deeplearning.ai

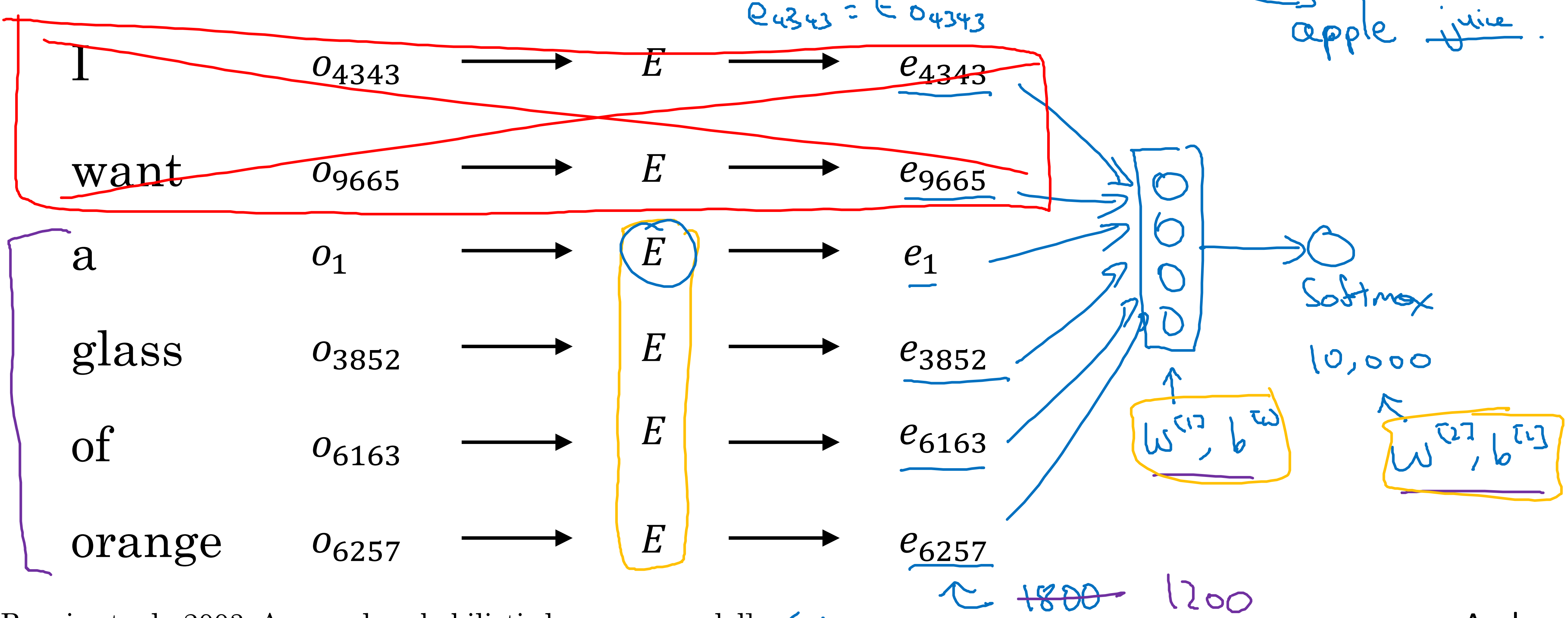
NLP and Word Embeddings

Learning word embeddings

Neural language model

4

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



$e_{4343} = E\, o_{4343}$

I o_{4343} → E → e_{4343}

want o_{9665} → E → e_{9665}

a o_{1} → E → e_{1}

glass o_{3852} → E → e_{3852}

of o_{6163} → E → e_{6163}

orange o_{6257} → E → e_{6257}

[Bengio et. al., 2003, A neural probabilistic language model] ←

Andrew Ng

Other context/target pairs

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

context

target

Context: Last 4 words.

- 4 words on left & right
- Last 1 word
- Nearby 1 word

a glass of orange ? to go alg with

orange ?

glass ?

skip gram

deeplearning.ai

NLP and Word Embeddings

Word2Vec

Skip-grams

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

Context	Target
orange	juice
orange	glass
orange	my

[Mikolov et. al., 2013. Efficient estimation of word representations in vector space.]

Model

Vocab size = 10,000k

$$x \xrightarrow{\qquad\qquad} \underline{y}$$

Context c ("orange") 6257 $\longrightarrow$ Target $\underline{t}$ ("juice") $\underline{\underline{4834}}$

$$o_c \to \boxed{E} \to e_c \to \underset{\text{Softmax}}{O} \to \hat{y}$$

$e_c = E o_c$

Softmax: $$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$$

θ_t = parameter associated with output t

$$\to \mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{10,000} y_i \log \hat{y}_i \leftarrow$$

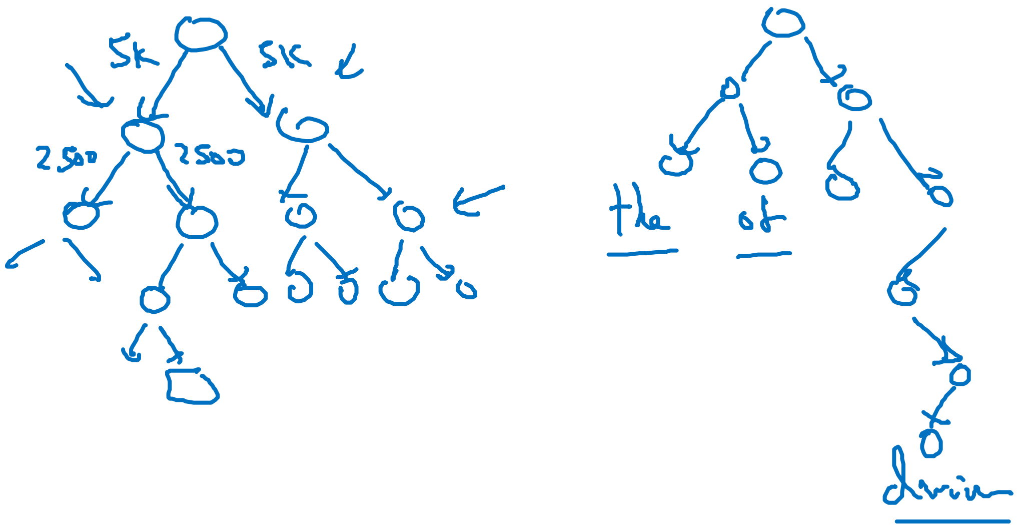
$$y = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow 4834$$

Problems with softmax classification

$$\underline{p(t|c)} = \frac{e^{\theta_t^T e_c}}{\boxed{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}}$$

Hierachical softmax.

$\log |V|$



How to sample the context c?

→ the, of, a, and, to, ...

→ orange, apple, durian

e_{durian}

t

$c \rightarrow t$

$P(c)$

Andrew Ng

deeplearning.ai

NLP and Word Embeddings

Negative sampling

Defining a new learning problem

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

x → y

Context	word	target?
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

k

k = 5-20 smaller. datasets

k = 2-5 larger dataset

[Mikolov et. al., 2013. Distributed representation of words and phrases and their compositionality]

Andrew Ng

Model

Softmax: $p(t|c) = \frac{e^{\theta_t^T e_c}}{\underline{\sum_{j=1}^{10,000}} e^{\theta_j^T e_c}}$

10,000-way softmax

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

x

y

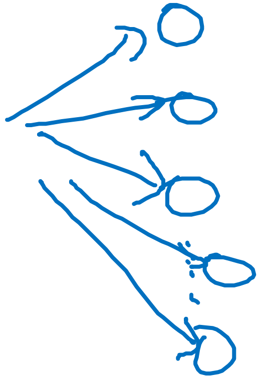
context	word	target?
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

k

c t y

Orange 6257

$o_{6257} \rightarrow E \rightarrow e_{6257}$



juice?

king

10,000

10,000 binary classification problem

$\underline{k+1}$

Andrew Ng

Selecting negative examples

x y

context	word	target?
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

↑ t

the , of, and, . . .

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=1}^{10,000} f(w_j)^{3/4}}$$

$$\frac{1}{|V|}$$

↑

NLP and Word Embeddings

GloVe word vectors

deeplearning.ai

GloVe (global vectors for word representation)

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

c, t

$X_{\underset{\overset{\uparrow}{c}}{i}\underset{\overset{\uparrow}{t}}{j}}$ = # times $\underset{\overset{\uparrow}{t}}{i}$ appears in context of $\underset{\overset{\uparrow}{c}}{j}$.

$\underbrace{X_{ij}}_{} = X_{ji}$ ←

[Pennington et. al., 2014. GloVe: Global vectors for word representation] ←

Model

$$\text{minimize} \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} \underbrace{f(X_{ij})}_{\text{weighting term}} \left(\theta_i^T e_j + \overset{t}{b_i} + \overset{c}{b_j'} - \log X_{ij}\right)^2$$

$\log X_{ij} \to 0$?

$$\text{“}\overset{t}{\theta_t^T}\ \overset{c}{e_c}\text{”}$$

$f(X_{ij}) = 0$ if $X_{ij} = 0$.

“$0 \log 0$” $= 0$

this, is, of, a, ...

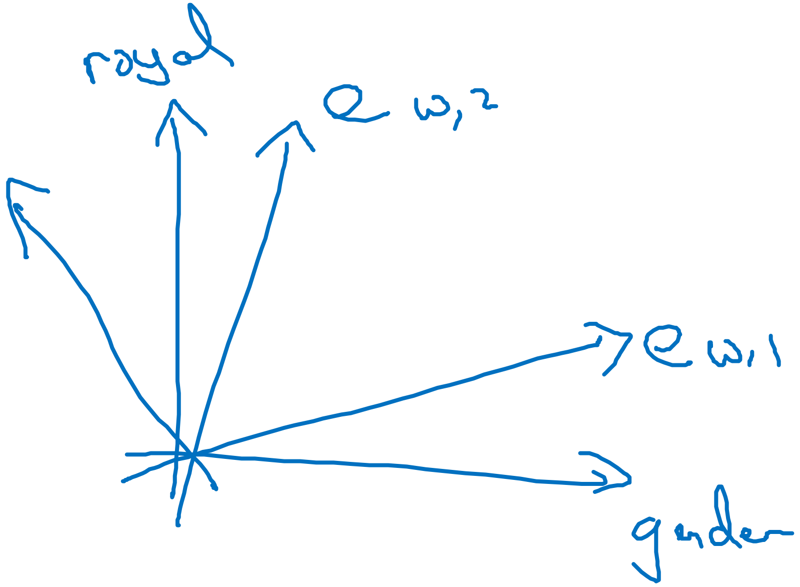
durian

θ_i, e_j are symmetric

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

A note on the featurization view of word embeddings

	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	0.70	0.69
Food	0.09	0.01	0.02	0.01



$$\text{minimize} \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(X_{ij})\left(\theta_i^T e_j + b_i - b_j' - \log X_{ij}\right)^2$$

$$(A\theta_i)^T (A^{-T} e_j) = \theta_i^T \cancel{A^T} \cancel{A^{-T}} e_j$$