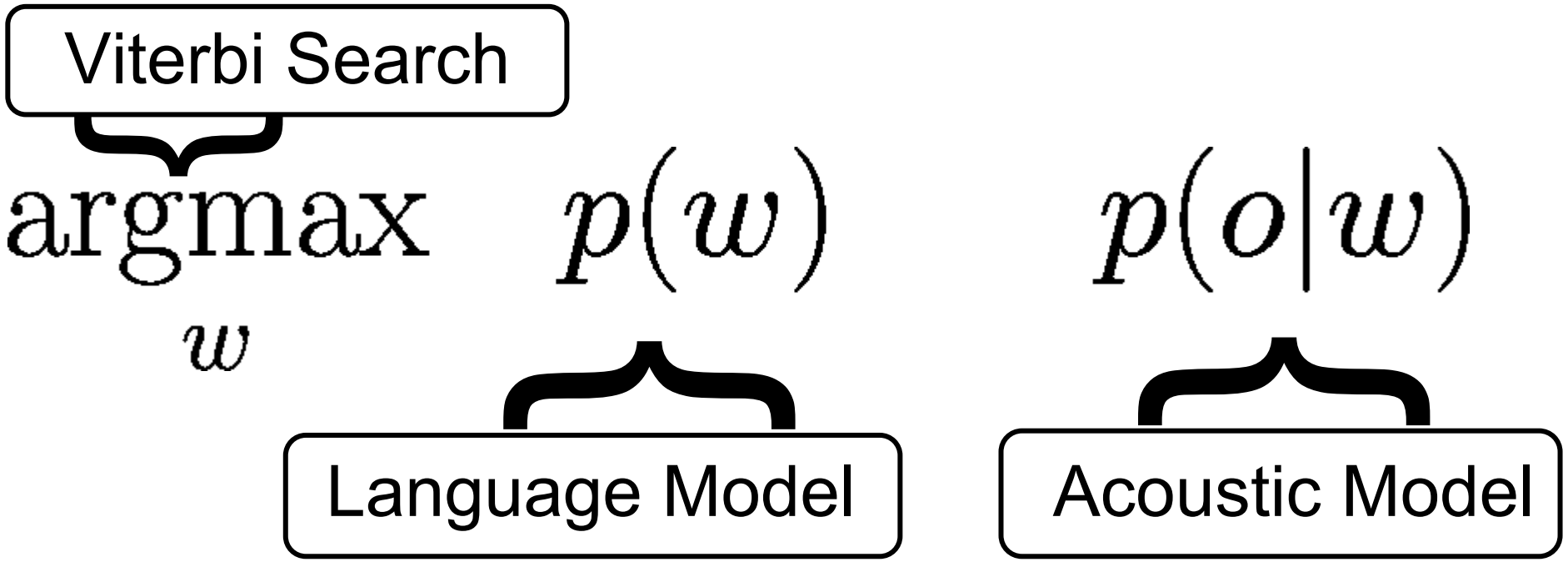
Deep Neural Networks for Acoustic Modeling in Speech Recognition

Hinton et al, Oct 2012

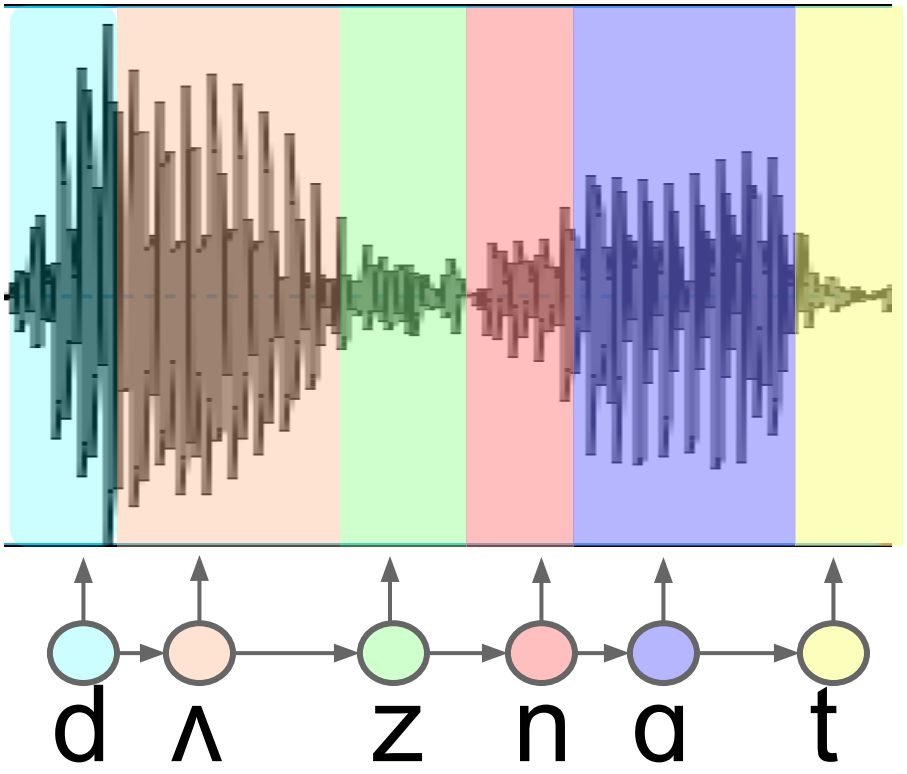
Outline

1. The Speech Problem
2. GMMs (- 2011)
3. Neural Nets (2011 - Today)
4. Tweaks
5. Why are DNNs better?
6. 2012-2015 Developments

Speech Recognition Problem



Speech Recognition Problem - HMM

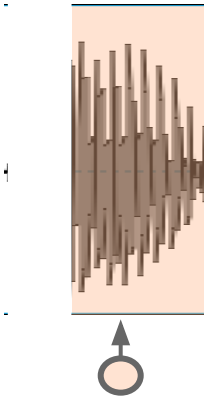


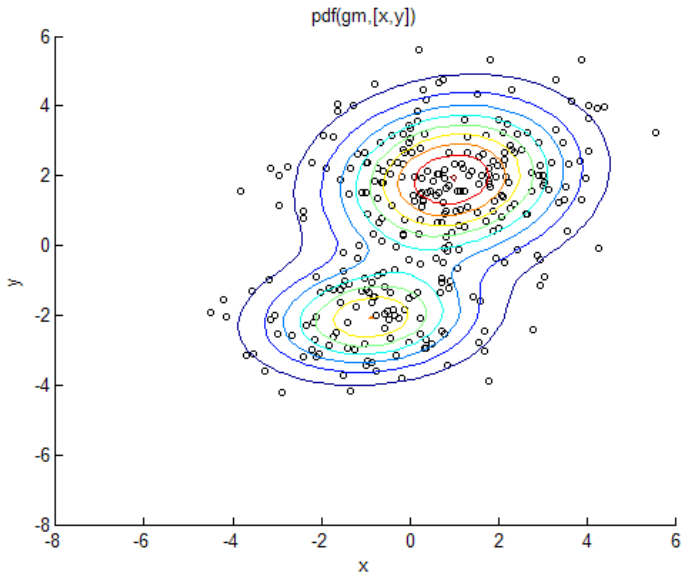
"does not"

$$p(o_i|s_i)p(s_i|s_{i-1})$$

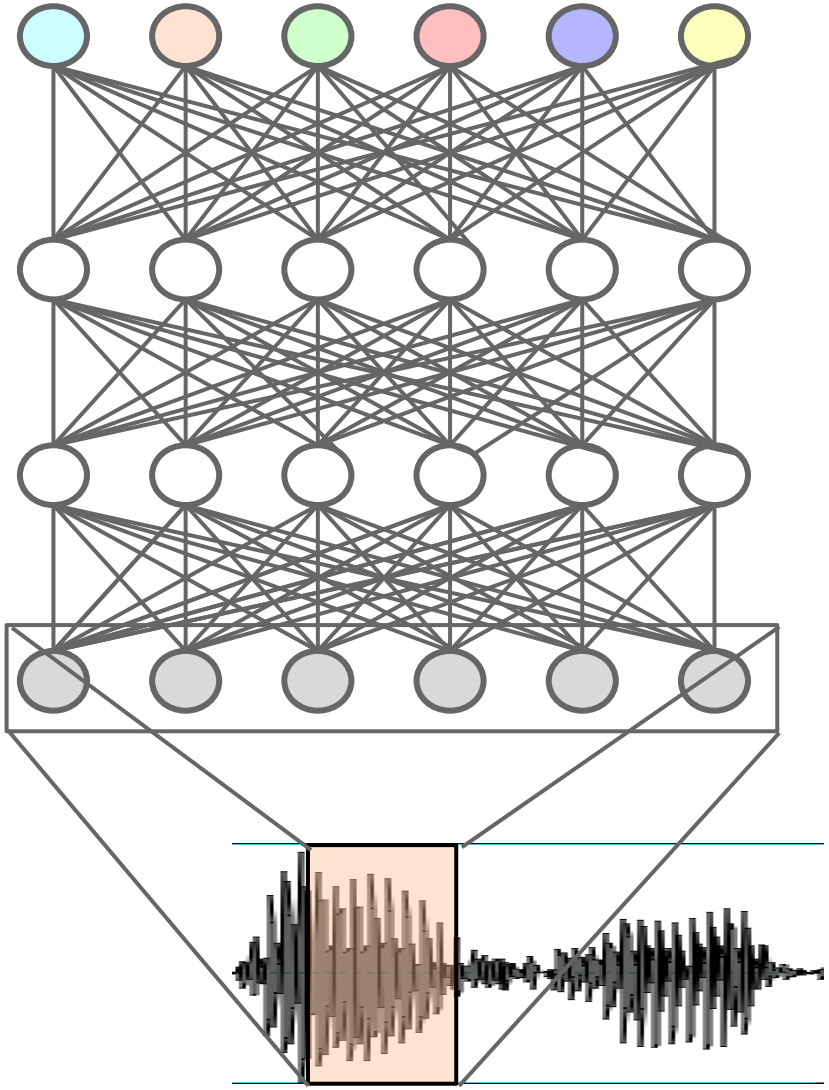
Acoustic Model - 1980s to 2011 - GMMs

- Easy to fit by plugging in to baum-welch (EM)
- Fast to train and decode
- Can fine-tune using discriminative post-training (MMI)





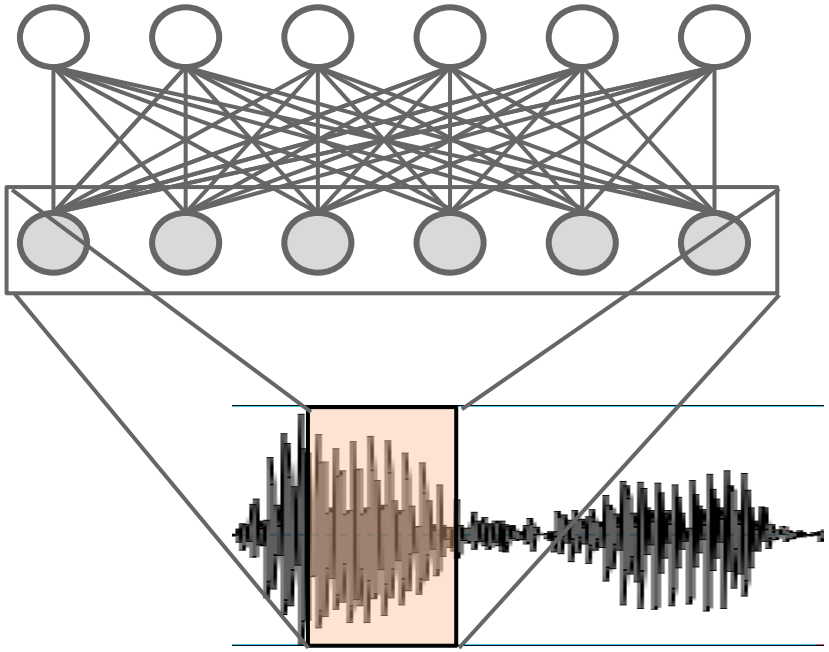
Acoustic Model - 2011-Today - Neural Nets



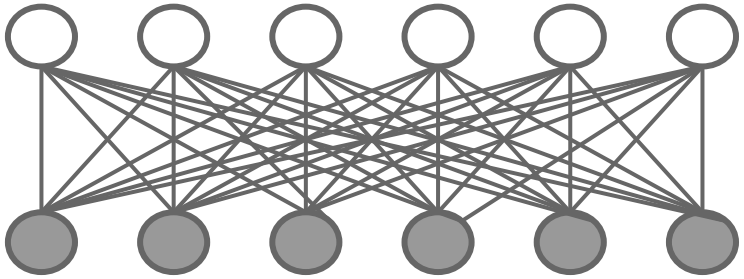
Neural Net - Pretraining V-H Units

Gaussian-Bernoulli RBM:

$$E(\mathbf{v},\mathbf{h}) = \sum_{i \in \text{vis}} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j \in \text{hid}} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij},$$



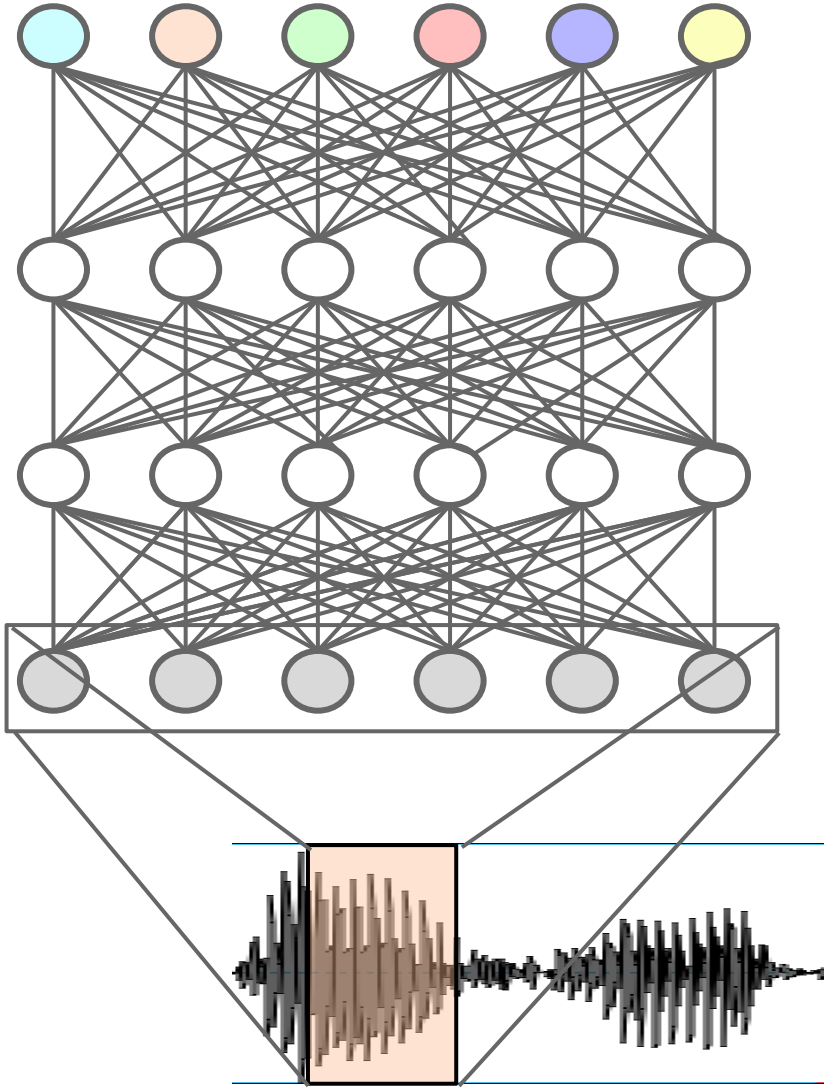
Neural Net - Pretraining H-H Units



RBM:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i \in \text{visible}} a_i v_i - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i,j} v_i h_j w_{ij},$$

Neural Net - Fine Tuning (Backprop)



Tweaks - Maximum Mutual Information Training

$$p(l_{1:T}|v_{1:T}) = p(l_{1:T}|h_{1:T})$$

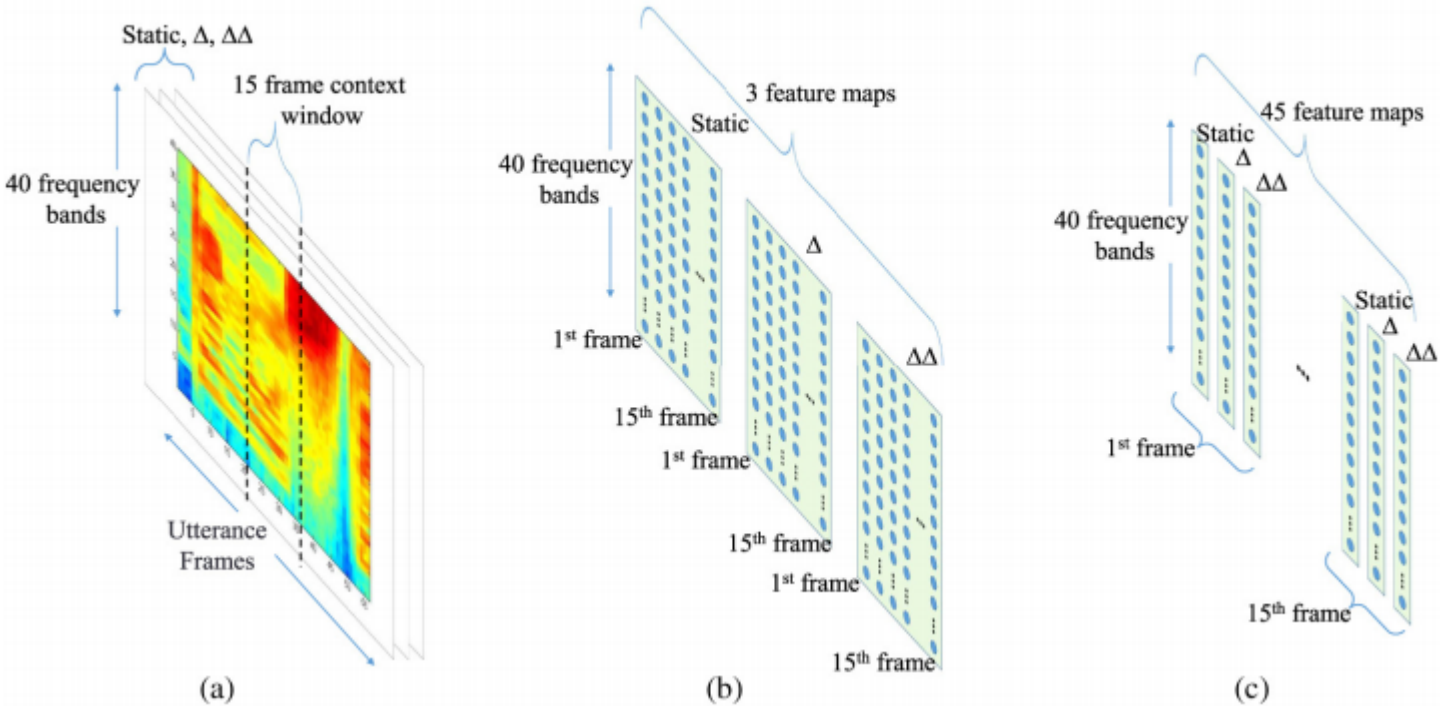
$$= \frac{\exp\left(\sum_{t=1}^{T} \gamma_{ij}\phi_{ij}(l_{t-1}, l_t) + \sum_{t=1}^{T}\sum_{d=1}^{D} \lambda_{l_t,d} h_{td}\right)}{Z(h_{1:T})}$$

transition probabilities (HMM)

agreement between activations + hidden units

more closely related to objective (sequence labeling)
~5% relative gain in accuracy

Tweaks - Convolutional Nets



~5% relative gain in accuracy

Source: “Convolutional Neural Networks for Speech Recognition” O. Abdel-Hamid et al, IEEE Transactions on Audio, Speech, and Language Processing, Oct 2014

Why are DNNs > GMMs?

Hinton's story:

- RBM is a "Product of experts" model, whereas GMM is a "Mixture of experts" model
  - *"Each param of a product model is constrained by a large fraction of the data"*
- DNNs can model simultaneous events, GMMs assume 1 mixture component generates observation
- DNNs benefit more from context frames

Experiments

[TABLE 4] WER IN % ON ENGLISH BROADCAST NEWS.

	50 H		430 H	
LVCSR STAGE	GMM-HMM BASELINE	AE-BN	GMM/HMM BASELINE	AE-BN
FSA	24.8	20.6	20.2	17.6
+fBMMI	20.7	19.0	17.7	16.6
+BMMI	19.6	18.1	16.5	15.8
+MLLR	18.8	**17.5**	16.0	**15.5**
MODEL COMBINATION	16.4		15.0	

Developments Since 2012

Better hardware means bigger training sets

- 2011 - several hundred hours of training data
- 2015 - 100,000 hours of training data (Baidu DeepSpeech, 10k + 90k synthesized)

“Currently, the biggest disadvantage of DNNs compared with GMMs is that it is much harder to make good use of large cluster machines to train them on massive data sets.”

> No longer true! (GTC 2015)

Developments Since 2012

Learn the features - raw filterbank/FFT energies outperform hand-engineered MFCCs/PLPs (esp. in noisy environments*)

Lots of hacks and tweaks:

- Dropout/ReLU/etc - no need for pretraining
- Recurrent nets
- Data augmentation - primarily the addition of noise

* “Deep Speech: Scaling up end-to-end speech recognition” - A. Hannum et al, 2015