Copyright Notice

NLP and Word Embeddings

deeplearning.ai

Word representation

Word representation

V = [a, aaron, …, zulu, <UNK>]

$|V| = 10{,}000$

1-hot representation

$$\begin{array}{cccccc}
\text{Man} & \text{Woman} & \text{King} & \text{Queen} & \text{Apple} & \text{Orange} \\
(5391) & (9853) & (4914) & (7157) & (456) & (6257) \\
\begin{bmatrix}0\\0\\0\\0\\\vdots\\1\\\vdots\\0\\0\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\0\\0\\\vdots\\1\\\vdots\\0\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\\vdots\\1\\\vdots\\0\\0\\0\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\0\\0\\\vdots\\1\\\vdots\\0\end{bmatrix} &
\begin{bmatrix}0\\\vdots\\1\\\vdots\\0\\0\\0\\0\\0\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\0\\0\\\vdots\\1\\\vdots\\0\end{bmatrix} \\
\uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\
O_{5391} & O_{9853} & & & &
\end{array}$$

I want a glass of orange juice______.

I want a glass of apple ? ______.

Featurized representation: word embedding

300

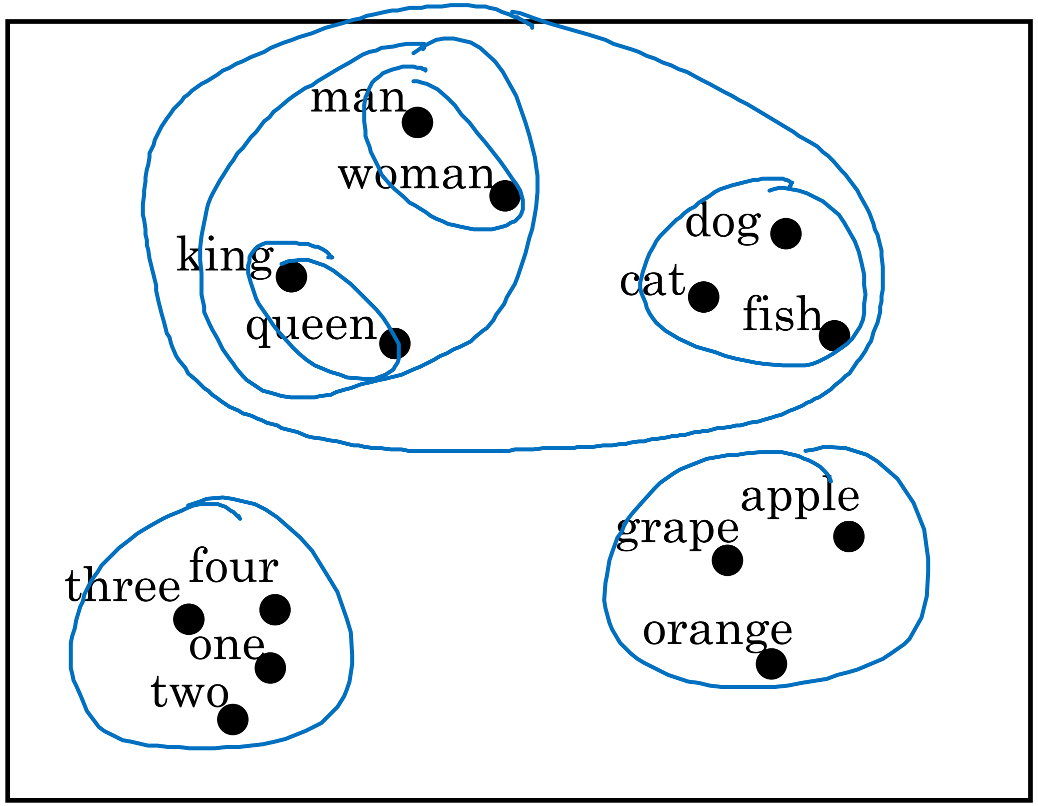
	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97
⋮	⋮	⋮				
size						
cost						
alive						
verb						

e_{5391} e_{9853}

I want a glass of orange juice.

I want a glass of apple juice.

Visualizing word embeddings

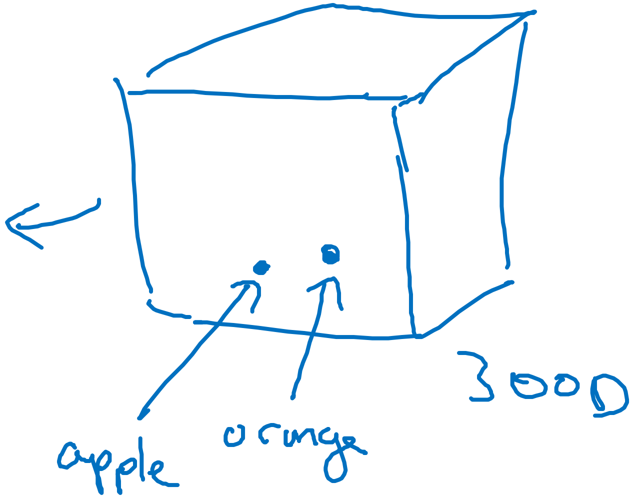


t-SNE

→ 300 D

↓

2D



[van der Maaten and Hinton., 2008. Visualizing data using t-SNE]

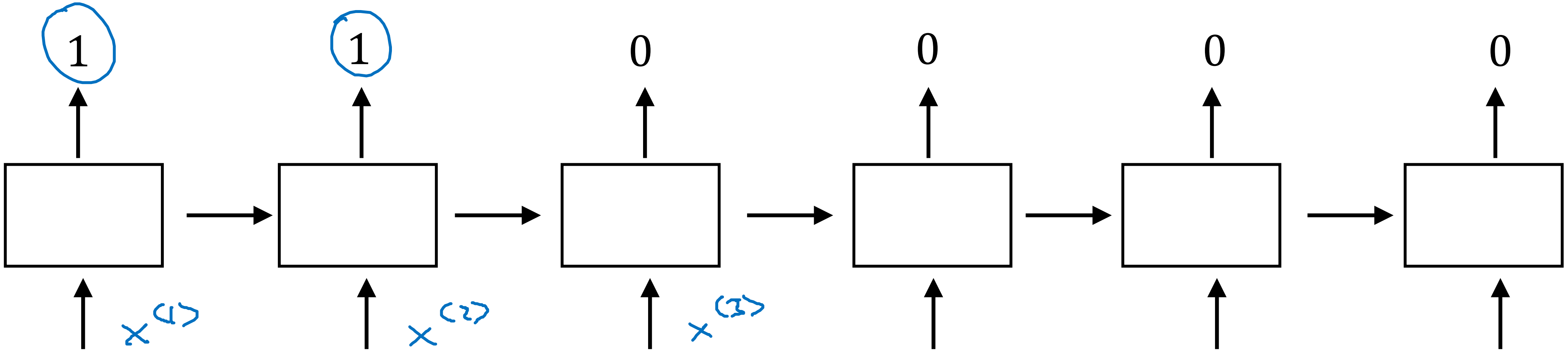
Andrew Ng

deeplearning.ai

NLP and Word Embeddings

Using word embeddings

Named entity recognition example



1 1 0 0 0 0

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$

Sally Johnson is an orange farmer

Robert Lin is ~~an~~ ~~apple farmer~~

→ 1B words – 100B words

a durian cultivator.

→ 100k words

BRNN

Transfer learning and word embeddings

Ⓐ

1. Learn word embeddings from large text corpus. (1-100B words)

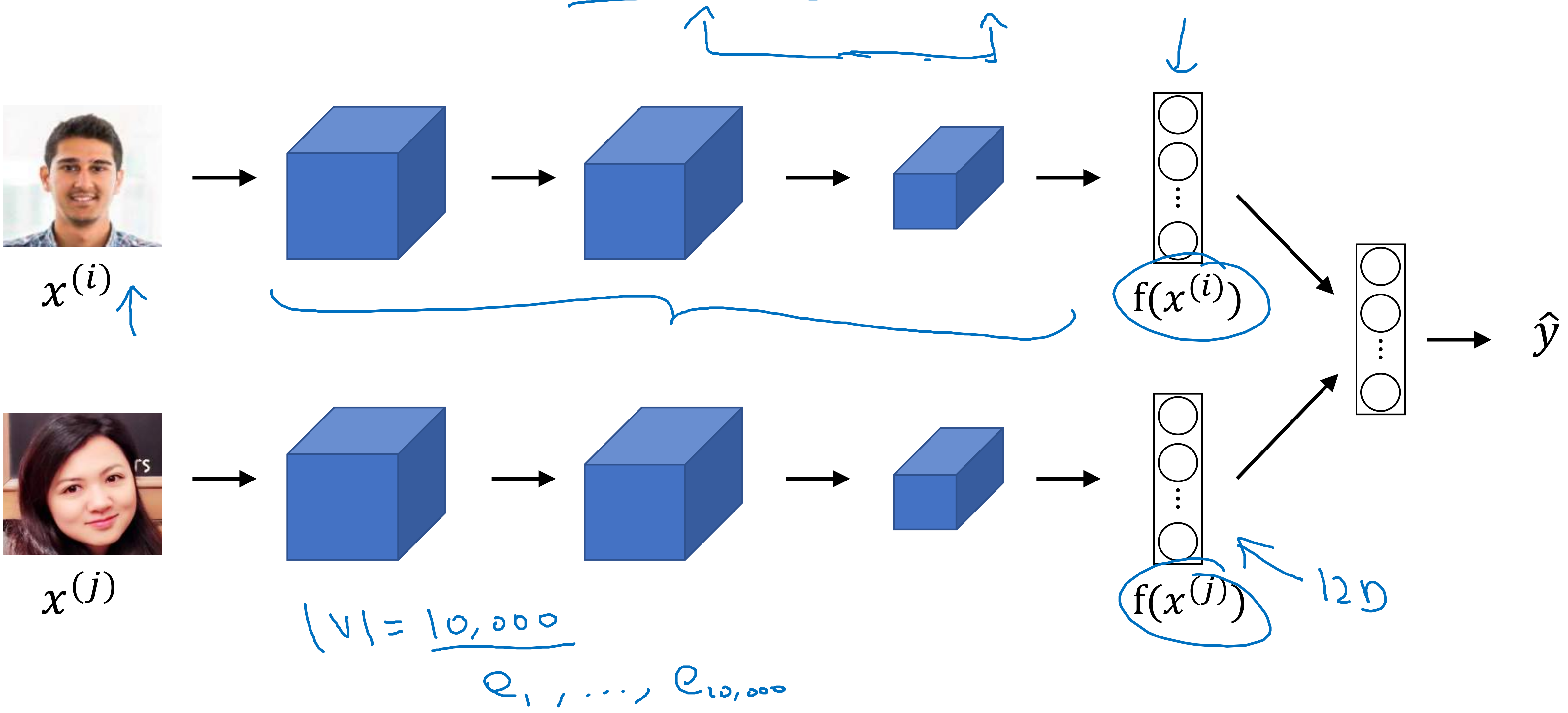
   (Or download pre-trained embedding online.)

Ⓑ

2. Transfer embedding to new task with smaller training set. (say, 100k words) →10,000 →300

3. Optional: Continue to finetune the word embeddings with new data.

Relation to face encoding (embedding) 128D



$|V| = \underline{10,000}$

$e_1, \ldots, e_{10,000}$

[Taigman et. al., 2014. DeepFace: Closing the gap to human level performance]

deeplearning.ai

NLP and Word Embeddings

Properties of word embeddings

Analogies

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	−1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97

e_{5391}

e_{man}

e_{woman}

Man → Woman as King → ? Queen

$$e_{man} - e_{woman} \approx e_{king} - e_{?}$$

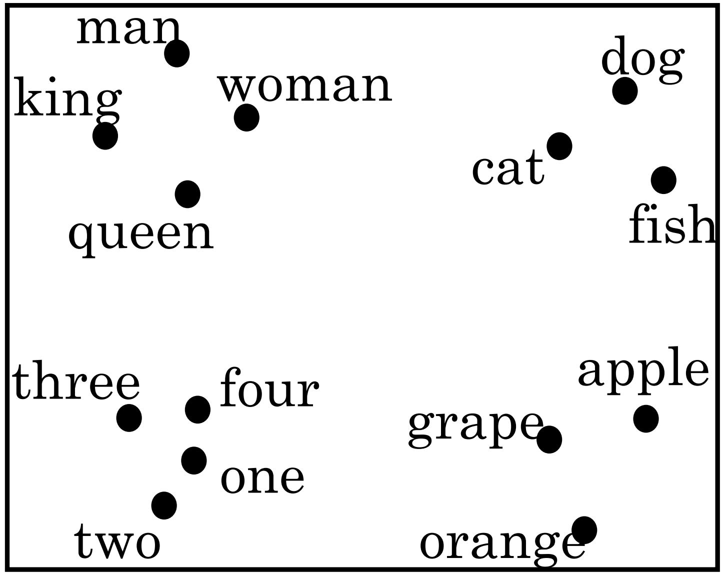
$\uparrow$ queen

$$e_{man} - e_{woman} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$e_{king} - e_{queen} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

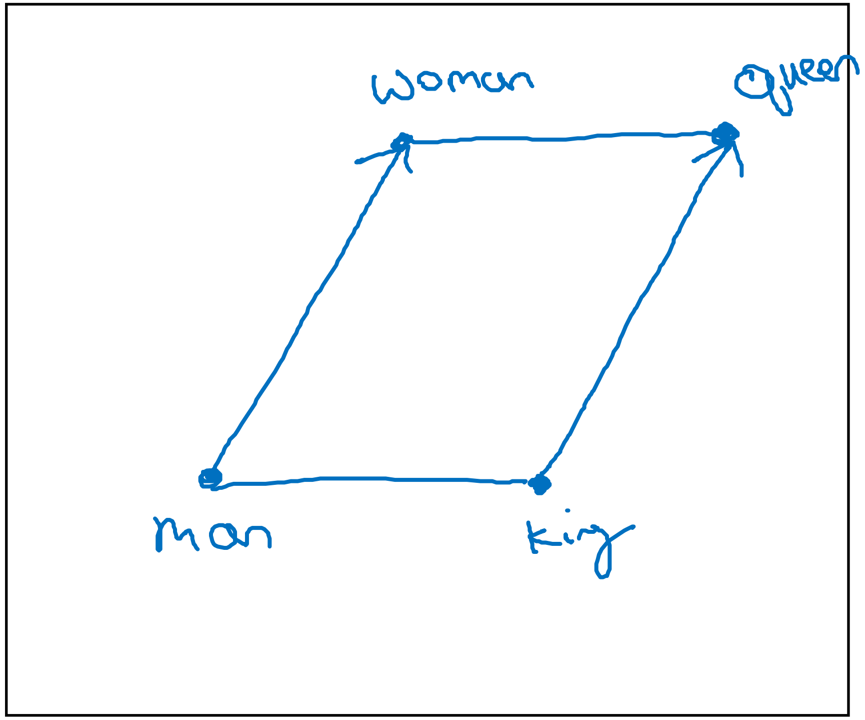
[Mikolov et. al., 2013, Linguistic regularities in continuous space word representations]

Analogies using word vectors



300D → 2D

t-SNE



300 D

$$\underbrace{e_{man} - e_{woman} \approx e_{king} - \cancel{e_?}}\ e_w$$

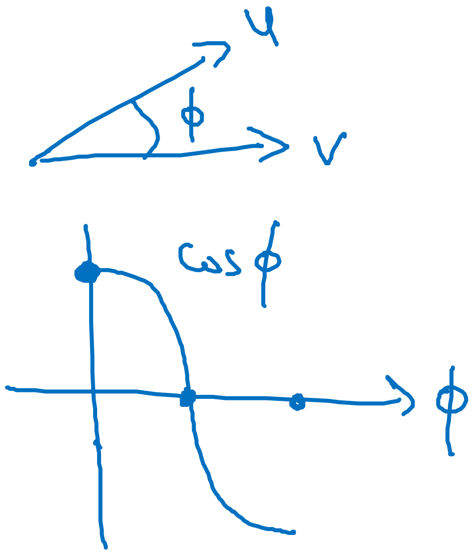
Find word w: $\arg\max_{w} \boxed{sim(\underline{e_w}, \underline{e_{king} - e_{man} + e_{woman}})}$

30 - 75%

Cosine similarity

$$\rightarrow \boxed{sim(e_w, e_{king} - e_{man} + e_{woman})}$$

$$\underline{sim(u, v)} = \frac{u^T v}{\|u\|_2 \|v\|_2}$$



$$\|u - v\|^2$$

Man:Woman as Boy:Girl

Ottawa:Canada as Nairobi:Kenya

Big:Bigger as Tall:Taller

Yen:Japan as Ruble:Russia

deeplearning.ai

NLP and Word Embeddings

Embedding matrix

Embedding matrix

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

$$\rightarrow \underline{E \cdot O_j = e_j}$$

= embedding for word j

In practice, use specialized function to look up an embedding.

→ Embedding

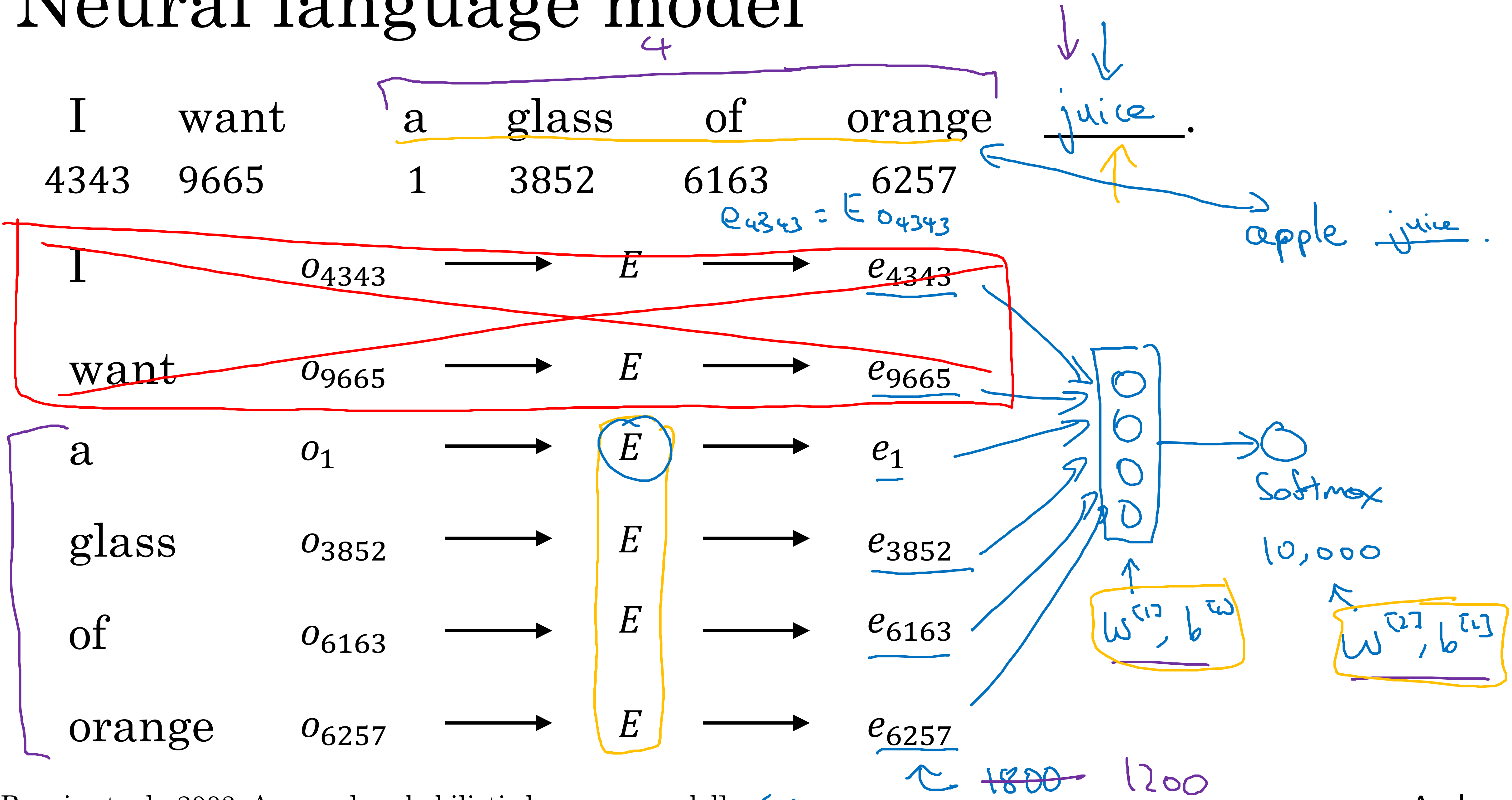
Andrew Ng

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	

apple juice .

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

I o_{4343} → E → e_{4343}

want o_{9665} → E → e_{9665}

a o_{1} → E → e_{1}

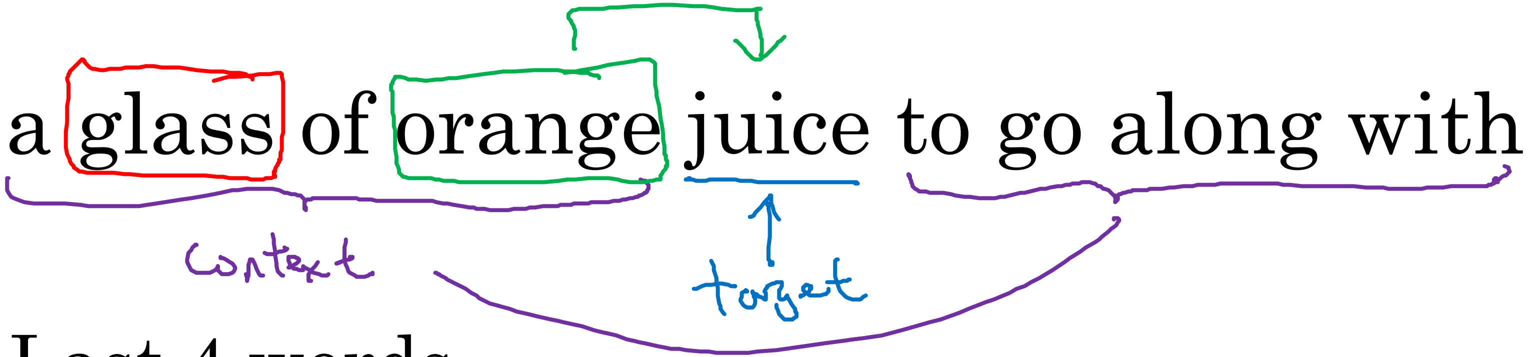
glass o_{3852} → E → e_{3852}

of o_{6163} → E → e_{6163}

orange o_{6257} → E → e_{6257}

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

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	a glass of orange ___?___ to go alg with
Last 1 word	orange ___?___
Nearby 1 word	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

$$x \longrightarrow \underline{y}$$

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

$$o_c \longrightarrow E \longrightarrow e_c \longrightarrow \underset{\text{Softmax}}{o} \longrightarrow \hat{y}$$

$$e_c = E o_c$$

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

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

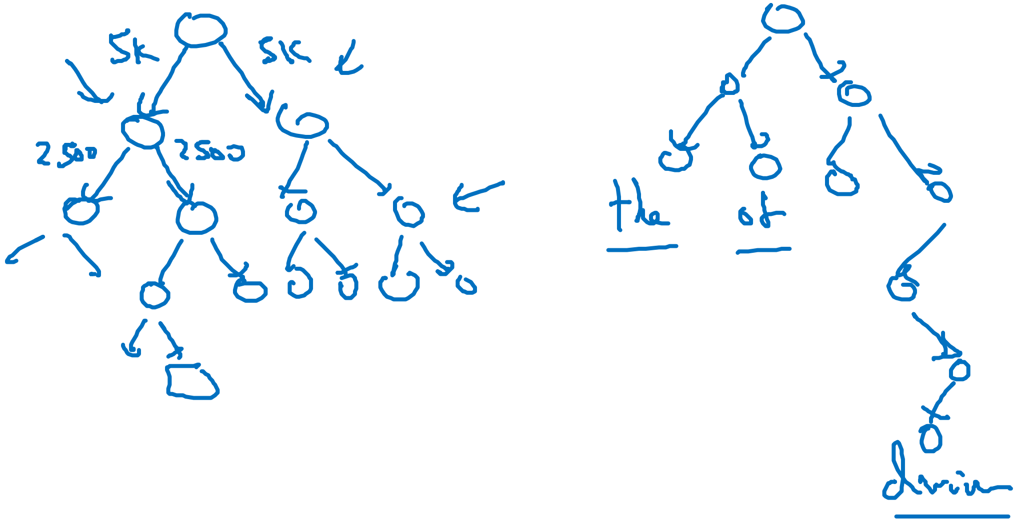
$$y = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow 4834$$

Problems with softmax classification

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

Hierarchical softmax.

log |V|



How to sample the context c?

→ the, of, a, and, to, ...

→ orange, apple, durian

durian

t

c → t

P(c)

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.

$x \longrightarrow y$

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]

Model

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

10,000-way softmax

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

x y

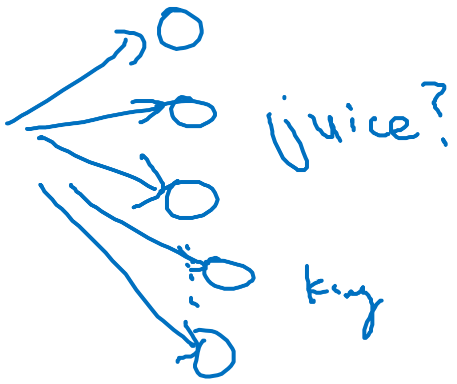
context	word	target?
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

k

c t y

Orange 6257

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



10,000

10,000 binary classification problem

$\underline{k+1}$

Andrew Ng

Selecting negative examples

$$\underbrace{\text{context} \quad \text{word}}^{x} \quad \underbrace{\text{target?}}^{y}$$

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

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

$$\underbrace{X_{ij}} = X_{ji} \leftarrow$$

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

Model

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

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

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

weighting term

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

“$0 \log 0$” $= 0$

this, is, of, a, ...

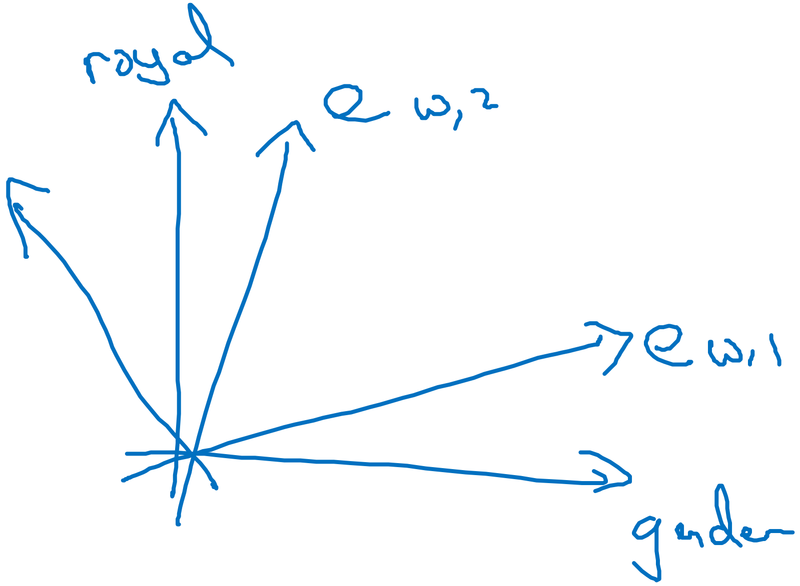
durian

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



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

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

deeplearning.ai

NLP and Word Embeddings

Sentiment classification

Sentiment classification problem

$$\underline{x} \longrightarrow y$$

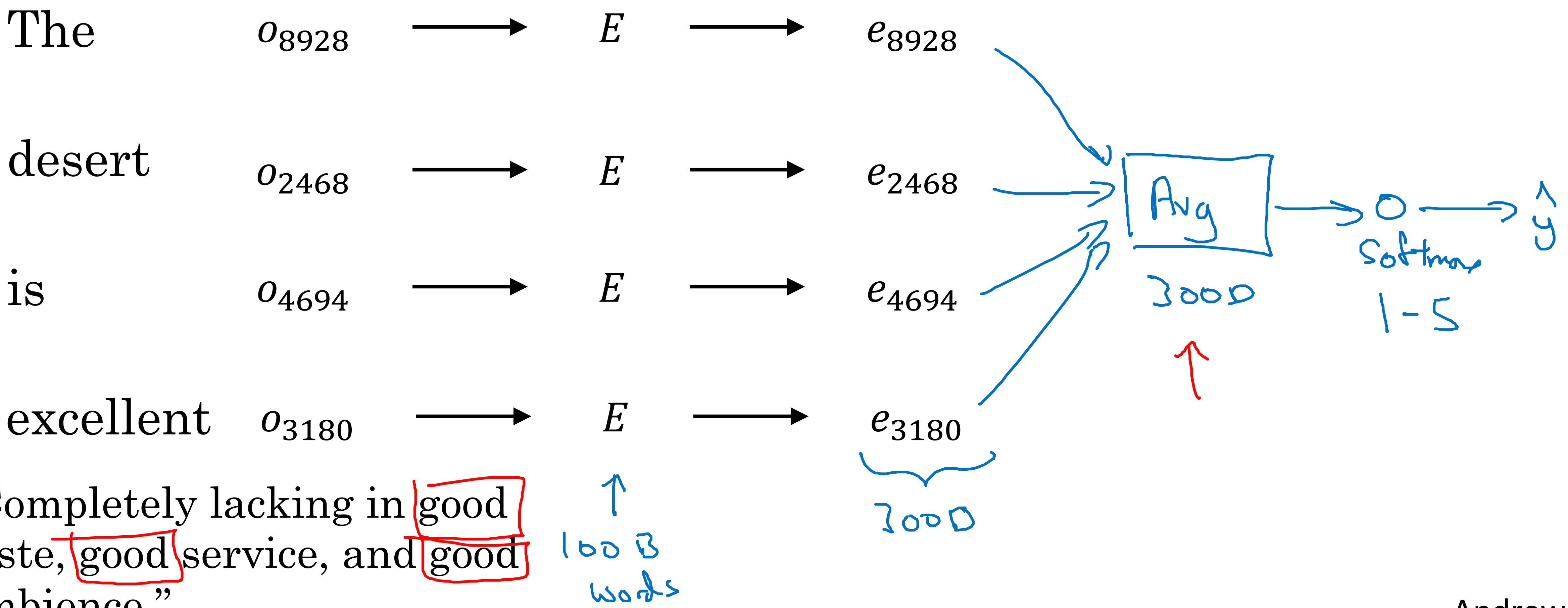
The dessert is excellent.	★★★★☆
Service was quite slow.	★★☆☆☆
Good for a quick meal, but nothing special.	★★★☆☆
Completely lacking in good taste, good service, and good ambience.	★☆☆☆☆

$\underline{10,000}$ → 100,000 words

Simple sentiment classification model

The dessert is excellent ★★★★☆

8928 2468 4694 3180



The $o_{8928} \longrightarrow E \longrightarrow e_{8928}$

desert $o_{2468} \longrightarrow E \longrightarrow e_{2468}$

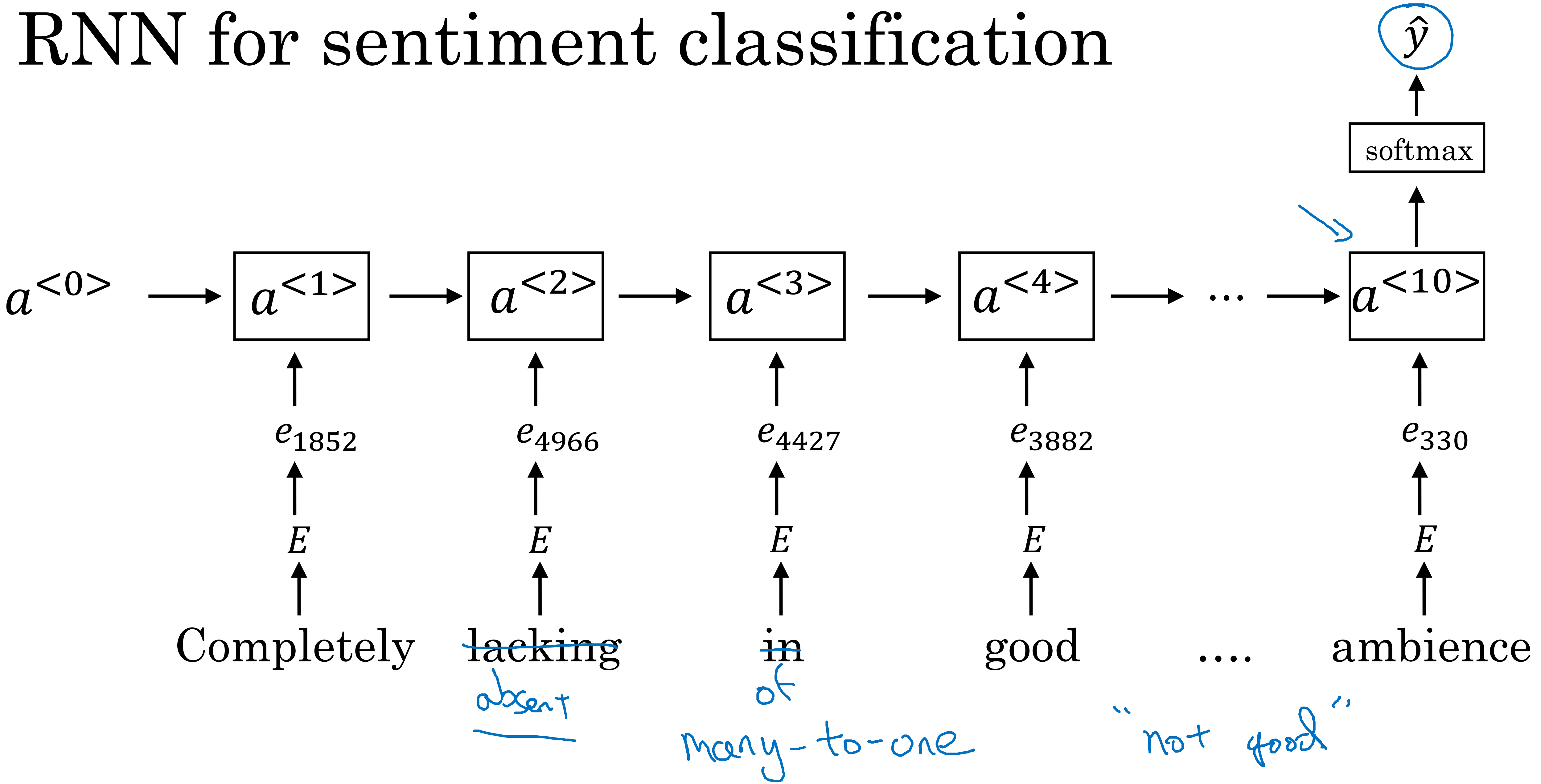
is $o_{4694} \longrightarrow E \longrightarrow e_{4694}$

excellent $o_{3180} \longrightarrow E \longrightarrow e_{3180}$

"Completely lacking in good taste, good service, and good ambience."

Andrew Ng

RNN for sentiment classification



$\hat{y}$

softmax

$a^{<0>}$ → $a^{<1>}$ → $a^{<2>}$ → $a^{<3>}$ → $a^{<4>}$ → … → $a^{<10>}$

e_{1852} e_{4966} e_{4427} e_{3882} e_{330}

E E E E E

Completely ~~lacking~~ ~~in~~ good …. ambience

absent

of

many-to-one

"not good"

Andrew Ng

NLP and Word Embeddings

Debiasing word embeddings

deeplearning.ai

The problem of bias in word embeddings

Man:Woman as King:Queen

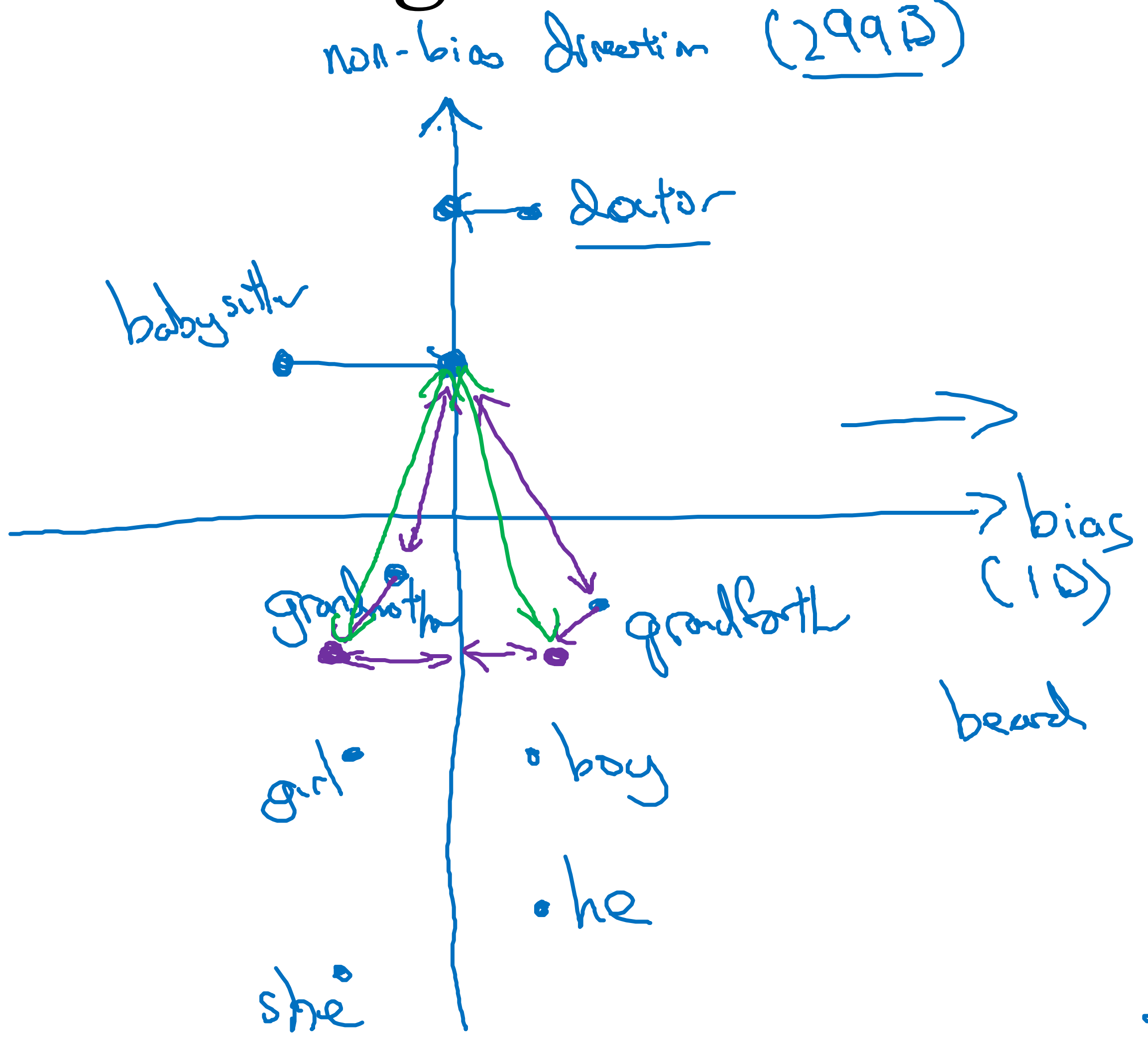
Man:Computer_Programmer as Woman: Homemaker ✗

Father:Doctor as Mother: Nurse ✗

Word embeddings can reflect gender, ethnicity, age, sexual orientation, and other biases of the text used to train the model.

[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]

Addressing bias in word embeddings



1. Identify bias direction.

$e_{he} - e_{she}$

$e_{male} - e_{female}$

⋮

→ average

2. Neutralize: For every word that is not definitional, project to get rid of bias.

3. Equalize pairs.

→ grandmoth – grandfath

girl – boy

[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]