Deep Neural Networks in HMM-based and HMM-free Speech Recognition

Andrew Maas

Collaborators: Awni Hannun, Peng Qi, Chris Lengerich, Ziang Xie, and Anshul Samar

Advisors: Andrew Ng and Dan Jurafsky

Outline

- HMM-based speech recognition with DNNs
- What makes HMM-DNN systems work and how can we improve?
- Recurrent DNNs for HMM-free recognition

HMM-GMM Speech Recognition

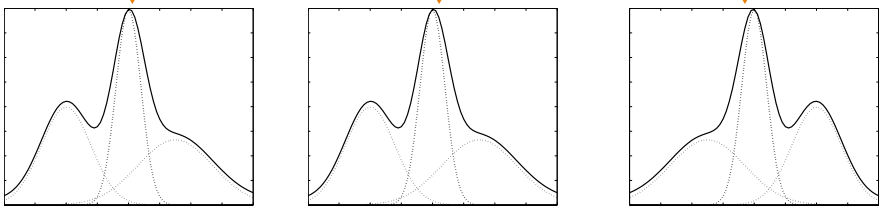
Transcription: Samson

Pronunciation: S – AE – M – S –AH – N

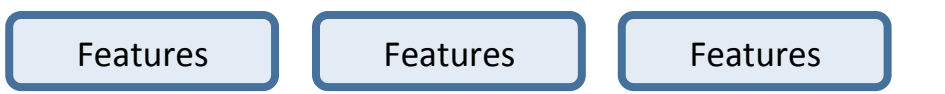
Sub-phones : 942 – 6 – 37 – 8006 – 4422 …

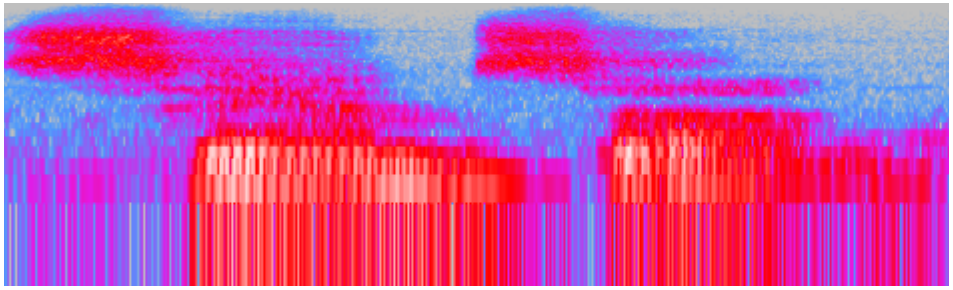
Hidden Markov Model (HMM):

Acoustic Model:



Audio Input:





GMM models:
P(x|s)
x: input features
s: HMM state

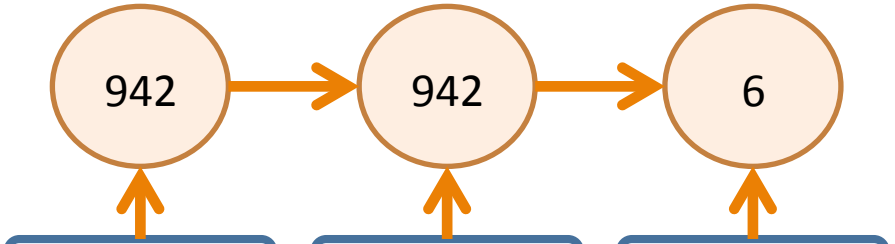
HMM-DNN (Hybrid) Recognition

Transcription: Samson

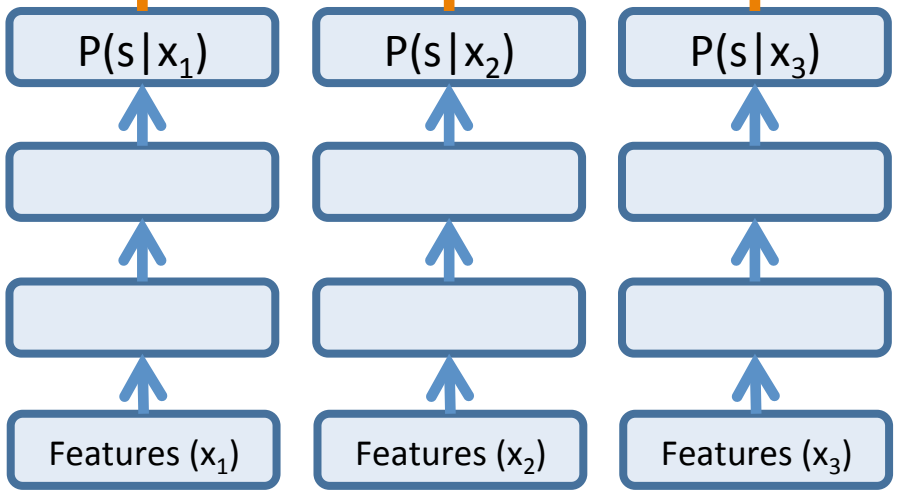
Pronunciation: S – AE – M – S –AH – N

Sub-phones : 942 – 6 – 37 – 8006 – 4422 …

Hidden Markov Model (HMM):



Acoustic Model:

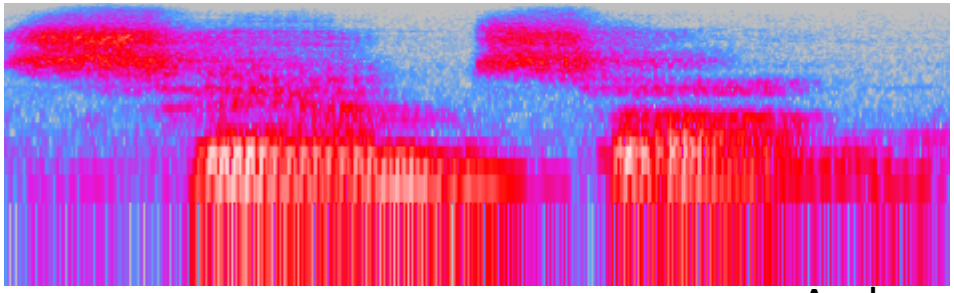


Use a DNN to approximate: $P(s|x)$

Apply Bayes' Rule: $P(x|s) = P(s|x) * P(x) / P(s)$

DNN * Constant / State prior

Audio Input:



Hybrid Systems now Dominate ASR

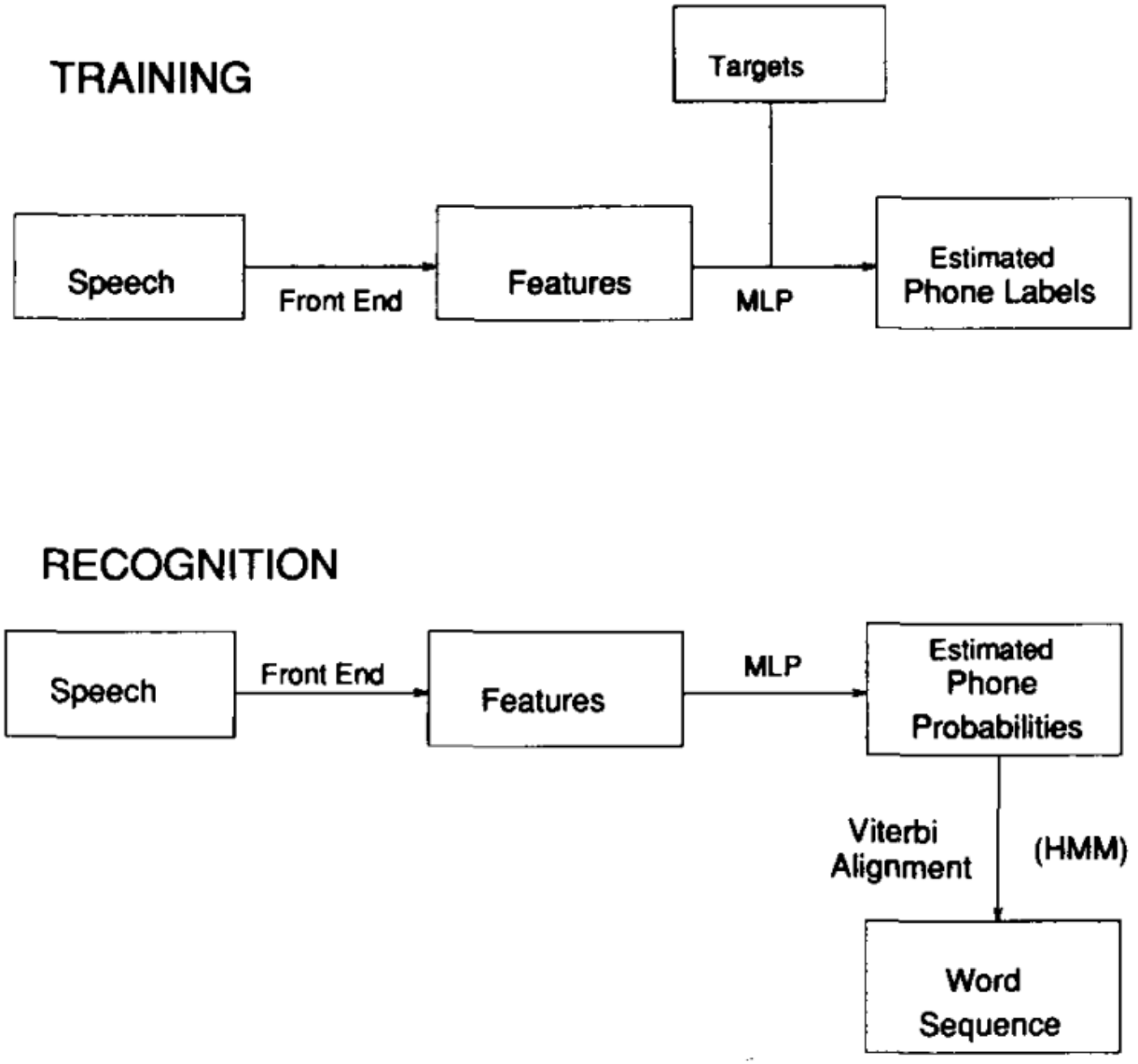
[TABLE 3] A COMPARISON OF THE PERCENTAGE WERs USING DNN-HMMs AND GMM-HMMs ON FIVE DIFFERENT LARGE VOCABULARY TASKS.

TASK	HOURS OF TRAINING DATA	DNN-HMM	GMM-HMM WITH SAME DATA	GMM-HMM WITH MORE DATA
SWITCHBOARD (TEST SET 1)	309	18.5	27.4	18.6 (2,000 H)
SWITCHBOARD (TEST SET 2)	309	16.1	23.6	17.1 (2,000 H)
ENGLISH BROADCAST NEWS	50	17.5	18.8	
BING VOICE SEARCH (SENTENCE ERROR RATES)	24	30.4	36.2	
GOOGLE VOICE INPUT	5,870	12.3		16.0 (>> 5,870 H)
YOUTUBE	1,400	47.6	52.3	

Hinton et al. 2012.

Andrew Maas. NLP Lunch. August 2014

Not Really a New Idea



Renals, Morgan, Bourland, Cohen, & Franco. 1994.

Hybrid MLPs on Resource Management

TABLE I
RESULTS USING THE THREE TEST SETS WITH THE PERPLEXITY 60 WORDPAIR GRAMMAR. (CI-MLP is the context-independent MLP-HMM hybrid system, CD-HMM is the full context-dependent Decipher system, and the MIX system is a simple interpolation between the CD-HMM and the CI-MLP.)

	% error		
Test Set	CI-MLP	CD-HMM	MIX
Feb 91	5.8	3.8	3.2
Sep 92a	10.9	10.1	7.7
Sep 92b	9.5	7.0	5.7

TABLE II
RESULTS USING THE THREE TEST SETS USING NO GRAMMAR (PERLPEXITY 991)

	% error		
Test Set	CI-MLP	CD-HMM	MIX
Feb 91	24.7	19.3	15.9
Sep 92a	31.5	29.2	25.4
Sep 92b	30.9	26.6	21.5

Renals, Morgan, Bourland, Cohen, & Franco. 1994.

Outline

- HMM-based speech recognition with DNNs
- **What makes HMM-DNN systems work and how can we improve?**
- Recurrent DNNs for HMM-free recognition

What's Different in Modern Systems?

- Context-dependent HMM states
- Fast computers = run many DNN experiments
- Deeper nets improve on shallow nets
- Architecture choices (easiest is replacing sigmoid)
- Pre-training *does not matter*. Initially we thought this was the new trick that made things work
- Many more model parameters (scaling up)

Modern Systems use **D**NNs and Senones

COMPARISON OF CONTEXT-INDEPENDENT MONOPHONE STATE LABELS AND CONTEXT-DEPENDENT TRIPHONE SENONE LABELS

# Hidden Layers	# Hidden Units	Label Type	Dev Accuracy
1	2K	Monophone States	59.3%
1	2K	Triphone Senones	68.1%
3	2K	Monophone States	64.2%
3	2K	Triphone Senones	69.6%

Criterion	Dev Accuracy	Test Accuracy
ML	62.9%	60.4%
MMI	65.1%	62.8%
MPE	65.5%	63.8%

Dahl, Yu, Deng & Acero. 2011.

Andrew Maas. NLP Lunch. August 2014

Depth Matters (Somewhat)

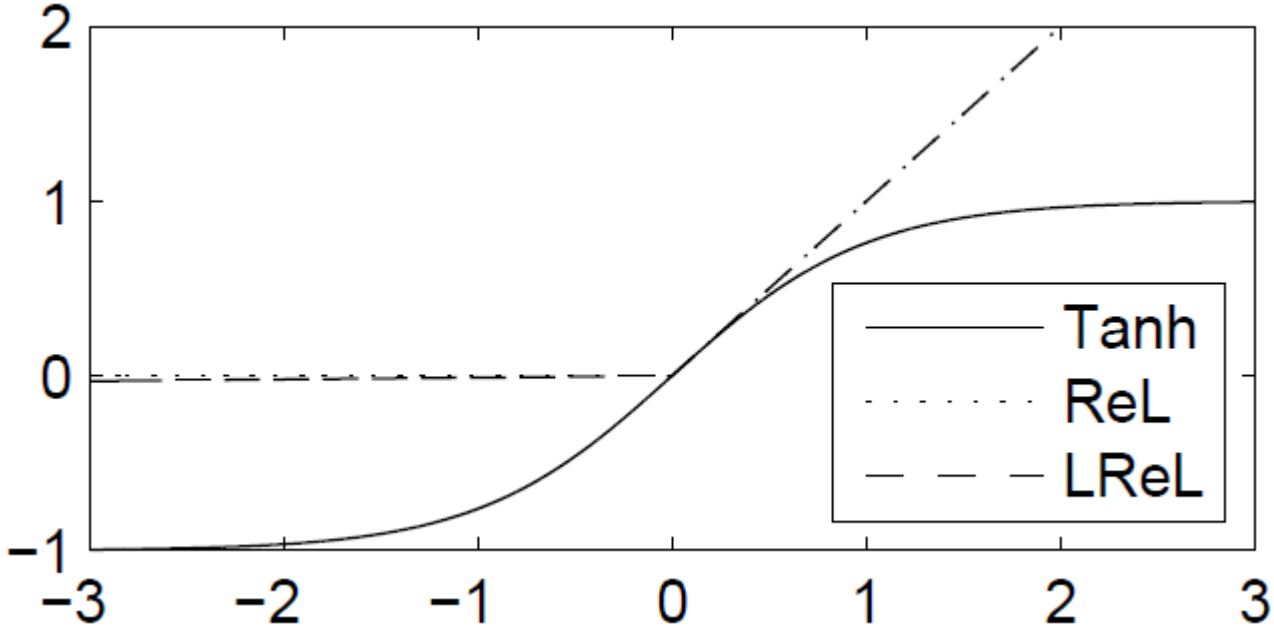
Table 1: Effect of CD-DNN-HMM network depth on WER (%) on Hub5'00-SWB using the 309-hour Switchboard training set. DBN pretraining is applied.

$L \times N$	WER	$1 \times N$	WER
1×2k	24.2	–	–
2×2k	20.4	–	–
3×2k	18.4	–	–
4×2k	17.8	–	–
5×2k	17.2	1×3772	22.5
7×2k	17.1	1×4634	22.6
9×2k	17.0	–	–
5×3k	17.0	–	–
–	–	1×16k	22.1

Warning! Depth can also act as a regularizer because it makes optimization more difficult. This is why you will sometimes see very deep networks perform well on TIMIT or other small tasks.

Yu, Seltzer, Li, Huang, Seide. 2013.

Architecture Choices: Replacing Sigmoids



Rectified Linear (ReL)

$$h^{(i)} = \max(w^{(i)T}x, 0) = \begin{cases} w^{(i)T}x & w^{(i)T}x > 0 \\ 0 & \text{else} \end{cases}$$

[Glorot et al, AISTATS 2011]

Leaky Rectified Linear (LReL)

$$h^{(i)} = \begin{cases} w^{(i)T}x & w^{(i)T}x > 0 \\ 0.01w^{(i)T}x & \text{else} \end{cases}$$

Andrew Maas. NLP Lunch. August 2014

Rectifier DNNs on Switchboard

Model	Dev CrossEnt	Dev Acc(%)
GMM Baseline	N/A	N/A
2 Layer Tanh	2.09	48.0
2 Layer ReLU	1.91	51.7
2 Layer LRelU	1.90	51.8
3 Layer Tanh	2.02	49.8
3 Layer RelU	1.83	53.3
3 Layer LRelU	1.83	53.4
4 Layer Tanh	1.98	49.8
4 Layer RelU	1.79	53.9
4 Layer LRelU	1.78	53.9

Maas, Hannun, & Ng,. 2013.

Rectifier DNNs on Switchboard

Model	Dev CrossEnt	Dev Acc(%)	Switchboard WER
GMM Baseline	N/A	N/A	25.1
2 Layer Tanh	2.09	48.0	21.0
2 Layer ReLU	1.91	51.7	19.1
2 Layer LRelU	1.90	51.8	19.1
3 Layer Tanh	2.02	49.8	20.0
3 Layer RelU	1.83	53.3	18.1
3 Layer LRelU	1.83	53.4	17.8
4 Layer Tanh	1.98	49.8	19.5
4 Layer RelU	1.79	53.9	17.3
4 Layer LRelU	1.78	53.9	17.3
9 Layer Sigmoid CE [MSR]	--	--	17.0
7 Layer Sigmoid MMI [IBM]	--	--	13.7

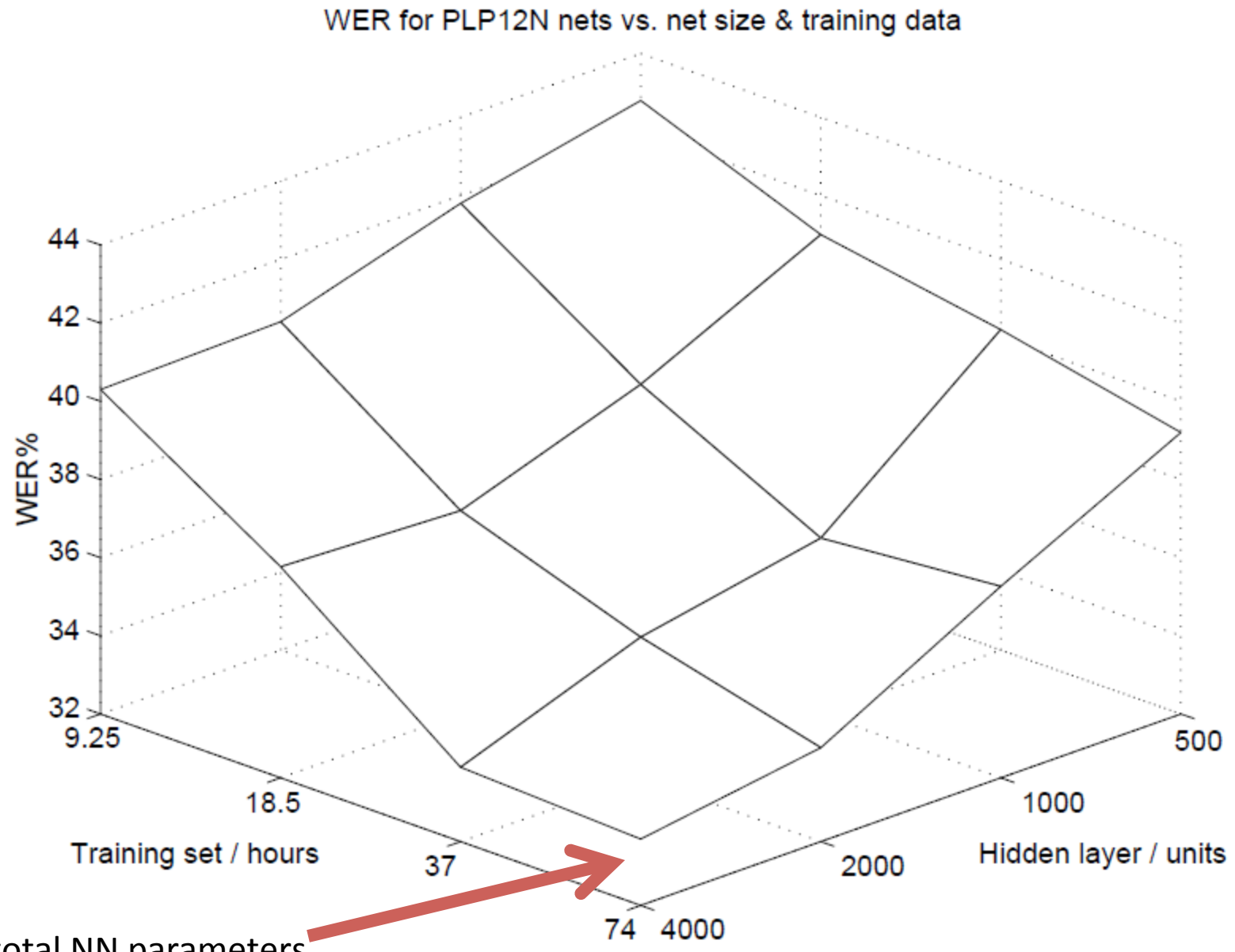
Maas, Hannun, & Ng,. 2013.

Rectifier DNNs on Switchboard

Model	Dev CrossEnt	Dev Acc(%)	Switchboard WER	Callhome WER	Eval 2000 WER
GMM Baseline	N/A	N/A	25.1	40.6	32.6
2 Layer Tanh	2.09	48.0	21.0	34.3	27.7
2 Layer ReLU	1.91	51.7	19.1	32.3	25.7
2 Layer LRelU	1.90	51.8	19.1	32.1	25.6
3 Layer Tanh	2.02	49.8	20.0	32.7	26.4
3 Layer RelU	1.83	53.3	18.1	30.6	24.4
3 Layer LRelU	1.83	53.4	17.8	30.7	24.3
4 Layer Tanh	1.98	49.8	19.5	32.3	25.9
4 Layer RelU	1.79	53.9	17.3	29.9	23.6
4 Layer LRelU	1.78	53.9	17.3	29.9	23.7
9 Layer Sigmoid CE [MSR]	--	--	17.0	--	--
7 Layer Sigmoid MMI [IBM]	--	--	13.7	--	--

Maas, Hannun, & Ng,. 2013.

Scaling up NN acoustic models in 1999



0.7M total NN parameters

[Ellis & Morgan. 1999]

Adding More Parameters 15 Years Ago

Size matters: An empirical study of neural network training for LVCSR. Ellis & Morgan. ICASSP. 1999.

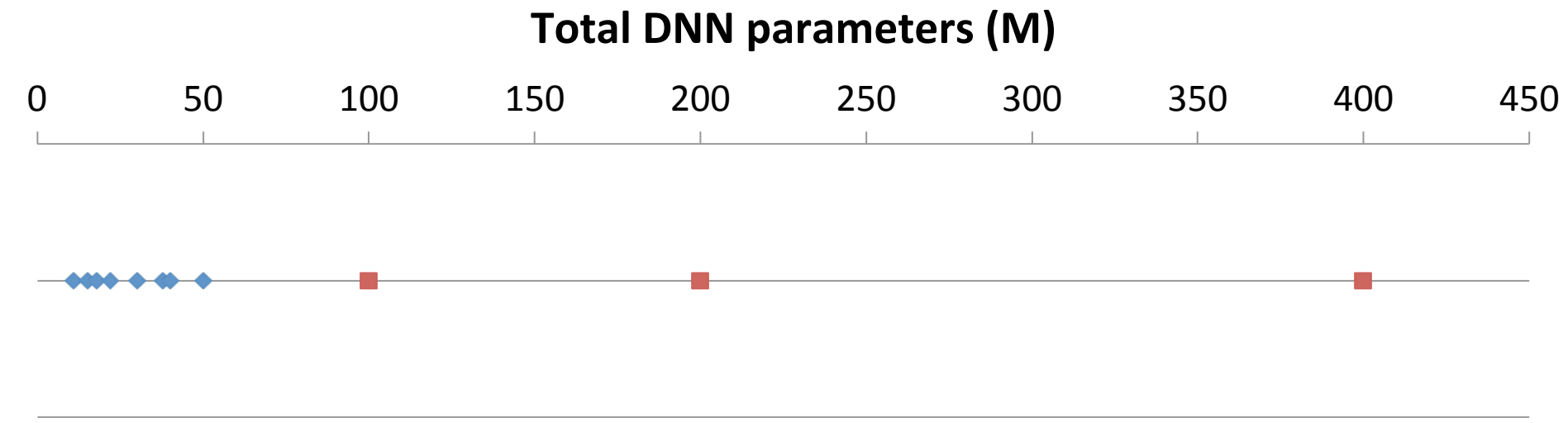
Hybrid NN. 1 hidden layer. 54 HMM states.

74hr broadcast news task

"…improvements are almost always obtained by increasing either or both of the amount of training data or the number of network parameters … We are now planning to train an 8000 hidden unit net on 150 hours of data … this training will require over three weeks of computation."

Adding More Parameters Now

- Comparing total number of parameters (in millions) of previous work versus our new experiments



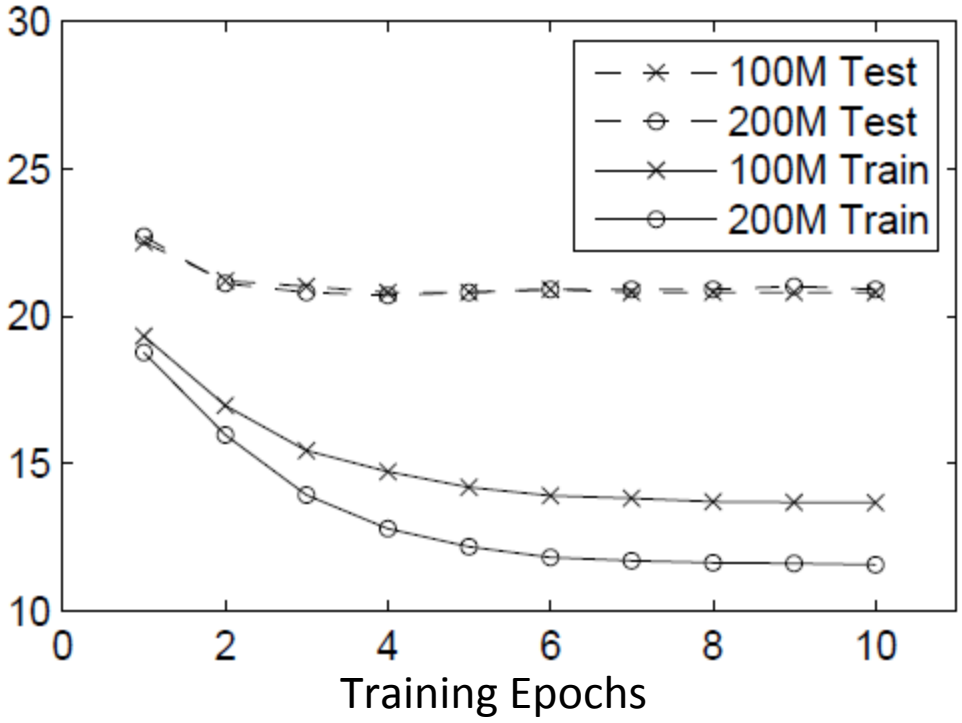
Experimental Framework

- 1-4 GPUs using the infrastructure of Coates et al (ICML 2013)
- Improved baseline GMM system. ~9,000 HMM states
- Fix DNN hidden layers to 5
- Vary total number of parameters, same number of hidden units in each hidden layer
- Evaluate input context of 21 and 41 frames
- Sizes evaluated: 36M, 100M, 200M
- Layer sizes: 2048, 3953, 5984
- Output layer:
  51% of all parameters for a 36M DNN
  6% of all parameters for a 200M DNN

Frame-level results

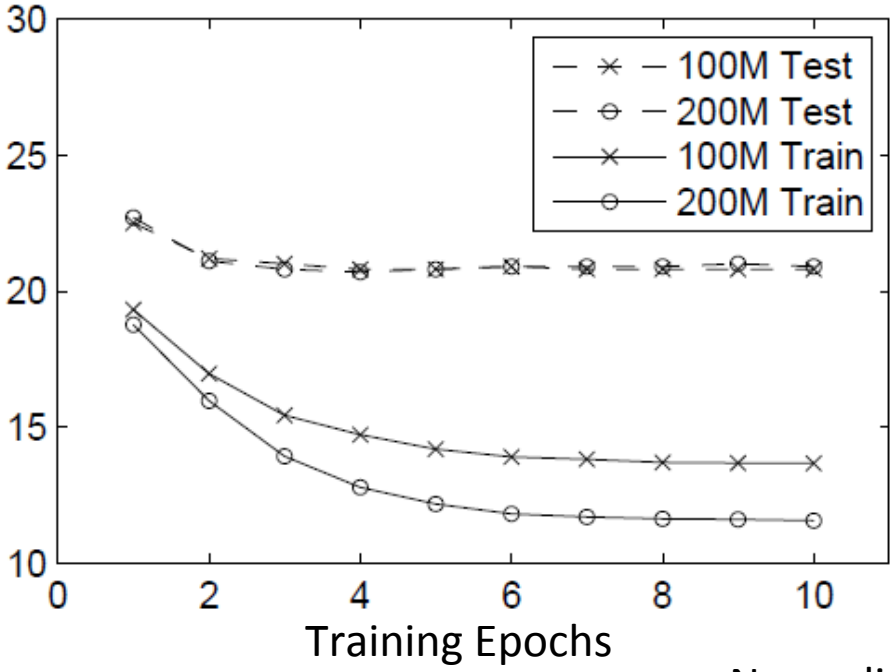
Model Size	Layer Size	Input	Dev CrossEnt	Dev Acc(%)
GMM Baseline	N/A	1	N/A	N/A
36M	2048	21	1.23	66.20
100M	3953	21	.77	78.56
100M	3953	41	.50	85.58
200M	5984	21	.51	86.06
200M	5984	41	.26	93.05

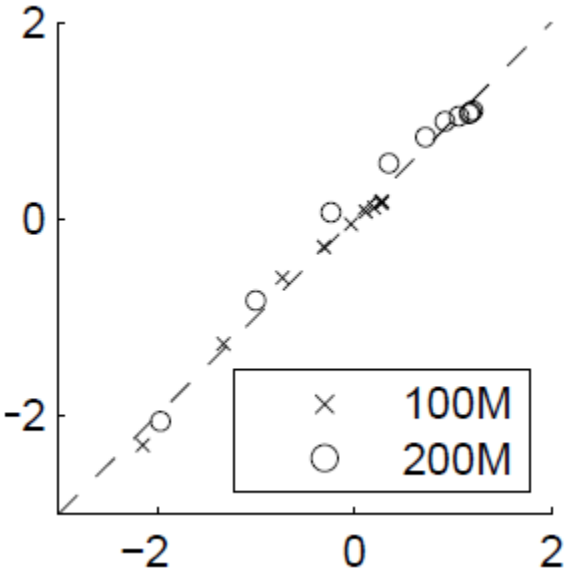
Word error rate during training



Andrew Maas. NLP Lunch. August 2014

Correlating frame-level metrics and WER





Normalize training set word accuracy and negative cross entropy separately. Strong correlation (r=0.992) suggests that cross entropy is a good predictor of *training set* WER

Generalization: Early realignment

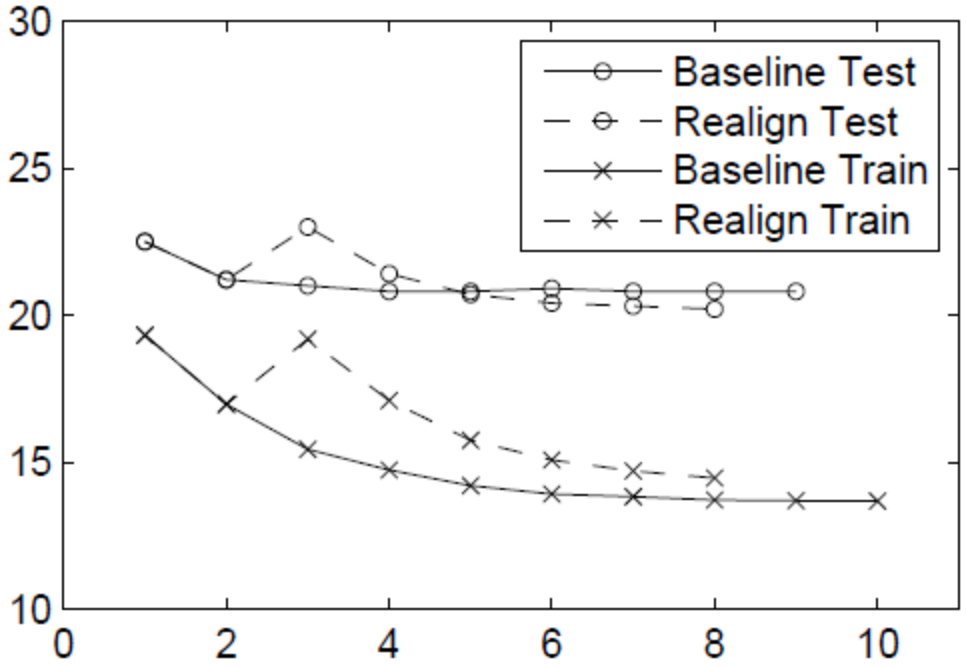


Figure 3. Word error rates on the training and test sets for LVCSR systems with DNN acoustic models trained with and without label realignment after epoch 2. A DNN which re-generates its training labels with a forced alignment early during optimization generalizes much better to test data than a DNN which converges to the original labels.

Generalization: Dropout

- During training randomly zero out hidden unit activations with probability p=0.5
- Cross validate over p in initial experiments
- Previous work found dropout helped on 50hr broadcast news but did not directly evaluate dropout with control experiments (Dahl et al 2013, Sainath et al 2013)

Generalization: Dropout

Model	SWBD	CH	EV
GMM Baseline	21.7	36.1	29.0
2048 Layer (36M)	15.1	27.1	21.2
2048 Layer (36M) DO	14.7	26.7	20.8
3953 Layer (100M)	14.7	26.7	20.7
3953 Layer (100M) DO	14.6	26.3	20.5
5984 Layer (200M)	15.0	26.9	21.0
5984 Layer (200M) DO	14.9	26.3	20.7

Increasing training set size

- Switchboard:
  Training data: 300hrs. 4,870 speakers
  Baseline GMM: 8,986 HMM states. 200k Gaussians
- Fisher:
  Training data: 2,000hrs. 23,394 speakers
  Baseline GMM: 7,793 HMM states. 300k Gaussians
  (Baseline kept a bit weak for comparison)
  Train on 300hr subset as well as full training set

Evaluating SWBD systems on Fisher

Model	Training hours	Dev CrossEnt	Dev Acc(%)	SWBD WER	CH WER	EV WER	FSH WER
SWBD GMM	300	N/A	N/A	21.7	36.1	29.0	33.9
SWBD 36M	300	1.23	66.2	15.1	27.1	21.2	25.1
SWBD 100M	300	0.77	78.5	14.5	27.0	20.8	25.3
SWBD 200M	300	0.51	86.0	15.0	26.8	20.9	25.9

Sample of Results

- 2,000 hours of conversational telephone speech
- Kaldi baseline recognizer (GMM)
- DNNs take 1 -3 weeks to train

Acoustic Model	Training hours	Dev CrossEnt	Dev Acc(%)	FSH WER
GMM	2,000	N/A	N/A	32.3
DNN 36M	300	2.23	49.9	24.2
DNN 200M	300	2.34	49.8	23.7
DNN 36M	2,000	1.99	53.1	23.3
DNN 200M	2,000	1.91	55.1	21.9

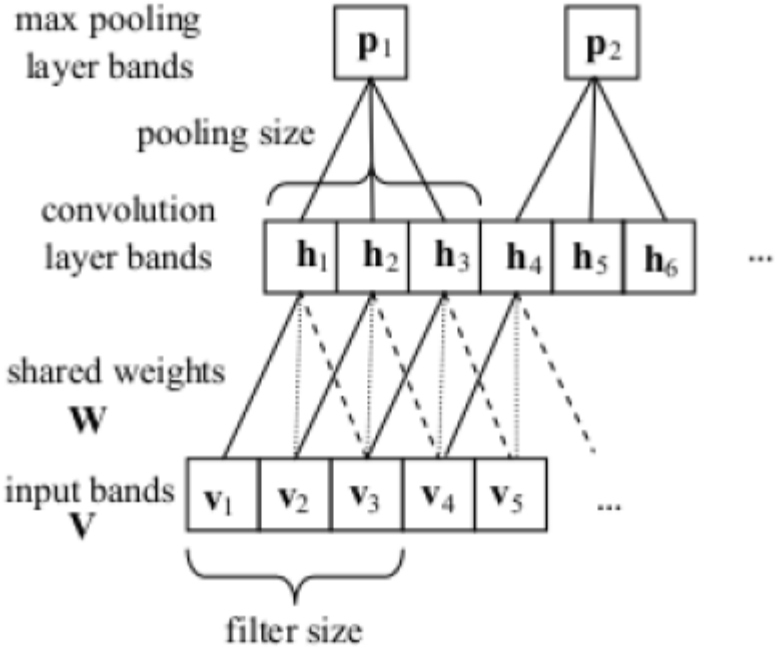
Maas, Hannun, Qi, Lengerich, Ng, & Jurafsky. ArXiv.

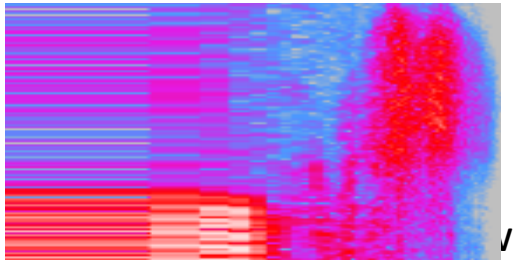
Current Work: What Aspects Matter?

- Number of layers
- Optimization algorithm
- Total number of DNN parameters
- Input features (is speaker adaptation necessary?)
- Training set size (we have more or less run out of academically available speech data)

Current Work: Convolutional Networks

- Slide your filters along the frequency axis of filterbank features
- Great for spectral distortions (eg. Short wave radio)

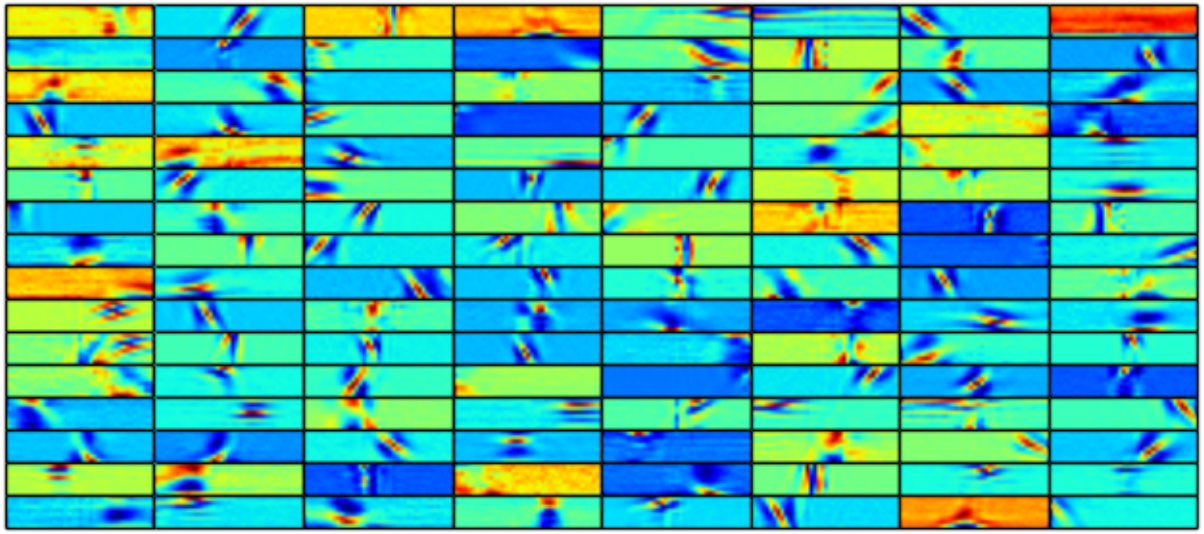




Sainath, Mohamed, Kingsbury, & Ramabhadran . 2013.

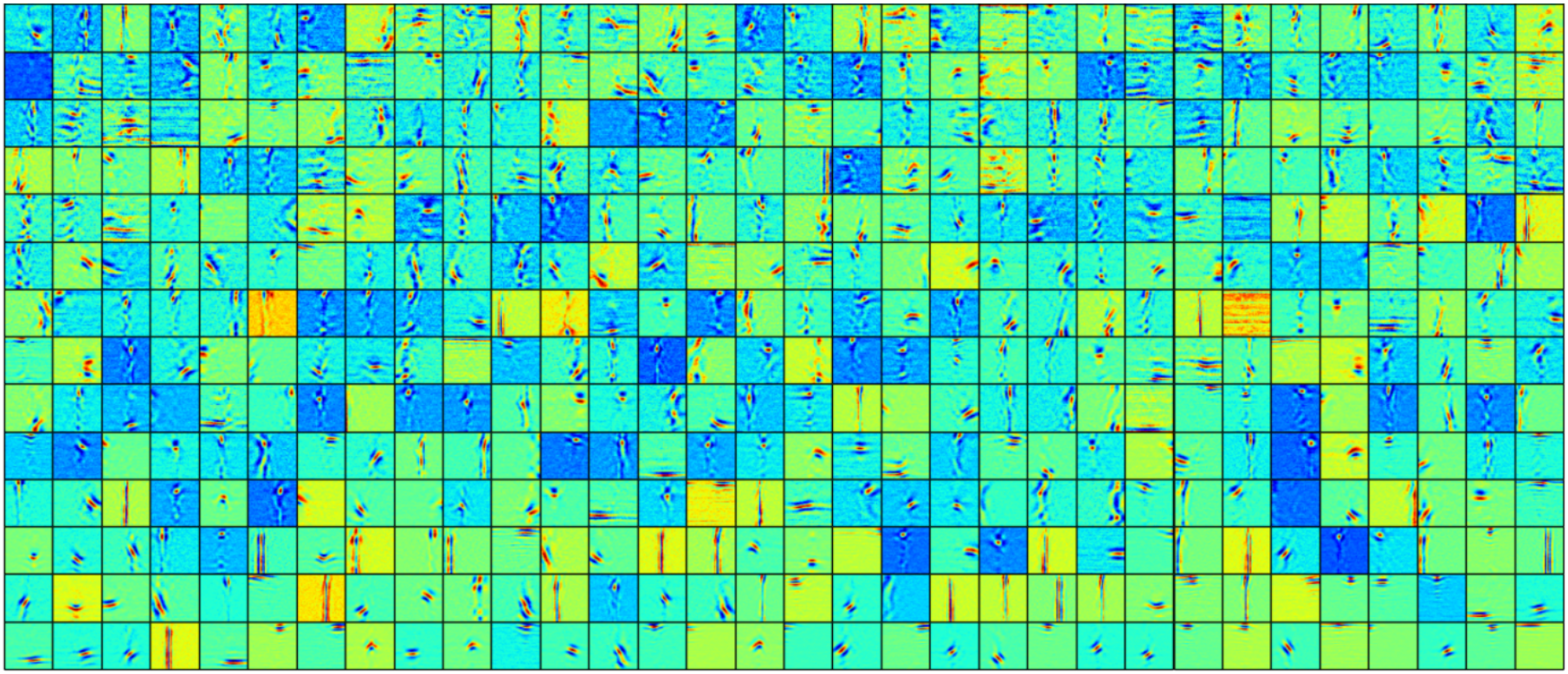
v Maas. NLP Lunch. August 2014

Learned Convolutional Features



Andrew Maas. NLP Lunch. August 2014

Learned DNN Features



Andrew Maas. NLP Lunch. August 2014

Outline

- HMM-based speech recognition with DNNs
- What makes HMM-DNN systems work and how can we improve?
- **Recurrent DNNs for HMM-free recognition**

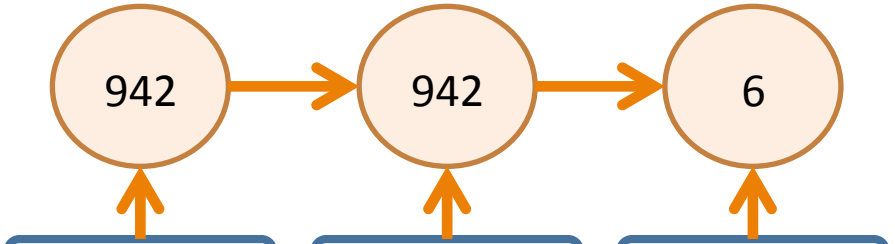
HMM-DNN (Hybrid) Recognition

Transcription: Samson

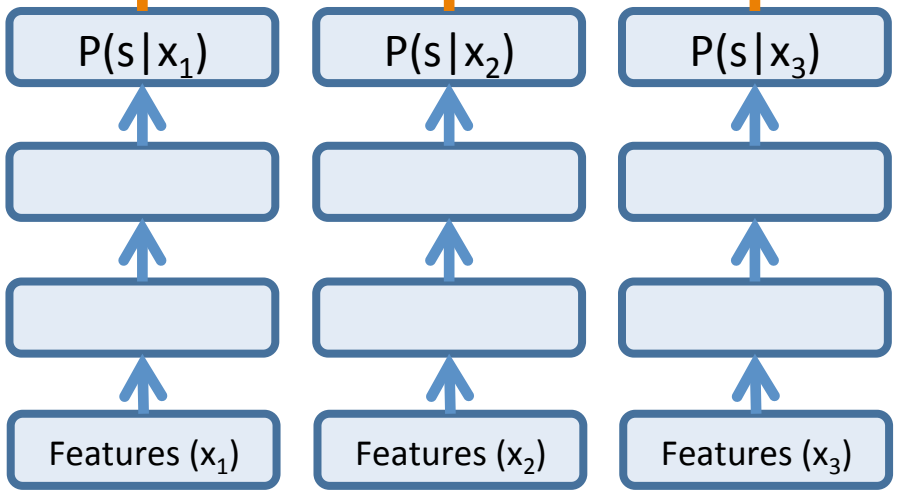
Pronunciation: S – AE – M – S –AH – N

Sub-phones : 942 – 6 – 37 – 8006 – 4422 …

Hidden Markov Model (HMM):



Acoustic Model:

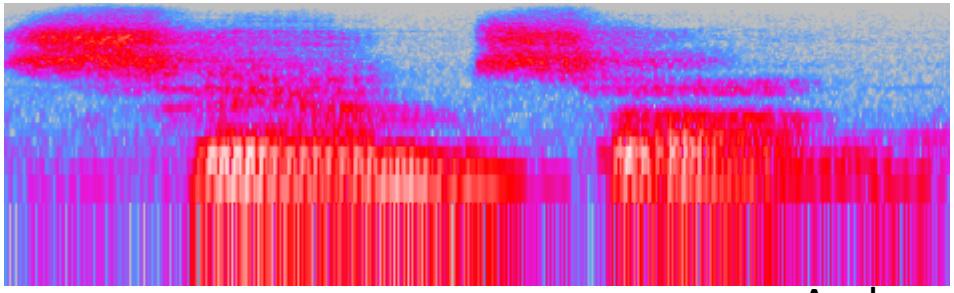


Use a DNN to approximate: $P(s|x)$

Apply Bayes' Rule: $P(x|s) = P(s|x) * P(x) / P(s)$

DNN * Constant / State prior

Audio Input:



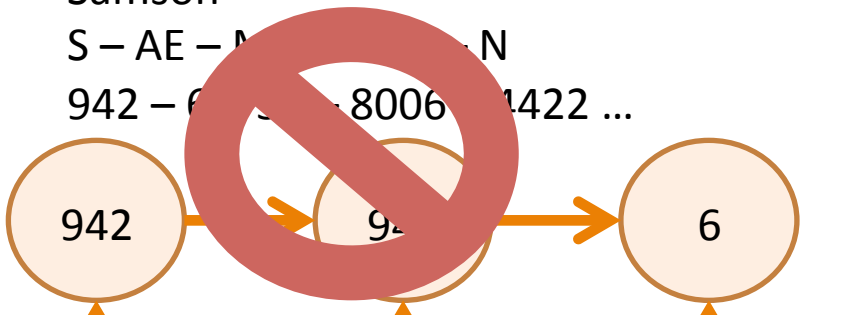
HMM-Free Recognition with CTC

Transcription: Samson

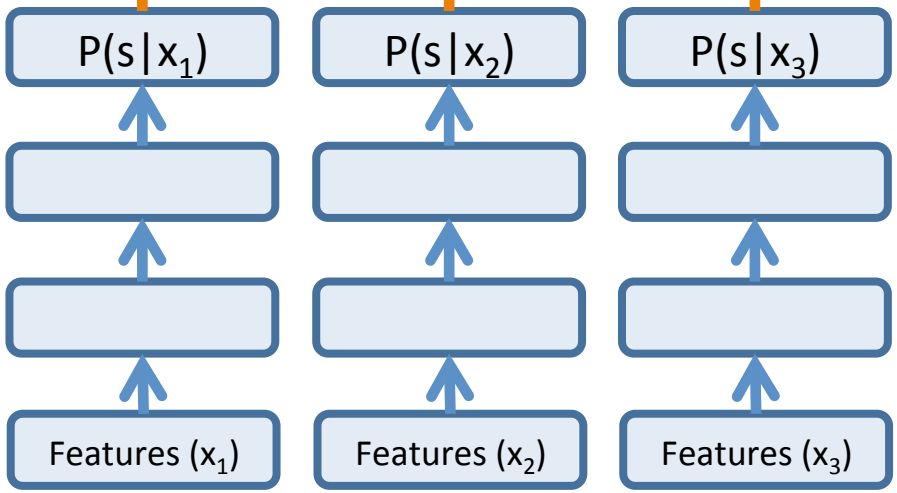
Pronunciation: S – AE – M – N

Sub-phones : 942 – 6 – 8006 – 1422 …

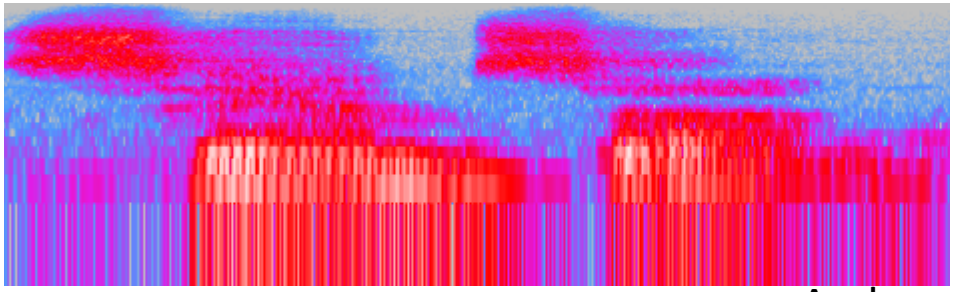
Hidden Markov Model (HMM):



Acoustic Model:



Audio Input:



(Graves & Jaitly, ICML 2014)

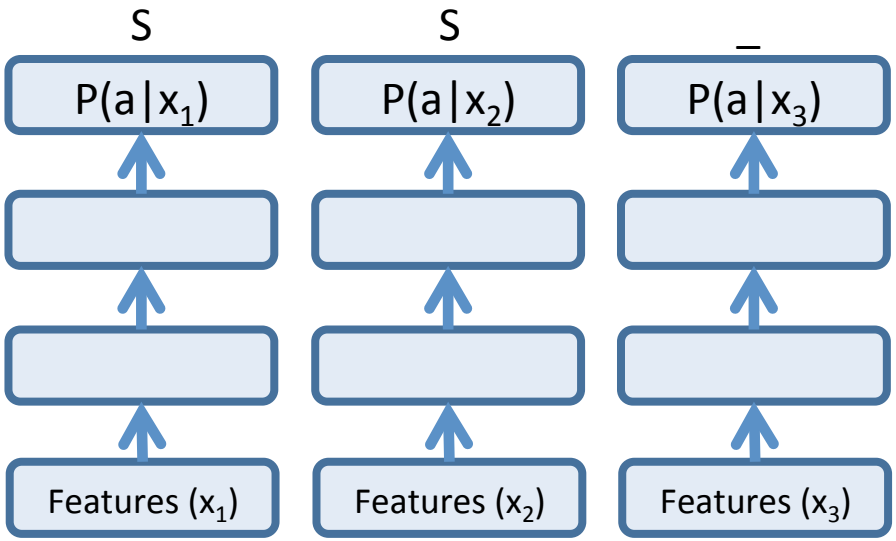
HMM-Free Recognition with CTC

Transcription: Samson

Characters: SAMSON

Collapsing function: SS___AA_M_S___O___NNNN

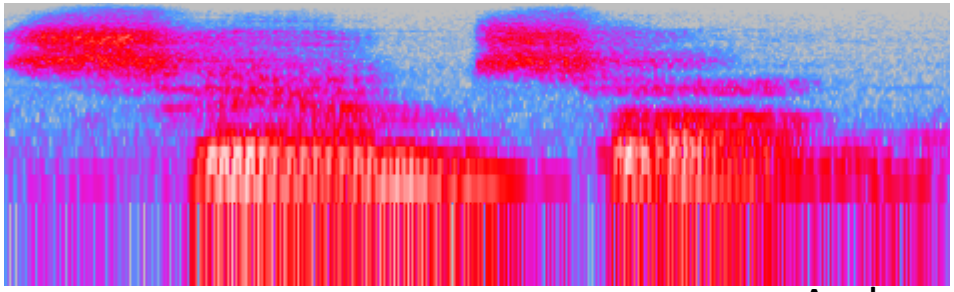
Acoustic Model:



Use a DNN to approximate: P(a|x)

The distribution over *characters*

Audio Input:



(Graves & Jaitly, ICML 2014)

Collapsing Example (WSJ Corpus)

Per-frame argmax:

_____
_____YY__EE_____TT_<><>_____
A_____<><>_RR__E_____HH_____B___II_____LLL__I_____TT_____AA_____TT_____
_IIO__N___<><>___CC_____RRR_U_____<><>_____II___SS<><>_____
___O_____NN_____HHH_A_____NNDDD<><>_____I_
_N___<>__THH_E_____<><>_____BB_UUII_____LLLL
DD____II____NNG_____<><>_____L___O___O_G__G___II____
NNG_____<><>____B___RR_II_____CK__S<><>_____
__P___LL__A_____SSTT_____EERR__<><>_____A___NND_<><>___B___LLL_UU____EE__P
P___R___I_____NNSS_<><>_____F_____OOU_____RRR_____<>_____
_____F_____OO__RRR__TT_Y____<><>_____T____WWW_OO_____<><><>____NN___EW
_____<><>_____B___E_
_____T_____I____N___<><>____E_____PP_____AA___RR___TT____MM_EE___NNNTSS
<><><>_____
_____

After collapsing:

YET<>A<>REHBILITATION<>CRU<>IS<>ONHAND<>IN<>THE<>BUILDING<>LOOGGING<>BRICKS<>PLAS
TER<>AND<>BLUEPRINS<>FOUR<>FORTY<>TWO<>NEW<>BETIN<>EPARTMENTS<>

Reference:

YET<>A<>REHABILITATION<>CREW<>IS<>ON<>HAND<>IN<>THE<>BUILDING<>LUGGING<>BRICKS<>P
LASTER<>AND<>BLUEPRINTS<>FOR<>FORTY<>TWO<>NEW<>BEDROOM<>APARTMENTS<>

More Example Results (WSJ)

YET<>A<>REHBILITATION<>CRU<>IS<>ONHAND<>IN<>THE<>BUILDING<>LOOGGING<>BRICKS<>PLASTER<>AND<>BLUEPRINS<>FOUR<>FORTY<>TWO<>NEW<>BETIN<>EPARTMENTS<>

YET<>A<>REHABILITATION<>CREW<>IS<>ON<>HAND<>IN<>THE<>BUILDING<>LUGGING<>BRICKS<>PLASTER<>AND<>BLUEPRINTS<>FOR<>FORTY<>TWO<>NEW<>BEDROOM<>APARTMENTS<>

THIS<>PARCLE<>GUNA<>COME<>BACK<>ON<>THIS<>ILAND<>SOM<>DAY<>SOO<>

THE<>SPARKLE<>GONNA<>COME<>BACK<>ON<>THIS<>ISLAND<>SOMEDAY<>SOON<>

TRADE<>REPRESENTIGD<>JUIDER<>WARANTS<>THAT<>THE<>U<>S<>WONT<>BACKCOFF<><>ITS<>PUSH<>FOR<>TRADE<>BARIOR<>REDUCTIONS<>

TRADE<>REPRESENTATIVE<>YEUTTER<>WARNS<>THAT<>THE<>U<>S<>WONT<>BACK<>OFF<>ITS<>PUSH<>FOR<>TRADE<>BARRIER<>REDUCTIONS<>

TREASURY<>SECRETARY<>BAGER<>AT<>ROHIE<>WOS<>IN<>AUGGRAL<>PRESSED<>FOUR<>ARISE<>IN<>THE<>VALUE<>OF<>KOREAS<>CURRENCY<>

TREASURY<>SECRETARY<>BAKER<>AT<>ROH<>TAE<>WOOS<>INAUGURAL<>PRESSED<>FOR<>A<>RISE<>IN<>THE<>VALUE<>OF<>KOREAS<>CURRENCY<>

CTC Objective Function

Labels at each time index are conditionally independent (like HMMs)

$$\Pr(\mathbf{a}|x) = \prod_{t=1}^{T} \Pr(a_t, t|x)$$

Sum over all time-level labelings consistent with the output label.

$$\Pr(y|x) = \sum_{a \in \mathcal{B}^{-1}(y)} \Pr(a|x)$$

Output label: AB

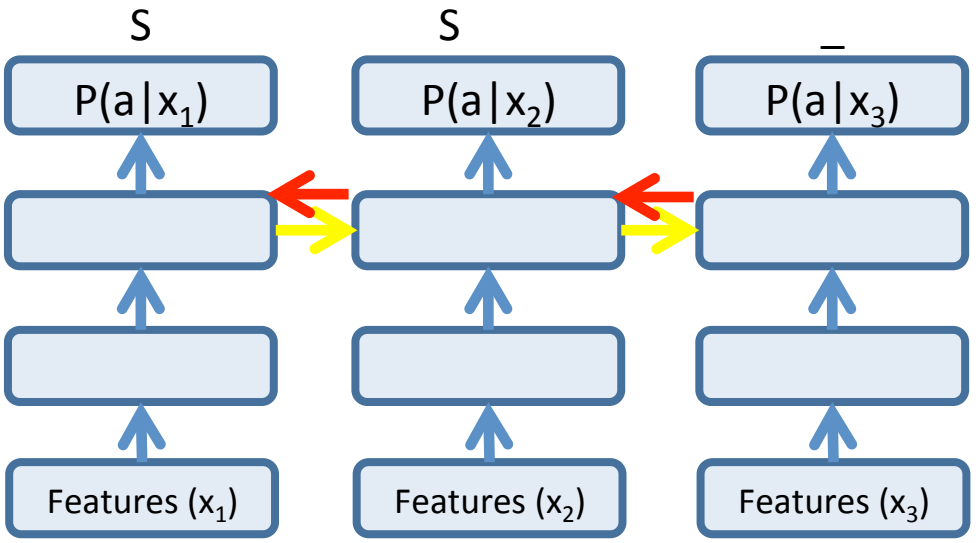
Time-level labelings: AB, _AB, A_B, … _A_B_

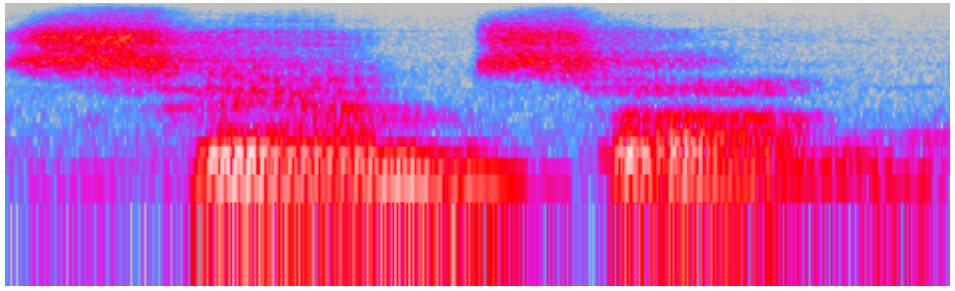
Final objective maximizes probability of true labels:

$$CTC(x) = -\log \Pr(y^*|x)$$

(Graves & Jaitly, ICML 2014)

Network Architecture: Recurrence Matters!





Architecture	CER
DNN	22
RNN	13
BRNN	10

Decoding with a Language Model

Language Model	WER	CER
None (our raw hypothesis)	35	10
Lexicon	24	8
Bigram LM	14	6

Current Work

- Decoding improvements (working on two approaches to get efficient decoding with higher-order n-grams)
- Running experiments on Switchboard (a much better modern ASR benchmark corpus)
- Many things to try! We may very well break free from 25 years of speech recognition dogma

Conclusions

- DNNs are great function approximators with acoustic inputs
- There exists a fairly simple training scheme for DNNs as used in standard HMM-DNN systems
- Getting WER gains from HMM-DNN systems from DNN improvements alone is difficult
- A direct acoustic to word mapping is possible with CTC
- Remains to be seen whether state-of-the-art is possible with CTC training, but looks promising

Questions?

- More CTC details:

Deep Learning Reading Group

Tomorrow @ 12pm in Gates 120