Skip-gram

Mikolov et al. 2013 "word2vec"

Learn 2 vectors for every word

word vector

context vector

Try to predict context given word

==

Inputs: a corpus of text

Output: $\vec{v}_w$, $\vec{c}_w$ for each word w in vocab

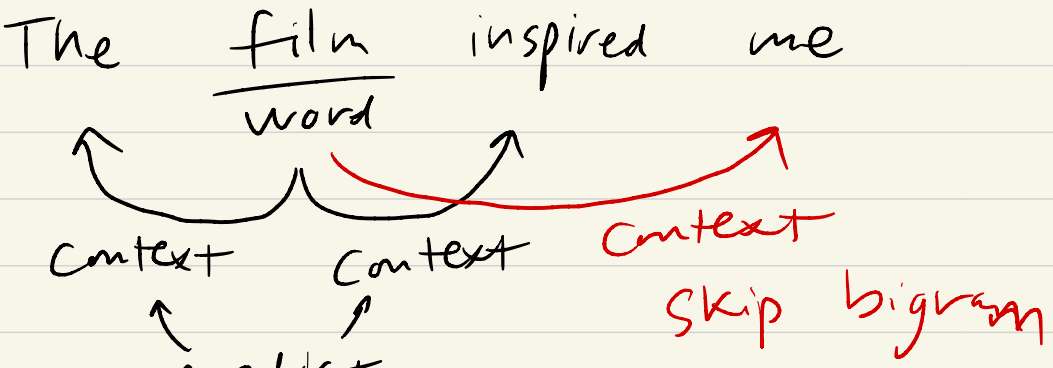
($\vec{v}_w$: word, $\vec{c}_w$: context)

(for A2: use $\vec{v}$, or $\vec{v} + \vec{c}$)

Hyperparameters: d ($50 \sim 300$)

window size k

Let $K=1$ $K=2$ (skip-gram includes $K=1$)



predict

word context

(film, The)

(film, inspired)

(film, me)

Training examples

Model (skip-gram)

$\bar{v}_x$

$P(\text{context} = y \mid \text{word} = x)$

$$= \frac{e^{\bar{v}_x \cdot \bar{c}_y}}{\sum_{y' \in V} e^{\bar{v}_x \cdot \bar{c}_{y'}}}$$

distribution over context words in V

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

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

randomly initialized

Training

(x, y) train exs

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

$P(\text{context} \mid \text{word} = \text{saw})$

2 outcomes (I, saw)

$e^{\bar{v}_{saw} \cdot \bar{c}_I} = e$ $e^{\bar{v}_{saw} \cdot \bar{c}_{saw}} = 1$

$$P(I \mid saw) = \frac{e}{e+1} \approx \frac{3}{4}$$

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

(2) How to minimize loss further by changing $\bar{c}$?

$c_I = [0\ 2] \Rightarrow c_I = [0\ 10]$

$\frac{e^{10}}{e^{10}+1} \approx 0.999$

(3) Why is $\bar{v} \neq \bar{c}$? Why two spaces?

(saw, saw) always be high!

word vector selects for words that are near it

noun → verbs

noun ↛ nouns

Problems with skip-gram

If we ran this training over 100M word corpus with $V = 30K$, what's going to be hard?

- polysemy: different word senses
  - different vector per sense?
  - train on a homogeneous corpus
  - context-dependent vectors (BERT, GPT)
- computation: $|V|$, d 50~300

$$P(y|x) = O(|V|d)$$

For training: do that $\times$ 100M

Two fixes

Skip-gram w/negative sampling

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

(word, ~sampled context) as fake data

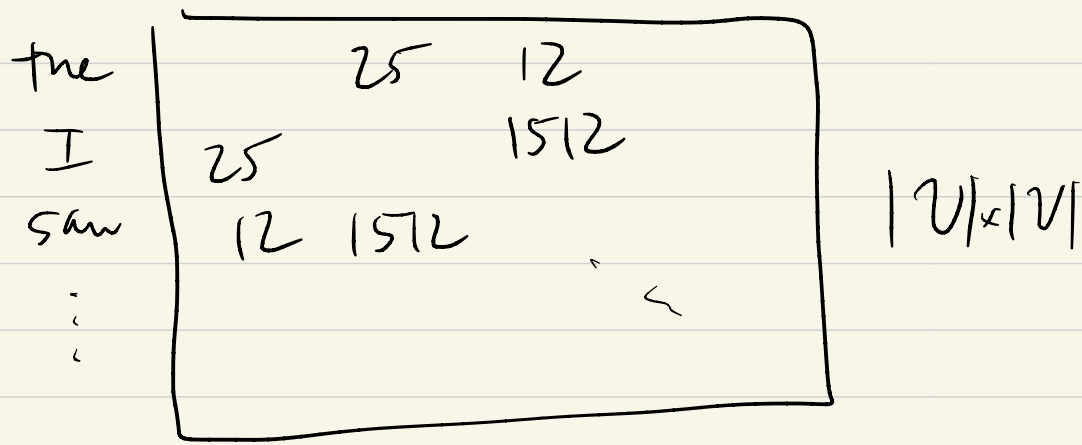
Learn a classifier

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

(film, buy) is this fake?

GloVe

Factorizes a matrix of (word, context) counts (K=1)



matrix factorization

$$(|V| \times v) \times (v \times |V|) \approx |V| \times |V|$$

Same as SG + SGNS