Distributional Semantics

LIN 313 Language and Computers
UT AustinFall 2025
Instructor: Gabriella Chronis

Admin

- Test grades posted Wednesday
- Reading "Man is to programmer as woman is to homemaker? Debiasing Word Embeddings" for Monday 11/3
  - focus on sections 1-4 ; skim the rest / don't worry about the math

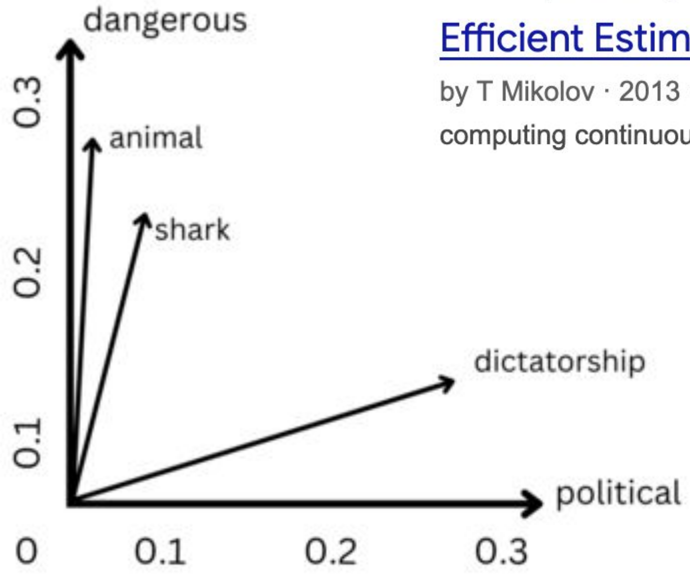
Overview

- introduction to distributional semantic spaces
    - the distributional hypothesis
    - syntagmatic vs. paradigmatic relationships between words
    - build a count-based co-occurence model of word meaning

The Semantic Space of a Neural Network

- When you train a neural network to predict the next word (or the missing word), the model learns some information about the meaning of the words.
- semantic space demo: https://projector.tensorflow.org/
- In a semantic space, the distance between two vectors is corresponds to the semantic similarity between them

Word2Vec (Mikolov et al. 2013)



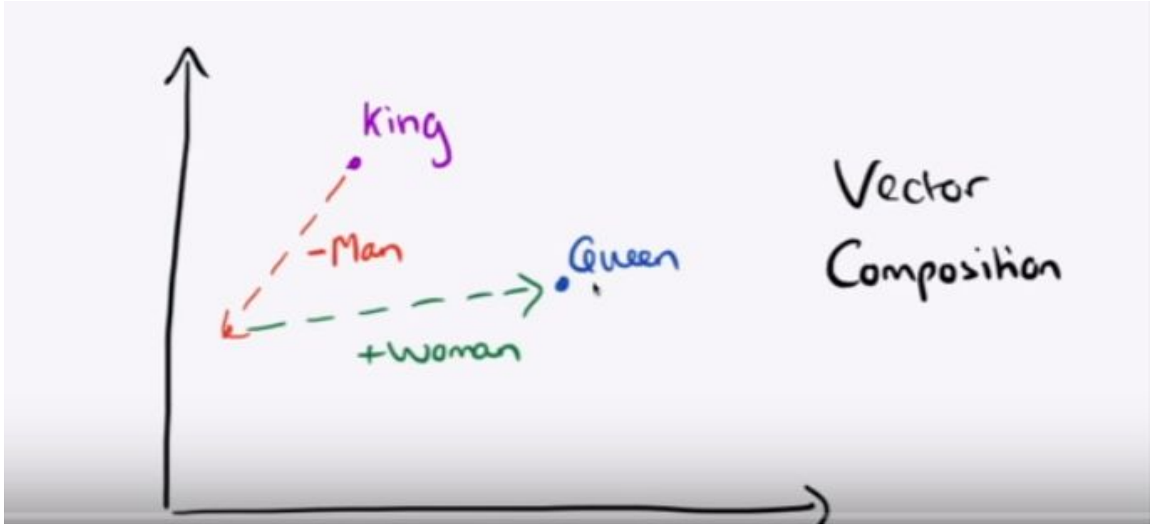
arXiv
https://arxiv.org › cs

[Efficient Estimation of Word Representations in Vector Space]

by T Mikolov · 2013 · Cited by 50129 — Abstract:We propose two novel model architectures for computing continuous vector representations of words from very large data sets.

Analogy Solving with Vectors

- If you take the vector for **king**, subtract the vector for **man,** add the vector for **queen**, you end up at a new point in space.
- When you look around, you find that the closest neighbor in that space is the vector for **queen**



Word2Vec learns geographic relationships

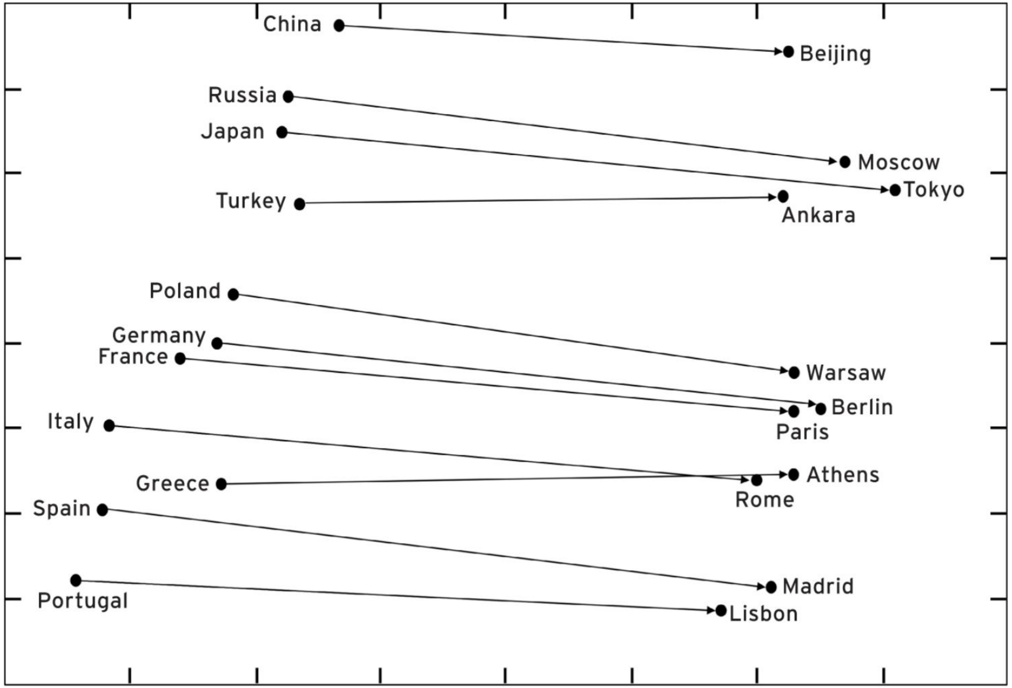


FIGURE 37: Two-dimensional representation of distances between word vectors for countries and word vectors for their capital cities

The Distributional Hypothesis

- Basic idea: words that occur in similar contexts have similar meanings
  - the **distribution** of a word is the sum of (linguistic) contexts in which it occurs

- Zellig Harris
  - *Methods in Structural Linguistics* (1951)
    - focused on doing linguistics with *data*
  - lifelong political activist - leftist zionist, outspoken critic of 1940s attacks on Palestinian Arabs
  - Notable students: Noam Chomsky, Aravind Joshi
- JR Firth
  - studied **collocations -** known to us as n-grams
  - "You shall know a word by the company it keeps." (1957:11)

Wugs, again

Consider the sentence:

The ___stelp___ scampered up the tree.

What is a stelp?

How do you know?

Semantic Relationships: syntagmatic & paradigmatic

It might help at this point to distinguish two kinds of meaning relations between words.

syntagmatic relationships between words arise because of a sequential ordering. Words in a sentence form a syntagm.

- think of the **syntax**, which links words with different functions in a sentence.
- words in syntagmatic relation are likely to occur in the **same sentence**

paradigmatic relationships between words are taxonomical. Words in a **paradigm** are related because they may be substituted for one another in certain contexts. They occur in the same slot in a sentence.

- think of a verb form paradigm in Spanish or a noun declension table in Latin a foreign language
- paradigm words are likely to occur in the **same "slot"** in the sentence

Syntagmatic or paradigmatic? (like everything to do with language, it's always both. But is one stronger?)

lilac - rose

- paradigmatic
- reasoning: we can put them in a list where each member of the list belongs to a shared semantic category, and can occupy the same slot in a sentence
    - "The bouquet smells of {lilac, rose, peony, lilies)

of - the

- syntagmatic
- reasoning: the words often occur in a fixed order in a phrase or sentence; no clear meaning relation
    - "The book of the month" "president of the chess club" "the name of the woman in the yellow hat"

ice - cream

- syntagmatic
- reasoning: the words most often co-occur in the fixed expression "ice cream"

word - sick

- paradigmatic?
- reasoning: depends on the context! (as all of these questions do). In their discourse particle sense, we can imagine them in a paradigm {word, sick, sweet, cool, dope, rad}

Count-based (Co-occurrence) semantic spaces

We want to build a representation of each word in a corpus based on its distribution over contexts. Words with similar distributions are similar!

Steps

1. create a table
2. make a row for each word
3. make a column for each context
    a. count how many times each word appears in each context, and mark it in the cell

Turning an n-gram model into a semantic vector space

See Distributional Semantics notebook

Disadvantages of purely count-based vectors

There are several disadvantages to this approach

- the vectors are ginormous
  - depending on what we count as a context (neighbor words, n-grams, skip-grams [more on these Wednesday], or whole sentences), the dimensions of a word vector can be anywhere from 20K to 200K and up! This is too big to be useful.
  - they are also very **sparse** vectors: they contain mostly 0s
- words are only similar if they appear in the same contexts
  - language is extremely variable. We want a model that knows words are similar if they appear in *similar* contexts.

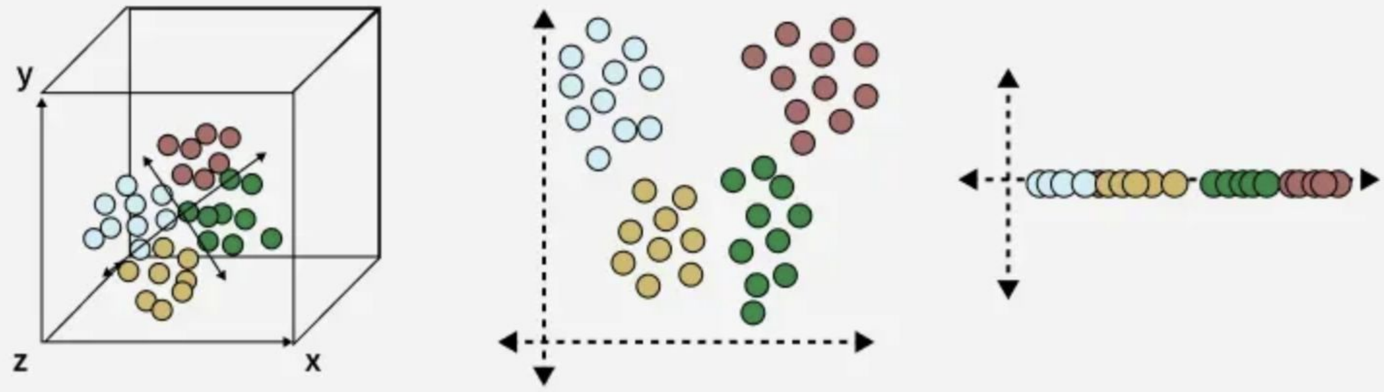
Dimensionality Reduction

There are many techniques for reducing the dimensionality of data

- Latent Semantic Analysis (Landauer & Dumais 1997) uses a matrix factorization technique called SVD (singular value decomposition)
- Other dimensionality reduction techniques
    - Principle Components Analysis (PCA)
    - UMAP
    - T-SNE
- The basic idea is to squish the bulk of the important information into fewer dimensions, while preserving the relationships between vectors as much as possible.

What is Dimensionality Reduction

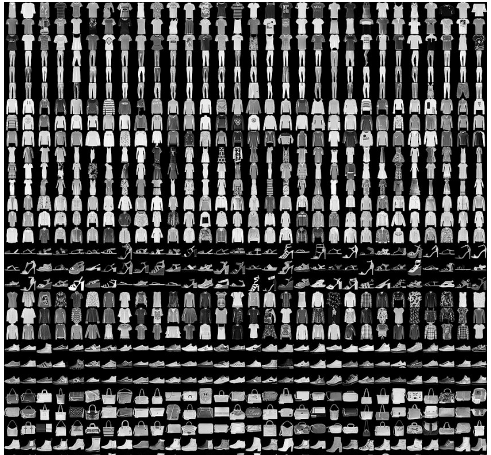
Dimensionality Reduction is the process of reducing the number of input variables (features) in a dataset while preserving as much important information as possible.

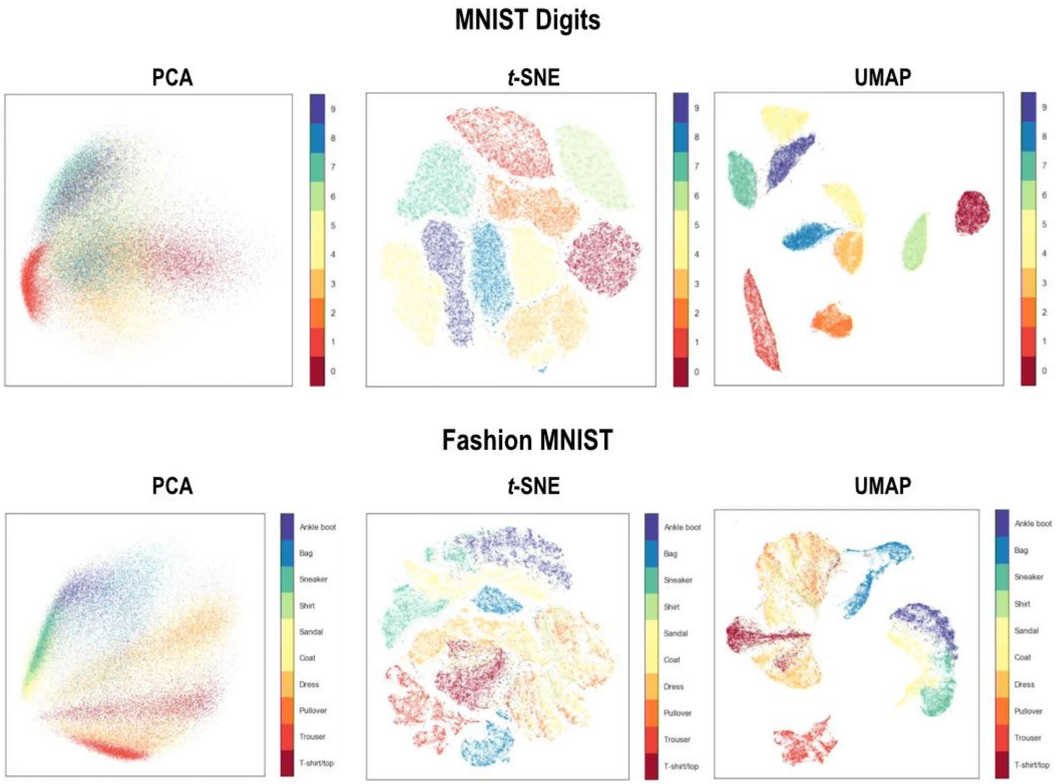


Different dimensionality reduction techniques

MNIST Digits = handwritten digits for character recognition tasks

MNIST Fashion:





Equivalence of Deep-learning embeddings and Count-based Models

- neural networks start with random embeddings and *tune* them during training.
- Q: Why do we bother training neural networks when if can just count and factorize?
- A: with neural networks we never need to hold all of our training data in memory at once (impossible with realistic large datasets!)

Neural Word Embedding as Implicit Matrix Factorization

Omer Levy
Department of Computer Science
Bar-Ilan University
omerlevy@gmail.com

Yoav Goldberg
Department of Computer Science
Bar-Ilan University
yoav.goldberg@gmail.com

Abstract

We analyze skip-gram with negative-sampling (SGNS), a word embedding method introduced by Mikolov et al., and show that it is implicitly factorizing a word-context matrix, whose cells are the pointwise mutual information (PMI) of the respective word and context pairs, shifted by a global constant. We find that another embedding method, NCE, is implicitly factorizing a similar matrix, where each cell is the (shifted) log conditional probability of a word given its context. We show that using a sparse *Shifted Positive PMI* word-context matrix to represent