Neural Models for NLP

Amit Awekar

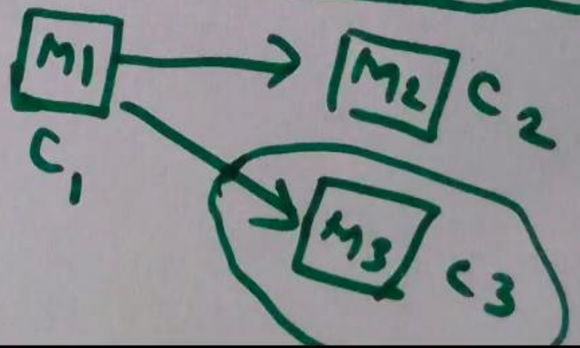
Language : System, Symbols, Convey meaning

Natural : Human-Human, Machine-Human, Ambiguity

Processing : Goal driven, Task oriented

Models : Machine Learning : Input → Output, Cost, Data Parameters Optimizer

Neural : Brain



Input	Output
1	2
2	6
3	10
40	38

Model

$y = a \cdot x + b$

Parameters

$a = \#\ 21$

$b = \#\ 1$

$M1 : y = x$

$M2 : y = 2x$

$M3 : y = x + 1$

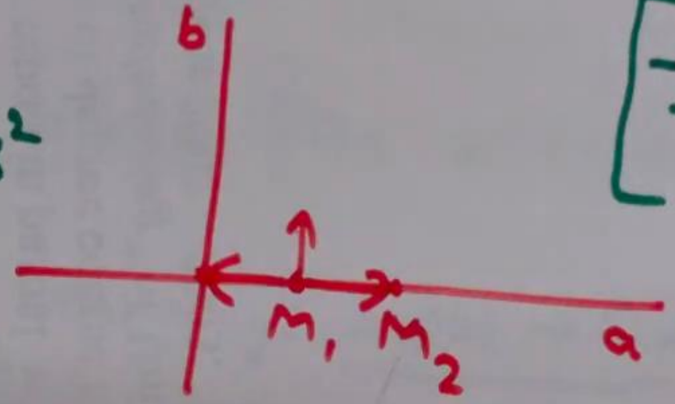
Optimizer

$$\left[\frac{-\partial C}{\partial a}, \frac{-\partial C}{\partial b}\right]$$

Cost Function

$C(M_1) = 1^2 + 4^2 + 7^2 + 28^2$

$$\sum_{i=1}^{d} (y_i - y_{im})^2$$



$C(M_2) = 0 + 2^2 + 4^2 + 18^2$

$C(M_3) = 0 + 3^2 + 6^2 + 27^2$

Step Size

1 unit

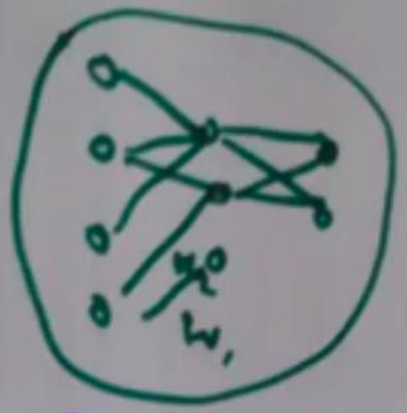
$$a = a - \left[\frac{\partial C}{\partial a} \cdot \text{Step Size}\right]$$

$$b = b - \left(\frac{\partial C}{\partial b} \cdot \text{Step Size}\right)$$

Data

Tweet 1 +ve 1
Tweet 2 -ve -1
⋮ ⋮
Tweet 10^6 +ve

Model



Parameters
$w_1 = 0.1$
w_2
⋮
w_{1000}

Model Sequence

M_1
M_2
⋮
M_{20}

Cost Function

$C(M_1)$

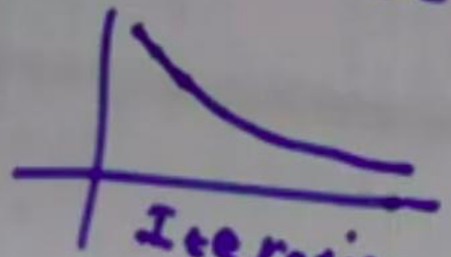
$$\sum_{i=1}^{n} (y_i - y_{im})^2$$

Optimizer

$$w_1 = w_1 - \frac{\partial C}{\partial w_1} \cdot \text{Step Size}$$

Back propogation

Step Size



Motivation for ANN

Neurons : 10^{11} : Total

10^{4} : Connections per neuron

Excited / Inhibited : States of neuron

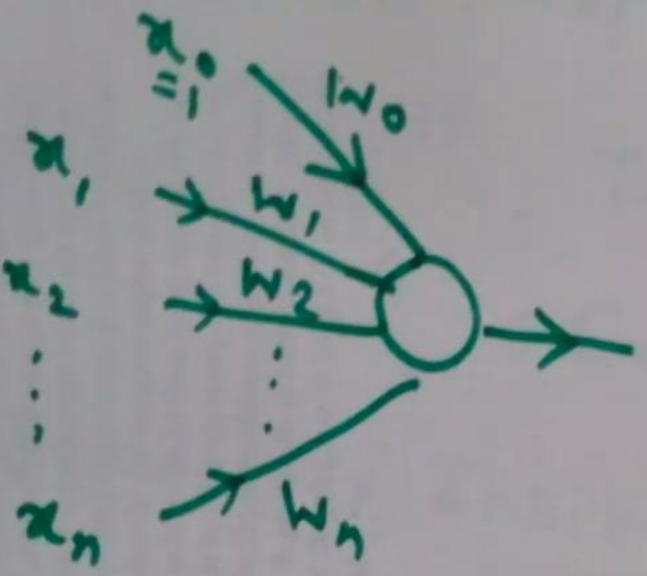
10^{-3} sec : Speed of neurn

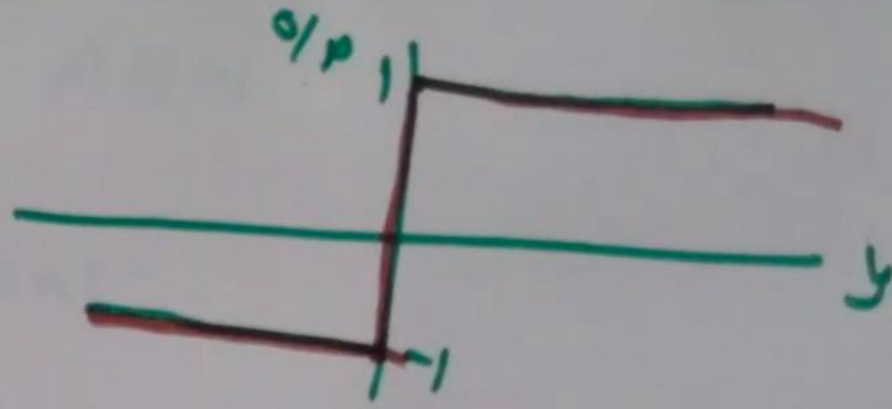
10^{-1} sec : Time required for many human decisions

Parallel

Distributed across multiple neurons

Perceptron





output = 1 if $y > 0$

-1 else

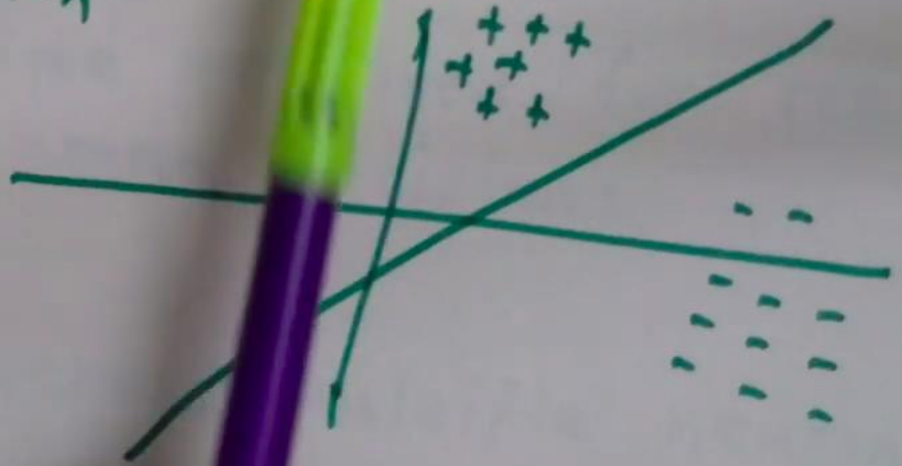
$w_0 + x_1 \cdot w_1 + x_2 \cdot w_2 \ldots + x_n \cdot w_n$

$$y = \sum_{i=0}^{n} \vec{x_i} \cdot \vec{w_i}$$

$$w_i = w_i + \underline{\Delta w_i}$$

$$\downarrow$$

$$\eta (t - o) x_i$$



Descent

$$o(\vec{x}) = \vec{w} \cdot \vec{x}$$

$$\frac{1}{2}\sum_{d \in D}(t_d - o_d)^2 \leftarrow E(\vec{w})$$

$$\nabla E(\vec{w}) \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1} \cdots \frac{\partial E}{\partial w_n}\right]$$

$$\vec{w} = \vec{w} + \Delta\vec{w}$$

$$\Delta\vec{w} = -\eta \cdot \nabla E(\vec{w})$$

$$w_i = w_i + \Delta w_i$$

$$\Delta w_i = -\eta \cdot \frac{\partial E}{\partial w_i}$$

$$\frac{1}{2}\sum_{d\in D}\frac{\partial}{\partial w_i}(t_d - o_d)^2$$

$$\frac{1}{2}\sum_d 2(t_d - o_d)\frac{\partial}{\partial w_i}(t_d - o_d)$$

$$\vec{w}\cdot\vec{x}_d$$

$$w_0 + w_1 x_1 + w_2 x_2 \cdots + w_n x_n$$

$$w_i x_{id}$$

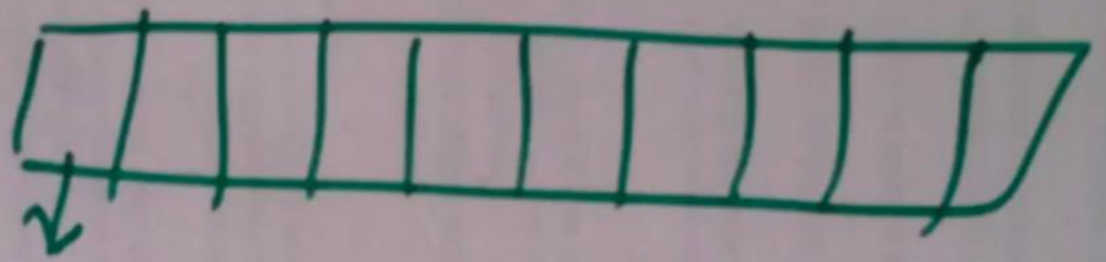
$$\sum_d (t_d - o_d)\, x_{i,d}$$

$$w_i = w_i + \boxed{\Delta w_i} \rightarrow -\eta \cdot \frac{\partial E}{\partial w_i}$$

$$\eta \cdot \sum_d (t_d - o_d)\, x_{i,d}$$

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2$$

$d_1 \rightarrow d_2$

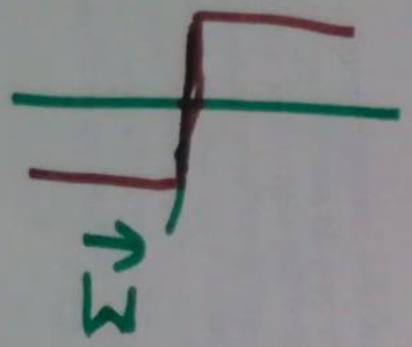


$$E_d(\vec{w}) = \frac{1}{2}(t_d - o_d)^2$$

$$\Delta w_i = n \cdot (t_d - o_d) \cdot x_{i,d}$$

$$E_{d_2}(\vec{w}) = \frac{1}{2}(t_{d_2} - o_{d_2})^2$$

Thresholded



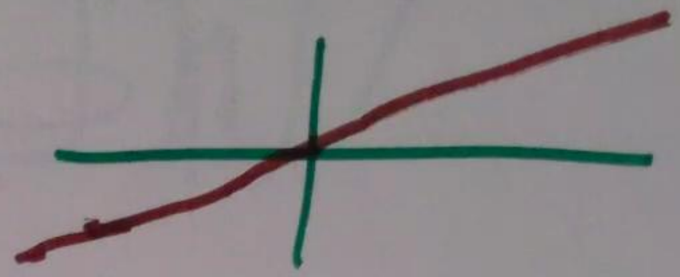
$\vec{w}$

Perceptron rule

$w_i = w_i + \Delta w_i$

$\Delta w_i = \eta \cdot (t_d - o_d) x_{id}$

Linearly separable



$\vec{w}$

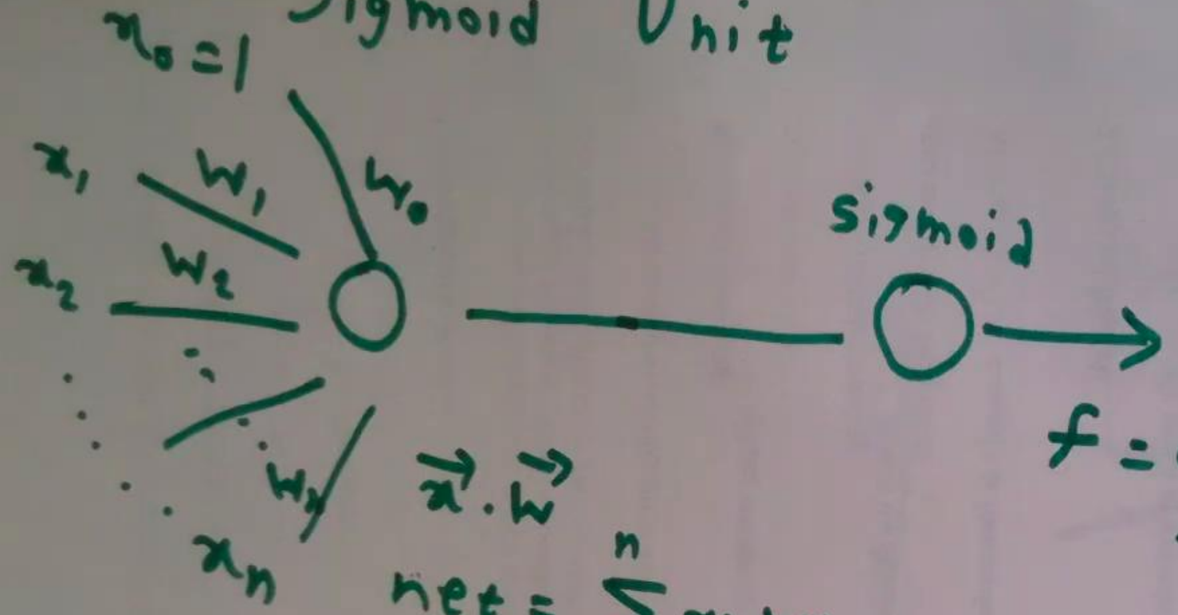
$$\Delta w_i = \eta \cdot \sum_{d \in D} (t_d - o_d) x_{id}$$

Stochastic Gradient Descent

$E(\vec{w})$ $\rightleftarrows$ $E_1(\vec{w})$, $E_2(\vec{w})$

$\Delta w_i = \eta \cdot (t - o) x_i$

Sigmoid Unit



$\vec{x} \cdot \vec{w}$

$$net = \sum_{i=0}^{n} x_i \cdot w_i$$

f = activation function

$$output = f(net)$$

$$sigmoid(net) = \frac{1}{1 + e^{-net}}$$

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}} = y$$

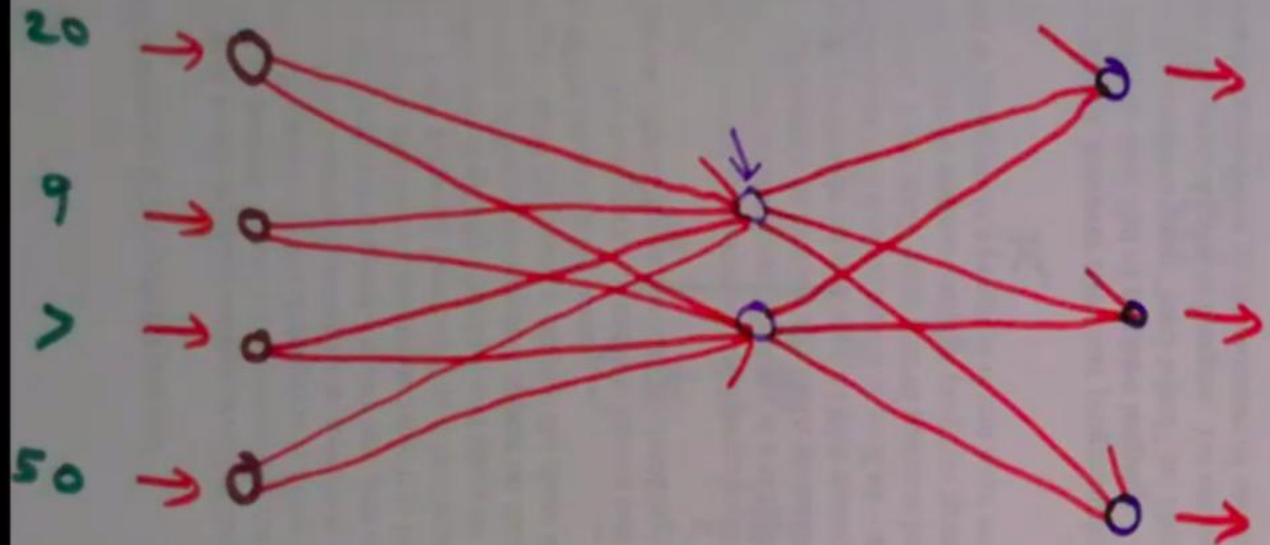
$$\frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) = \frac{-1}{(1+e^{-x})^2}\frac{d}{dx}(1+e^{-x})$$

$$\frac{+1}{(1+e^{-x})^2} \times e^{-x}$$

$$\frac{1e^{-x}+1-1}{(1+e^{-x})^2} = \frac{e^{-x}+1}{(1+e^{-x})^2} - \frac{1}{(1+e^{-x})^2}$$

$$y - y^2$$

n_{in} 4 n_{hidden} 2 n_{out} 3



W_{ji}

x_{ji}

O_{dk}

t_{dk}

$\vec{x}, \vec{t}$

$$W_{ji} = W_{ji} + \Delta W_{ji}$$

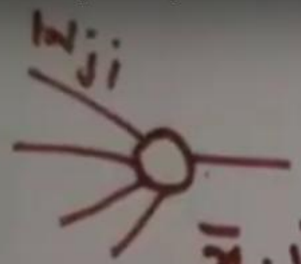
$$\Delta W_{ji} = \eta \cdot \delta_j \cdot x_{ji}$$

Output layer : k

$$\delta_k = O_k (1 - O_k)(t_k - O_k)$$

Hidden layer : h

$$\delta_h = O_h (1 - O_h) \sum_{k \in Output} \delta_k \cdot W_{hk}$$



$\vec{x} \cdot \vec{w}$

net

Sigmoid(net) → Output

$$w_{ji} = w_{ji} + \Delta w_{ji}$$

$$\Delta w_{ji} = -\eta \cdot \frac{\partial E}{\partial w_{ji}} \quad \frac{\partial E}{\partial net_j} \quad \frac{\partial net_j}{\partial w_{ji}} \rightarrow \frac{\partial w_{ji} \cdot x_{ji}}{\partial w_{ji}} \rightarrow x_{ji}$$

$$E_d(\vec{w}) = \frac{1}{2} \sum_{k \in output} (t_k - o_k)^2$$

$$\delta_j = -\frac{\partial E}{\partial net_j}$$

$$\frac{\partial}{\partial net_k}$$

$$= \delta_{net_k}$$

$$o_k$$

$$(t_k - o_k)^2$$

$$-\frac{\partial o_k}{\partial net_k} \cdot \frac{\partial E}{\partial o_k}$$

$$\frac{1}{2}\sum_{k \in output} (t_k - o_k)^2$$

$$-o_k(1-o_k) \cdot \frac{1}{2}\frac{\partial}{\partial o_k}(t_k - o_k)^2$$

$$\frac{1}{2} \cdot 2(t_k - o_k) \cdot \frac{\partial (t_k - o_k)}{\partial o_k}$$

$$o_k(1-o_k) \cdot (t_k - o_k)$$

$$\partial net_j$$

$$\frac{\partial net_k}{\partial net_k} \qquad net_k \qquad o_k$$

$$-\frac{\partial o_k}{\partial net_k} \cdot \boxed{\frac{\partial E}{\partial o_k}} \qquad \frac{1}{2}\sum_{k \in output} (t_k - o_k)^2 \qquad (t_k - o_k)^2$$

$$-o_k(1-o_k) \cdot \frac{1}{2}\frac{\partial}{\partial o_k}(t_k - o_k)^2$$

$$\frac{1}{2} \cdot 2(t_k - o_k) \cdot \frac{\partial (t_k - o_k)}{\partial o_k}$$

$$o_k(1-o_k) \cdot (t_k - o_k)$$

$$w_{ji} = w_{ji} + \Delta w_{ji}$$

$$\Delta w_{ji} = -\eta \cdot \frac{\partial E}{\partial w_{ji}}$$

$$-\eta \cdot \frac{\partial E}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ji}} x_{ji}$$

$$\delta_j = \sum_{k \in Downstream(j)} \frac{\partial E}{\partial net_k} \cdot \frac{\partial net_k}{\partial net_j}$$

$$x_{ji} \cdot \eta \cdot \sum \delta_k \cdot \frac{\partial o_j}{\partial net_j}$$

$$x_{ji} \cdot \eta \cdot o_j \cdot (1 - o_j) \sum \delta_k \cdot w_{kj}$$

$$\delta_j$$

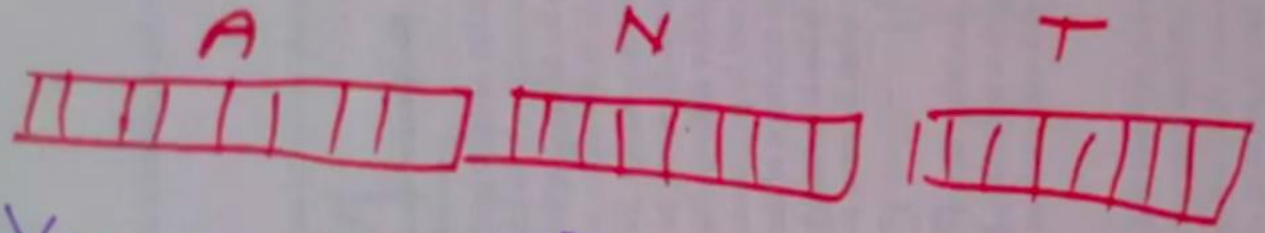
Data Representation : Vector

Characters, Words, Phrases, Sentences, Paragraph,

Document, Section, Book / Collection, Corpus

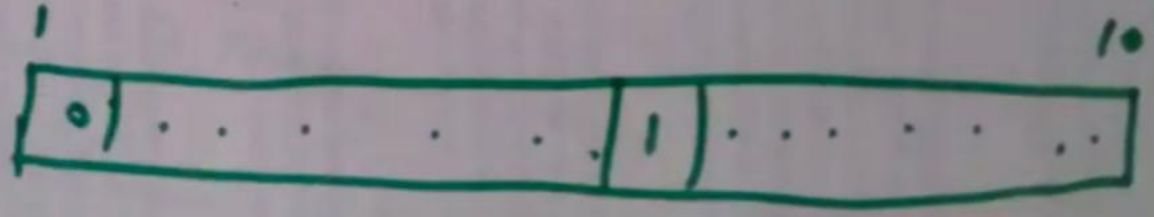
Audio

Characters: ASCII (8 bits) UNICODE (16 bits)



Vocab size: 10^5

One hot encoding:



CBoW

Given: Contex

Predict: Word

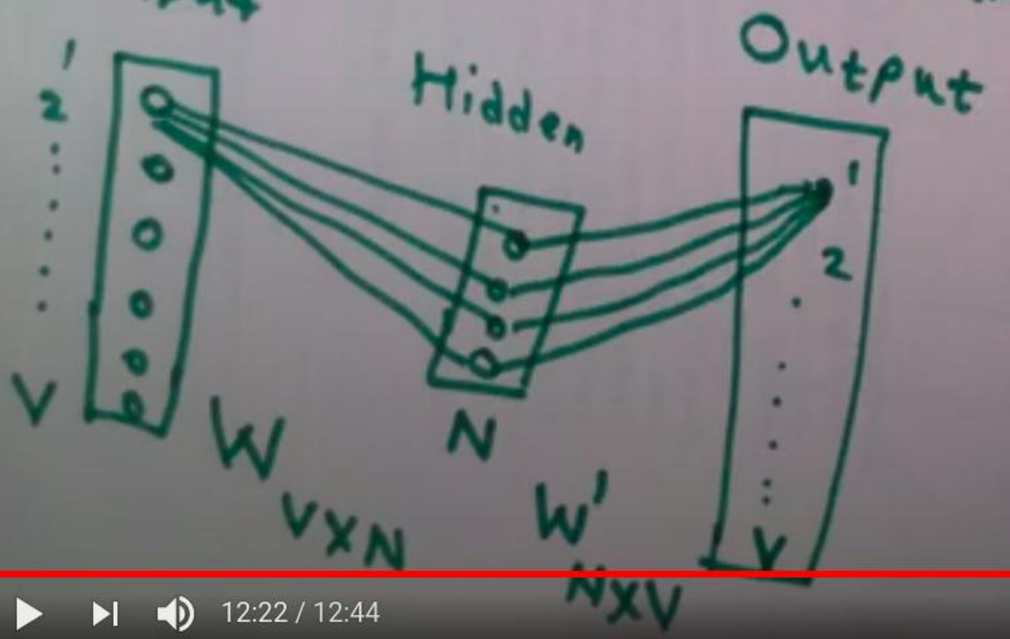
Skip Gram

Given: Word

Predict: Content

IIT (1) Guwahati (2) is (3) located (4) in (5) north (6) eastern

State (1) of (2) Assam.



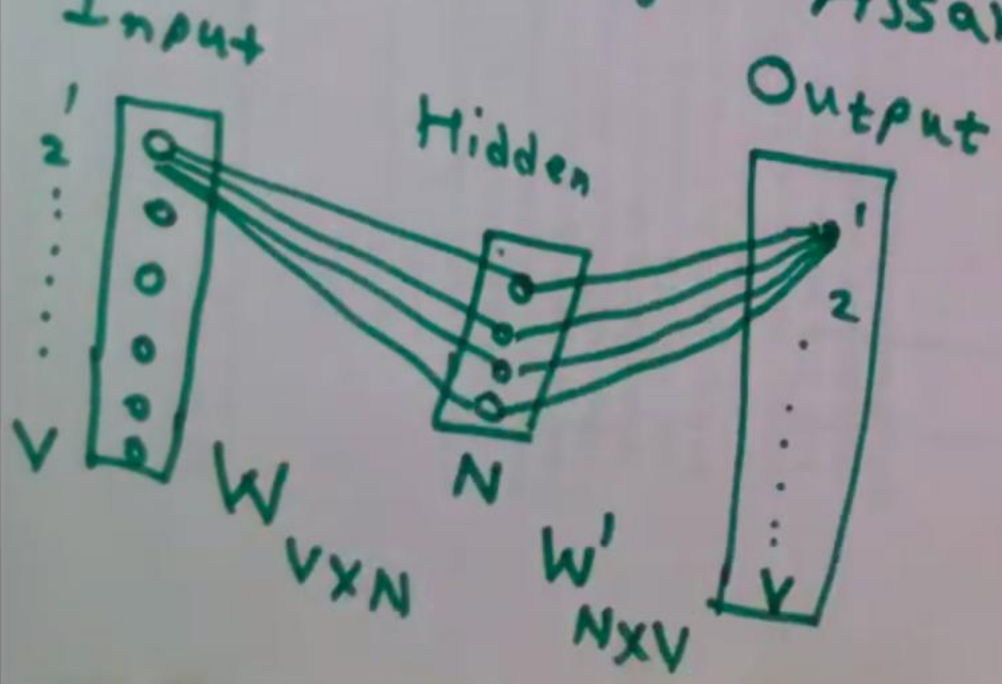
Predict: Word

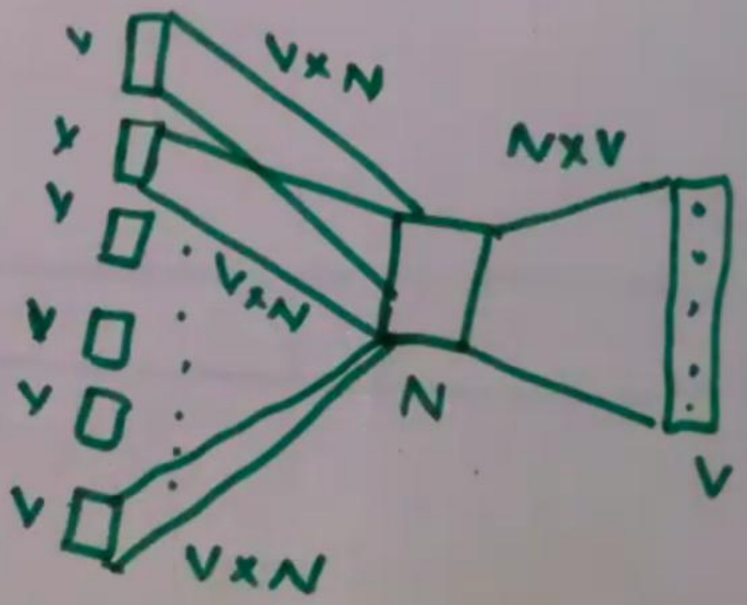
Given: Word

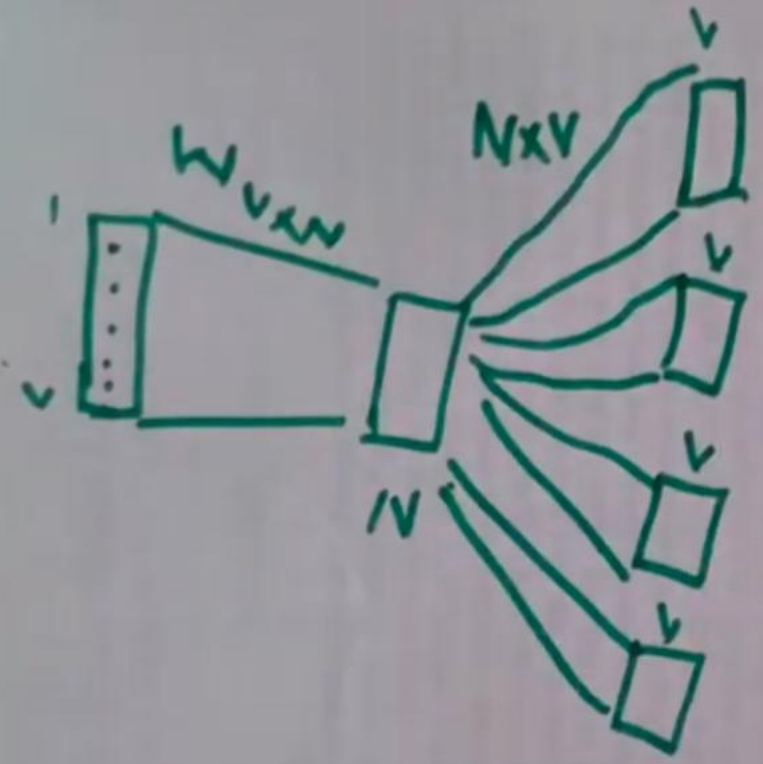
Predict: Content

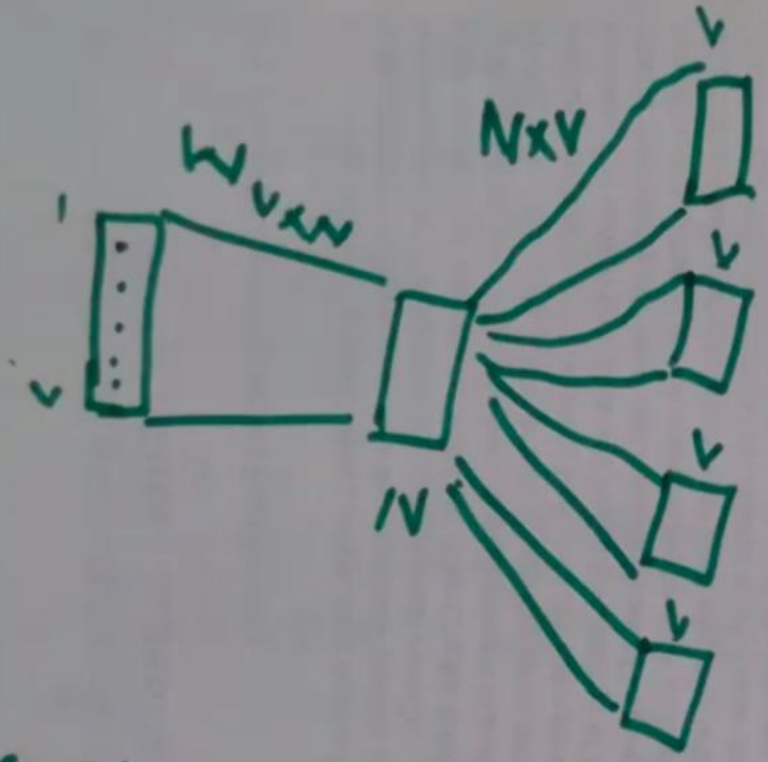
IIT (1) Guwahati (2) is (3) located (4) in (5) north eastern (6)

State (1) of (2) Assam.






V
W V×N
N×V
V
V
V
V
N
V



$$P(w_i \mid w_j)$$

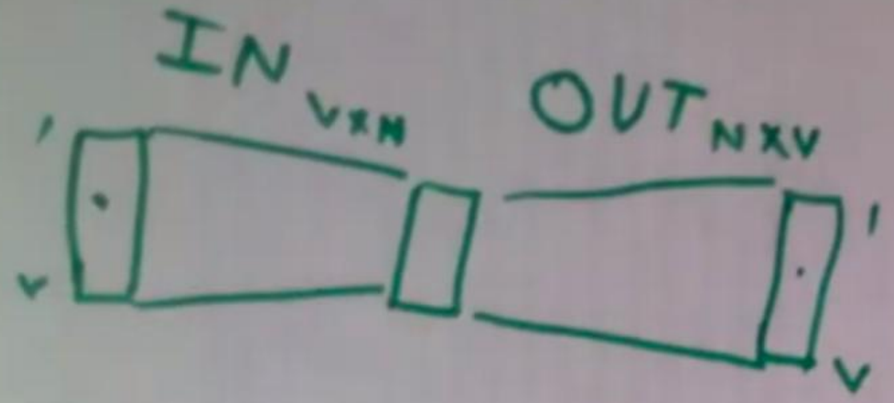
$$P(w_2 \mid w_j)$$

$$P(w_3 \mid w_j)$$

$$\prod_{j=1}^{T} \prod_{i=1}^{k} P(w_i \mid w_j) = L(\theta)$$

$$F(\theta) = -\frac{1}{T} \log(L(\theta)) = -\frac{1}{T} \sum_{j=1}^{T} \sum_{i=1}^{k} P(w_i \mid w_j, \theta)$$

$P(w_j | \underline{w_i}, \Theta)$



Input
located
located
.
.
.

$\underline{v_i}$

Output
Guwahati
in
.
.
.

u_j

Output Value

$w_j : v_i \cdot u_j : e^{v_i \cdot u_j}$

$w_{j+1} : v_i \cdot u_{j+1} : e^{v_i \cdot u_{j+1}}$

Soft Max

$$w_j : \frac{e^{v_i \cdot u_j}}{\sum_{k=1}^{V} e^{v_i \cdot w_k}}$$

$$p(w_{j^*} \mid w_i, \theta) = \frac{\exp(u_{j^*})}{\sum_{j'=1}^{V} \exp(u_{j'})} = y_{j^*} \qquad E = -\log(y_{j^*})$$

$$-t_j + y_j$$

$$-u_{j^*} + \log \sum_{j'=1}^{V} \exp$$

$$\frac{\partial E}{\partial w'_{ij}} = \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial w'_{ij}} \qquad h_i$$

$$\frac{\partial u_{j^*}}{\partial u_j} = \begin{cases} 0 & \text{if } j \neq j^* \\ 1 & \text{if } j = j^* \end{cases}$$

$$w'_{ij} = w'_{ij} - \eta \cdot \frac{\partial E}{\partial w'_{ij}}$$

$$\frac{\partial \log \sum_{j'=1}^{V} \exp(u_{j'})}{\partial u_j}$$

$$\frac{1}{\sum_{j'=1}^{V} \exp(u_{j'})} \sum_{n=1}^{V} \frac{\partial \exp(u_n)}{\partial u_j}$$

$$\rightarrow \frac{\exp(u_j)}{\sum_{j'=1}^{V} \exp(u_{j'})}$$

$$w'_{ij} = w'_{ij} + \Delta w'_{ij}$$

$$-\eta \cdot h_i (y_j - t_j)$$

$$e_j$$

$$w'_{ij} - \eta \cdot e_j \cdot h_i$$

$$W_{ki} = W_{ki} + \underline{\Delta W_{ki}}$$

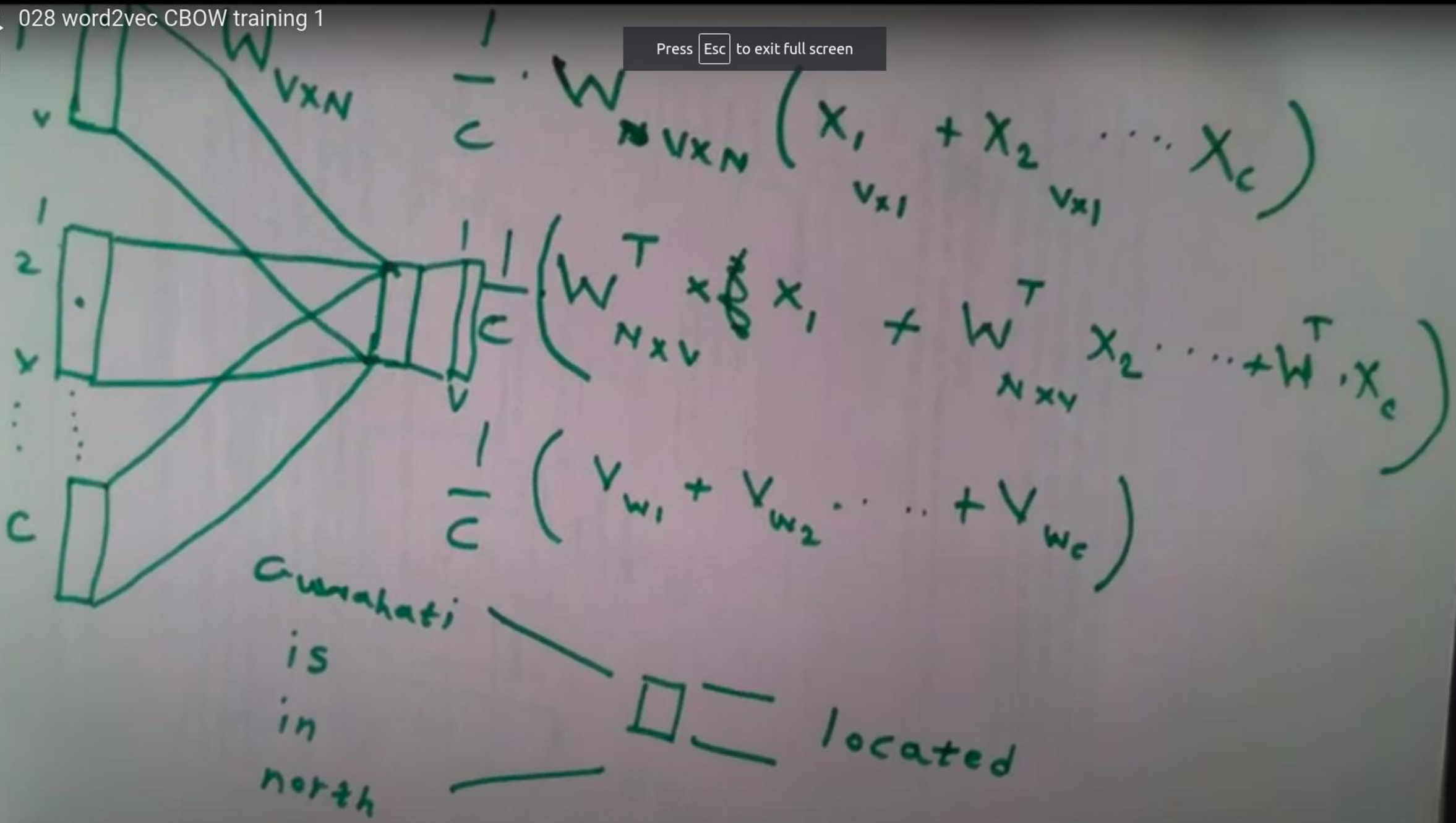
$$-\eta \frac{\partial E}{\partial W_{ki}} \qquad -\eta \cdot x_k \cdot EH_i$$

$$\frac{\partial E}{\partial W_{ki}} = \boxed{\frac{\partial E}{\partial h_i}} \boxed{\frac{\partial h_i}{\partial W_{ki}}} \; x_k$$

$$\downarrow$$

$$\sum_{j=1}^{V} \frac{\partial E}{\partial u_j} \cdot \boxed{\frac{\partial u_j}{\partial h_i}} \; w'_{ij}$$

$$\boxed{\sum_{j=1}^{V} e_j \cdot w'_{ij}}$$



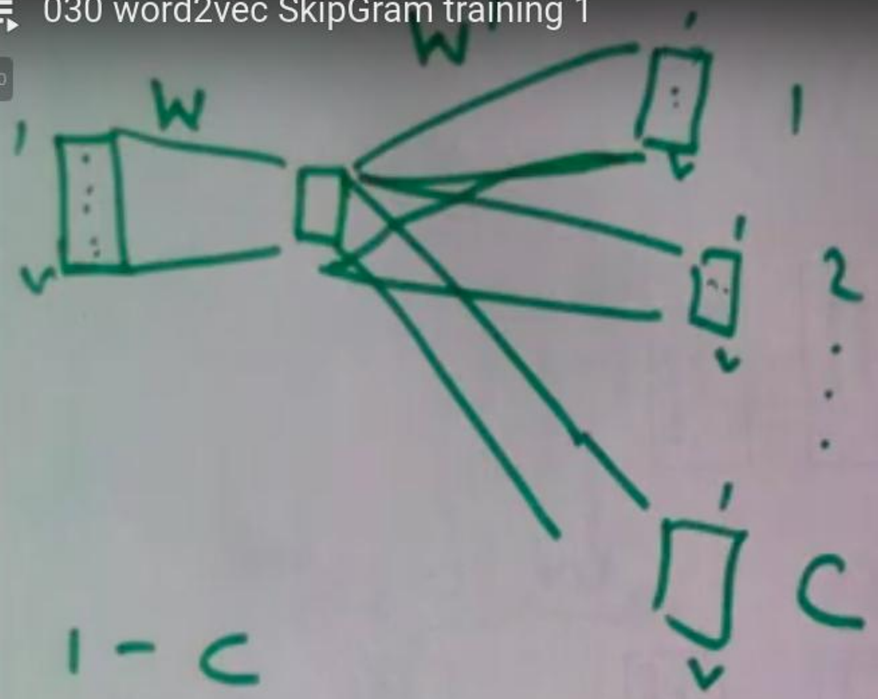
$$E = -\log\left(p(w_{j*} \mid w_1, w_2 \cdots w_c)\right)$$

$$= -u_{j*} + \log \sum_{j'=1}^{V} \exp(u_{j'})$$

$$w_{ki} = w_{ki} + \Delta w_{ki}$$

$$-\eta \cdot \frac{\partial E}{\partial w_{ki}}$$

$$-\eta \cdot x_k \cdot EH_i \cdot \frac{1}{c}$$



$1 - C$

$W_{V \times N} \cdot X_{V \times 1}$

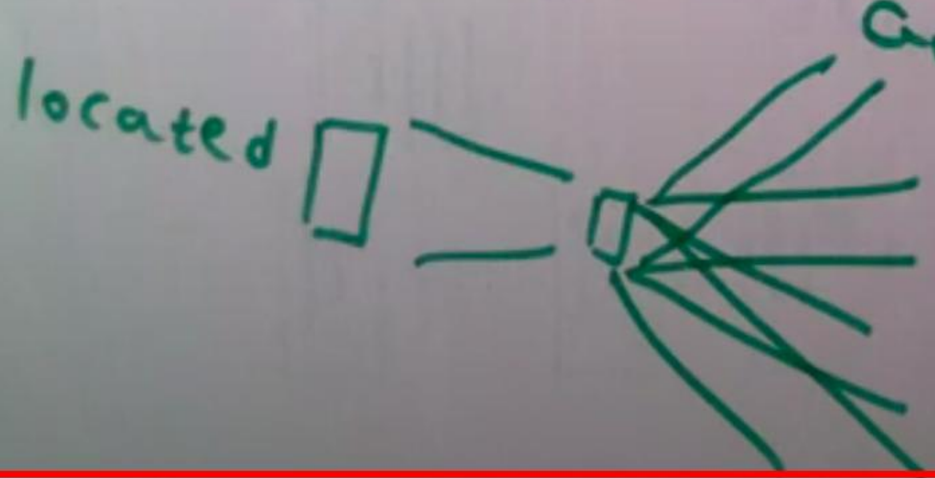
$W^T_{N \times V} \cdot X_{V \times 1} : h$

$N \times 1$

v_{w_i}

u_j

$$P(w_{c,j} | w_i) = y_{c,j} = \frac{\exp(u_{c,j})}{\sum_{j'=1}^{V} \exp(u_{j'})}$$



$u_{cj} = u_j = h \cdot v'_j$

$$E = -\log \prod_{c=1}^{C} p(w_{oc} \mid w_i) \rightarrow$$

$$\sum_{c=1}^{C} \log p(w_{oc} \mid w_i)$$

$$\frac{\exp(u_{cj_c^*})}{\sum_{j'=1}^{V} \exp(u_{j'})}$$

$$= -\sum_{c=1}^{C} u_{j_c^*} + C \cdot \log \sum_{j'=1}^{V} \exp(u_{j'})$$

$$\Delta w'_{ij} = -\eta \cdot \frac{\partial E}{\partial w'_{ij}} = -\eta \cdot \sum_{c=1}^{C} \frac{\partial E}{\partial u}$$

$$\frac{\partial E}{\partial u_{jc}} = y_{cj} - t_{cj} = e_{cj}$$

$$EI_j = \sum_{c=1}^{C} e_{cj}$$

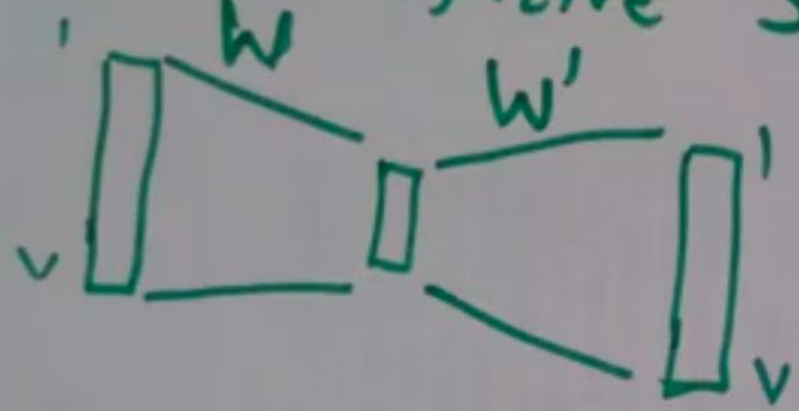
$c = 1$
Guwahati
j = Guwahati
$t_j = 1$
else 0

$c = 2$
is
j = is
$t_j = 1$
else 0

$$\Delta w'_{ij} = -\eta \cdot \left(\sum_{c=1}^{C} e_{cj} \right) \cdot h_i$$

$$-\eta \cdot h_i \cdot EI_j$$

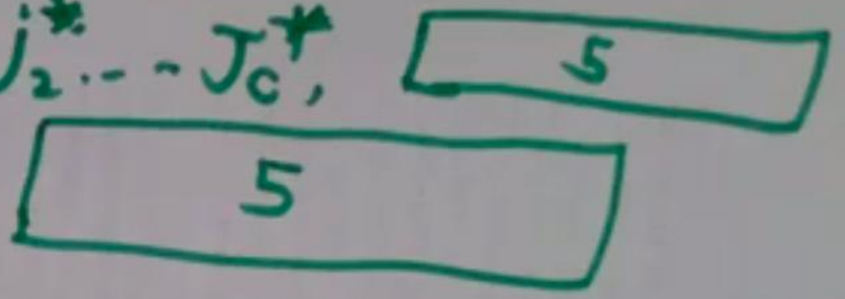
Negative Sampling



$J_1^*, J_2^*, \ldots, J_C^*$

j^*

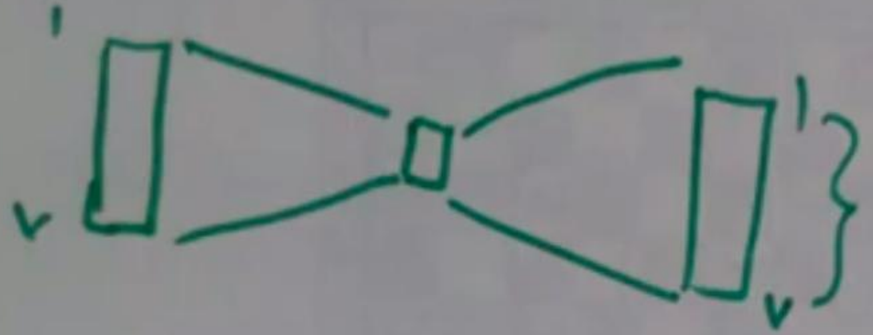
10^6

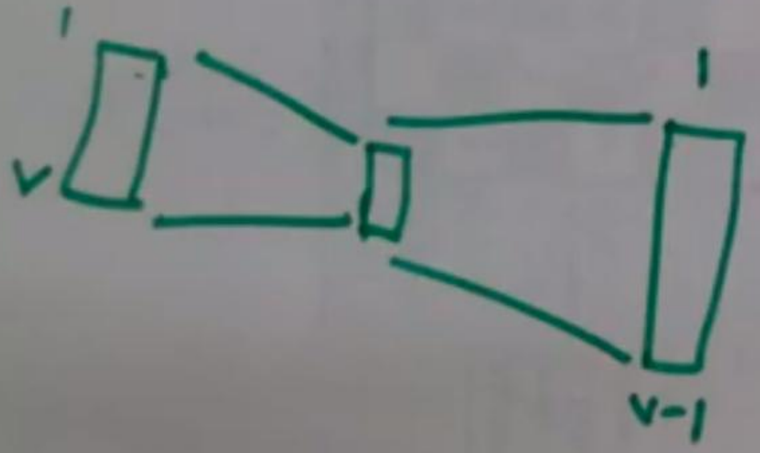


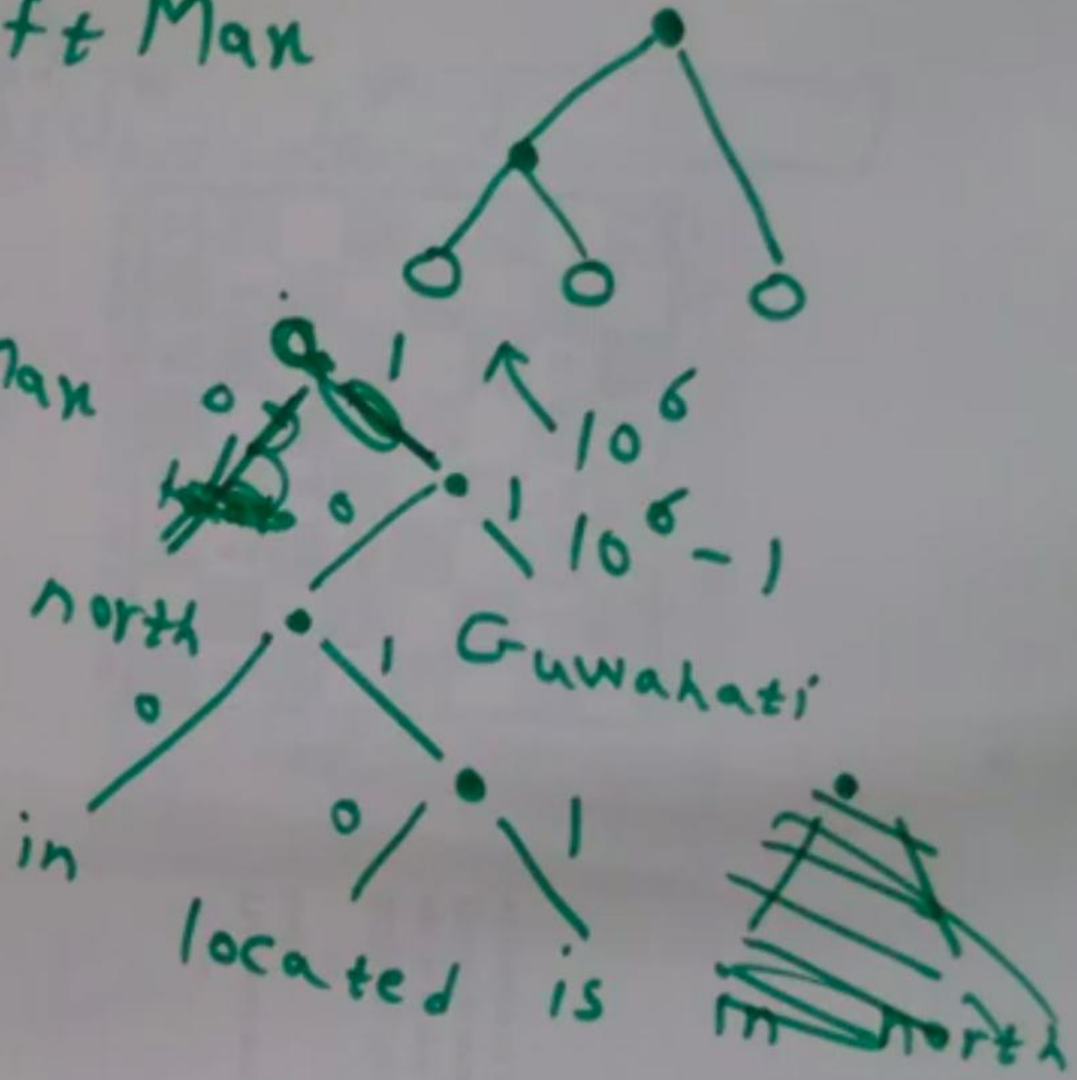
$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=1}^{v} f(w_j)^{3/4}}$$

$\frac{6}{10^6}$

Hierarchical Soft Max



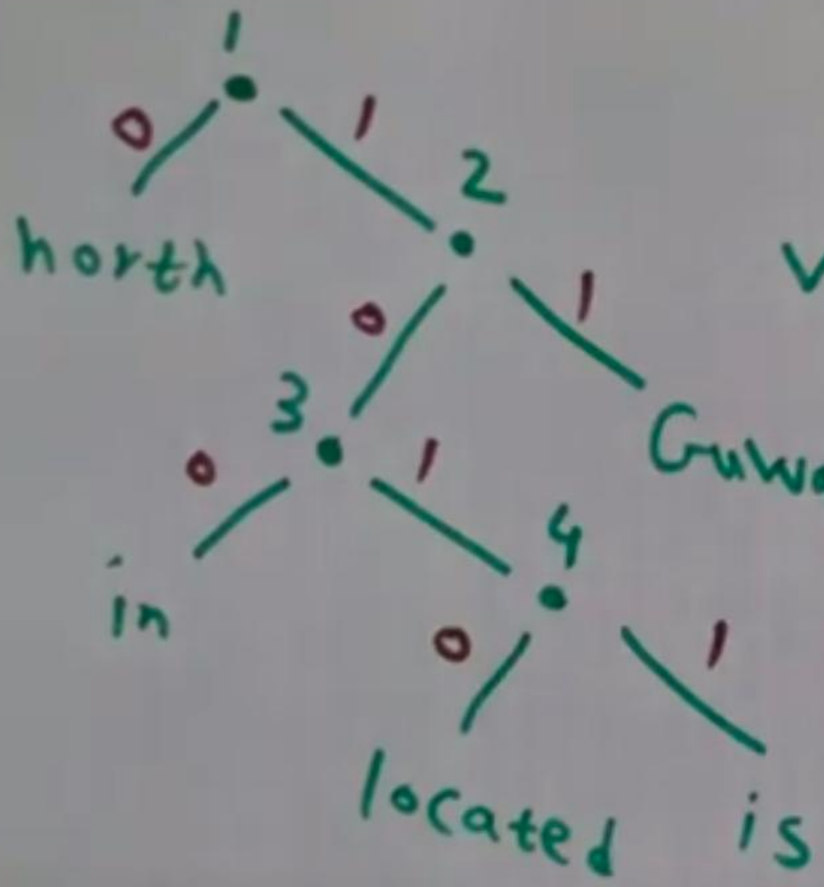


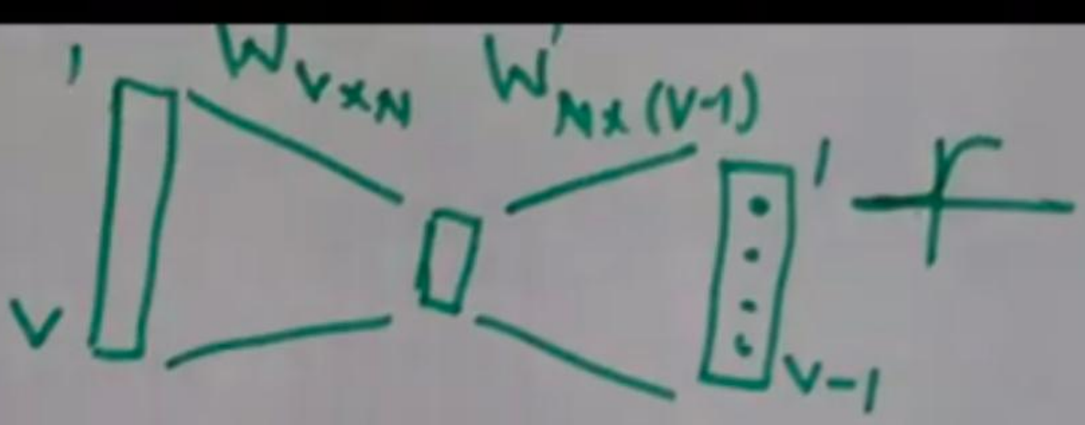


north: 0 is: 1011

in: 100

Guwahati: 11





north: 0

in : 100

	1	2	3	4
north	0			
in	1	0	0	
located	1	0	1	0
Guwahati	1	1		
is	1	0	1	1

Phrases

→ Times of India

Indian Express

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i\ w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}$$

Evaluation of Word Vectors

Objective

- Training speed
- Memory requirement
- Large training data
- Downstream task

Subjective

- Word analogies

vector(King)
− vector(Male)
+ vector(Female)

vector(X)

bigger
− big
+ small

smaller

Country : Capital
India : New Delhi
USA : ?

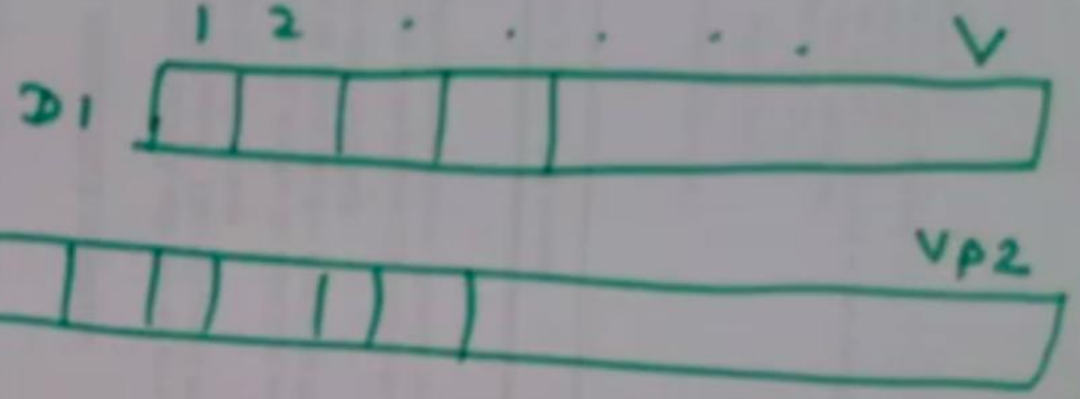
Bigger : Big :: Smaller : Small
King : Male :: Queen : Female

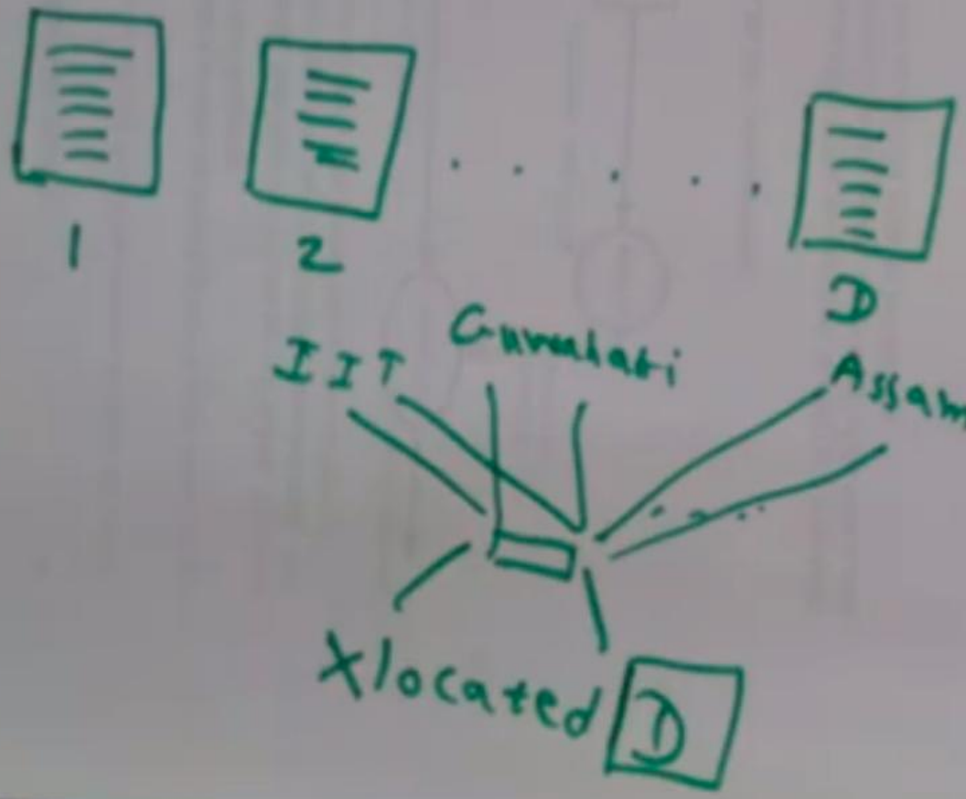
Documents

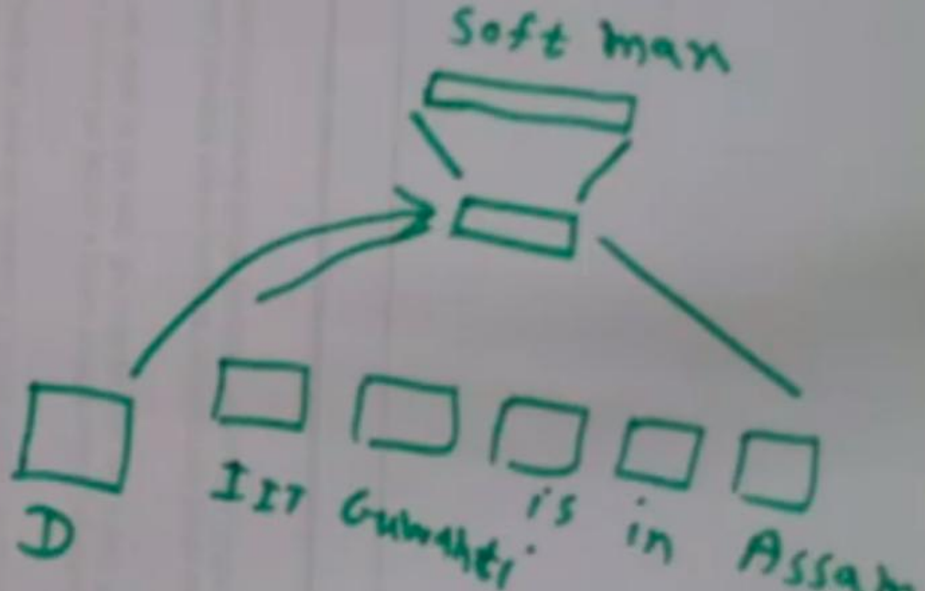
Bag of words : tf idf

Bag of n-grams:

doc2vec



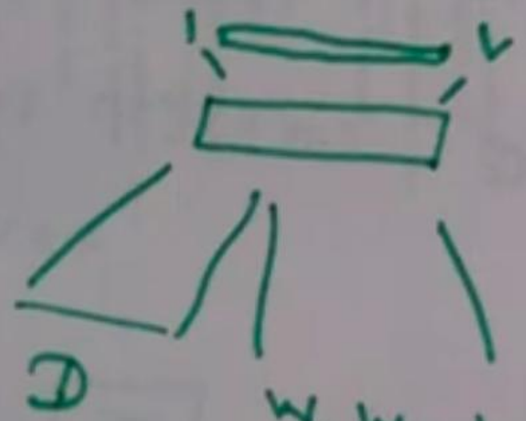




PV-DM: Paragraph Vector Distributed Memory

N: Paragraphs / Documents : p
V: Vocabulary size : q

$(N \times p) + (V \times q)$



	D_1	D_2	D_3
1	0.7	0.5	0.1
2	0.9	0.6	0.2

≈ 2

$0.7 \rightarrow 0.7001$
$0.9 \rightarrow 0.89999$

$W_1, W_2 \ldots W_c$

Finance

bank
money
lending
interest
:
player
is
cost

Sports

bat
ball
money
:
:
is
this

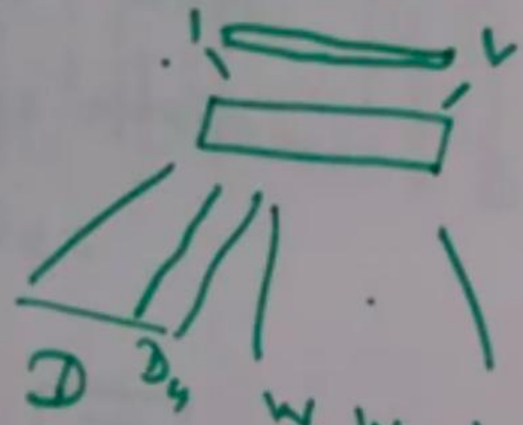
Computer science

algorithm
cost
:
:
:
is
this

Paragraph Vector Distributed Memory

N: Paragraphs / Documents : p
V: Vocabulary size : q

$(N \times p) + (V \times q)$



	D_1	D_2	D_3	D_4
1	0.7	0.5	0.1	0.6
2	0.9	0.6	0.2	0.2

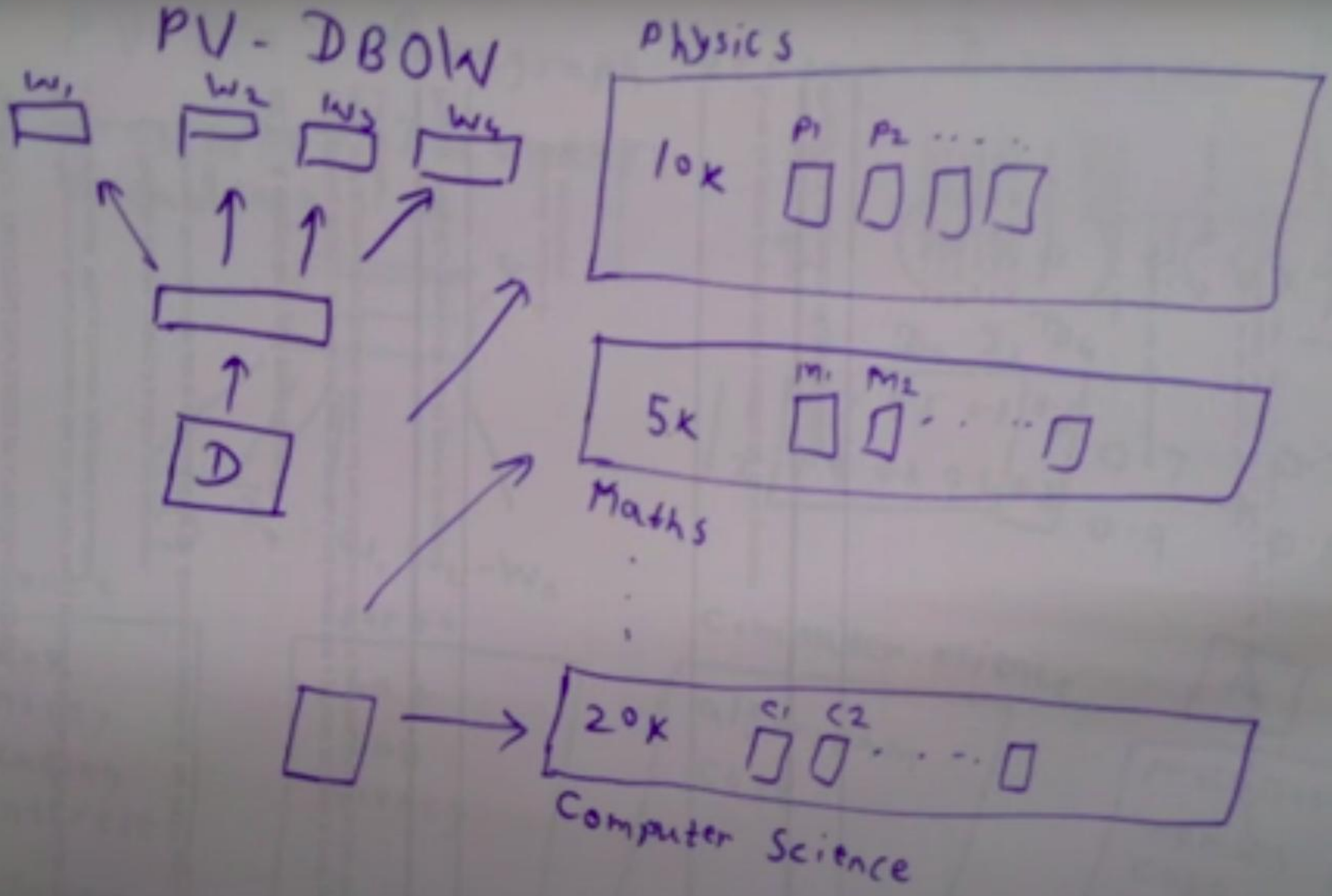
0.7 → 0.7001
0.9 → 0.89999

Finance
bank
money
lending
interest
:
player
is
cost

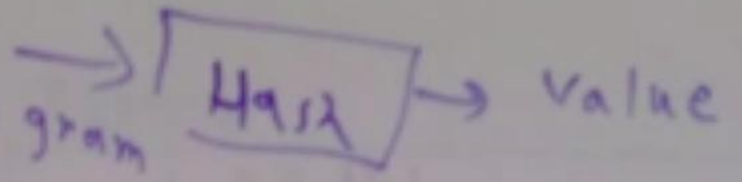
Sports
bat
ball
money
:
is
this

computer science
algorithm
cost
:
is
this

D_4
processor
memory
cost
network
:
is
this


PV-DBOW
W1
W2
W3
W4
D
Physics
10k
P1
P2
5k
M1
M2
Maths
20k
C1
C2
Computer Science

Fast Text

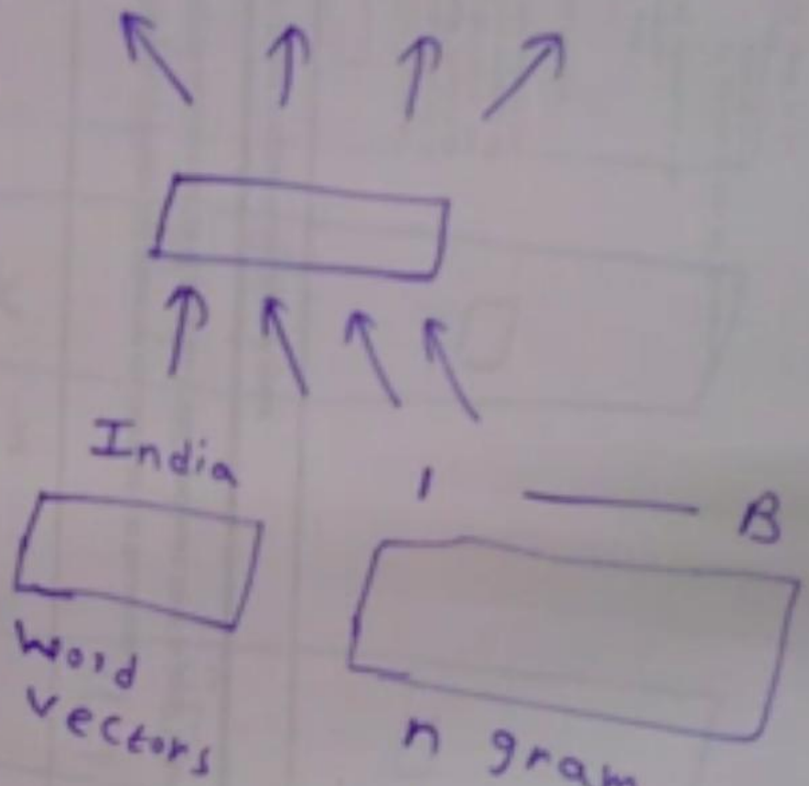


1 - B

India Indian

3 gram

1	<In	<In
2	Ind	Ind
3	ndi	ndi
4	dia	dia
5	ia>	ian
		an>



India → Asia Population Software Education