CS371N Lecture 7
Word Embeddings

Skip-gram

Mikolov et al. 2013 "word2vec"

Learn 2 vectors for every word
- word vector
- context vector

Try to predict context given word

Inputs: corpus of text

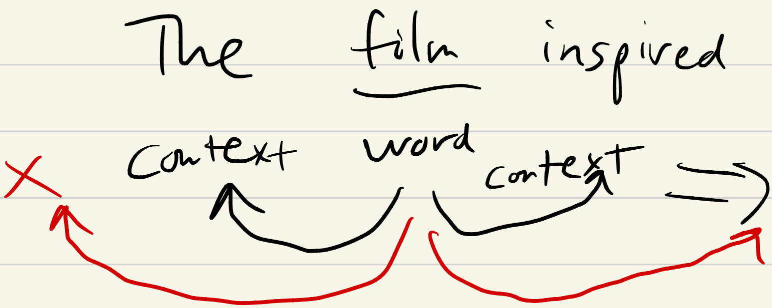
Outputs: $\bar{v}_w$, $\bar{c}_w$ for each word w in vocab V
(↑ word, ↑ context)

(In A2: you are given just one vector)

Hyperparameters: d dimension (50 - 300)

window size K

Turn a sentence into (word, context) pairs



(film, The)

(film, inspired)

K=2: Look 2 words away

(film, me)

Loop over words

from offset $\in \{-K, -K+1 \ldots -1, 1, \ldots K\}$

form pair (word, word + offset)

Model (skip-gram)

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

(y ↦ $\bar{c}_y$, x ↦ $\bar{v}_x$)

distribution over contexts

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

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

randomly initialize

Training (word, context) examples (x = word, y = context)

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

Ex Corpus = I saw $k=1$

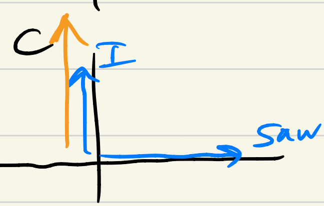
vocab = {I, saw} $d=2$

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

① Let $\bar{c}_{saw} = [1,0]$

$\bar{c}_I = [0,1]$

what is P(context | w = saw)?



2 outcomes (I, saw)

$$P(I \mid saw) = \frac{e^{\bar{v}_{saw} \cdot \bar{c}_I}}{e^{\bar{v}_{saw} \cdot \bar{c}_I} + e^{\bar{v}_{saw} \cdot \bar{c}_{saw}}} = \frac{e}{e + 1} \approx \frac{3}{4}$$

$$P(saw \mid saw) \approx \frac{1}{4}$$

② How to minimize loss further by changing $\bar{c}$? $c_I = [0, 10] \Rightarrow \frac{e^{10}}{e^{10}+1}$

③ Why do we need two spaces?
Why $\bar{V} \neq \bar{C}$?

If one space: $P(\text{saw} | \text{saw})$ has to be high! $\bar{V}_{saw} \cdot \bar{V}_{saw}$

Problems with skip-gram

Suppose we have a 100M word corpus
Vocab size = 30k vector dim $d = 300$
What's hard here?

$k=1$: 200M pairs
Each $P(\cdot | \cdot) = O(|V| d)$
$200M \cdot O(|V| d)$

Fixes

① Skip-gram w/ negative sampling (SGNS)

Take (word, context) pairs as "real" data

(word, ~sampled context) as "fake" data

Learn classifier

$$P(\text{real} \mid y, x) = \frac{e^{\bar{v}_x \cdot \bar{c}_y}}{1 + e^{\bar{v}_x \cdot \bar{c}_y}}$$

SG: 30K denom.

SGNS: 1 positive + 10 sampled negs. = 11

② GloVe

Factorizes a matrix of (word, context) counts

context \ word	the	I	saw	...
the		25	12	
I	25		1512	
saw	12	1512		
...				

$= M$

matrix factorization

$$V^T \quad C \quad = M$$

$(d \times |V|) \, (d \times |V|) \quad (|V| \times |V|)$

Gives the same solution as SG/SGNS