KU LEUVEN

Ontwerp, analyse en implementatie van een convolutionair neuraal netwerk voor gelijktijdige spraak en beeldherkenning

Overview

-

1. Problem sketch
2. Current research
3. Neural networks
4. Dataset: TCDTIMIT
5. Objectives
6. Lipreading
7. Speech (audio)
8. Sensor fusion

KU LEUVEN

1. Problem sketch

Speech recognition is useful:

- Automatic subtitles
- Assisting hearing impaired
- Human-computer interation (Siri)
- International meetings (translations)
-
- Until now: mostly audio
- Use images (lipreading) → robustness, performance

KU LEUVEN

2. Research

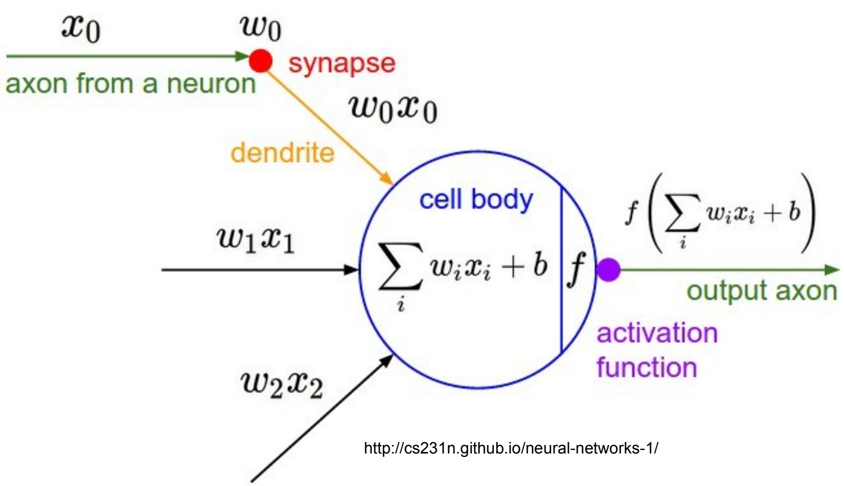
- In the past:
  - Mostly audio SR
  - Acoustic model: formants, fricatives,...
  - Record sounds, statistical correlation of spectrals
    - → Hidden Markov Models (HMM)
  - Language model on top
  - Often limited in scope (eg. Phone support)

2. Research

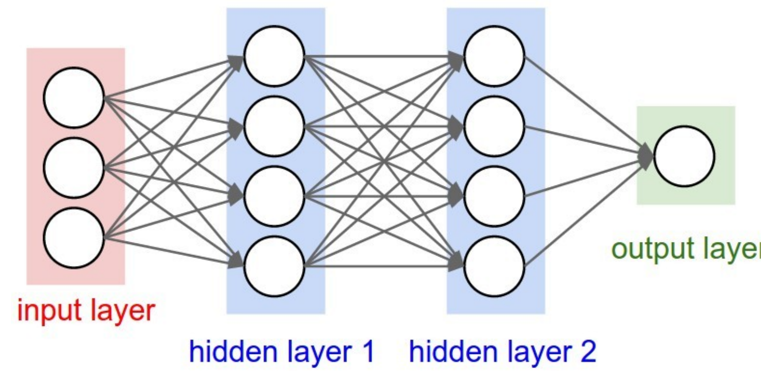
- Current:
-
  - Still mostly audio
  - Acoustic model: formants, fricatives,…
  - Record sounds, statistical correlation of spectrals
    - → Convolutional Neural Networks
    -
  - Language model on top (possibly DNN?)
  - Much broader in scope (Siri, Cortana, SR 'in the wild')

3. Neural Networks

-
- Simple units with nonlinear output function



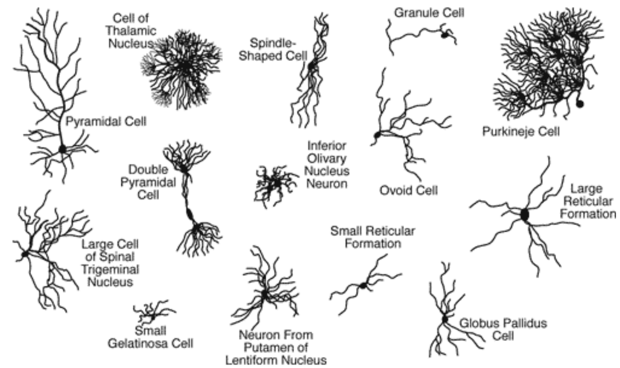
http://cs231n.github.io/neural-networks-1/



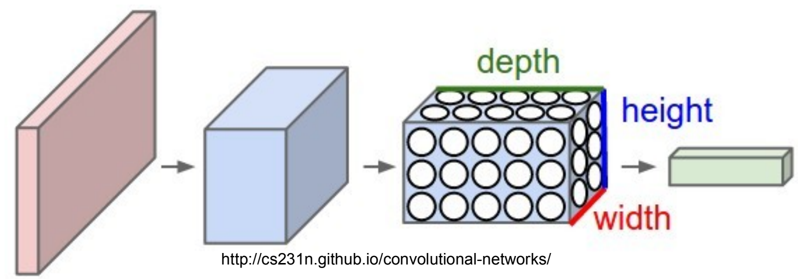
http://cs231n.github.io/neural-networks-1/

3. Neural Networks

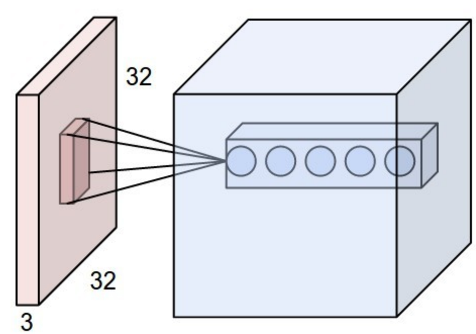
- Goal: Pattern Recognition → high-dimensional input data
- Fully connected Nns don’t scale
  - We want to reduce # parameters
- Brain also uses specialized neurons
- → Convolutional Neural Networks
- Layers in 3D ≈ trainable filters
- parameter sharing + pooling
- Layer types: CONV, ReLu, Pool, FC,...



http://www.mind.ilstu.edu/curriculum/neurons_intro/neurons_intro.php



http://cs231n.github.io/convolutional-networks/



http://cs231n.github.io/convolutional-networks/

4. Dataset: TCDTIMIT

- Alternatives:
  - GRID: large dataset, but small vocabulary
  - VidTIMIT: small dataset
  - Many non-public databases (Google etc)
- TCDTIMIT:
  - Many speakers, high quality
  - Continuous speech, good coverage of phonemes and visemes. (TIMIT)
  - Available to other researchers.
  - Content:
    - 2255 sentences from TIMIT
    - 59 volunteers (98 sentences each)
    - 3 professional lipspeakers (377 sentences each)
    - ~25 phonemes/sentence
    - Total: 235k phoneme examples; ~ 6k each

Harte, N.; Gillen, E., "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," Multimedia, IEEE Transactions on , vol.17, no.5, pp.603,615, May 2015 doi: 10.1109/TMM.2015.2407694

4. Dataset: TCDTIMIT

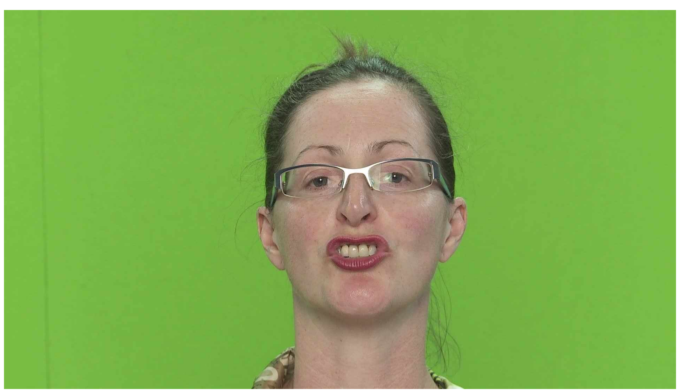
- Issues downloading & extracting
- Lacking documentation
- Very little support
- Files missing
- After processing:
  - time mismatch phoneme- frame
  - frames missing
  - Other issues
- → write own software to extract data from videos
  → make available for other researchers using that database

KU LEUVEN

4. Dataset: TCDTIMIT

- Goal: labeled frames of phoneme pronounciation
- SW pipeline:
  - Extract phoneme time information
  - Extract frames
  - Remove invalid frames
  - Extract faces, mouths
  - Grayscale and compress
  - Pickle for simple loading in Python

Frame	Phoneme
16	sil
34	sh
37	iy
40	hh
44	ae
45	d
47	y
49	uh
...	...



Sa1.mp4 (60MB)

38 x
sa1_34_sh.jpg (2KB)

4. Dataset: TCDTIMIT

-
- Train/test/validation set splits:
- For each speaker:
  - 80% training set
  - 10% validation set
  - 10% test set
  - Baseline results from database paper (using HMM)

	Split 1 (Table 4.1)		Split 2	
	Train set	Test set	Train set	Test set
%correct	42.33	46.78	41.18	46.97
%accuracy	36.50	34.77	35.53	35.61

Harte, N.; Gillen, E., "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," Multimedia, IEEE Transactions on , vol.17, no.5, pp.603,615, May 2015 doi: 10.1109/TMM.2015.2407694

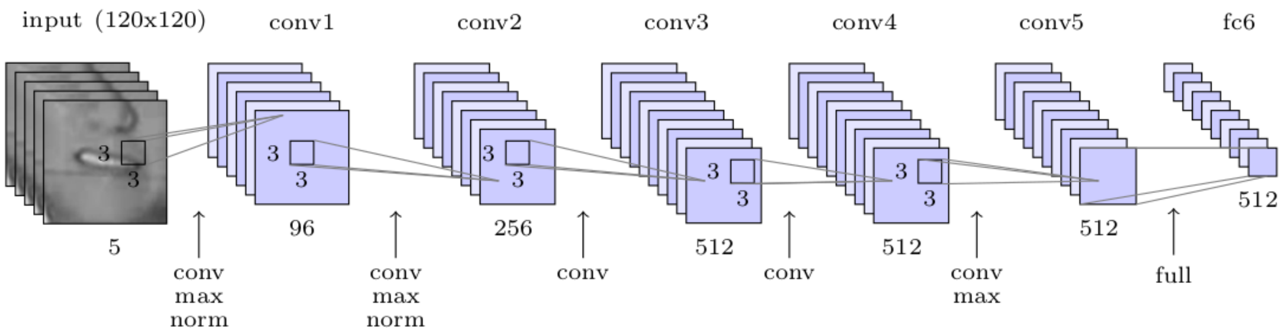
5. Objectives

- Combine lipreading and audio to achieve:
  - Better performance (we use more information)
  - Better robustness (low quality recording, background noise,…)
    - → use best information source available
  - Work on phonemes, not words or sentences
    - Simpler; also smaller networks needed
    - Language independent (if you have a dataset)
    - Possible to put language model on top

KU LEUVEN

6. Lipreading

- Most SR research focused on Audio (phonemes)
- Limited correllation lips ↔ sound (aspirated or not,…) → map to visemes
- Here, just phonemes used for lipreading (possible information loss + solve ambiguity by language model)
- Classification problem: 39 phonemes
- Networks tested: 1) CIFAR 10 8 layer network
    - 2) ResNet 50 layerscifar
    - 3) simple 6-layer ConvNet
- No time-aspect (yet)



6. Lipreading: Google WLAS

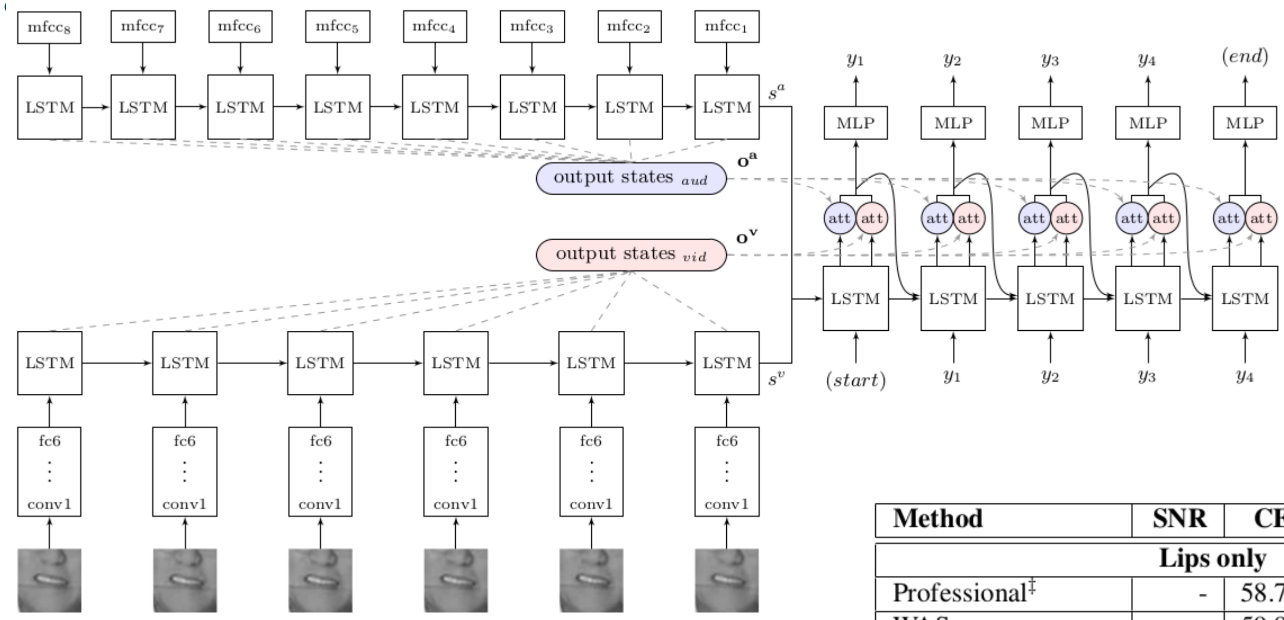
-
- Nov 2016
- Goal: transcribe videos of mouth motion to characters
- Beats a professional lip reader on videos from BBC television
- 3 parts, merged with alignment mechanism

$$
\begin{aligned}
s^v, \mathbf{o}^v &= \texttt{Watch}(\mathbf{x}^v) \\
s^a, \mathbf{o}^a &= \texttt{Listen}(\mathbf{x}^a) \\
P(\mathbf{y}|\mathbf{x}^v, \mathbf{x}^a) &= \texttt{Spell}(s^v, s^a, \mathbf{o}^v, \mathbf{o}^a)
\end{aligned}
$$

https://www.youtube.com/watch?v=5aogzAUPilE&feature=youtu.be

KU LEUVEN

6. Lipreading: Google WLAS



$$s^v, \mathbf{o}^v = \mathtt{Watch}(\mathbf{x}^v)$$
$$s^a, \mathbf{o}^a = \mathtt{Listen}(\mathbf{x}^a)$$
$$P(\mathbf{y}|\mathbf{x}^v, \mathbf{x}^a) = \mathtt{Spell}(s^v, s^a, \mathbf{o}^v, \mathbf{o}^a)$$

Method	SNR	CER
Lips only		
Professional‡	-	58.7%
WAS	-	59.9%
WAS+CL	-	47.1%
WAS+CL+SS	-	44.2%
WAS+CL+SS+BS	-	42.1%
Audio only		
LAS+CL+SS+BS	clean	16.2%
LAS+CL+SS+BS	10dB	33.7%
LAS+CL+SS+BS	0dB	59.0%
Audio and lips		
WLAS+CL+SS+BS	clean	13.3%
WLAS+CL+SS+BS	10dB	22.8%
WLAS+CL+SS+BS	0dB	35.8%

6. Lipreading: results Google network

- Train and test on lipspeakers:

validation error rate:	57.58%
test error rate:	56.68%

- T[illegible]s:

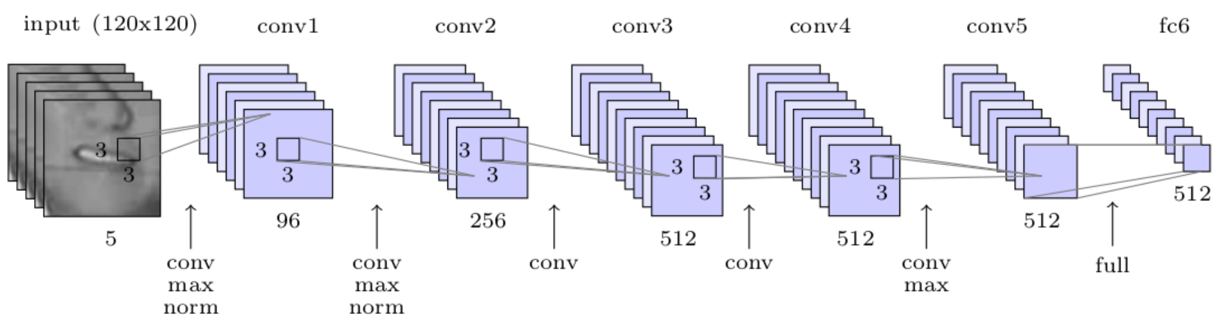
validation error rate:	74.48%
test error rate:	73.53%

- Train on lipspeakers, test on volunteers:

validation error rate:	64.6%
test error rate:	[illegible]1.6%

- Train on volunteers, test on lipspeakers:

validation error rate:	57.58%
test error rate:	56.68%



Conclusion similar to TCDTIMIT paper:
"Visual and audio-visual baseline results on the non-lipspeakers were low overall. Results on the lipspeakers were significantly higher."

KU LEUVEN

6. Lipreading: results CIFAR10

- Train and test on lipspeakers:

validation error rate:	57.05%
test error rate:	57.83%

- Trained and test on volunteers:

validation error rate:	74.48%
test error rate:	72.76%

- Training takes about 10x longer than on Google network (500s/epoch)
- Performance not better
- Some more layers, more parameters
- Good network for lipreading

http://www.cs.toronto.edu/~kriz/cifar.html

6. Lipreading: results ResNet50

- Train and test on lipspeakers:

validation error rate:	61.95%
test error rate:	62.45%

- Trained and test on volunteers:

validation error rate:	74.48%
test error rate:	72.76%

- Training takes about 5x longer than on Google network (500s/epoch)
- Performance not better
- Many more layers, more complex architecture with more parameters
- -> not well suited for lipreading

https://github.com/KaimingHe/deep-residual-networks

KU LEUVEN

6. Lipreading: demo

- Take picture
- Extract face, mouth, convert to grayscale and resize to 120x120x1
- Reshape image for evaluation
- Evaluate, print phoneme predictions

```
1) python preprocessImage.py -i testImages/image.jpg
2) python evaluateImage.py -i testImages/image_mouth_gray_resized.jpg
                           -m results/ResNet50/allLipspeakers/allLipspeakers.npz
```

KU LEUVEN

7. Audio SR

-
- Two-layer LSTM architecture, MFCC as input
-
- Train with noise to make more robust
-
- Two layer LSTM

8. Combining audio and visual

-
- SR: inherent time aspect
- Lipreading: mostly time-independent, could benefit from limited time aspect
-
- Audio and video synchronized thanks to labeled dataset
  - -> possible to combine feature vectors
-
- 'Late fusion': combine output sequences (weighting)
- Weighting determined by:
  - performance of seperate models
  - S/N of audio (if known)
  - Quality of video/image (resolution, face angle, lighting,...)
  -
- Analyse performance:
  - Different amounts of audio and/or image noise
  - Comparison audio only/visual only/ audio-visual

KU LEUVEN