w_1 [0.03 0.05 −0.1 0.8 . .]

w_2 [0.2 0.8 −0.3 . .]

w_3 [ ]

d-dim

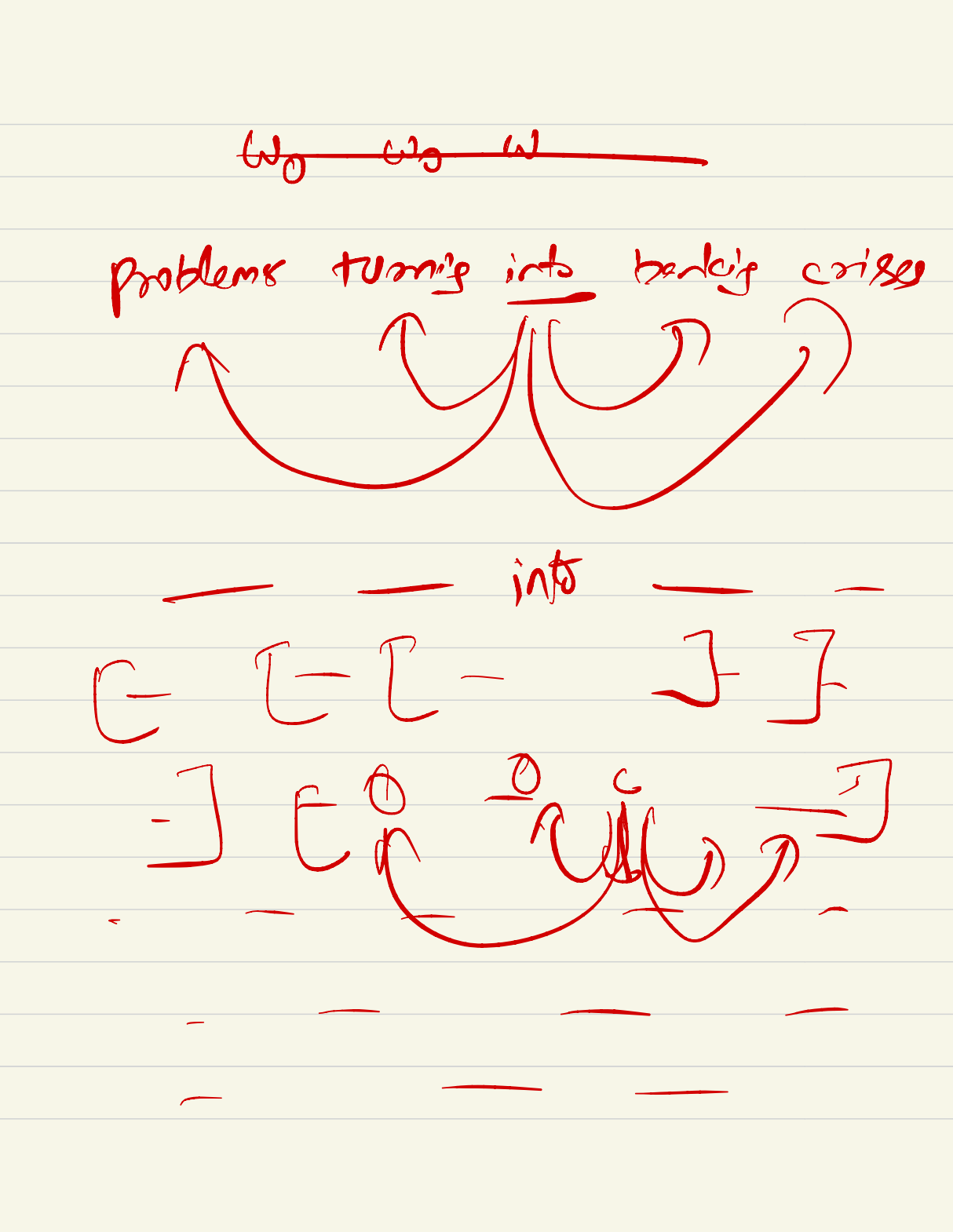
$$\frac{\vec{w_1} \cdot \vec{w_2}}{|\vec{w_1}||\vec{w_2}|}$$

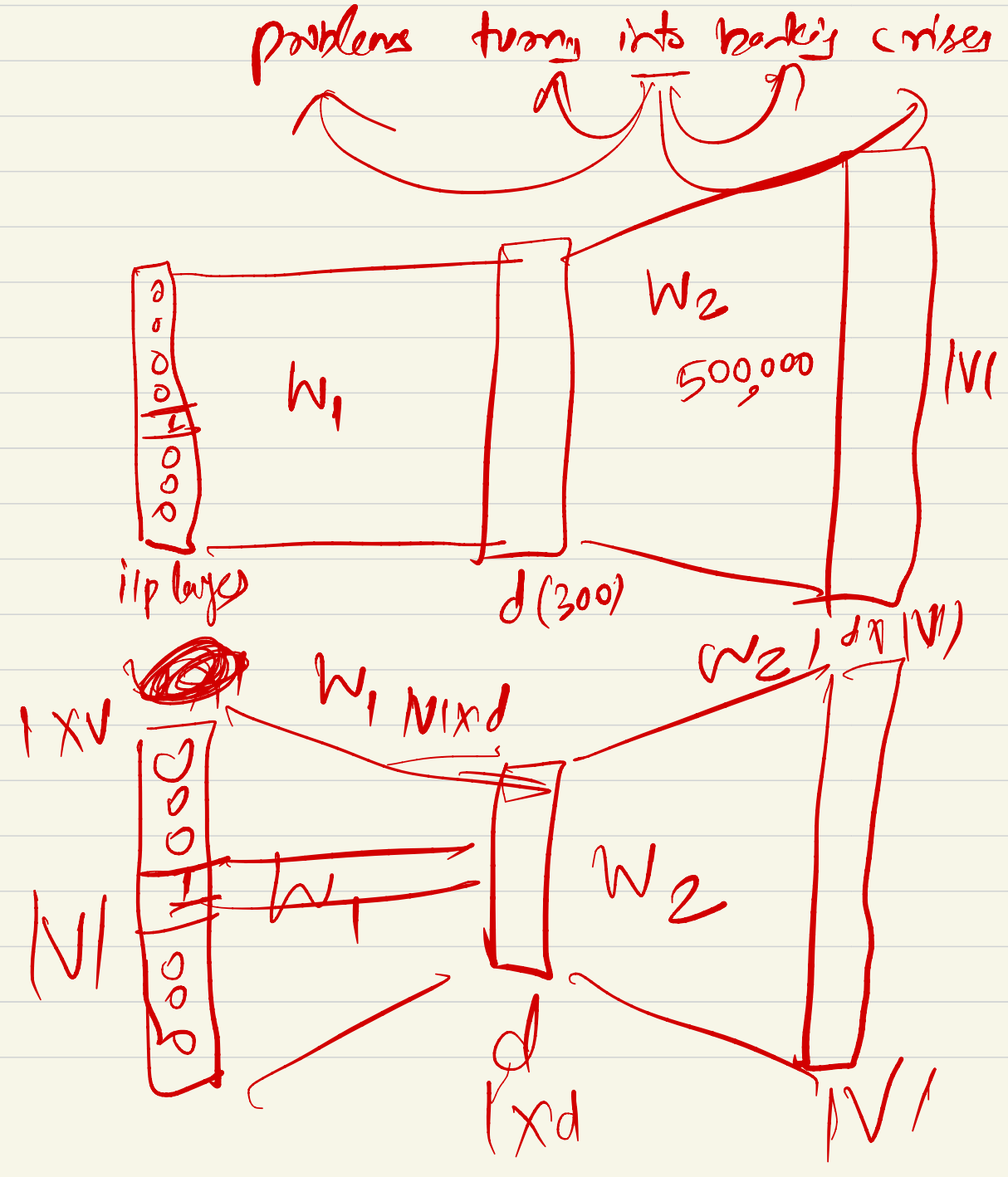
w_1 [−0.2 −0.3 . .]$_{-d}$

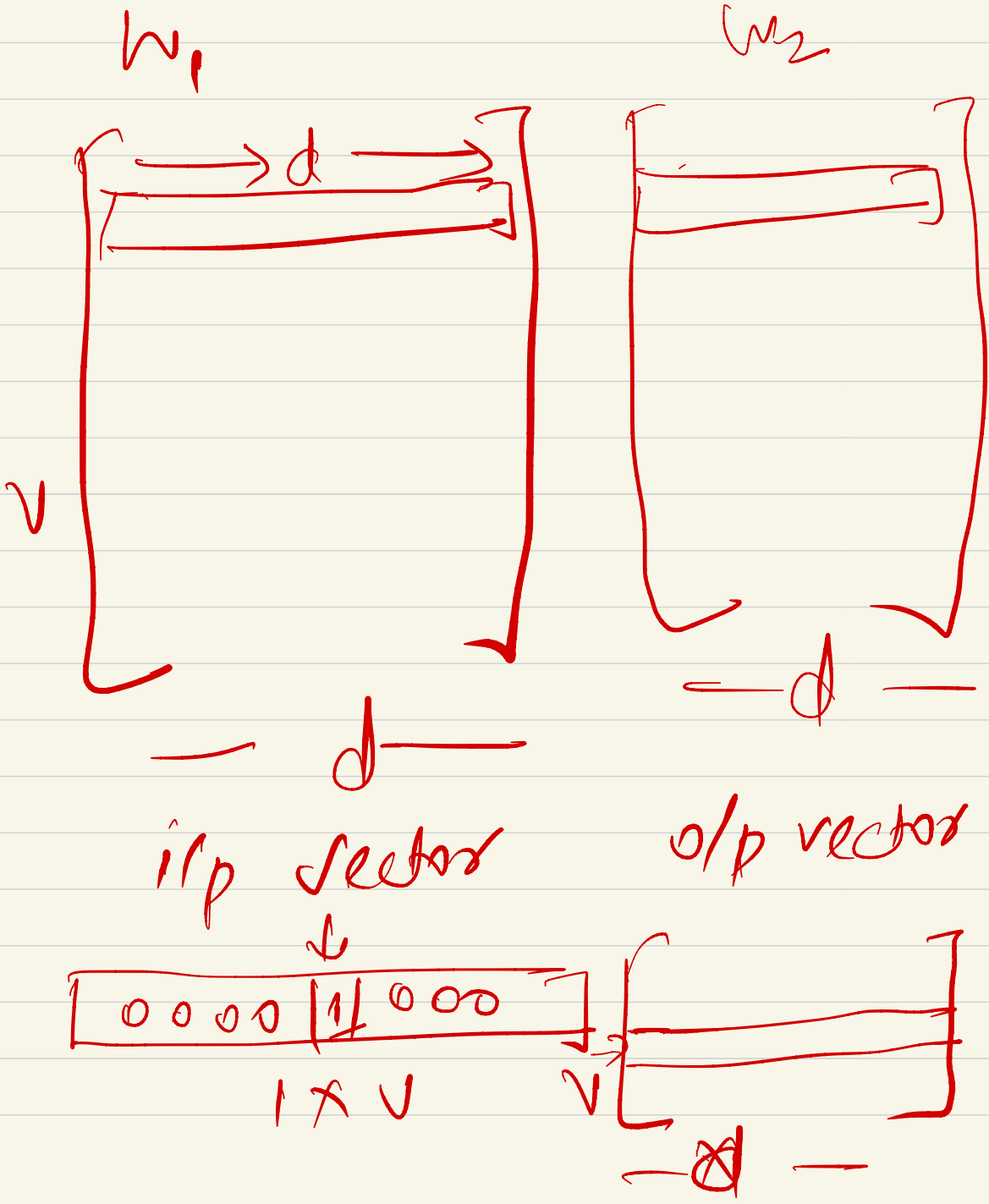
w_2 [ ]$_{-d}$

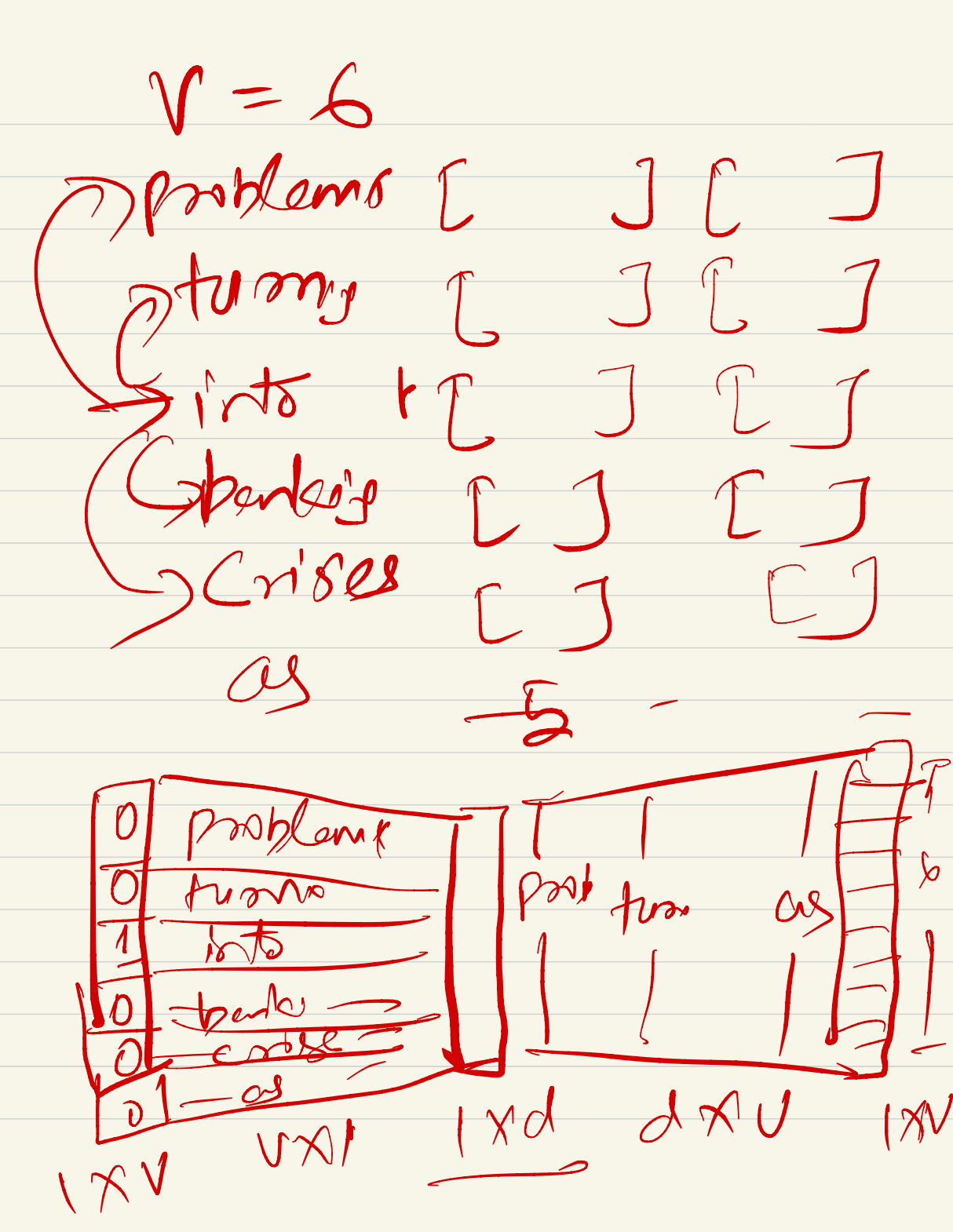
w_2 [ ]$_{-d}$

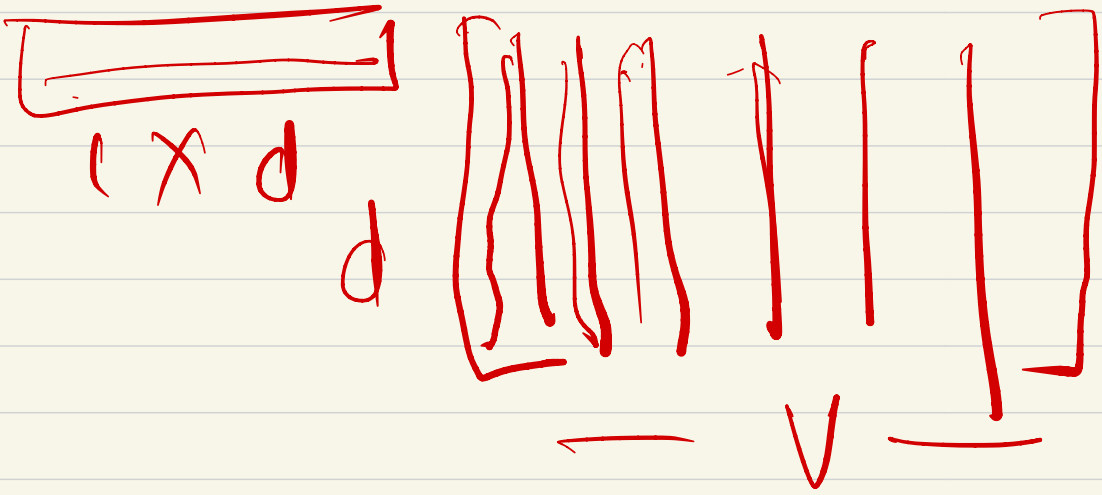
$d = 100/300$

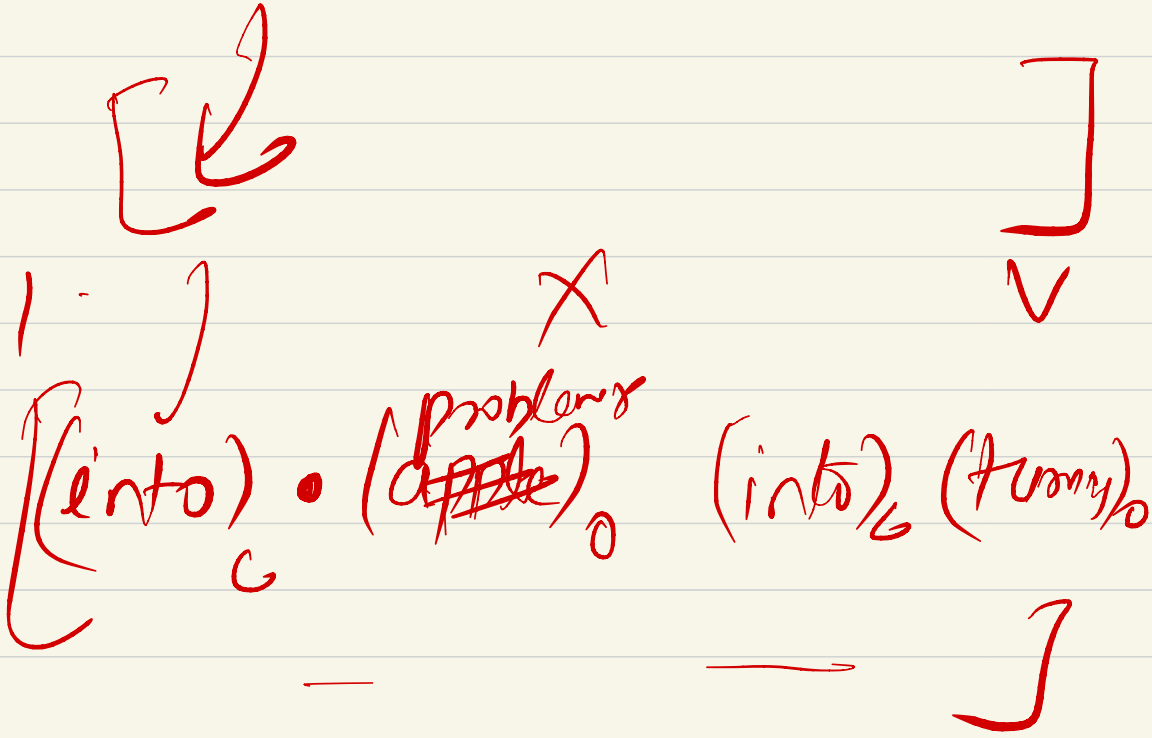

problems turning into banking crises
into

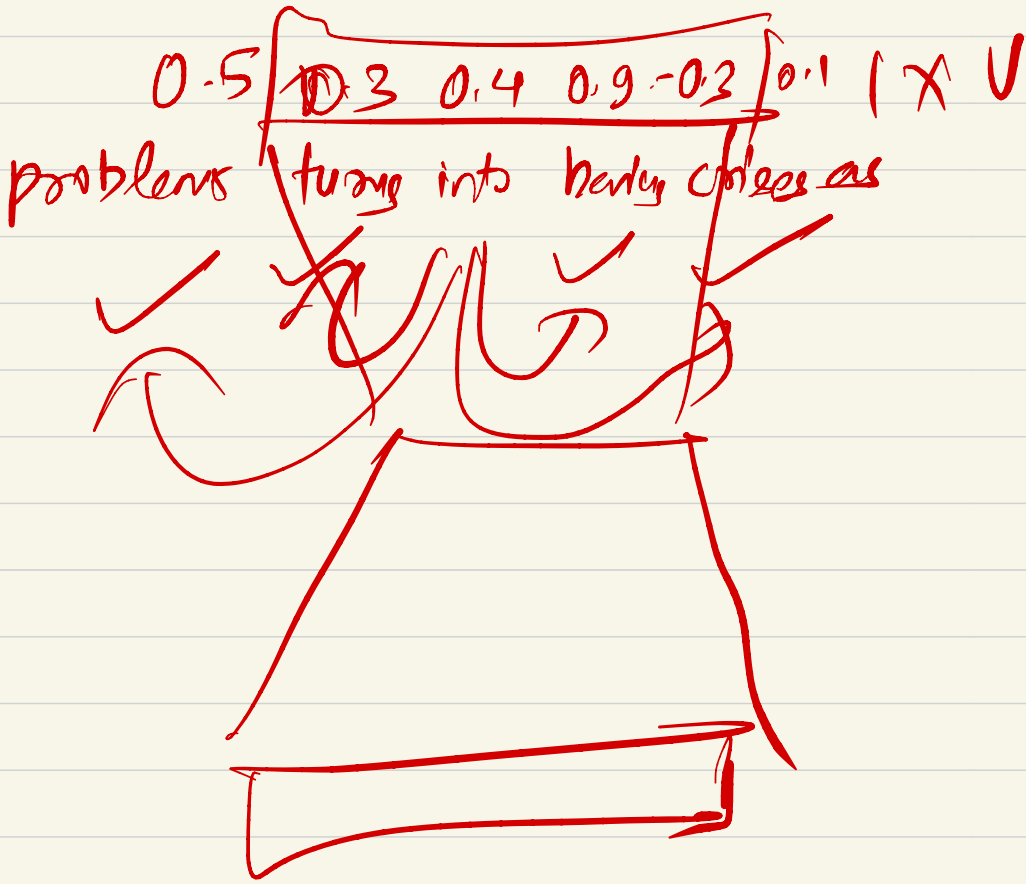

problems turning into banking crises
W1
W2
500,000
|V|
i/p layer
d (300)
1 x V
W1 N x d
W2
|V|
d
1 x d
|V|


W1
W2
d
V
d
d
i/p vector
o/p vector
0 0 0 0 1 1 0 0 0
1 x V
V
d


V = 6
problems
turning
into
banking
crises
as
0
0
1
0
0
1


x
d
d
V


into
Problems
into
turn



Pred. [0.5 0.3 0.4 0.9 -0.3 0.1]

GT [1 0 0 0 0 0]

$$-\sum_i p_i \log q_i$$

$$p = [0 \; 0 \; 0 \; 1 \; 0 \; 0 \; 0]$$

$$Q = [q_1 \; q_2 \; \cdots \; q_i \; \cdots]$$

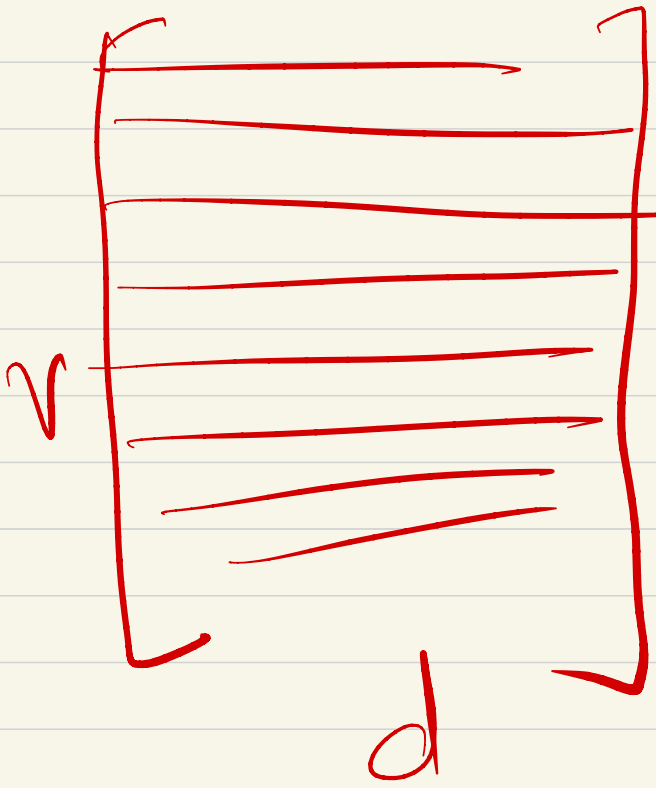
$$-\log q_i$$

$$\log \frac{1}{q_i}$$

Softmax

x_1 — — x_n

$$\frac{e^{x_1}}{\sum e^{x_i}} \quad \frac{e^{x_2}}{\sum e^{x_i}} \quad \frac{e^{x_n}}{\sum e^{x_i}}$$


V
d
Center
word vectors

passed through relU activation function

context window

2-d wv

in [1 -1] [1 1] [-2 1] [0 1] [1 0]

out

V = 5

d = 2

e^{-3} e^{-1} e^{5} e^{1} e^{-2}

$-\log \frac{e^{1}}{(e^{-3}+\ldots)}$

2 2 2 2 2

1×5 [-3 -1 5 1 -2] o/p layer

5×2 Out vectors

$-\log\left(\frac{e^{-1}}{e^{-3}+e^{1}+e^{5}+e^{1}+e^{-2}}\right)$

2

1×2 [-2 1] hidden

2×5 In vectors

through

1×5 [0 0 1 0 0] I/p layer

5

Cross-Entropy Error

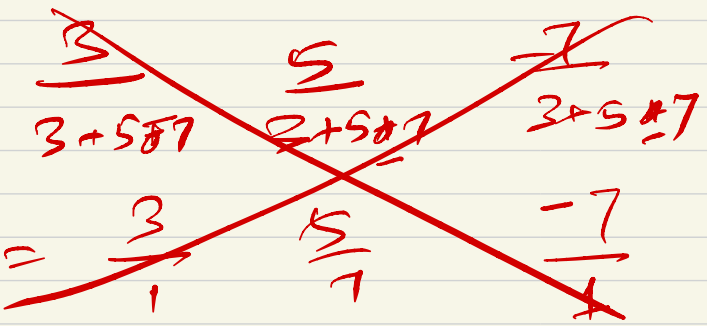
P = [0 1 0 0 0] →

Model pred. q = [q_1 q_2 q_3 q_4 q_5]

$-\sum_i p_i \log q_i = -\log(q_2)$

Softmax

3 5 -7



Softmax $\frac{e^3}{e^3+e^5+e^{-7}}$ $\frac{e^5}{e^3+e^5+e^{-7}}$ $\frac{e^{-7}}{e^3+e^5+e^{-7}}$

1-hot

hotel = [0 1 — 0]

motel [— 0 — 1 — 0]

chalk = [-0.1 — 0 —]

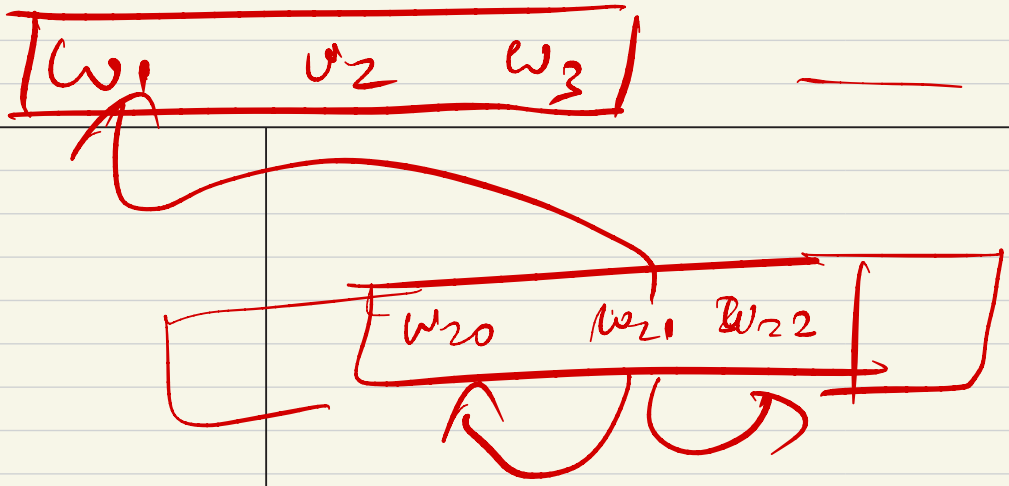
hotel = [0.3 -0.5]

motel = [0.4 -0.3]

chalk = [-0.1 0.3]

0.27

-0.45



d 300 100

$\sum_{n}$

w_{100}

man : woman :: uncle : ??

man	[0.3 0.7]
woman	[0.5 0.3]
uncle	[0.9 0.1]

man − woman ≃ uncle − x

x ≃ uncle + woman − man

≃ [1.1 −0.3]

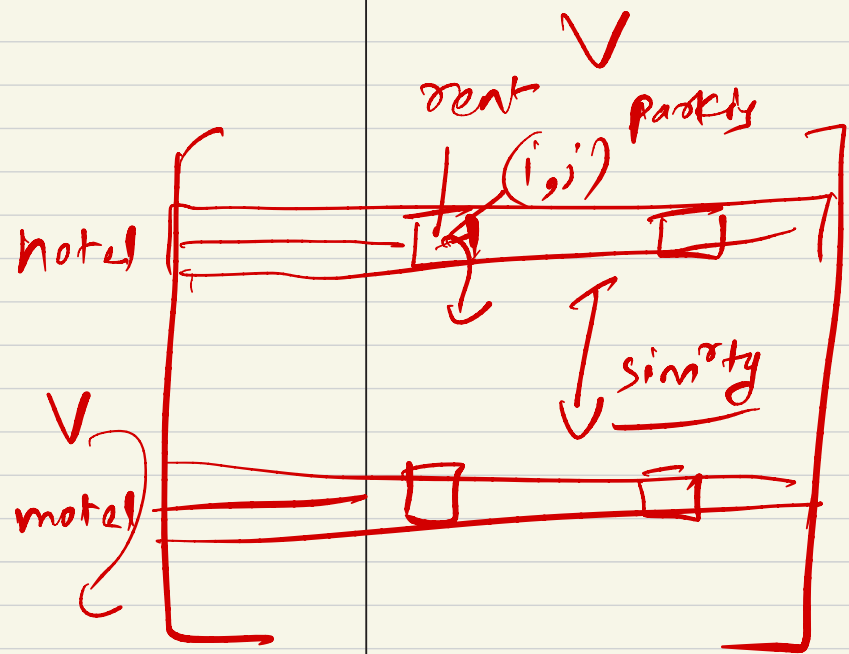
V | words

$\underset{x}{\text{argmax}}$ [1.1 −0.3] x

x = aunt

V words in vocab

co-occ.



1m

N dim

2

~200 words

500,000

w_1 [ ]

w_2 [ ]

w_V [ ]

$\vec{w}_i \cdot \vec{w}_j \approx \text{cooc}(i,j)$

100

W E_1 → H_1

E_2 H_2

En Wiki | Hi Wiki

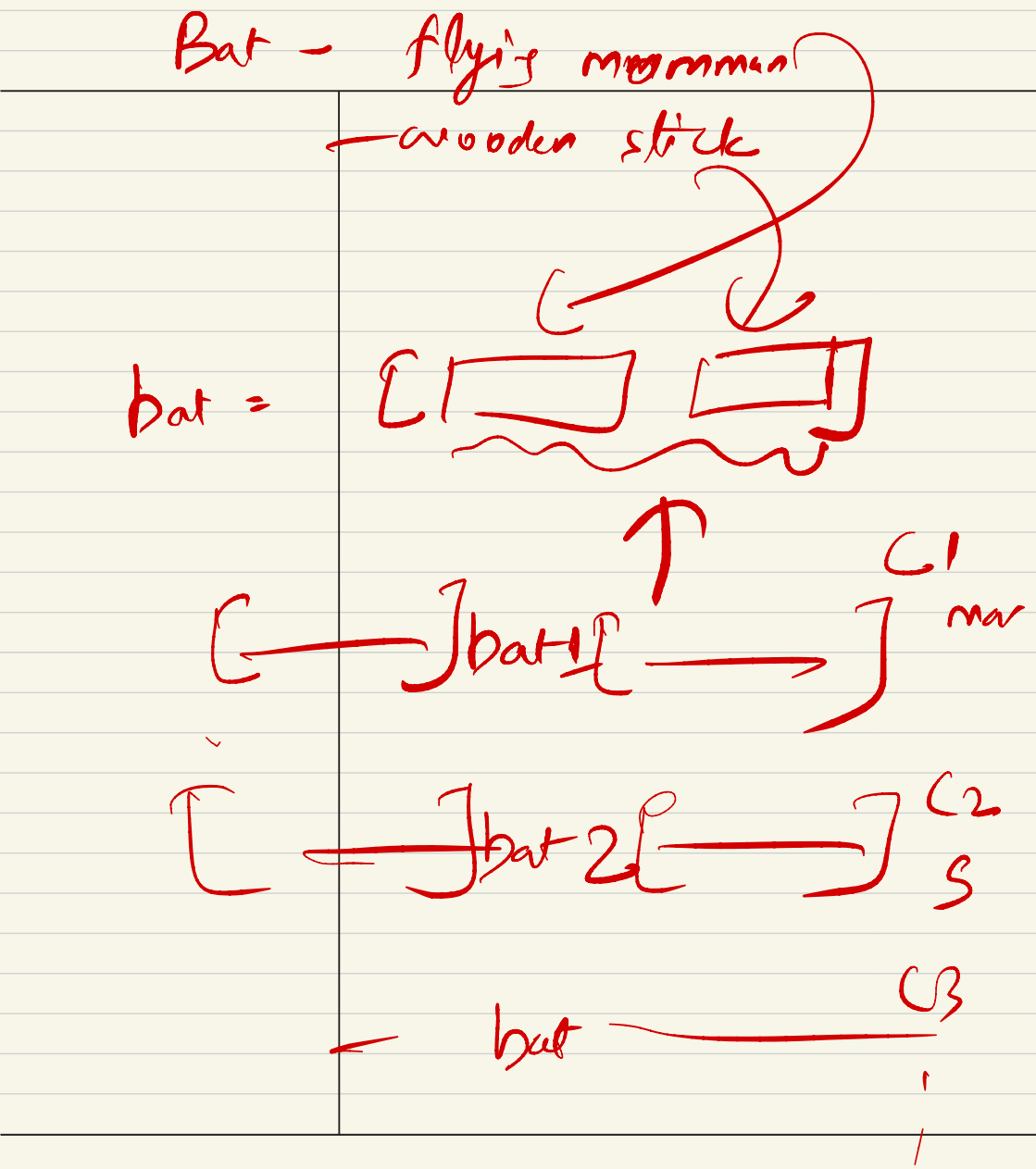
night [ ] | रात [ ]

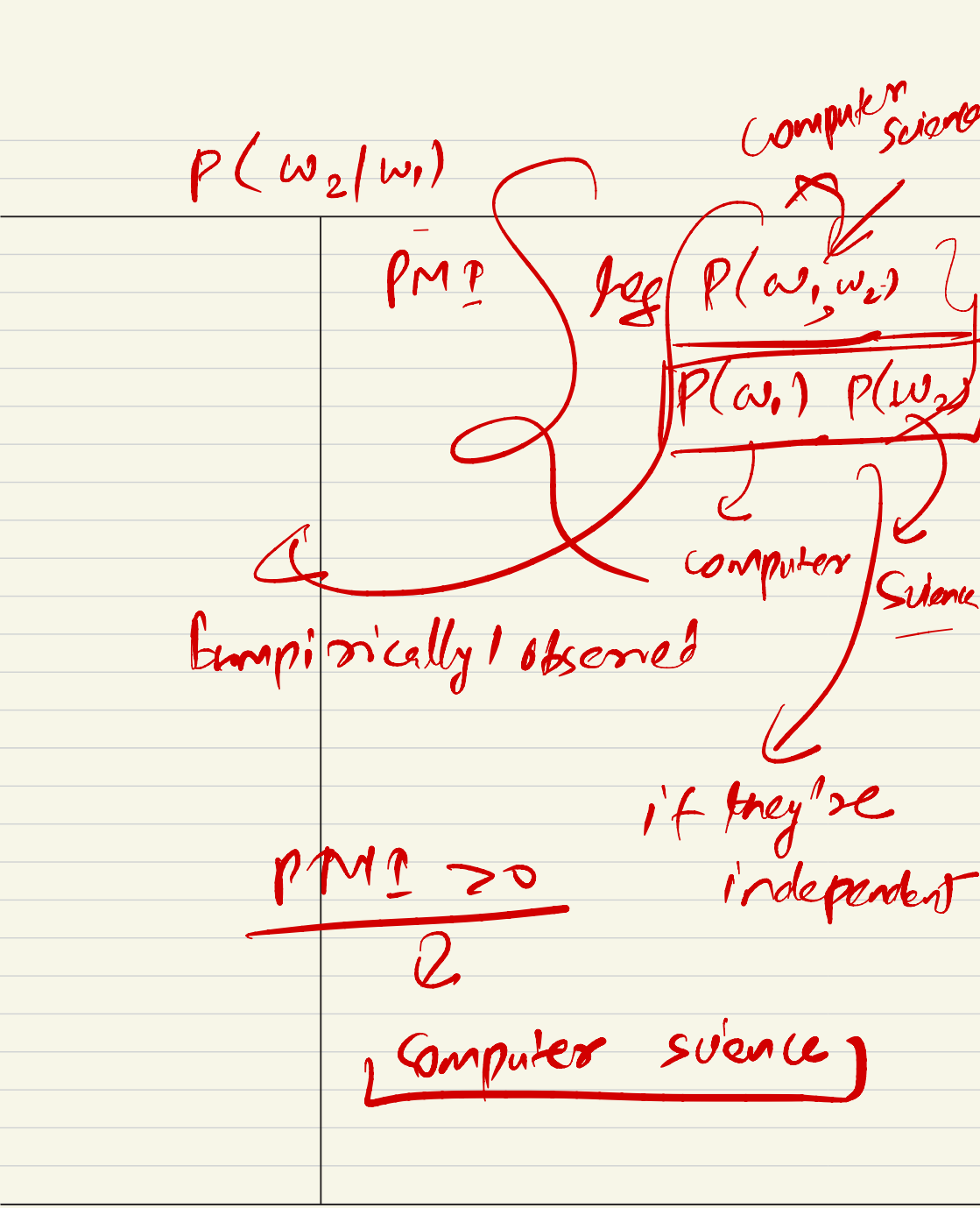
noon [ ] | दोपहर [ ]

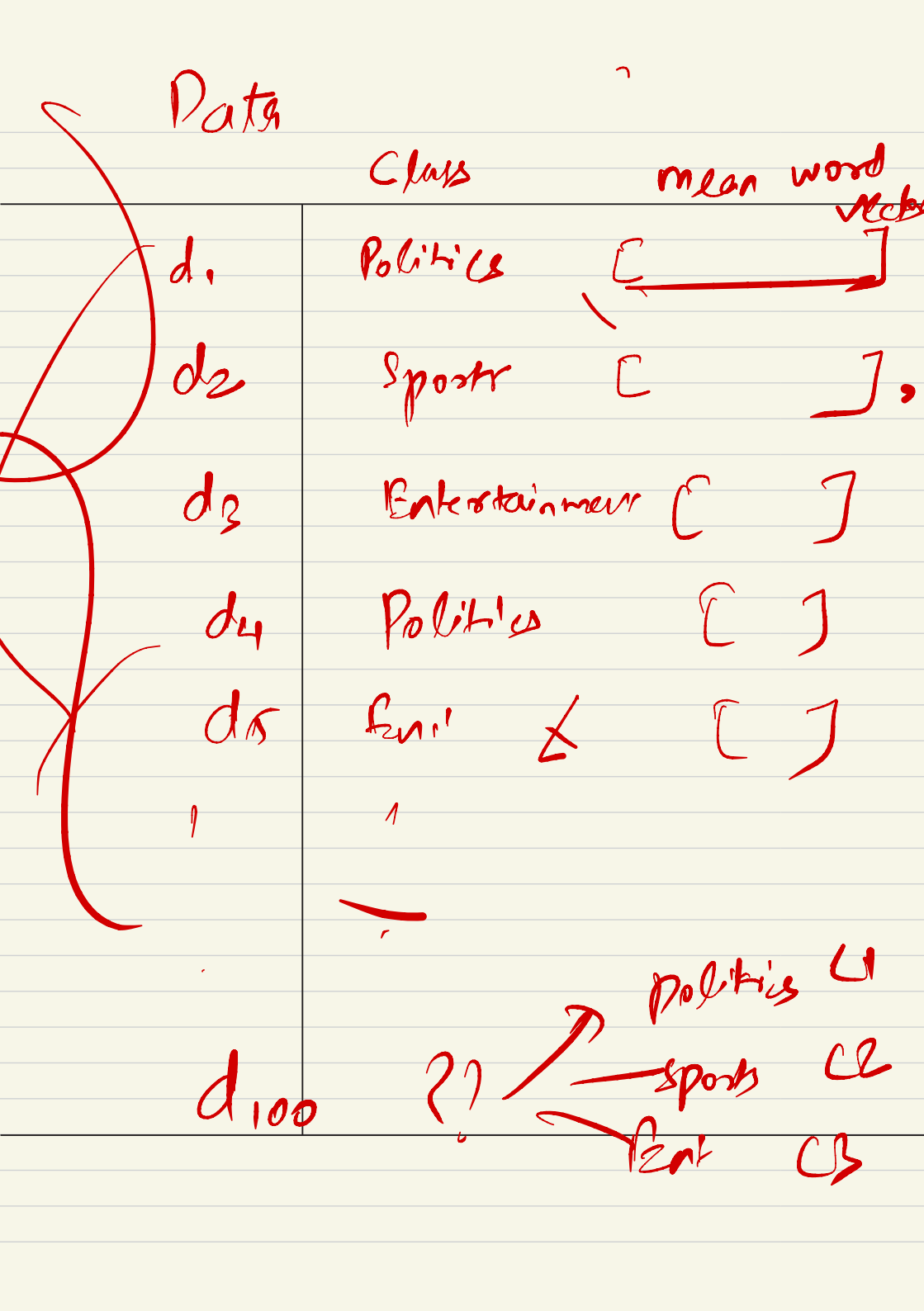
[Earthquake] → भूकम्प

mix & merge

~~Jumble~~ Hi & En


Bat - flying mammal
wooden stick
bat =
bat1
C1
mar
bat 2
C2
S
bat
C3


computer science
$P(w_2|w_1)$
PMI
log
$P(w_1, w_2)$
$P(w_1)\ P(w_2)$
computer
Science
Empirically observed
if they're independent
$PMI \geq 0$
2
computer science


Data
Class
mean word vecto
d1
Politics
d2
Sports
d3
Entertainment
d4
Politics
d5
Ent.
d100
??
Politics C1
Sports C2
Ent. C3

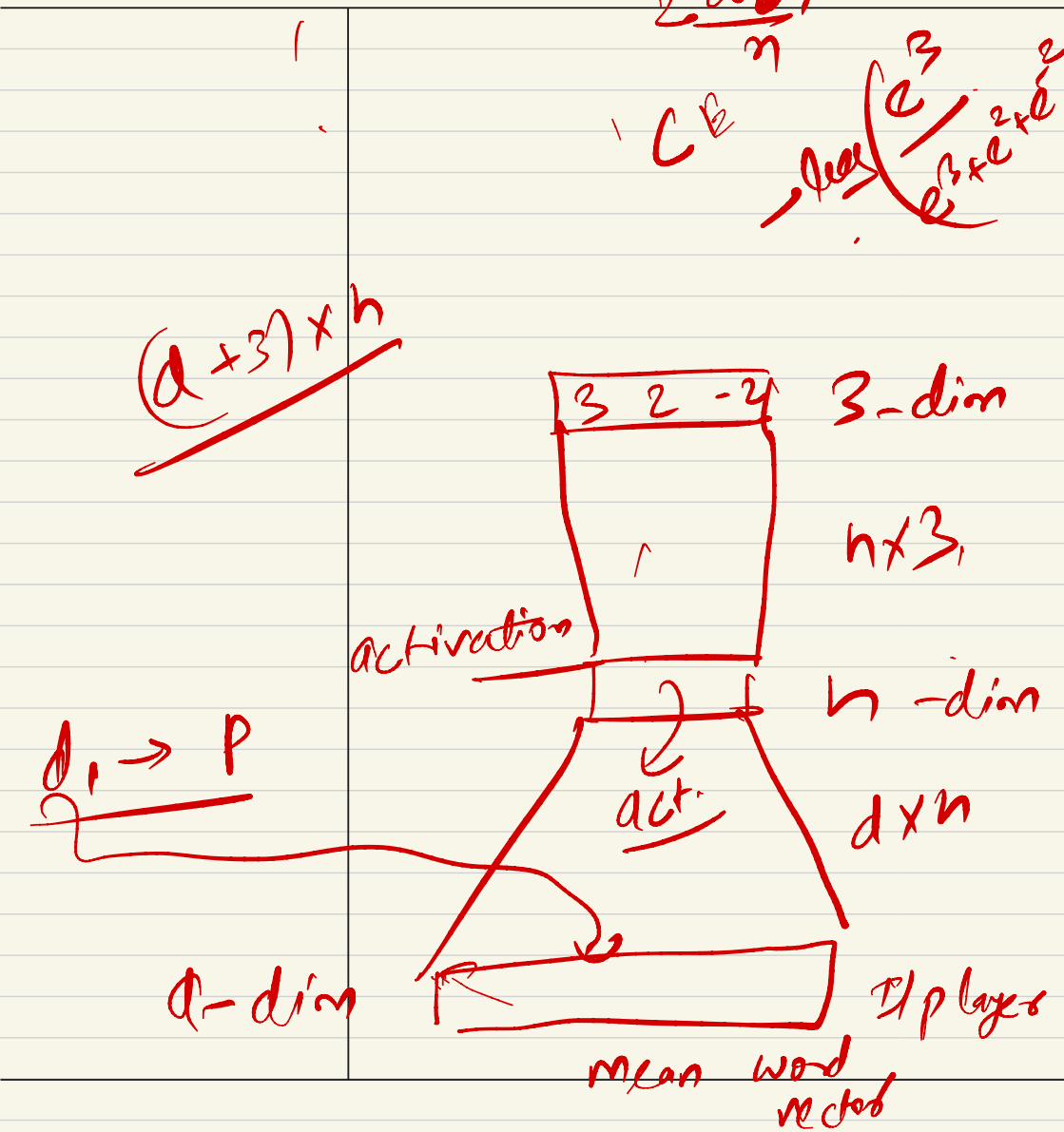
[ ] d_1

d_2

S unique

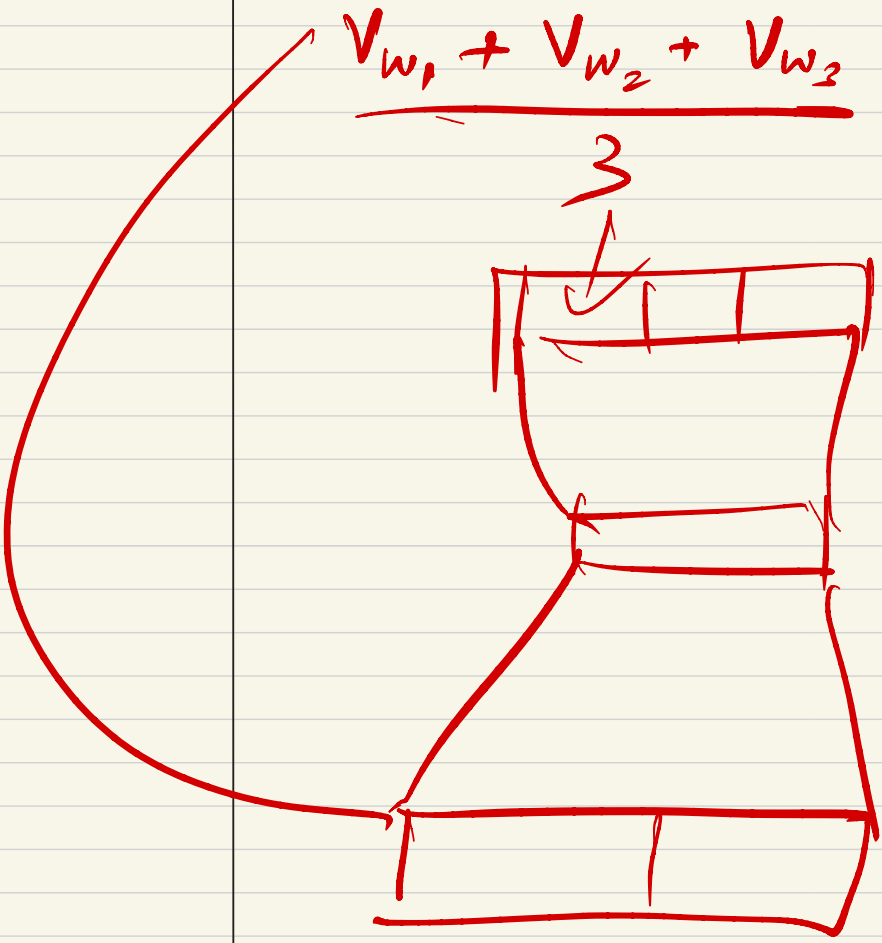
P words in d

$\frac{\sum w_i}{n}$



w_1 w_2 w_3

$$\frac{V_{w_1} + V_{w_2} + V_{w_3}}{3}$$

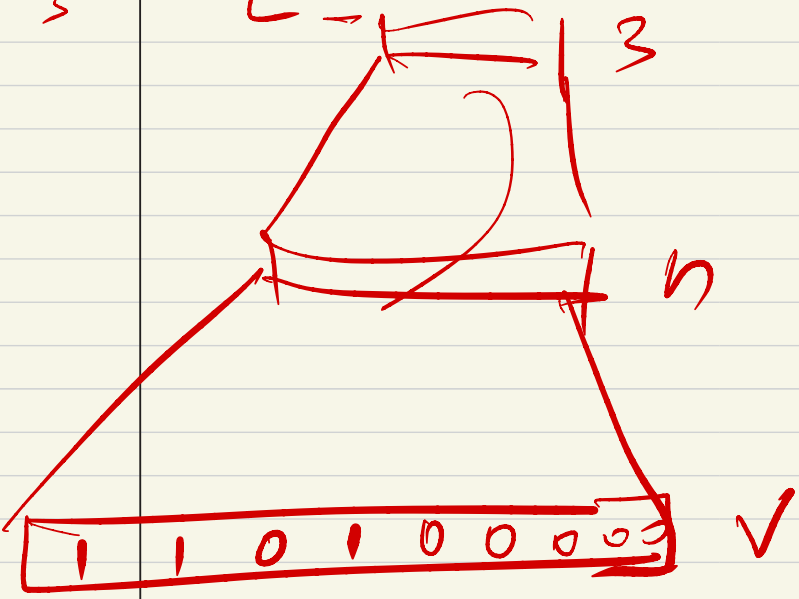


V

$w_1 = [0\ 1\ 0\ 0\ 0\ 0]$

$w_2 = [0\ 0\ 0\ 1\ 0\ 0]$

$w_3 = [1\ 0\ 0\ 0\ 0\ 0]$



multi-hot

$[1\ 1\ 0\ 1\ 0\ 0]$

multi-hot

10000 words

apple 1

doc.

3

apple eat

eat 500

drive

[1-0-1-0-1-0-] 10000

500

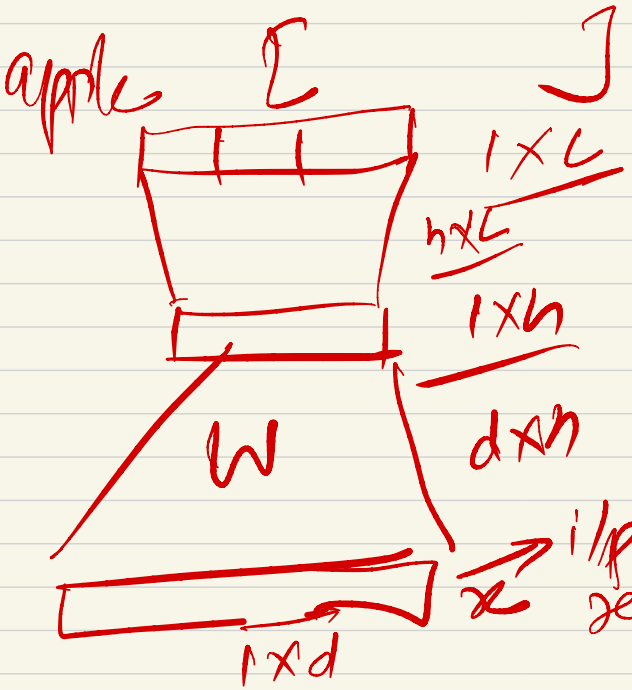
Zebra 10000

Word Vector
≡ Word Embeddings

Word2 Vec 300d/100d

Glove

3



$\sigma(Wx+b)$

Multi-hot vector

P(the students opened their books)

= P(the) P(students | the)

P(opened | the students)

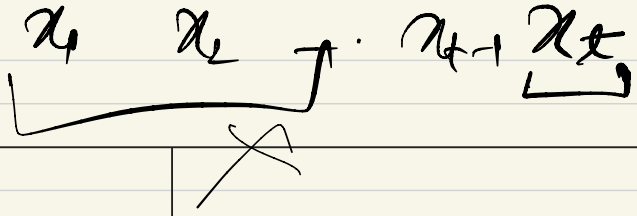
P(their | the students opened)

P(books | the student opened their)

$\frac{\#\ \text{the}}{\sum_{w} \#\ w}$

$\frac{C(\text{the, student})}{C(\text{the})}$

$\frac{C(\text{the students opened})}{C(\text{the students})}$



$P(x_t \mid x_1 \ldots x_{t-1})$

1st order ≈ Markov

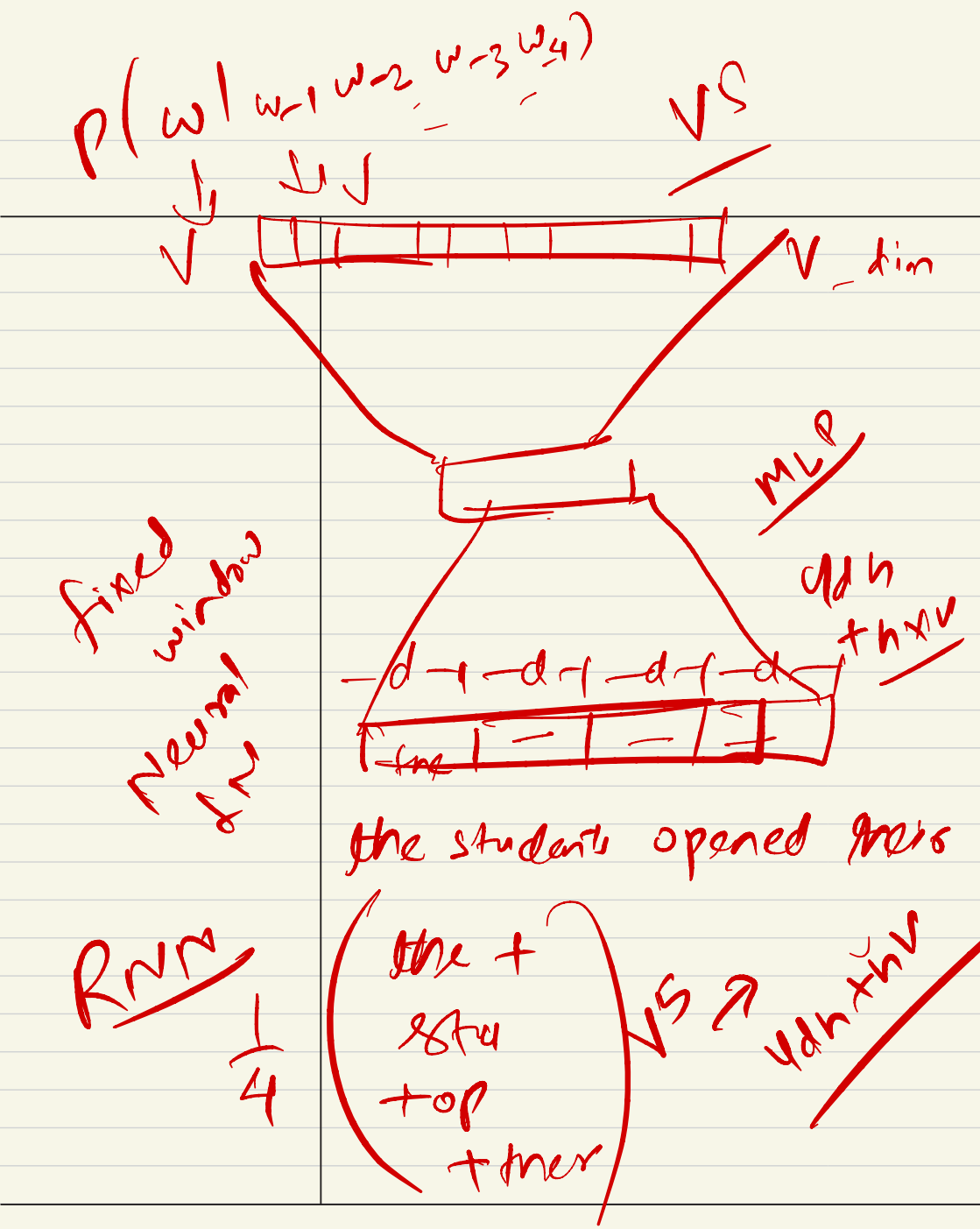
$P(x_t \mid x_{t-1})$

P(the students opened their books)

= P(the) P(students) ... P(books)

Unigram LM

bigram LM P(the) P(students | the)


P(w | w-1 w-2 w-3 w-4)
VS
V
V
V_dim
MLP
fixed window neural LM
the students opened their
RNN
1/4
the + stu + op + ther
VS

S_1

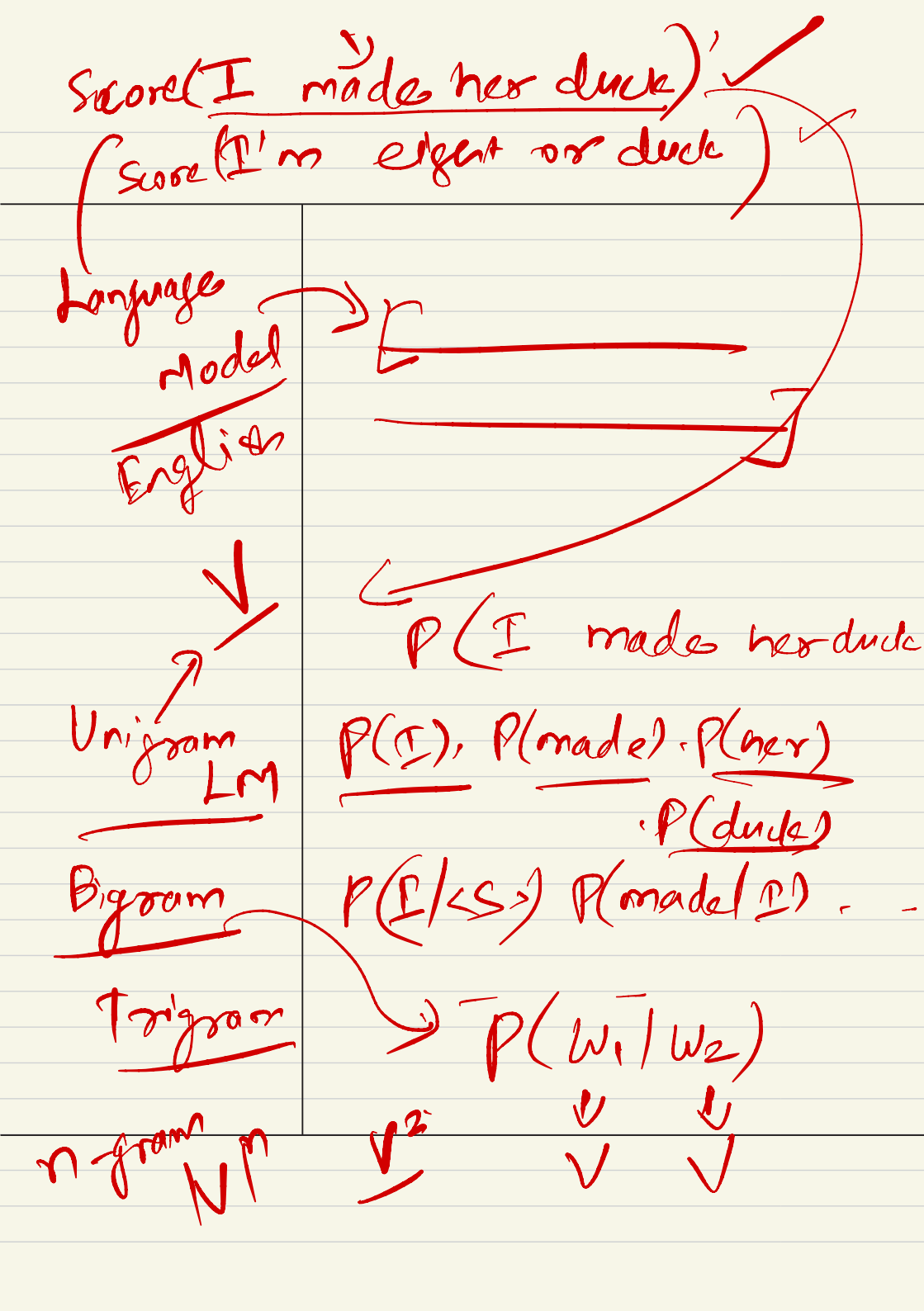
S_2

d_1

d_2

mean embedding

$$\left(\frac{w_1 + w_2 - - \quad w_n}{n}\right)\left(\frac{w_1' + \ldots w_m'}{m}\right)$$


Score(I made her duck)
(Score(I'm eight or duck
Language model
English
P(I made her duck
Unigram LM
P(I) P(made) P(her)
P(duck)
Bigram
P(I/<s>) P(made/I)
Trigram
P(w1/w2)
n gram
V^2
V
V

$P(\text{I made her duck})$

$= P(\text{I}) \cdot P(\text{made} \mid \text{I})$

$P(\text{her} \mid \text{I made}) =$

$P(\text{duck} \mid \text{made her})$

3

$$\frac{C(\text{I made her})}{C(\text{I made})}$$

$P(\text{I'm super or duck})$

I made her duck.

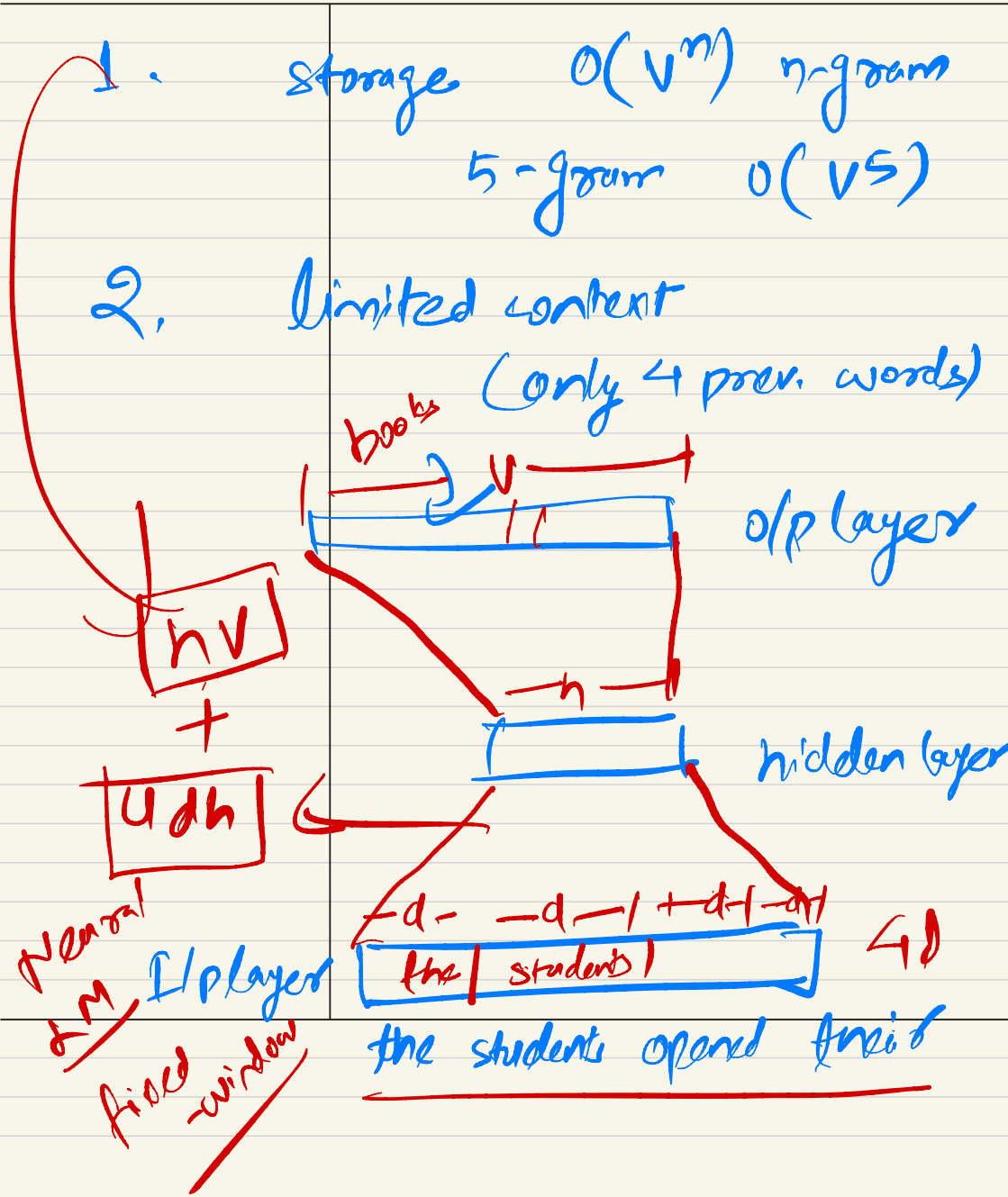
$\arg\max_{w \in V} P(w \mid \text{made her}) \rightarrow$ w?

Problems

1. storage $O(V^n)$ n-gram

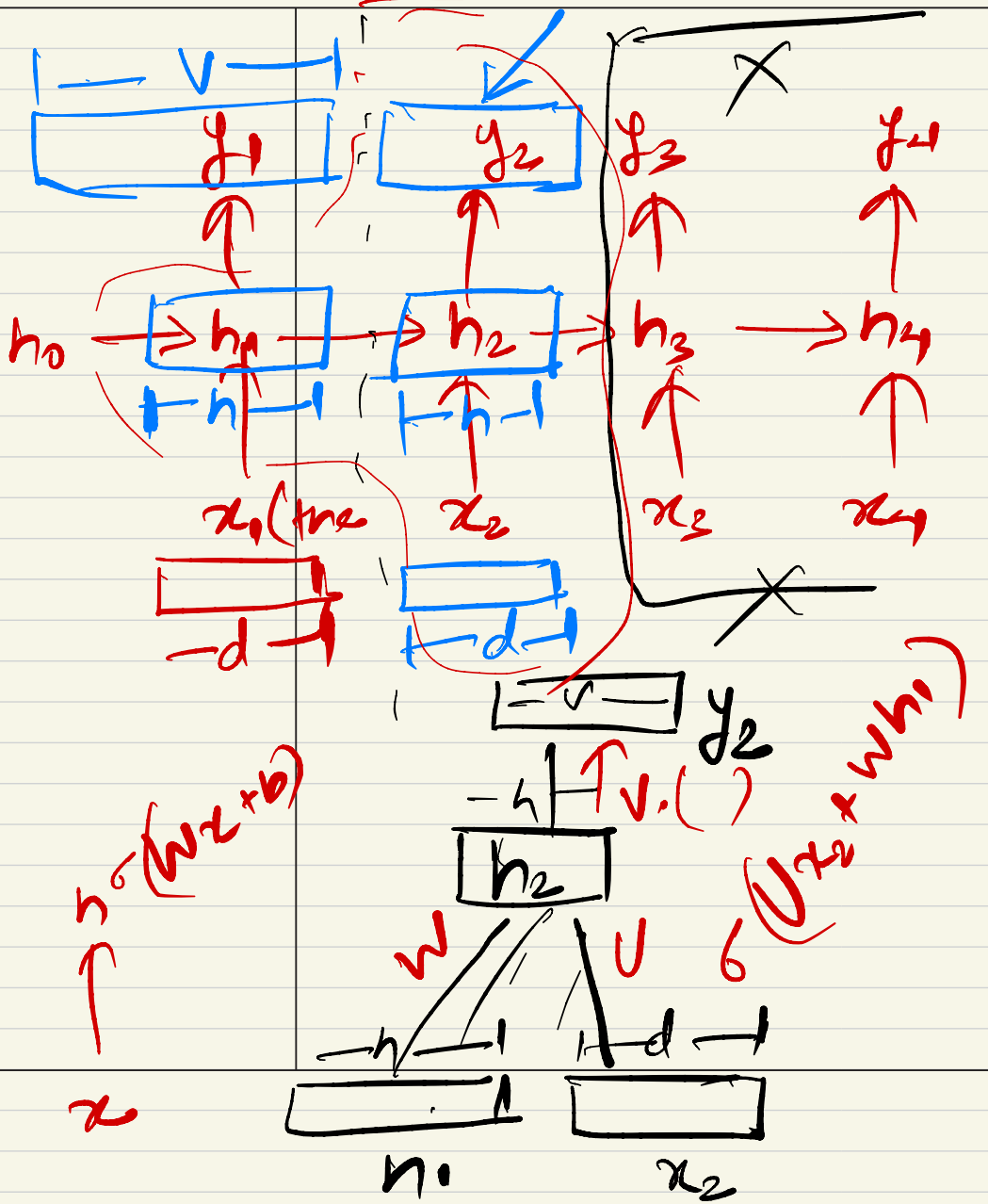
5-gram $O(V^5)$

2. limited context (only 4 prev. words)

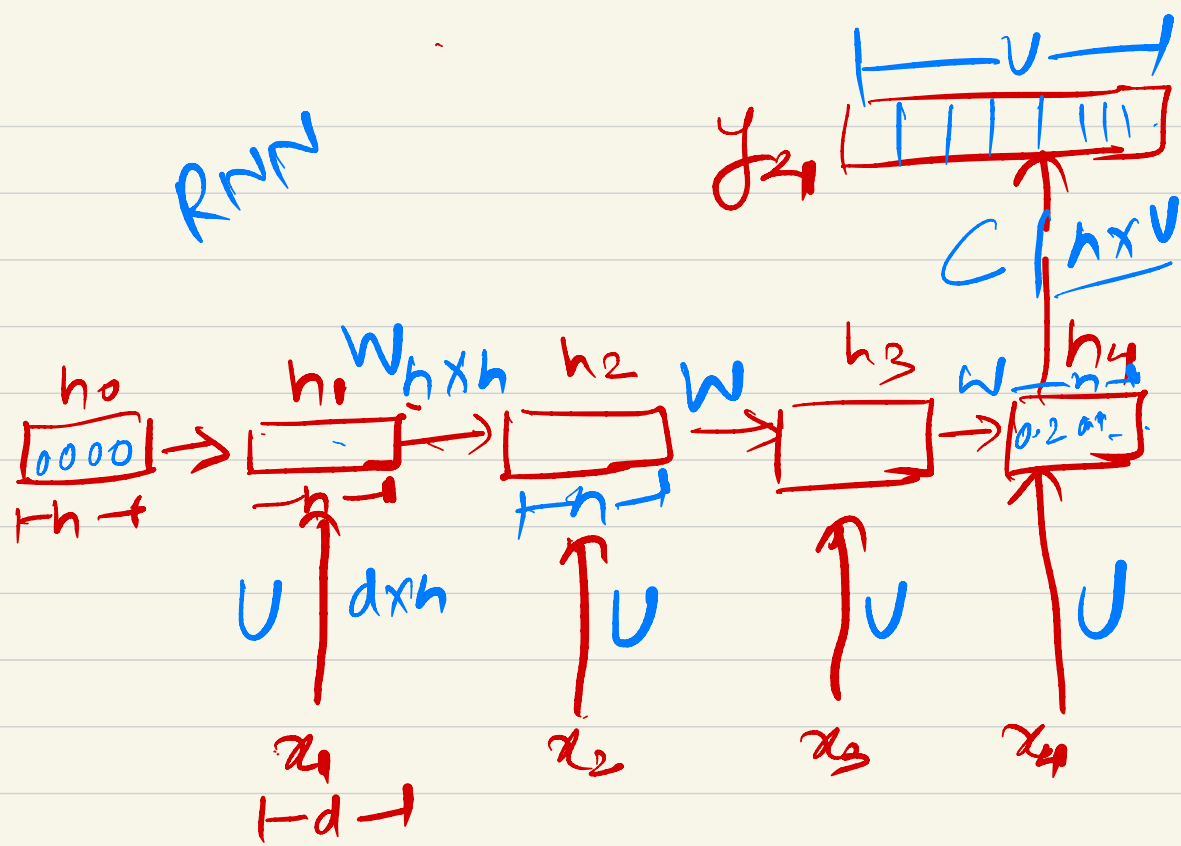


the students opened their ___

RNN



RNN



U, W, C

dxh + hxh + hxV

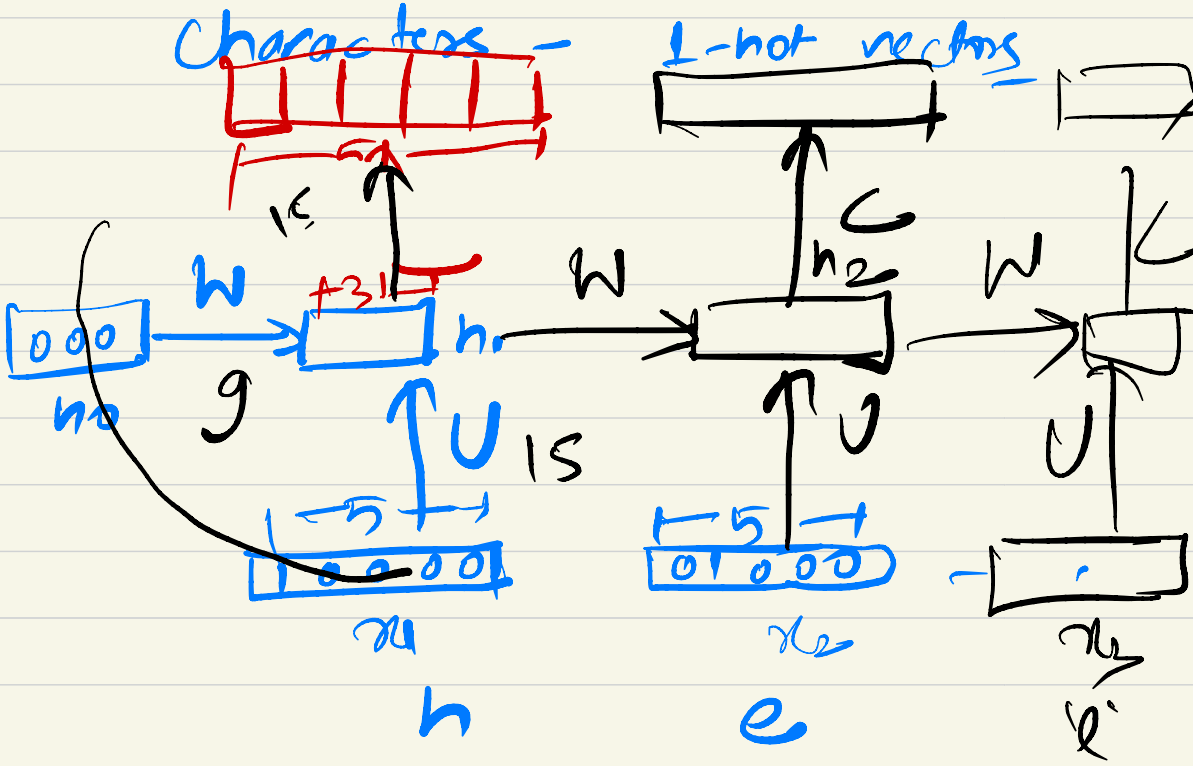
RNN for next characters prediction

5 chars = ['h', 'e', 'l', 'o', '#']

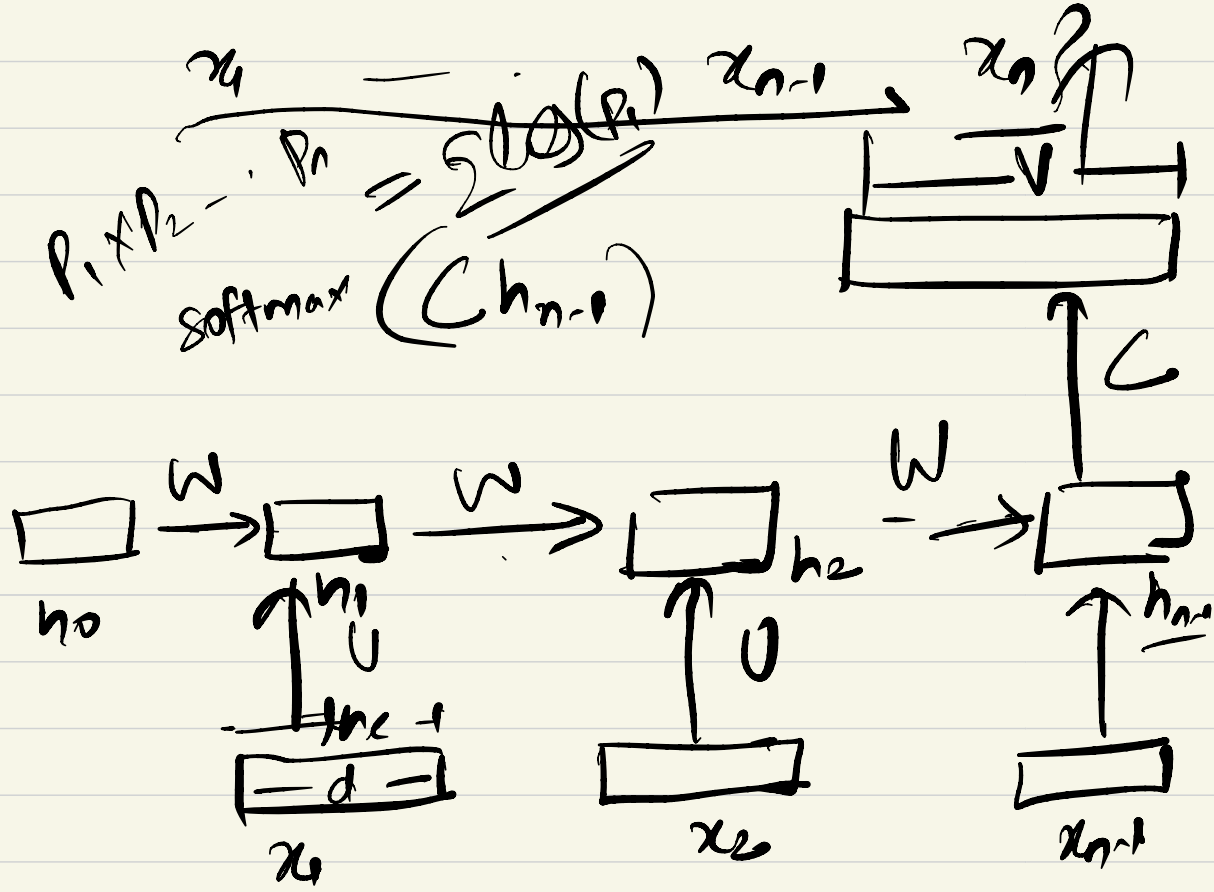
↳ RNN

Recurrent state - 3-dim

$h = 3$

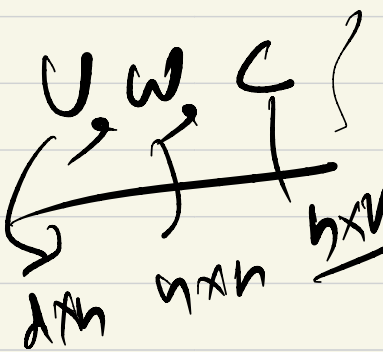


RNN LM

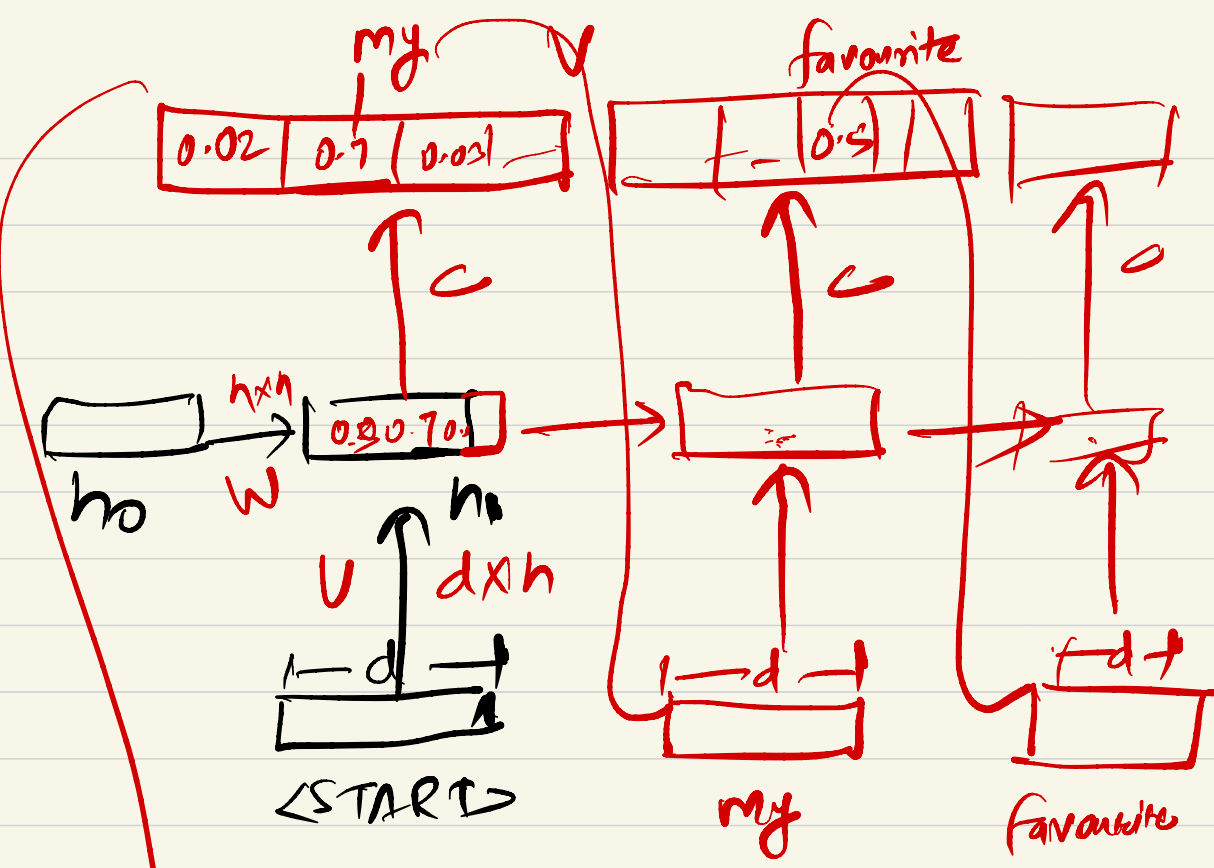


$h_1 = \tanh(Wh_0 + Ux_1)$

$W_{n \times n}$ $U_{d \times n}$



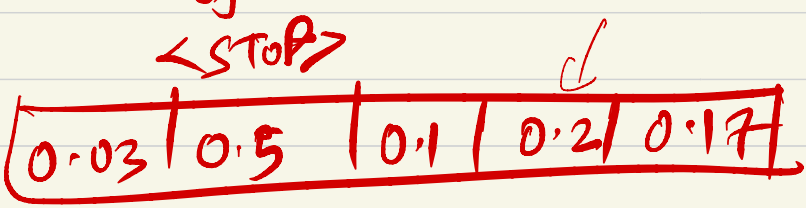
BPTT



Generation

of

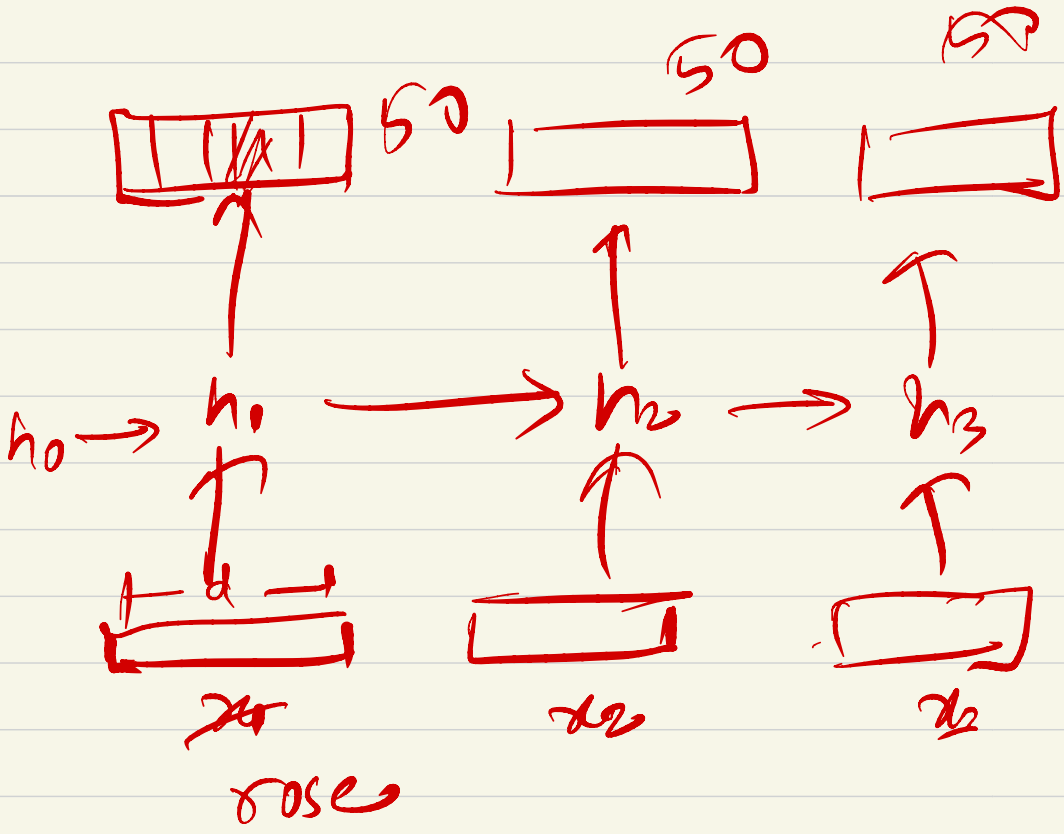
<STOP>



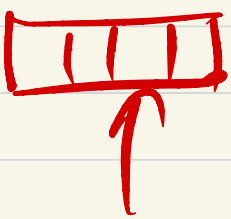
Sampling

RNN for POS Tagging → 50

The stuffed cat knocked - -



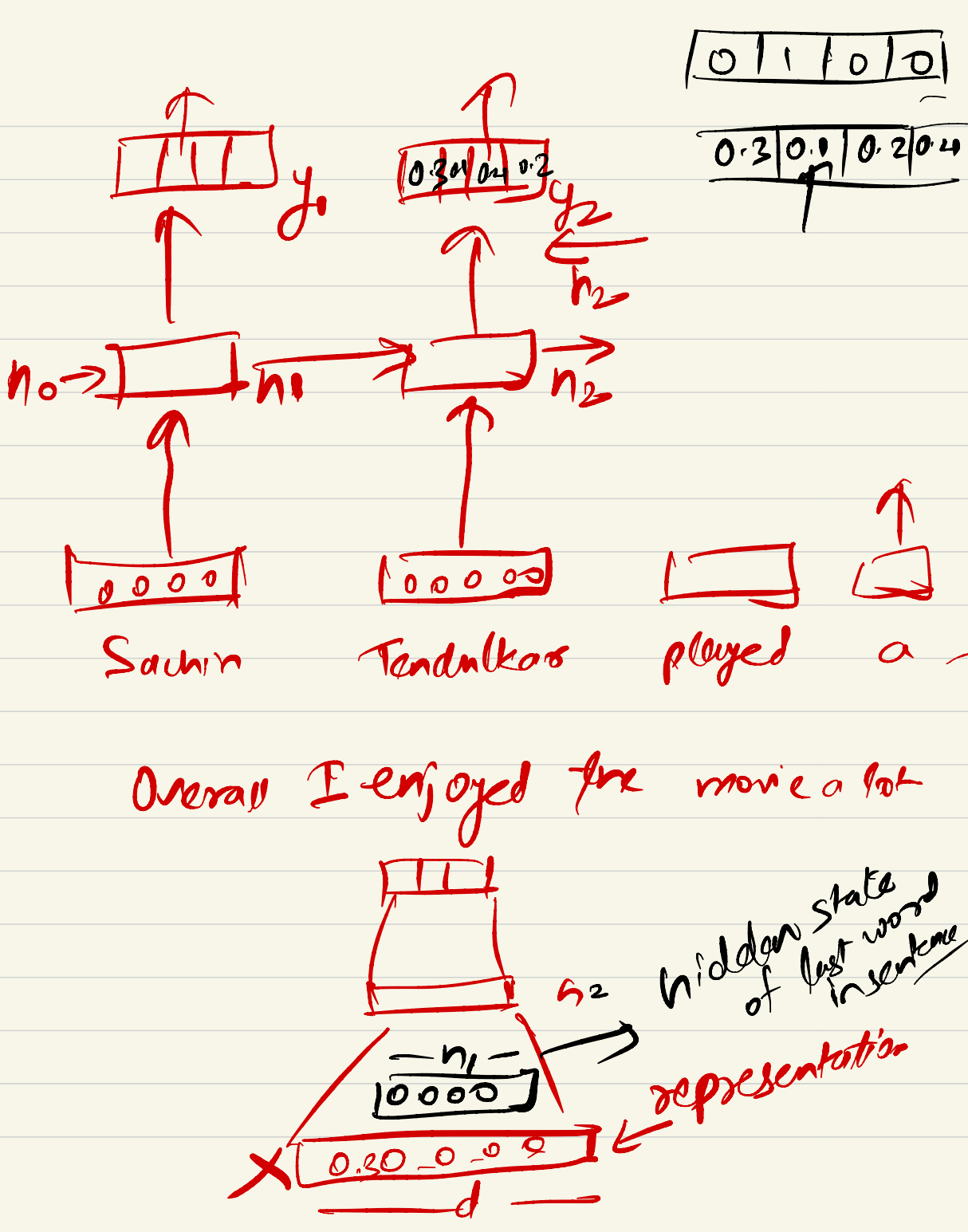
Name of a Person
Location
Organization
Other+1

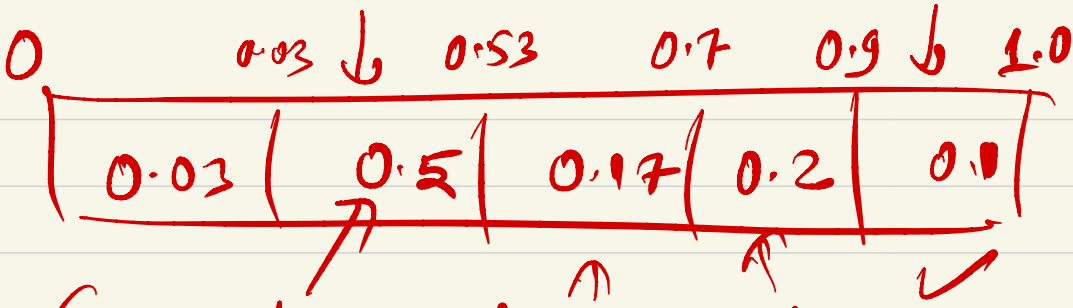


w_1 w_2 w_3 ...

Name Name

Sachin Tendulkar played a brilliant knock.


0 1 0 0
0.3 0.1 0.2 0.4
y_0
y_1
y_2
h_0
h_1
h_2
Sachin
Tendulkar
played
a
Overall I enjoyed the movie a lot
s_2
hidden state of last word in sentence
h_1
representation
X
d



Generate a random number between [0 – 1]

GRU

Update gate $u^{(t)} = \sigma(W_u h_{t-1} + U_u x_t + b_u)$

Reset gate $r^{(t)} = \sigma(W_r h_{t-1} + U_r x_t + b_r)$

$\tilde{h}_t = \tanh\left(W_h \left(r^{(t)} \circ h^{(t-1)}\right) + U_h x_t + b_h \right)$

$h_t = u_t \circ \tilde{h}_t + (1 - u_t) \circ h_{t-1}$

→ select useful parts of prev. hidden state

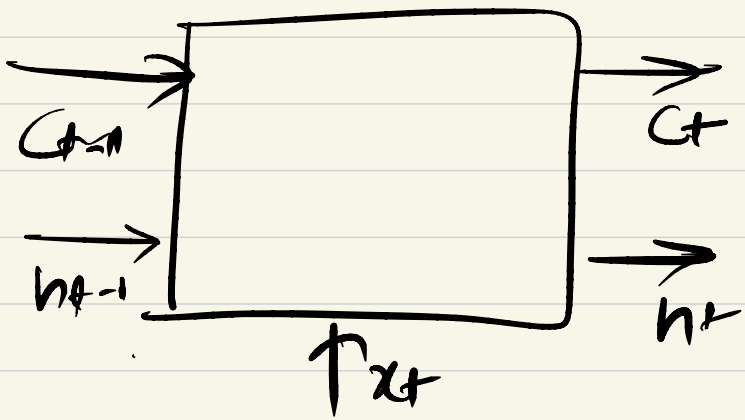
→ What is kept from prev. hidden state & what is ignored.

$u_t = 0$ old content

$r^{(t)} = 0 \Rightarrow \tilde{h}_t$ → only new content

$u_t = 1 \Rightarrow$ only new

LSTM

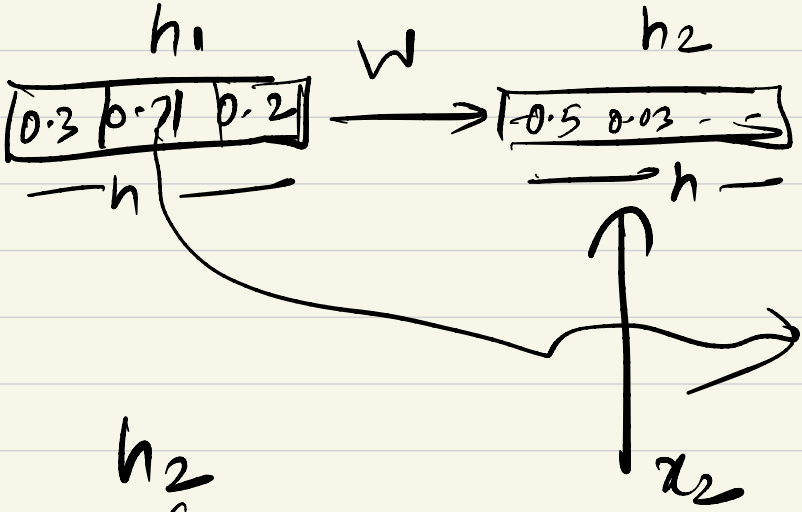


forget gate input gate

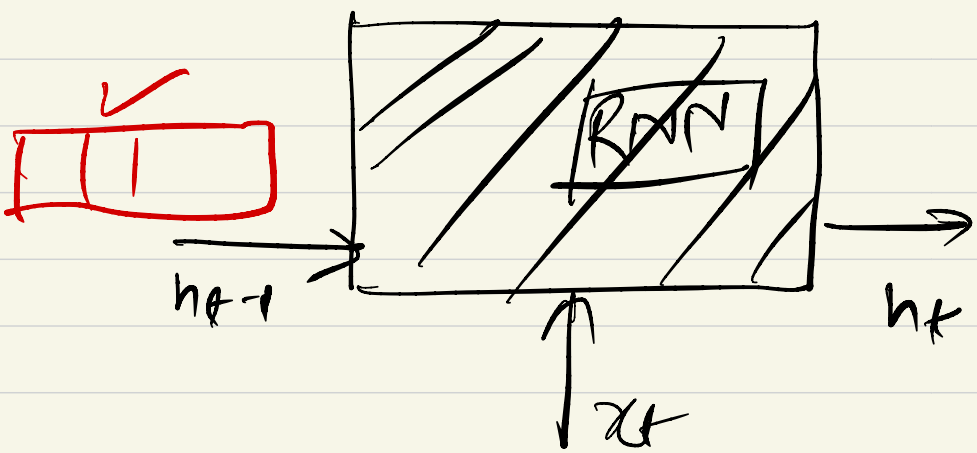
$$f_t = \sigma(W_f h_{t-1} + b_f x_t + b_f)$$
$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$
$$O_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$

$$\tilde{C}_t = \tanh(W_c h^{(t-1)} + U_c x^{(t)} + b_c)$$
$$C_t = f^{(t)} \circ C^{(t-1)} + i^{(t)} \circ \tilde{C}^{(t)}$$

$$\boxed{h^{(t)} = O^{(t)} \circ \tanh C^{(t)}}$$

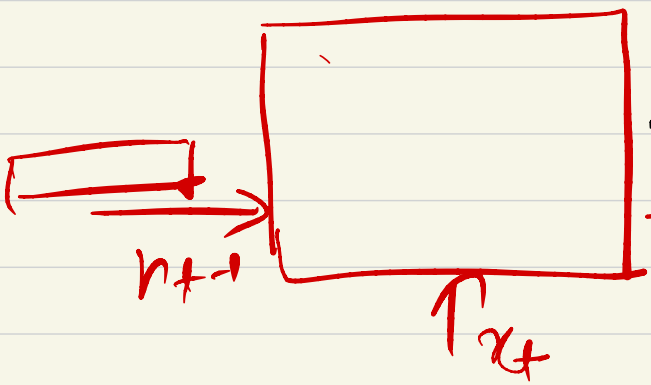


$$h_2 = \sigma(W h_1 + U x_2)$$



Vanilla - $\sigma(U x_t + W h_{t-1})$

GRU



h^{t-1} [0.5 0.3 0.2]

$u^{(t)} = [0 \; 1 \; 1]$

$r^{(t)} = [1 \; 0 \; 1]$

$h^{(t)} = [0.5 \; \ldots]$ (new)

Gates

Update

Reset

$$u^{(t)} = \sigma(W_u h_{t-1} + U_u x_t + b_u)$$

$u^{(t)} = 0$ → old info

$r^{(t)} = 0$ $u^{(t)} = 1$ → New info

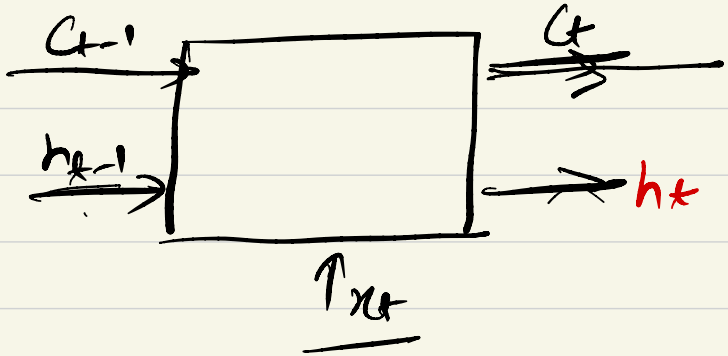
[0 1 0 1 0 0 1]

$$r^{(t)} = \sigma(W_r h_{t-1} + U_r x_t + b_r)$$

$$\tilde{h}_t = \tanh(W_h(r^{(t)} \circ h_{t-1}) + U_h x_t + b_h)$$

$r^{(t)} = [1 \; 0 \; 0 \; 1 \; 0 \; 1 \; 1 \; \ldots]$

$$h_t = (1 - u^{(t)}) \circ h_{t-1} + u^{(t)} \circ \tilde{h}_t$$



I/p Gate $i^{(t)} = \sigma(W_i h_{t-1} + U_i x_t + b_i)$

forget Gate $f^{(t)} = \sigma(W_f h_{t-1} + U_f x_t + b_f)$

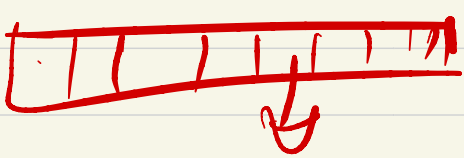
Output Gate $o^{(t)} = \sigma(W_o h_{t-1} + U_o x_t + b_o)$

$$\tilde{C}^{(t)} = \tanh(W_c h_{t-1} + U_c x_t + b_c)$$

$$C^{(t)} = f^{(t)} \circ C_{t-1} + i^{(t)} \circ \tilde{C}^{(t)}$$

$$h^{(t)} = o^{(t)} \circ \tanh C^{(t)}$$

$C^{(t)}$



off | 0 0 0 0 1 0 0 0 |

ELMO ✓

BERT ✓

CS 224n

{XLNet}

{NLP with Deep Learning}

Retrieval Based

Chatbot

~~GAN~~

Generative