Deep learning

Dr. Aissa Boulmerka
a.boulmerka@centre-univ-mila.dz

2023-2024

CHAPTER 9
NLP AND WORD EMBEDDINGS

Word representation

V = [a, aaron, …, zulu, <UNK>]　　　$|V| = 10000$

1-hot representation

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

I want a glass of orange __juice__.

I want a glass of apple ___?___.

Featurized representation: word embedding

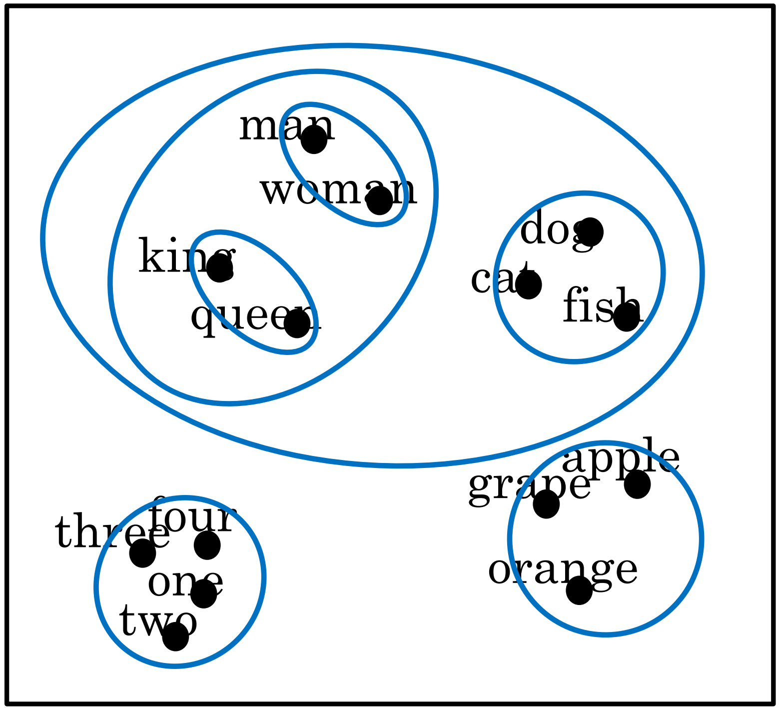
	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97
⋮						
Size						
Cost						

300

I want a glass of orange __juice__.

I want a glass of apple__juice__.

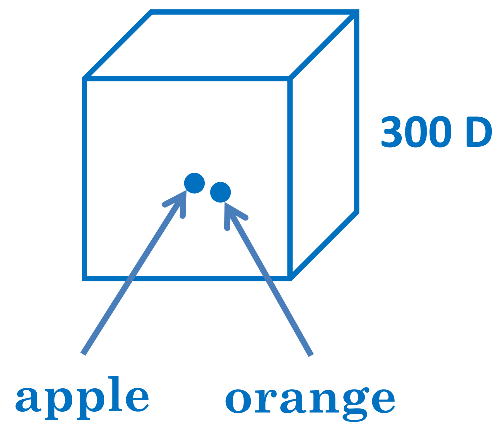
Visualizing word embeddings



t-SNE

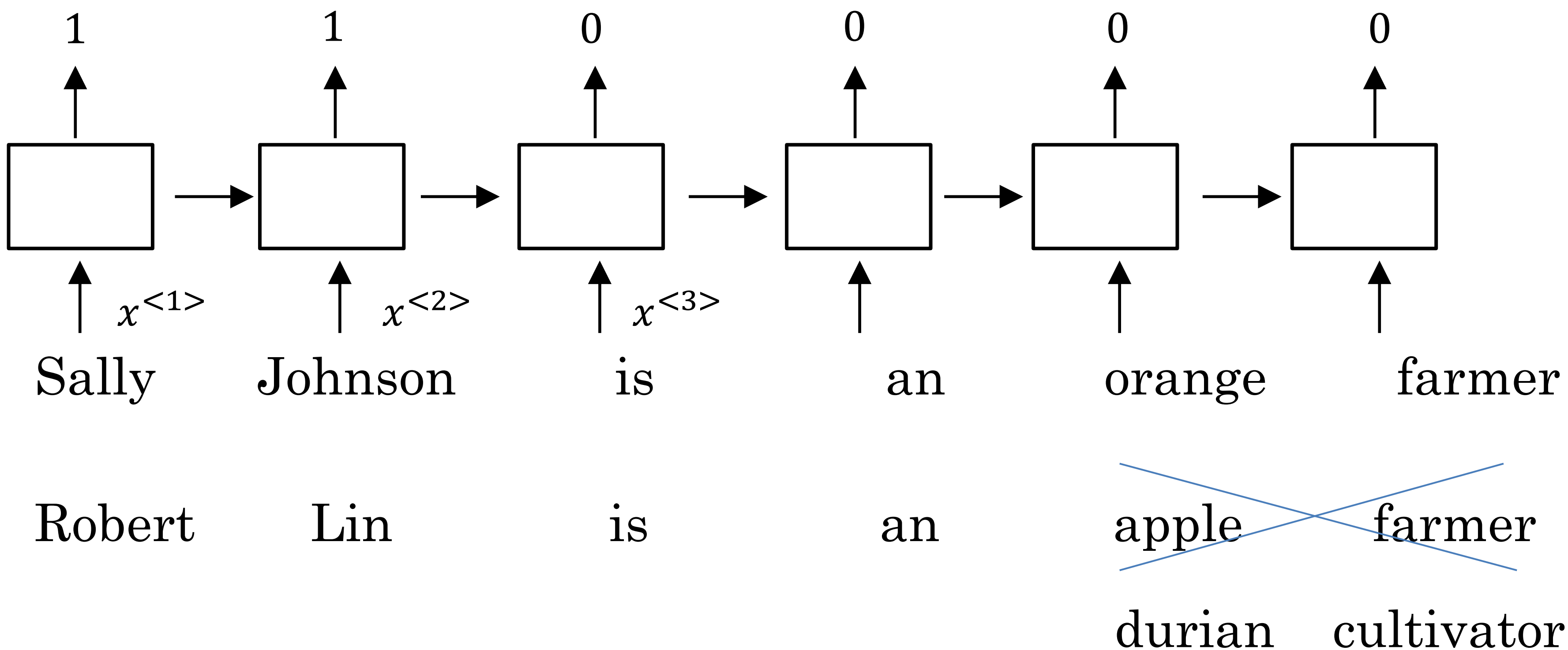
300 D

2 D



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

Named entity recognition example



BRNN

Transfer learning and word embeddings

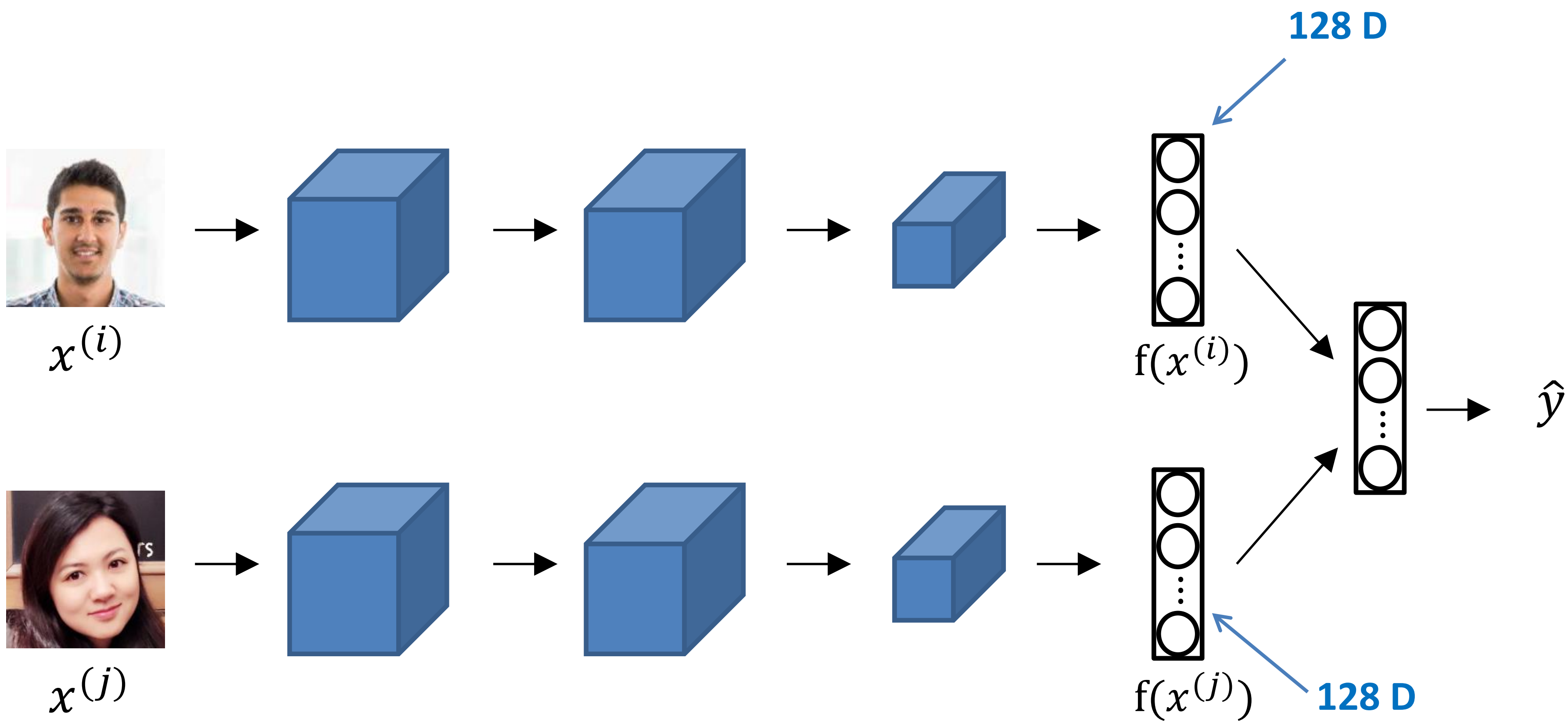
1. Learn word embeddings from large text corpus. (1-100B words)

   (Or download pre-trained embedding online.)

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

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

Relation to face embedding



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

Analogies

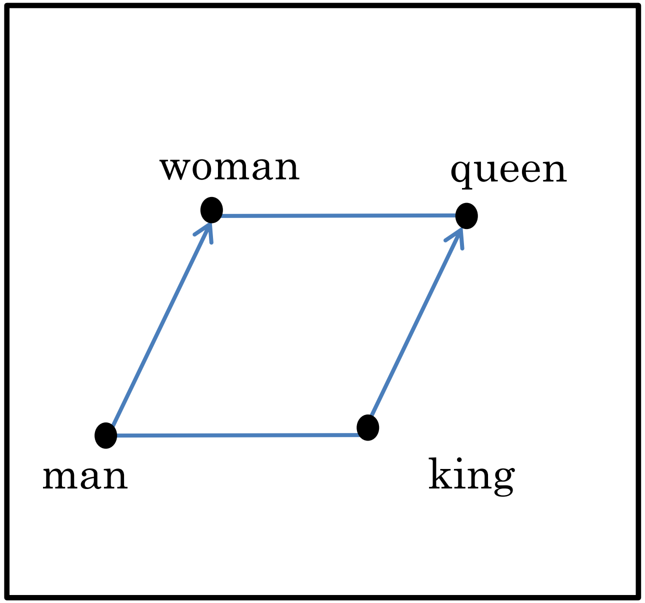
	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97

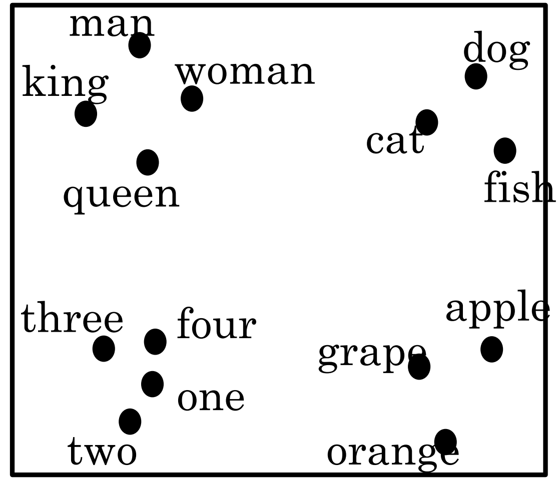
$$e_{man} - e_{woman} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad e_{king} - e_{qeen} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{e_{man} - e_{woman} \approx e_{king} - e_{qeen}}$$

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

Analogies using word vectors





t-SNE

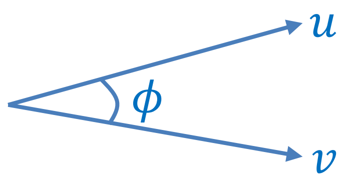
$$e_{man} - e_{woman} \approx e_{king} - e_{w} \quad ?$$

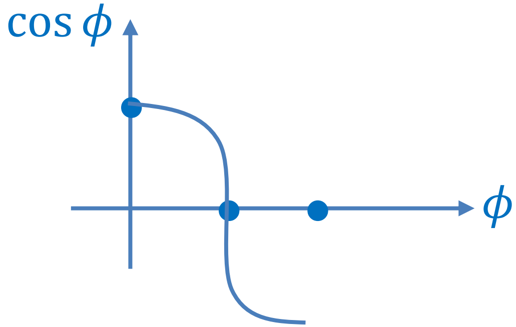
$$\boldsymbol{Find\ word\ w:\ \underset{w}{argmax}\ sim(e_w, e_{king} - e_{man} + e_{woman})}$$

Cosine similarity

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

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





Man:Woman as Boy:Girl

Ottawa:Canada as Nairobi:Kenya

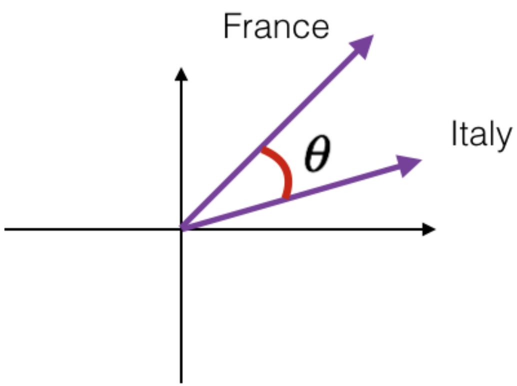
Big:Bigger as Tall:Taller

Yen:Japan as Ruble: Russia

Cosine similarity

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

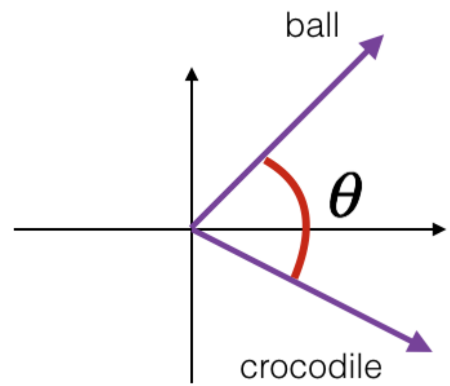
$$sim(u, v) = \frac{u^T v}{\|u\|_2 \|v\|_2}$$



France and Italy are quite similar

θ is close to 0°

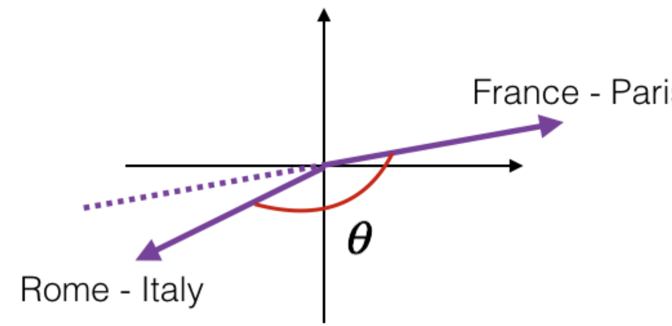
$\cos(\theta) \approx 1$



ball and crocodile are not similar

θ is close to 90°

$\cos(\theta) \approx 0$

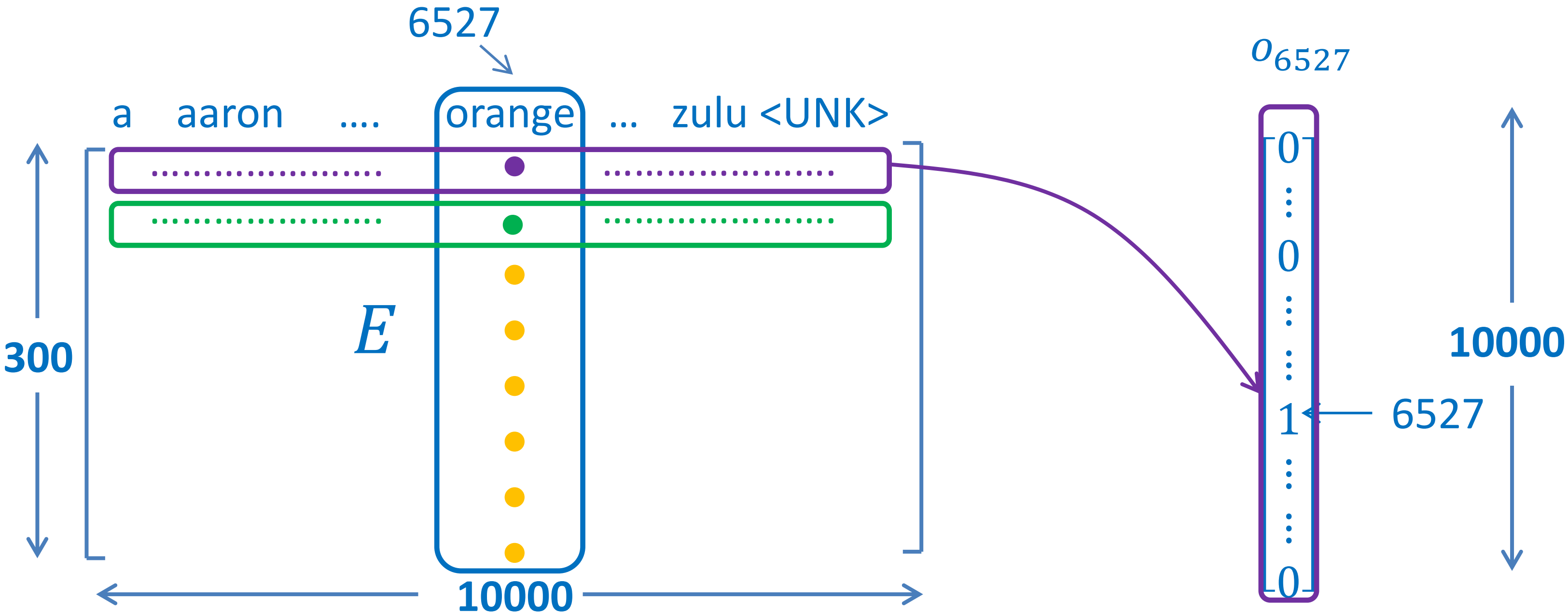


the two vectors are similar but opposite
the first one encodes (city - country)
while the second one encodes (country - city)

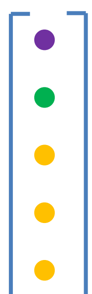
θ is close to 180°

$\cos(\theta) \approx -1$

Embedding matrix



$$\underset{(300,\ 10K)}{E} \, . \, \underset{(10K,1)}{o_{6257}} =$$



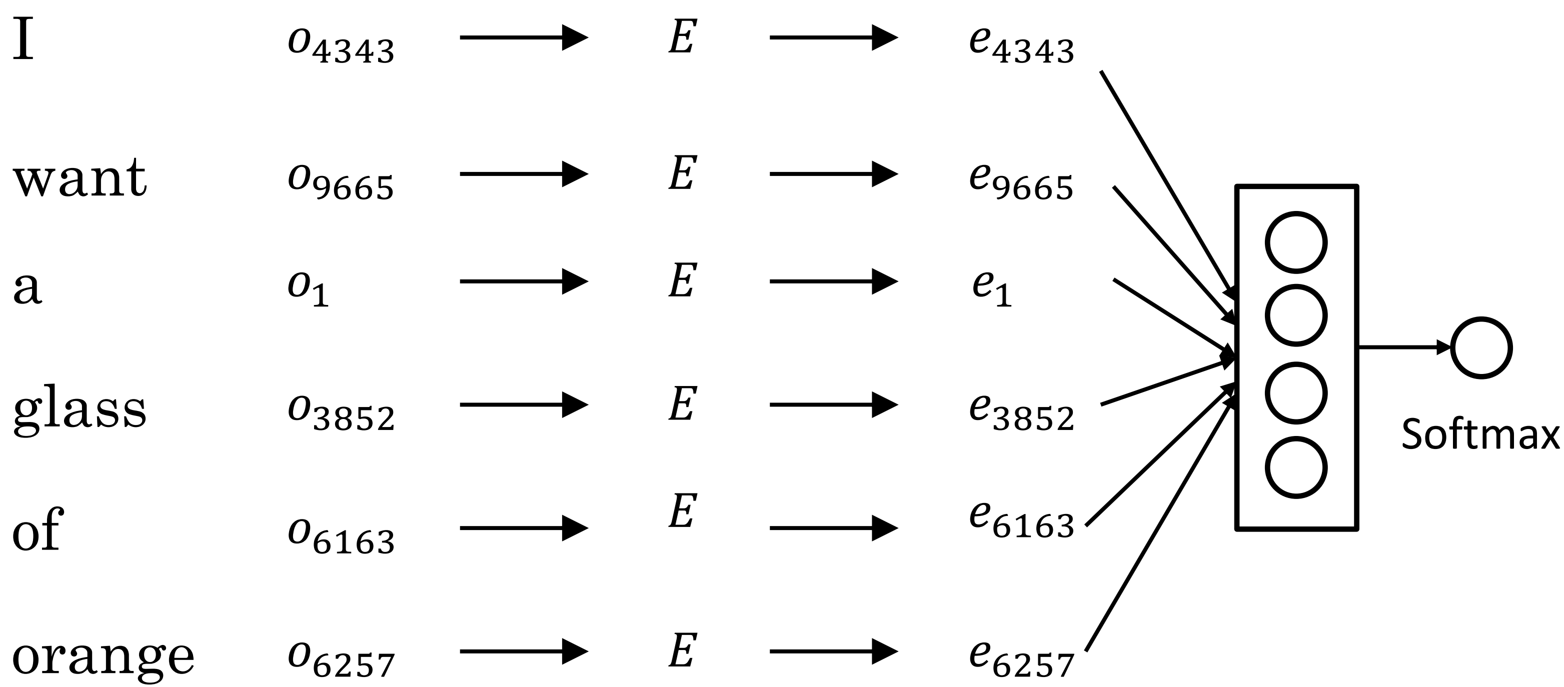
$$= \underset{(300,\ 1)}{e_{6257}} \Longrightarrow E\,.\,o_j = e_j$$ = **embedding for word j**

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

Neural language model

I want a glass of orange ______.

4343 9665 1 3852 6163 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: Last 4 words.

4 words on left & right

Last 1 word

Nearby 1 word

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 = 10k

$$x \longrightarrow y$$

Context c («orange») 6257   Target t («juice») 4834

$$o_c \longrightarrow E \longrightarrow e_c \longrightarrow o \longrightarrow \hat{y}$$

$e_c = E o_c$ Softmax

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

θ_t: parameter associated with the output t

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{10000} y_i \log \hat{y}_i$$

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

Problems with softmax classification

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

Very slow to compute !!!!

How to sample the context c?

NEGATIVE SAMPLING

Defining a new learning problem

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

$$\underbrace{\text{Context} \quad \text{Word}}_{x} \longrightarrow \underbrace{\text{Target?}}_{y}$$

Context	Word	Target?
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

k

$k = 5 - 20$: *smaller dataset*

$k = 2 - 5$: *larger dataset*

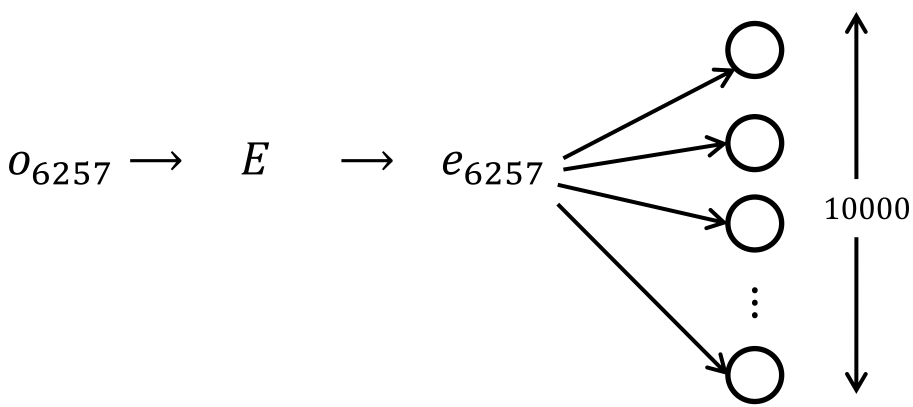
[Mikolov et. al., 2013. Distributed representation of words and phrases and their compositionality]

Model

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

Sotmax: 10000 classes

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



10000 binary classification problem

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

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

Selecting negative examples

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

$$p(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=1}^{10000} f(w_j)^{3/4}}$$

GLOVE WORD VECTORS

GloVe (global vectors for word representation)

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

c, t

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

$X_{ij} = X_{ji}$

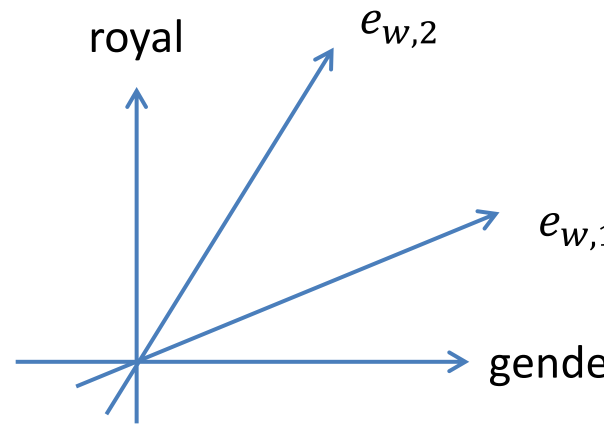
[Pennington et. al., 2014. GloVe: Global vectors for word representation]

Model

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

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



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

SENTIMENT CLASSIFICATION

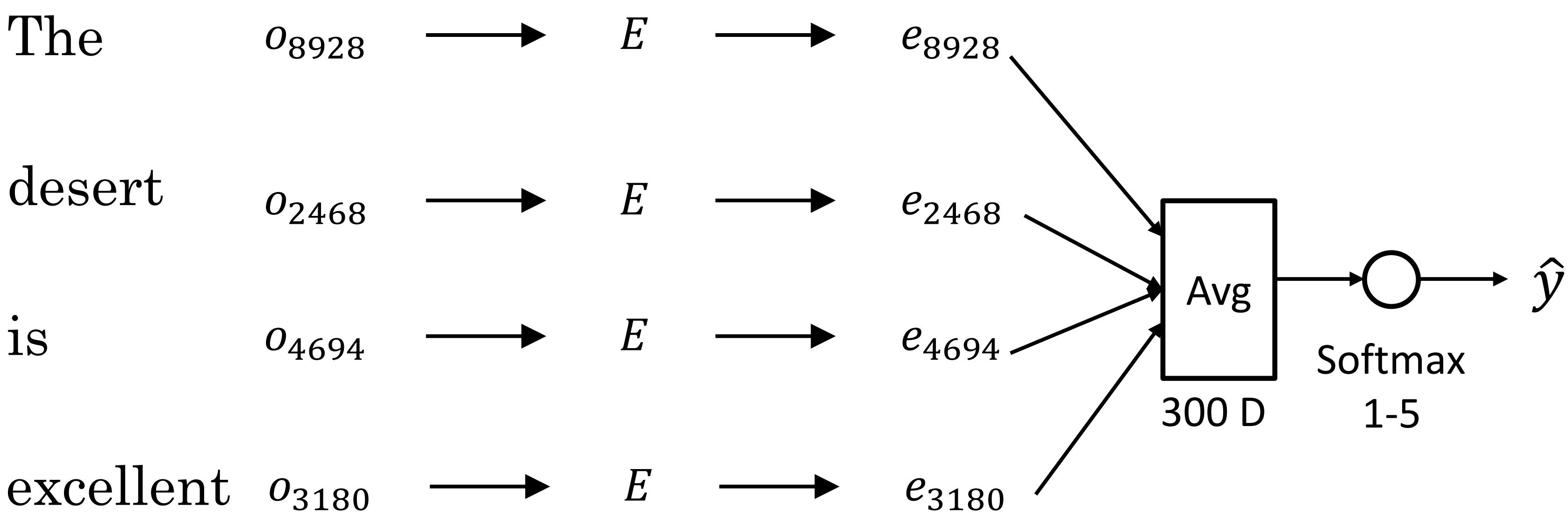
Sentiment classification problem

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.	★☆☆☆☆

Simple sentiment classification model

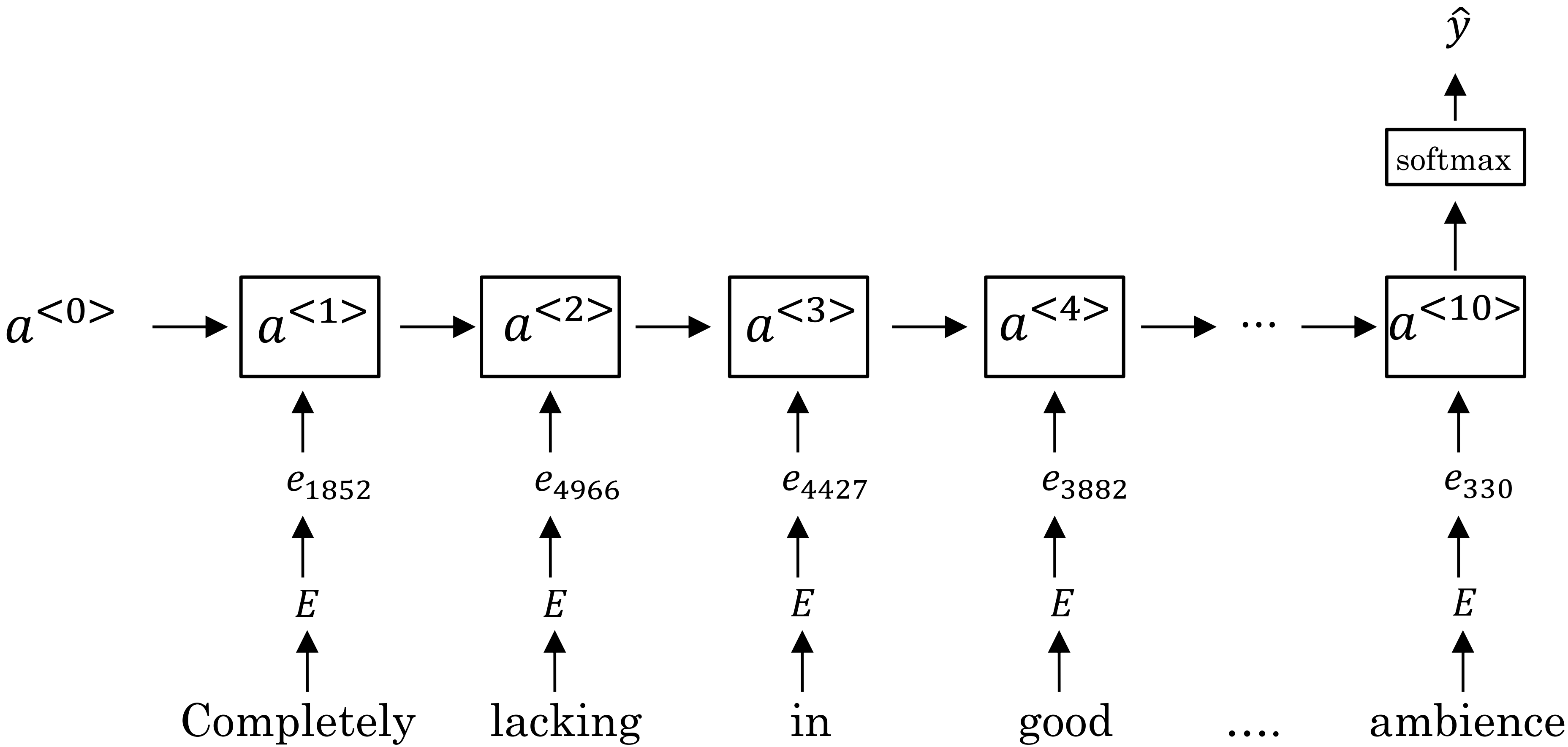
The dessert is excellent

8928 2468 4694 3180



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

RNN for sentiment classification



Many-to-one

DEBIASING WORD EMBEDDINGS

The problem of bias in word embeddings

Man:Woman as King:Queen

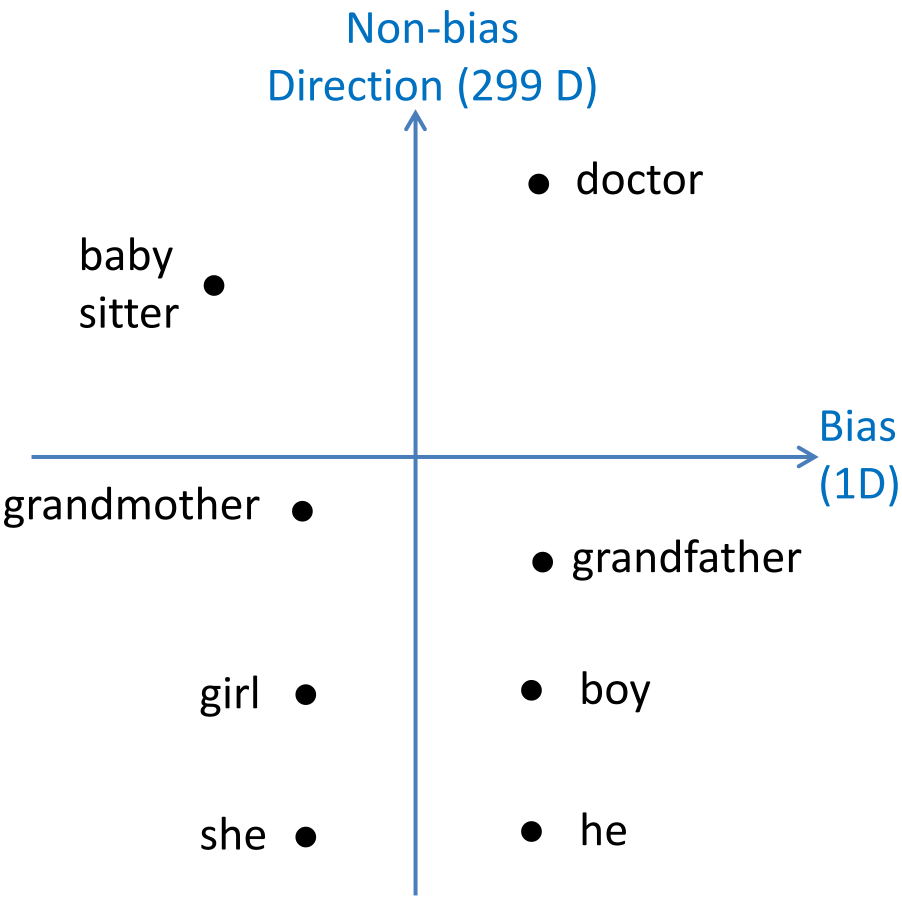
Man:Computer_Programmer as Woman: Homemaker

Father:Doctor as Mother: Nurse

Word embeddings can reflect gender, ethnicity, age, 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.

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

References

- Andrew Ng. Deep learning. Coursera.
- Geoffrey Hinton. Neural Networks for Machine Learning.
- Kevin P. Murphy. Probabilistic Machine Learning An Introduction. MIT Press, 2022.
- MIT Deep Learning 6.S191 (http://introtodeeplearning.com/)