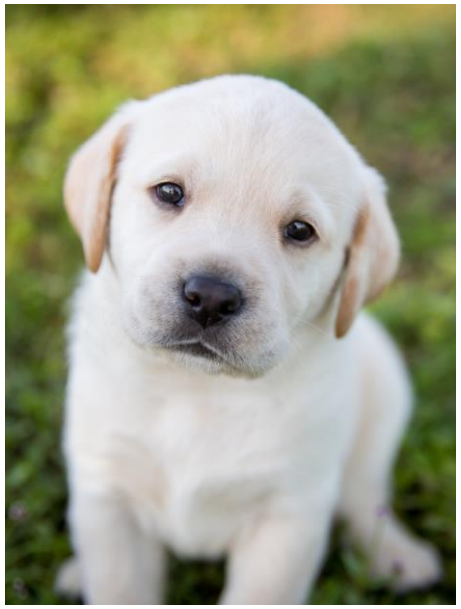
manipal PROlearn

Embeddings in text

By

Mirza Rahim Baig

Term vs concept



"Dog"

What is a dog anyway?

Are all dogs the same?

Will its qualities change if you start calling it a rose?

- 'terms' in various languages

Terms are **linguistic representations** of **concepts - merely symbols!**

BTW do dogs bark in different languages?

Recall: Semantic associations between words

Synonyms

- "Sidewalk" and "Footpath"

Antonyms

- Example: "hot" and "cold"

Homonymy

- dog bark vs tree bark

Polysemy

- The magazine is interesting, vs.
- The magazine fired its staff

Meronyms and Holonyms

- "cockpit" is a part of a "plane"

Hypernyms and Hyponyms

- "dog" is the hyponym of "animal"

Can semantic associations be captured in BOW models?

Distributional Semantics

'You shall know a word by the company it keeps'.

- John Firth, 1957

Guess the meaning of the term 'furwala'-

"I adopted a young Persian furwala a month back. Like all furwalas, it loves to scratch its back and hates water, but unlike other furwalas, it miserably fails at catching a mouse"

Words with similar meanings tend to be used in similar contexts

Distributional Semantics

Distribution over occurrence contexts

- Term x context matrix
- Rows represent terms, columns represent context vectors
- E.g. LSA

Distribution over occurrence of terms

- Square matrix showing co-occurrence between terms
- Symmetric matrix
- E.g. word2vec, GloVe
- Co-occurrence if the terms occur in the same context

Note: Both approaches have very high dimensionality for the vectors for the words

Word vectors

Measuring similarity between co-occurrence vectors?

- **Norm**
  - Euclidean if p = 2

$$L_p(t_i, t_j) = \sqrt[p]{\sum_{k=1}^{n}(|f_{j,k} - f_{i,k}|)^p}$$

- **Cosine similarity**
  - Between 0 and 1 since counts are non-negative
  - Concerned more about the angle

$$cos(t_i, t_j) = \frac{t_i \cdot t_j}{\|t_i\|_2 \|t_j\|_2} = \frac{\sum_{k=1}^{n} f_{i,k} \cdot f_{j,k}}{\sqrt{\sum_{k=1}^{n} f_{i,k}^2}\sqrt{\sum_{k=1}^{n} f_{j,k}^2}}$$

Hands on

Word Embeddings

Raw representations had - sparsity, dependence, noise terms

Embeddings -

- Much lower dimensional representation (50 - 500)
- Process of mapping terms to lower space while preserving their distributional semantics
- E.g. SVD, SkipGram, CBOW

Stanford DL for NLP Lectures on Youtube

word2vec

Key idea: prediction objective (prediction in local neighbourhood)

- Predict between every word and its context words

Two algorithms -

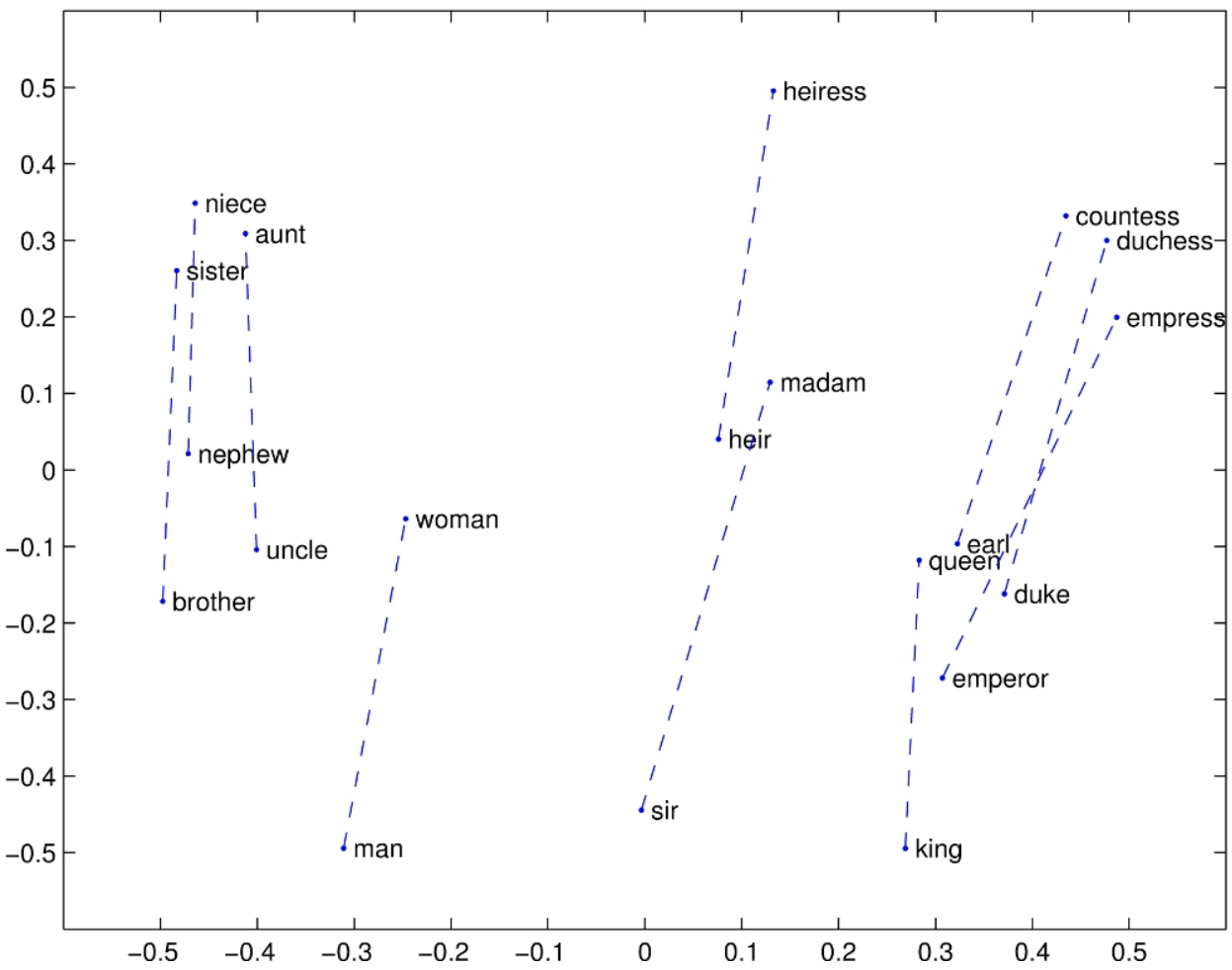
1. Skip grams
2. Continuous bag of words

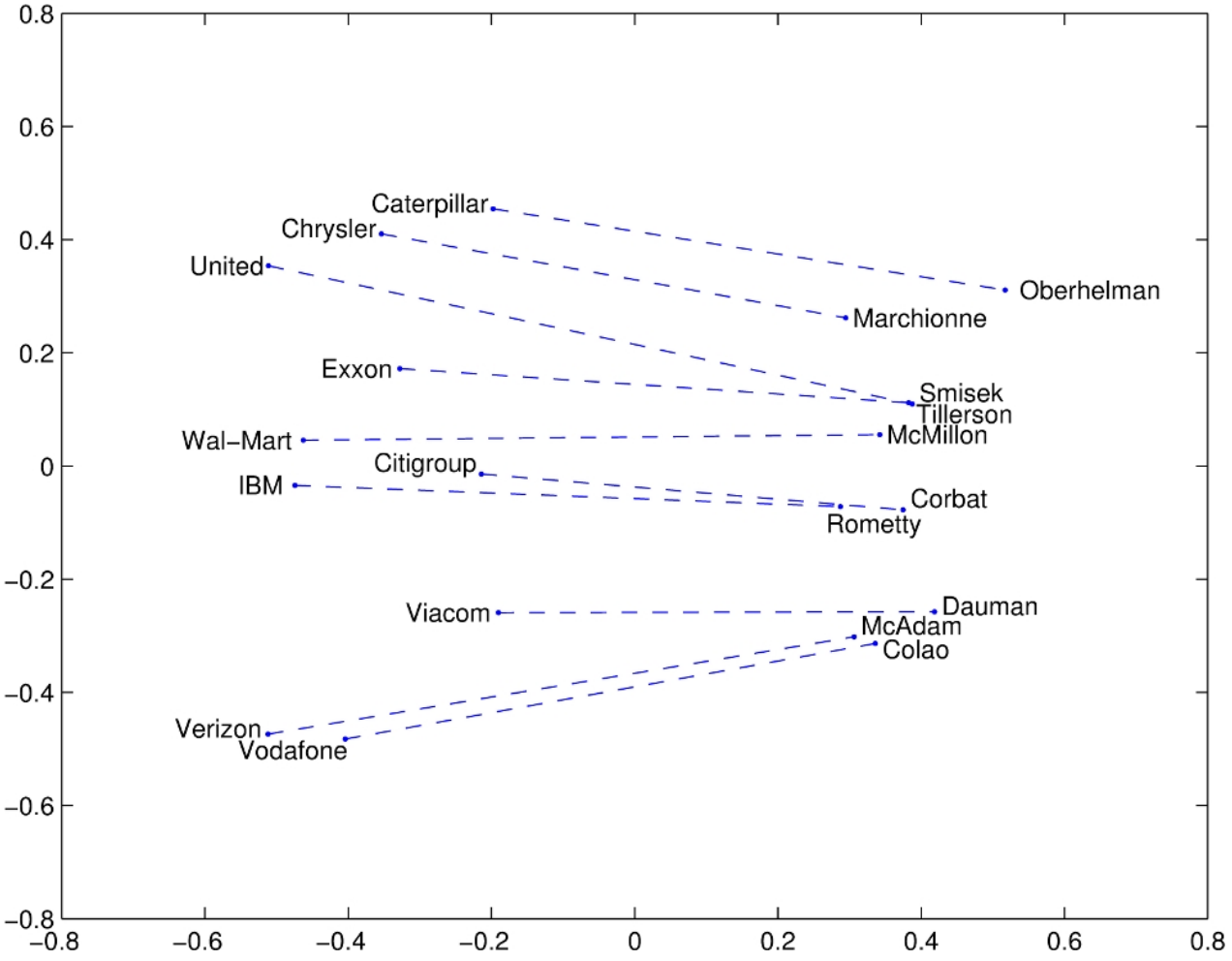
Two training algorithms -

1. Hierarchal softmax
2. Negative sampling

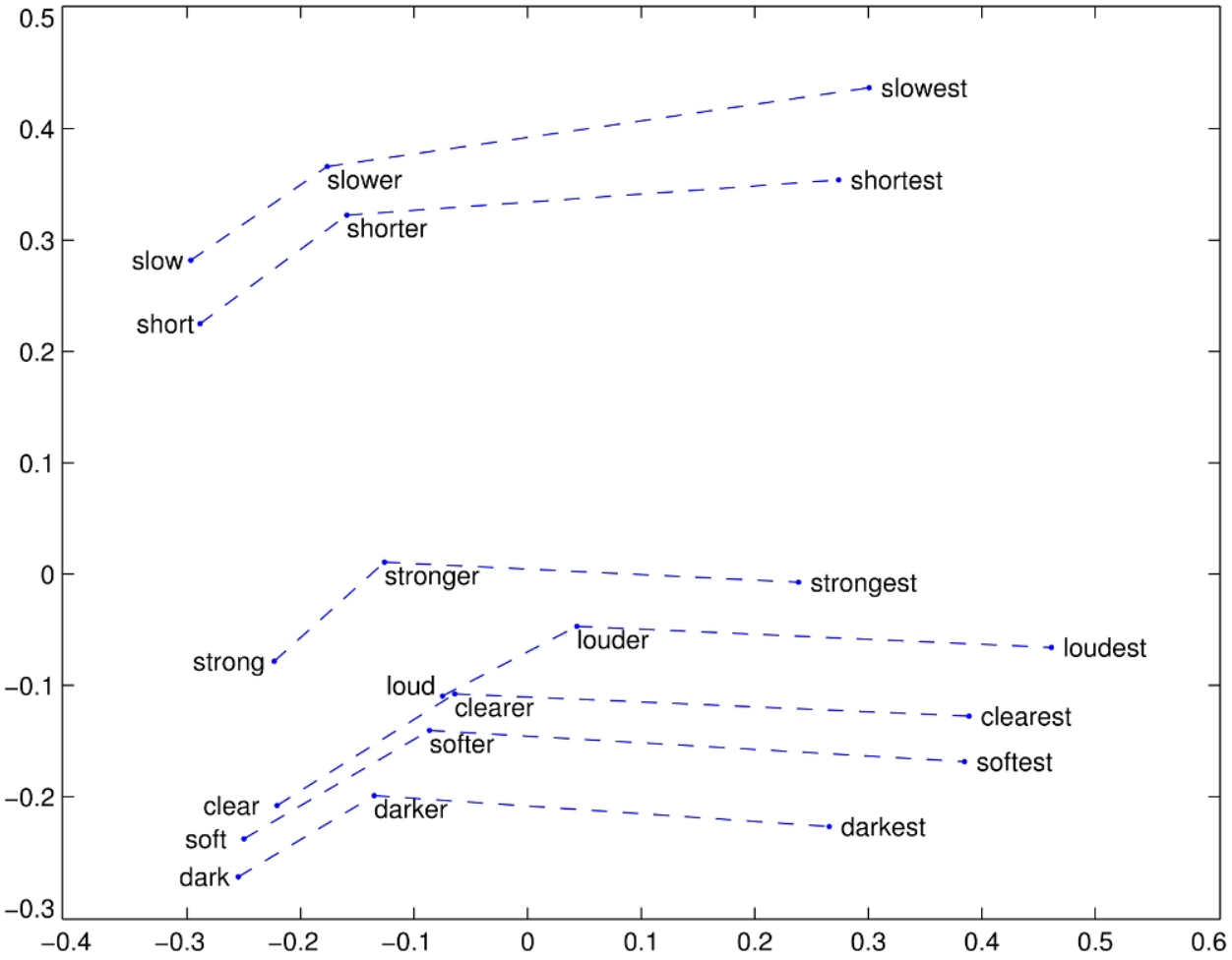
Capture semantic regularities: Demo

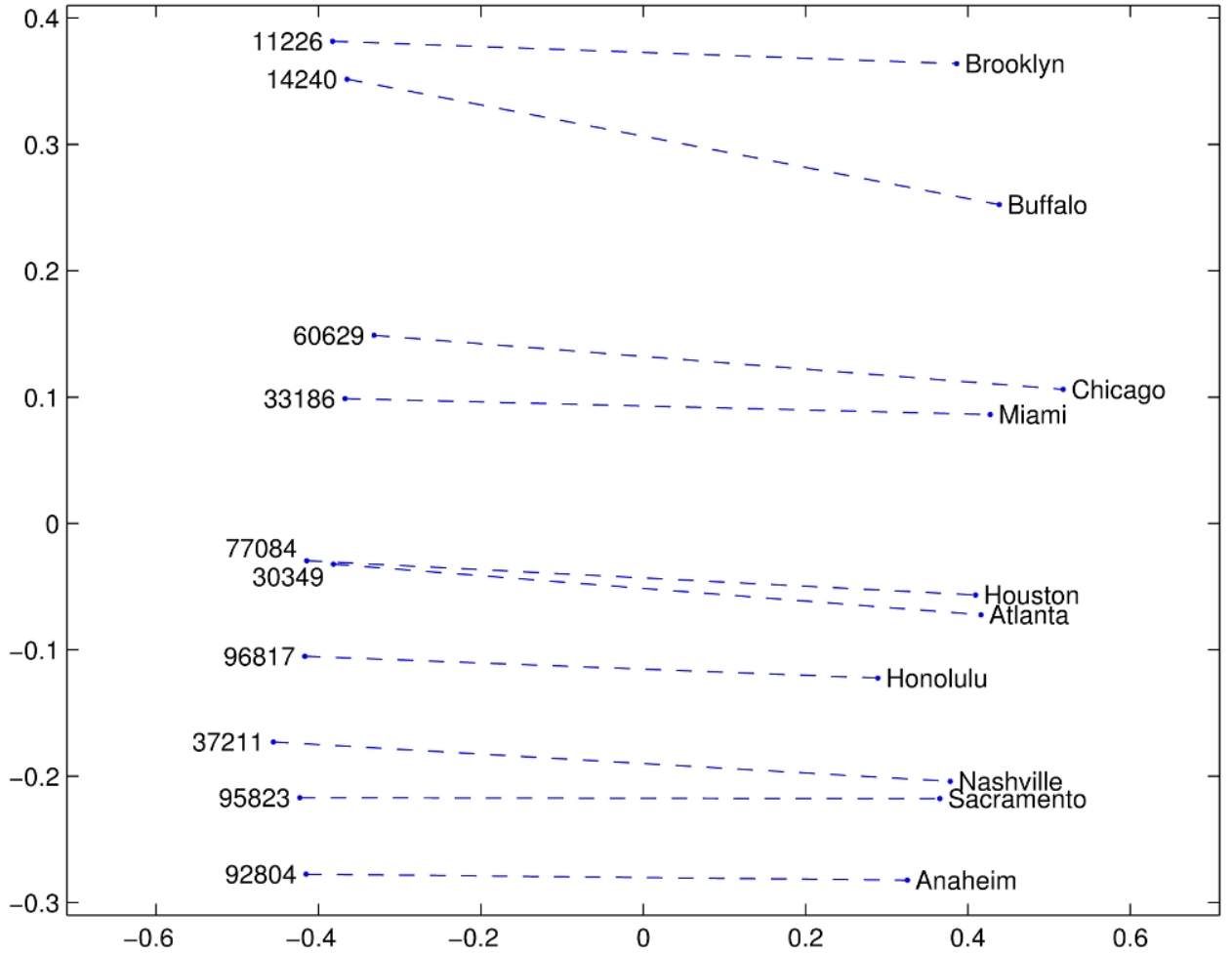
Semantic regularities captured in word vectors



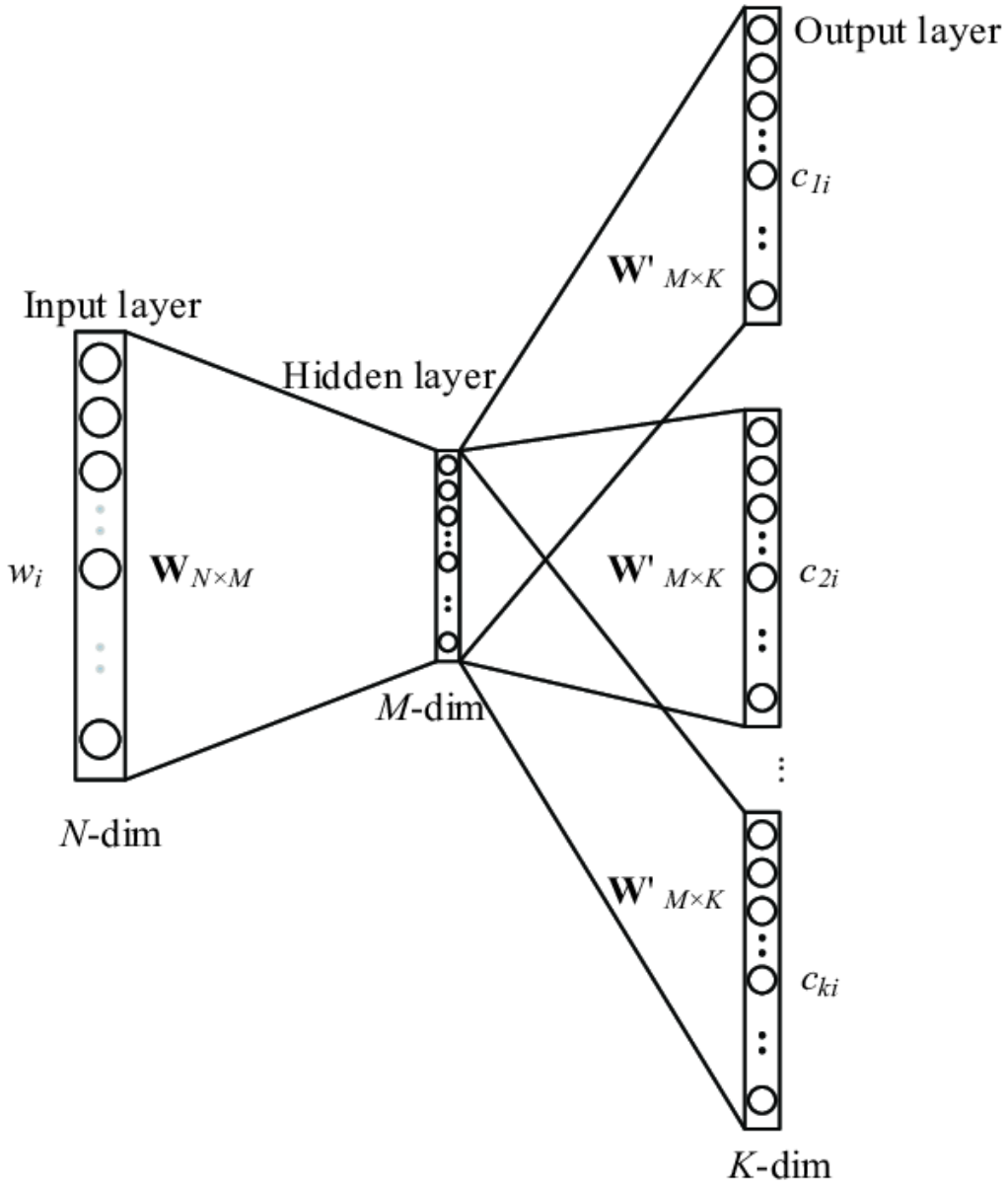


Semantic regularities captured in word vectors





Skipgram

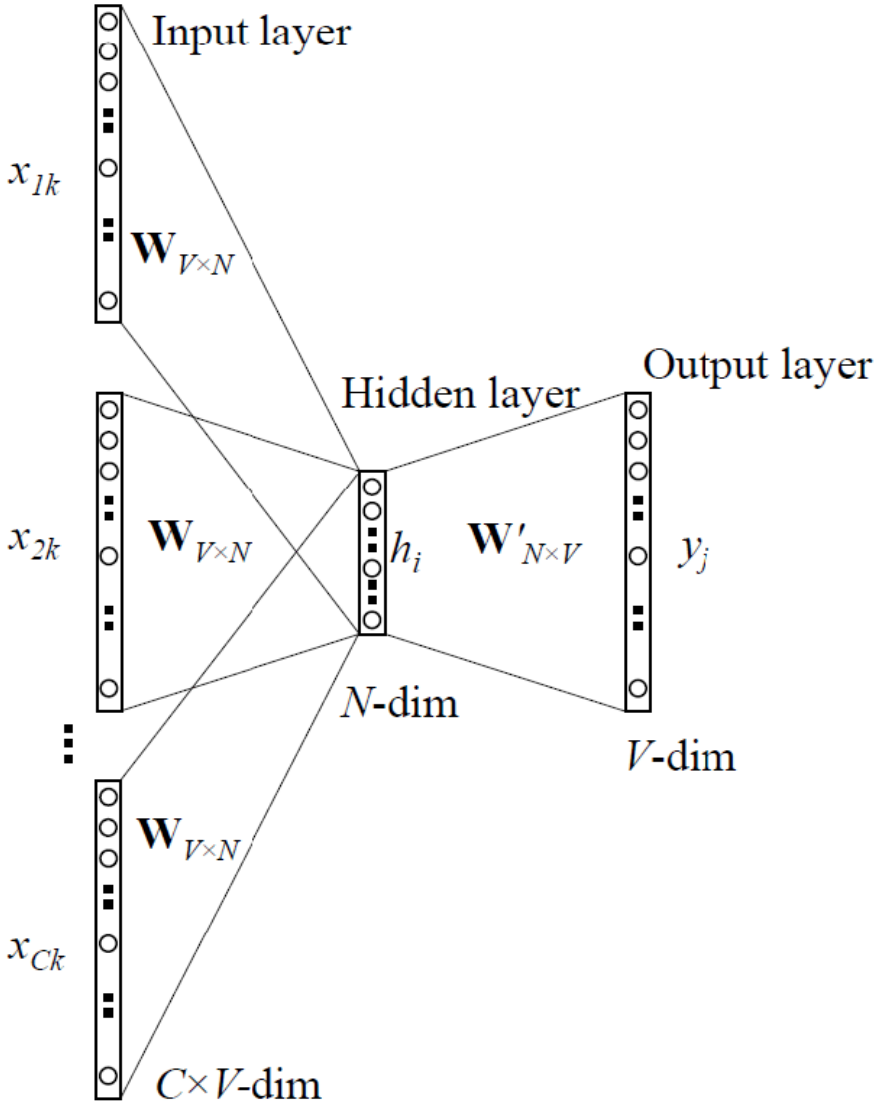


$$\arg\max_{w^c \in \mathcal{P}(V^c)} \sum_{i=1}^{|V|} \log P(w_i^c | w_i)$$

$$J'(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \neq 0}} p(w_{t+j} | w_t ; \theta)$$

- works well with small amount of training data
- represents rare words or phrases well

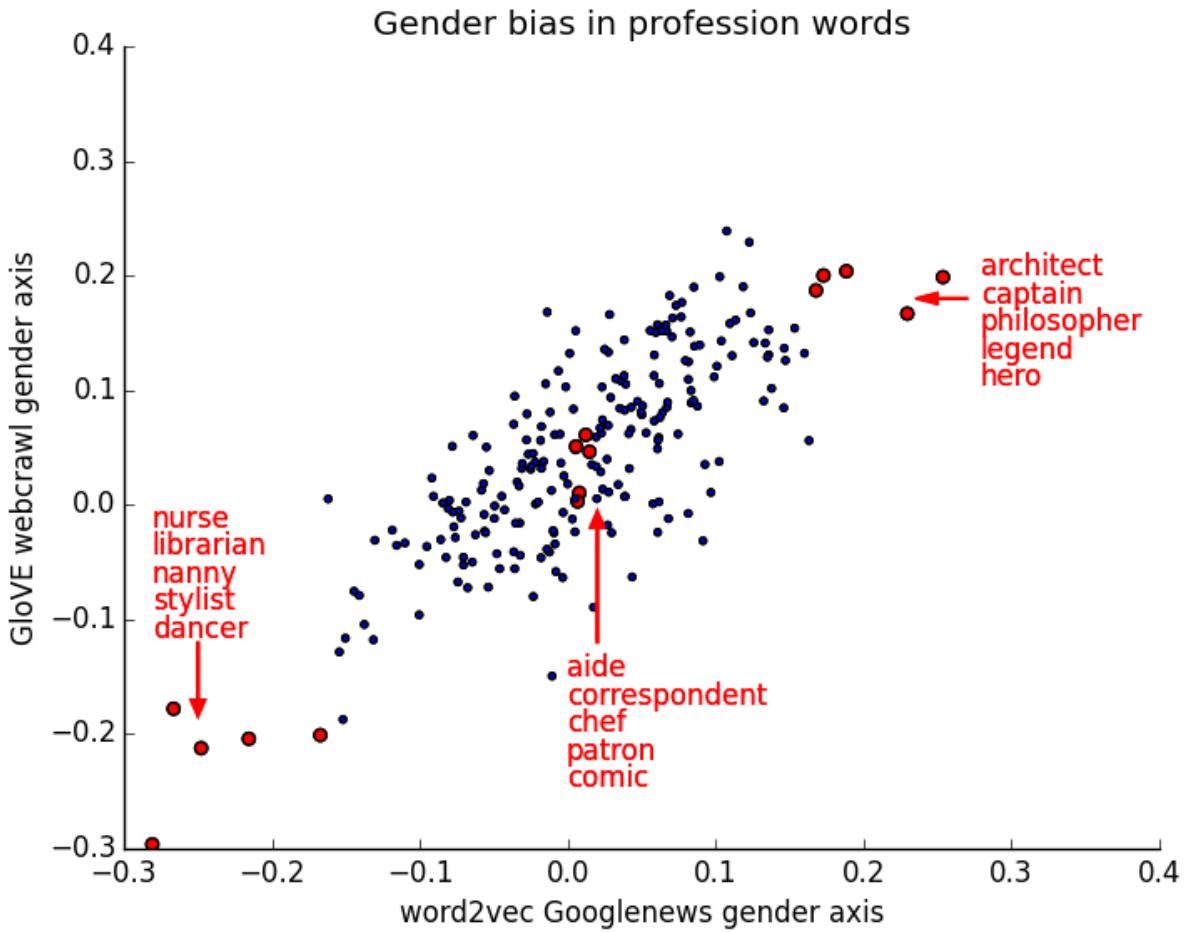
CBOW



- several times faster to train than skip-gram
- slightly better accuracy for frequent words

Hands on

As sexist/racist/classist etc. as the input data



GloVe

- Works on the co-occurrence information
- GloVe is count based

- Performs very well on word analogy tasks
- Good

Link to the GloVe project page
Paper on GloVe

Word2vec vs Glove

Both work on the word co-occurrence information
Word2vec is prediction based model, Glove is count based

Very little difference in performance
Glove is a little easier to parallelize

manipal PROlearn

THANK YOU