

CS60010: Deep Learning

Spring 2023

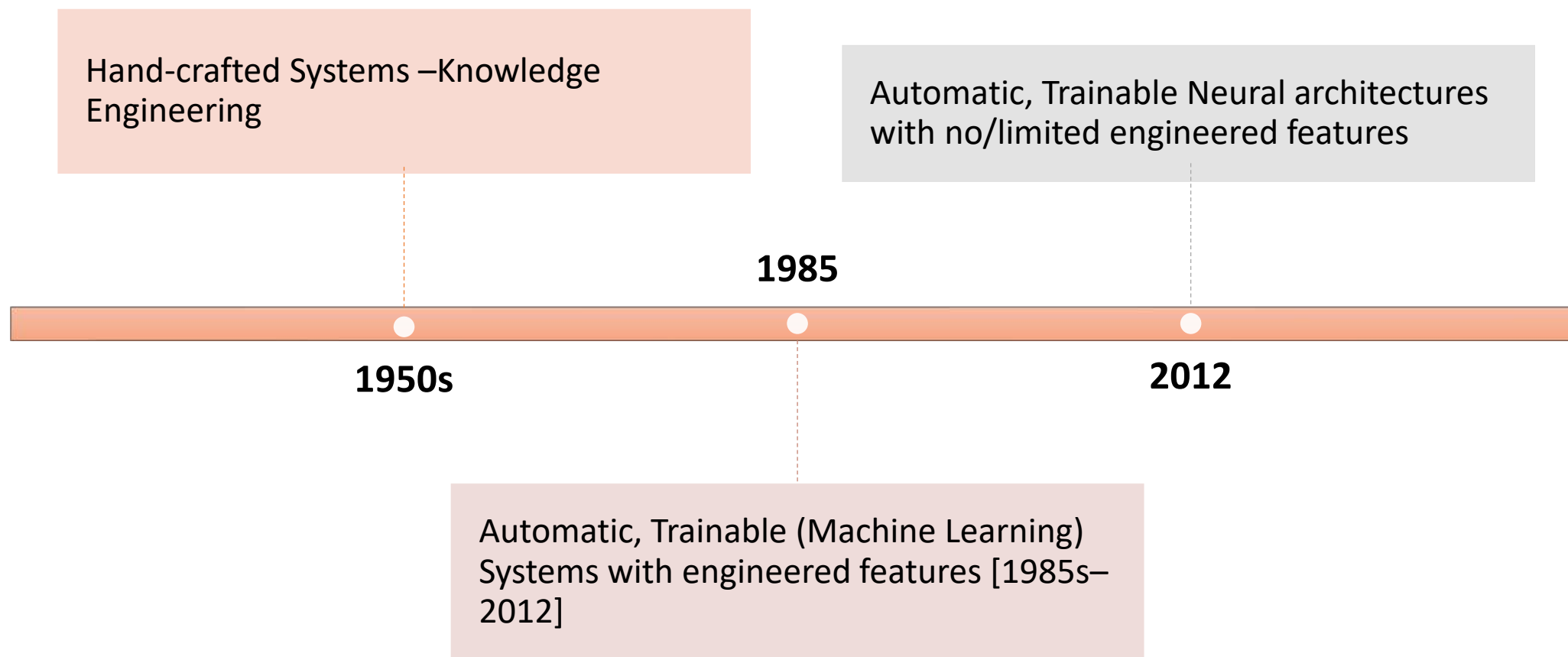
Sudeshna Sarkar

NLP Word Embedding

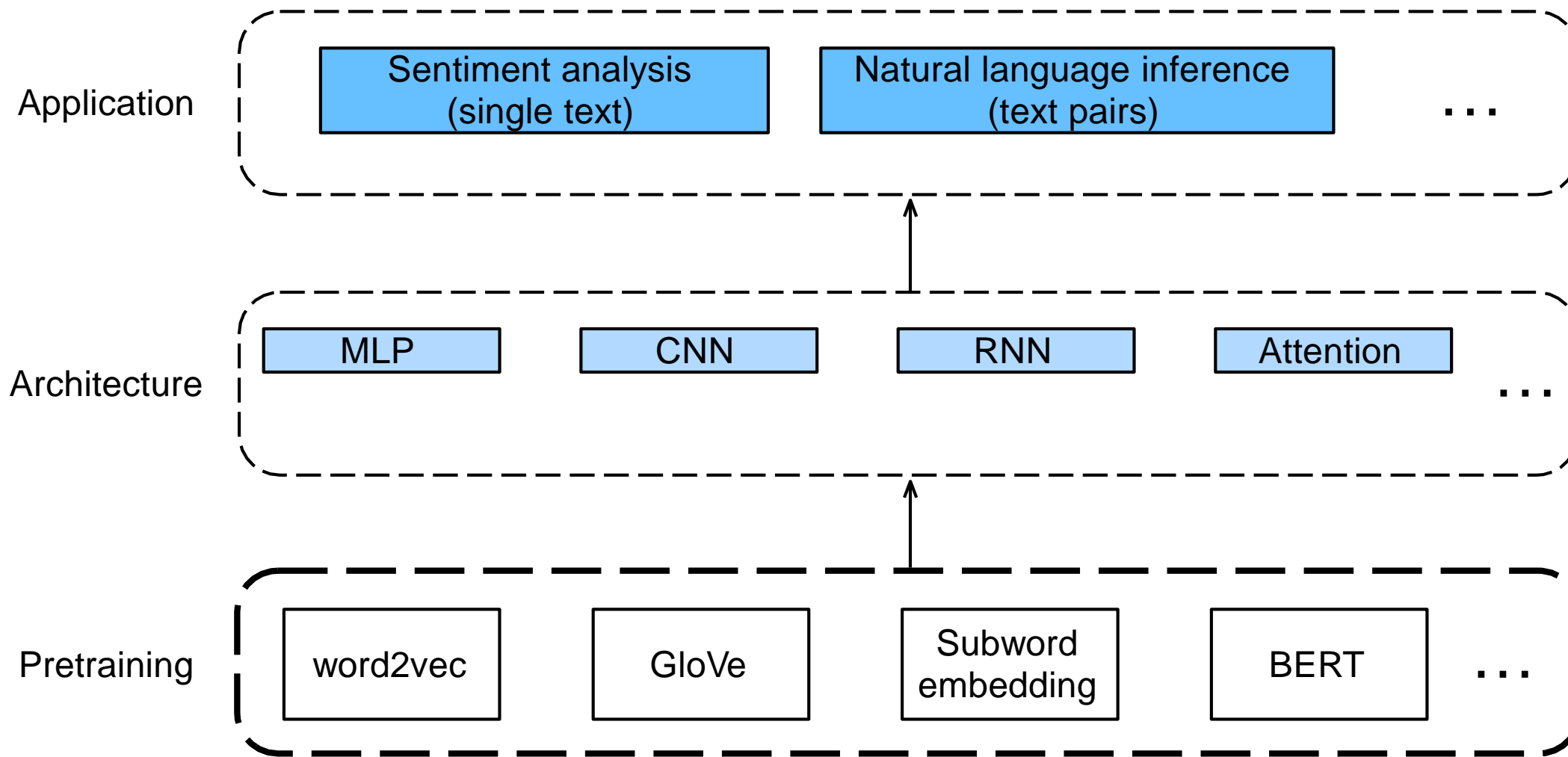
Sudeshna Sarkar

9 Mar 2023

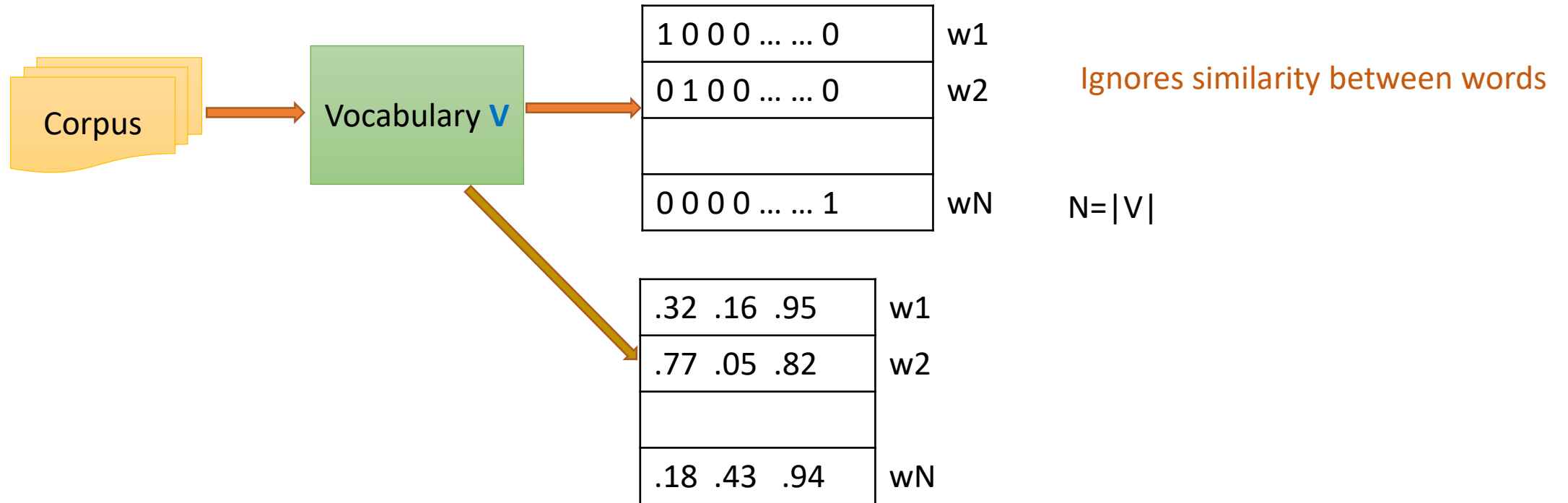
Three Generations of NLP



NLP Roadmap



Word Representations



Word2vec: Represent each word with a low-dimensional dense vector

Model more generalizable

Word2Vec

Input: A large corpus of text, V, d

- V : a vocabulary
- d : a dimensionality (i.e. 50, 300)
- Text Corpora
 1. Wikipedia
 2. Twitter
 3. Common Crawl

$$\begin{aligned} v_{\text{cat}} &= \begin{pmatrix} -0.224 \\ 0.130 \\ -0.290 \\ 0.276 \end{pmatrix} & v_{\text{dog}} &= \begin{pmatrix} -0.124 \\ 0.430 \\ -0.200 \\ 0.329 \end{pmatrix} \\ v_{\text{the}} &= \begin{pmatrix} 0.234 \\ 0.266 \\ 0.239 \\ -0.199 \end{pmatrix} & v_{\text{language}} &= \begin{pmatrix} 0.290 \\ -0.441 \\ 0.762 \\ 0.982 \end{pmatrix} \end{aligned}$$

Output: A vector for every element of V of dimension d

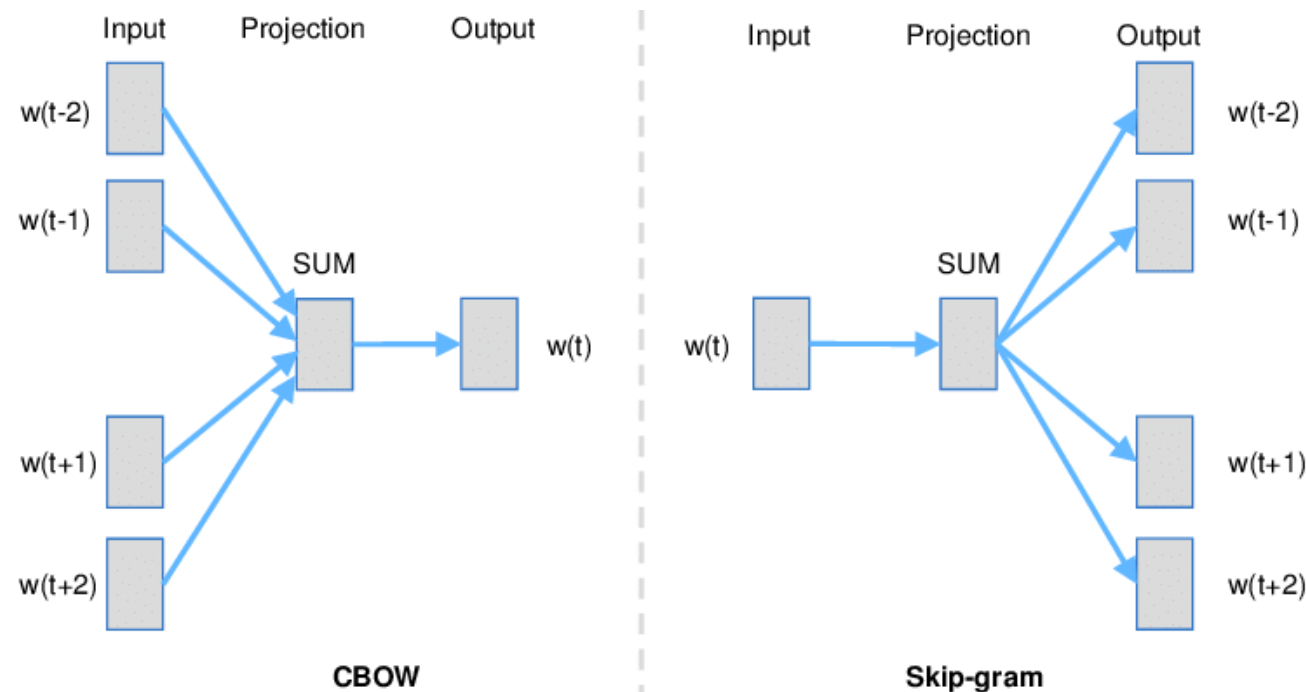
Word2vec Representations

“You shall know a word by the company it keeps”

Key idea: Predict surrounding words of every word

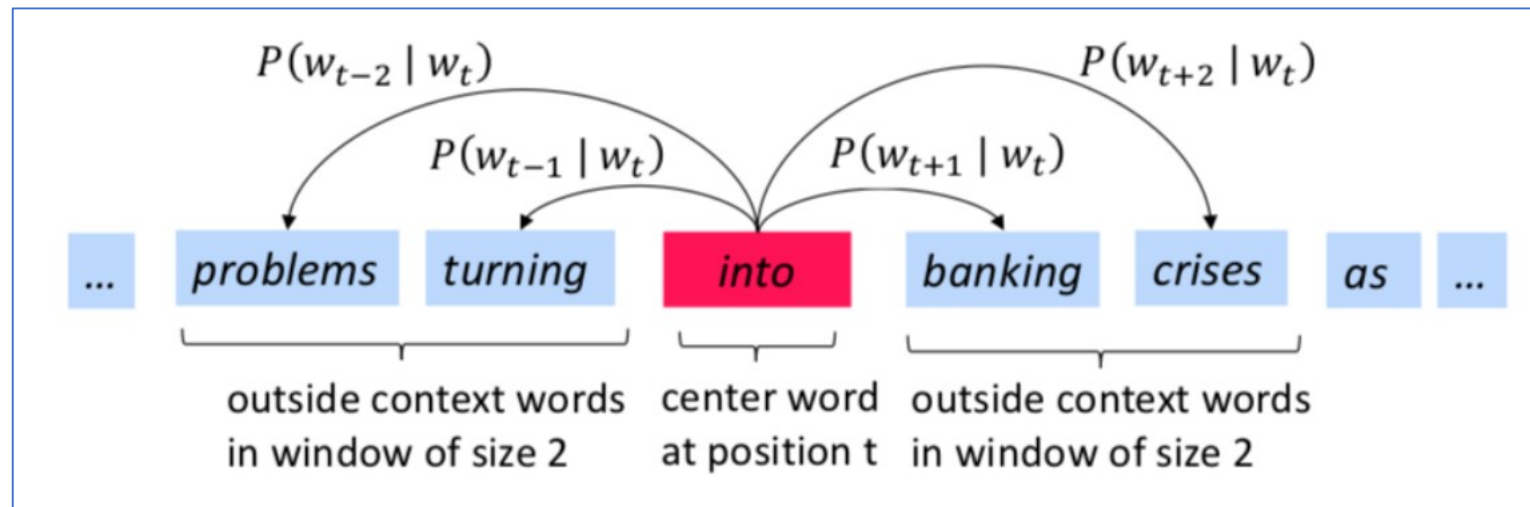
Assign each word a vector such that similar words have similar vectors

1. CBOW: $P(\text{Word} | \text{Context})$
2. Skipgram: $P(\text{Context} | \text{Word})$



Word2vec : Skip-grams

Given a word, predict the words you expect to see in the context.



Compute over entire corpus, moving center word.

Word2vec Representations

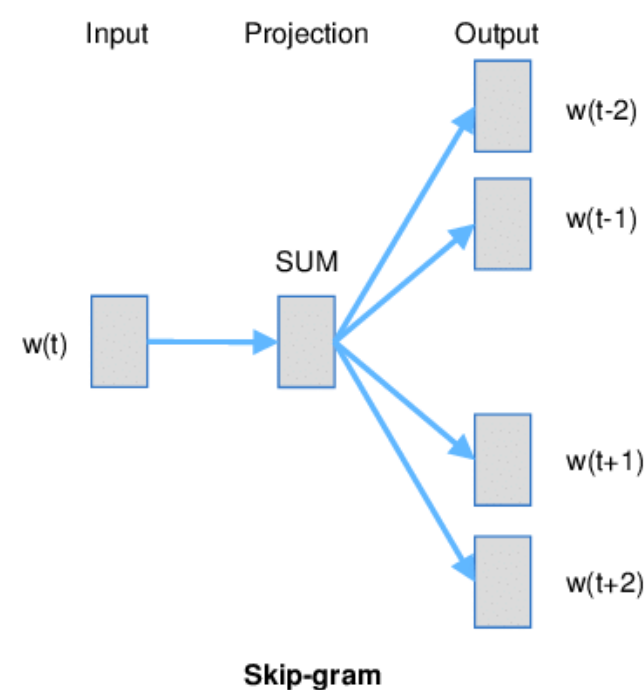
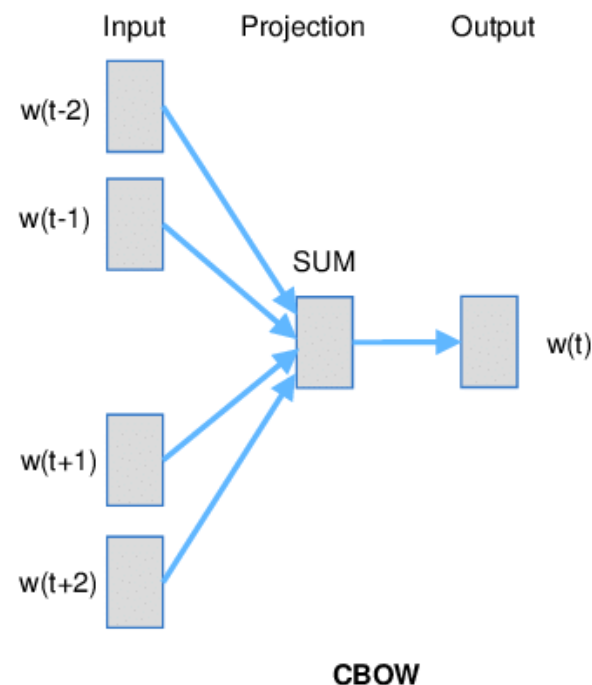
1. CBOW: $P(\text{Word} | \text{Context})$
2. Skipgram: $P(\text{Context} | \text{Word})$

Train a classifier on a binary **prediction** task:

Is w likely to show up near "*desert*"?

Take the learned classifier weights as the word embeddings

Use running text as implicitly supervised training data.



Skip-grams - Loss

Given a word, predict the words you expect to see in the context

$$L_{\theta} = \frac{1}{T} \times \sum_{i=1}^T \sum_{\substack{j \in \{i-m \dots i+m\} \\ j \neq i}} \log(p_{\theta}(w_j | w_i))$$

Overall loss, average of all tokens in the dataset, negative log probability

Minimizing loss

- We want to minimize the objective function.

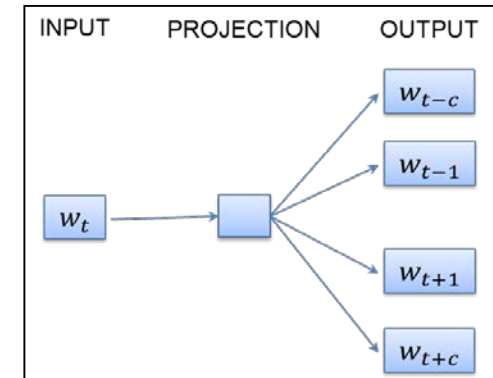
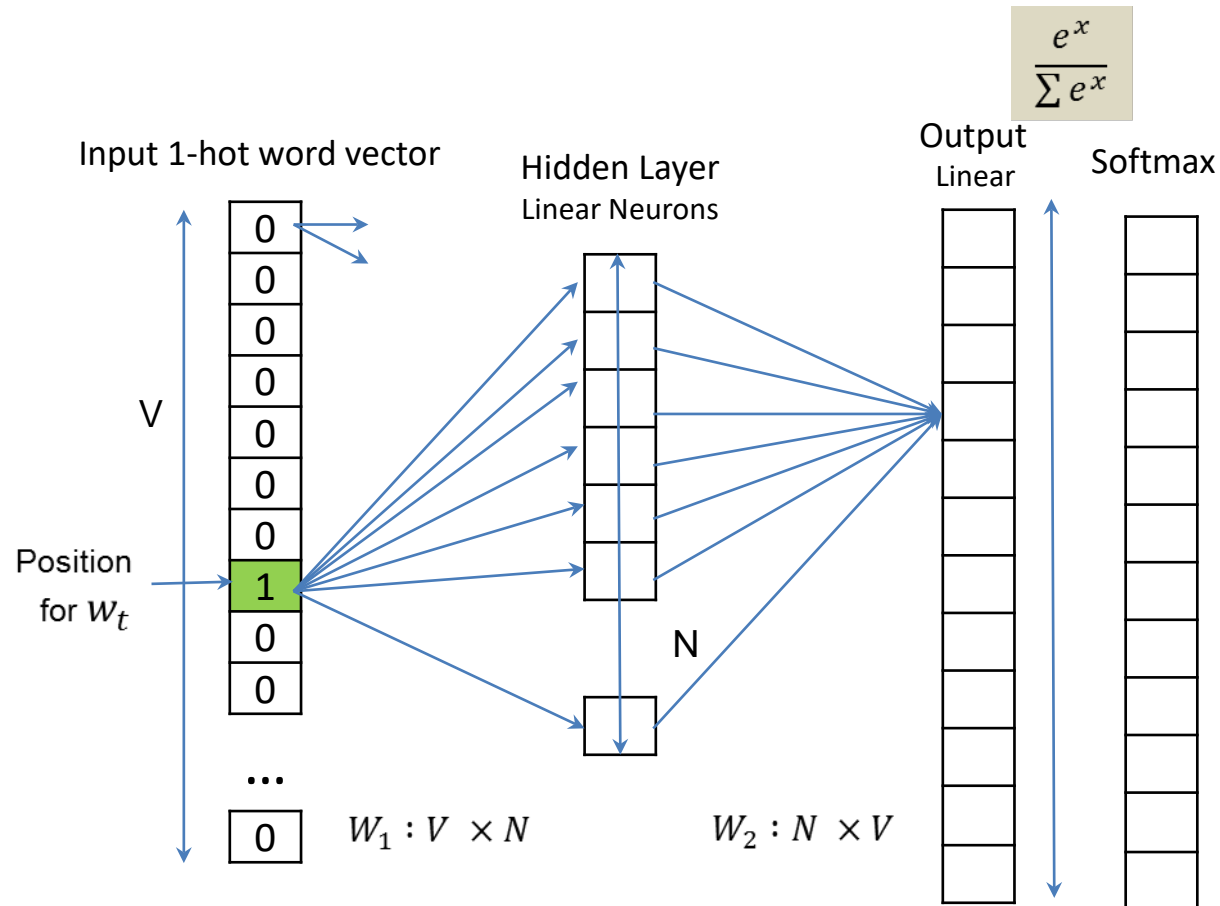
$$L_{\theta} = \frac{1}{T} \times \sum_{i=1}^T \sum_{\substack{j \in \{i-m \dots i+m\} \\ j \neq i}} \log(p_{\theta}(w_j | w_i))$$

- Two sets of vectors per word:
 - v_w for centre word
 - u_w for context word
- For a centre word c and a context word o :

Only parameters to this model are the two sets of embeddings

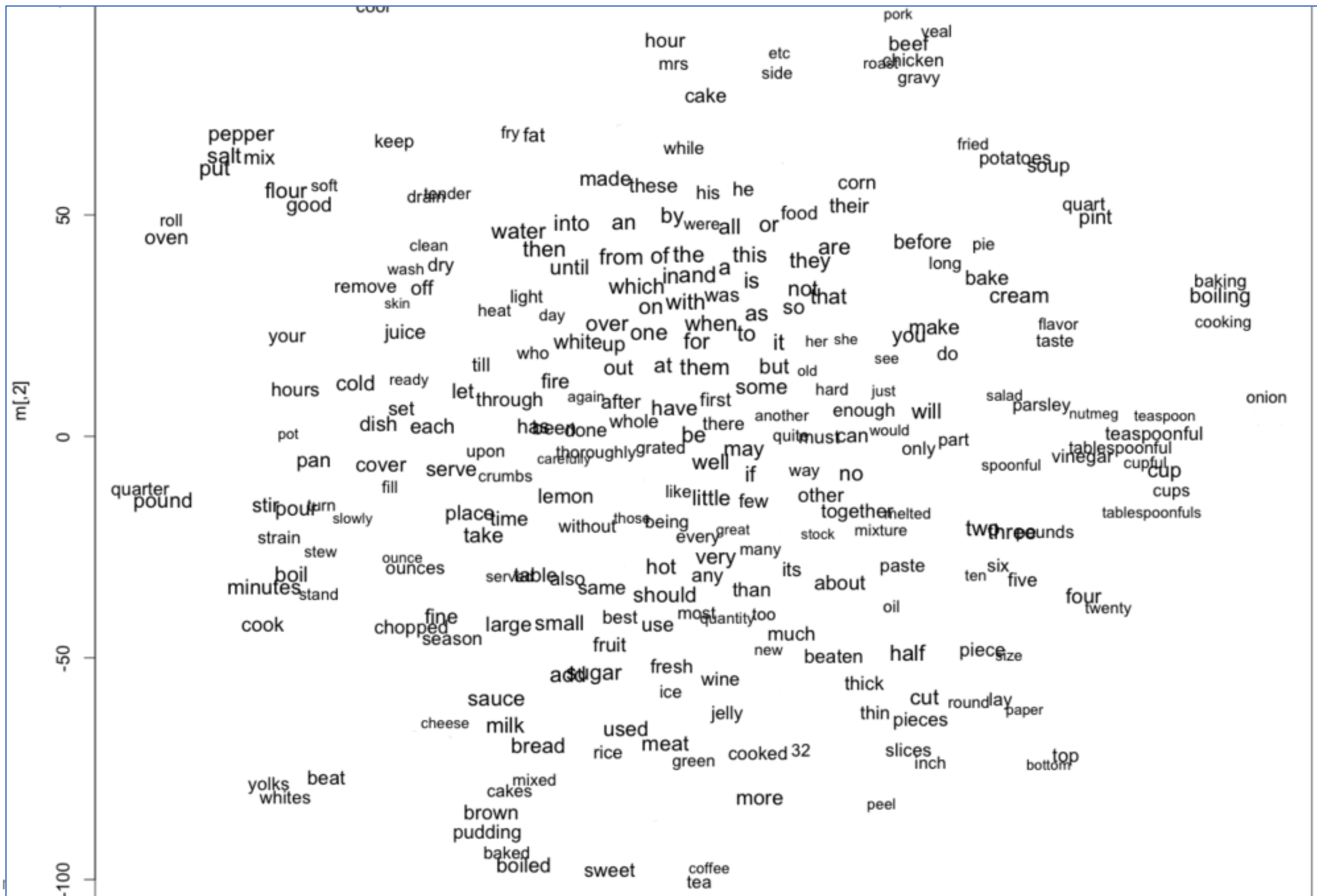
$$p(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

Skipgram Model

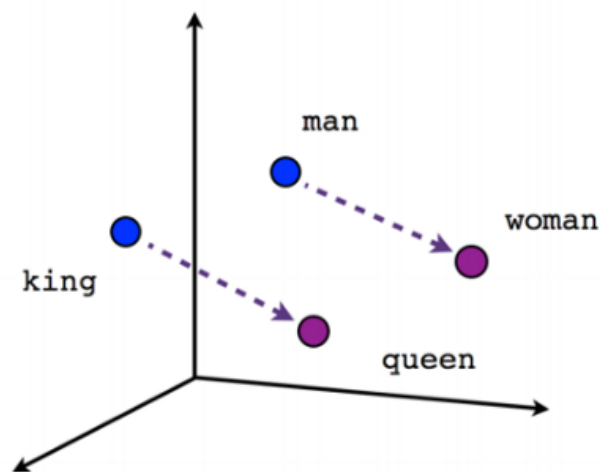


Probability that the word in a context position is w_i

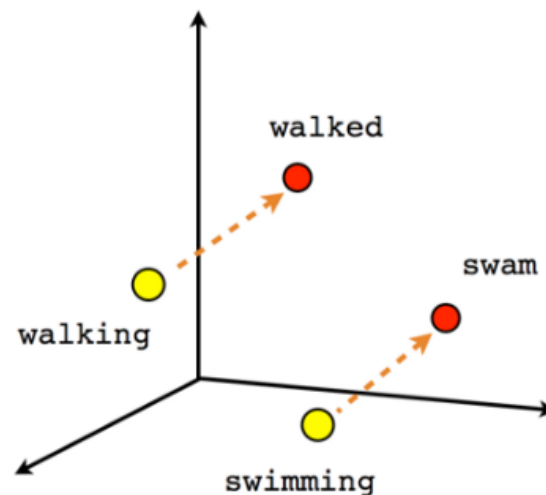
Learn with stochastic gradient decent: $\theta_{t+1} = \theta_t - \eta \nabla_{\theta} L_{\theta}$



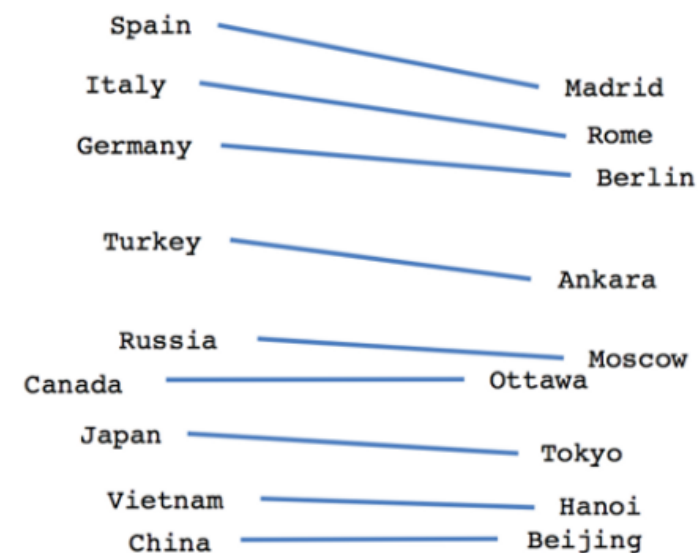
Embeddings capture relational meaning!



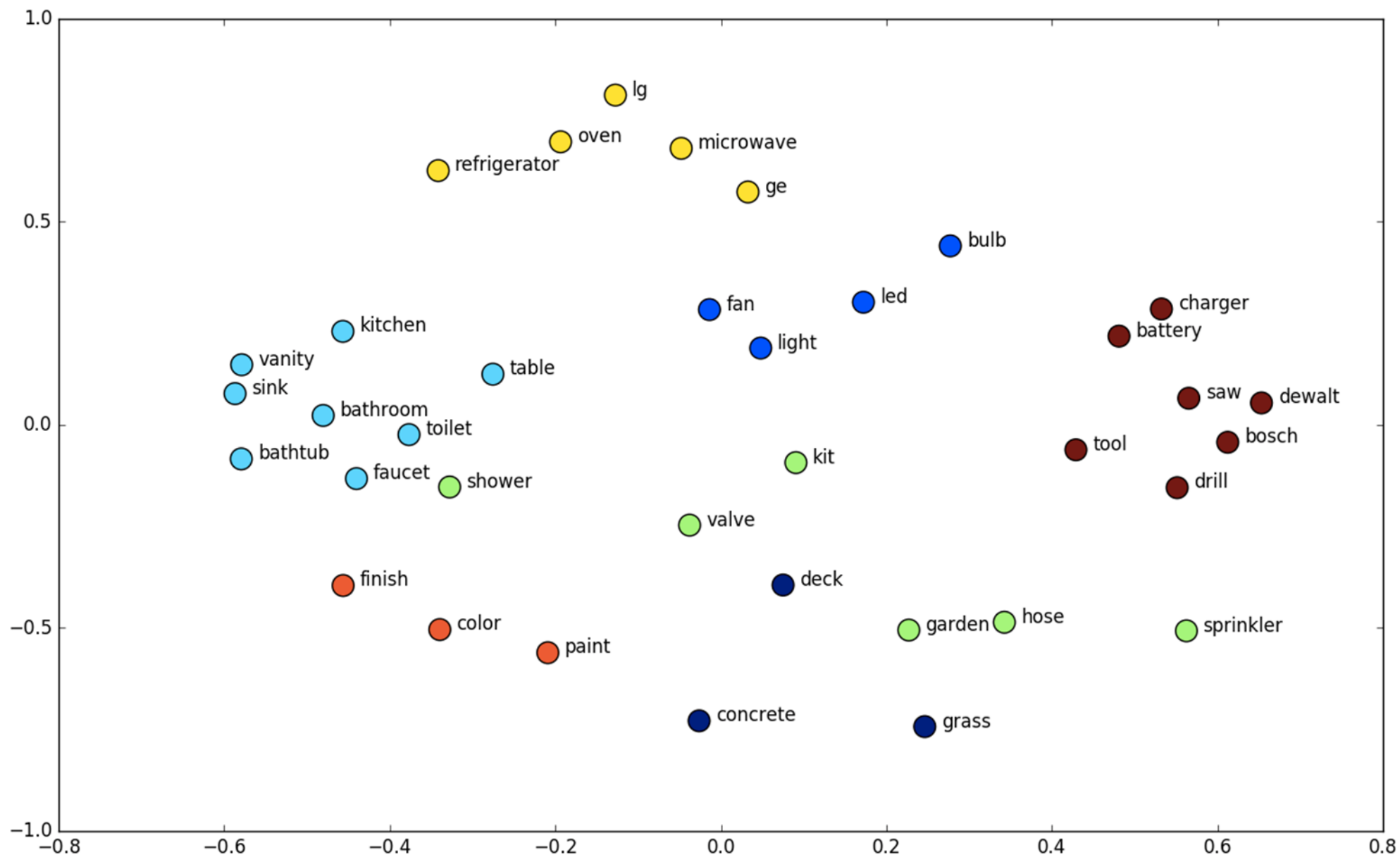
Male-Female




Verb tense



Country-Capital



Problem of Shallow Word Representation

 Mongabay.com

Upset about Amazon fires last year? Focus on deforestation this year (commentary)

That is, the recent deforestation surge fueled the 2019 Brazilian Amazon fires. The fires were in fact a lagging indicator of recent deforestation. Such information ...

1 week ago



 Business Insider

How to restart your Amazon Fire Stick in 3 different ways - Business Insider

The Amazon Fire Stick is one of the most popular media streaming devices available right now. Not only can you shop on Amazon using a Fire Stick, but you can ...

10 hours ago

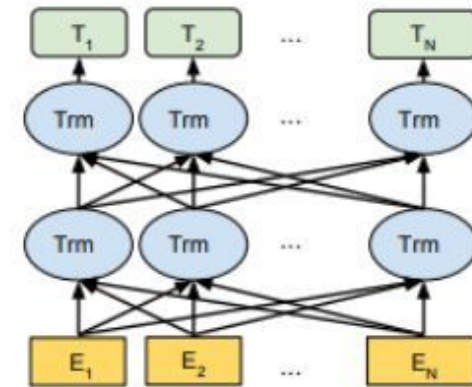


Contextualized Word Vectors

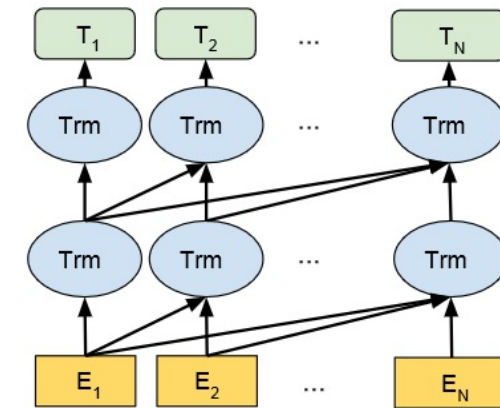
Incorporating context into word embeddings
a watershed idea in NLP

- BERT: Bidirectional Encoder Representations from Transformers (BERT, 2018)
- GPT-2/3

Led to significant improvements on virtually every NLP task.



BERT Architecture



GPT