Deep Learning for NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

Lecture 9 – Convolutional neural networks

- **Dr. Steffen Eger**
- **Niraj D Pandey**
- **Wei Zhao**



- Natural Language Learning Group (NLLG)
- Technische Universität Darmstadt

Last lectures

TECHNISCHE UNIVERSITÄT DARMSTADT

- RNNs - Recurrent Neural Nets

- **Today:** Classifying sentences

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Problem 1: Variable-sized input
  standard MLP always expect the same input size

- Problem 2: Relevance of words
  “to” and “a” are not very informative, but content words like “kidnapping” are important for most tasks **independent of their position in the input**

- Problem 3: MLP may have too many parameters (“too complex models”) in certain situations

Today

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Convolution and pooling
2. Convolutional networks for NLP

- Lecture based on/inspired by:
  - Lecture Videos by Richard Socher and Andrej Karpathy

https://w tch?v=vYJtZwoO9Rw

h vatch?v=AQirPKrAyDg

Idea of convolution

TECHNISCHE UNIVERSITÄT DARMSTADT

*"A convolutional neural network is designed to identify **indicative local predictors** in a large structure, and combine them to produce a fixed size vector representation of the structure, capturing these local aspects that are most informative for the prediction task at hand."*

Yoav Goldberg

Convolution in image recognition

TECHNISCHE UNIVERSITÄT DARMSTADT

Example by Richard Socher:

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

$*$

1	0	1
0	1	0
1	0	1

?	?	?
?	?	?
?	?	?

Image representation *f*

Filter *g*
(also known as kernel)

Convolved image representation

Convolved features

Apply the filter to the image, move over different filter regions:

1_1	1_0	1_1	0	0
0_0	1_1	1_0	1	0
0_1	0_0	1_1	1	1
0	0	1	1	0
0	1	1	0	0

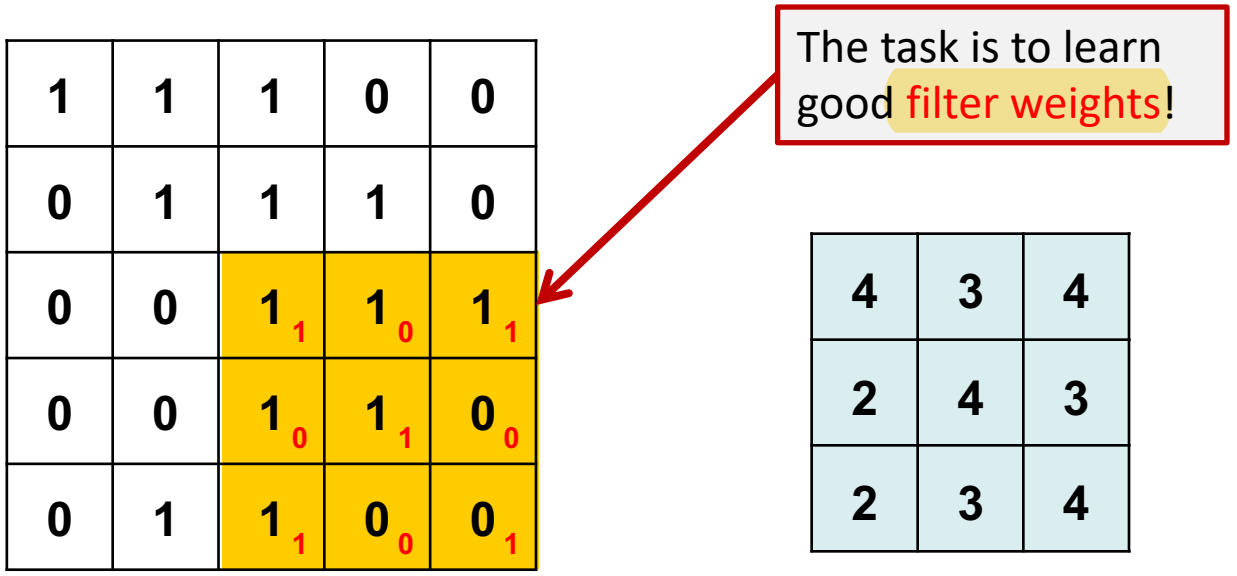
Convolution operation

4	3	4
2	4	3
2	3	4

Convolved features!

Convolved features

Apply the filter to the image, move over different filter regions :



Convolution operation

Convolved features!

And in texts?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Sentiment classification
  - The **movie** was **really good**.
  - We saw this **really good movie**.
  - The **movie**, which we saw yesterday with all the colleagues in this tiny movie ~~theatre~~ next to the bridge, was (despite my expectations) **really good**.

→ For this task, position information does not really matter.

Dimensionality

TECHNISCHE UNIVERSITÄT DARMSTADT

- Advantages of text flow:
  - Usually only one dimension

  → as opposed to two dimensions (or even three) in images

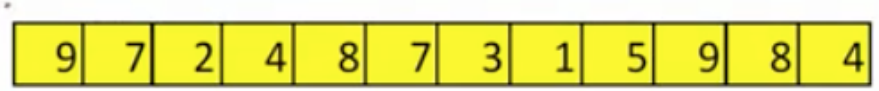
- Convolutional networks in NLP are also called **time-delay neural networks** (TDNN)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean …

The 1d convolution operation

TECHNISCHE UNIVERSITÄT DARMSTADT

$$(f * g)[i] = \sum_{m=-M}^{M} f[i-m]g[m]$$



- * is the convolution operator
- f is the input representation
- i is the current position in the input representation
- M is the window size
- $g[m]$ is the weight for an input element with distance m to the current input
  → g it is also often referred to as the *filter* (or *kernel*)

- Careful! You will find many terminological alternatives in the literature:
  - *w* (for weights) instead of *g*, *n* or *t* (for time) instead of *i*, *a* (for age) instead of *m*, …

Convolutional Layer in NLP

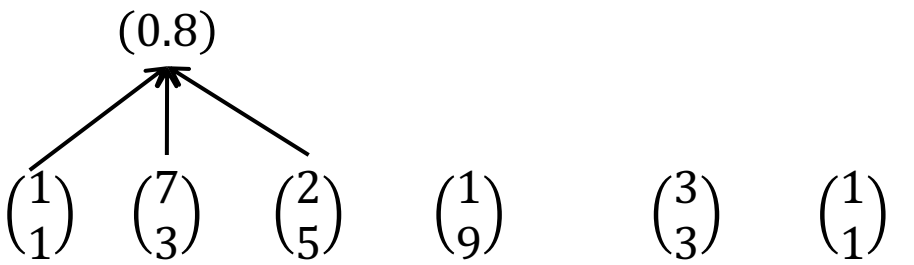
TECHNISCHE UNIVERSITÄT DARMSTADT

- Input sentence $\boldsymbol{x}_{i:i+n}$ is a concatenation of the word vectors $\boldsymbol{x}_i, \dots, \boldsymbol{x}_{i+n} \in \boldsymbol{R}^d$
- Convolutional filter: $\boldsymbol{w} \in \boldsymbol{R}^{hd}$ h is *filter size*
- Convolution operation: $\boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h}$

(Or any other non-linearity)

$$c_i = \tanh(\ \boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h} + b\ )$$

$h = 3$
$d = 2$



Convolutional Layer in NLP

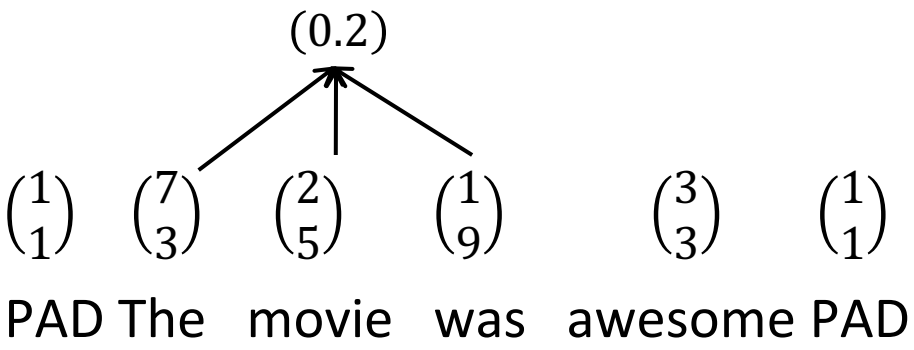
TECHNISCHE UNIVERSITÄT DARMSTADT

- Input sentence $\boldsymbol{x}_{1:n}$ is a concatenation of the word vectors $\boldsymbol{x}_i \in \boldsymbol{R}^d$
- Convolutional filter: $\boldsymbol{w} \in \boldsymbol{R}^{hd}$
- Convolution operation: $\boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h}$

h is *filter size*

$$c_i = \tanh(\ \boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h} + b\ )$$

$h = 3$
$d = 2$



Convolutional Layer in NLP

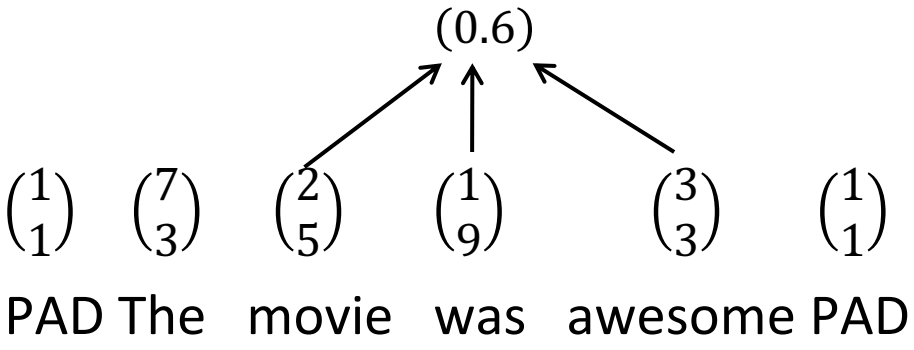
TECHNISCHE UNIVERSITÄT DARMSTADT

- Input sentence $\boldsymbol{x}_{1:n}$ is a concatenation of the word vectors $\boldsymbol{x}_i \in \boldsymbol{R}^d$
- Convolutional filter: $\boldsymbol{w} \in \boldsymbol{R}^{hd}$
- Convolution operation: $\boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h}$

h is *filter size*

$$c_i = \tanh(\ \boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h} + b\ )$$

$h = 3$
$d = 2$



Convolutional Layer in NLP

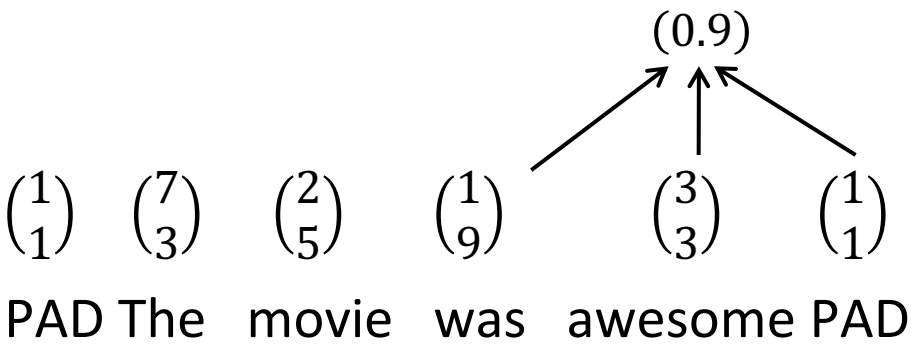
TECHNISCHE UNIVERSITÄT DARMSTADT

- Input sentence $\boldsymbol{x}_{1:n}$ is a concatenation of the word vectors $\boldsymbol{x}_i \in \boldsymbol{R}^d$
- Convolutional filter: $\boldsymbol{w} \in \boldsymbol{R}^{hd}$
- Convolution operation: $\boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h}$

h is *filter size*

$$c_i = \tanh(\ \boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h} + b\ )$$

$h = 3$
$d = 2$



Different viewpoints on convolution

TECHNISCHE UNIVERSITÄT DARMSTADT

- Is convolution for text in 1d or in 2d?
- Can also interpret that each input $f[i-m]$ lies in R^d
- and each weight $g[m]$ also lies in R^d

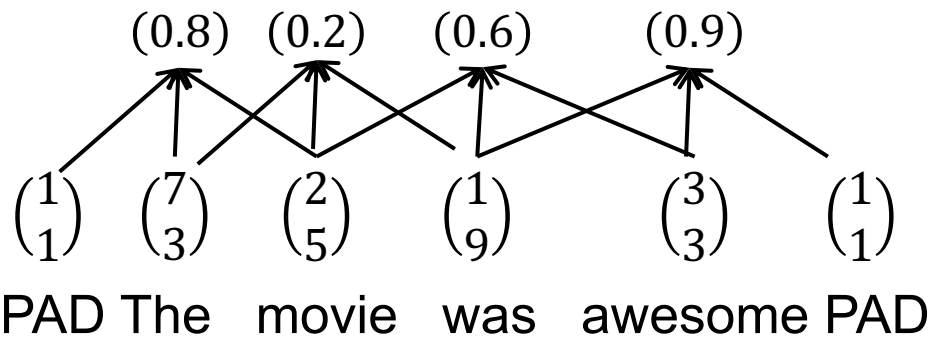
$$(\boldsymbol{f} * \boldsymbol{g})[\boldsymbol{i}] = \sum_{\boldsymbol{m}} \boldsymbol{f}[\boldsymbol{i}-\boldsymbol{m}]\boldsymbol{g}[\boldsymbol{m}]$$

Properties of convolutional networks

TECHNISCHE UNIVERSITÄT DARMSTADT

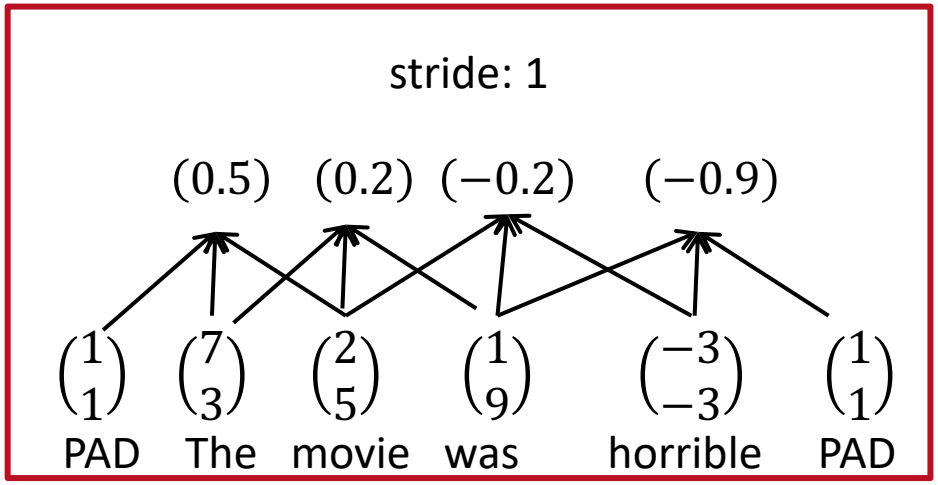
- Not every input is connected to every output in the following layer
  → **sparse connectivity** (vs **fully-connected/dense** layers)

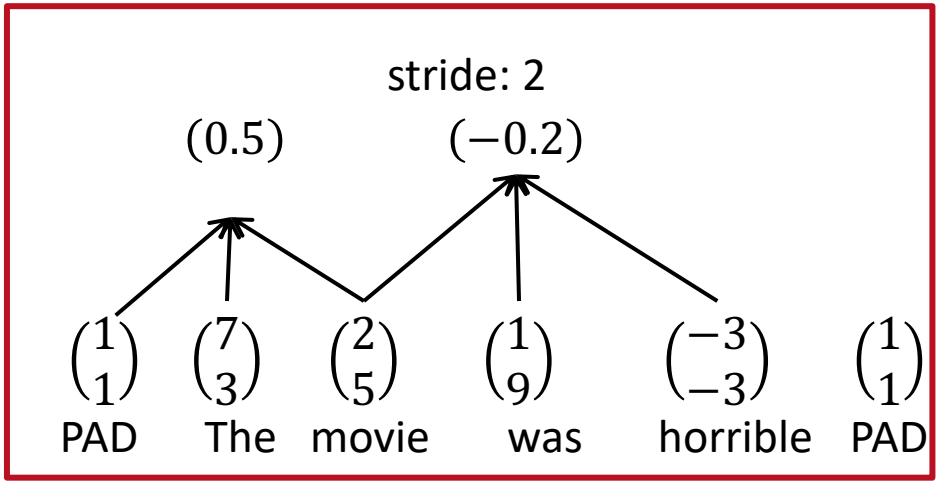
- For each window, we use the same weights and bias values
  → **parameter sharing**



Stride

- The stride specifies the steps size for moving over the sentence
- In NLP, stride 1 is commonly used
- In computer vision, other values might be more useful





Dense layer vs. Convolutional Layer

TECHNISCHE UNIVERSITÄT DARMSTADT

- In principle, a convolutional layer could handle variable-sized inputs
- But in practice, it handles fixed-sized input, just like in an MLP
  - We usually pad with zeros so that all sequences in our data have the same length
  - Sometimes we also truncate

Dense layer vs. Convolutional Layer

TECHNISCHE UNIVERSITÄT DARMSTADT

- So, the main difference to our known dense layers is really **parameter-sharing** and **sparse connectivity**

- Why are these two properties important?

Pooling layer

TECHNISCHE UNIVERSITÄT DARMSTADT

- Another new building block: **pooling layer**

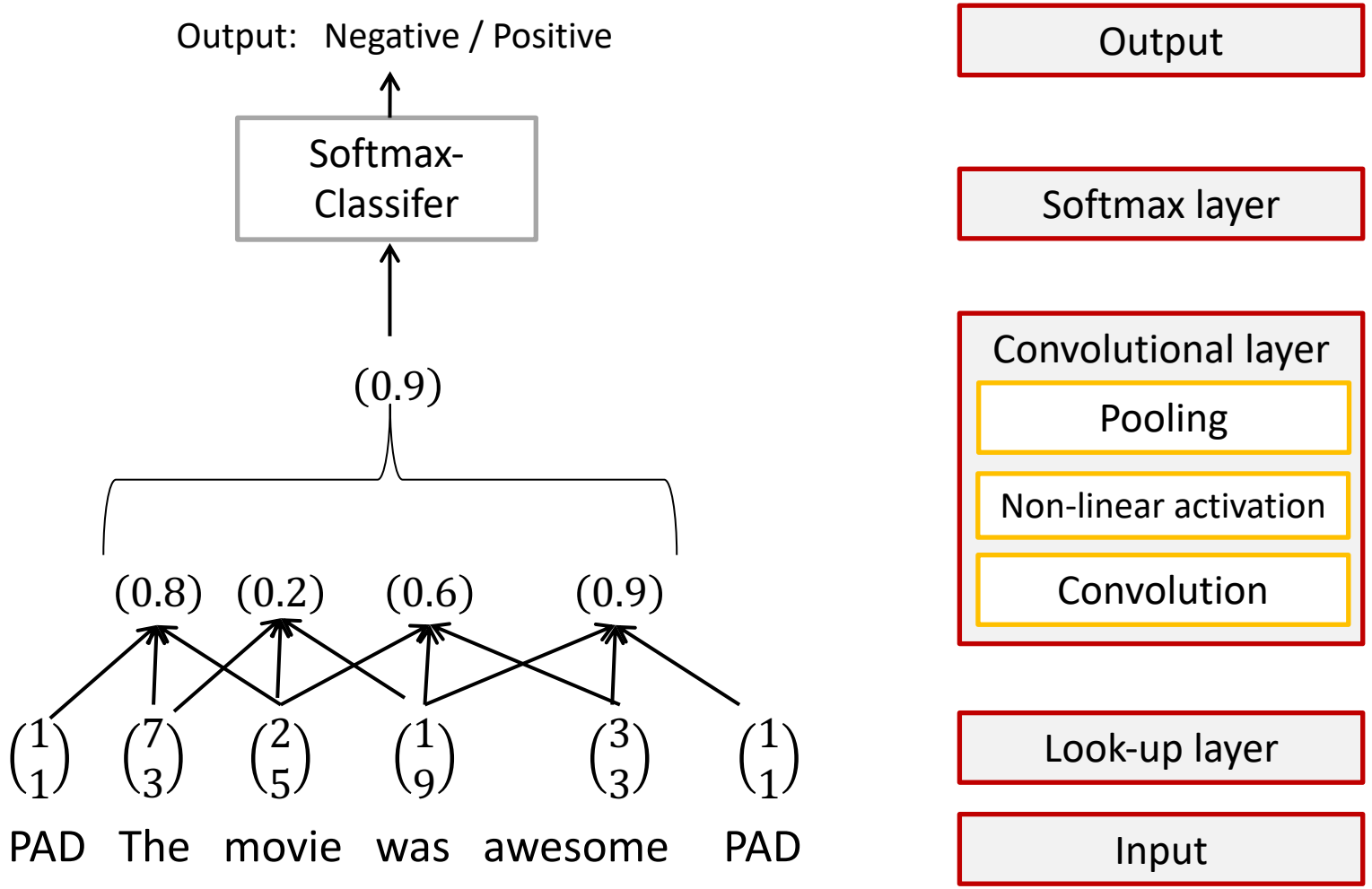
  → Idea: capture the most important activation

- Let $c_1, c_2, \ldots \in \boldsymbol{R}$ denote the output values for our convolutional filter
- Compute the output o for a **max-over-time pooling** layer:

$$o = \max_i c_i$$

- Max-over-time pooling is most common in NLP. You can also find min-pooling and mean-pooling in other areas. Could also use some other averaging
- Note that there are no associated weights

Classification with convolution and pooling



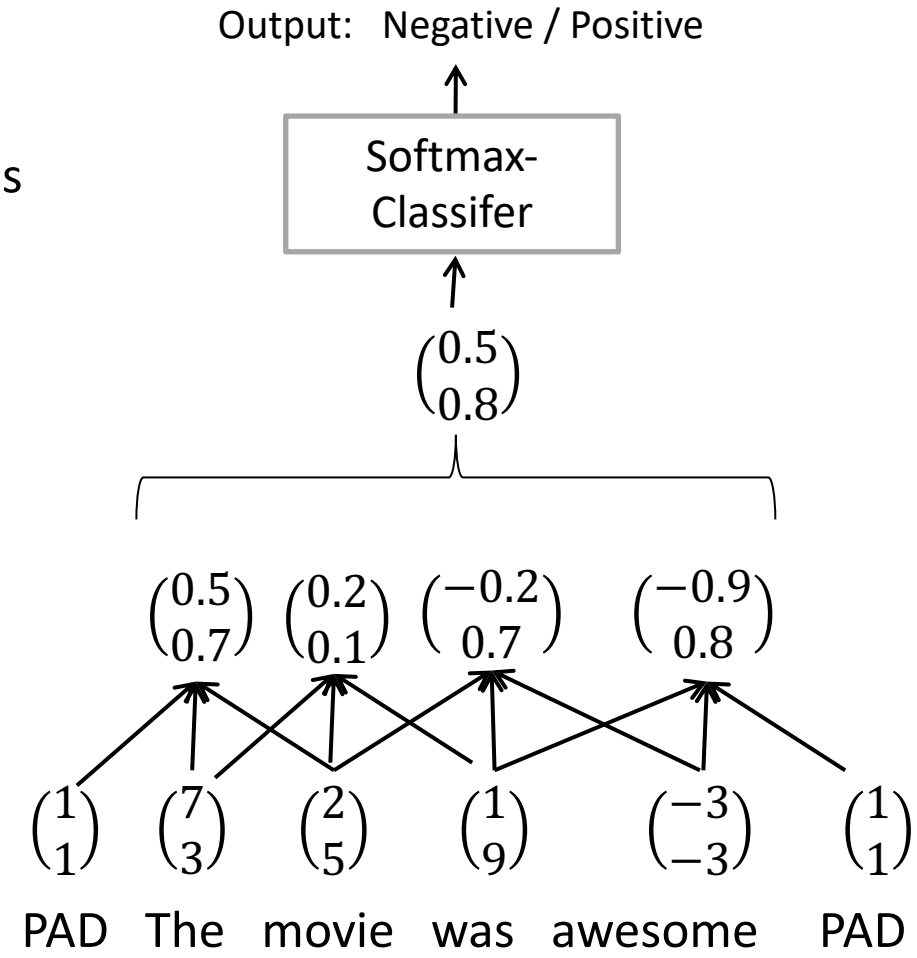
Multiple filters

TECHNISCHE UNIVERSITÄT DARMSTADT

- Usually we have **many** filters (hundreds or thousands), not just one
- They may be of same or of different sizes

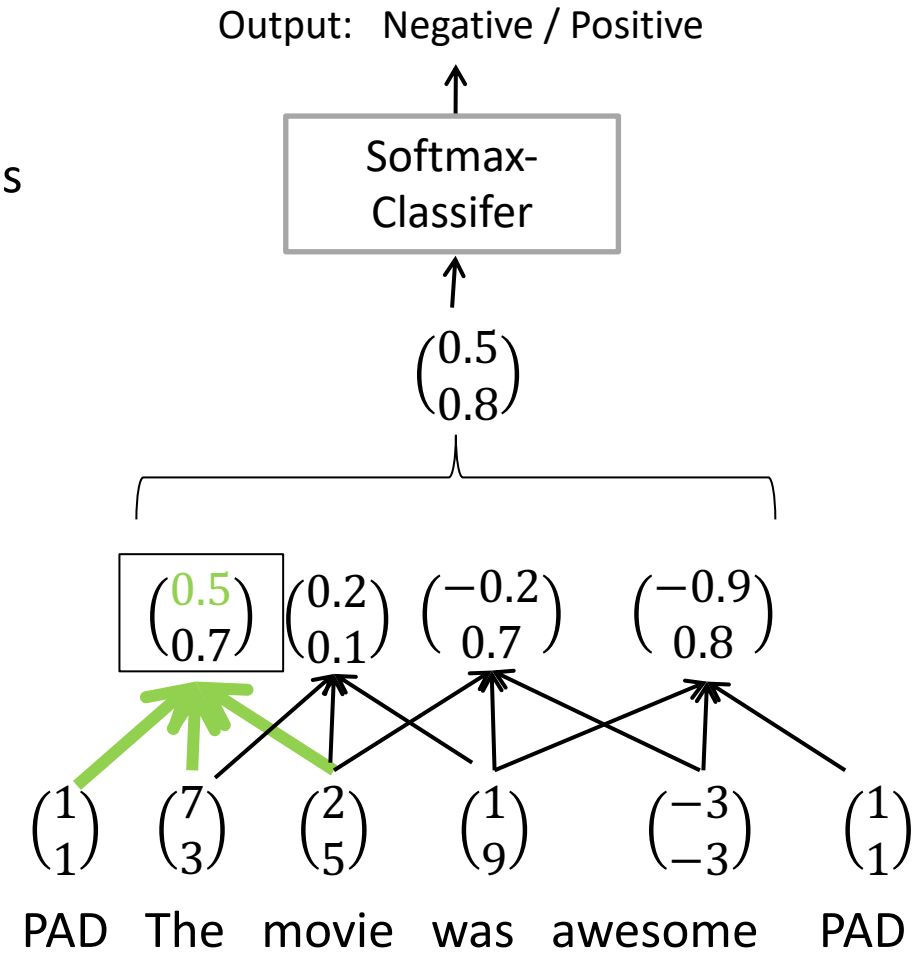
Multiple filters

- Further filters.
- The convolved representation is often called a feature representation.



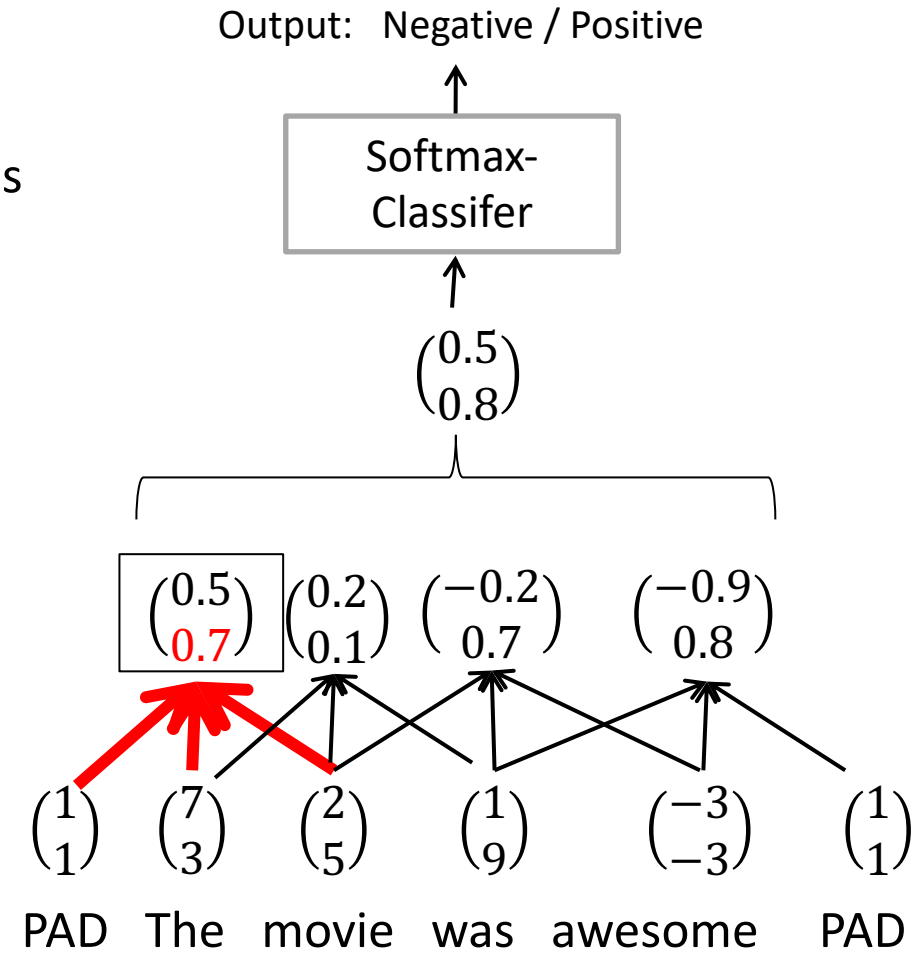
Multiple filters

- Further filters.
- The convolved representation is often called a feature representation.

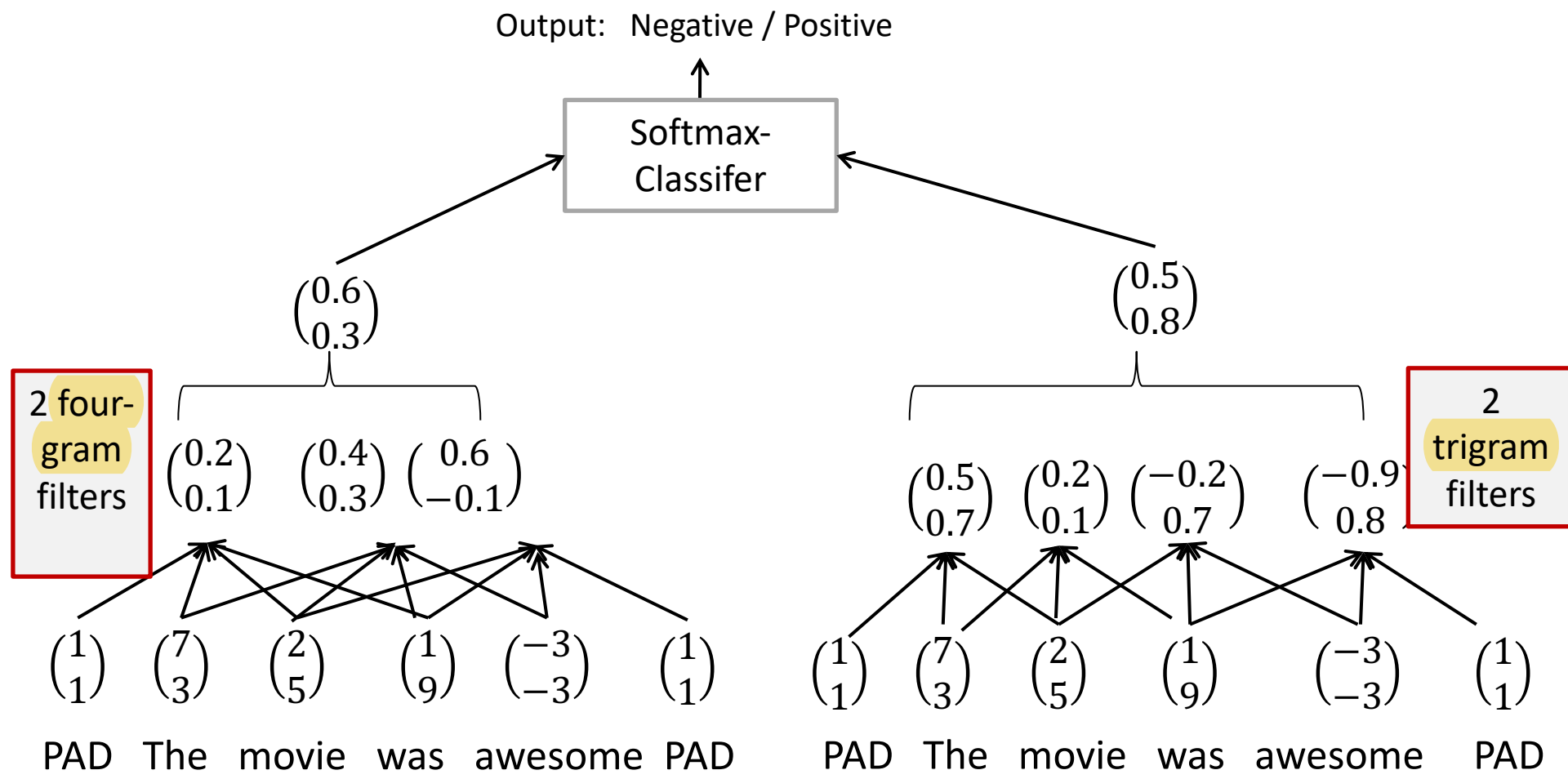


Multiple filters

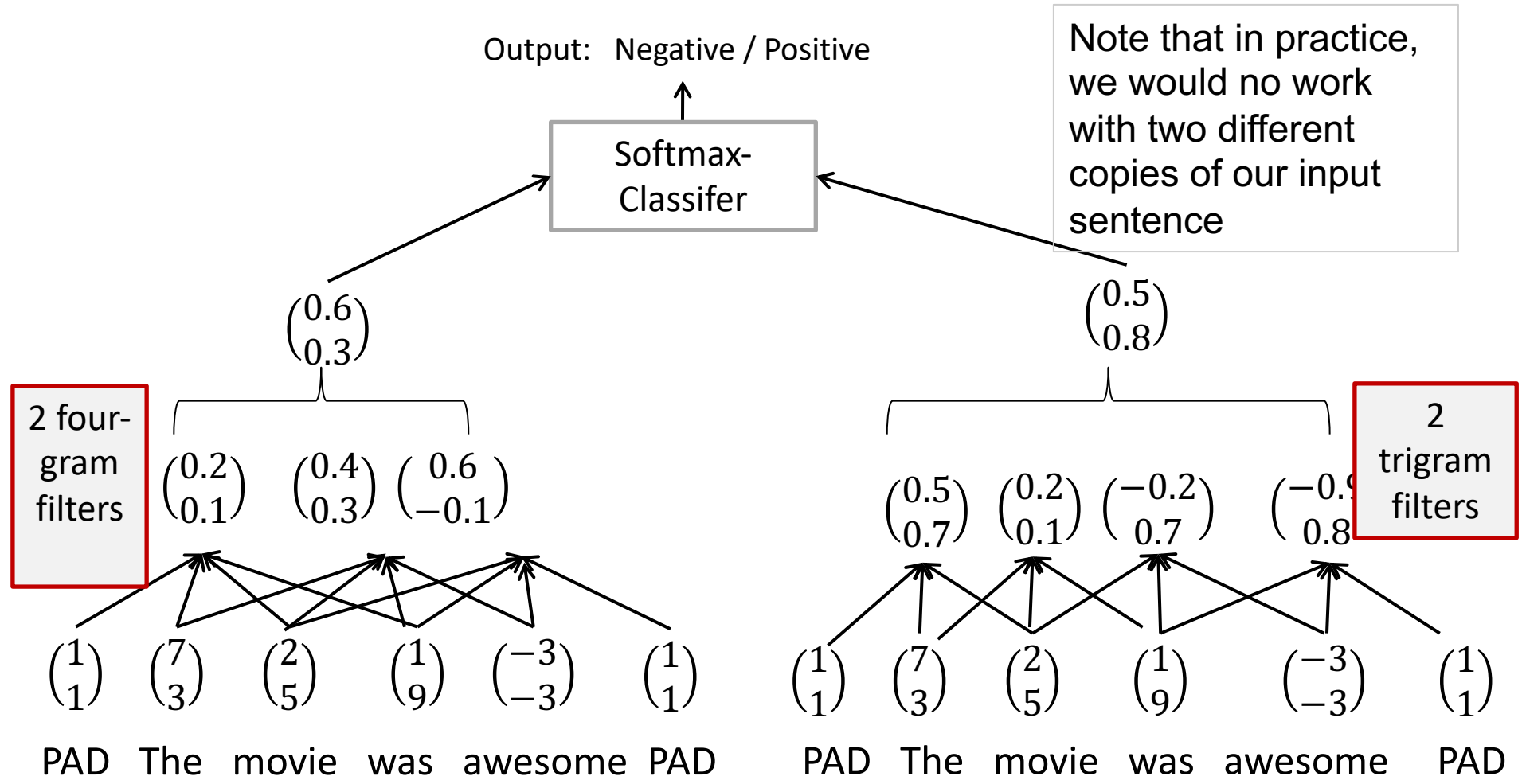
- Further filters.
- The convolved representation is often called a feature representation.



Combining different n-gram sizes



Combining different n-gram sizes



Properties of pooling

TECHNISCHE UNIVERSITÄT DARMSTADT

- Idea: Extracting relevant features independent of their position in the input
- Problems:
  - Output remains the same if a feature occurs once or multiple times

    The music was great, but the cast was horrible, the plot was horrible and the costumes were horrible.

  - Order of features is not considered

    I don‘t love it, I hate it. vs *I don‘t hate it, I love it.*

Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Convolution and pooling

2. **Convolutional networks for NLP**

Convolutional networks for NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

- Sentence classification
  - Kalchbrenner, Grefenstette, Blunsom, 2014: *A Convolutional Neural Network for Modelling Sentences*
  - Kim, 2014: *Convolutional Neural Networks for Sentence Classification*
  - Zhang and Wallace, 2016: *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification*
- Semantic role labeling (SRL)
  - The SENNA framework
    Collobert and Weston, 2011: *Natural Language Processing (Almost) from Scratch*
- Character-based approach
  - Zhang, Zhao, LeCun 2015: *Character-level Convolutional Networks for Text Classification*
  - Kim et al., 2016: *Character-aware Neural Language Models*

Sentence classification tasks

TECHNISCHE UNIVERSITÄT DARMSTADT

- Sentiment classification of movie reviews, product reviews, tweets

- Question classification into 6 question types: person, location, numeric information,…

- Classifying sentences into subjective / objective

“I feel this work is not novel enough.“

- Classifying whether a sentence is ironic

What is a word?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Convolutional approach was first developed for images (group pixels together)
- Our unit: words

pneumonoultramicroscopicsilicovolcanoconiosis

lung very small silicon/quartz volcano dust disease

„Quarzstaublunge”

- In Chinese: 乒乓球拍卖完了。

乒乓 /球拍 /卖完了。

ping-pong racket sold out

“The ping-pong rackets have been sold out.”

乒乓球 /拍卖 /完了。

ping-pong ball auction finish

“The auction of the ping-pong ball has been finished.”

Character-based approach

TECHNISCHE UNIVERSITÄT DARMSTADT

- Characters as units, smaller vocabulary:
  - 70 characters:26 English letters, 10 digits, 33 other characters and the new line character.

```
abcdefghijklmnopqrstuvwxyz0123456789
-,;.!?:’’’/\|_@#$%^&*~‘+-=<>()[]{}
```

  - plus uppercase letters, if required

- No embeddings, just one-hot vectors
  - Blanks and all other characters are all-zero vectors

Character-based approach for text classification

TECHNISCHE UNIVERSITÄT DARMSTADT

- Classifying
  - articles into topics
  - reviews into positive/negative

- Approach can compete with state-of-the-art, but requires huge datasets
  - Increase the training set by replacing words with their synonyms using a thesaurus

- Fascinating idea, might lead to more language-independent models, but currently word embeddings work better for most tasks
- Has also been used with LSTMs -> next lecture

Summary

TECHNISCHE UNIVERSITÄT DARMSTADT

- Convolutional networks can deal with variable sized input
  - Sparse connectivity, parameter sharing
  - Narrow vs wide convolution
- Pooling makes it possible to focus on most relevant features
  - Max-over-time pooling
- Convolutional networks for NLP
  - Sentence classification
  - Character-based approaches

References

- The lectures referenced on slide 4
- www.deeplearningbook.org
- Kalchbrenner, Grefenstette, Blunsom, (2014): A Convolutional Neural Network for Modelling Sentences, *arXiv preprint arXiv:1404.2188*
- Kim(2014): Convolutional Neural Networks for Sentence Classification, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*: 1746–1751.
- Zhang and Wallace, 2016: A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification, *arXiv preprint arXiv:1510.03820*.
- Collobert and Weston (2011): Natural Language Processing (Almost) from Scratch, in *The Journal of Machine Learning Research 12*: 2493-2537.
- Zhang, Zhao, LeCun (2015): Character-level Convolutional Networks for Text Classification, in *Advances in Neural Information Processing Systems: 649-657*.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014): Dropout: A simple way to prevent neural networks from overfitting, in *The Journal of Machine Learning Research, 15(1)*: 1929-1958.
- Prechelt, L. (1998): Early stopping-but when?, in *Neural Networks: Tricks of the trade*: 55-69, Springer Berlin Heidelberg.
- Conneau et al. (2016), „Very Deep Convolutional Networks for Text Classification“