Word Embeddings → low dim rep of words capturing similarity

$\sim 30-500$

Distributional hypothesis: "You shall know a word by the company it keeps"

Mikolov et al 2013 word2vec

each word w → $\bar{v}_w$ word vector
$\bar{c}_w$ context vector

predict each word's context given that word

Skip Gram

Input: large corpus of sentences
output: $\bar{v}_w, \bar{c}_w$ for each word w
Hyperparams: word vectors dim d $(\sim 50-300)$
window size k $(k > 1)$ — neighbors of each word taken up to k positions away

the film inspired — film → inspired, film → the

skip-gram: context | word → $P(\text{context}=y \mid \text{word}=x) = \dfrac{\exp(\bar{v}_x \cdot \bar{c}_y)}{\sum_{y' \in V} \exp(\bar{v}_x \cdot \bar{c}_{y'})}$

sum over vocab → $\sum_{y' \in V}$

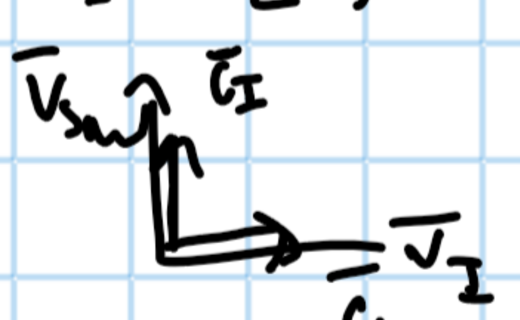
if $\bar{v}_x$ is ~ to $\bar{c}_y$, y is likely to be in x's context

$\bar{v}$ and $\bar{c}$ are model params

$2 \cdot |V| \times d$

ex: corpus = I saw

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



word	context
I	saw
saw	I

if $\bar{c}_{saw} = [1, 0]$ and $\bar{c}_I = [0, 1]$

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

$\exp(v_{saw} \cdot c_I)$ ≈ 3 $\quad \exp(v_{saw} \cdot c_{saw})$ → 0 → 1

$P(\text{context}=I \mid \text{word}=saw) = \frac{3}{4}$ $\quad$ saw|saw $= \frac{1}{4}$

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

(x,y) pairs in data

init params randomly

"impossible" to get $P \to 1$