CBOW Model:

There is a lot of advancement in AI recently

V : Vocab $\quad |V|$: size of vocab. $\quad |V| = 10{,}000$.

word	position
There :	57
is :	1058
a :	352

57 1058 352 . . . . . .

$$\text{OHE}_{\text{there}} = [\overset{0}{0}, 0, 0, \ldots, \overset{56}{1}, 0, 0, \ldots, 0]$$

$$\vec{a}_{\text{OHE}} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}_{|V|}$$

Context Window: $k = 2$

There is a lot of $\boxed{\text{advancement}}$ in AI recently.

target

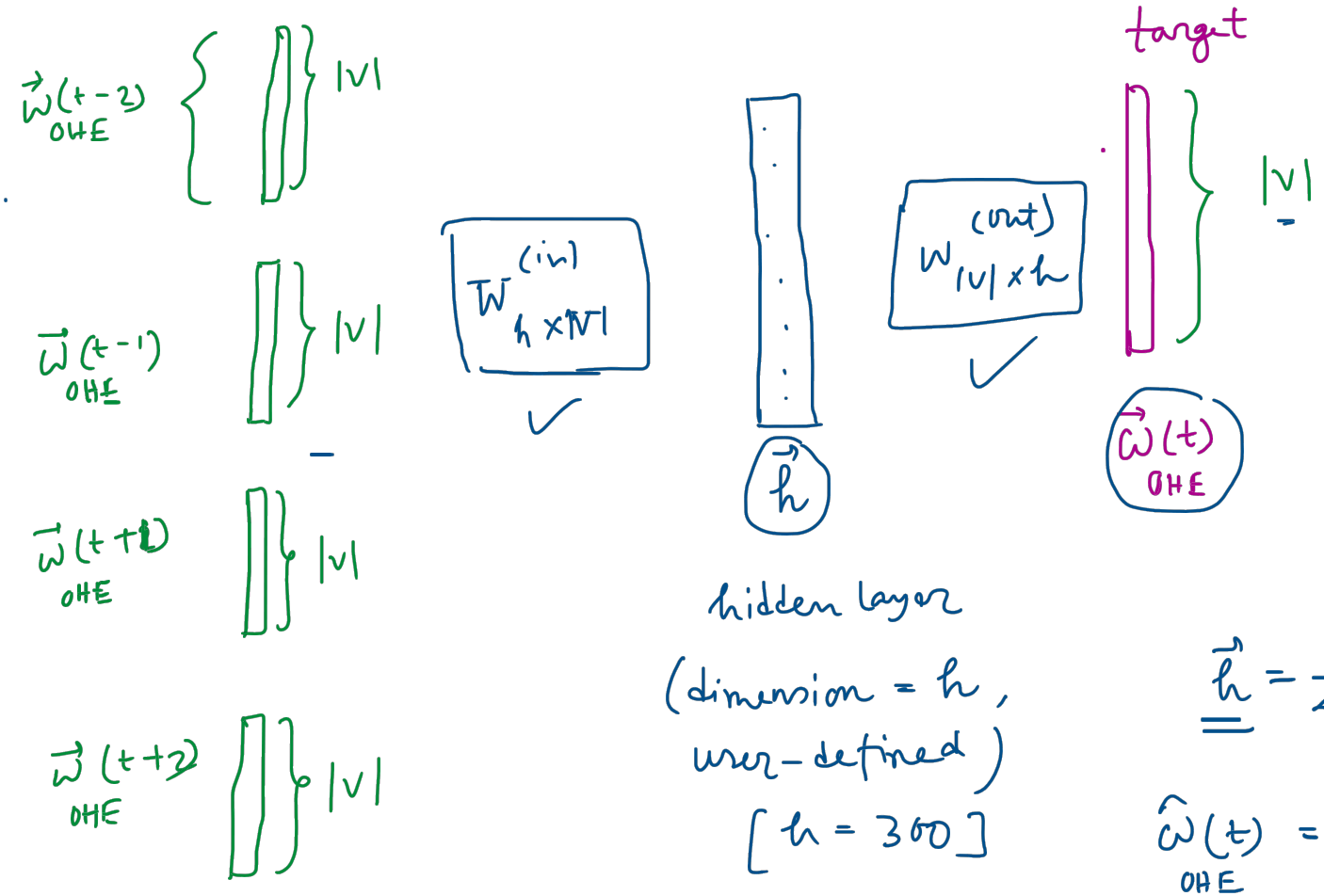
$w(t-2) \quad w(t-1) \quad w(t) \quad w(t+1) \quad w(t+2)$

$$|\vec{w}_{\text{OHE}}(t)| = |V|$$

x	y
there, is, lot, of	a
is, a, of, adv	lot
a, lot, adv, in	of
⋮	⋮

Shallow NN (1 hidden layer)

$K = 2$



hidden layer (dimension = h, user-defined) [$h = 300$]

$$\vec{h}_{t-2} = f\left(W_{h \times |V|}^{(in)} \cdot \vec{w}_{OHE}^{(t-2)}\right)$$

$$\vec{h}_{t-1} = f\left(W_{h \times |V|}^{(in)} \cdot \vec{w}_{OHE}^{(t-1)}\right)$$

$$\vdots$$

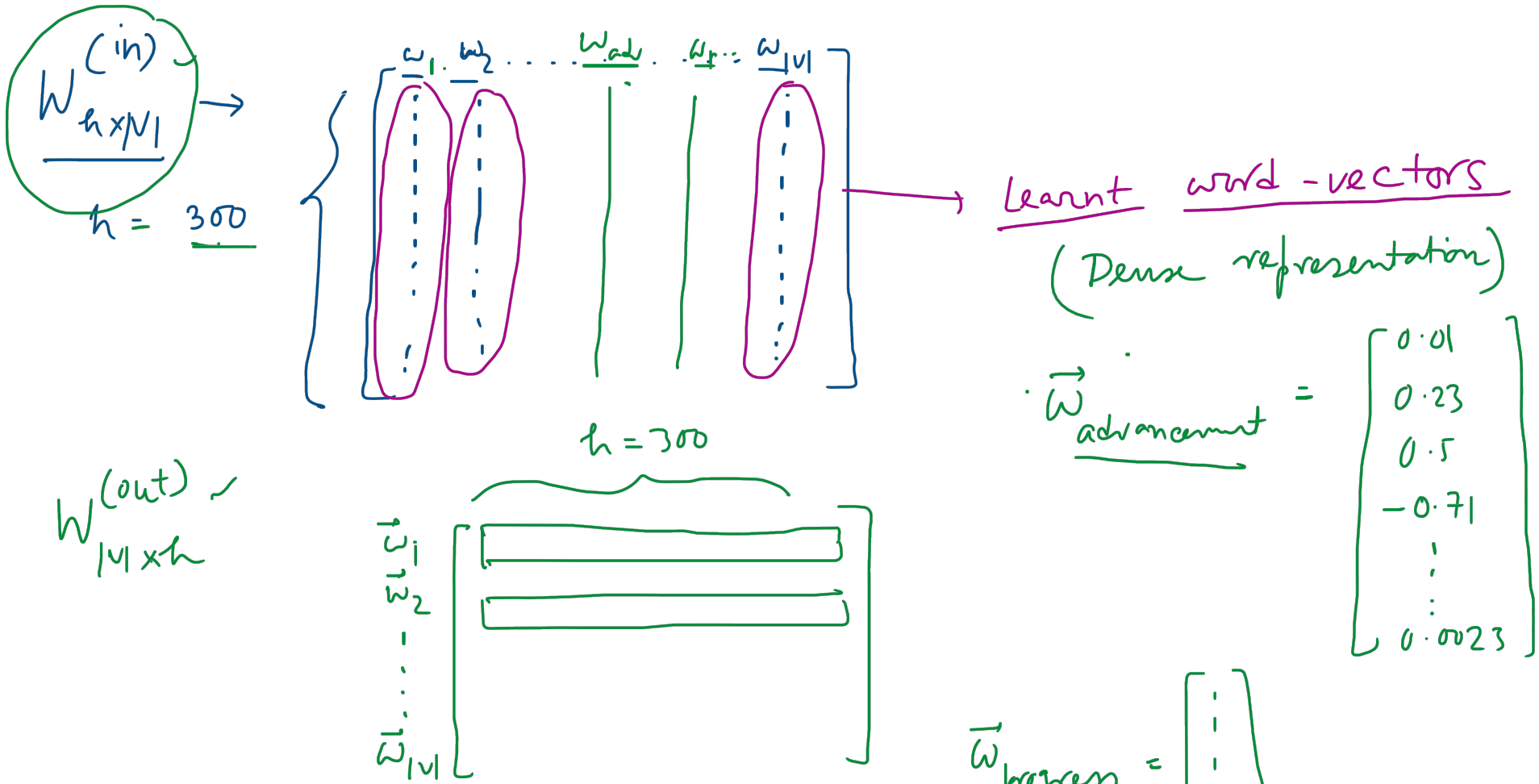
$$\vec{h}_{t+1} = f\left(W_{h \times |V|}^{(in)} \cdot \vec{w}_{OHE}^{(t+1)}\right)$$

$$\vec{h}_{t+2} = f\left(W_{h \times |V|}^{(in)} \cdot \vec{w}_{OHE}^{(t+2)}\right)$$

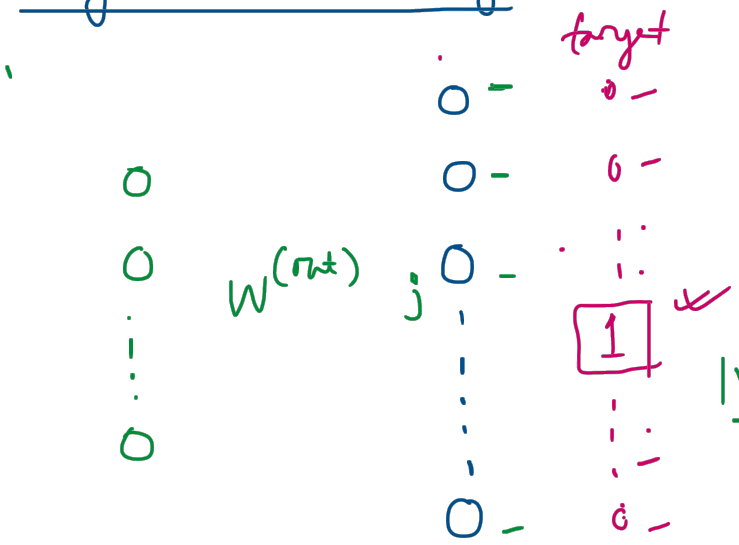
$$\vec{h} = \frac{1}{4}\left(\vec{h}_{t-2} + \vec{h}_{t-1} + \vec{h}_{t+1} + \vec{h}_{t+2}\right)$$

$$\hat{w}_{OHE}^{(t)} = \text{Softmax}\left(W_{|V| \times h}^{(out)} \cdot \vec{h}\right)$$

$$\mathcal{L}\left(\vec{w}^{(t)}, \hat{w}^{(t)}\right)$$


$W^{(in)}_{h \times |V|}$
$h = 300$
w_1 w_2 $w_{aardvark}$ w_{zebra} $w_{|V|}$
Learnt word-vectors
(Dense representation)
$\vec{w}_{advancement}$ = [0.01, 0.23, 0.5, −0.71, ⋮, 0.0023]
$W^{(out)}_{|V| \times h}$
$h = 300$
$\vec{w}_1$ $\vec{w}_2$ ⋮ $\vec{w}_{|V|}$
$\vec{w}_{program}$ = [⋮]
$\cos(\vec{w}_{advancement}, \vec{w}_{program}) \simeq 1$

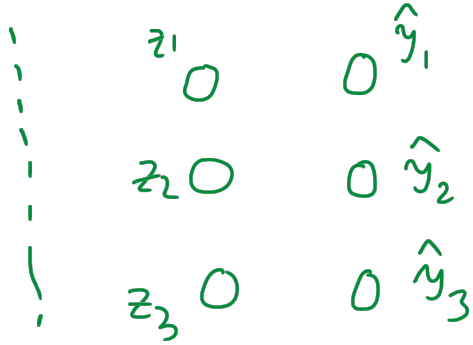
Negative Sampling :-



logits = $W^{(out)} \cdot h$

output = softmax(Logits)

$|V|$ many exponential to calculate the softmax output.



$\hat{y} = \text{softmax}(z)$

$$\hat{y}_j = \frac{e^{z_j}}{\sum_{i=1}^{n} e^{z_i}}$$

$\rightarrow$ n-exp

hidden layer (h)

output layer $|V|$

$$\hat{y}_j = \frac{e^{z_j}}{\sum_{i=1}^{|V|} e^{z_i}}$$

number of negative samples

large corpus $\rightarrow$ 2-3 words.
mid " $\rightarrow$ 3-5 "
small " $\rightarrow$ 5-10 "

$|V| = 10000$ (say)

I select the positive target and some negative words (negative sampling)

	$u(w)$
w_1 :	1000
w_2 :	780
:	:
:	:
:	:

(Unigram) distribution

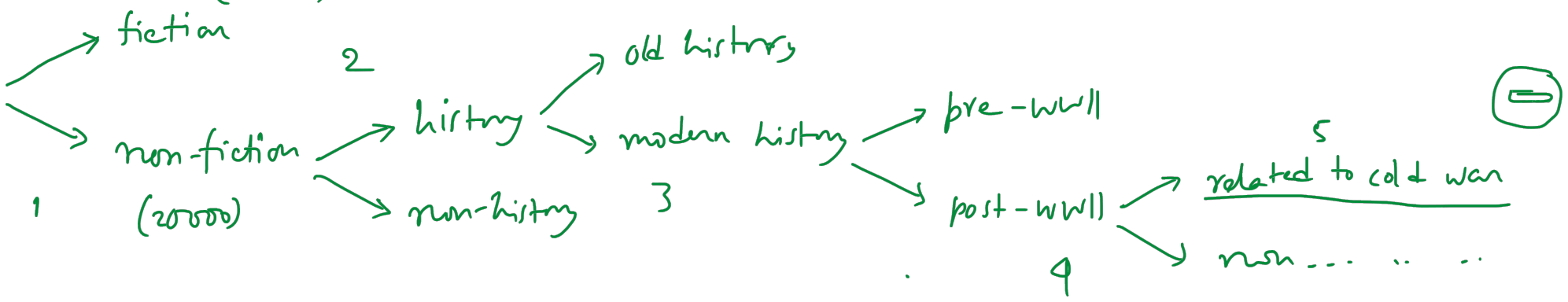
$u(w)^{3/4}$

	$u(w)$	$[u(w)]^{1/2}$	$[u(w)]^{\alpha}$
w_1 :	10000	100	
w_2 :	900	30	

$\left(\frac{\log_2 |V|}{w}\right)$

Hierarchical Softmax :- Book in a Library

(10000)



$P(\text{path}) = \prod p(\text{splits})$