Convolutional Neural Network based Speech Emotion Recognition

Michael Neumann

Institut für maschinelle Sprachverarbeitung

October 21, 2016

Overview

Introduction - Emotion Recognition

- Why recognize emotions?

Introduction - Emotion Recognition

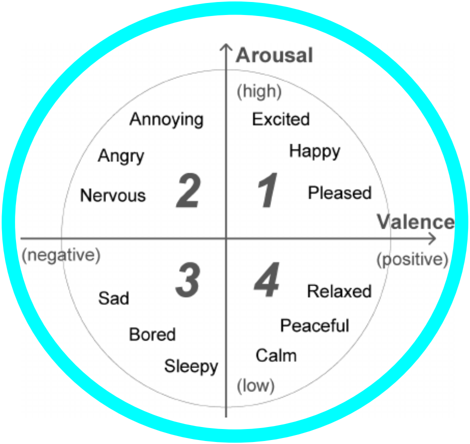
- Why recognize emotions?
  $\rightarrow$ Human-Machine-interaction: react more naturally

Introduction - Emotion Recognition

- Why recognize emotions?
  → Human-Machine-interaction: react more naturally
- Methods of classification:
  - Emotional categories (e.g. 'sad', 'happy', 'angry', 'neutral')
  - Valence and Arousal dimensions

Introduction - Emotion Recognition

- Why recognize emotions? $\rightarrow$ Human-Machine-interaction: react more naturally
- Methods of classification:
  - Emotional categories (e.g. 'sad', 'happy', 'angry', 'neutral')
  - Valence and Arousal dimensions



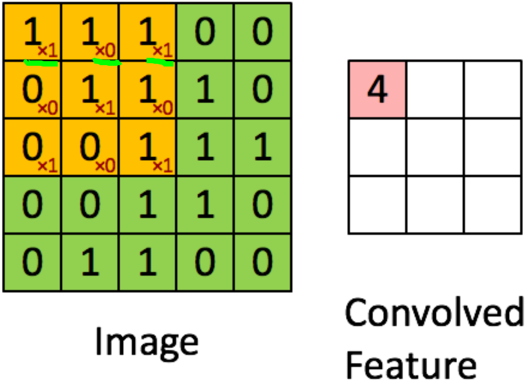
Introduction - Deep Learning

- Deep Learning (DL) has become state-of-the-art for many tasks (e.g. speech recognition, computer vision)
- Convolutional neural networks (CNNs) originate from computer vision
- Successfully used for speech data recently

Institut für Maschinelle Sprachverarbeitung

Introduction - Deep Learning

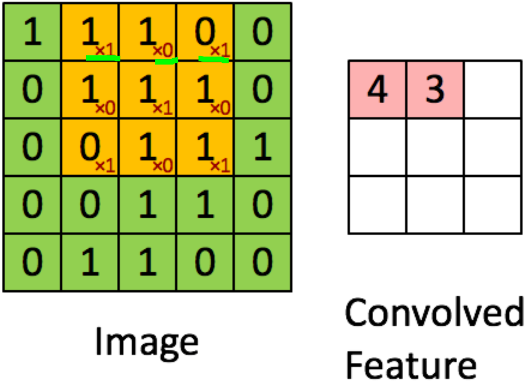
How convolution works



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Introduction - Deep Learning

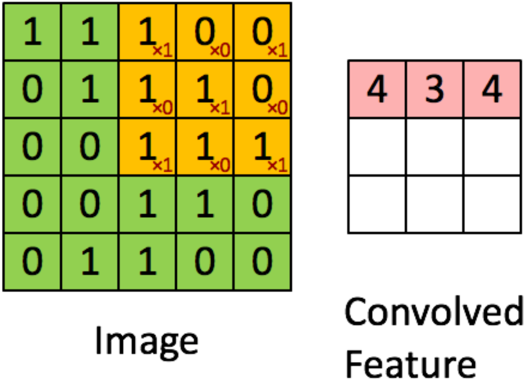
How convolution works



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Introduction - Deep Learning

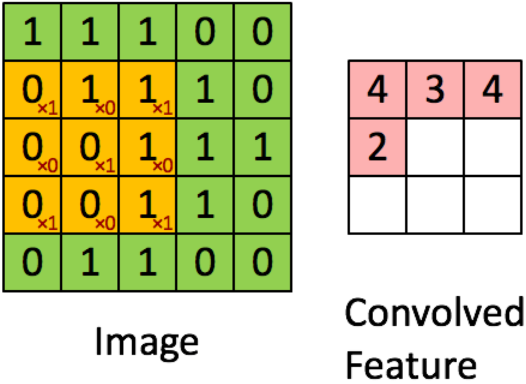
How convolution works



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Introduction - Deep Learning

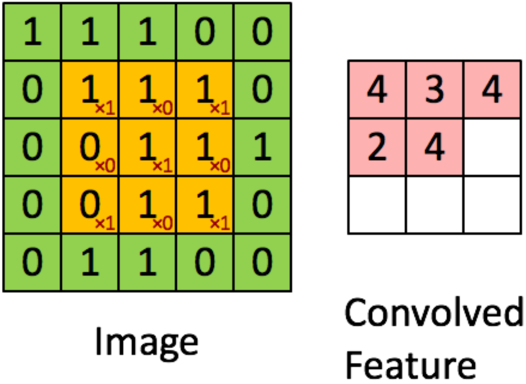
How convolution works



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Introduction - Deep Learning

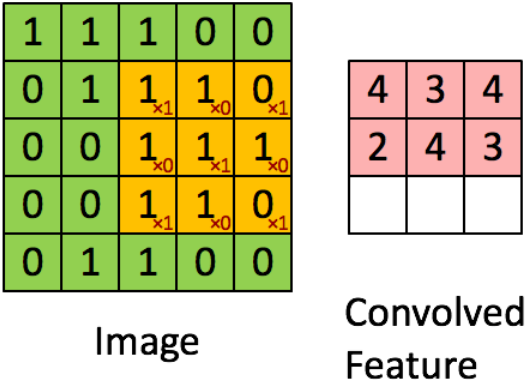
How convolution works



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Introduction - Deep Learning

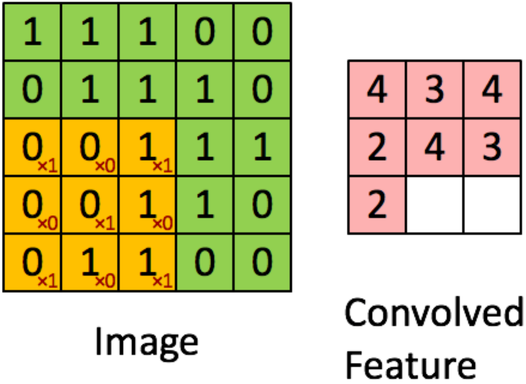
How convolution works



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Introduction - Deep Learning

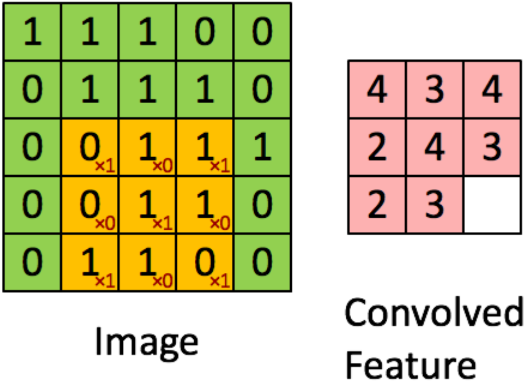
How convolution works



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Introduction - Deep Learning

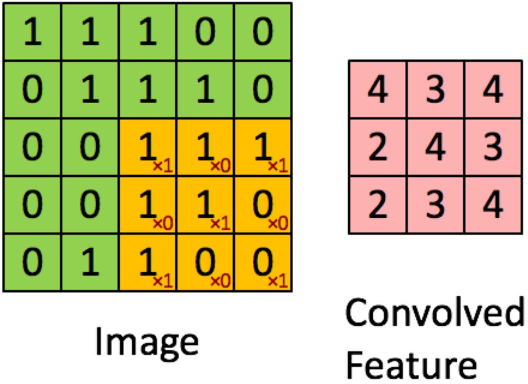
How convolution works



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Introduction - Deep Learning

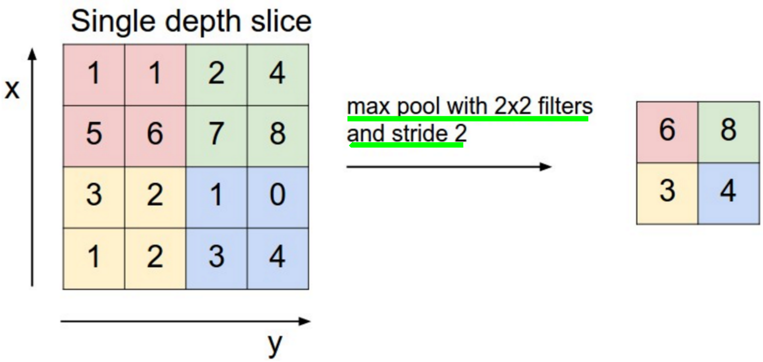
How convolution works



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

Introduction - Deep Learning

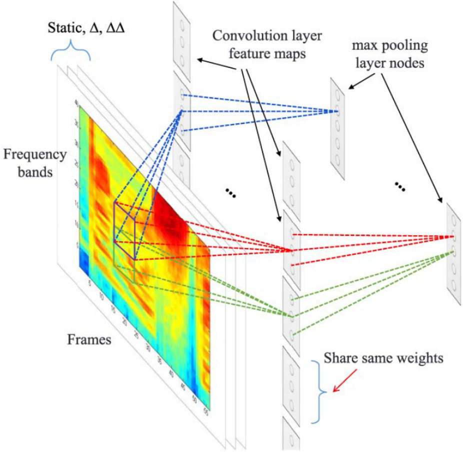
How pooling works



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

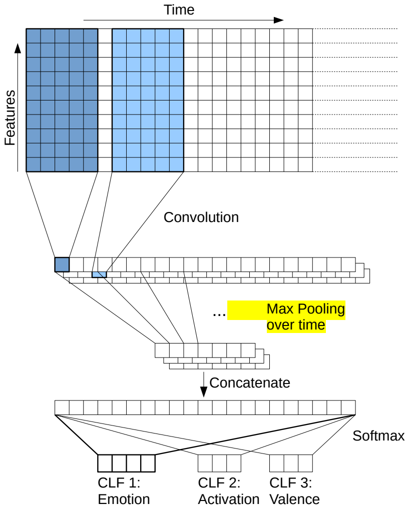
Introduction - Deep Learning

CNN for speech



http://recognize-speech.com/acoustic-model/knn/comparing-different-architectures/convolutional-neural-networks-cnns

CNN Model



- Simple CNN with one convolutional and one pooling layer
- Multi-task Learning: Consider Activation/ Valence information
- Cost function:

$$J = (1 - \alpha - \beta) \cdot J_{CLF1}$$
$$+\alpha \cdot J_{CLF2}$$
$$+\beta \cdot J_{CLF3}$$

Experimental Setup

- Input Features:
  - Logarithmic power of Mel-frequency bands (logMel)
  - Mel frequency cepstral coefficients (MFCC)
  - extended Geneva minimalistic acoustic parameter set (eGeMAPS)

Experimental Setup

- Input Features:
  - Logarithmic power of Mel-frequency bands (logMel)
  - Mel frequency cepstral coefficients (MFCC)
  - extended Geneva minimalistic acoustic parameter set (eGeMAPS)
- Dataset:
  - Interactive Emotional Dyadic Motion Capture (IEMOCAP) database
  - 5,531 utterances

Experimental Setup

- Experiment 1:
  - Performance with different feature sets using single-task and multi-task learning
  - Which input features are most suitable?
  - Does multi-task learning improve results?

Experimental Setup

- Experiment 1:
  - Performance with different feature sets using single-task and multi-task learning
  - Which input features are most suitable?
  - Does multi-task learning improve results?
- Experiment 2:
  - Train and test the model with decreasing signal length
  - How big is the performance impact with shorter utterance snippets?

Experimental Results

Experiment 1

Features	Single-task Accuracy	Multi-task Accuracy
log Mel	56.01	57.26
MFCC	56.07	56.13
eGeMAPS	55.71	56.73

Experimental Results

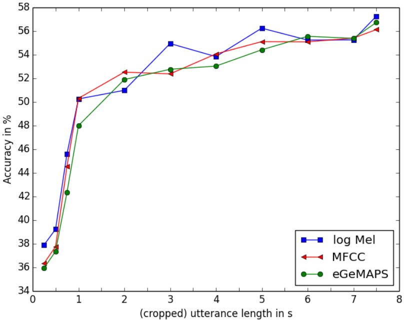
Experiment 1

Features	Single-task Accuracy	Multi-task Accuracy
log Mel	56.01	57.26
MFCC	56.07	56.13
eGeMAPS	55.71	56.73

- No great performance differences between feature sets
- Multi-task learning improves performance

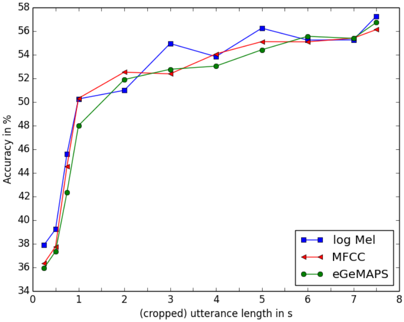
Experimental Results

Experiment 2



Experimental Results

Experiment 2



- Best accuracy with longest signal
- Only slight performance decrease until 1s
- Relatively short snippet of 3s can be sufficient

Conclusions

Conclusions

- Similar performance despite differences in input features

Conclusions

- Similar performance despite differences in input features

→ Network architecture more important than choice of features

Conclusions

- Similar performance despite differences in input features

$\rightarrow$ Network architecture more important than choice of features

- Multi-task learning improves performance slightly

Conclusions

- Similar performance despite differences in input features

→ Network architecture more important than choice of features

- Multi-task learning improves performance slightly
- Prediction can be performed based on the first 3 sec. (with slight performance loss)

CNN based Speech Emotion Recognition

Thanks for your attention.