Very Deep Multilingual Convolutional Neural Networks for LVCSR

Tom Sercu[1,2] Christian Puhrsch[2] Brian Kingsbury[1] Yann LeCun[2]

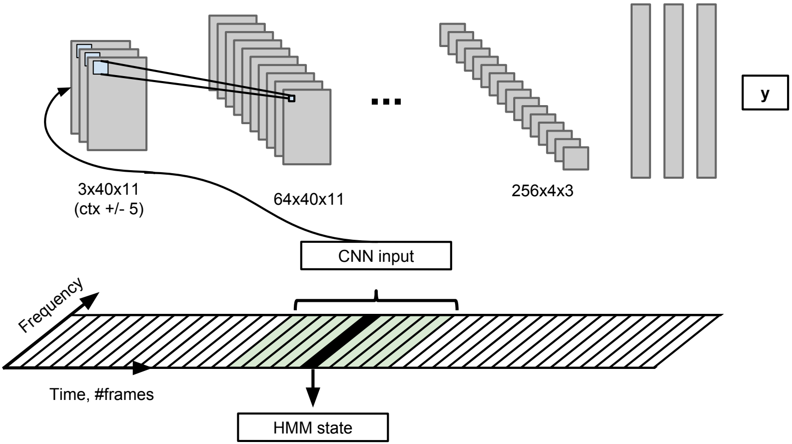
[1]IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598, U.S.A.

[2]Center for Data Science, Courant Institute of Mathematical Sciences, New York University

Convolutional Neural Networks for LVCSR

Hybrid, acoustic model on logmel features

[Abdel-Hamid et al., 2012] [Sainath et al., 2013]



Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

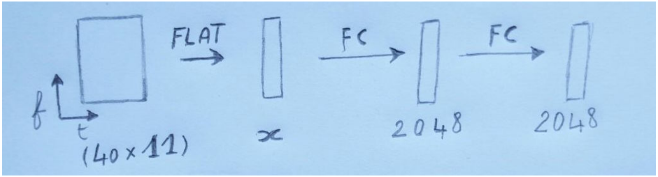
Of course CNNs!

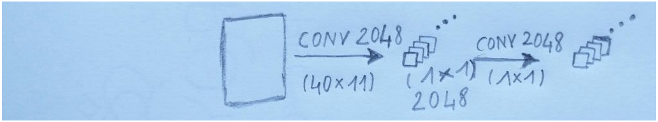
Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

Of course CNNs!

- DNN is a specific type of CNN





Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

Of course CNNs!

- DNN is a specific type of CNN
- Why not keep spatial (time, frequency) resolution?
  - Efficient parametrization
  - Increased depth

Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

Of course CNNs!

- DNN is a specific type of CNN
- Why not keep spatial (time, frequency) resolution?
  - Efficient parametrization
  - Increased depth

But...

Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

Of course CNNs!

- DNN is a specific type of CNN
- Why not keep spatial (time, frequency) resolution?
  - Efficient parametrization
  - Increased depth

But...

- ... the CNN assumptions are broken!
  - Images: good feature detectors are translation invariant
  - Speech: translation invariance in time, frequency?

Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

Of course CNNs!

- DNN is a specific type of CNN
- Why not keep spatial (time, frequency) resolution?
  - Efficient parametrization
  - Increased depth

But...

- ... the CNN assumptions are broken!
  - Images: good feature detectors are translation invariant
  - Speech: translation invariance in **time**, frequency?

Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

Of course CNNs!

- DNN is a specific type of CNN
- Why not keep spatial (time, frequency) resolution?
  - Efficient parametrization
  - Increased depth

But...

- ... the CNN assumptions are broken!
  - Images: good feature detectors are translation invariant
  - Speech: translation invariance in **time**, **frequency**?

Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

Of course CNNs!

- DNN is a specific type of CNN
- Why not keep spatial (time, frequency) resolution?
  - Efficient parametrization
  - Increased depth

But...

- ... the CNN assumptions are broken!
  - Images: good feature detectors are translation invariant
  - Speech: translation invariance in **time**, **frequency**?
- ... aren‘t recurrent networks more powerful?

What just happened in Computer Vision?

VGG Convolutional Neural Networks

IMAGENET

What just happened in Computer Vision?

VGG Convolutional Neural Networks

IMAGENET

- til 2011: Handcrafted + SVM

What just happened in Computer Vision?

VGG Convolutional Neural Networks

IMAGENET

- til 2011: Handcrafted + SVM
- 2012: Alexnet: GPUs, ReLU

What just happened in Computer Vision?

VGG Convolutional Neural Networks

IMAGENET

- til 2011: Handcrafted + SVM
- 2012: Alexnet: GPUs, ReLU
- 2013: Clarifai, Overfeat

What just happened in Computer Vision?

VGG Convolutional Neural Networks

IMAGENET

- til 2011: Handcrafted + SVM
- 2012: Alexnet: GPUs, ReLU
- 2013: Clarifai, Overfeat
- 2014: GoogleNet, VGG net

What just happened in Computer Vision?

VGG Convolutional Neural Networks

- til 2011: Handcrafted + SVM
- 2012: Alexnet: GPUs, ReLU
- 2013: Clarifai, Overfeat
- 2014: GoogleNet, VGG net
- 2015: Residual Networks

What just happened in Computer Vision?

VGG Convolutional Neural Networks

IMAGENET

- til 2011: Handcrafted + SVM
- 2012: Alexnet: GPUs, ReLU
- 2013: Clarifai, Overfeat
- 2014: GoogleNet, **VGG net**
- 2015: Residual Networks

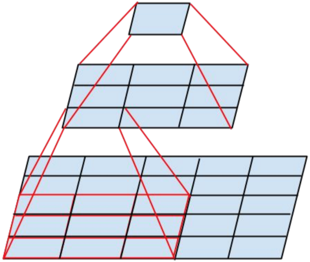
What just happened in Computer Vision?

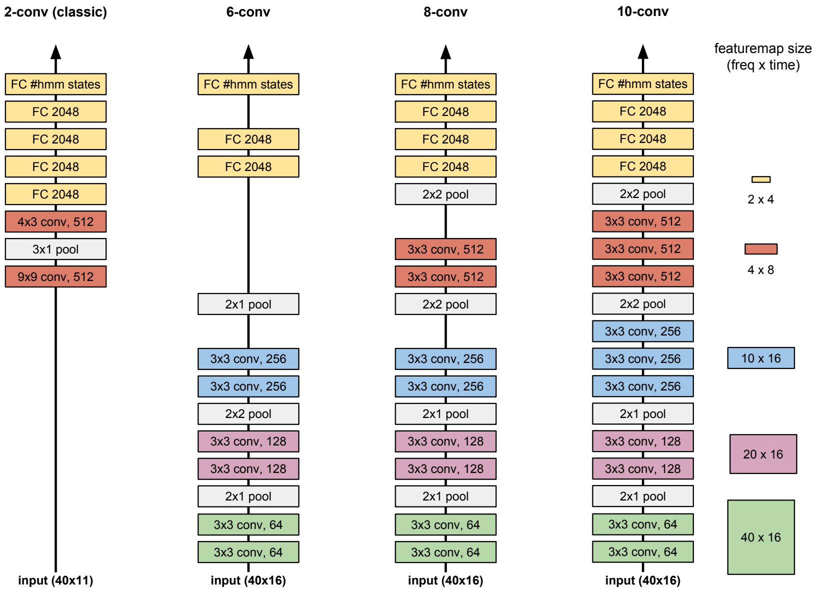
VGG Convolutional Neural Networks

[Simonyan and Zisserman, 2014]

IMAGENET

- til 2011: Handcrafted + SVM
- 2012: Alexnet: GPUs, ReLU
- 2013: Clarifai, Overfeat
- 2014: GoogleNet, **VGG net**
- 2015: Residual Networks




2-conv (classic)
6-conv
8-conv
10-conv
featuremap size
(freq x time)
FC #hmm states
FC 2048
4x3 conv, 512
3x1 pool
9x9 conv, 512
input (40x11)
2x1 pool
3x3 conv, 256
2x2 pool
3x3 conv, 128
3x3 conv, 64
input (40x16)
3x3 conv, 512
2 x 4
4 x 8
10 x 16
20 x 16
40 x 16

Results on 300-h switchboard - CE

	WER (CE)
Classic 512 [Soltau et al., 2014]	13.2
Classic+AD+Maxout [Saon et al., 2015]	12.6
Classic 256 ReLU (Ada+SGD)	
6 conv (Ada+SGD)	
8 conv (SGD)	
10 conv (SGD)	

Results on 300-h switchboard - CE

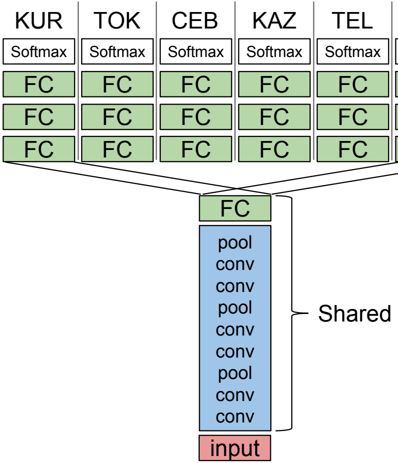
	WER (CE)
Classic 512 [Soltau et al., 2014]	13.2
Classic+AD+Maxout [Saon et al., 2015]	12.6
Classic 256 ReLU (Ada+SGD)	13.8
6 conv (Ada+SGD)	13.1
8 conv (SGD)	
10 conv (SGD)	

Results on 300-h switchboard - CE

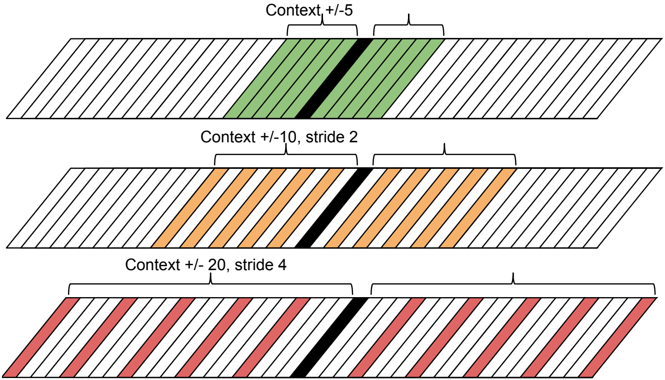
	WER (CE)
Classic 512 [Soltau et al., 2014]	13.2
Classic+AD+Maxout [Saon et al., 2015]	12.6
Classic 256 ReLU (Ada+SGD)	13.8
6 conv (Ada+SGD)	13.1
8 conv (SGD)	11.9
10 conv (SGD)	**11.8**

Multilingual CNN

Model



Multiscale Features



Optimization and tricks

- Optimization
  - Fast: Adam + SGD finetuning
  - Better: Pure SGD (with nesterov acceleration)
- Unbalanced data: sample from $p_i = \frac{f_i^\gamma}{\sum_j f_j^\gamma}$.
- Start from random initialization
  $[-a, a]$ where $a = (\mathrm{kW} \times \mathrm{kH} \times \mathrm{numInputFeatureMaps})^{-\frac{1}{2}}$.

Acknowledgements

Thank you to ...

- My colleagues at IBM
- The torch developers for an amazing deep learning tool
- Christian Szegedy for the figure of slide 3
- The IARPA Babel program

This effort uses the very limited language packs from IARPA Babel Program language collections IARPA-babel205b-v1.0a, IARPA-babel207b-v1.0e, IARPA-babel301b-v2.0b, IARPA-babel302b-v1.0a, IARPA-babel303b-v1.0a, and IARPA-babel304b-v1.0b. This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., and Penn, G. (2012).
Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition.
In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4277–4280. IEEE.

Sainath, T. N., Mohamed, A.-r., Kingsbury, B., and Ramabhadran, B. (2013).
Deep convolutional neural networks for lvcsr.
In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8614–8618. IEEE.

Saon, G., Kuo, H.-K. J., Rennie, S., and Picheny, M. (2015).
The ibm 2015 english conversational telephone speech recognition system.
arXiv preprint arXiv:1505.05899.

Simonyan, K. and Zisserman, A. (2014).
Very deep convolutional networks for large-scale image recognition.
arXiv preprint arXiv:1409.1556.

Soltau, H., Saon, G., and Sainath, T. N. (2014).
Joint training of convolutional and non-convolutional neural networks.
to Proc. ICASSP.

Conclusion

Overview

Very deep convolutional networks

- Small 3×3 kernels
- Multiple convs before pooling
- Best arch: 10 convs, 14 total
- 10.6% improvement over classic CNNs (300h, CE)

Multilingual training

Shared convolutional layers

Multiscale features

Same computation, more context