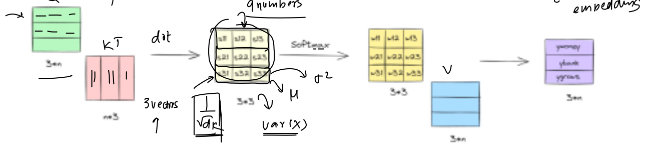
(1,3) (1,3)

(3 vector)

9 vector-vector dot pro

(1,3)

512 dim



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

$\sqrt{3}$

Scaled Dot product

$[QK^T]$
matrix matrix
dot product
vectors dot product

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

$\frac{QK^T}{\sqrt{d_k}}$ → unstable gradient
↳ $\sqrt{d_k}$

Softmax$\left(\frac{QK^T}{\sqrt{3}}\right)V$

$\frac{1}{\sqrt{d_k}}$

$\left[\frac{1}{\sqrt{d_k}}\right]$ scale ?

why ?

↳ Dot-product ka Nature

1) [1, 2, 3] — [3, 2, 1]
2) [ ] - [ ]
3) [ ] — [ ]
4) [ ] — [ ]
5 [ ] — [ ]

f g h i . j → σ_2^2

1) [1, 2] — [3, 2]
2) [ ] — [ ]
3) [ ] — [ ]
4) [ ] — [ ]
5) [ ] — [ ]

a b c d e → σ_1^2

low dimensional → dot product → low variance

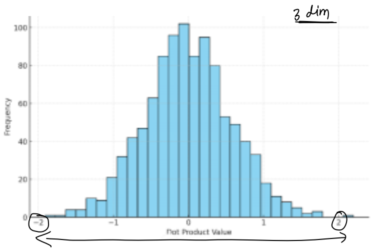
high dim → dot product → high variance

code

3 dim

1000 pair of vectors

1) — x — → —
2) — x — → —
⋮
1000 — — →

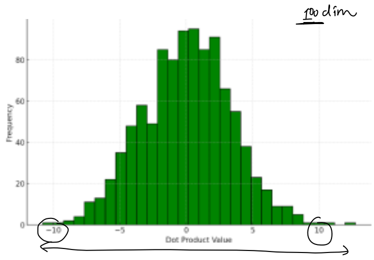


100 dim

1000 pairs of vector

1) — x —
2) — x —
⋮
1000) — x —

0 0 0 0 0 0

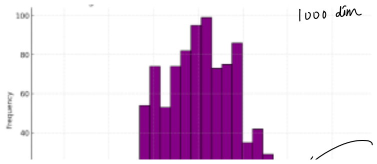


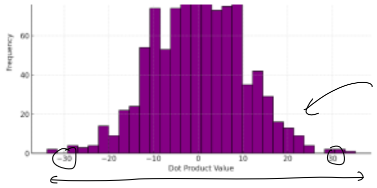
1000 dim

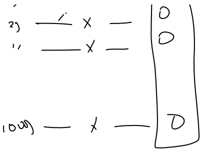
1000 pair of vectors

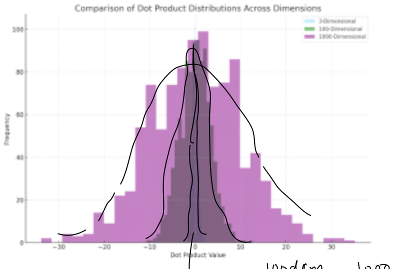
1) — x —
2) — x —
3) — x —

0 0 0







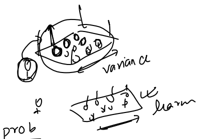

Comparison of Dot Product Distributions Across Dimensions

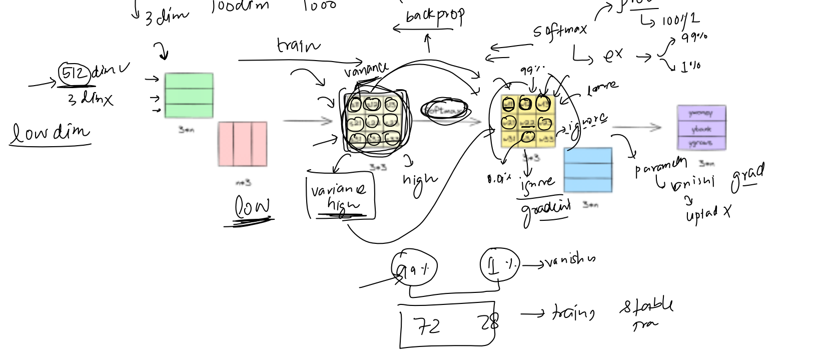
3 dim 100dim 1000

vector

dim → variance
↳ high

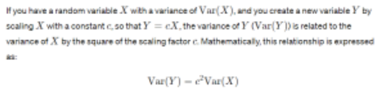
high variance
↓
it's a problem





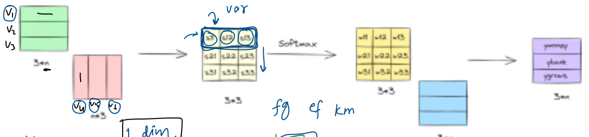
cach c

$$\left\{\begin{array}{l} X \rightarrow Var(x) \\ Y \rightarrow cX \\ \quad \hookrightarrow c^2 Var(x) \end{array}\right\}$$


If you have a random variable X with a variance of $\mathrm{Var}(X)$, and you create a new variable Y by scaling X with a constant c, so that $Y = cX$, the variance of Y ($\mathrm{Var}(Y)$) is related to the variance of X by the square of the scaling factor c. Mathematically, this relationship is expressed as:

$$\mathrm{Var}(Y) = c^2\mathrm{Var}(X)$$

$\begin{bmatrix} V_1 \odot V_4 \\ V_1 \odot V_5 \end{bmatrix}$ dim ↑ ↳ variance ↑ → $\boxed{\text{math quantify}}$



fg ef km

V_1 — V_4, V_5, V_6

$\boxed{1\ dim}$

[a]

3

[b] 2 → ab (6) 6 → random var → x

[c] 1 → ac (3) 3 → variance (sample) ↓ expected var → $\boxed{Var(x)}$

[d] 4 → ad (12) 12

2dim

$\textcircled{X}$ → var(X)

cX → $c^2 Var(x)$

[a b]

[c d] → $\dfrac{ac + bd}{\sqrt{2}}$ → 2 Var(X)

[e f] → $\dfrac{ae + bf}{\sqrt{2}}$ 18 → y → Var(y) Var(y) ≃ 2 Var(x)

[g h] → $\dfrac{ag + bh}{\sqrt{2}}$ 2 → Var(X)

Var(y) > var(X) Yes No

y → Var(y) → 2Var(x) var

$\dfrac{y}{\sqrt{2}} \rightarrow \dfrac{1}{2} Var(y) \rightarrow \dfrac{1}{\cancel{2}} \cancel{2}\, Var(x) \rightarrow Var(x)$

2dim → Var(X)

$\boxed{\dfrac{1}{\sqrt{d_K}}}$

Var(z) > Var(y) > Var(x)

3 dim [a b c]

[d e f] → $\dfrac{ad + bc + cf}{\sqrt{3}}$

[g h i] → $\dfrac{ag + bh + ci}{\sqrt{3}}$ → z

[k l m] → $\dfrac{ak + bl + cm}{\sqrt{3}}$ → Var(X)

Var(z) ≃ 3 Var(x)

z → var(z) ≃ 3 Var(x)

$\dfrac{1}{\sqrt{3}} z \rightarrow \left(\dfrac{1}{\sqrt{3}}\right)^2 Var(z) = \dfrac{1}{\cancel{3}} \cancel{3} Var$

d dim → d Var(x)
↳ varina dim = 1

dim → varian linear

1 dim → Var(x) → Var(x)

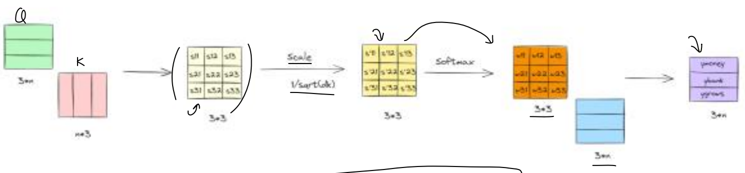
2 dim → 2 Var(x) → Var(x)

3 dim → 3 Var(x) → Var(x)

⋮

d dim → d Var(x) → Var(x)

[_ - - - - -] → var(X) →
↑ d dim



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

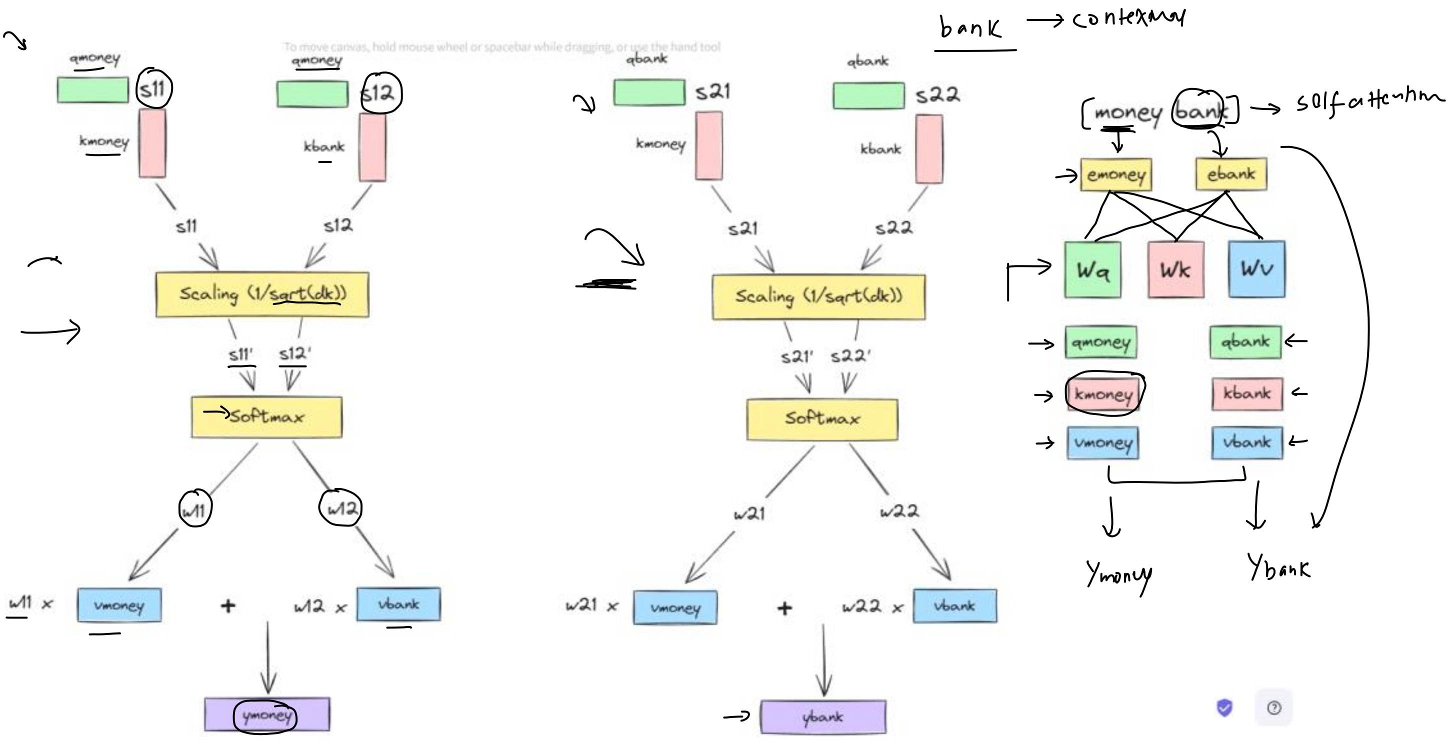
↑ transformers

What is d_k

28 February 2024 16:59

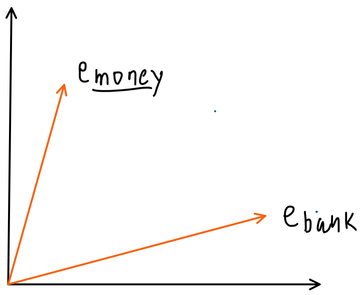
Recep

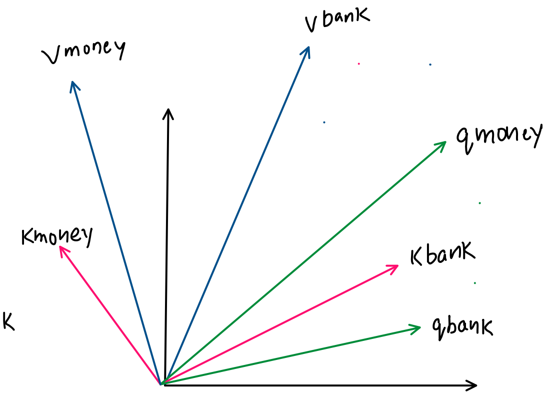
08 March 2024 15:14

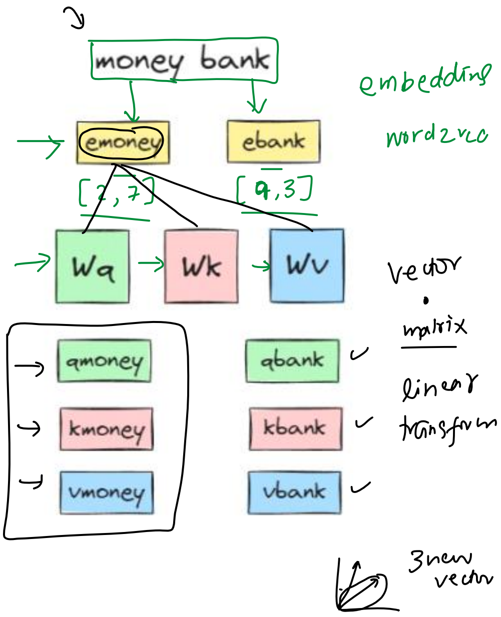


Geometric Intuition

08 March 2024 15:16



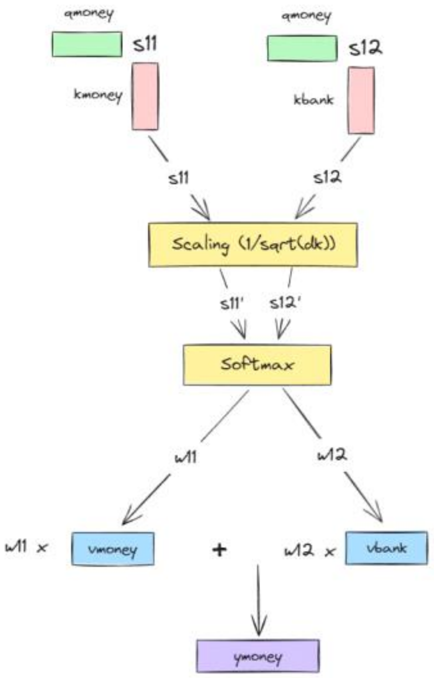


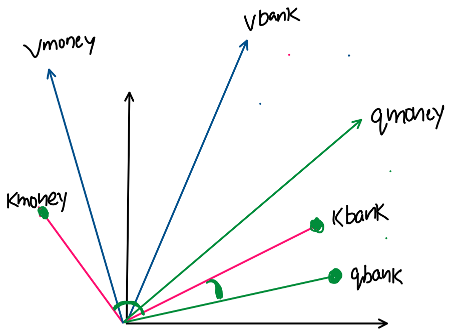


2x2 2x2 2x2

✓ $W_q = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ $W_K = \begin{bmatrix} 3 & 4 \\ 5 & 1 \end{bmatrix}$ $W_v = \begin{bmatrix} 4 & 1 \\ 2 & 1 \end{bmatrix}$

* All values are hypothetical





Dot Product

$S_{21} = \boxed{10}$ $S_{22} = \frac{\boxed{32}}{\sqrt{2}}$

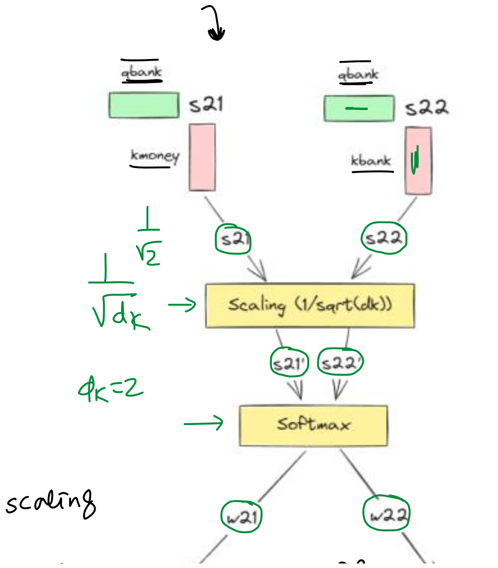
Scaling

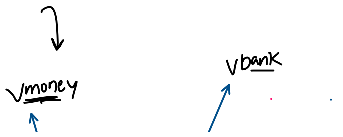
$S'_{21} = \boxed{\frac{10}{\sqrt{2}}} = 7.09$ $S'_{22} = 22.69$

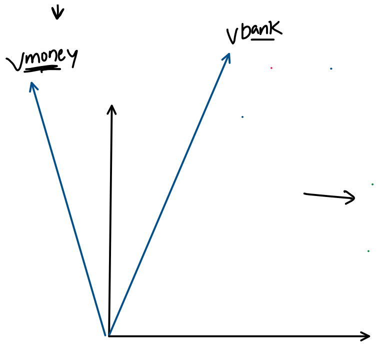
Softmax

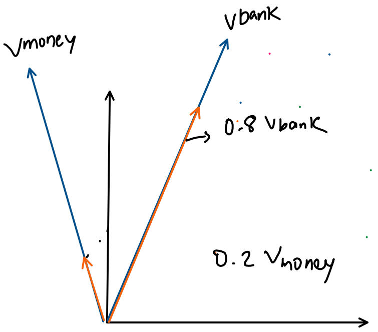
$\boxed{1}$

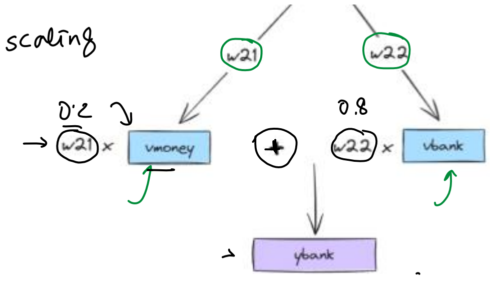
$W_{21} = \boxed{0.2}$ $W_{22} = \boxed{0.8}$

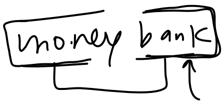






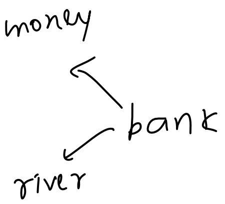


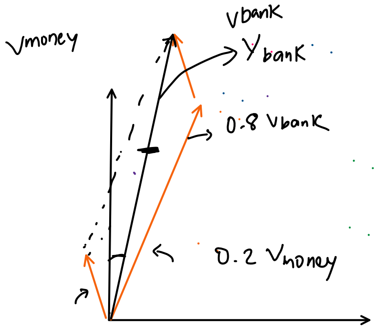


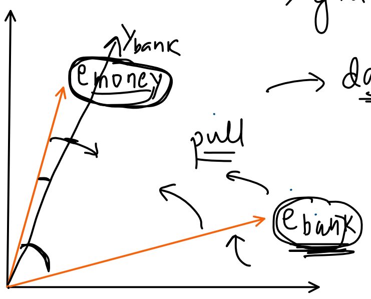


self attention

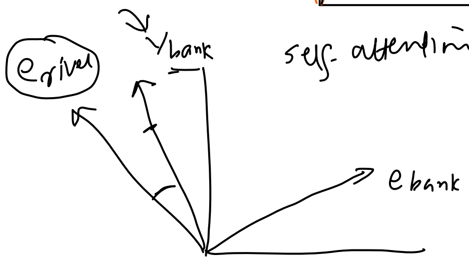
→ gravity







→ dataset



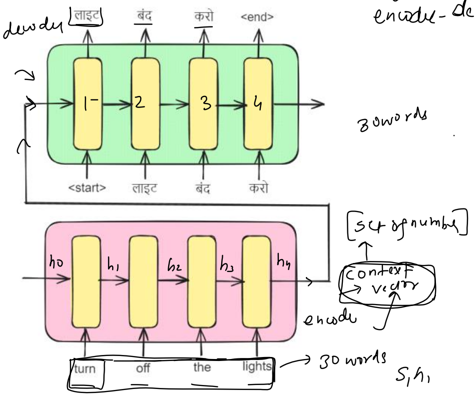
self- attention

river bank

Recap of Attention

11 March 2024 17:09

long trans → NN
↓
encoder-decoder



30words

→ 30 words

s_1h_1

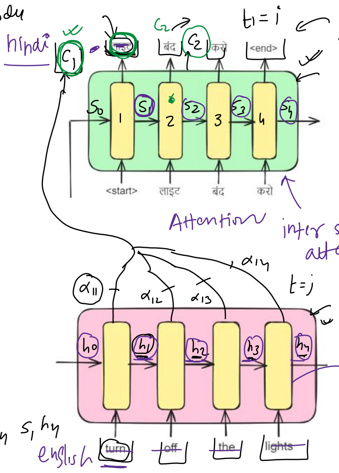
e_{11} s_1h_2

e_{12} s_1h_3

e_{13}

e_{14} s_1h_4

Attention



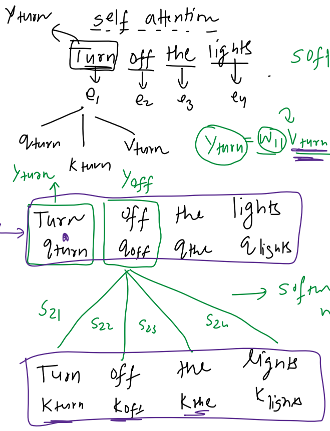
$$c_i = \sum \alpha_{ij} h_j \checkmark$$

$$\alpha_{ij} = \text{softmax}(e_{ij}) \checkmark$$

$$e_{ij} = s_i^T h_j \checkmark$$

inter sequence attention

s_i → query
h_j → key
h_j → value



softmax(s_{11})

s_{12} / s_{13} / s_{14}

$$y_{turn} = w_{11} v_{turn} + w_{12} v_{off} + w_{13} v_{the} + w_{14} v_{lights}$$

y_{off}

→ softmax $\overset{v_{turn}}{w_{21}}$ $\overset{v_{off}}{w_{22}}$ $\overset{v_{the}}{w_{23}}$ $\overset{v_{lights}}{w_{24}}$

intra sequence
↓
self

self

$$c_i = \sum \alpha_{ij} h_j$$ → weights

$$\alpha_{ij} = \text{softmax}(e_{ij})$$ → alingment score

$$\boxed{e_{ij} = s_i^T h_j}$$

→ Luong attention

↓

self attention