PHONE RECOGNITION USING RESTRICTED BOLTZMANN MACHINES

Abdel-rahman Mohamed and Geoffrey Hinton

Department of Computer Science, University of Toronto

ABSTRACT

For decades, Hidden Markov Models (HMMs) have been the state-of-the-art technique for acoustic modeling despite their unrealistic independence assumptions and the very limited representational capacity of their hidden states. Conditional Restricted Boltzmann Machines (CRBMs) have recently proved to be very effective for modeling motion capture sequences and this paper investigates the application of this more powerful type of generative model to acoustic modeling. On the standard TIMIT corpus, one type of CRBM outperforms HMMs and is comparable with the best other methods, achieving a phone error rate (PER) of 26.7% on the TIMIT core test set.

Index Terms— phone recognition, restricted Boltzmann machines, distributed representations.

1. INTRODUCTION

A state-of-the-art Automatic Speech Recognition (ASR) system typically uses Hidden Markov Models (HMMs) to model the sequential structure of speech signals, with local spectral variability modeled using mixtures of Gaussian densities. Many methods have been proposed for relaxing the very strong conditional Independence assumptions of standard HMMs (e.g. [1],[2]).

In this work, we propose using variants of Restricted Boltzmann Machines (RBMs)[3] to model the spectral variability in each phone. Unlike HMMs, RBMs use a distributed hidden state that allows many different features to cooperatively determine each output frame, and the observations interact with the hidden features using an undirected model. An RBM is a bipartite graph in which visible units that represent observations are connected to hidden units using undirected weighted connections. The hidden units learn non-linear features that allow the RBM to model the statistical structure in the vectors of visible states.

RBMs have been used successfully for hand-written character recognition [3, 4], object recognition [5], collaborative filtering [6] and document retrieval. By conditioning on previous observations, RBMs can be used to model high-dimensional, sequential data and they have proved to be very successful for modeling motion capture data [7].

Several different RBM architectures are described in section 2 and ways of training them are described in section 4. Section 3 describes how to perform phone recognition using a trained RBM. Sections 5 and 6 compare the performance of different RBM architectures and training methods. The performance of the best type of RBM is also compared to other state-of-the-art acoustic modeling techniques.

2. RESTRICTED BOLTZMANN MACHINES

An RBM is a particular type of Markov Random Field (MRF) that has one layer of stochastic visible units and one layer of stochastic hidden units. There are no visible-visible or hidden-hidden connections but all visible units typically have connections to all hidden units [figure 1-(a)]. The weights on the connections and the biases of the individual units define a probability distribution over the state vectors, $\mathbf{v}$ of the visible units via an energy function. We consider RBM's with Bernoulli hidden units and Gaussian visible units that have a fixed variance of 1. The energy of the joint configuration $(\mathbf{v}, \mathbf{h})$ is given by [8]:

$$E(\mathbf{v}, \mathbf{h}; \theta) = \sum_{i=1}^{\mathcal{V}} \frac{(v_i - b_i)^2}{2} - \sum_{i=1}^{\mathcal{V}} \sum_{j=1}^{\mathcal{H}} w_{ij} v_i h_j - \sum_{j=1}^{\mathcal{H}} a_j h_j \tag{1}$$

where $\theta = (\mathbf{w}, \mathbf{b}, \mathbf{a})$ and w_{ij} represents the symmetric interaction term between visible unit i and hidden unit j while b_i and a_j are their bias terms. $\mathcal{V}$ and $\mathcal{H}$ are the numbers of visible and hidden units. The probability that the model assigns to a visible vector $\mathbf{v}$ is:

$$p(\mathbf{v}; \theta) = \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h}; \theta)}}{\sum_{\mathbf{u}} \sum_{\mathbf{h}} e^{-E(\mathbf{u}, \mathbf{h}; \theta)}} \tag{2}$$

Since there are no hidden-hidden or visible-visible connections, the conditional distributions $p(\mathbf{v}|\mathbf{h})$ and $p(\mathbf{h}|\mathbf{v})$ are factorial and are given by

$$\begin{aligned} p(h_j = 1|\mathbf{v}; \theta) &= \sigma(\sum_{i=1}^{\mathcal{V}} w_{ij} v_i + a_j) \\ p(v_i = 1|\mathbf{h}; \theta) &= \mathcal{N}(\sum_{j=1}^{\mathcal{H}} w_{ij} h_j + b_i, 1) \end{aligned} \tag{3}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ and $\mathcal{N}(\mu, V)$ is a Gaussian.

2.1. The conditional RBM

The Conditional RBM (CRBM)[7] is a variant of the standard RBM that models vectors of sequential data by considering the visible variables in previous time steps as additional, conditioning inputs. Two types of directed connections are added; autoregressive connections from the past n frames of the visible vector to the current visible vector, and connections from the past n frames of the visible vector to the hidden units as in figure 1-(b). Given the data vectors at times $t, t-1, ..., t-n$ the hidden units at time t are conditionally independent. One drawback of the CRBM is that it ignores future frames when inferring the hidden states, so it does not do backward smoothing. Performing backward smoothing correctly in a CRBM would be intractable because, unlike an HMM, there are exponentially many possible hidden state vectors, so it is not possible to work with the full distribution over hidden vectors when the hidden units are not independent.

If we are willing to give up on the ability to generate data sequentially from the model, the CRBM can be modified to have both autoregressive and visible-hidden connections from a limited set of future frames as well as from a limited past. So we get the interpolating CRBM (ICRBM) [figure 1-(c)]. The directed, autoregressive connections from temporally adjacent frames ensure that the ICRBM does not waste the representational capacity of the non-linear hidden units by modeling aspects of the central frame that can be predicted linearly from the adjacent frames.

3. USING RBM'S FOR PHONE RECOGNITION

A context window of successive frames of feature vectors is used to set the states of the visible units of the RBM. To train the RBM to model the joint distribution of a set of frames and the $\mathcal{L}$ possible phone labels of the last or central frame, we add an extra "softmax" visible unit that has $\mathcal{L}$ states, one of which has value 1. The energy function becomes:

$$E(\mathbf{v}, \mathbf{l}, \mathbf{h}; \theta) = -\sum_{i=1}^{\mathcal{V}}\sum_{j=1}^{\mathcal{H}} w_{ij} h_j v_i - \sum_{k=1}^{\mathcal{L}}\sum_{j=1}^{\mathcal{H}} w_{kj} h_j l_k - \sum_{j=1}^{\mathcal{H}} a_j h_j - \sum_{k=1}^{\mathcal{L}} c_k l_k + \sum_{i=1}^{\mathcal{V}} \frac{(v_i - b_i)^2}{2} \quad (4)$$

$$p(l_k = 1|\mathbf{h}; \theta) = \text{softmax}(\sum_{j=1}^{\mathcal{H}} w_{kj} h_j + c_k) \quad (5)$$

And $p(\mathbf{l}|\mathbf{v})$ can be computed exactly using

$$p(\mathbf{l}|\mathbf{v}) = \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{l},\mathbf{h})}}{\sum_{\mathbf{l}}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{l},\mathbf{h})}} \quad (6)$$

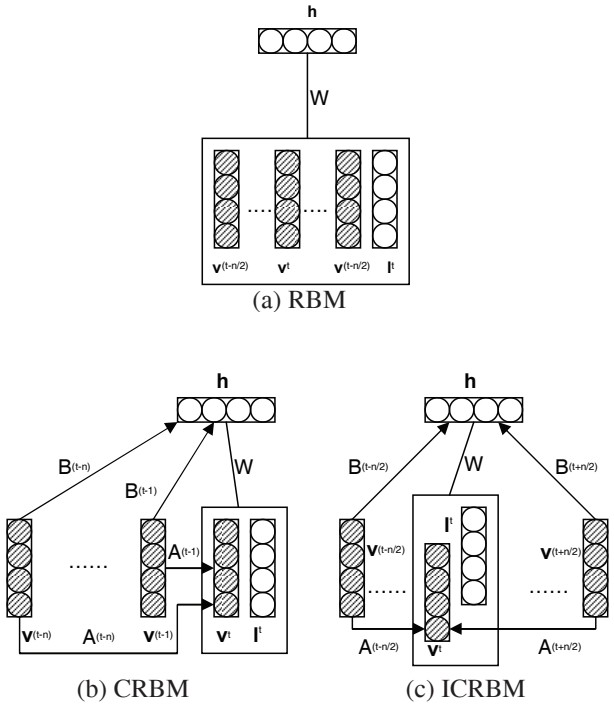

Fig. 1. Different RBM architectures. (a) shows an RBM that models the joint density of the label and all the frames in the window. (b) and (c) show two types of conditional RBM that model the joint density of the label and a single frame conditional on the other frames in the window.

The value of $p(\mathbf{l}|\mathbf{v})$ can be computed efficiently by utilizing the fact that the hidden units are conditionally independent. This allows the hidden units to be marginalized out in a time that is linear in the number of hidden units. To generate phone sequences, the values of $log\ p(\mathbf{l}|\mathbf{v})$ per frame are fed to a Viterbi decoder.

4. RBM TRAINING

Following the gradient of the joint likelihood function of data and labels, the update rule for the visible-hidden weights is

$$\Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \quad (7)$$

The expectation $\langle v_i h_j \rangle_{data}$ defines the frequency with which the visible unit v_i and the hidden unit h_j are on together and $\langle v_i h_j \rangle_{model}$ is the expectation with respect to the distribution defined by the model. the term $\langle . \rangle_{model}$ takes exponential time to compute exactly so the Contrastive Divergence (CD) approximation to the gradient is used instead [3]. The new update rule becomes:

$$\Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_1 \quad (8)$$

where $\langle . \rangle_1$ represents the expectation with respect to the distribution of samples from running a Gibbs sampler ini-

tialized at the data for one full step. The update rule for the CRBM visible-hidden undirected weights is the same as above but for the weights of the directed connections it is different. For the autoregressive visible-visible links it is

$$\Delta A_{ij}^{(t-q)} = v_i^{(t-q)}(\langle v_j^t \rangle_{data} - \langle v_j^t \rangle_1) \quad (9)$$

where $A_{ij}^{(t-q)}$ is the weight from unit i at time $(t-q)$ to unit j. For the visible-hidden directed links it is

$$\Delta B_{ij}^{(t-q)} = v_i^{(t-q)}(\langle h_j^t \rangle_{data} - \langle h_j^t \rangle_1) \quad (10)$$

4.1. Discriminative and hybrid training of an RBM

Since the log conditional probability, $log\ p(\mathbf{l}|\mathbf{v})$, can be computed exactly, the gradient can also be computed exactly. If the correct label is m, the update rule for the visible-hidden weights is

$$\Delta w_{ij} = v_i \sigma \left(a_j + w_{jm} + \sum_{i=1}^{\mathcal{V}} w_{ij} v_i \right) - v_i \sum_{k=1}^{\mathcal{L}} p(l_k = 1|\mathbf{v}) \sigma \left(a_j + w_{jk} + \sum_{i=1}^{\mathcal{V}} w_{ij} v_i \right) \quad (11)$$

To avoid model overfitting, we follow the gradient of a hybrid function $f(\mathbf{v}, \mathbf{l})$ which contains both generative and discriminative components.

$$f(\mathbf{v}, \mathbf{l}) = \alpha\ log\ p(\mathbf{l}|\mathbf{v}) + log\ p(\mathbf{v}|\mathbf{l}) \quad (12)$$

where $log\ p(\mathbf{v}|\mathbf{l})$ works as a regularizer and is learned by using the original labels with the reconstructed data to infer the states of the hidden units at the end of the sampling step. The α parameter is used to control the emphasis given to the discriminative component in the objective function. Since the original labels are used during hidden layer reconstruction for evaluating the gradient of $log\ p(\mathbf{v}|\mathbf{l})$, the label biases are updated using the gradient of $log\ p(\mathbf{l}|\mathbf{v})$ only.

5. EVALUATION SETUP

All phone recognition experiments were performed on the core test set of the TIMIT corpus [1]. All SA records were removed as they could bias the results. A development set of 50 speakers was used for model tuning. The speech was analyzed using a 25-ms Hamming window with a 10-ms fixed frame advance. In all the experiments, we represented the speech using 12th-order Mel frequency cepstral coefficients (MFCCs) and energy, along with their first and second temporal derivatives. The data were normalized to have zero mean and unit variance. We used 183 target class labels (i.e., 3 states for each one of the 61 phones). Forced alignment was used to produce frame labels using a baseline HMM system. After decoding, starting and ending silences were removed and the 61 phone classes were mapped to a set of 39 classes as in [9] for scoring. All of our experiments used a bigram language model over phones, estimated from the training set.

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

6. EVALUATIONS

6.1. CD training of three types of RBM

The three types of RBM shown in figure 1 were trained using a window of 11 frames as the visible states. 2000 hidden units were used for all architectures. Table 1 shows the phone error rate (PER) when the RBMs were trained generatively.

Table 1. *The PER of different RBM architectures.*

RBM	CRBM	ICRBM
36.9%	42.7%	39.3%

The ICRBM produces a lower PER than the CRBM, presumably because the near future is more relevant than the more distant past. The unconditional RBM performs the best, probably because modeling the joint density of the entire window reduces overfitting more effectively than only modeling one frame conditional on the other frames. In the conditional RBMs, the relation between frames captured by the autoregressive connections influences what the hidden units learn, but it does not directly help in decoding as $p(\mathbf{l}|\mathbf{v})$ does not depend on the autoregressive connections.

6.2. Hybrid training of the three types of RBM

Generatively trained network parameters were used to initialize hybrid training. The mean square error (MSE) between the original and predicted targets, was measured every epoch on the development set. If the MSE reduction between two epochs was less than 0.5%, the learning rate was halved. Training stopped when no improvement was observed. The parameter α in equation (12) was tuned to maximize the PER on the development set. The best RBM model achieved 27.5% PER while the best ICRBM model achieved 26.7%. The discriminative component of the hybrid gradient forces the ICRBM to extract non-linear features from the context that are more useful for predicting the label. It has more capacity for these features than the unconditional RBM because it does not have to model the contextual frames or the linear dependencies of the modeled frame on the context.

6.3. Comparison with other models

Since a feedforward neural network is quite similar, it was compared to the ICRBM model. A feedforward neural network with 2000 hidden units and an input window of

11 frames was trained twice using backpropagation; once from random weights and once from the generatively trained weights of the unconditional RBM. The learning rate was reduced in the same way as in RBM. Table 2 shows that the ICRBM outperformed both feedforward models, probably because the generative component of the hybrid training greatly reduces overfitting.

Table 2. *PER of the ICRBM compared to the NN model.*

NN (random weights)	NN (RBM weights)	ICRBM
28.7%	28.3%	26.7%

A two-tailed Matched Pairs Sentence-Segment Word Error (MAPSSWE) significance test [10] was conducted with the null hypothesis that there is no performance difference between the ICRBM and the feedforward neural net models using the NIST sc_stats tool. The test finds a significant difference at the level of p=0.05. Table 3 compares the results achieved by the ICRBM model to other proposed models.

Table 3. *Reported results on TIMIT core test set*

Method	PER
Conditional Random Field [11]	34.8%
Large-Margin GMM [12]	28.2%
CD-HMM [2]	27.3%
ICRBM (this paper)	**26.7%**
Augmented conditional Random Fields [2]	26.6%
Recurrent Neural Nets [13]	26.1%
Monophone HTMs [1]	24.8%
Heterogeneous Classifiers [14]	24.4%

7. CONCLUSIONS

In this work, several variants of Restricted Boltzmann Machines were investigated for acoustic modeling. They all used Gaussian visible units to represent MFCC coefficients, a softmax visible unit to represent labels, and Bernoulli hidden units. The hidden features learned by each RBM define a joint probability distribution over MFCC coefficients and discrete labels. Three architectures were evaluated: the unconditional RBM, the conditional CRBM, and the interpolating conditional ICRBM. Generative and discriminative update rules were investigated and a hybrid that blends the two gradients worked best. Using this hybrid, the ICRBM achieved significantly better results than a feedforward neural network model with the same architecture.

8. REFERENCES

[1] L. Deng and D. Yu, "Use of differential cepstra as acoustic features in hidden trajectory modelling for phonetic recognition," in *Proc. ICASSP*, 2007, pp. 445–448.

[2] Hifny Y. and Renals S., "Speech recognition using augmented conditional random fields," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 2, pp. 354–365, 2009.

[3] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, pp. 1771–1800, 2002.

[4] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[5] V. Nair and G. E. Hinton, "3-d object recognition with deep belief nets," in *Proc. NIPS*, 2009.

[6] Salakhutdinov R. R, A. Mnih, and G. E. Hinton, "Restricted Boltzmann machines for collaborative filtering," in *Proc. ICML*, 2007, pp. 791–798.

[7] G. W. Taylor, G. E. Hinton, and S. Roweis, "Modeling human motion using binary latent variables," in *Proc. NIPS*, 2007.

[8] M. Welling, M. Rosen-Zvi, and G. E. Hinton, "Exponential family harmoniums with an application to information retrieval," in *Proc. NIPS*, 2005.

[9] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 37, no. 11, pp. 16411648, 1989.

[10] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, pp. 532–535.

[11] J. Moris and E. Fosler-Lussier, "Combining phonetic attributes using conditional random fields," in *Proc. Interspeech*, 2006, pp. 597–600.

[12] F. Sha and L. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden markov models," in *Proc. ICASSP*, 2007, pp. 313–316.

[13] A. Robinson, "An application to recurrent nets to phone probability estimation," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 298–305, 1994.

[14] A. Halberstadt and J. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," in *Proc. ICSLP*, 1998.