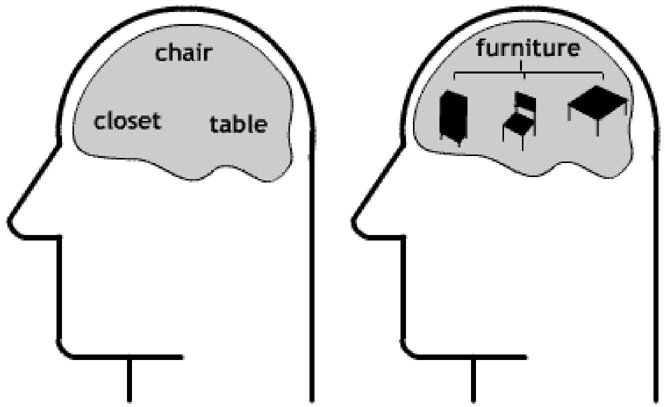
Learning Representations

Mastodon – Aresos – le 28-11-2014

Learning representations

- Objective: learning robust and meaningful representations from data
- Handcrafted versus learned representation
  - Very often complex to define what are good representations
- General methods that can be used for
  - Different application domains
  - Multimodal data
  - Multi-task learning
  - Learning the latent factors behind the data generation
  - Unsupervised feature learning
- Several families of techniques
  - Algebraic and statistical models
  - (Deep) Neural networks

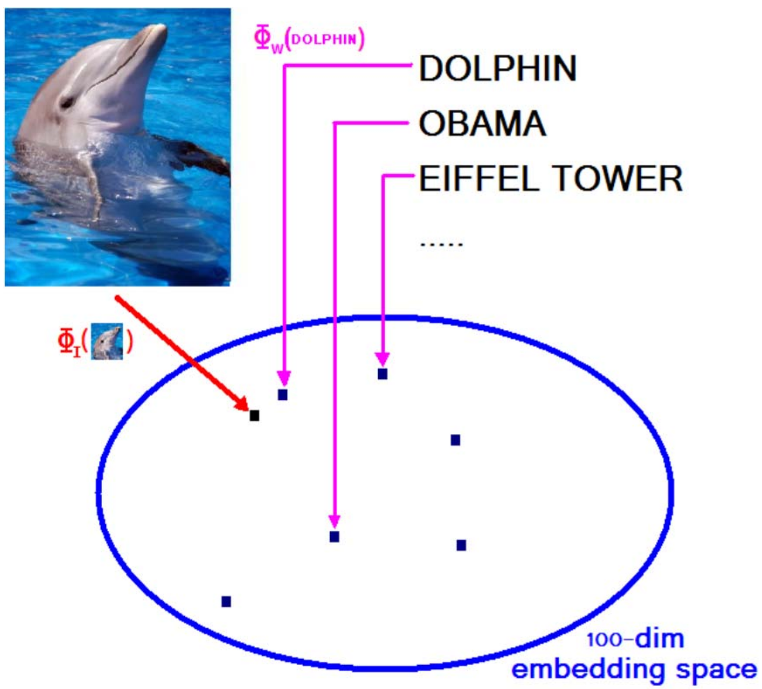


Learning representations - Success story

- Very active recent domain, technology adopted (sometimes already operational) by big actors (Google, Facebook, Msoft ..)
- Success in many academic benchmarks for a large series of different problems
  - Image / scene labeling
  - Speech recognition
  - Natural language processing
  - Language translation
  - etc

Google image search (Bengio et al. 2010, 2011)

- Learning to represent multiple modalities in the same space



Learn $\Phi_I(\cdot)$ and $\Phi_W(\cdot)$ to optimize precision@k.

Illustrations with Neural Networks
Multi-modal information embedding (Socher et al. 2013 – Stanford NLP group)

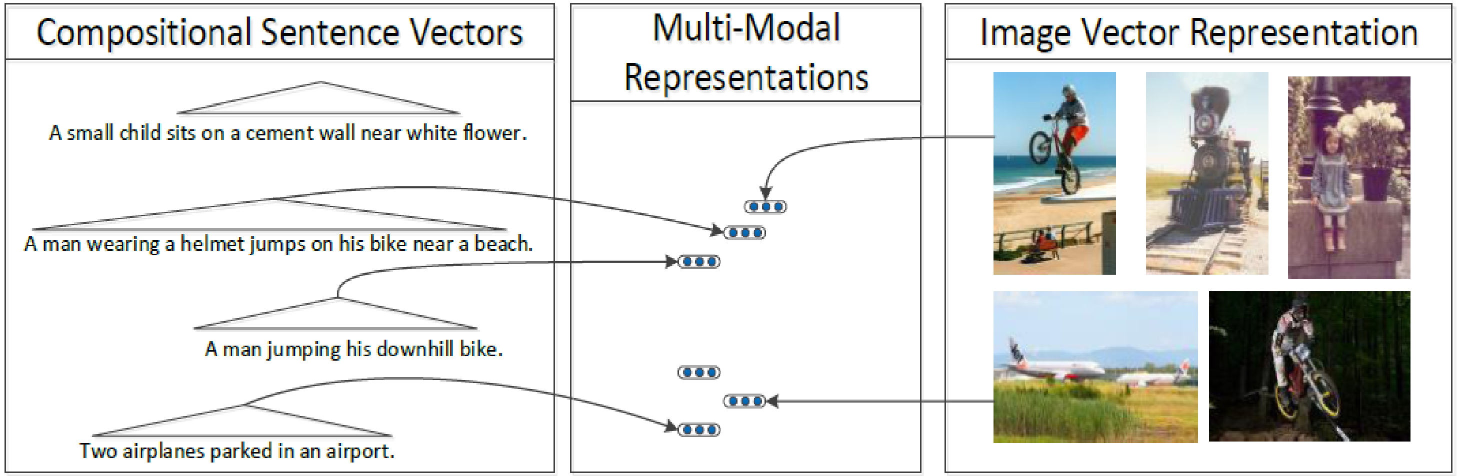
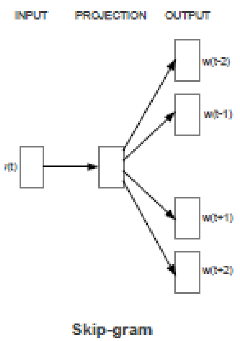


Figure 1: The DT-RNN learns vector representations for sentences based on their dependency trees. We learn to map the outputs of convolutional neural networks applied to images into the same space and can then compare both sentences and images. This allows us to query images with a sentence and give sentence descriptions to images.

Learning Language models (Mikolov et al. 2013)

- Simple neural networks language model and
  - Word2Vec software



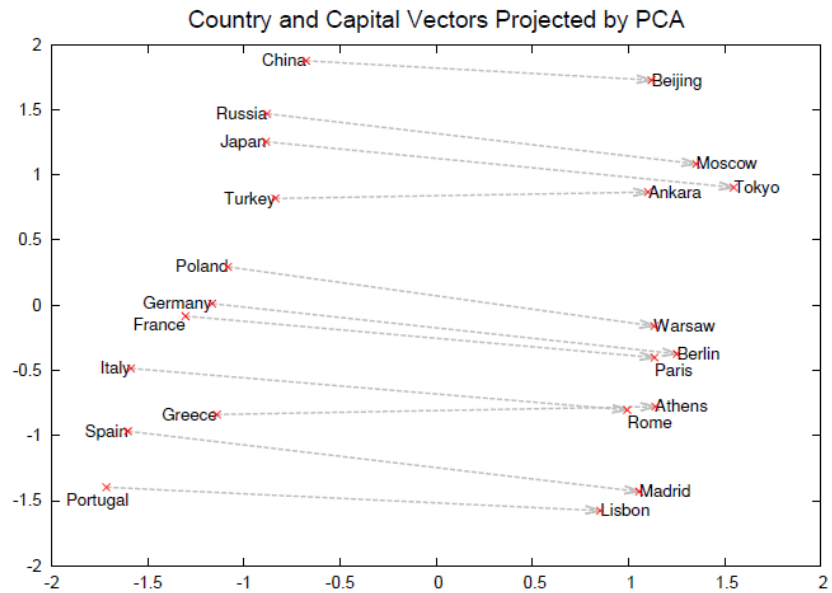


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

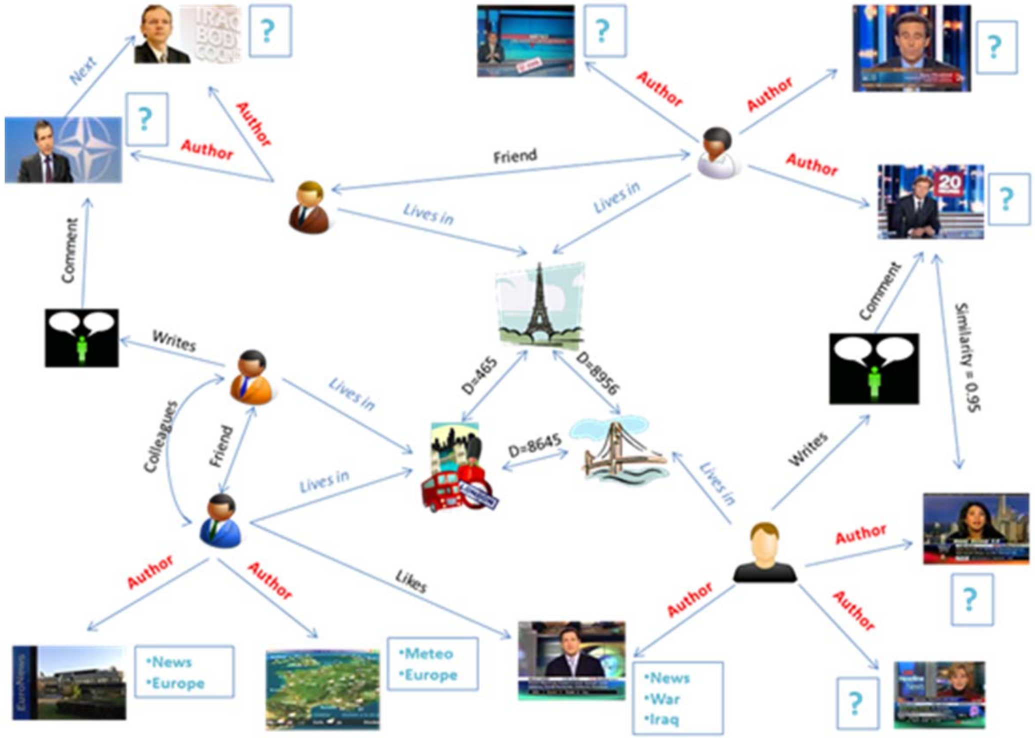
- Analogical reasoning
  - Paris – France + Italy = Rome
  - Discovery of 3 way relations
    - Semantic, syntaxic

Work done in Aresos on representation learning

- Social networks
  - Relational classification
  - Information diffusion
- Natural language processing
  - Semantic compositionality

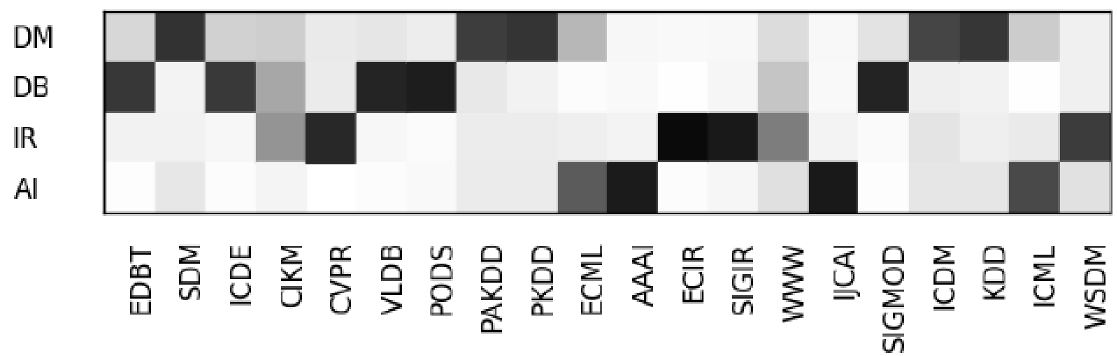
Relational Classification

- Label the items with corresponding tags, classes, …



Relational classification

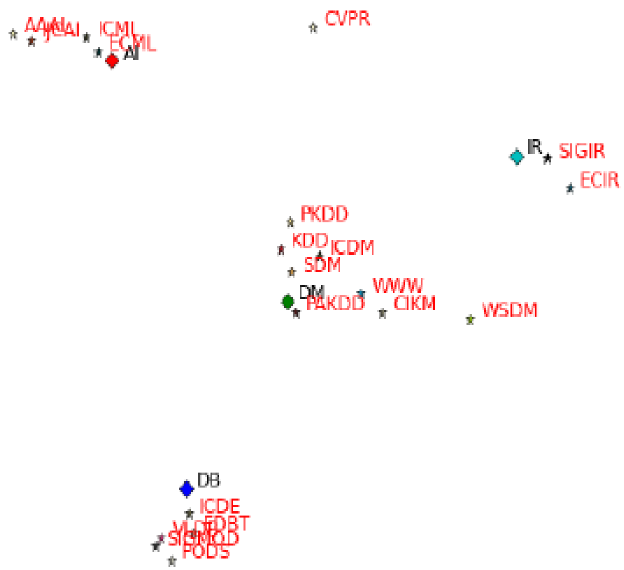
- Correlations among labels from different node types
  - Exemple from DBLP data (Author domains x Conferences)
    - Authors: 4 labels
    - Conferences: 20 labels
    - *P(author domain| conference)* reveals correlations



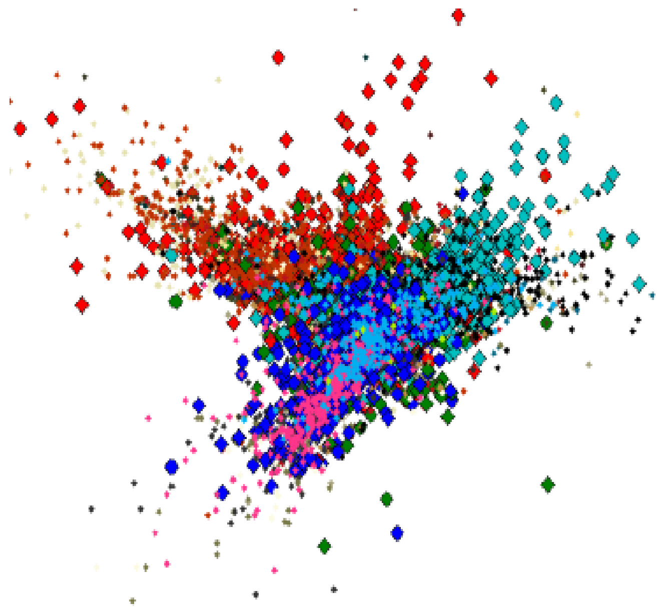
Relational classification

- Latent space PCA for DBLP

Class centroids



All nodes, 1 color = 1 class

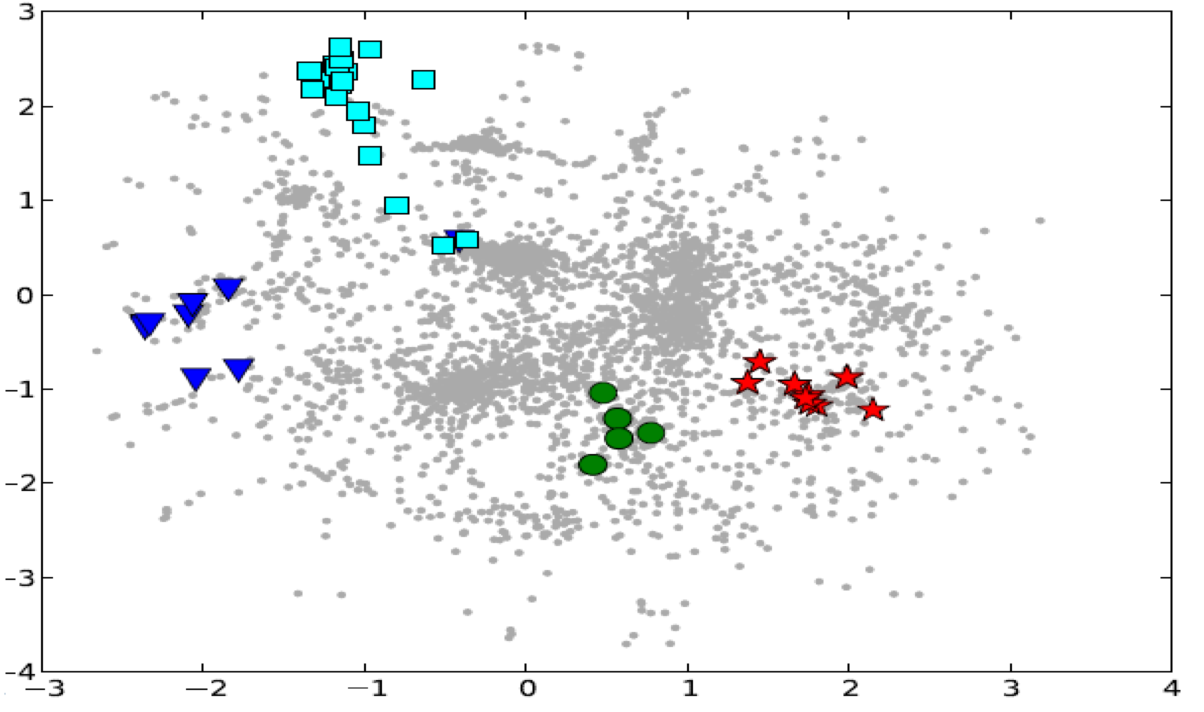


Information diffusion

- Objective
  - Predict information diffusion cascades
- State of the art
  - Several models, often inspired from earlier work in epidemiology and social science
    - Graph propagation model
      - e.g. Independent Cascade model, Linear Threshold model
      - General assumptions
        - Close world, nodes do operate the same way, no features (e.g. content) associated to nodes
- Proposition
  - Learn propagation models directly from observed cascades, without any network assumption
    - All the factors influencing this diffusion (External influences, node roles, etc) are directly extracted from the data
  - Propagation is modeled in a latent space, where the diffusion process follows a simple formalization
    - Map the discrete problem onto a continuous space
    - The latent space is learned from the cascade data
    - Benefits: inference (here diffusion prediction) is extremely fast

Information diffusion

- Illustration
  - Digg dataset
    - User post stories that are digged
    - Diggs = likes
    - Cascades = stories and diggs
    - 1 digg = contamination
    - 1 month crawling
    - 5 k users, 71 k links
    - 150 k Training cascades
    - 66 k test cascades
  - Latent space of size 2
  - User clusters
  - Color points = 4 observed
    Test cascades

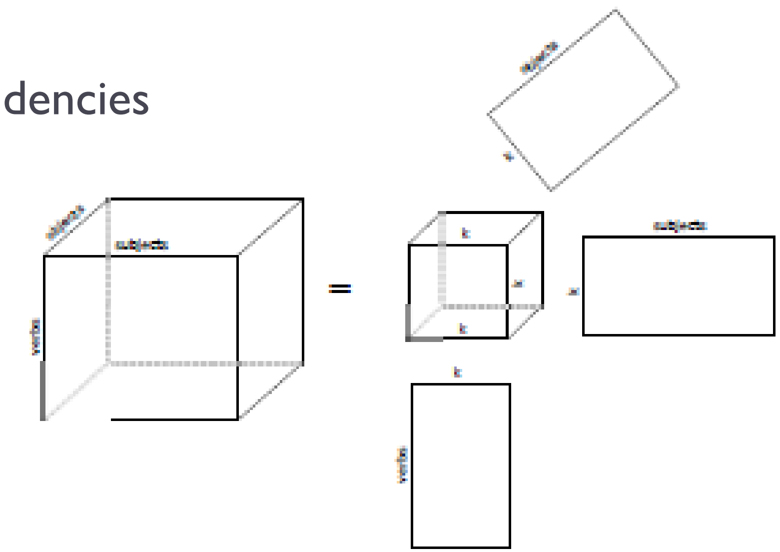


Natural language processing
Tensor model of semantic compositionality (Van de Cruys et al.)

- Compositionality
  - the meaning of a complex expression is a function of the meaning of the parts and the way they are combined
- Distributional hypothesis of meaning
  - Words that appear in the same context tend to be semantically similar
- How to reconcile the principe of compositionality with distributional semantics
- Objective
  - Model compositionality as a multi-way interaction between latent factors learned from the data
  - Task
    - Learn three way interactions (verb, subject, object) VSO from knowledge basis

Natural language processing
Tensor model of semantic compositionality (Van de Cruys et al.)

- Method
  - Make use of tensor decomposition to capture 3 ways dependencies



- What is it good for
  - Learn knowledge basis in a continuous representation form
    - E.g. Freebase, Wikipedia, ect
  - Infer new knowledge
    - Analogical reasoning properties
  - Firt step towards complex NLP/ RI tasks
    - E.g. Question Answering
  - Evaluation
    - Compute similarity scores between SVO phrases

Representation learning + phylomemies

- Internship 2014
- Internship 2014-2015