Deep Learning of Text Representations

Fatih Uzdilli

21.01.2015

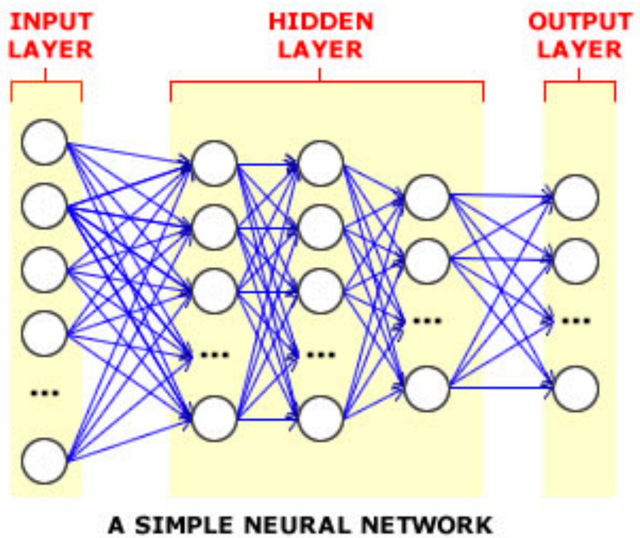
Outline

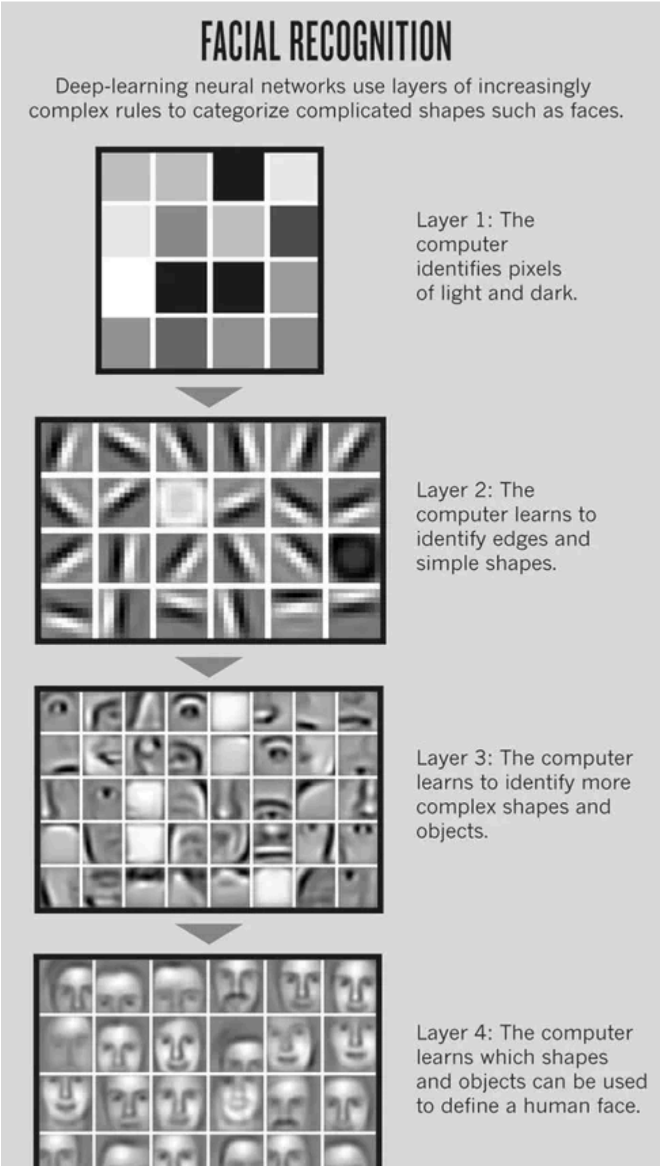
- Deep Learning & Text-Analysis
- Word Representations
- Compositionality
- Results

What is the role of deep learning in text-analysis?

What is deep learning?

- Deep learning algorithms learn multiple levels of representation of increasing complexity/abstraction from raw sensory inputs.





What are raw sensory inputs?

- Images: Intensity for each pixel
- Audio: Aplitude at each time point
- Text: ???

Machine learning (for text) until now

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

- Human designed representation and input features
- Machine Learning => often linear models, just optimizing weights

Good input features are essential for successful ML!

(feature engineering = 90% of effort in industrial ML)

Bag-of-Words Feature

- The most simple approach for text
- Also called: Unigram
- 80/20-rule‘s perfect example

- How To:
  - Vector of lenght |Vocabulary|
  - Every index represents one word
  - For each word occuring in text: set value at index of word to 1

- Can‘t distinguish:

+ White blood cells destroying an infection

- An infection destroying white blood cells.

Some state-of-the-art features for sentiment detection for tweets

Zürcher Hochschule für Angewandte Wissenschaften

School of Engineering

InIT Institut für angewandte Informationstechnologie

- N-gram (n=1-4)
- N-gram with lemma (n=1-4)
- N-gram with removing middle word(s) (n=1-4)
- N-gram using word clusters (n=1-4)
- Substrings-n-grams
- POS-n-grams (n-grams with middle words replaced by POS-tag)
- Encoding negation context into words
- Number of all capitalized words
- Number of hashtags
- Number of POS-tags
- Number of words in a negated context
- Number of elongated words
- Text ends with punctuation
- Length of longest continous punctuation
- Is last word in negative words list
- Is last word in positive words list
- Sentiment lexicon score for last token
- Total sentiment lexicon score of all tokens
- Maximum sentiment lexicon score of all tokens
- Number of tokens having positive sentiment lexicon score

Stats about resulting feature vector:

- Vector-Size: 2.1Mio
- Avg Non-Zero-Values: ~1100
- => Very very sparse

Promlem 1: Handcrafting Features

- Problem:
  - Handcrafting features is time-consuming
  - Needs experience, you have to be good
  - Has to be done for each task/domain

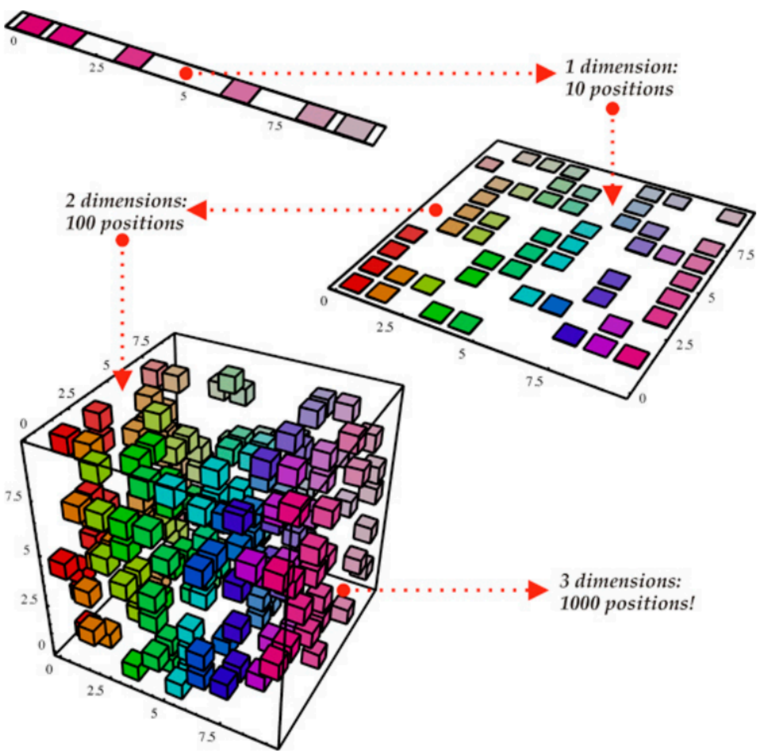
- Alternative:

  Representation Learning:

  let the machine learn good feature representations

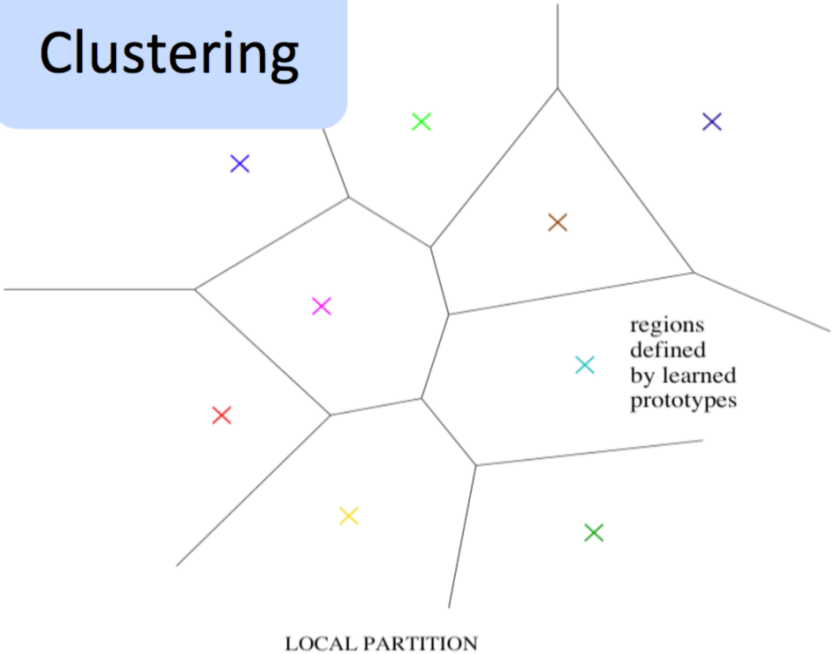
Problem 2: Curse of Dimensionality

- Problem:
  - Current Natural-Language-Processing systems are fragile because of their atomic symbol representations
    - „He is smart“ vs. „he is brilliant“
  - Curse of dimensionality: to generalize, we need all relevant variations => more dimensions than variations available!

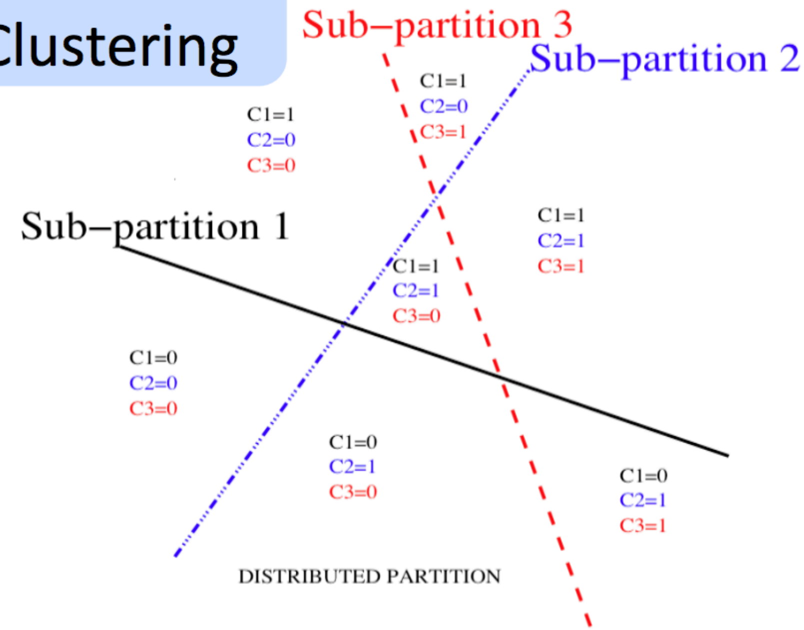


We need Distributed Representations

Clustering



Multi-Clustering

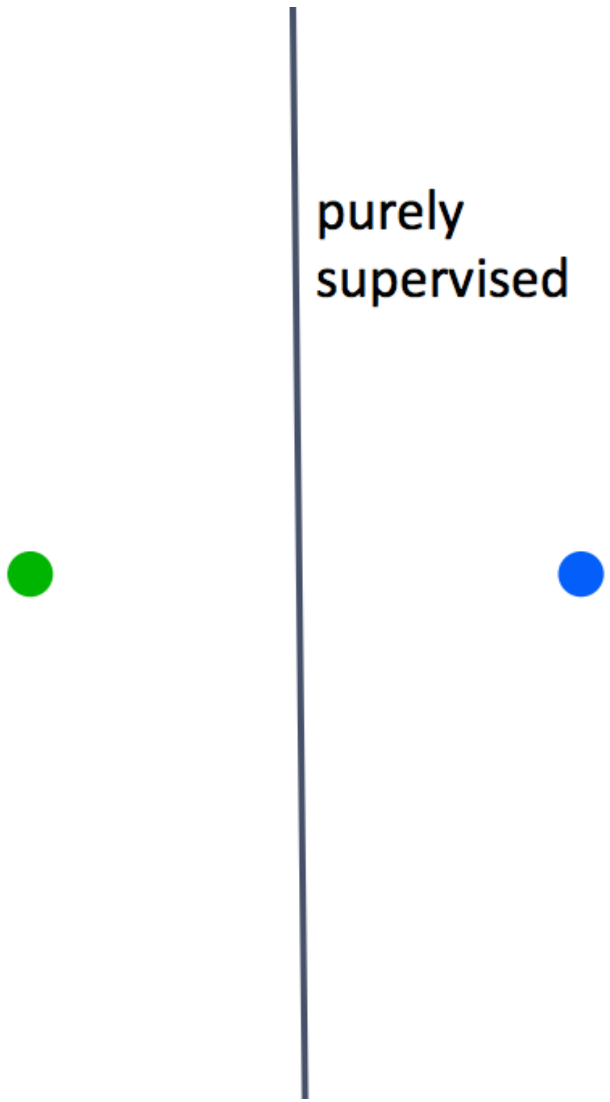


- Distributed Representations:
  - represent multiple dimensions of similarity, non-mutually exclusive features => exponentially large set of distinguishable configuration

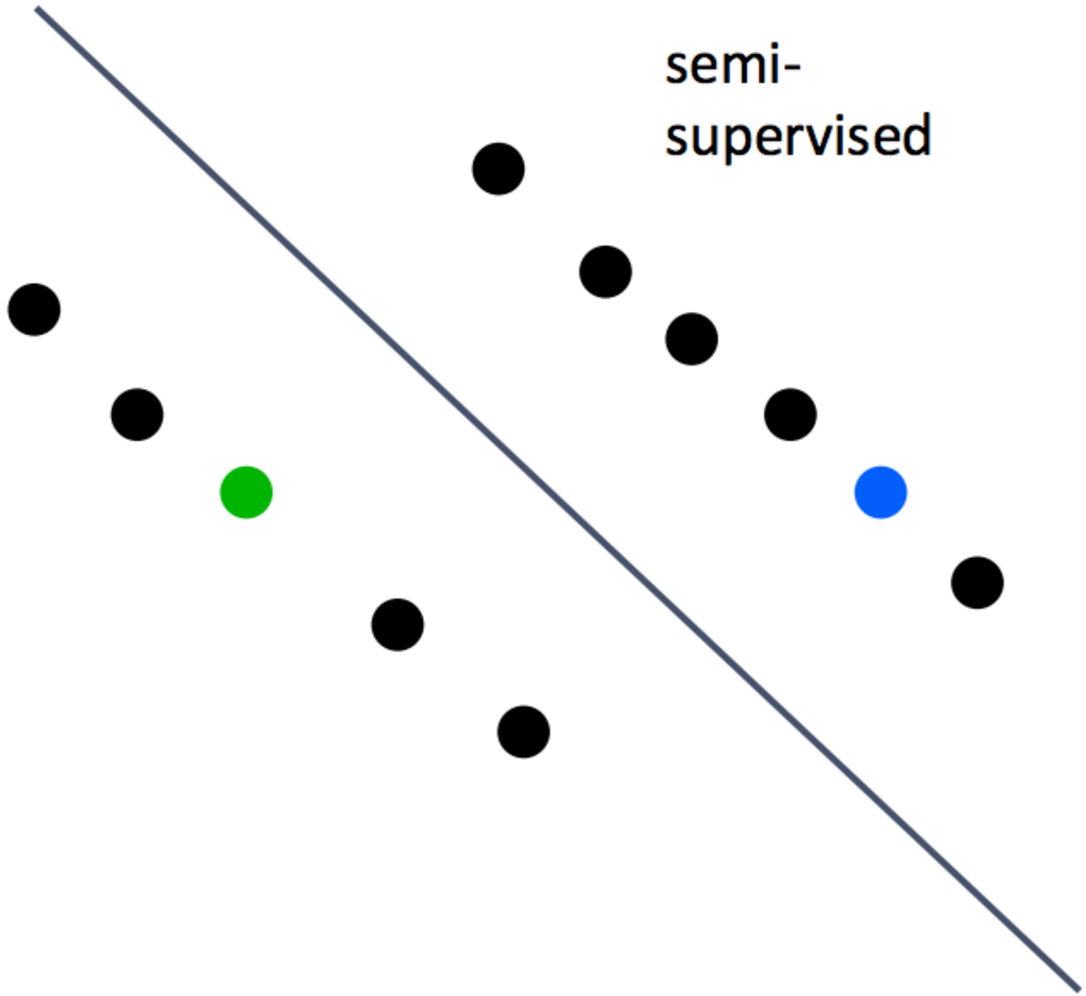
Problem 3: Not enough labeled data

- Problem:
  - Most methods require labeled training data (i.e. supervised learning) but almost all data is unlabeled
- Alternative:
  - Unsupervised feature learning

Purely subervised setup



Semi-Supervised setup



Let‘s start with word representations

Neural Word Embeddings as a Distributed Representation

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering
InIT Institut für angewandte Informationstechnologie

- Using large amount of data
- Similar idea to soft clustering models like LSI, LDA
- Allows adding more supervision from multiple tasks
  → can become more meaningful
- Word is represented as a dense vector

$$\text{linguistics} = \begin{bmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{bmatrix}$$

Word Embeddings Visualization

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

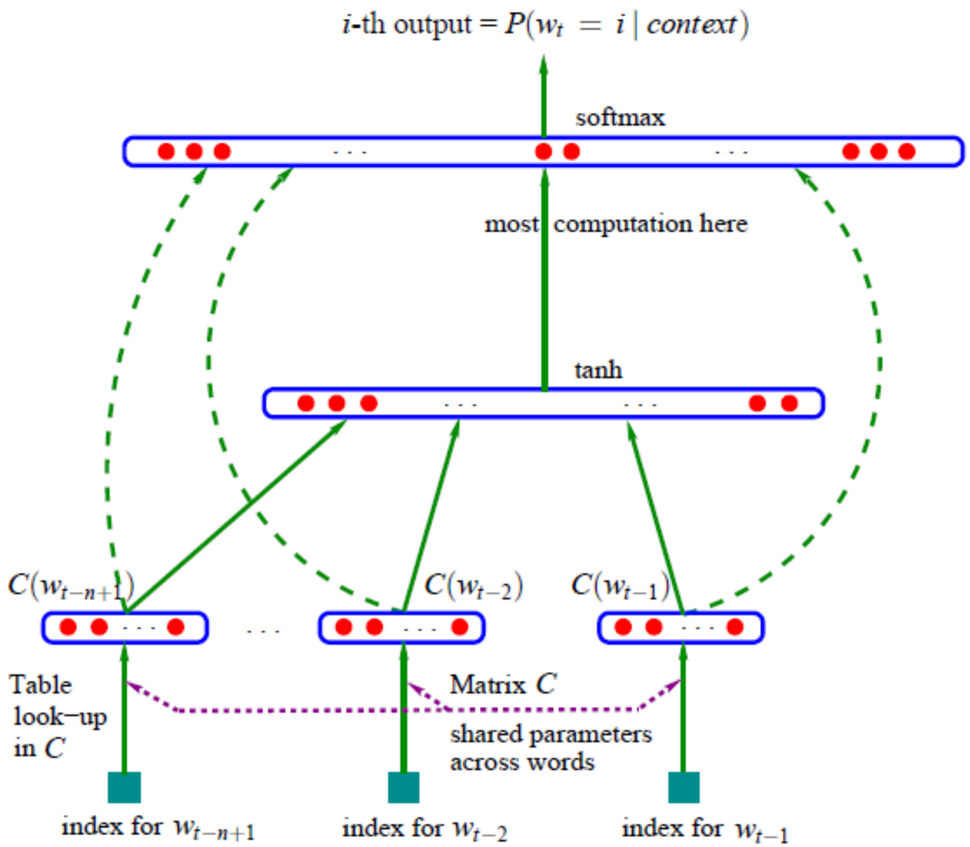
Vector Operations for Analogy Testing

- Syntactically:
  - $X_{apple} - X_{apples} \approx X_{car} - X_{cars} \approx X_{family} - X_{families}$
- Semantically:
  - $X_{shirt} - X_{clothing} \approx X_{chair} - X_{furniture}$
  - $X_{switzerland} - X_{zurich} + X_{istanbul} \approx X_{turkey}$

A Neural Probabilistic Language Model

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering
InIT Institut für angewandte Informationstechnologie

- Y. Bengio, 2003
- Task: Given a sequence of words in a window, predict next word
- Input vectors also involved in backpropagation

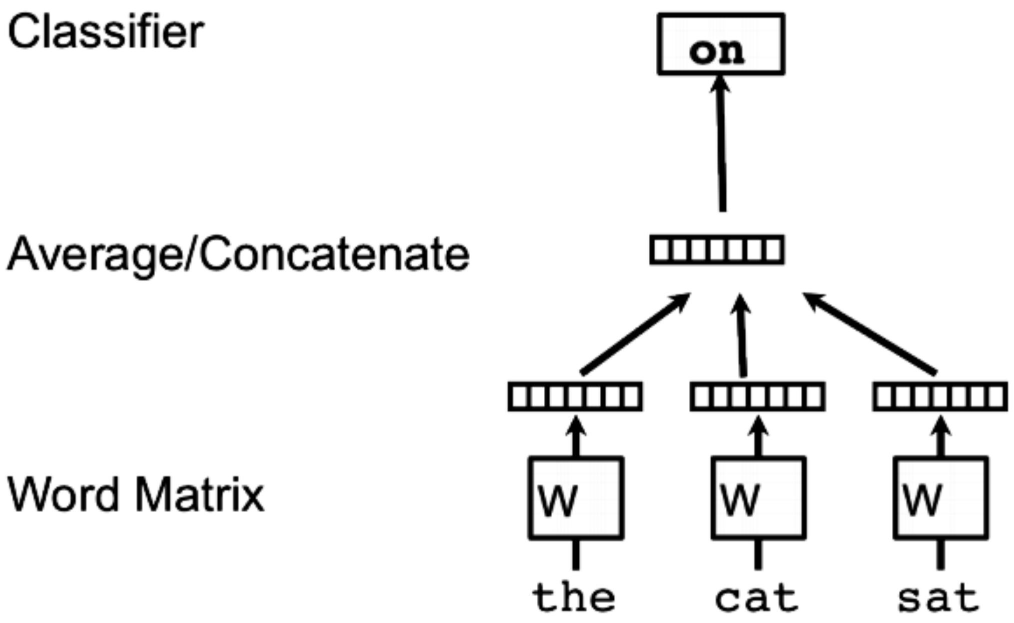


← Hierarchical Softmax
Moring & Bengio, 2005

Speedup: Number of output nodes to update every step shrinked to log(n)

Word2Vec (Continuous Bag-Of-Words)

- Mikolov et. al, 2013
- Speedup: Neural language model with hidden layer removed



Speedup+Quality-Improvement:
- Negative Sampling

Quality-Improvement:
- Subsampling

Still in use:
- Hierarchical Softmax

Other approaches for word embeddings

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering
InIT Institut für angewandte Informationstechnologie

- Predict whether a given sequence exists in nature (Collobert et al. 2011)
  - Example:
    - the cat chills on a mat ✔
    - the cat chills hello a mat ✗
  - Negative examples created by replacing middle word in a window with random word

- Other ideas:
  - Add standard nlp tasks such as POS-Tagging, Named Entity Recognition (NER) etc.

Word2Vec Demo

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

DEMO

Let‘s go to a higher level: Compositionality

Phrase representations by summing up word vectors

- $X_{south} + X_{africa} =$ „X_{south_africa}"
- $\rightarrow X_{south} + X_{africa} - X_{africa} + X_{europe} \approx X_{germany}$

- $X_{new} + X_{york} =$ „X_{new_york}"

- Works well for up to 3-grams

What about whole sentences?

Sentence Representation is difficult because of the varying size!!

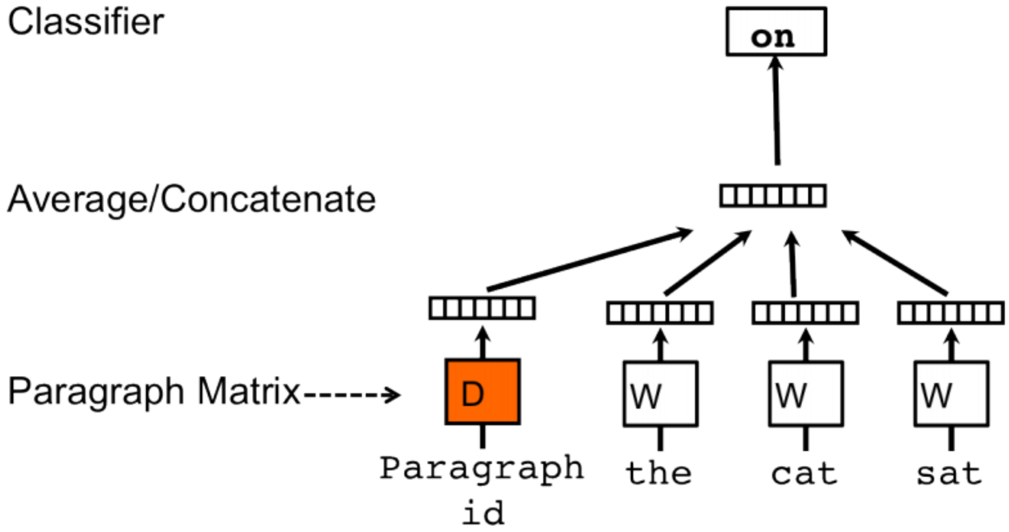
No final answer for sentences yet

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

- Some Mentionable Approaches:
  - Convolutional Networks with Max-Over-Time-Pooling (Collobert&Weston, 2008, 2011)
  - Recursive Neural Networks (Socher et. al, 2010)
  - Recursive Autoencoder (Socher et. al, 2011)
  - Recursive Neural Tensor Networks (Socher et. al, 2013)
  - Paragraph Vector based on Word2Vec (Le&Mikolov, 2014)
  - Convolutional Networks with with various pooling schemes, regularizations (Kim, 2014) (Kalchbrenner et. al., 2014)

Paragraph Vector

- Le & Mikolov, 2014
- Add an additional vector for each sentence/document during word2vec training to learn vectors for sentence/document.

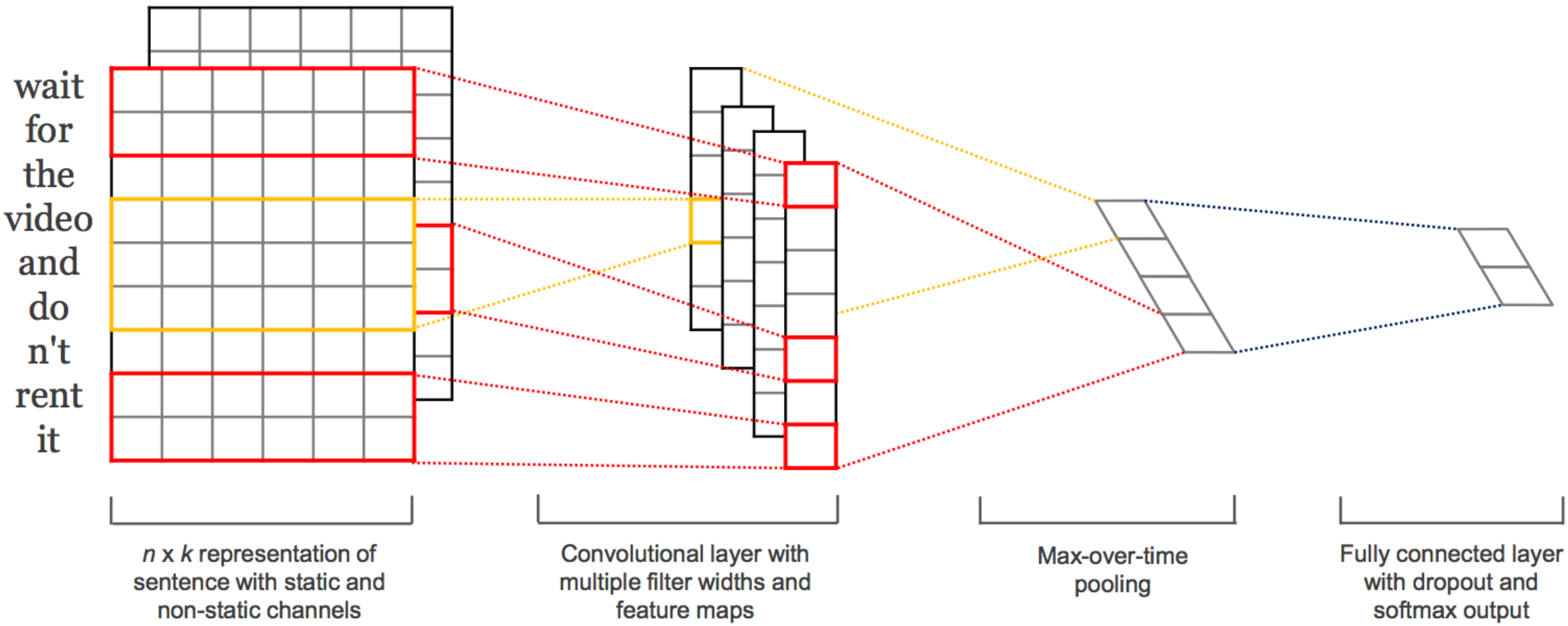


Slow on test-time!

Convolutional Neural Networks for Sentence Classification

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

- Yoon Kim, 2014



Results

Results on SemEval2014 Shared Task 9

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

Sentiment (3-class)-Classification Task on Twitter Data

	Deep Learning Part	Classical Features Part	Final Score
Best System	-	70.96	70.96
Coooolll	66.86	67.07	70.14
Think Positive	67.04	-	67.04

For practical uses deep learning has been just a provider of one additional feature !

Results on Sentiment Treebank Dataset

Sentiment of movie review sentences, label provided on each sub-phrase for training

	5 class (++, +, o, -, --)	2 class (+, -)
Bag-Of-Words + SVM	40.0	82.2
Feature Engineered **Twitter**-Sentiment-Classifier (2013)	43.7	84.1
RAE (Socher et. al., 2011)	43.2	82.4
MV-RNN (Socher et al., 2012)	44.4	82.9
RNTN (Socher et al., 2013)	45.7	85.4
DCNN (Kalchbrenner et al., 2014)	48.5	86.8
Paragraph-Vec (Le and Mikolov, 2014)	**48.7**	87.8
CNN (Kim, 2014)	47.4	**88.1**

2014‘s results of deep learning systems seem to be useful in their own

BUT: Deep learning systems with good results are difficult to reproduce.

Summary

- Semi-supervised distributed representation learning shows future direction
- Word representation „easy“
- Sentence representation still ongoing issue
- Good results difficult to reproduce

Further Reading

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

- 2013 – Mikolov et. al - Efficient Estimation of Word Representations in Vector Space: http://arxiv.org/pdf/1301.3781.pdf
- 2014 – Quoc & Mikolov - Distributed Representations of Sentences and Documents: http://cs.stanford.edu/~quocle/paragraph_vector.pdf
- 2014 – Y. Kim – Convolutional Neural Networks for Sentence Classification: http://emnlp2014.org/papers/pdf/EMNLP2014181.pdf