1 - hot representation

t - SNE (300 D → 2D) for visualizing high-D embedding

non-linear dimensionality reduction

Why word embedding?

for example, a model have not seen durian, but have seen other fruit such as orange, embedding can draw similarity. This enable process large unlabelled datasets.

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) → 10,000 → 300

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

Some similarity with Face Recognition (encoding)

Properties: for example $e_{man} - e_{woman} \approx e_{king} - e_?$

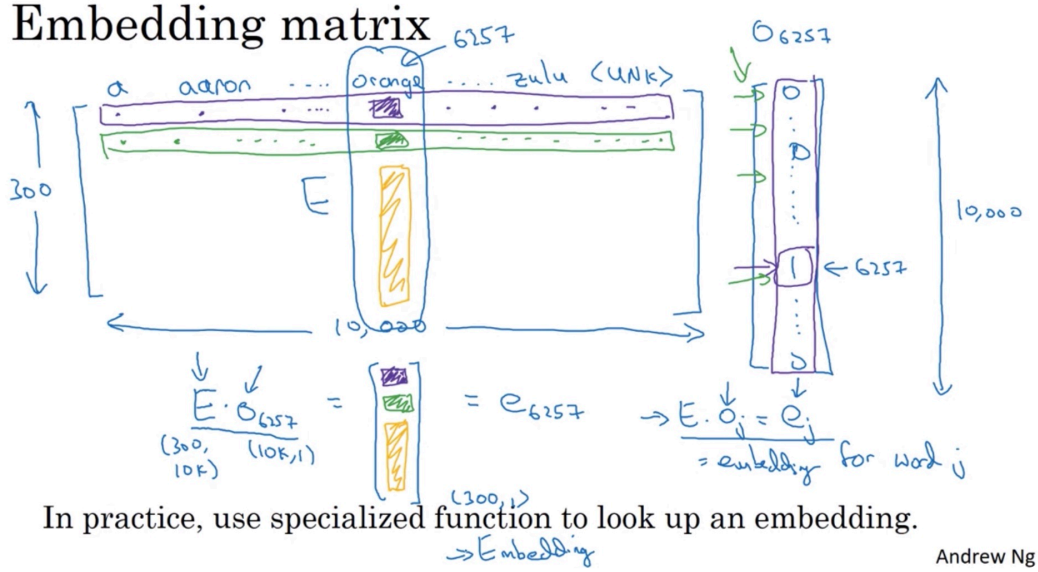
The vector difference is similar for the 2 pair of words

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

cosine similarity

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

Embedding Matrix

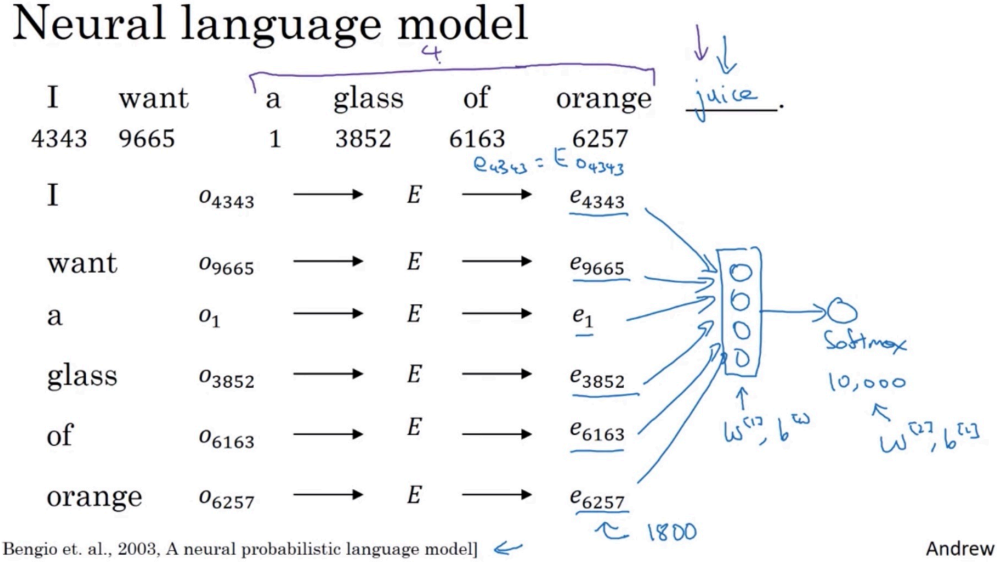


In practice, use specialized function to look up an embedding.

Andrew Ng

Because matrix multiplication with sparse matrix inefficient

Neural language model



Bengio et. al., 2003, A neural probabilistic language model] ←

Andrew

For BRNN, using nearby one word instead of 4 or n-words is surprising effective as well

Word2Vec

skip gram : Given context word, predict random forward / backward target word

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

Context	Target
orange	juice
orange	glass
orange	my

Model

Vocab size = 10,000k

Context c ("orange") 6257 → Target t ("juice") 4834

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

one-hot mapping $e_c = E O_c$

Softmax: $p(t|c) = \frac{e^{\theta_t^T e_c}}{\left[\sum_{j=1}^{10,000} e^{\theta_j^T e_c}\right]}$ θ_t = parameter associated with output t

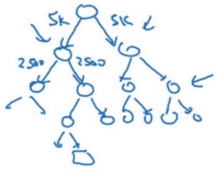
extensive, slow

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

$\hat{y}$ = 10,000-dim vector, 4834th entry = 1

→ solution: hierarchical softmax

common words placed shallower and rare words are deeper



$\log|V|$ usually not perfectly balanced.

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	0
orange	book	0
orange	the	0
orange	of	0

k = 5-20 smaller dataset

k = 2-5 larger dataset

Softmax → binary classification.

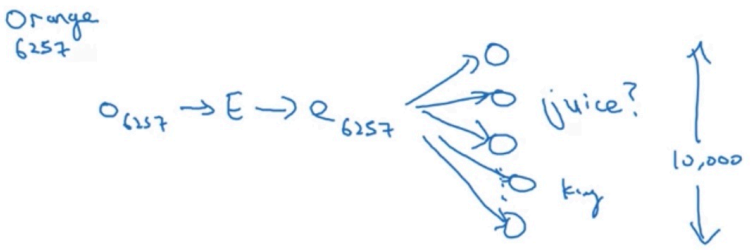
Softmax $p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$

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

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

And

Negative Sampling - continued



Instead of training 10,000 examples, train on 1 positive example and 4 randomly chosen negative example.

10,000 softmax ⇒ 10,000 binary classification ⇒ k+1 binary

Therefore this is a lot cheaper to train.

How to Sample?

① $P(w_i)$ → getting a lot of "the, of, and, ..."
↳ Sample According to Empirical frequency. (How often

② $\frac{1}{|V|}$
↳ uniform random distribution

③ The best. combine ① and ②.

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

GloVe → global vectors for word representation

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

$$\min \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(X_{ij})(\theta_i^T e_j - \log X_{ij})^2$$

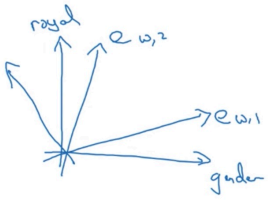
weighting term

θ_i, e_j are symetric $\frac{e_w + \theta_w}{2}$

$f(X_{ij}) = 0$ if $X_{ij} = 0$ ⇒ Stop words (the, and, is)

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	←



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

$[(A\theta_i)^T (A^{-T} e_j) = \theta_i^T A^T A^{-T} e_j]$ Andrew Ng

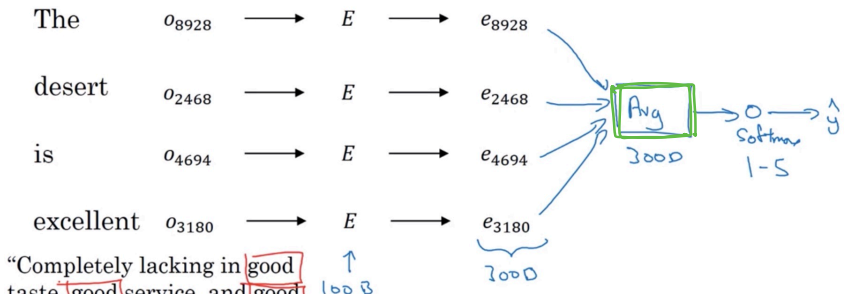
why sometimes the component is not human interpretable

Sentiment classification (lack of labelled training data)

Simple sentiment classification model

The dessert is excellent ★★★★☆

8928 2468 4694 3180



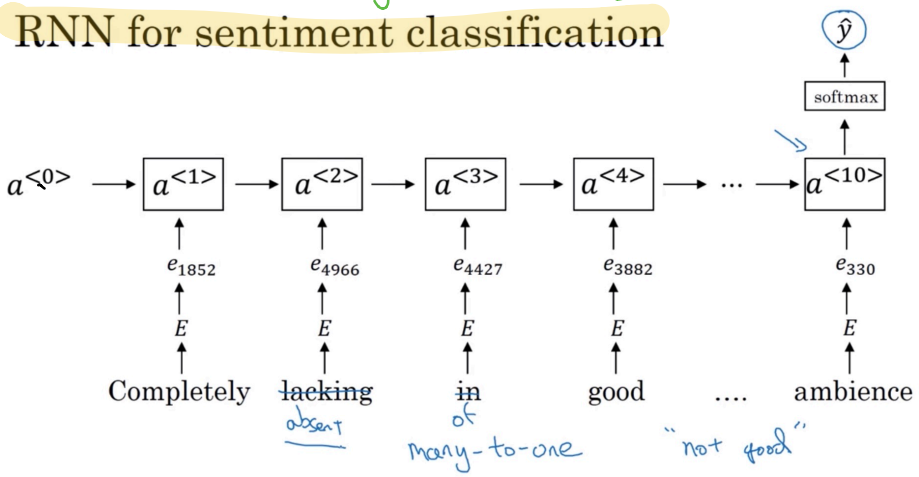
"Completely lacking in good taste, good service, and good ambience."

100B words

→ average skewed by # of times "good" are mentioned

Andrew Ng

RNN for sentiment classification



Debiasing Word Embeddings

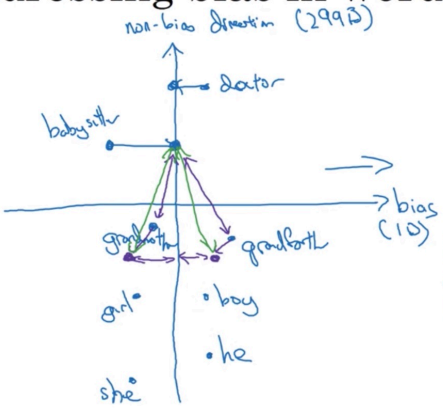
Eg. Man:Woman as King:Queen

Man:Computer_Programmer as Woman:Homemaker ✗

Father:Doctor as Mother:Nurse ✗ Gender Stereotype

Word embeddings can reflect gender, ethnicity, age, sexual orientation, and other biases of the text used to train the model.

Addressing bias in word embeddings



1. Identify bias direction.
   - $e_{he} - e_{she}$
   - $e_{male} - e_{female}$
   - ...
   - → average — or SVD
2. Neutralize: For every word that is not definitional, project to get rid of bias. — Gender-specific. Babysitter, doctor not definitional. reduce one component.
3. Equalize pairs. → grandmother (girl) — grandfather (boy) — Bin classifier / Hard pick

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

1. **Suppose you learn a word embedding for a vocabulary of 10000 words. Then the embedding vectors should be 10000 dimensional, so as to capture the full range of variation and meaning in those words.**

- (●) True
- ( ) False

Incorrect

No, the dimension of word vectors is usually smaller than the size of the vocabulary. Most common sizes for word vectors ranges between 50 and 400.

9. **Suppose you have a 10000 word vocabulary, and are learning 500-dimensional word embeddings.The GloVe model minimizes this objective:**

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

Which of these statements are correct? Check all that apply.

- [x] θ_i and e_j should be initialized to 0 at the beginning of training.

This should not be selected

The variables should not be initialized to 0 at the beginning of training.

- [ ] θ_i and e_j should be initialized randomly at the beginning of training.
- [x] X_{ij} is the number of times word i appears in the context of word j.

Correct

- [x] The weighting function $f(.)$ must satisfy $f(0) = 0$.

Correct

The weighting function helps prevent learning only from extremely common word pairs. It is not necessary that it satisfies this function.