Copyright Notice

These slides are distributed under the Creative Commons License.

DeepLearning.AI makes these slides available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite DeepLearning.AI as the source of the slides.

For the rest of the details of the license, see https://creativecommons.org/licenses/by-sa/2.0/legalcode

NLP and Word Embeddings

Word representation

deeplearning.ai

Word representation

V = [a, aaron, …, zulu, <UNK>]

$|V| = 10,000$

1-hot representation

Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	1	0
0	0	⋮	0	⋮	0
⋮	0	1	0	0	0
1	⋮	⋮	⋮	0	0
⋮	1	0	1	0	1
0	⋮	0	⋮	0	⋮
0	0	0	0	0	0

O_{5391} O_{9853}

I want a glass of orange juice.

I want a glass of apple _____ ?.

Andrew Ng
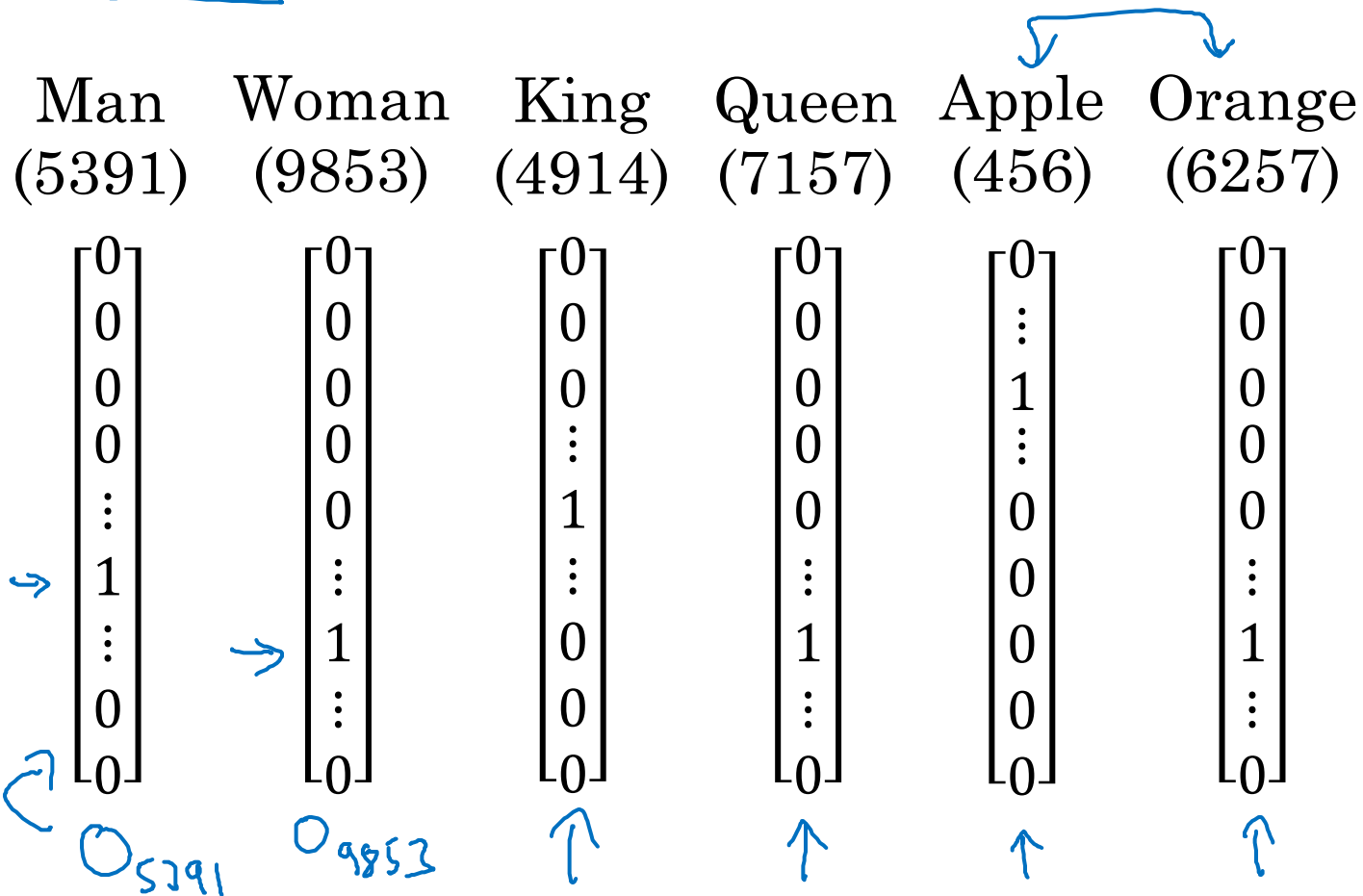
Featurized representation: word embedding

	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

300

size
cost
alive verb

e_{5391} e_{9853}

I want a glass of orange juice.

I want a glass of apple juice.
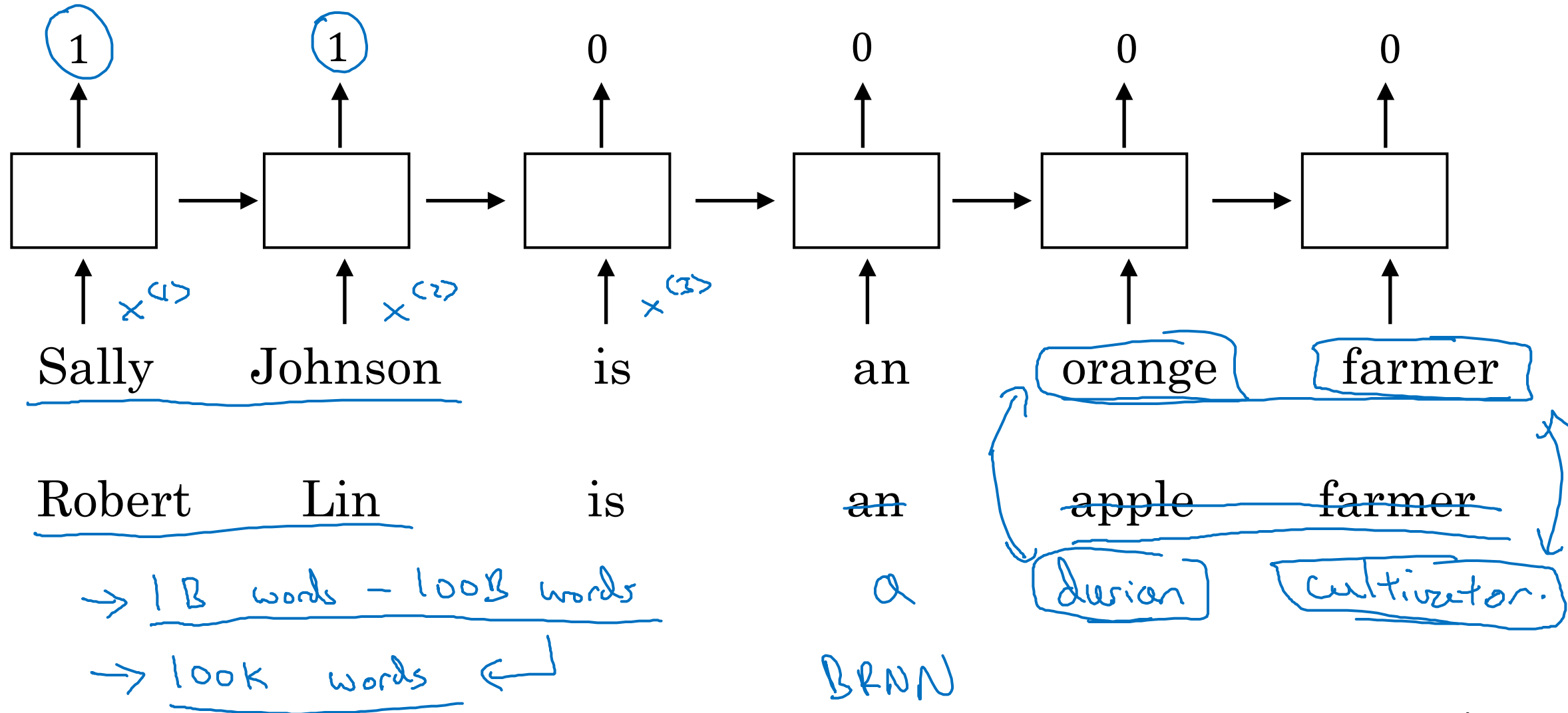
Andrew Ng

Visualizing word embeddings

t-SNE

→ 300 D
↓
2 D

300 D

apple   orange

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

Andrew Ng

NLP and Word Embeddings

Using word embeddings

deeplearning.ai

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

→ 100k words

a

durian   cultivator.

BRNN
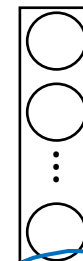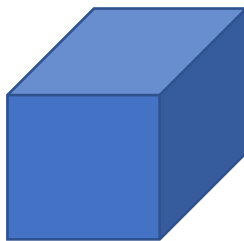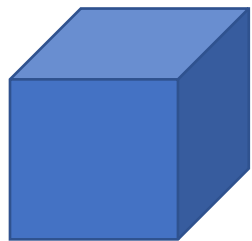
Andrew Ng

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) $\rightarrow 10,000$ $\rightarrow 300$

3. Optional: Continue to finetune the word embeddings with new data.
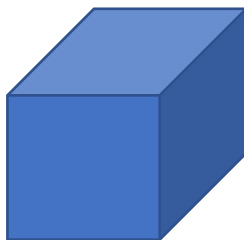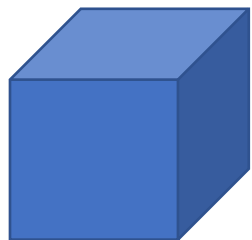
Relation to face encoding (embedding)

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

Andrew Ng

NLP and Word Embeddings

Properties of word embeddings

deeplearning.ai

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 $\rightarrow$ Woman   as   King $\rightarrow$? 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]   Andrew Ng

Analogies using word vectors
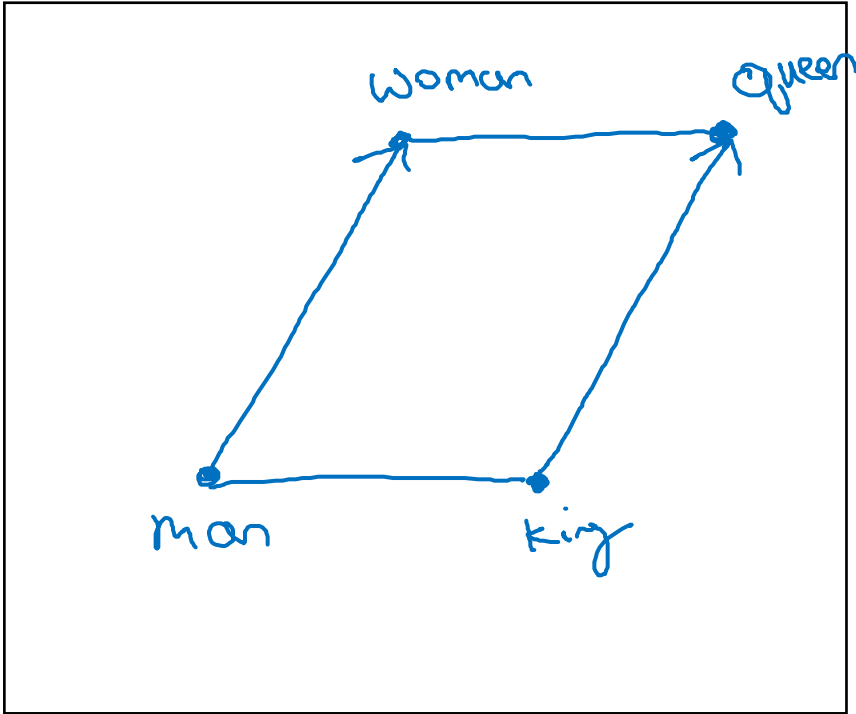
3000 → 2D
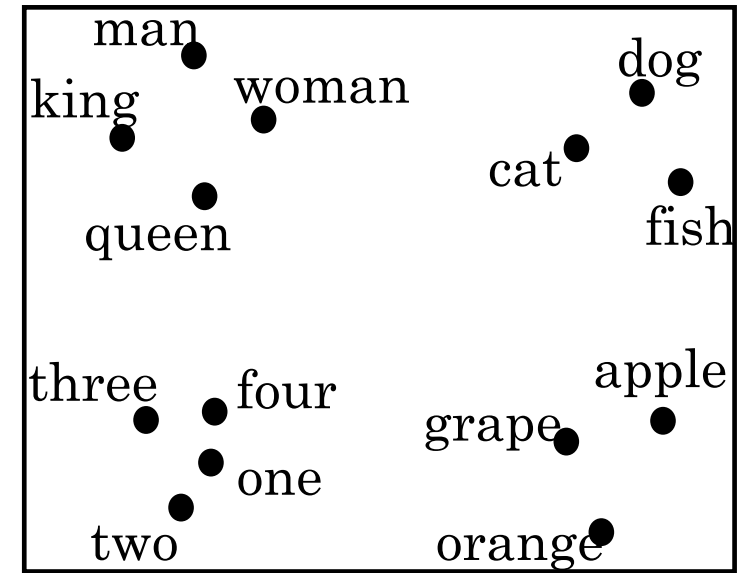
t-SNE

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

300 D

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

30 - 75%

Andrew Ng

Cosine similarity

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

$$Sim(u,v) = \frac{u^T v}{||u||_2 ||v||_2}$$
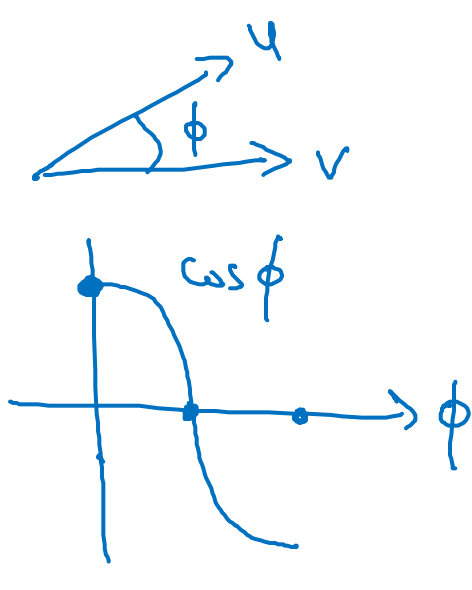
$$||u-v||^2$$

Man:Woman as Boy:Girl

Ottawa:Canada as Nairobi:Kenya

Big:Bigger as Tall:Taller

Yen:Japan as Ruble:Russia

NLP and Word Embeddings

Embedding matrix

deeplearning.ai

Embedding matrix

6257

a   aaron   ....   orange   ..... zulu ⟨UNK⟩

O_{6257}

300

E

10,000

6257

10,000

6257

$E \cdot O_{6257} =$ $= e_{6257}$

(300, 10K)   (10K, 1)

(300, 1)

$\rightarrow E \cdot O_j = e_j$
= embedding for word j
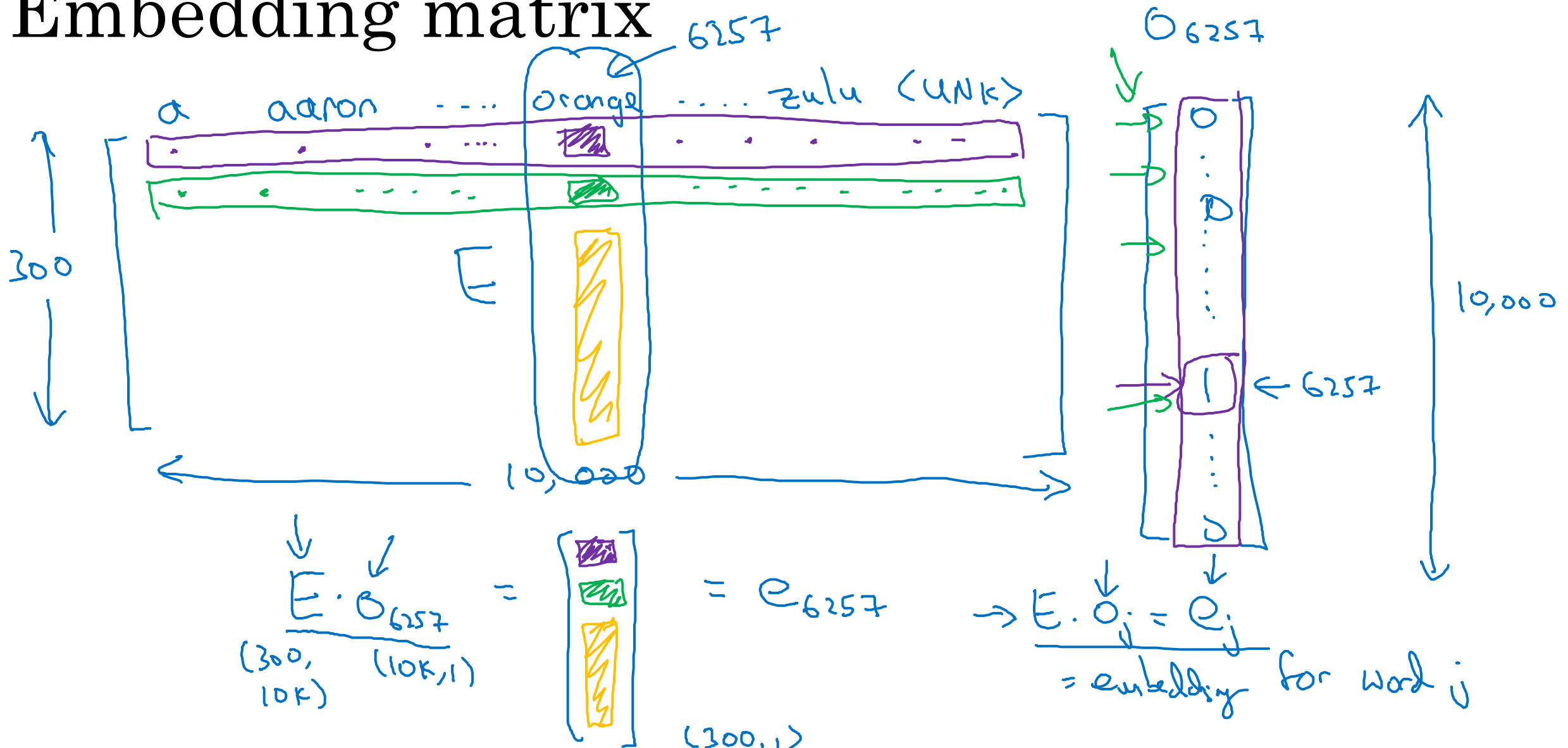
→ Embedding

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

deeplearning.ai

NLP and Word Embeddings

Learning word embeddings

Neural language model

I   want   a   glass   of   orange   <u>juice</u>.

4343 9665   1   3852   6163   6257

apple juice.

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

I	o_{4343}	$\rightarrow$	E	$\rightarrow$ e_{4343}
want	o_{9665}	$\rightarrow$	E	$\rightarrow$ e_{9665}
a	o_1	$\rightarrow$	E	$\rightarrow$ e_1
glass	o_{3852}	$\rightarrow$	E	$\rightarrow$ e_{3852}
of	o_{6163}	$\rightarrow$	E	$\rightarrow$ e_{6163}
orange	o_{6257}	$\rightarrow$	E	$\rightarrow$ e_{6257}

Softmax

10,000

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

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

1800 → 1200

[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

Andrew Ng

NLP and Word Embeddings

Word2Vec

deeplearning.ai

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.]

Andrew Ng

Model

Vocab size = 10,000k

x

y

Context c ("orange")   →   Target t

("juice")

6257

4834

O_c → E → e_c → Softmax ⊙ → $\hat{y}$

$e_c = E O_c$
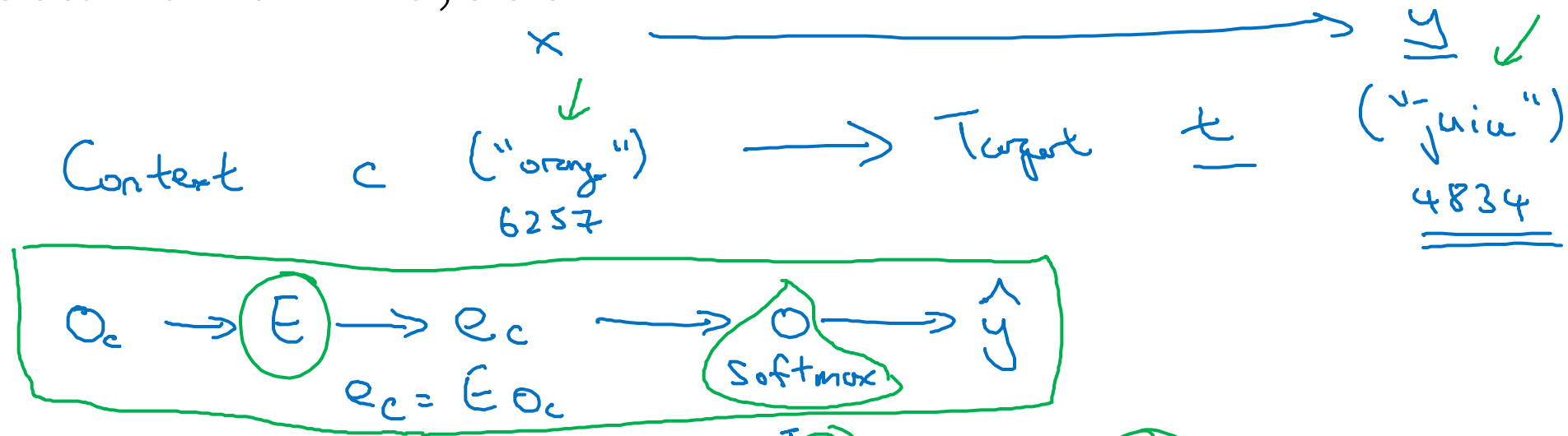
Softmax: $p(t|c) = \dfrac{e^{\Theta_t^T e_c}}{\sum\limits_{j=1}^{10,000} e^{\Theta_j^T e_c}}$

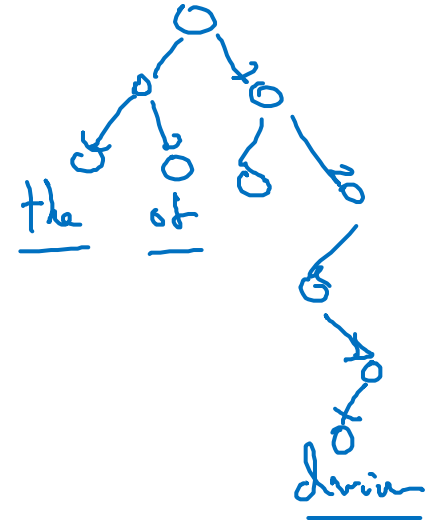Θ_t = parameter associated with output t

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

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

Andrew Ng

Problems with softmax classification

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

Hierarchial   softmax.

$\log |v|$

5k   5k

2500   2500

the   of

durian
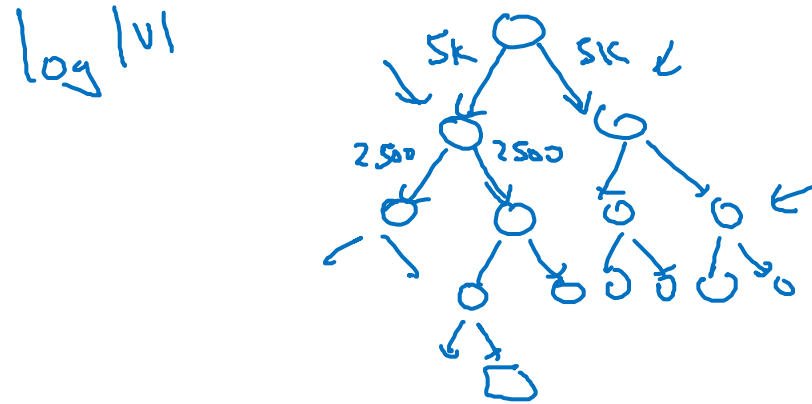
How to sample the context c?

$\rightarrow$ the, of, a, and, to, ...

$\rightarrow$ orange, apple, durian

2 durian

t

$c \rightarrow t$

$P(c)$

NLP and Word Embeddings

deeplearning.ai

Negative sampling

Defining a new learning problem

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

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

$k = 5-20$ smaller. datasets
$k = 2-5$ larger dataset

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

Andrew Ng
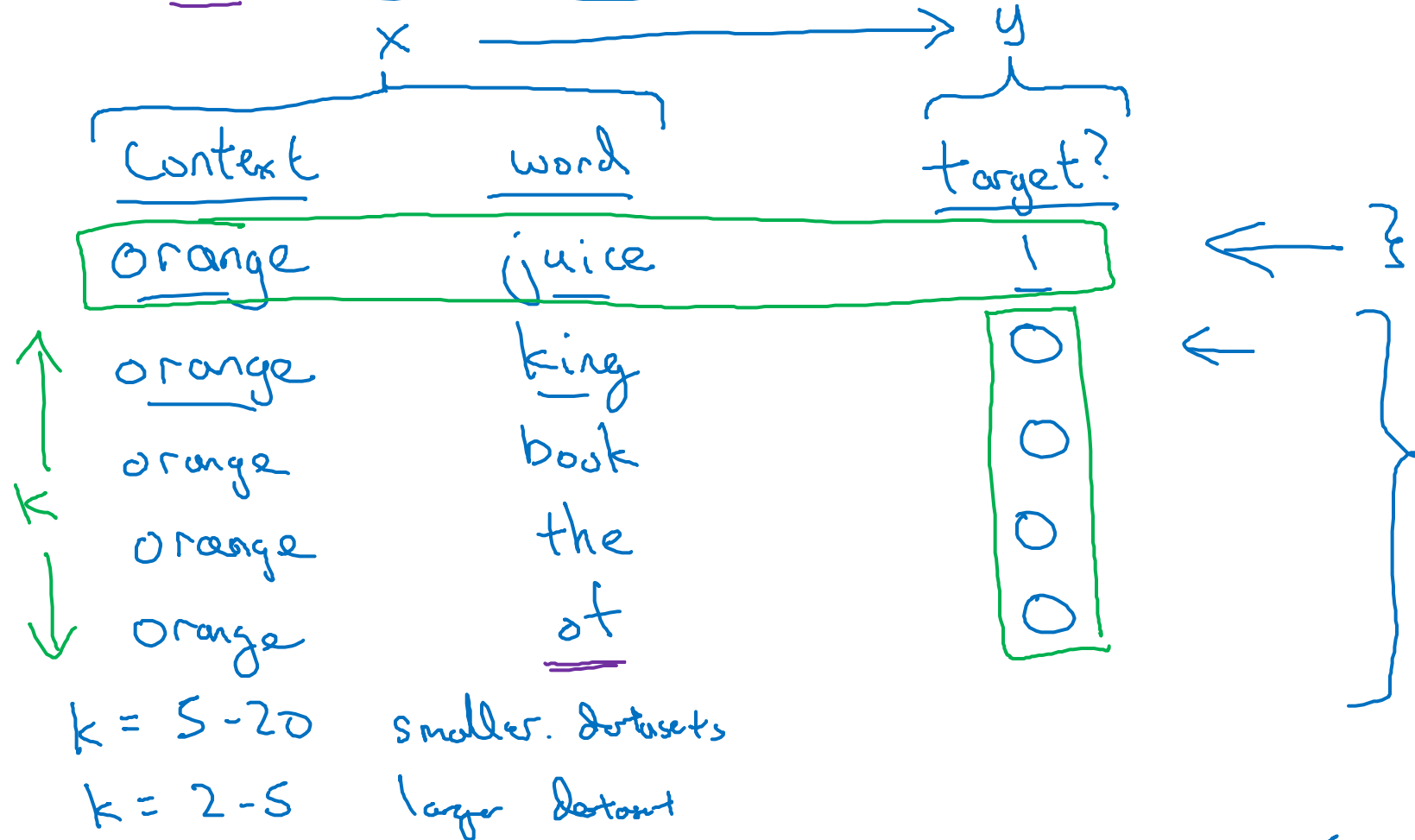
Model

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

$P(y=1|c,t) = \sigma\left(\theta_t^T e_c\right) \leftarrow$

x

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

y

c t y

Orange
6257

$o_{6257} \rightarrow E \rightarrow e_{6257}$ juice?   king
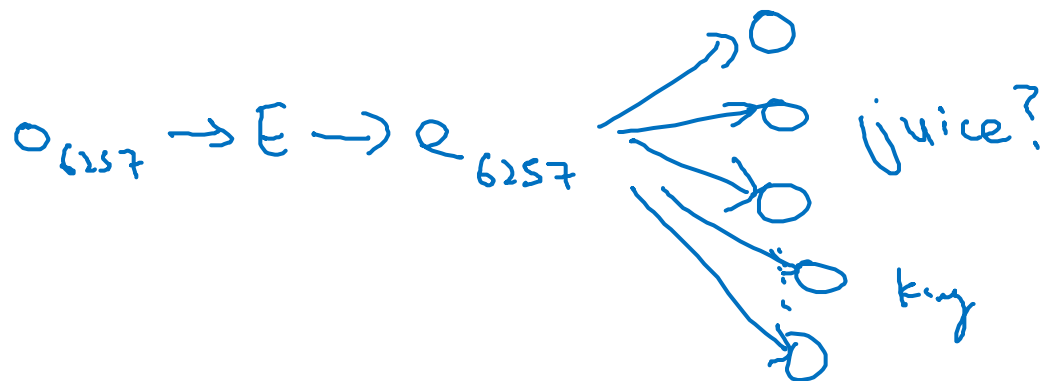
10,000

10,000 binary classification problem

$k+1$

Andrew Ng

Selecting negative examples

$\underbrace{\qquad\qquad}_{x}$ $\underbrace{\qquad}_{y}$

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

the, of, and, ...

$\uparrow$
x

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

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

Andrew Ng

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.

$$c, t$$

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

$$X_{ij} = X_{ji}$$

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

Andrew Ng

Model

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

$t \quad c$ (over $\Theta_i^T \, e_j$)

$t \quad c$

"$\Theta_t^T \, e_c$"

> 0?

weighting term
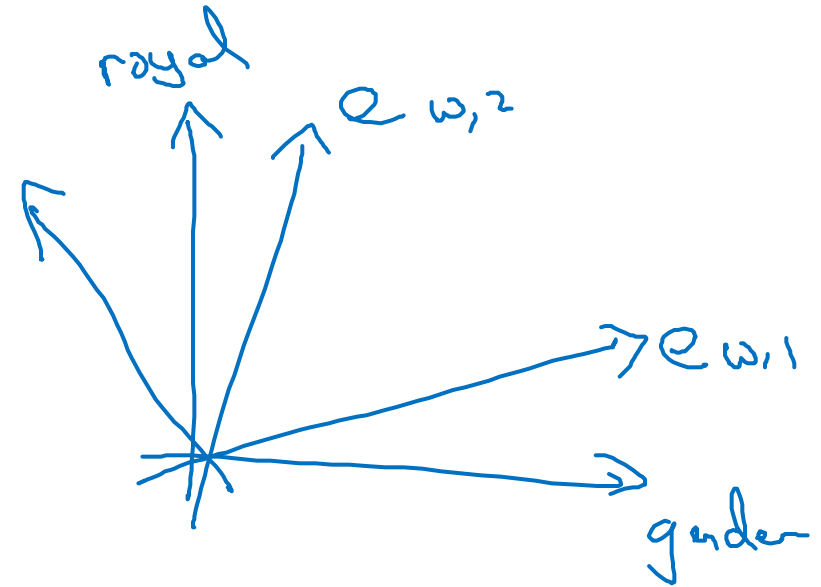
$f(X_{ij}) = 0$ at $X_{ij} = 0$.   "$0 \log 0$" $= 0$

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

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

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

Andrew Ng

Sentiment classification problem

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. ★☆☆☆☆

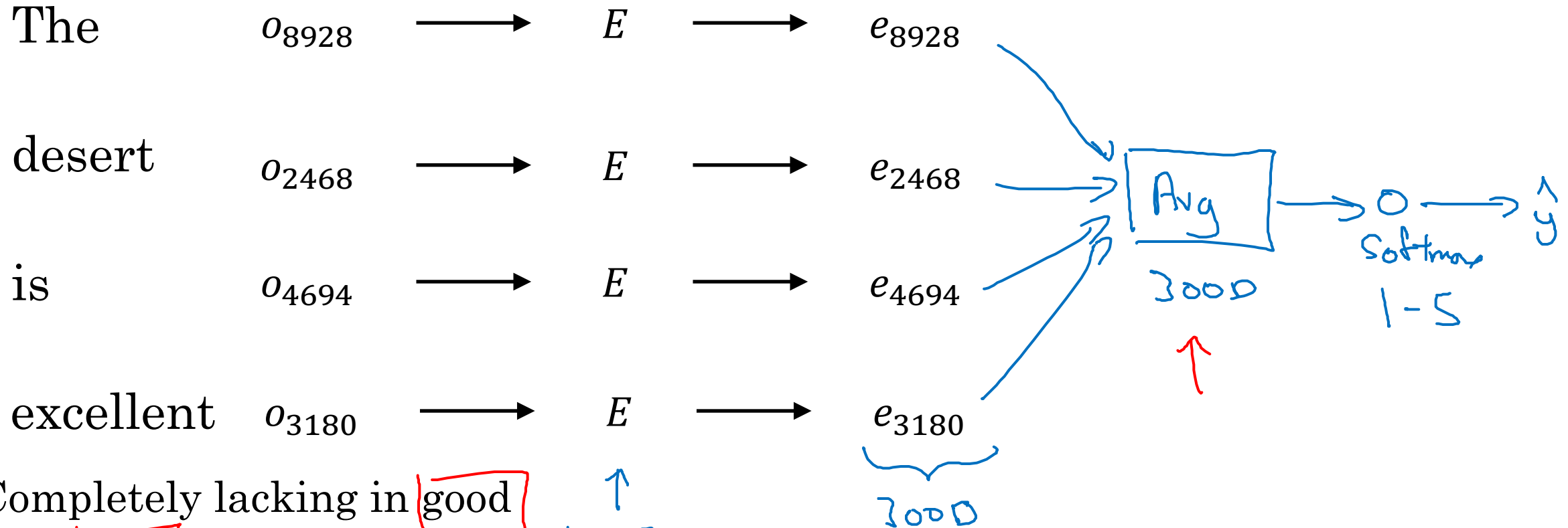10,000 → 100,000 words

Andrew Ng

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}

Avg → O → $\hat{y}$

Softmax

300D

1-5

300D

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

100 B words

Andrew Ng

RNN for sentiment classification

$\hat{y}$

softmax

$a^{<0>} \rightarrow a^{<1>} \rightarrow a^{<2>} \rightarrow a^{<3>} \rightarrow a^{<4>} \rightarrow \dots \rightarrow a^{<10>}$

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"
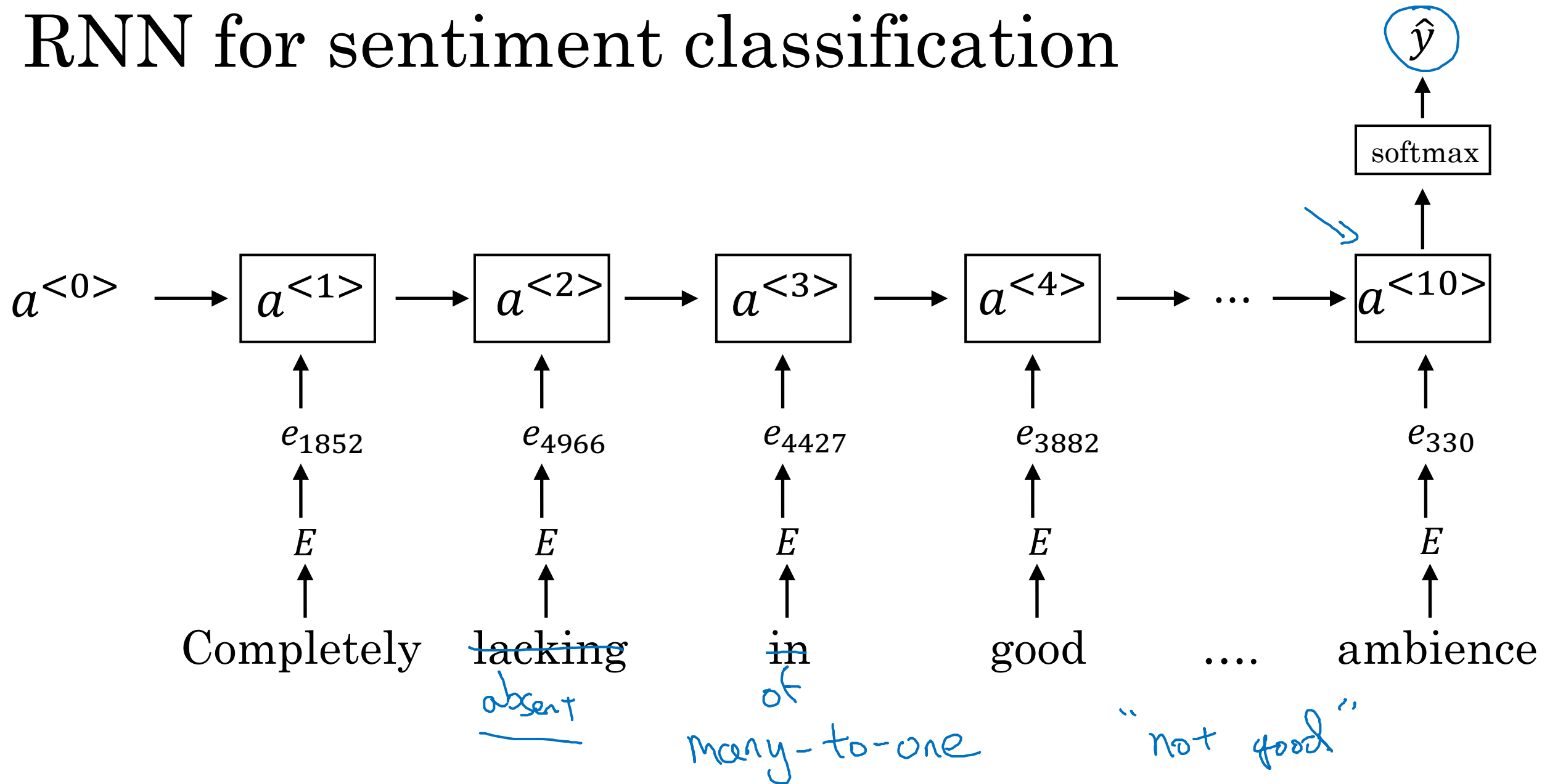
NLP and Word Embeddings

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 ✗

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]

Andrew Ng

Addressing bias in word embeddings
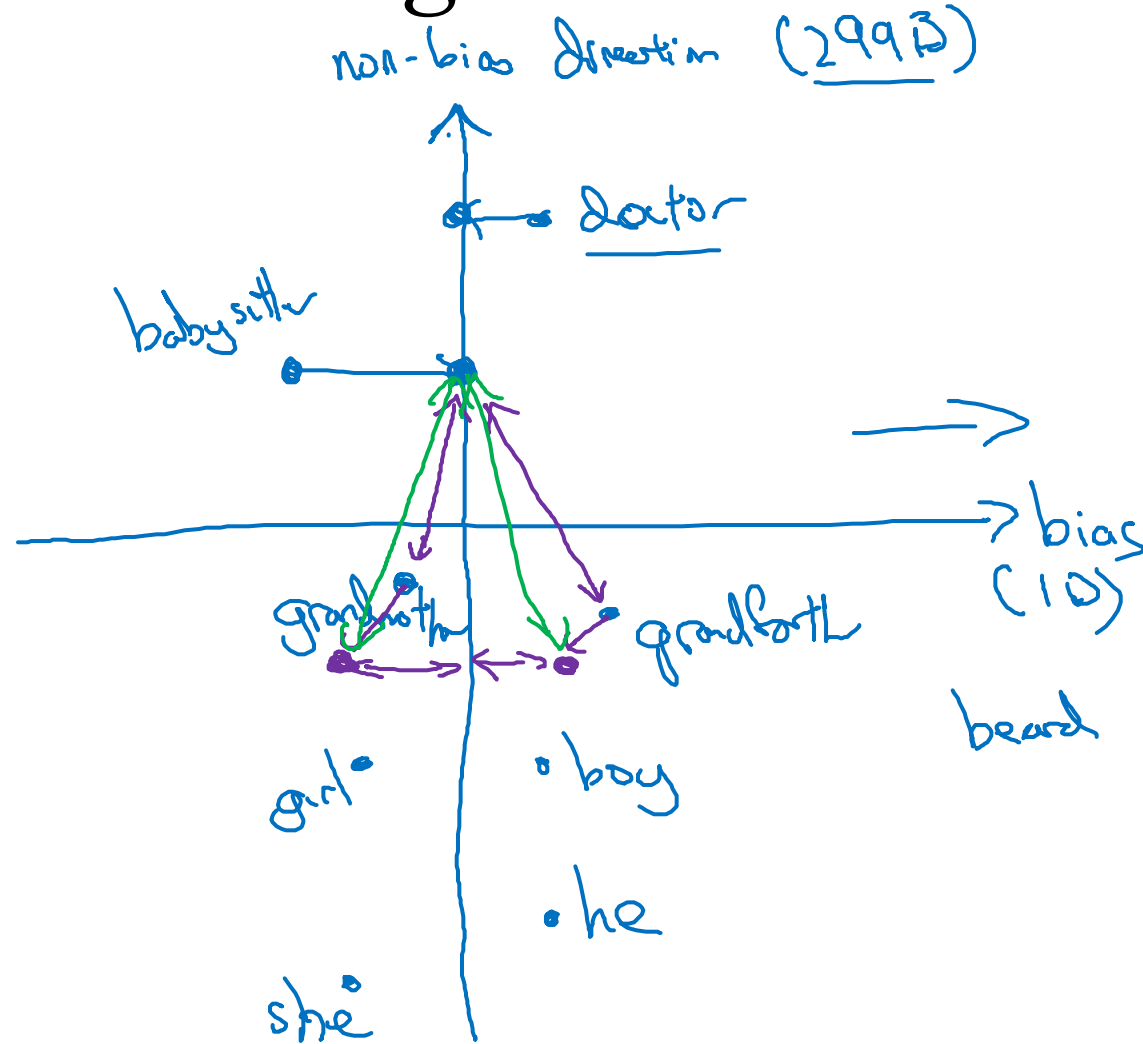
1. Identify bias direction.

$$\begin{cases} e_{he} - e_{she} \\ e_{male} - e_{female} \\ \vdots \end{cases} \rightarrow \text{average}$$

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

3. <u>Equalize pairs.</u>

$$\rightarrow \left\{ \begin{array}{ccc} \text{grandmoth} & - & \text{grandfath} \\ \text{girl} & & \text{boy} \end{array} \right\}$$

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

Andrew Ng