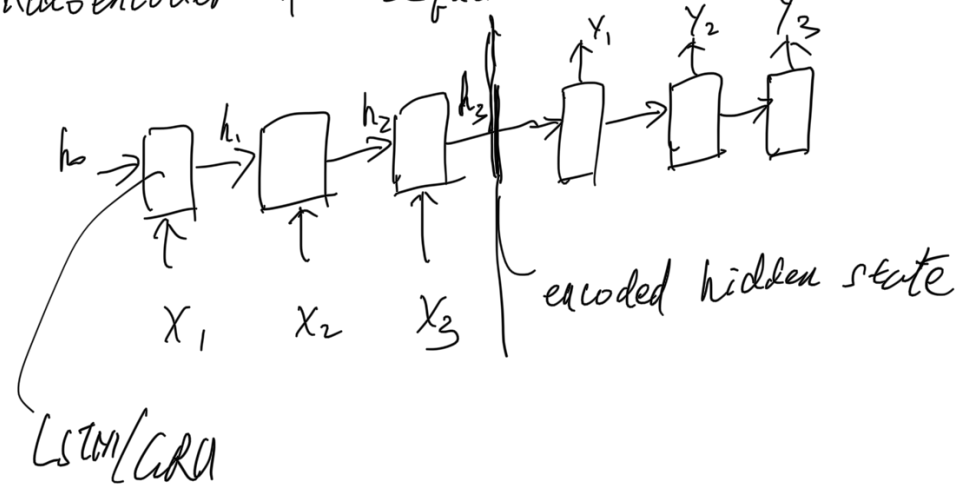
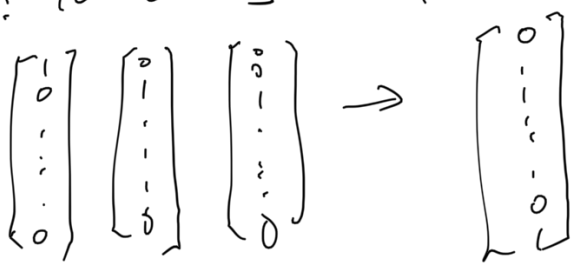
Lecture 10

Autoencoder for Sequential Data



Dict: 10 words ⇒ represent all words with 10-D vectors



Representation can not be very long — not efficient

relationships between words { similarity / distance
analogies.

Embedding

Co-occurrence Matrix with Singular Value Decomposit (SVD)

We learn Deep Learning Use A-lot

$X =$

	we	learn	deep	learning	use	a-lot
we	0	1	1	0	1	1
learn	1	0	1	1	0	1
deep	1	1	0	1	0	0
learning	0	1	1	0	0	0
use	1	0	0	1	0	0
a-lot	1	1	0	0	0	0

we learn deep learning

we use deep learning a lot

learning deep learning

we learn a lot deep learning

$$X = U S V^T$$

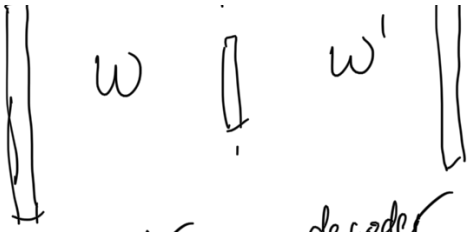
similar to PCA

Word2Vec :

- Continuous bag of words (CBOW)
- skip gram

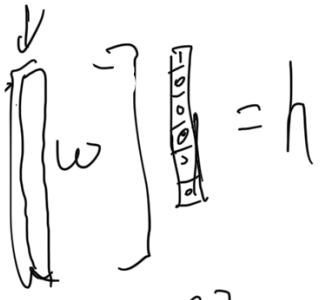
prepare data : corpus, Vocabulary, context length

CBOW :

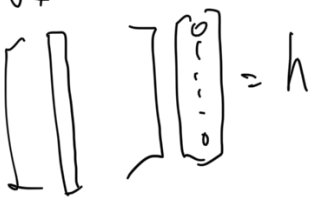


x

$h = Wx$, $y = W'h$

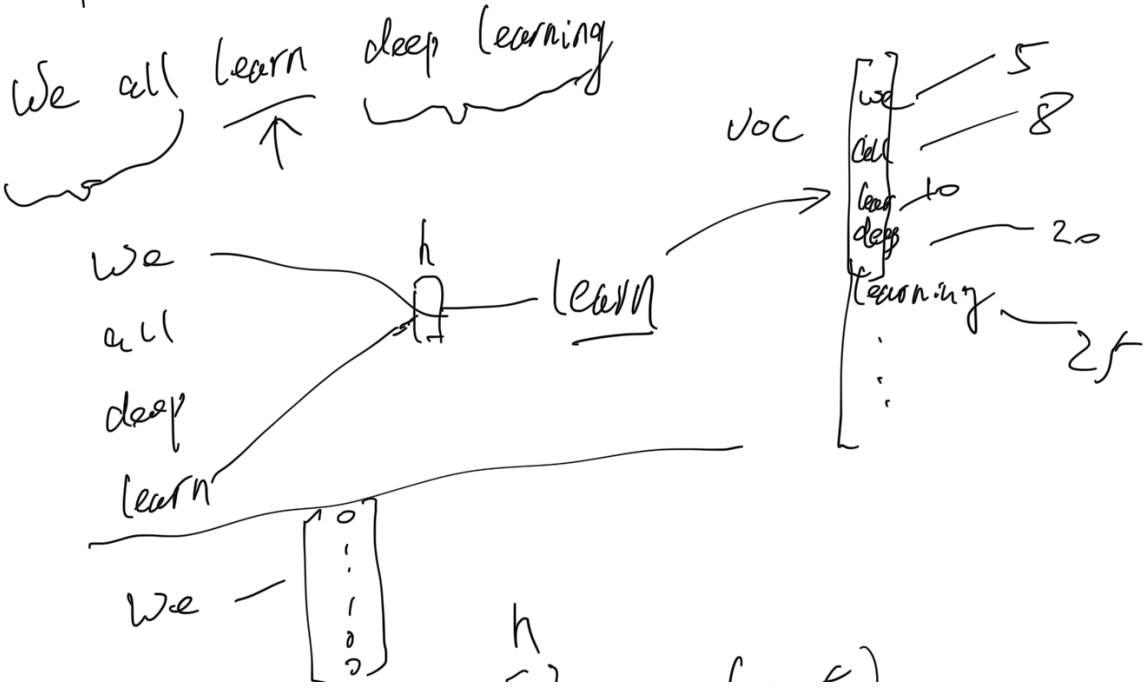


$h = W(:, 1)$



word in Voc, index is i $\Rightarrow x =$  i-element

if word j in Voc is a neighbor of word i, related.



$$[W]\begin{bmatrix}0\\ \vdots\\ 1\\ 0\\ \vdots\end{bmatrix} = [\;] = W(:,5)$$

we

$$W\,[\;] \rightarrow W(:,8)$$

cell

$$W\,[\;] \rightarrow W(:,20)$$

deep

$$W\,[\;] \rightarrow W(:,25)$$

learning

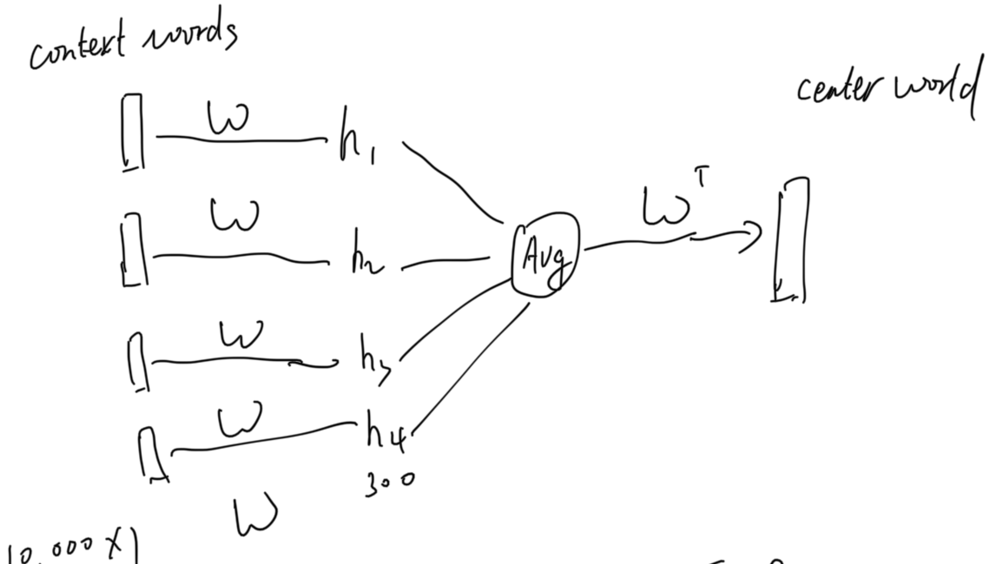
$$\Downarrow$$

$$h$$

$$W^T h = y = \begin{bmatrix}0\\0\\ \vdots\\ 1\\0\\0\\0\end{bmatrix}$$ 10-element

softmax

$$\min(\hat{y} - y)^2$$



$$y = W^T h$$

$h = W X$

300×1 $\quad 10{,}000 \times 1$ $\quad 300 \times 10{,}000$

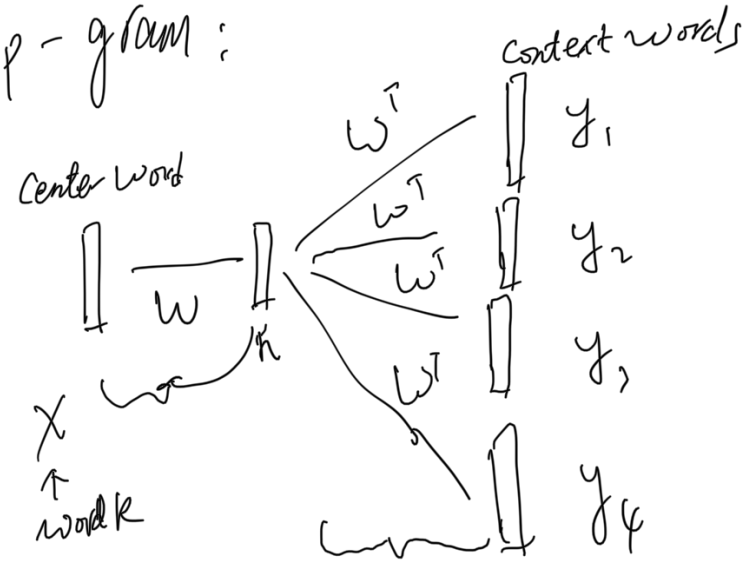
$10{,}000 \times 300$ $\quad 300 \times 1$ $\quad 10{,}000 \times 1$

After training,

for each word

$h = W X$

skip-gram:

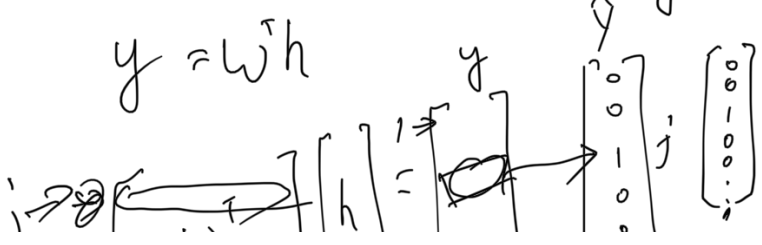


$h = W X$

$\begin{bmatrix} | \\ W \\ | \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = W(:, k)$. X is Voc word k

$\hat{y}$ — ground truth is one-hot

$y = W' h$



W [ ][ ] [ ] [:]

$W^T(j,:)\, h = y(j)$

$T = 10,000$

maximize Softmax

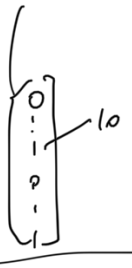
$$\frac{e^{W^T(j,:)\,h}}{\sum_{i=1}^{I} e^{W^T(i,:)\,h}}$$

$$\sum_{j}^{4} \frac{e^{W^T(j,:)\,h}}{\sum_{i=1}^{I} e^{W^T(i,:)\,h}}$$

, $j \in$ words nearby

voc word k

we learn deep learning



voc