ConvNets for Speech

NYU Lab presentation

Tom Sercu

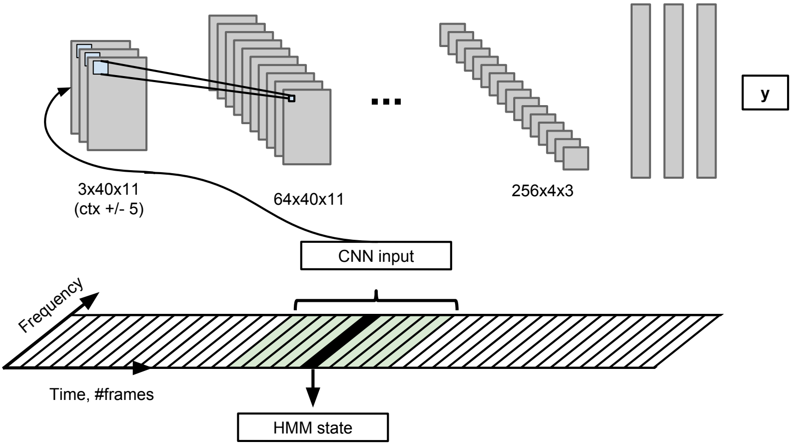
Joint work with Christian Puhrsch Brian Kingsbury Yann LeCun Vaibhava Goel

- Very Deep Multilingual Convolutional Neural Networks for LVCSR http://arxiv.org/abs/1509.08967
- Advances in Very Deep Convolutional Neural Networks for LVCSR http://arxiv.org/abs/1604.01792
- The IBM 2016 English Conversational Telephone Speech Recognition System http://arxiv.org/abs/1604.08242

Convolutional Neural Networks for LVCSR

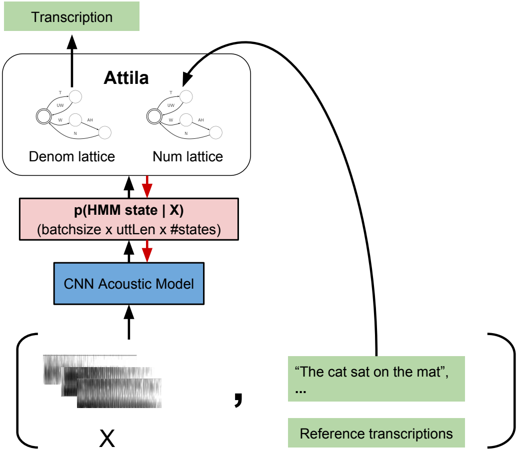
NN-HMM Hybrid, acoustic model on logmel features

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



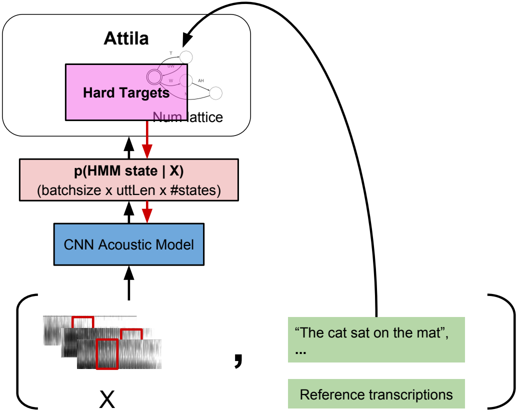
NN-HMM Hybrid speech recognition system

A sloppy picture



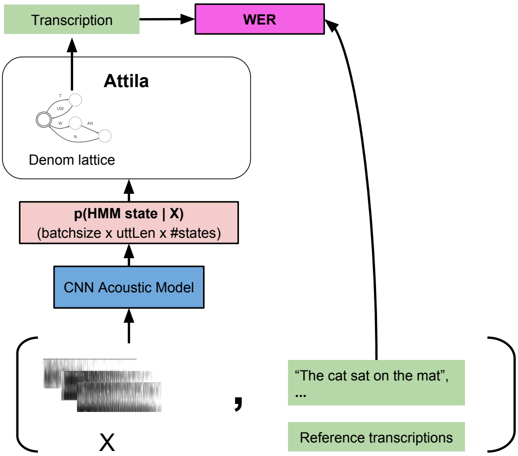
NN-HMM Hybrid speech recognition system

XE Cross-Entropy Training



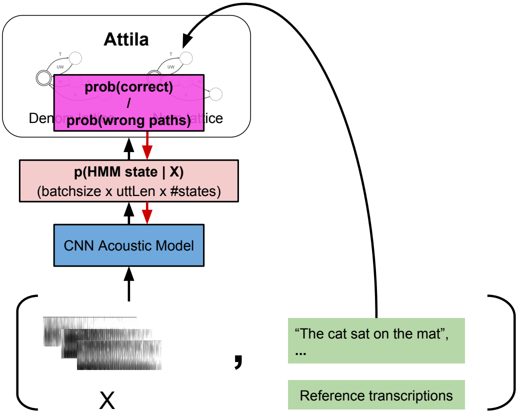
NN-HMM Hybrid speech recognition system

Decoding: getting a WER score



NN-HMM Hybrid speech recognition system

ST Sequence Training



Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

Of course CNNs!

- In speech, all NNs are used convolutional anyway

Convolutional Neural Networks for LVCSR

Why CNN is the right acoustic model

Of course CNNs!

- In speech, all NNs are used convolutional anyway
- So 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!

- In speech, all NNs are used convolutional anyway
- So 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!

- In speech, all NNs are used convolutional anyway
- So 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!

- In speech, all NNs are used convolutional anyway
- So 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!

- In speech, all NNs are used convolutional anyway
- So 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!

- In speech, all NNs are used convolutional anyway
- So 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?

Computer Vision is not Speech Recognition

ImageNet vs Switchboard

	ImageNet	SWB-1 300h	SWB 2000h
# frames/images	1.2M	100M	720M
# classes	1k	8.2k	32k
image size	224×224	40×23	
Class imbalance	No prob	Huge (25% silence)	
Learn Invariance	Viewpoint Illumination Partial obs	Speaker var (Pitch, Accent) Structured Noise, ...	

What just happened in Computer Vision?

VGG Convolutional Neural 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

IMAGENET

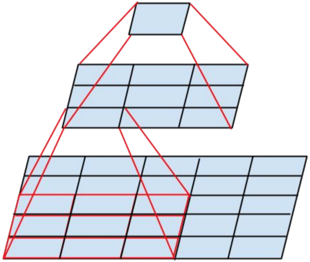
- til 2011: Handcrafted + SVM
- 2012: Alexnet: GPUs, ReLU
- 2013: Clarifai, Overfeat
- 2014: GoogleNet, **VGG net**
- 2015: Residual Networks

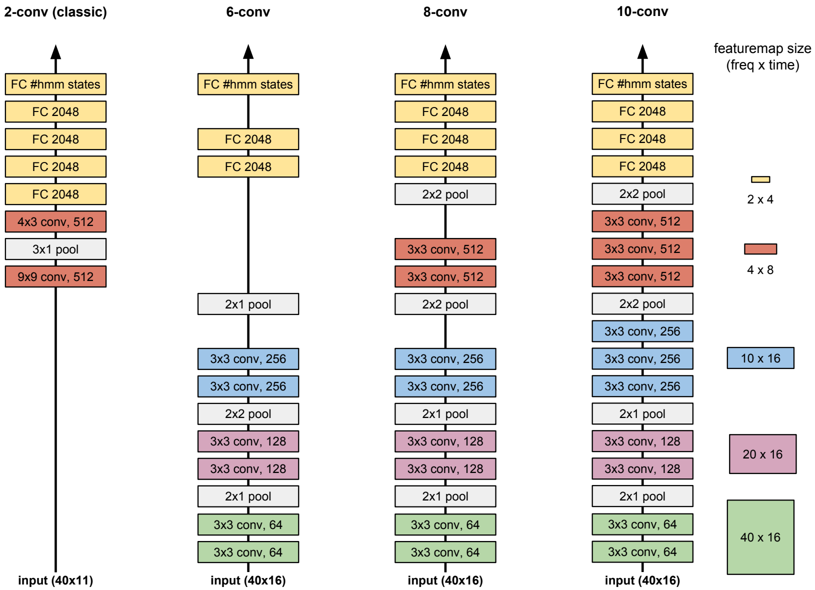
What just happened in Computer Vision?

VGG Convolutional Neural Networks

[Simonyan and Zisserman, 2014]

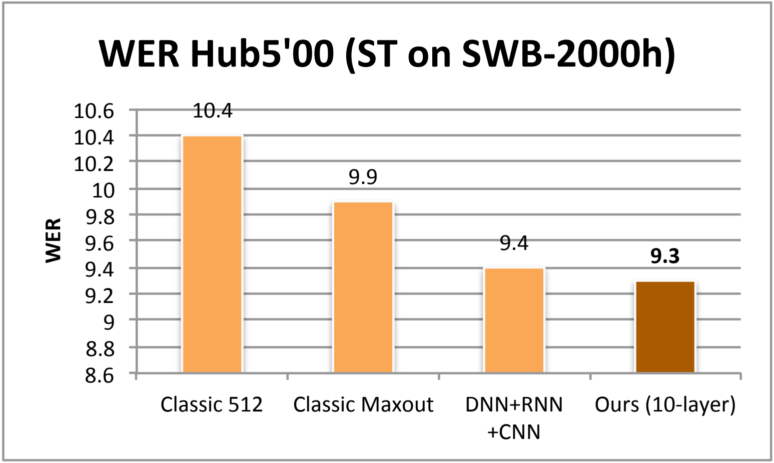
- til 2011: Handcrafted + SVM
- 2012: Alexnet: GPUs, ReLU
- 2013: Clarifai, Overfeat
- 2014: GoogleNet, **VGG net**
- 2015: Residual Networks




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

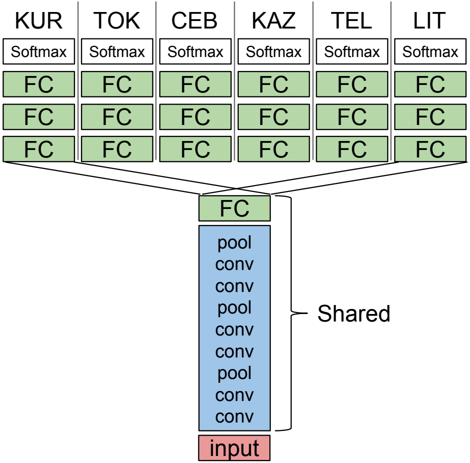
Result on switchboard

A first look



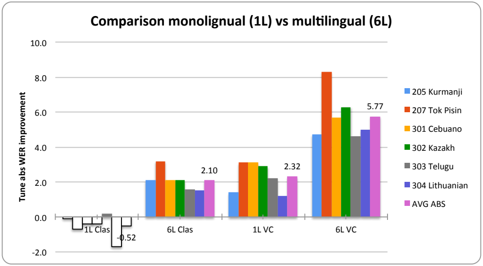
Multilingual CNN

BABEL - Leveraging many small data sets



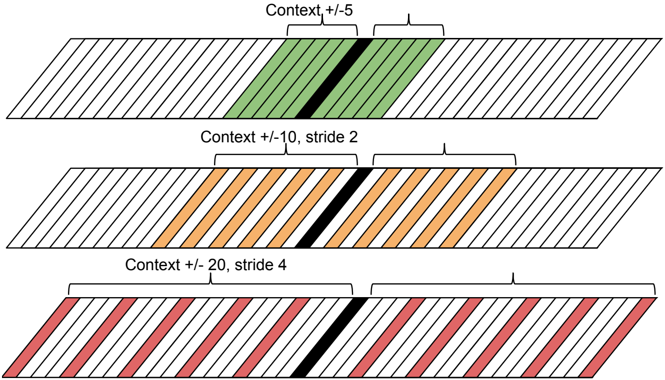
Multilingual CNN

BABEL - Leveraging many small data sets



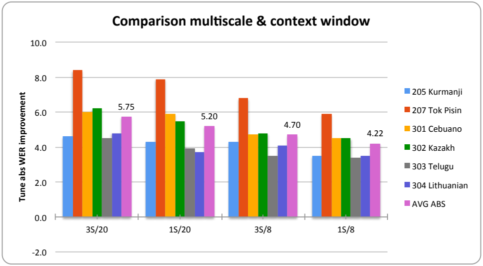
	DNN	1L Clas	6L Clas	1L VC	6L VC
KUR	82.7	82.8	80.6	81.3	78
TOK	62.6	63.3	59.4	59.5	54.3
CEB	76.3	76.7	74.2	73.2	70.6
KAZ	77.3	77.7	75.2	74.4	71
TEL	87.0	86.8	85.4	84.8	82.4
LIT	71.0	72.7	69.5	69.8	66
IMPR	0.00	-0.52	2.10	2.32	**5.77**

Multiscale Features



Multiscale Features

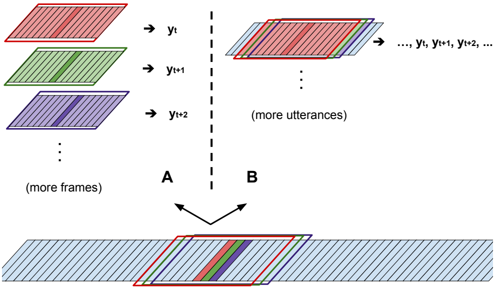
Results on BABEL



	DNN	3S/20	1S/20	3S/8	1S/8
KUR	82.7	78.1	78.4	78.4	79.2
TOK	62.6	54.2	54.7	55.8	56.7
CEB	76.3	70.3	70.4	71.6	71.8
KAZ	77.3	71.1	71.8	72.5	72.8
TEL	87.0	82.5	83.1	83.5	83.6
LIT	71.0	66.2	67.3	66.9	67.5
IMPR	0.00	**5.75**	5.20	4.70	4.22

How will we process a full utterance?

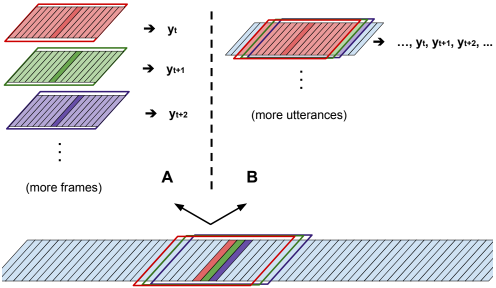
Sequence Training, test time



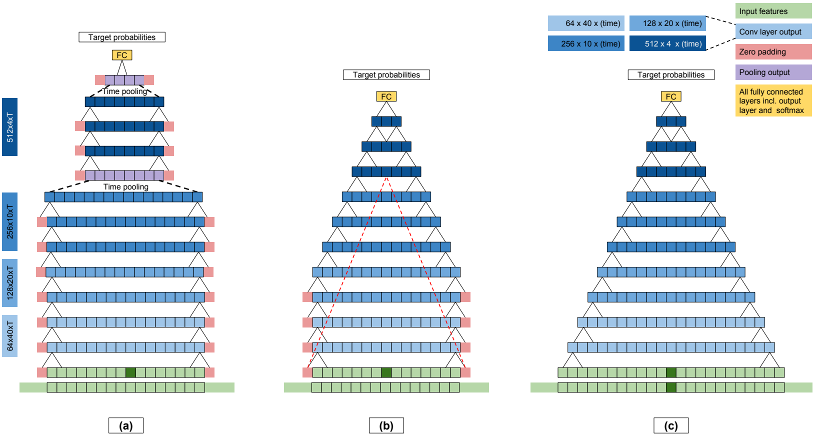
- A: Spliced evaluation, like during Cross-Entropy training
- B: Efficient evaluation

How will we process a full utterance?

Sequence Training, test time



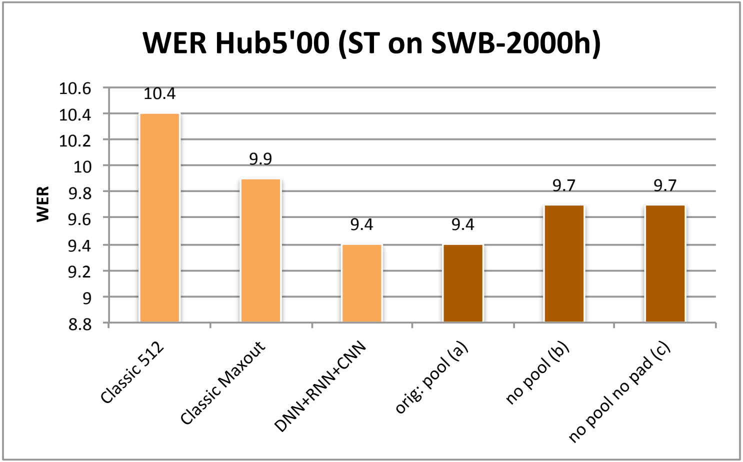
- A: Spliced evaluation, like during Cross-Entropy training
- B: Efficient evaluation - **possible with any model?**



Target probabilities
FC
Time pooling
Time pooling
512x4xT
256x10xT
128x20xT
64x40xT
64 x 40 x (time)
128 x 20 x (time)
256 x 10 x (time)
512 x 4 x (time)
Input features
Conv layer output
Zero padding
Pooling output
All fully connected layers incl. output layer and softmax
(a)
(b)
(c)

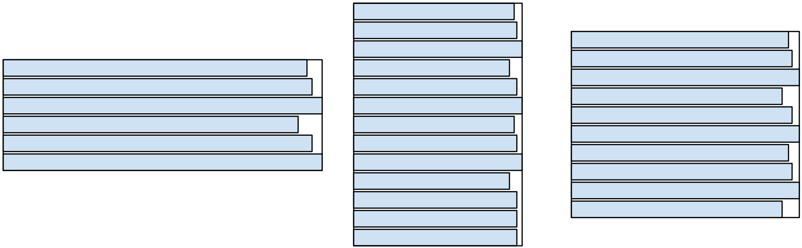
Result on switchboard

Performance hit from architectural constraint



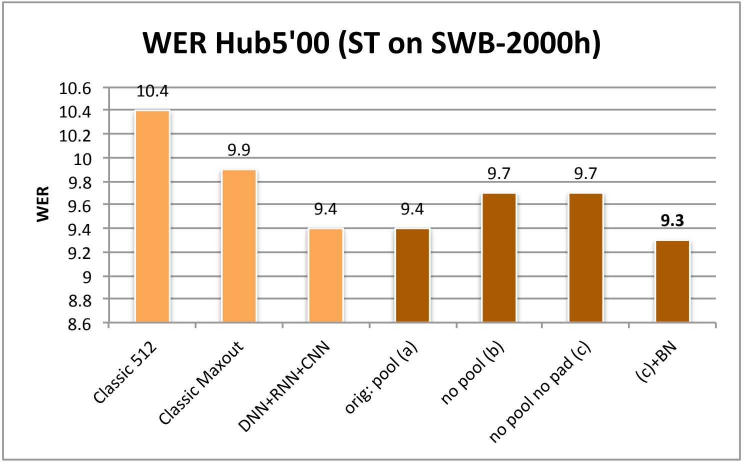
Batch Normalization

- Cross-Entropy training: No Problem.
  Regular Spatial BatchNorm
- During Sequence Training:
  - Spliced: GPU mem is full with 1 utterance
  - Efficient: stack multiple utts in a batch



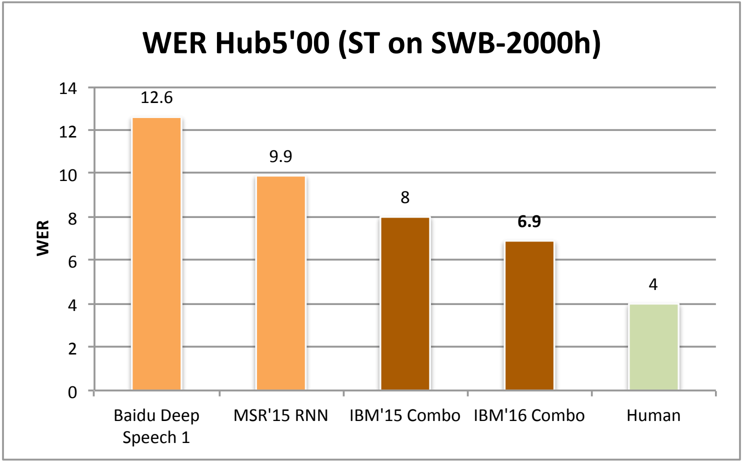
Result on switchboard

Getting performance back with Batch Normalization



Result on switchboard

With all bells and whistles (ensemble, big LM)



Training details

- 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 $a = (\mathrm{kW} \times \mathrm{kH} \times \mathrm{numInputFeatureMaps})^{-\frac{1}{2}}$.

Analysis

Objective mismatch

Some objectives we don't care about

- Frame-level cross-entropy $\mathcal{L}_u = -\sum_t \log y_{ut}(s_{ut})$
- CTC for E2E training of RNNs $\mathcal{L}_u = \sum_{\pi \in B^{-1}(l_u)} \prod_{t=1}^{T_u} y_{\pi_t}^t$
- Expected Sentence Error, e.g. MMI: $\mathcal{L}_u = \log \frac{p(X|S_u)P(W_u)}{\sum_W p(X|S)P(W)}$

Analysis

Objective mismatch

Some objectives we don't care about

- Frame-level cross-entropy $\mathcal{L}_u = -\sum_t \log y_{ut}(s_{ut})$
- CTC for E2E training of RNNs $\mathcal{L}_u = \sum_{\pi \in B^{-1}(l_u)} \prod_{t=1}^{T_u} y_{\pi_t}^t$
- Expected Sentence Error, e.g. MMI: $\mathcal{L}_u = \log \frac{p(X|S_u)P(W_u)}{\sum_W p(X|S)P(W)}$

What we do care about

- Word Error Rate

Analysis

Objective mismatch

Some objectives we don't care about

- Frame-level cross-entropy $\mathcal{L}_u = -\sum_t \log y_{ut}(s_{ut})$
- CTC for E2E training of RNNs $\mathcal{L}_u = \sum_{\pi \in B^{-1}(l_u)} \prod_{t=1}^{T_u} y_{\pi_t}^t$
- Expected Sentence Error, e.g. MMI: $\mathcal{L}_u = \log \frac{p(X|S_u)P(W_u)}{\sum_W p(X|S)P(W)}$

What we do care about

- Word Error Rate (?) – to publish papers

Analysis

Objective mismatch

Some objectives we don't care about

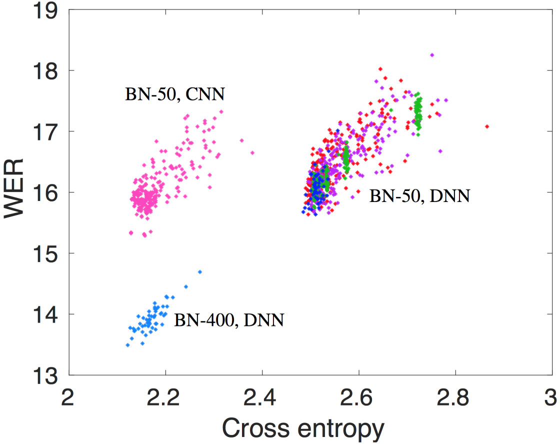
- Frame-level cross-entropy $\mathcal{L}_u = -\sum_t \log y_{ut}(s_{ut})$
- CTC for E2E training of RNNs $\mathcal{L}_u = \sum_{\pi \in B^{-1}(l_u)} \prod_{t=1}^{T_u} y_{\pi_t}^t$
- Expected Sentence Error, e.g. MMI: $\mathcal{L}_u = \log \frac{p(X|S_u)P(W_u)}{\sum_W p(X|S)P(W)}$

What we do care about

- Word Error Rate (?) – to publish papers
- Real life usability
  - Certain words are more important: weighted word error rate?
  - Segmentation into utterances, silence detection
  - Domain mismatch: noise, accents

Analysis

Objective mismatch - How well aligned are XE and WER?



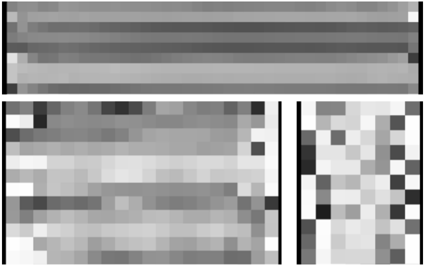
E van den Berg, B Ramabhadran, M Picheny, "Neural network training variance and performance evaluation in speech"

Analysis

- Expect filters to be sensitive to certain frequency regions?

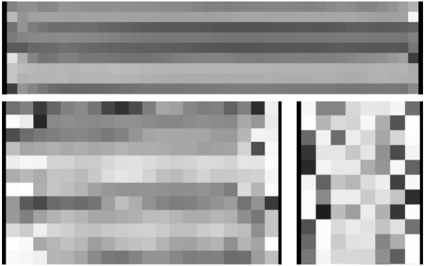
Analysis

- Expect filters to be sensitive to certain frequency regions?



Analysis

- Expect filters to be sensitive to certain frequency regions?



- Help the filters to be sensitive to certain frequency regions
  - Bias per-frequency
  - . . . or batchnorm statistics per (featuremap, frequency)

Acknowledgements

Thank you to . . .

- Collaborators at NYU and IBM
- The torch developers
- 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