Copyright Notice

NLP and Word Embeddings

deeplearning.ai

Word representation

Word representation

V = [a, aaron, ..., zulu, <UNK>]

$|V| = 10,000$

1-hot representation

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

I want a glass of orange juice.

I want a glass of apple ______.

It willln't be able to think that orange and apple are somewhat closel

Featurized representation: word embedding

	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}				

300

I want a glass of orange juice.

I want a glass of apple juice.

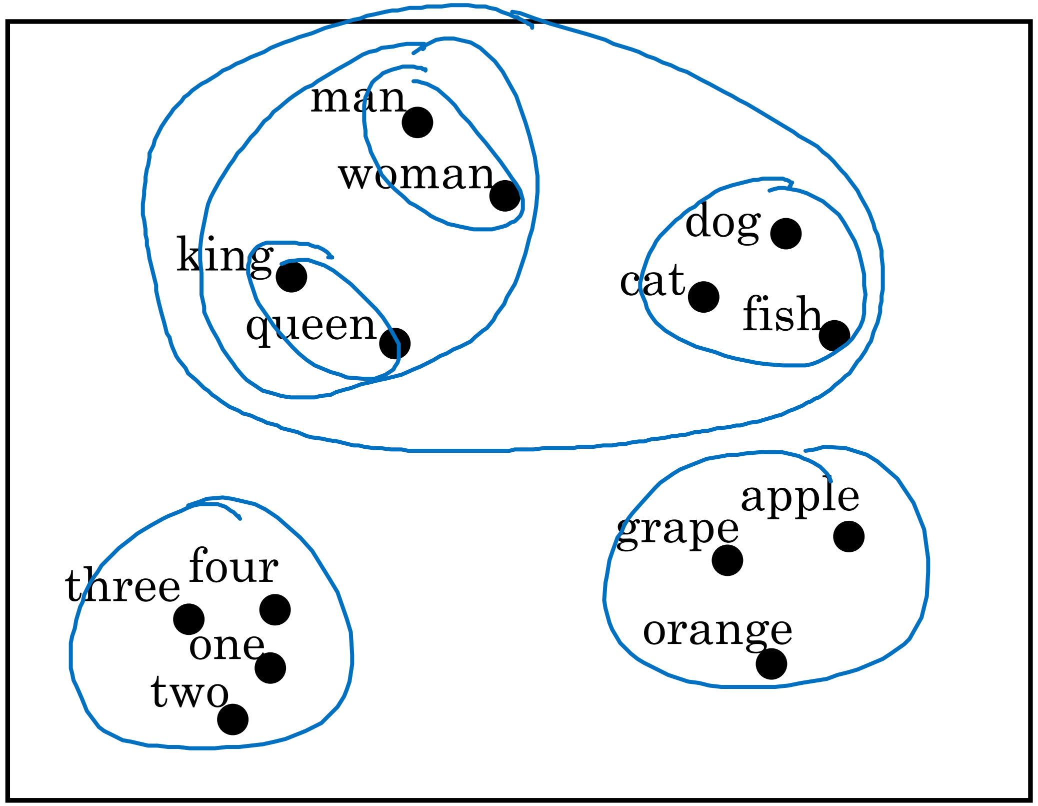
Andrew Ng

Visualizing word embeddings

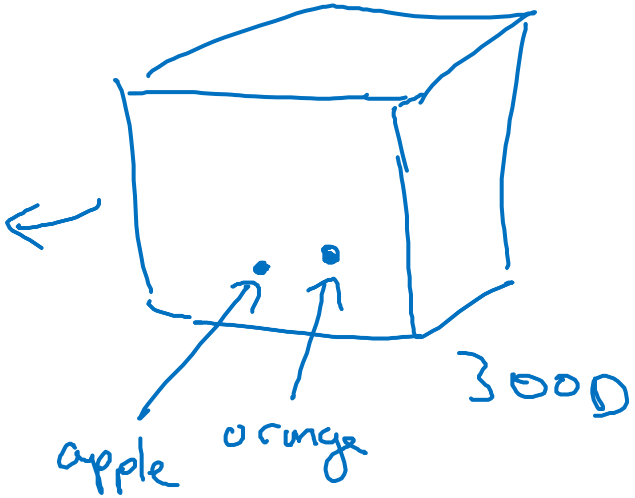
→ 300 D

↓

2 D



t-SNE



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

Andrew Ng

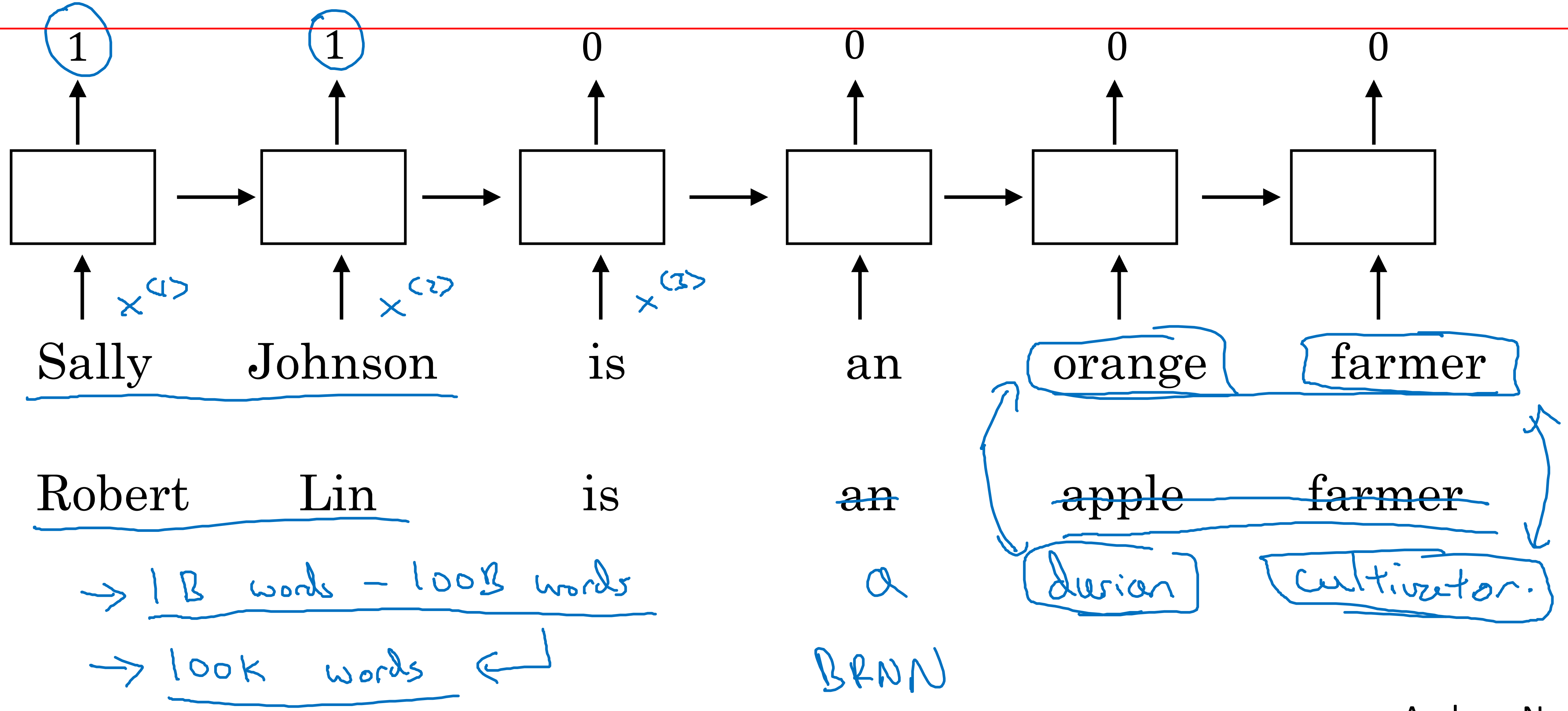
NLP and Word Embeddings

deeplearning.ai

Using word embeddings

Named entity recognition example

Let's say our model figured it out that Sally Joohnson is a name . Using the closeness of the apple and juice model can also figure it out that Robert Lin is



Andrew Ng

Transfer learning and word embeddings

A

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

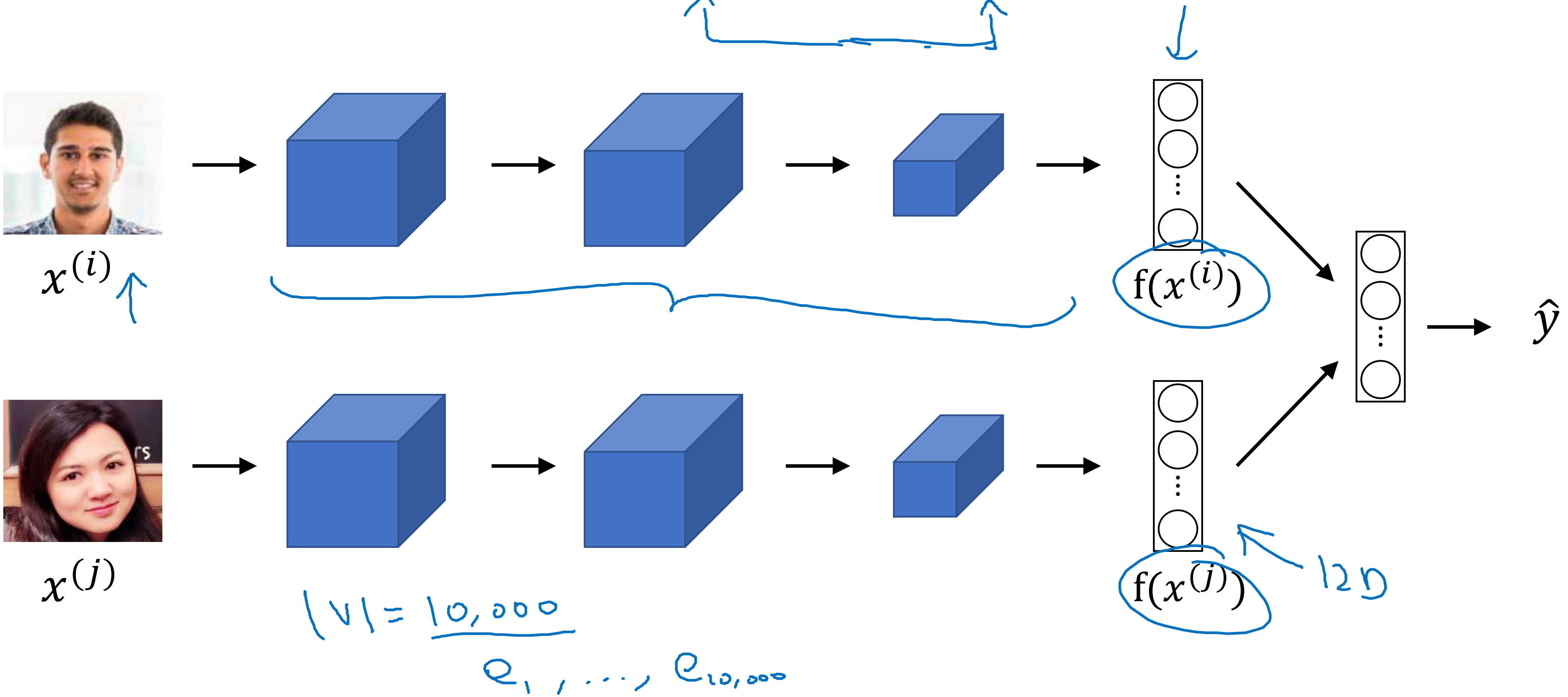
   (Or download pre-trained embedding online.)

B

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)



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

Andrew Ng

Using featurized representation can also help us to learn the analogies. Ma

NLP and Word Embeddings

deeplearning.ai

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

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

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

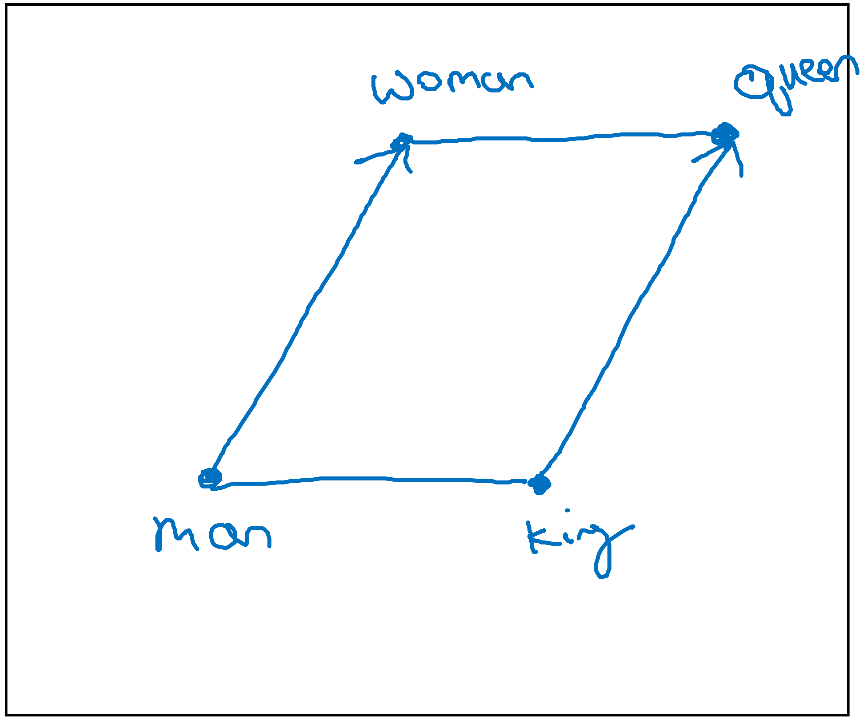
$$e_{king} - e_{queen} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

You see difference simila

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

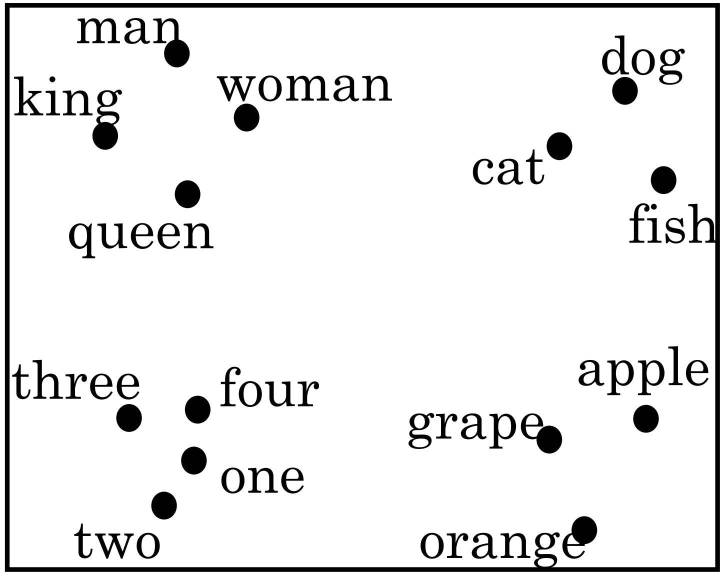
Andrew Ng

Analogies using word vectors



300 D

300D → 2D



t-SNE

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

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

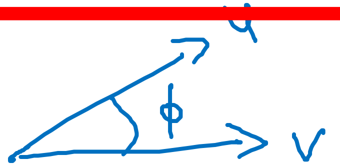
30 - 75%

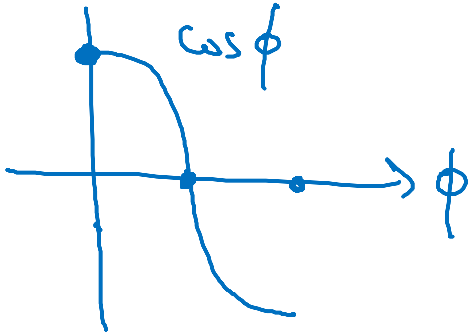
Cosine similarity

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

find the word w that maximises the similarity of king with

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





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

Man:Woman as Boy:Girl

Ottawa:Canada as Nairobi:Kenya

Big:Bigger as Tall:Taller

Yen:Japan as Ruble:Russia

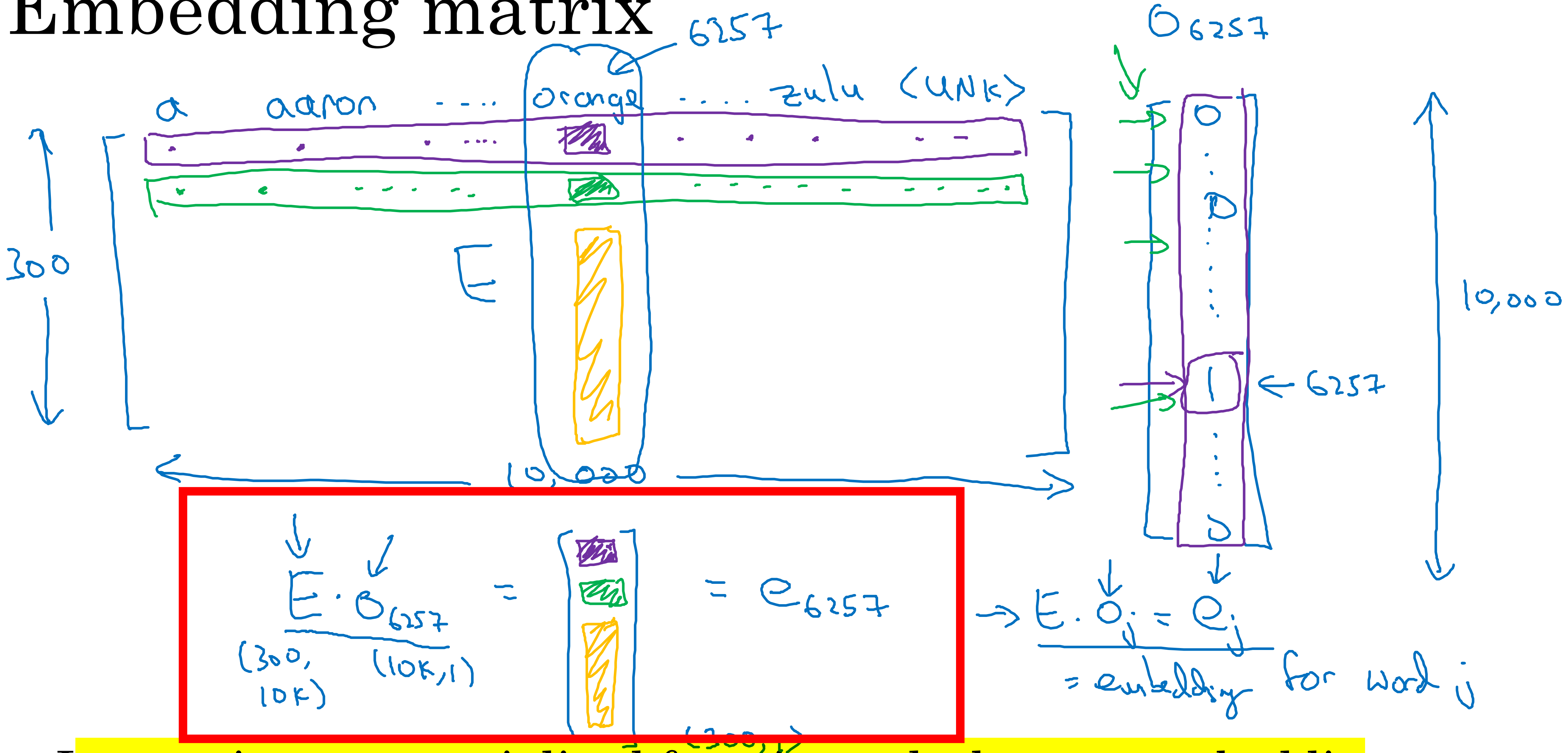
deeplearning.ai

NLP and Word Embeddings

Embedding matrix

Let's learn 300 parameter for each of the word so that if our corpus is 10000 words our embedding matr

Embedding matrix



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

→ Embedding

Andrew Ng

Each of the e is a 300 dimensional vector and we can feed all of them to a NN having a softmax word and the softma

NLP and Word Embeddings

deeplearning.ai

Learning word embeddings

Neural language model

4

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

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

apple juice.

I	o_{4343}	$\longrightarrow E \longrightarrow$	e_{4343}
want	o_{9665}	$\longrightarrow E \longrightarrow$	e_{9665}
a	o_{1}	$\longrightarrow E \longrightarrow$	e_{1}
glass	o_{3852}	$\longrightarrow E \longrightarrow$	e_{3852}
of	o_{6163}	$\longrightarrow E \longrightarrow$	e_{6163}
orange	o_{6257}	$\longrightarrow E \longrightarrow$	e_{6257}

Softmax

10,000

$W^{[1]}, b^{[1]}$

$W^{[2]}, b^{[2]}$

~~1800~~ 1200

If you want to predict on the bas

[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

skip gram

a glass of orange ? to go along with

orange ?

glass ?

Andrew Ng

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

randomly choose context and randomly cho

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

Model

Vocab size = 10,000k

$x \rightarrow y$

Context c ("orange") 6257 $\rightarrow$ Target t ("juice") 4834

$$O_c \rightarrow E \rightarrow e_c \rightarrow O \rightarrow \hat{y}$$

Softmax

$$e_c = E O_c$$

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

θ_t = parameter associated with output t

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

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

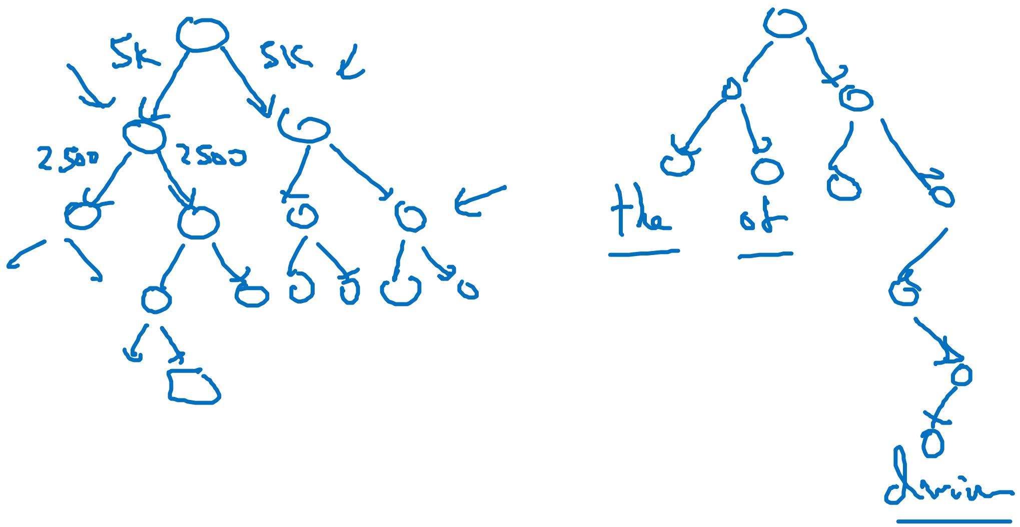
Andrew Ng

Problems with softmax classification

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

log |V|

Hierarchical softmax.



very large computation and very slow .

How to sample the context c?

t

→ the, of, a, and, to, ...

c → t

→ orange, apple, durian

P(c)

durian

Andrew Ng

NLP and Word Embeddings

deeplearning.ai

Negative sampling

We take a pair of context and words and try to predict wheter these can act as a context and target pairs

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

$k = 5\text{-}20$ smaller datasets

$k = 2\text{-}5$ larger dataset

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

Model

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

10,000-way softmax

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

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

x   y

k

c   t   y

Orange 6257

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

juice?

king

10,000

10,000 binary classification problem

k+1

Andrew Ng

Selecting negative examples

$\overbrace{\text{context} \quad \text{word}}^{x}$		$\overbrace{\text{target?}}^{y}$
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

t

the, of, and, ...

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

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

Andrew Ng

NLP and Word Embeddings

deeplearning.ai

GloVe word vectors

GloVe (global vectors for word representation)

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

c, t

X_{ij} = # times j appears in context of i.

c t

t

c

$\boxed{X_{ij} = X_{ji}}$

depends upon our definition of the target word . If target withing +-10 words of the

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

Model

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

$t \quad c \quad \to 0?$

"$\theta_t^T e_c$"

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

this, is, of, a, ...

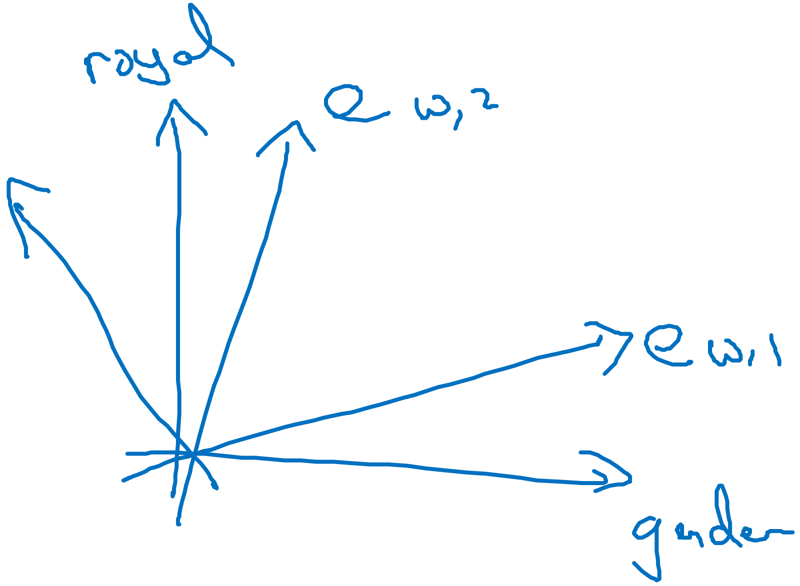
durian

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



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

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

Andrew Ng

deeplearning.ai

NLP and Word Embeddings

Sentiment classification

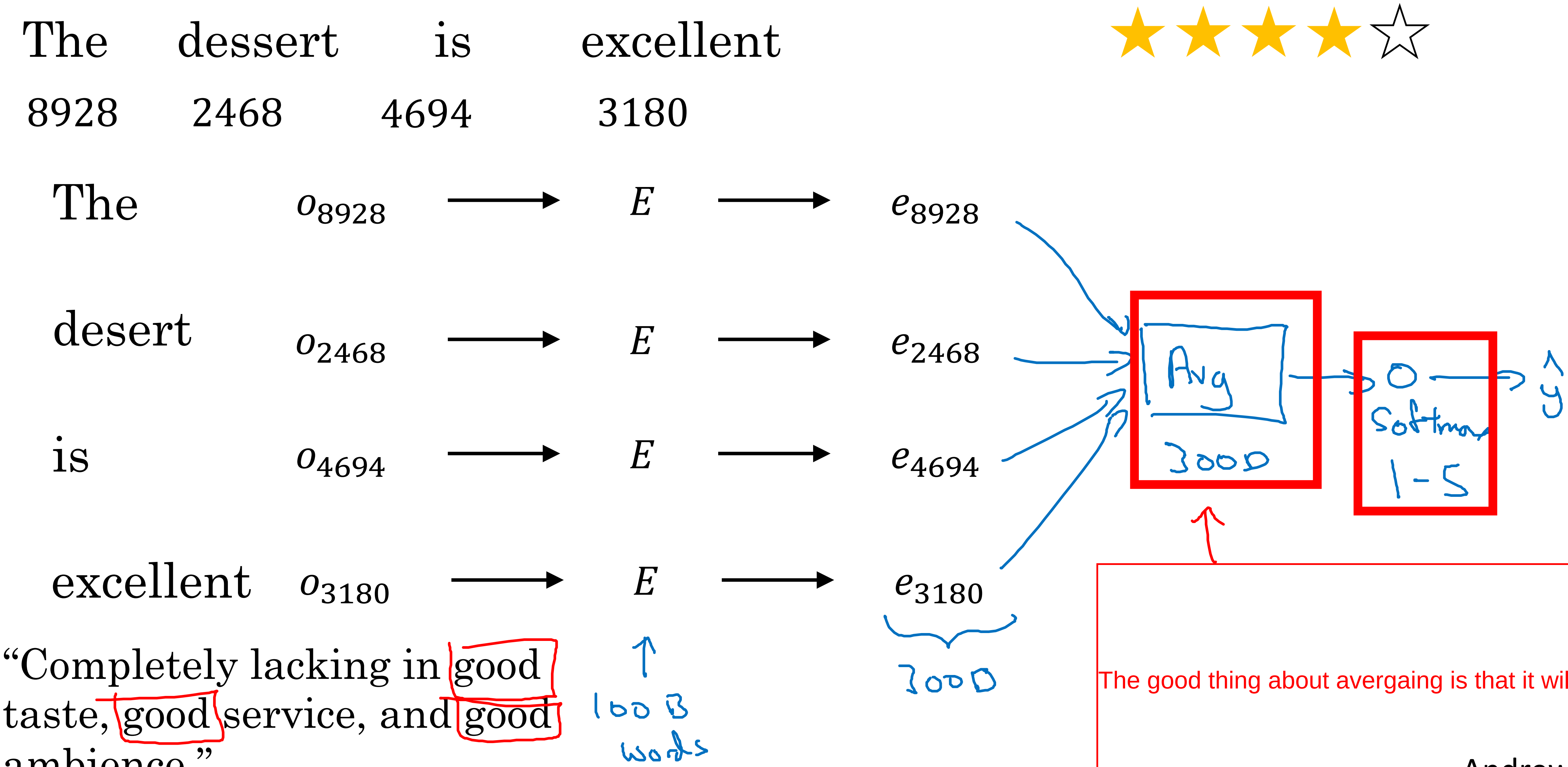
Sentiment classification problem

$\underline{x} \longrightarrow y$

x	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.	★☆☆☆☆

$\underline{10{,}000} \rightarrow 100{,}000$ words

Simple sentiment classification model

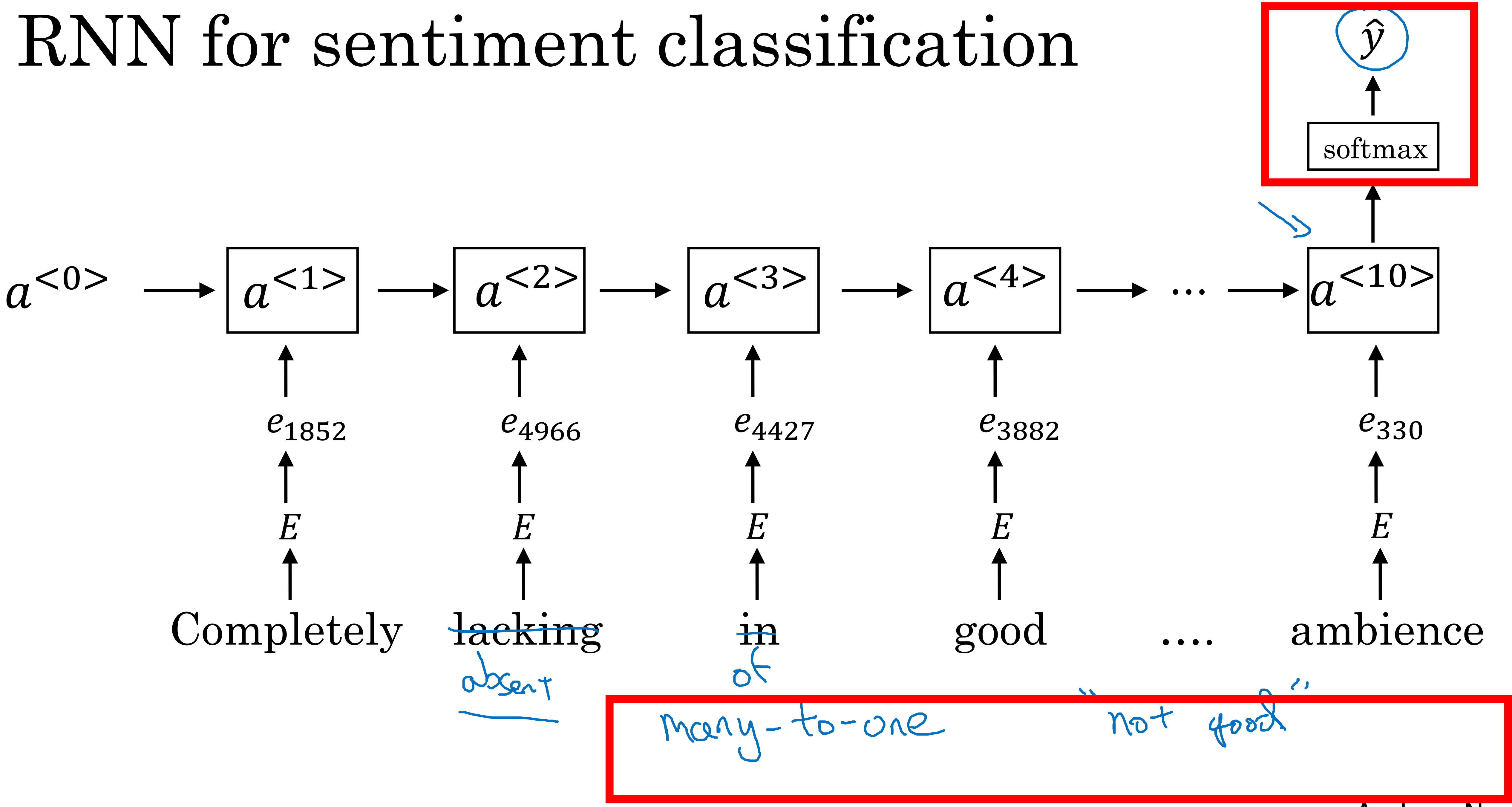


"Completely lacking in good taste, good service, and good ambience."

The good thing about avergaing is that it will alwa

Andrew Ng

RNN for sentiment classification



We want our model to be not bias to gender.

NLP and Word Embeddings

deeplearning.ai

Debiasing word embeddings

The problem of bias in word embeddings

Man:Woman as King:Queen

Man:Computer_Programmer as Woman:Homemaker

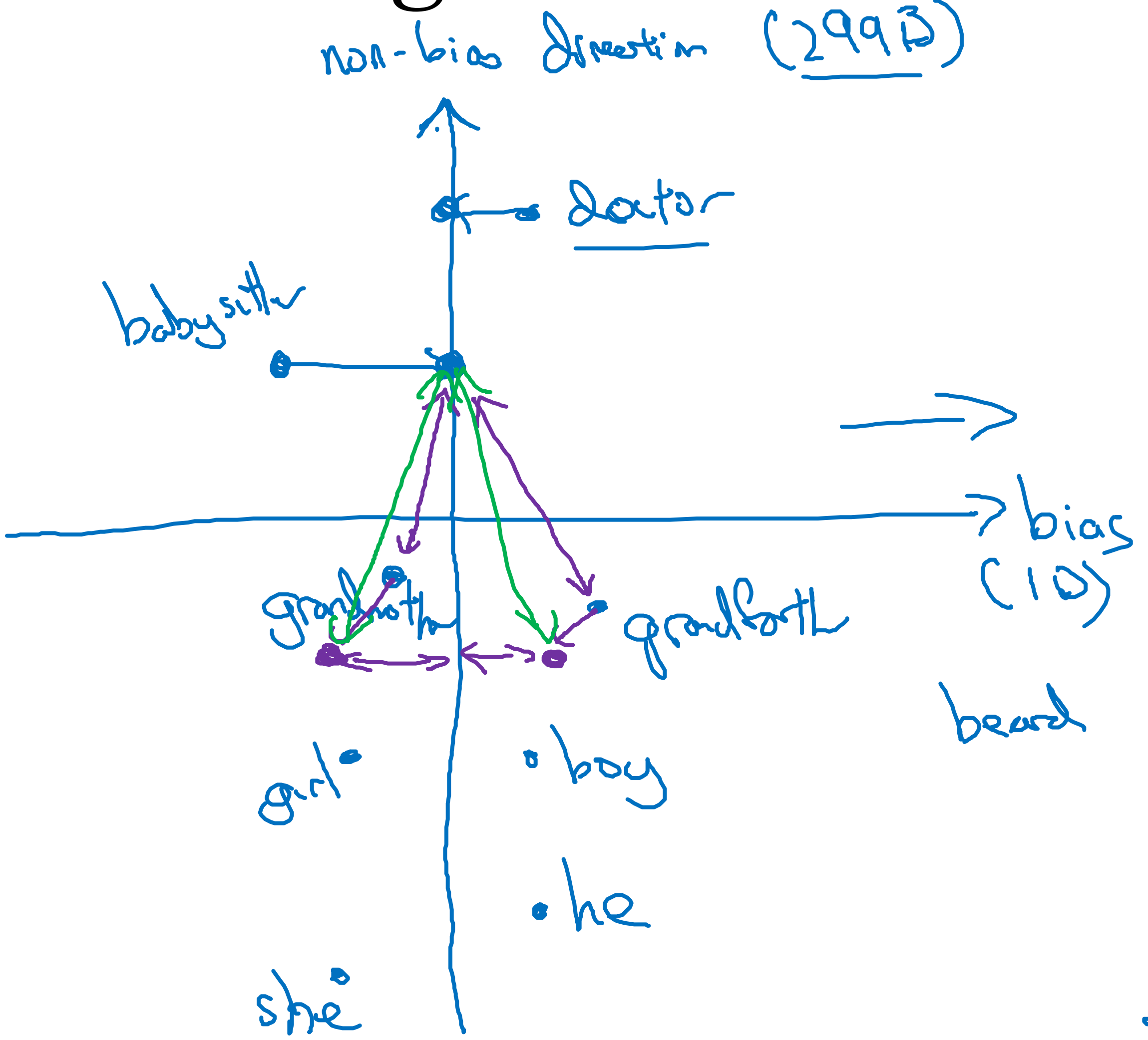
Father:Doctor as Mother:Nurse

We don't want these gender biases in out model.

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.

$e_{he} - e_{she}$

$e_{male} - e_{female}$

$\vdots$

average

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

3. Equalize pairs.

grandmother – grandfather

girl – boy

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

Andrew Ng