Deep Learning for Text

From Word Embeddings to Convolutional Neural Networks

Martin Jaggi

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SPINNINGBYTES

SwissText Conference, 8th June 2016

Natural Language Processing

- Numerous applications with huge impact:
  - Search - access to information
  - Question answering - access to knowledge
  - Machine translation - bridge multi-linguality
  - Machine reading & summarization - essence of text
  - Conversational agents - talk the talk
- ... we are only at the beginning!

[ETH NLP Class, Thomas Hofmann 2016]

Outline

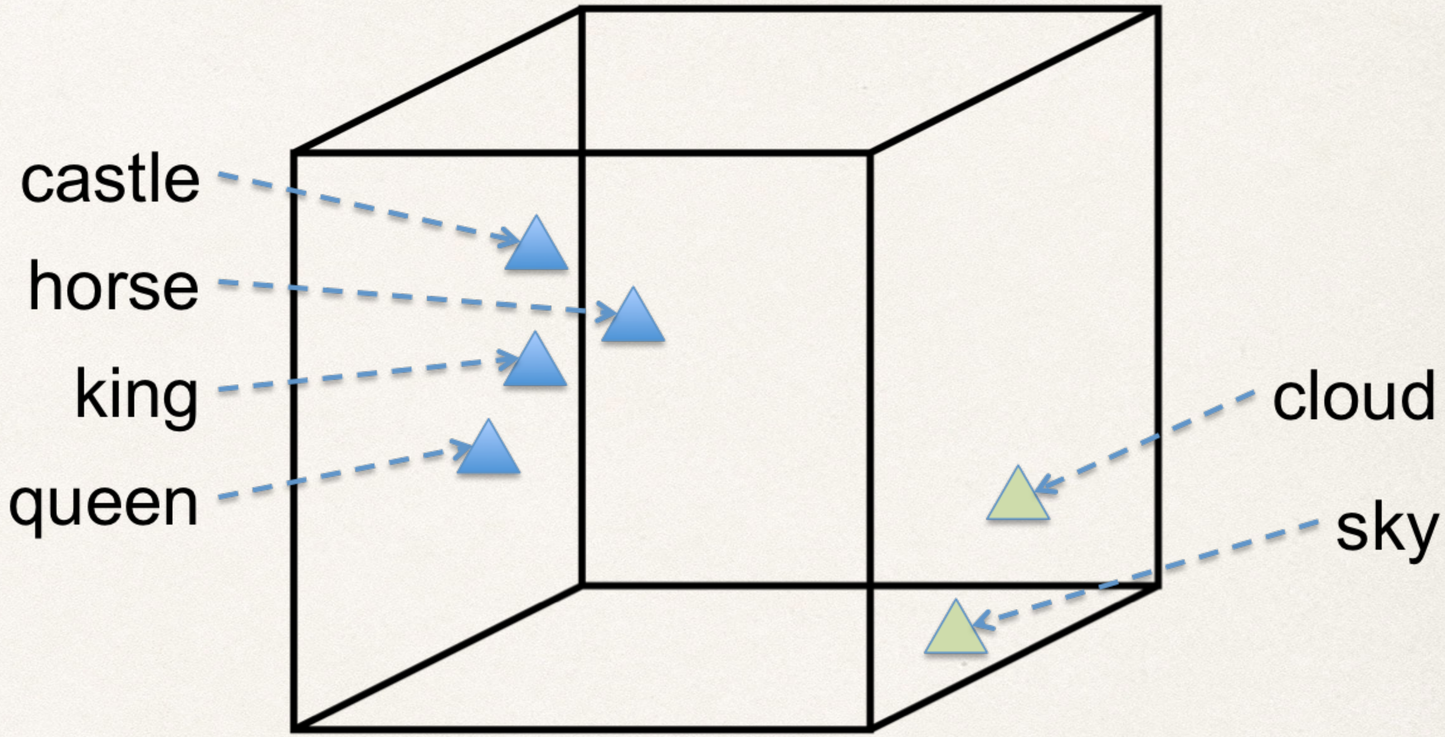
- Semantic Text Representations
  - Word Embeddings
  - Document Embeddings
  - Applications of Machine Learning to Text

From Words to Features

- Bag of words representation

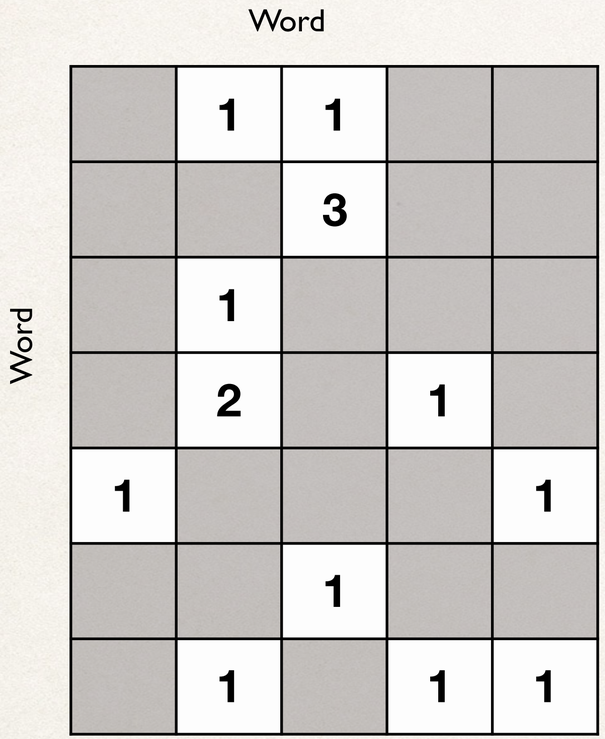
$$i \quad \longrightarrow \quad (0, .., 1, .., 0) \in \mathbb{R}^{1M}$$

Word Embeddings



$$i \longrightarrow \boldsymbol{v}_i \in \mathbb{R}^{50}$$

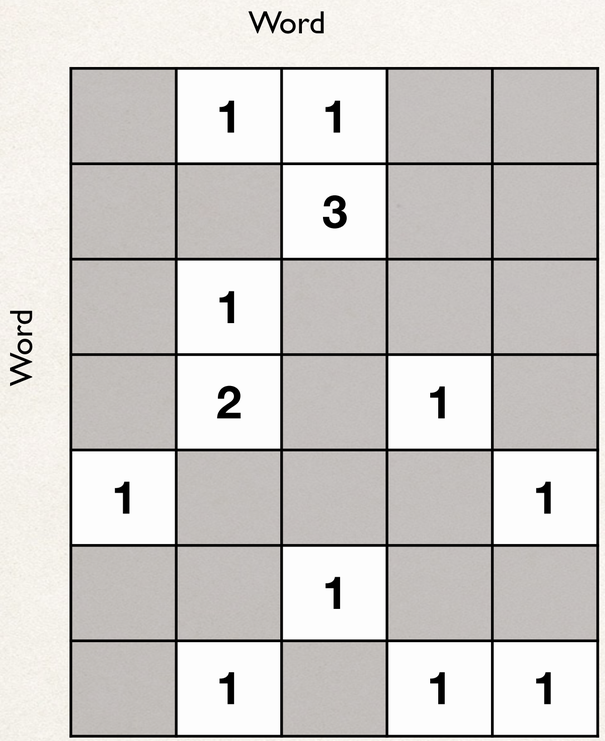
Word Embeddings



explain co-occurence *i,j*
by means of

$$\boldsymbol{v}_i^\top \boldsymbol{v}_j$$

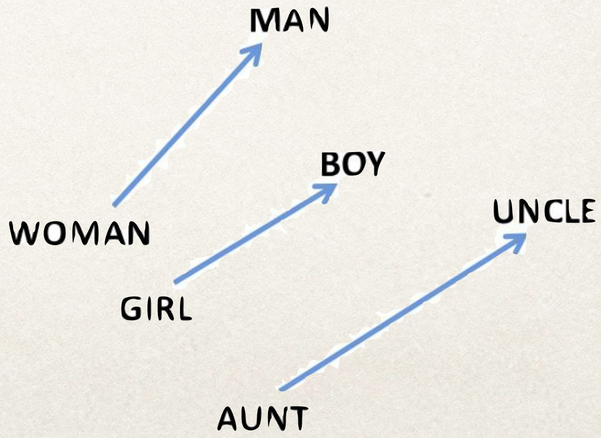
Word Embeddings



$\approx$

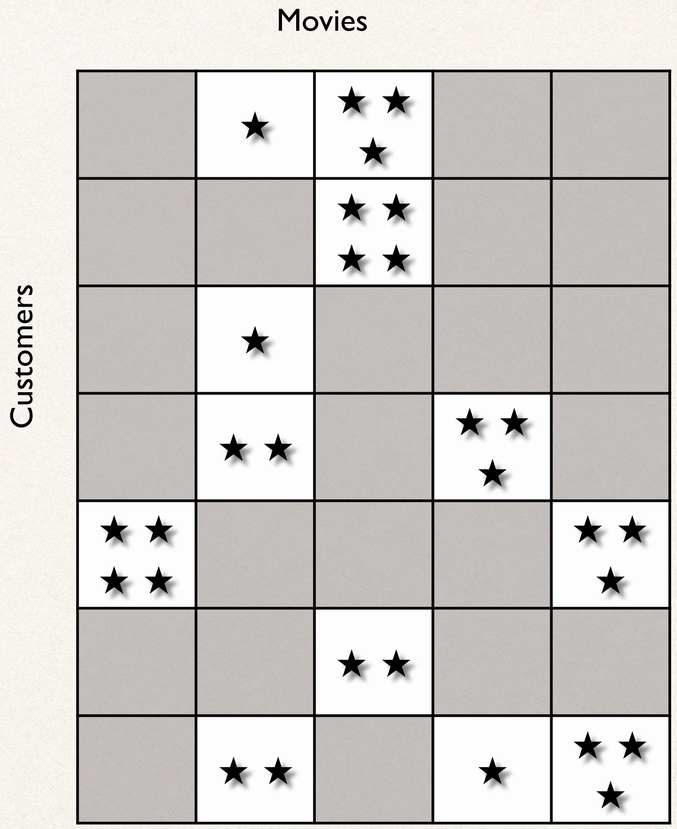
V $V^\top$

SVD, PLSA etc
word2vec, gloVe



spinningbytes.com/demos

Word Embeddings



$$\approx UV^T$$

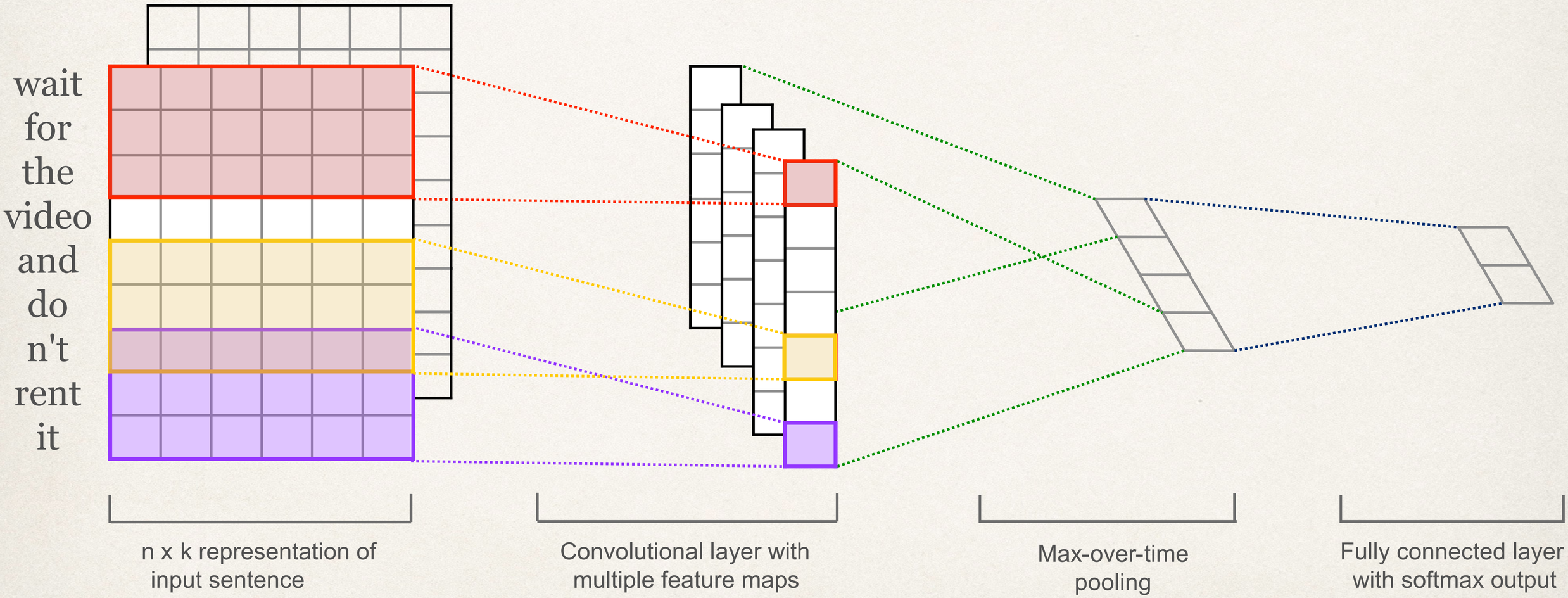
Word Embeddings - Summary

- Very successful new variation of an old theme
- State of the art feature representations for words
- Not related to deep learning
- Parallelization still challenging
- Limited to represent words or short n-grams

Document Embeddings

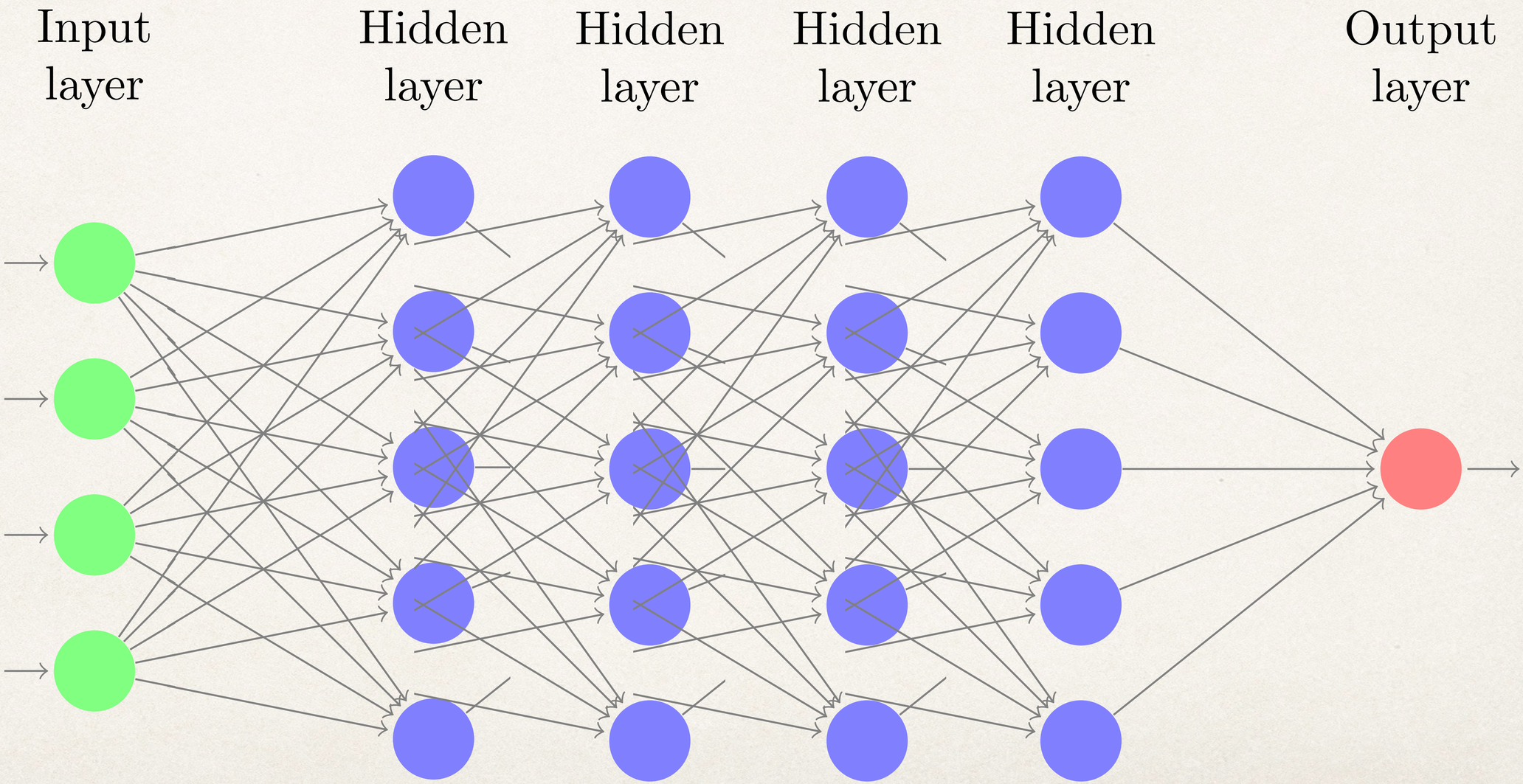
- How to represent a sequence of words?

A Two-Stage Approach



adapted from [Kim 2014]

Neural Networks



Alternative Document Embeddings

- Convolutional Neural Networks (CNN)
- Long Short-Term Memory (LSTM) Networks
- paragraph2vec / doc2vec

Application: Sentiment Classification

- A state-of-the art system for text classification
- Two ETH Master Theses by

  Jan Deriu & Maurice Gonzenbach

ETH zürich zh aw

Results

- **SemEval** Competition
  - running since 1998
  - new set of manually annotated tweets every year
- Our Entries in the Sentiment Competition
  - 2016 **1st place** (Convolutional NN, ensemble)
  - 2015 8th place (SVM, lexica, ensemble)
  - 2014 8th place (SVM, lexica, ensemble)

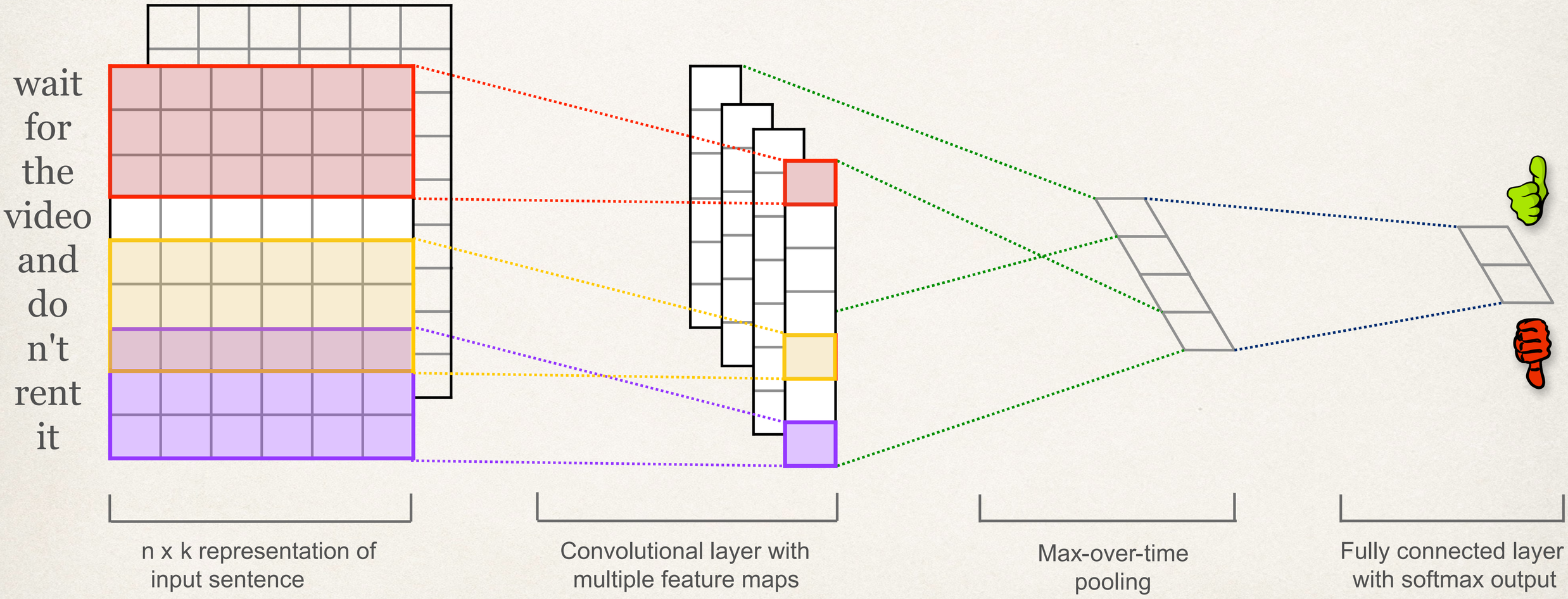
Results

- **SemEval** Competition
  - running since 1998
  - new set of manually annotated
- Our Entries in the Sentiment Comp
  - 2016 **1st place** (Convolutional N
  - 2015 8th place (SVM, lexica, ens
  - 2014 8th place (SVM, lexica, ens

negative	*neutral*	But i wanna wear my Concords tomorrow though but i don't feel like it
positive	*neutral*	Gonna watch Grey's Anatomy all day today and tomorrow(:
negative	*neutral*	@CoachVac heey do you know anything about UVA's fallll fest loll they
neutral	*neutral*	@DustyEf when that sun is high in that Texas sky, I'll be buckin it to co
neutral	*positive*	Up 20 points in my money league with Vernon Davis and L. Fitz still to
neutral	*positive*	DEEJAYING this FRIDAY in THE FIRST CHOP it's CHRIS actual SMITH w
negative	*negative*	The Rick Santorum signing that was scheduled for tomorrow at the Boo
positive	*neutral*	@dreami9 lol yep looks like it! Was after El Clasico on Sunday. I didn't
neutral	*neutral*	Back in Stoke on Trent for the 2nd time today!
neutral	*neutral*	First Girls Varsity Basketball Game tomorrow at 6:00 pm Then Football
neutral	*neutral*	#UFC lightweights @Young__Assassin VS @jamievarner set for TUF 16
neutral	*neutral*	@OOOOO_WEEEE slide thru sometime this weekend ill have somethin
negative	*negative*	@DannyB618 Sure absolutely-- I meant out of the Bachmann, Perry, S
negative	*negative*	@RichardGordon48 re Levein discussion on Wed. Can't keep changing
neutral	*neutral*	Today In History November 02, 1958 Elvis gave a party at his hotel bef
neutral	*positive*	Hustle cause you got to then kick back n party everyday like its Fri
positive	*positive*	I can't sleep. Way too exited about Vancouver tomorrow! I'm like a kid

#	System	2013 Tweet	2013 SMS	2014 Tweet	2014 Tweet sarcasm	2014 Live-Journal	2015 Tweet	2016 Tweet
1	SwissCheese	0.700_{4}	0.637_{2}	0.716_{4}	0.566_{1}	0.695_{7}	0.671_{1}	$\mathbf{0.633}_{1}$
2	SENSEI-LIF	0.706_{3}	0.634_{3}	0.744_{1}	0.467_{8}	0.741_{1}	0.662_{2}	$\mathbf{0.630}_{2}$
3	UNIMELB	0.687_{6}	0.593_{9}	0.706_{6}	0.449_{11}	0.683_{9}	0.651_{4}	$\mathbf{0.617}_{3}$
4	INESC-ID	0.723_{1}	0.609_{6}	0.727_{2}	0.554_{2}	0.702_{4}	0.657_{3}	$\mathbf{0.610}_{4}$
5	aueb.twitter.sentiment	0.666_{7}	0.618_{5}	0.708_{5}	0.410_{17}	0.695_{7}	0.623_{7}	$\mathbf{0.605}_{5}$
6	SentiSys	0.714_{2}	0.633_{4}	0.723_{3}	0.515_{4}	0.726_{2}	0.644_{5}	$\mathbf{0.598}_{6}$
7	I2RNTU	0.693_{5}	0.597_{7}	0.680_{7}	0.469_{6}	0.696_{6}	0.638_{6}	$\mathbf{0.596}_{7}$
8	INSIGHT-1	0.602_{16}	0.582_{12}	0.644_{15}	0.391_{23}	0.559_{23}	0.595_{16}	$\mathbf{0.593}_{8}$
9	TwiSE	0.610_{15}	0.540_{16}	0.645_{13}	0.450_{10}	0.649_{13}	0.621_{8}	$\mathbf{0.586}_{9}$
10	ECNU (*)	0.643_{9}	0.593_{9}	0.662_{8}	0.425_{14}	0.663_{10}	0.606_{11}	$\mathbf{0.585}_{10}$
11	NTNUSentEval	0.623_{11}	0.641_{1}	0.651_{10}	0.427_{13}	0.719_{3}	0.599_{13}	$\mathbf{0.583}_{11}$
12	MDSENT	0.589_{19}	0.509_{20}	0.587_{20}	0.386_{24}	0.606_{18}	0.593_{17}	$\mathbf{0.580}_{12}$
	CUFE	0.642_{10}	0.596_{8}	0.662_{8}	0.466_{9}	0.697_{5}	0.598_{14}	$\mathbf{0.580}_{12}$
14	THUIR	0.616_{12}	0.575_{14}	0.648_{11}	0.399_{20}	0.640_{15}	0.617_{10}	$\mathbf{0.576}_{14}$
	PUT	0.565_{21}	0.511_{19}	0.614_{19}	0.360_{27}	0.648_{14}	0.597_{15}	$\mathbf{0.576}_{14}$
16	LYS	0.650_{8}	0.579_{13}	0.647_{12}	0.407_{18}	0.655_{11}	0.603_{12}	$\mathbf{0.575}_{16}$
17	IIP	0.598_{17}	0.465_{23}	0.645_{13}	0.405_{19}	0.640_{15}	0.619_{9}	$\mathbf{0.574}_{17}$
18	UniPI	0.592_{18}	0.585_{11}	0.627_{17}	0.381_{25}	0.654_{12}	0.586_{18}	$\mathbf{0.571}_{18}$
19	DIEGOLab16 (*)	0.611_{14}	0.506_{21}	0.618_{18}	0.497_{5}	0.594_{20}	0.584_{19}	$\mathbf{0.554}_{19}$
20	GTI	0.612_{13}	0.524_{17}	0.639_{16}	0.468_{7}	0.623_{17}	0.584_{19}	$\mathbf{0.539}_{20}$
21	OPAL	0.567_{20}	0.562_{15}	0.556_{23}	0.395_{21}	0.593_{21}	0.531_{21}	$\mathbf{0.505}_{21}$
22	DSIC-ELIRF	0.494_{25}	0.404_{26}	0.546_{26}	0.342_{29}	0.517_{24}	0.531_{21}	$\mathbf{0.502}_{22}$
23	UofL	0.490_{26}	0.443_{24}	0.547_{25}	0.372_{26}	0.574_{22}	0.502_{25}	$\mathbf{0.499}_{23}$
	ELiRF	0.462_{28}	0.408_{25}	0.514_{28}	0.310_{33}	0.493_{25}	0.493_{26}	$\mathbf{0.499}_{23}$
25	ISTI-CNR	0.538_{22}	0.492_{22}	0.572_{21}	0.327_{30}	0.598_{19}	0.508_{24}	$\mathbf{0.494}_{25}$
26	SteM	0.518_{23}	0.315_{29}	0.571_{22}	0.320_{32}	0.405_{28}	0.517_{23}	$\mathbf{0.478}_{26}$
27	Tweester	0.506_{24}	0.340_{28}	0.529_{27}	0.540_{3}	0.379_{29}	0.479_{28}	$\mathbf{0.455}_{27}$
28	Minions	0.489_{27}	0.521_{18}	0.554_{24}	0.420_{16}	0.475_{26}	0.481_{27}	$\mathbf{0.415}_{28}$
29	Aicyber	0.418_{29}	0.361_{27}	0.457_{29}	0.326_{31}	0.440_{27}	0.432_{29}	$\mathbf{0.402}_{29}$
30	mib	0.394_{30}	0.310_{30}	0.415_{31}	0.352_{28}	0.359_{31}	0.413_{31}	$\mathbf{0.401}_{30}$
31	VCU-TSA	0.383_{31}	0.307_{31}	0.444_{30}	0.425_{14}	0.336_{32}	0.416_{30}	$\mathbf{0.372}_{31}$
32	SentimentalITists	0.339_{33}	0.238_{33}	0.393_{32}	0.288_{34}	0.323_{34}	0.343_{33}	$\mathbf{0.339}_{32}$
33	WR	0.355_{32}	0.284_{32}	0.393_{32}	0.430_{12}	0.366_{30}	0.377_{32}	$\mathbf{0.330}_{33}$
34	CICBUAPnlp	0.193_{34}	0.193_{34}	0.335_{34}	0.393_{22}	0.326_{33}	0.303_{34}	$\mathbf{0.303}_{34}$

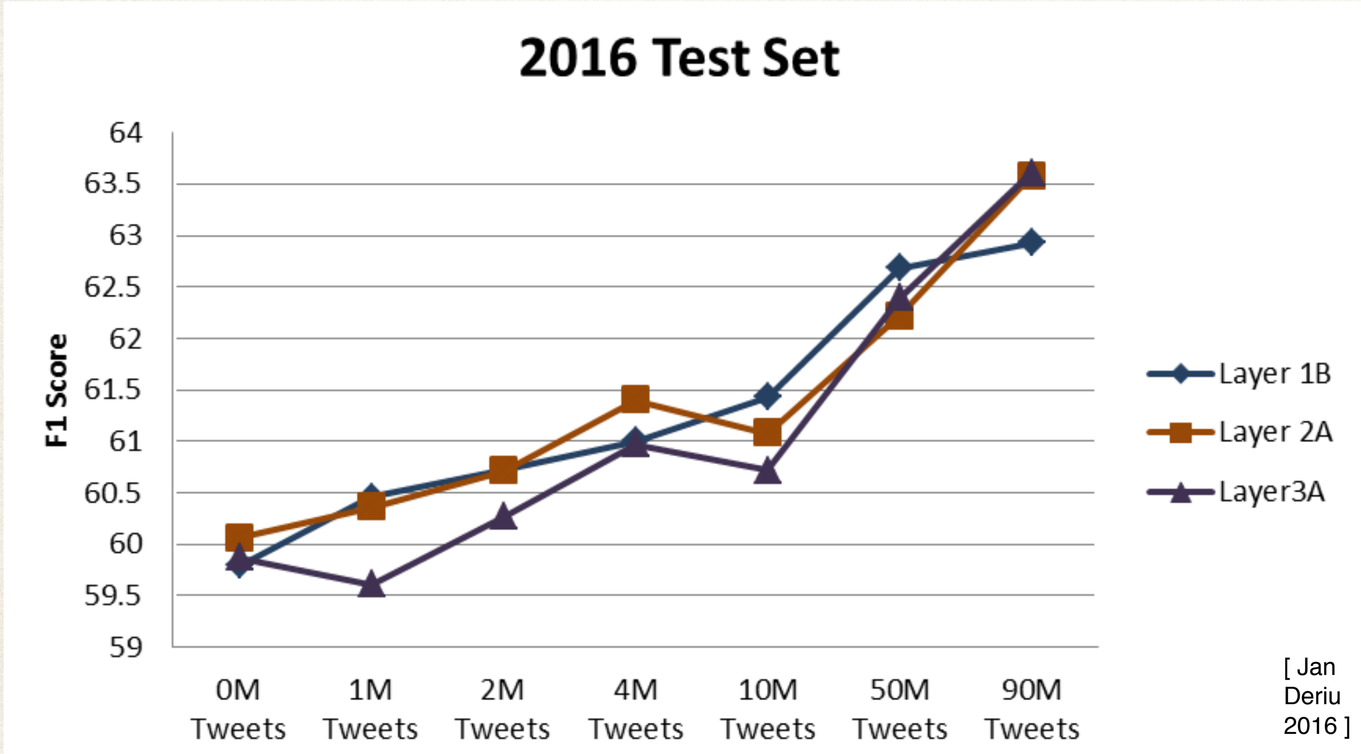
Convolutional Neural Network (CNN)



adapted from [Kim 2014]

Distant Supervision

- 90M tweets containing :) or :(



[Jan Deriu 2016]

System based on
(Severyn & Moschitti, 2015)

Evolving Word Embeddings

- Backpropagation changes both NN weights & word embeddings

[Jan Deriu 2016]

Outlook

- Sentence / document embeddings are useful for many tasks
- Medical applications
- Depth of the NNs?
- **Un-**supervised training?

References

- Many online resources, open source frameworks etc, active community
- Master Theses Jan Deriu & Maurice Gonzenbach
- *SwissCheese at SemEval-2016 Task 4: Sentiment Classification Using an Ensemble of Convolutional Neural Networks with Distant Supervision*

Thanks

Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, Dominic Egger, Pascal Julmy, Leon Derczynski, Mark Cieliebak