CS 378 Lecture 7: Word Embeddings

[add pronouns to Zoom]

Announcements

- A2 out
- Readings updated (bolding)
- Mid-semester survey
- Download A2 code for this lecture

Today

- Intro to word embeddings
- Explore embeddings
- Skip-gram: model + training
- Revisit DANs

Word Embeddings

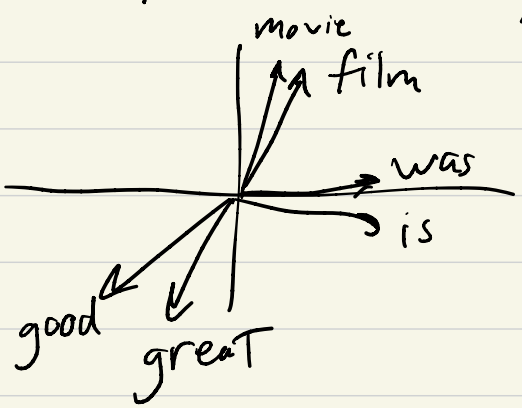
Movie was good → [1 movie, 1 good, 1 was, ...]

= [1 movie] + [1 good] + [1 was]

Each word is a $|V|$-len vector
w/ a single 1

film is great
movie was good $\Rightarrow$ dot prod = 0

Word embs: low-dimensional representations
(50–300)
that capture similarity



Distributional Hypothesis

JR Firth 1957: "you shall know a word by the company it keeps"

I watched the [movie]
I watched the [film]
The [film] inspired me
The [movie] inspired me

movie and film can show up in similar contexts

Are movie + film always substitutable?
polysemy: one word has multiple senses

Mikolov 2013: word2vec
Learn word + context vectors for each word
Attempt to predict context given word

Embedding properties

sim(good, bad) ≈ 0.8

Skip-gram

Input: corpus of sentences
Output: $\vec{v}_w, \vec{c}_w$ for each word w in the vocab
(what people use: $\vec{v}_w$ OR $\vec{c}_w$ OR $\vec{v}_w + \vec{c}_w$)

Hyperparameters: dimension d
window size K

Let $K=1$

The film inspired

look K words in each direction

Form (word, context pairs)

(film, The)
(film, inspired)

V = vocab

Skip-gram model

$$P(\text{context} = y \mid \text{word} = x) = \frac{e^{\bar{v}_x \cdot \bar{c}_y}}{\sum_{y' \in V} e^{\bar{v}_x \cdot \bar{c}_{y'}}}$$

$y \in V$

params: vectors $\bar{v}$ $|V| \times d$

$\bar{c}$ $|V| \times d$

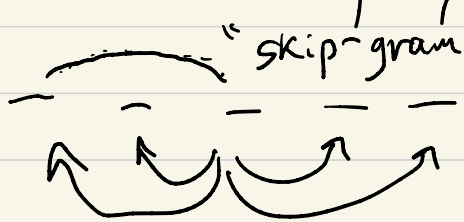
Training Take our corpus

Get (x, y) pairs

word context

Maximize $\sum_{(x,y)} \log P(\text{context} = y \mid \text{word} = x)$

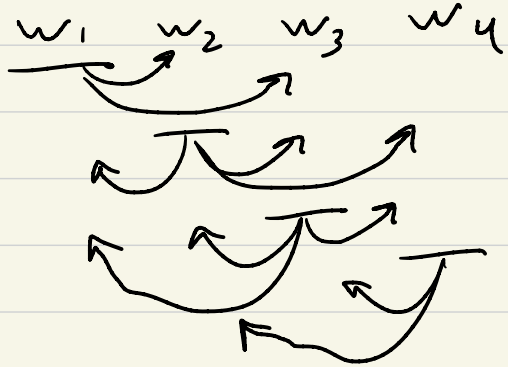
Randomly initialize $\bar{v}, \bar{c}$, use SGD



$K = 2$

$K = 3$: go 3 out, etc

K=2

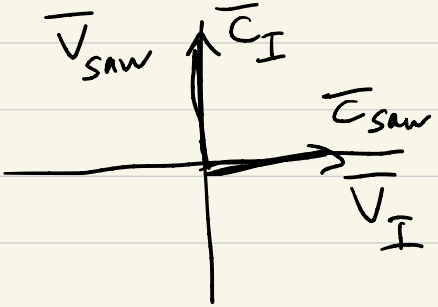


10 pairs as training data

Ex Corpus = I saw K=1

vocab = {I, saw}

Assume $\bar{V}_I = [1,0]$ $\bar{V}_{saw} = [0,1]$



① $\bar{C}_I = [0,1]$ $\bar{C}_{saw} = [1,0]$

$$P(\text{context} \mid \text{word}=\text{saw}) = \begin{cases} \text{I}: \dfrac{e^{\bar{V}_{saw}\cdot\bar{C}_I}}{e^{\bar{V}_{saw}\cdot\bar{C}_{saw}} + e^{\bar{V}_{saw}\cdot\bar{C}_I}} \\ \text{saw}: 1/4 \end{cases}$$

$3/1+3 = 3/4$

(2) $(\text{word}=\underset{\text{saw}}{\text{I}}, \text{context}=\underset{\text{I}}{\text{saw}})$ ★

$\bar{C}_{saw} = [100, 0]$ $\bar{C}_{I} = [0, 100]$

$$\frac{e^{100}}{1+e^{100}} = 0.99999\ldots$$

Maximizing likelihood is "impossible"!

(saw, I)
(saw, you)

can't assign prob 1 to each

skip-gram is slow

for each example: $O(|V| \cdot d)$ multiplies