Skip-gram

Input: large corpus of sentences

Output: $\bar{v}_w$, $\bar{c}_w$ for each word type w

Hyperparams: word vector dim d ($\sim 50-300$)
window size k (assume $k=1$)

The film inspired

word	context
film	inspired
film	The

Take all neighbors of each word token up to k positions away

Skip-gram: probabilistic model of context | word

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

$\bar{v}, \bar{c}$ model params

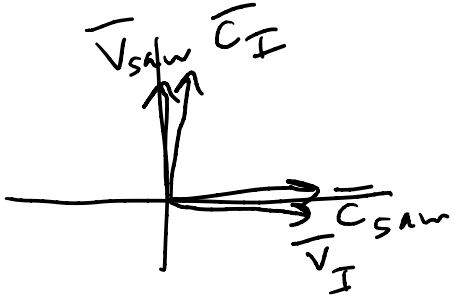
If $\bar{v}_x$ is similar to $\bar{c}_y$, y is likely to be in x's context

sum over vocab $\rightarrow y' \in V$

$2 \cdot |V| \times d$ params in model

Ex Corpus = I saw

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



word	context
I	saw
saw	I

If $\bar{c}_{saw} = [1,0]$ and $\bar{c}_I = [0,1]$, what is $P(\text{context} \mid \text{word} = \text{saw})$?

$\exp(\bar{v}_{saw} \cdot \bar{c}_I) \approx 3$ $\exp(\bar{v}_{saw} \cdot \bar{c}_{saw}) \approx 1$ ↑ Vocab

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

Training

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

pairs in data

"Impossible" problem: cannot drive $P \to 1$

Initialize params randomly