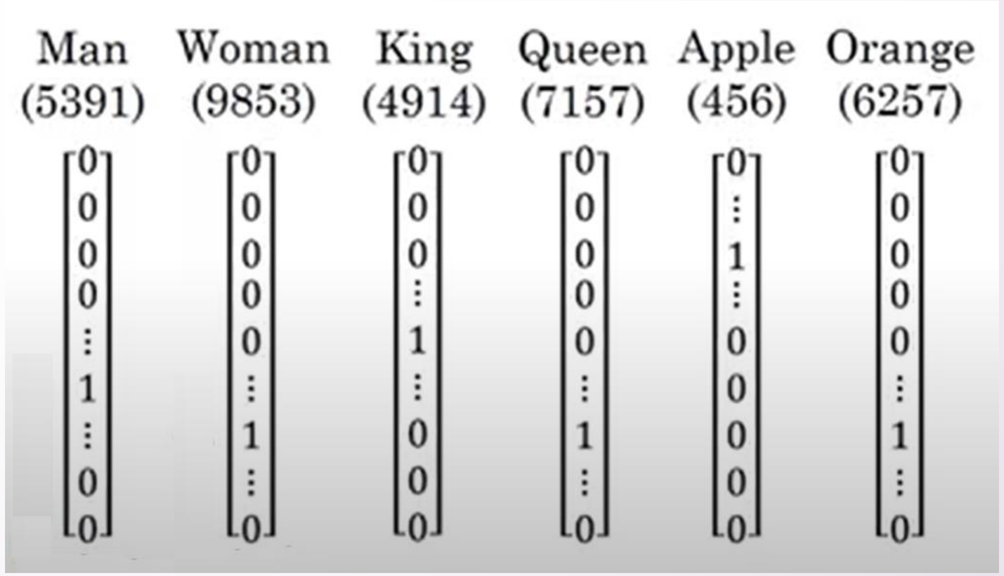
Word2vec and Negative Sampling

Ali Vefghi

Supervisor: Dr. Rahmati

word representation

- 1-hot representation : O(“Man”)=[0, 0, … , 1, 0, … ,0]
- Dict = [a, aaron, …., zulu, <UNK>]
- | Dict | = 10000
- | O(“Man”) | = 10000



word representation

- Weaknesses:
- Can’t generalize across words. For example cant get the relation between apple and orange.

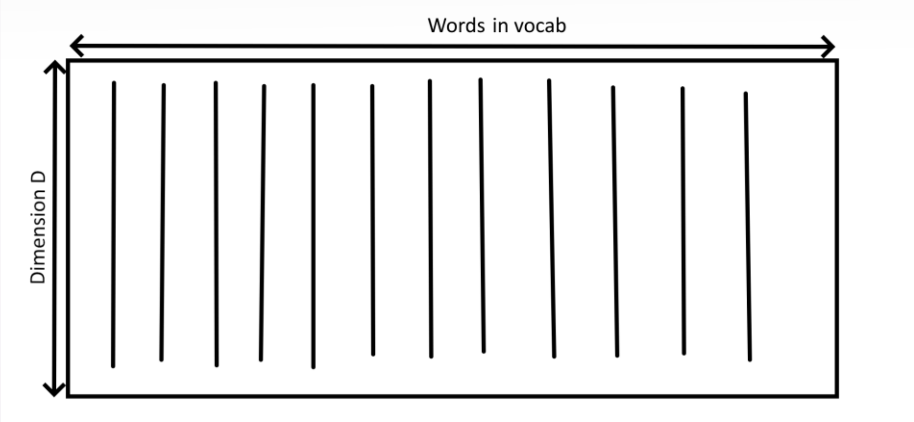
I want a glass of orange ------------. (juice)

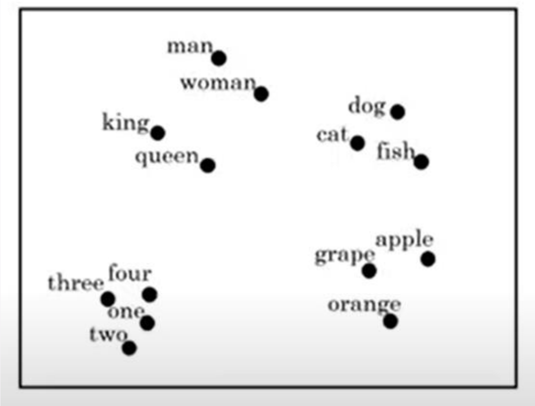
I want a glass of apple ------------. (juice)

- Too long, sparse and inefficient

Featurized representation: word embedding

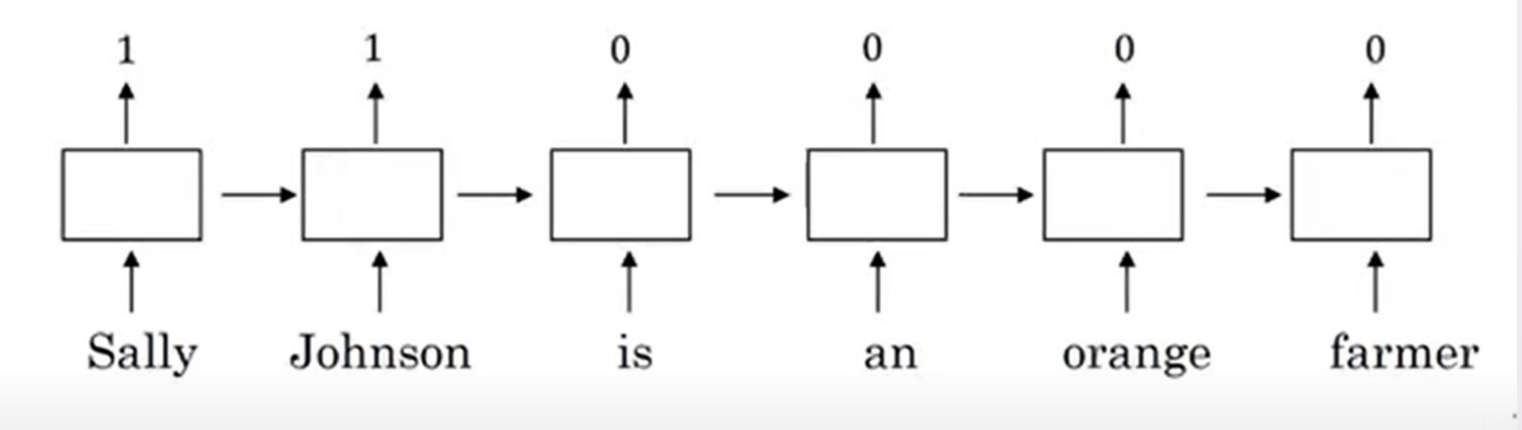
	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97





Named entity recognition example

- Extracting the names
- sally johnson is a person from orange farmer ?



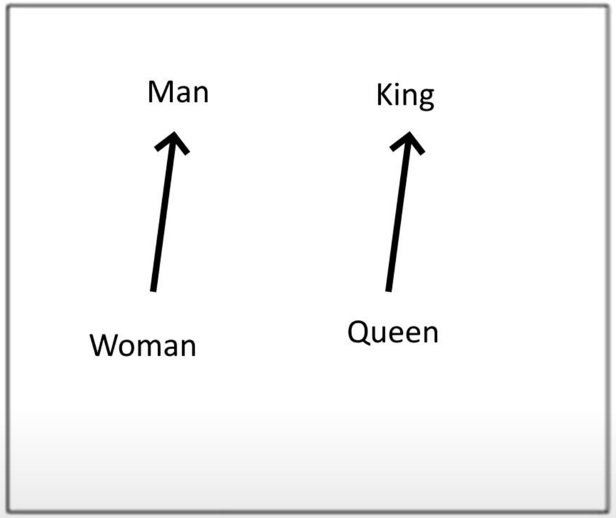
Transfer learning and word embedding

- 1. Learn word embeddings from a 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

Cos similarity



Man:Woman as Boy:Girl

Ottawa:Canada as Nairobi:Kenya

Big:Bigger as Tall:Taller

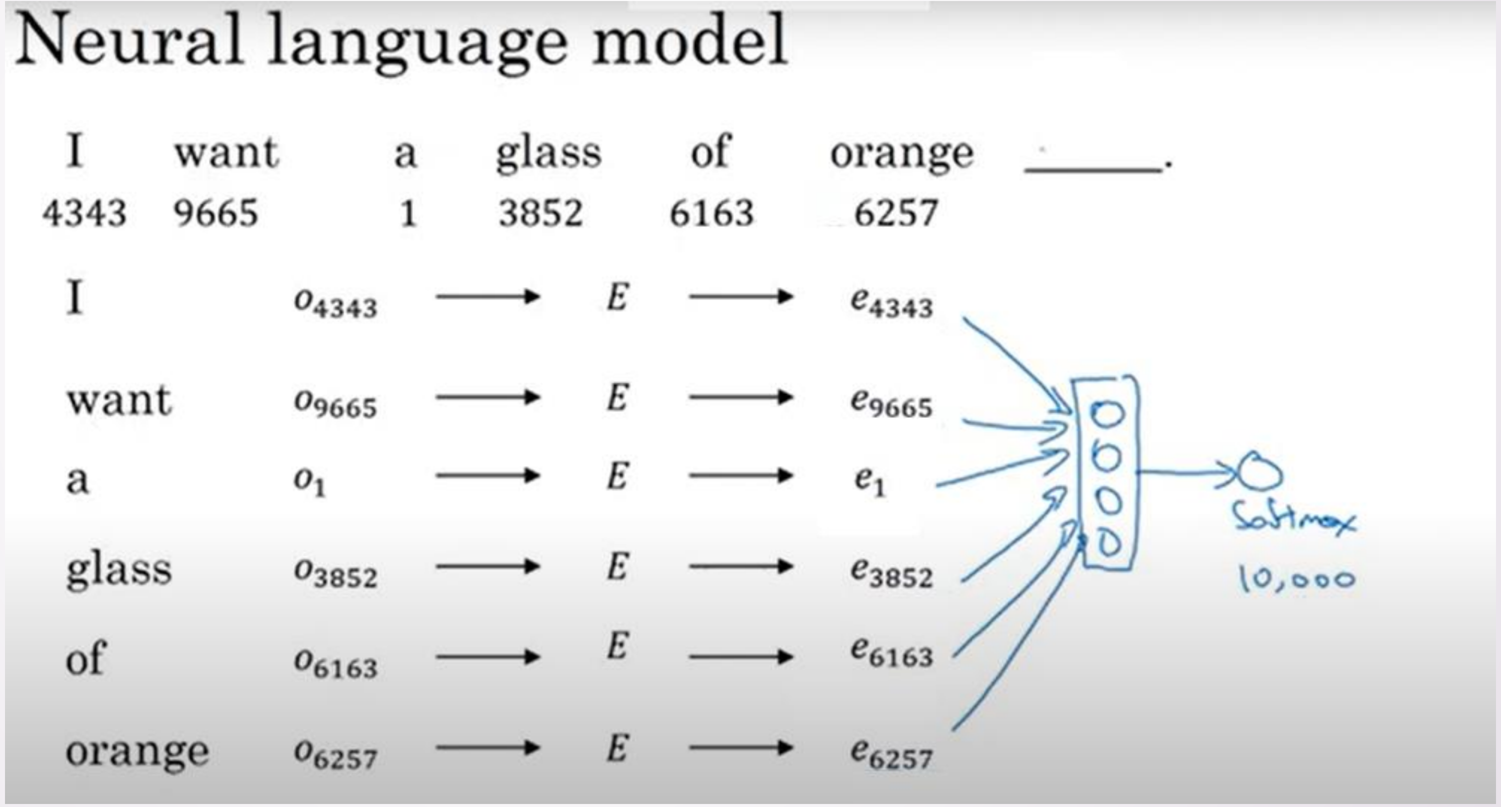
Yen:Japan as Ruble:Russia

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

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

- Need distance to learn with models

Simple neural network Abstract



Context/target pairs

- I want a glass of orange (juice:target) to go along with my cereal.

Context = last 4 words : a glass of orange ----

Or

4 words on left and right : a glass of orange ---- to go along with

Or

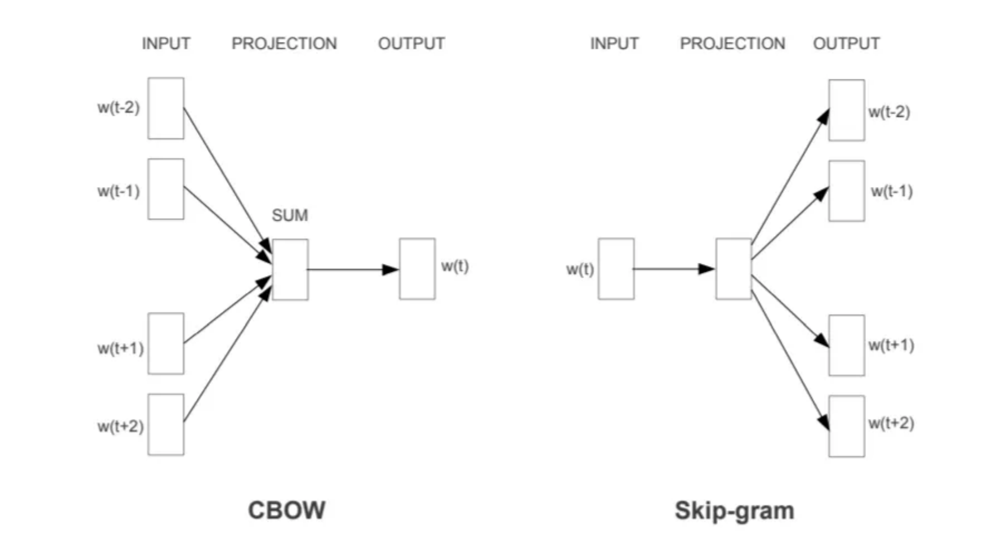
Last 1 word : orange ----

Or

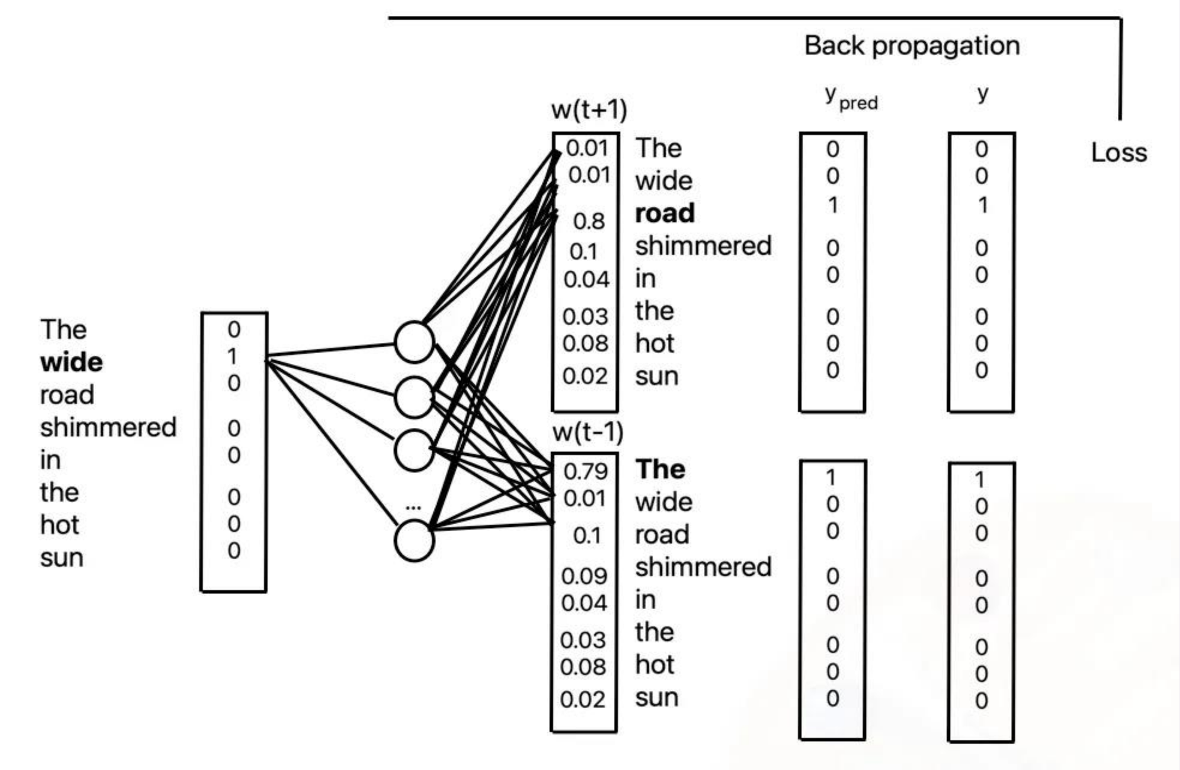
Nearby 1 word (skip-gram) : glass ? ---- ? ?

Word2vec

- **Main idea:** The words that appear near each other should have similar word vectors.
- **Skip-gram**: works well with a small amount of the training data, represents well even rare words or phrases.
- **CBOW**: several times faster to train than the skip-gram, slightly better accuracy for the frequent words.



Skip-grams



Back propagation
y_{pred}
y
Loss
w(t+1)
0.01 The
0.01 wide
0.8 road
0.1 shimmered
0.04 in
0.03 the
0.08 hot
0.02 sun
w(t-1)
0.79 The
0.01 wide
0.1 road
0.09 shimmered
0.04 in
0.03 the
0.08 hot
0.02 sun
The
wide
road
shimmered
in
the
hot
sun

Model

- Vocab size = 10000K
- Mapping from context c to a target t

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} \log\, p(w_{t+j}|w_t)$$

- e(c) -> softmax -> y^

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log p(c|w) = \sum_{(w,c)\in D} (\log e^{v_c \cdot v_w} - \log \sum_{c'} e^{v_{c'} \cdot v_w})$$

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

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

- Tetha(t) parameter associated with output t controlling the distribution

Problems

- For every p(t|c) we calculate the sum in the denomitaor

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

- Hierarchical softmax

Negative sampling

- Maximizing the similarity of the words in the same context
- Minimizing it when they occur in different contexts

- We do not need to update the entire output weight matrix

Defining a new learning problem

- I want a glass of orange juice to go along with my cereal
- Second word is random from dictionary
- Of is positive but it is ok to get it negative

- How to choose k (number of neg samples)
- K = 5-20 small datasets
- K = 2-5 large datasets

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

Model

- 1 giant 10000 way softmax -> 10000 binary classification problem

If we let $\sigma(x) = \frac{1}{1+e^{-x}}$ we get:

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log \frac{1}{1+e^{-v_c \cdot v_w}} + \sum_{(w,c)\in D'} \log(\frac{1}{1+e^{v_c \cdot v_w}})$$

$$= \arg\max_{\theta} \sum_{(w,c)\in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c)\in D'} \log \sigma(-v_c \cdot v_w)$$

- How do you choose negative examples?
- Heuristic:

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

For more information

word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method

Yoav Goldberg and Omer Levy
{yoav.goldberg,omerlevy}@gmail.com

February 14, 2014

Thank you