Skip-Gram

Input: Large corpus of sentences

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

Hyperparams: word vector dimension d (~ 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

- **Skipgram model**: Probabilistic model of context given a word

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

sum over all vocab

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

$|V| \times d$ dimensions

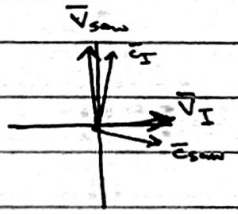
(2 · # of params in model)

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

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

$\exp(\bar{v}_{saw} \cdot \bar{c}_I) \quad \exp(\bar{v}_{saw} \cdot \bar{c}_{saw})$

$* 3 \quad = 1$ (b/c orthogonal)

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

- Training Skip-gram

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

"Impossible Problem"
cannot drive $P \to 1$

* Initialize parameters randomly (unlike naural nets)
  ↳ Iterating over data, model pulls similar vectors together over time

* other Word-Embedding Methods on "seg-19.pdf"