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 ?.

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
⋮ size cost alive verb	⋮	⋮				
	e_{5391}	e_{9853}				

300

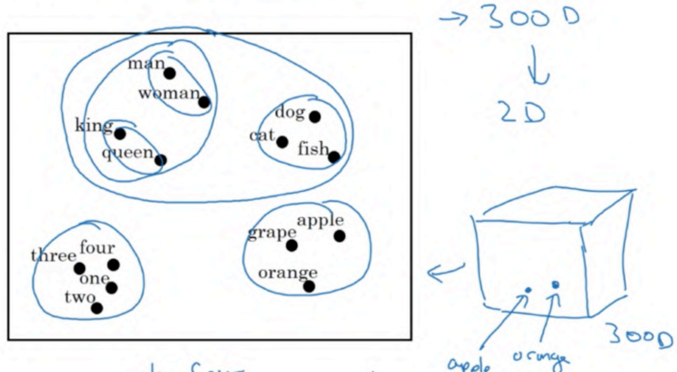
I want a glass of orange juice.

I want a glass of apple juice.

Andrew Ng

Visualizing word embeddings

→ 300 D ↓ 2D

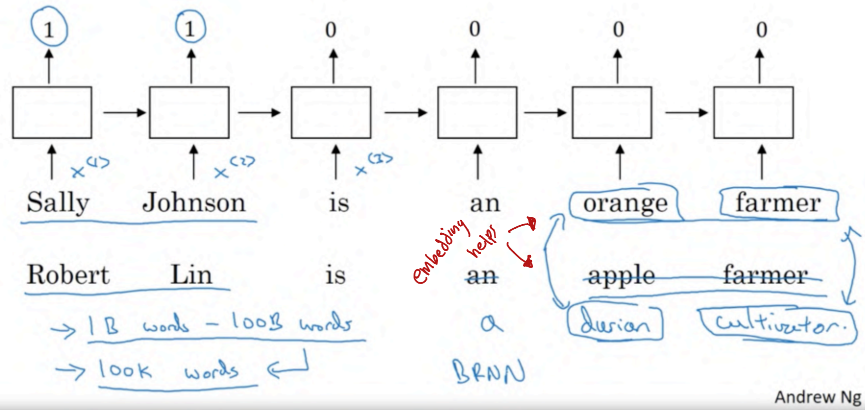


t-SNE

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

Andrew Ng

Named entity recognition example



Andrew Ng

Transfer learning and word embeddings

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

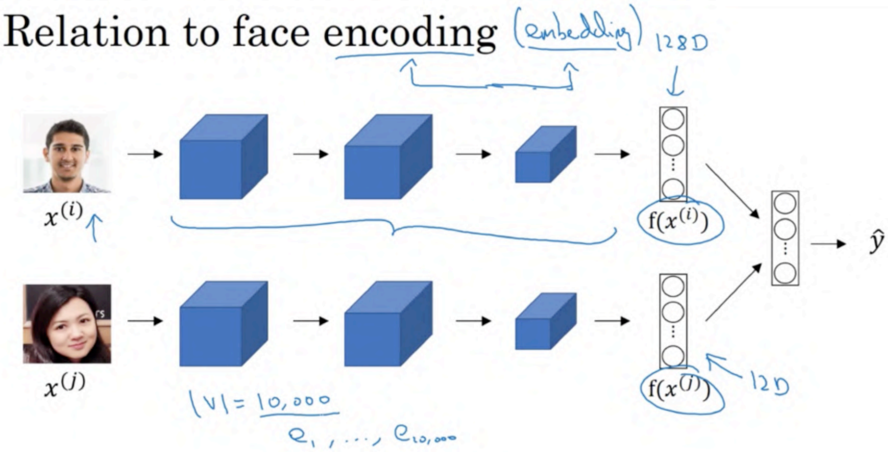
   (Or download pre-trained embedding online.)

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.

Andrew Ng

Relation to face encoding (embedding) 128D



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

Andrew Ng

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} $\qquad e_{woman}$

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

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

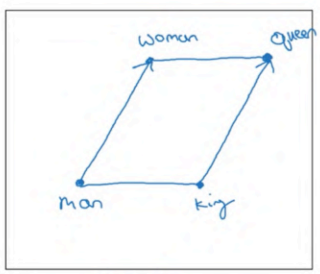
Man → Woman as King → ? Queen

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

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

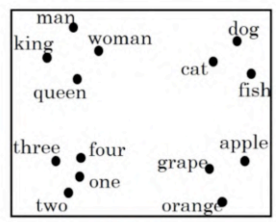
Andrew Ng

Analogies using word vectors



300 D

3000 → 2D



t-SNE

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

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

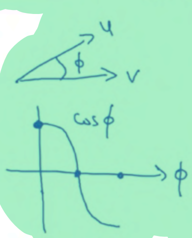
30 - 75%

Andrew Ng

Cosine similarity

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

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



$\|u - v\|^2$

Man:Woman as Boy:Girl

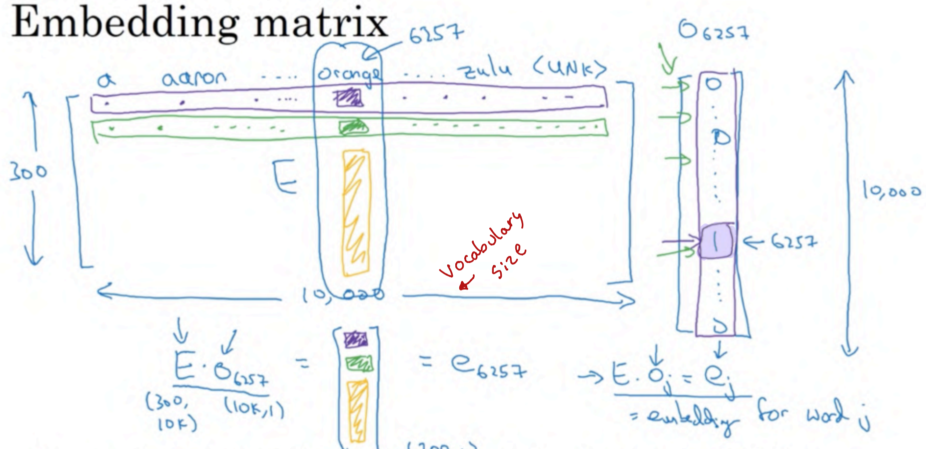
Ottawa:Canada as Nairobi:Kenya

Big:Bigger as Tall:Taller

Yen:Japan as Ruble:Russia

Andrew Ng

Embedding matrix



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

→ Embedding

Andrew Ng

Neural language model

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

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 a glass of orange ? to go along with
- Last 1 word orange ?
- Nearby 1 word glass ? skip gram

Andrew Ng

Skip-grams

randomly pick a word within a window

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

Context	Window	Target
orange	±1	juice
orange	±2	glass
orange	±6	my

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

Andrew Ng

Model

Vocab size = 10,000k

$x \longrightarrow y$

Content c ("orange") 6257 $\longrightarrow$ Target t ("juice") 4834

lots of parameters

one-hot vector

embedded vector for context word

$O_c \rightarrow E \rightarrow e_c \longrightarrow O \longrightarrow \hat{y}$ (softmax)

$e_c = E O_c$

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

target word given context word

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

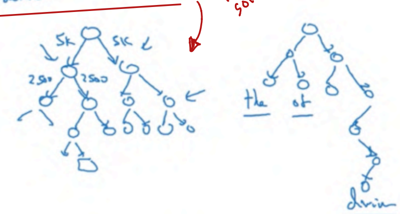
$$p(t|c) = \frac{e^{\theta_t^T e_c}}{\boxed{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}}$$

slow

Hierarchical softmax

possible solution

$\log |V|$



How to sample the context c?

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

$\rightarrow$ orange, apple, durian

e_{durian}

$P(c)$

t

$c \rightarrow t$

Andrew Ng

Defining a new learning problem

Negative Sampling

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

It's really to try to distinguish between these two types of distributions from which you might sample a pair of words.

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

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

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

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

Andrew Ng

Model

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

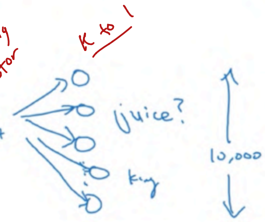
This technique is called negative sampling because what you're doing is, you have a positive example, the orange and then juice. And then you will go and deliberately generate a bunch of negative examples, negative samplings, hence, the name negative sampling, with which to train four more of these binary classifiers.
Play video starting at 8 minutes 54 seconds and follow transcript8:54
And on every iteration, you choose four different random negative words with which to train your algorithm on.

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

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

Orange 6257

$O_{6257} \rightarrow E \rightarrow e_{6257}$ embedding vector



10,000 binary classification problem

$k+1$

Andrew Ng

Selecting negative examples

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

the, of, and, ...

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

Andrew Ng

GloVe (global vectors for word representation)

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

c, t

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

$\underset{c\ \ t}{\uparrow\uparrow}$ $\quad i \uparrow t \quad j \uparrow c$

$X_{ij} = X_{ji}$

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

Andrew Ng

Model

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

0?

t c

"$\theta_t^T e_c$"

How related are they? The two words..

if $X_{ij}=0$, then log(0) is undefined so

weighting term

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

this, is, of, a, ...

durian

"0 log 0" = 0

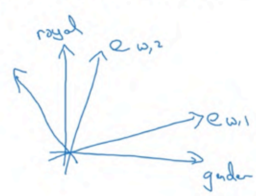
θ_i, e_j are symmetric

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

Andrew Ng

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



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

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

Andrew Ng

GloVe (global vectors for word representation)

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

c, t

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

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

Andrew Ng

2. Correction in "GloVe word vectors" slide 4 (8:11):

typo:

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

corrected:

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

Sentiment classification problem

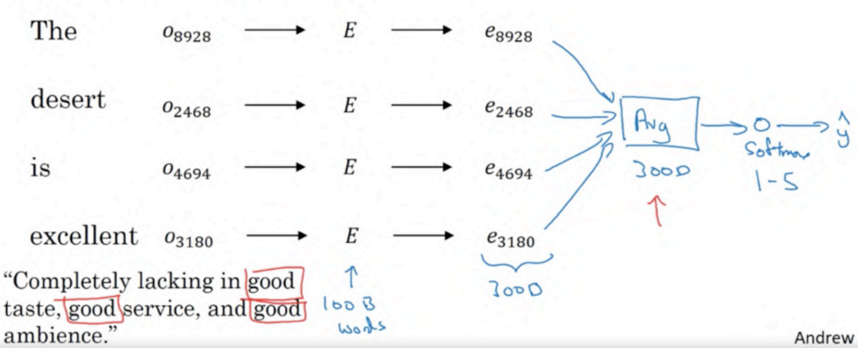
$x \longrightarrow y$

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.	★☆☆☆☆

10,000 → 100,000 words

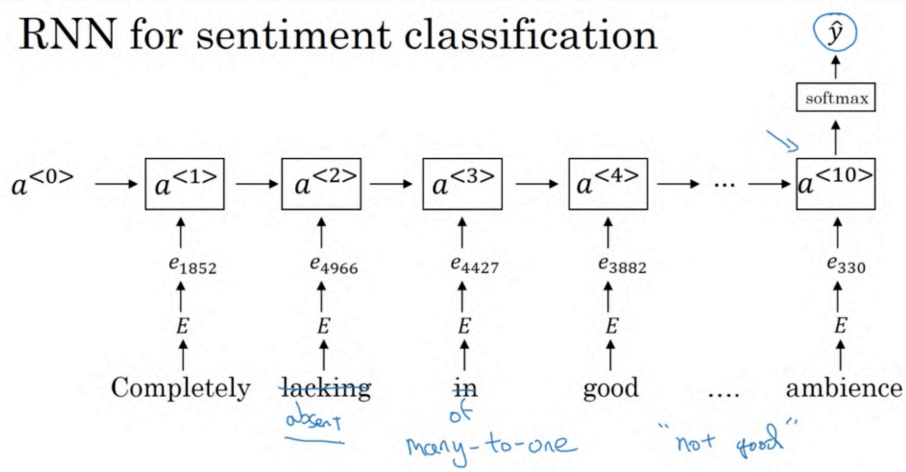
Simple sentiment classification model

The	dessert	is	excellent	★★★★☆
8928	2468	4694	3180	



"Completely lacking in $\boxed{\text{good}}$ taste, $\boxed{\text{good}}$ service, and $\boxed{\text{good}}$ ambience."

RNN for sentiment classification



The problem of bias in word embeddings

Man:Woman as King:Queen

Man:Computer_Programmer as Woman: Homemaker ✗

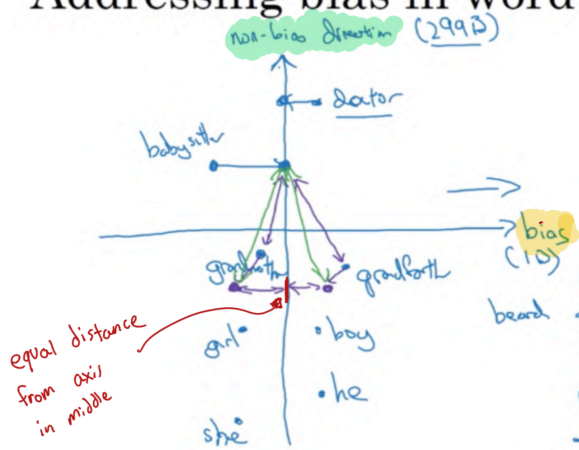
Father:Doctor as Mother: Nurse ✗

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.

   $e_{he} - e_{she}$

   $e_{male} - e_{female}$

   ⋮

   → average

2. Neutralize: For every word that is not definitional, project to get rid of bias. like Doctor or Babysitter (Not Grandma)

   has gender

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

Want only difference in their embedding to be gender

so both should be same exact similarity or exactly same difference