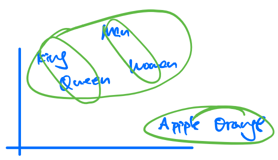
Word Embeddings

- 1-hot representation
- word embedding

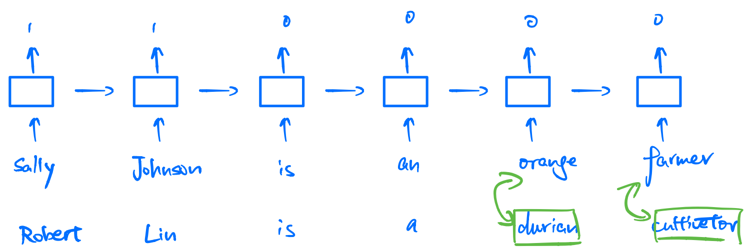
e.g.	Man (5391)	Woman (9853)	King (4914)	Queen (7157)
Gender	-1	1	-0.95	0.97
Royal	0.01	0.02	0.93	0.95
Age	0.03	0.02	0.7	0.69
⋮				
	e_{5391}	e_{9853}		

300

300D → 2D Visualization



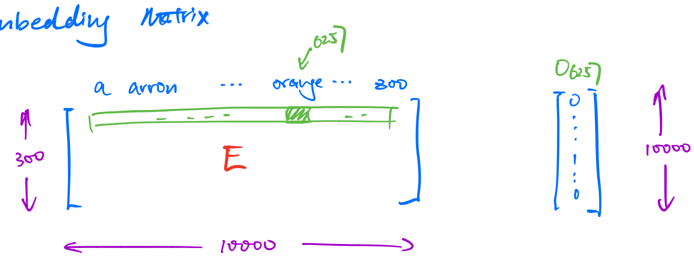
NER example



⇒ transfer learning

1. learn word embeddings from large text corpus (1-100B words)
   (or download pre-trained embedding)
2. transfer embedding to new task w/ smaller training set
   (say 100 k words)
3. Optional. Fine-tune word embeddings w/ new data

Embedding Matrix



$$\underset{(300,10k)}{E} \cdot \underset{(10k,1)}{O_{6257}} = \underset{(300,1)}{\begin{bmatrix} \cdot \end{bmatrix}} = e_{6257}$$

Learning word embeddings

e.g.

$\overbrace{\text{a glass of orange}}^{4}$

I	want	a	glass	of	orange	___ .
4343	9665	1	3852	6163	6257	

I $\quad O_{4343} \rightarrow E \rightarrow e_{4343}$

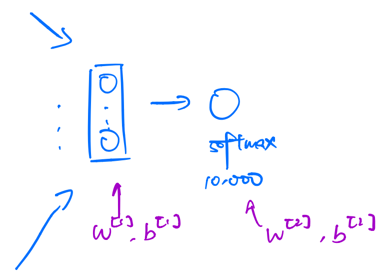
want $\quad e_{9665}$

a $\quad e_1$

glass $\quad e_{3852}$

of $\quad e_{6163}$

orange $\quad e_{6257}$



$300 \times 6 = \cancel{1800} \rightarrow 1200$ window

backprop to learn E

— Context : last 4 words

4 words on left & right

last 1 word

nearby 1 word

Word 2 Vec

Skip - Gram

e.g. I want a glass of orange juice to go along my cereal.

Context	target
orange	juice
orange	glass
orange	my

← sampled from nearby window

Model :

Vocab size = 10,000 K

Context c ("orange") → Target t ("juice")
6257 4834

$O_c \rightarrow E \rightarrow e_c \rightarrow \text{softmax} \rightarrow \hat{y}$ ← one-hot

$e_c = E O_c$

softmax : $P(t|c) = \dfrac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10000} e^{\theta_j^T e_c}}$ θ_t = param associated w/ output t

$$\mathcal{L}(\hat{y}, y) = - \sum_{i=1}^{10000} y_i \log \hat{y}_i$$

Problem :

Calc denominator of softmax is slow

Solution :

hierachical softmax

negative sampling

Negative Sampling

e.g. I want a glass of orange juice to go along my cereal.

Context (x)	word (x)	target ? (y)
orange	juice	1
orange	king	0
orange	book	0
orange	the	0

$\uparrow k \downarrow$

$k = 5 - 20$ smaller dataset

$k = 2-5$ larger dataset

Softmax : $$p(t \mid c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10000} e^{\theta_j^T e_c}}$$ } 10000-way softmax

$$\Rightarrow \quad P(y=1 \mid c, t) = \sigma(\theta_t^T e_c)$$

much faster

$$O_{6257} \rightarrow E \rightarrow e_{6257} \rightarrow \begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \end{matrix}$$

juice?

10,000 logistic regression

Select negative sampling

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=1}^{10000} f(w_j)^{3/4}}$$ ← between uniform distribution & word distribution

GloVe word vectors

e.g. I want a glass of orange juice to go along my cereal.

c, t

x_{ij} = # times j (t) appears in context of i (c)

$x_{ij} = x_{ji}$

symmetric $\Rightarrow e_w^{(final)} = \frac{e_w + \theta_w}{2}$

Model :

$$\text{minimize} \sum_{i=1}^{10000} \sum_{j=1}^{10000} f(x_{ij}) (\theta_i^T e_j + b_i + b_j' - \log x_{ij})^2$$

weighting term $f(x_{ij}) = 0$ if $x_{ij} = 0$