Deep Architectures for Human Computer Interaction

Athanasios K. Noulas
Intelligent Systems Laboratory Amsterdam
University of Amsterdam
1095 TM, Amsterdam
The Netherlands
anoulas@science.uva.nl

Ben J.A. Kröse
Intelligent Systems Laboratory Amsterdam
University of Amsterdam
1095 TM, Amsterdam
The Netherlands
krose@science.uva.nl

ABSTRACT

In this work we present the application of Conditional Restricted Boltzmann Machines in Human Computer Interaction. These provide a well suited framework to model the complex temporal patterns produced from humans in the audio and video modalities. They can be trained in a semi-supervised fashion and produce high accuracy, on-line inference results, while dealing with multiple concept groups. We present the results acquired in the task of speaker detection.

Categories and Subject Descriptors

H.1.2 [**Information Systems Applications**]: Models and Principles,User/machine Systems

Keywords

conditional Restricted Boltzmann Machines, deep architectures, audio-visual fusion

1. INTRODUCTION

If we want to create social robots and agents we have to equip them with the ability to analyse multi-modal information streams. More specifically, if we want to achieve natural Human Computer Interaction (HCI) we need machines that can analyse the audio-visual streams generated from humans and extract high order concepts, such as emotions or facial expressions for example. These concepts exhibit high variability and map to very noisy features, and the most promising results come from approaches that try to model this mapping using probabilistic methods.

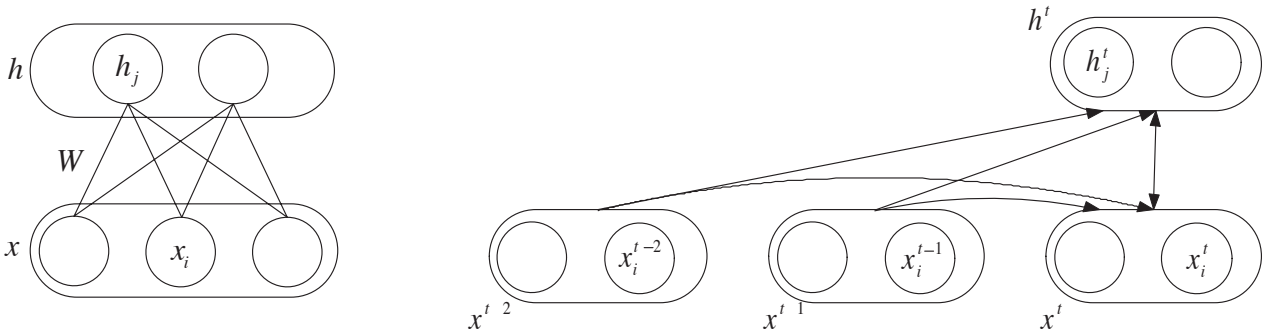
Probabilistic modelling of this kind of information streams is a formidable task. First, the audio and video modalities produce complex temporal patterns whose dynamics cannot be analytically estimated, but rather should be learnt from the data. Second, complex models that can capture the underlying process are often unusable in practice, since on-line inference is computationally extremely expensive. Finally, we need large quantities of training data to learn the parameters of such a complicated pattern, but the models tend to generalise very badly in tasks involving the same data but different objectives.

Different approaches have been applied in audiovisual information streams in HCI. Some researchers try to capture arising patterns using a complex generative model, in order to represent the complex physical mechanisms that generated the data. Others prefer a discriminative approach in order to boost their classification accuracy, on a specific domain. In this work we adapt the model introduced in the work of Sutskever et. al, [7], called the Conditional Restricted Boltzmann Machine (cRBM), which is a much more suitable model for the task at hand.

A cRBM is an extension of the Restricted Boltzmann Machine (RBM). In short, the RBM is an energy based model, which can model complex relationships between observations and higher order concepts. Multiple RBMs can be stack on top of each other and learnt one at a time. The final *deep architecture*, will convey all the available information through different layers but capture a different view of this information in each one of the layers [1]. We can now tune the parameters of the whole architecture to optimise the task at hand. When the temporal dimension of the data is taken into consideration, by *conditioning* inference at each point on previous observations, we move to a cRBM.

There exists an efficient learning algorithm for all RBM-based models, called Contrastive Divergence [4], which is a local approximation of the gradient of the data likelihood function and performs very well in practice. Furthermore, the first part of the learning procedure, where the structure of the observed data is extracted, is completely unsupervised. The model can generate data following the observations distribution after this point. We can then extend it to semi-supervised learning using some training data to tune the parameters towards the desired objective (or objectives). In our case we model the audio-visual streams generated by interacting persons. Successful HCI should allow similar interaction between humans and computers. We demonstrate the ability of the system to determine whether a person is speaking or not, but we can extend our objective to emotion detection or gesture recognition by tweaking the output of the unsupervised phased using properly labelled data examples.

A further advantage of such models is the fact that exact



The Restricted Boltzmann Machine A conditional RBM

Figure 1: On the left we can see the structure of the Restricted Boltzmann Machine. There are undirected connections from all the nodes of hidden layer h to all the nodes of visible layer x. On the right side we see a conditional RBM where the hidden and visible layers of time slice t receive incoming connections from the visible layers of the previous time slices.

on-line inference is fast, because all the variables we want to infer are conditionally independent given the observations. During inference the behaviour of a deep architecture is very similar to that of a neural network. Thus, real time performance of an agent utilising a deep architecture is feasible.

In the next section we present the proposed cRBM framework. In section 3 we discuss the application of such a model on multi-modal data coming from the face region of a person. Section 4 contains the details of our experimental setup and results compared to those achieved by the methods proposed in [6] and [2]. We conclude the paper with a section discussing future directions in the application of cRBMs on multi-modal information streams.

2. BUILDING DEEP ARCHITECTURES

In this section we briefly review the process of building a deep architecture for temporal data modelling. Firstly, we describe the probabilistic interpretation of the RBM as well as the way CD is applied for learning its parameters. Secondly, we present how we extend from the RBM to the cRBM. Thirdly, we describe how we can build deep architectures using a cRBM as building block. We conclude this section describing how we can use deep architectures to perform task specific inference.

2.1 Restricted Boltzmann Machine

A RBM is an energy-based model, which means that the probability distribution over the variables of interest is defined through an energy function. The undirected graphical model of the RBM is composed from a layer of observable variables $x = \{x(i)\}$ and a layer of hidden variables $h = \{h(j)\}$. The graphical model of the RBM is visible in figure 1. There are connections between different layers but no connections within a layer. The log likelihood of a probability of a given configuration is defined by:

$$-\log P(x,h) = -a^\top x - b^\top h - h^\top W x + const \qquad (1)$$

where the vectors a and b contain the biases of the corresponding nodes, and matrix W the weights connecting hidden to visible nodes. Note that we didn't write down the explicit form of the normalization term in order to express our inability to compute it in general.

The main advantage of such an architecture is that the posterior over the variables of a layer factorizes completely. Thus,

$$p(x|h) = \prod_i p(x_i|h) \quad and \quad p(h|x) = \prod_j p(h_j|h) \qquad (2)$$

Typically, RBMs use binary variables for both the hidden and visible nodes. In our case we have real valued data and therefore we use continuous real valued nodes with Gaussian noise [10] for the visible layer. The conditional distributions are then given by:

$$\begin{aligned} p(x_i|h) &= \mathcal{N}\left(c_i + \textstyle\sum_j h_k w_{ij}, 1\right) \\ p(h_j|x) &= \sigma\left(b_j + \textstyle\sum_i x_i w_{ij}\right) \end{aligned} \qquad (3)$$

with $\sigma(\cdot)$ being the logistic function.

The parameters W, b and h of the Energy function can be obtained by maximum likelihood learning but that is very slow. Instead, we can apply CD learning and , use the following update rules:

$$\begin{aligned} \Delta w_{ij} &\propto \langle x_i h_j \rangle_{data} - \langle x_i h_j \rangle_{recon} \\ \Delta a_i &\propto \langle x_i \rangle_{data} - \langle x_i \rangle_{recon} \\ \Delta b_j &\propto \langle h_j \rangle_{data} - \langle h_j \rangle_{recon} \end{aligned} \qquad (4)$$

We refer the interested reader to [4] regarding the origin of these rules and a thorough comparison of CD with ML learning. Intuitively, CD can be seen as an indirect way to make the model distribution identical to the data distribution. We obtain the model parameters that would distort the data reconstruction as little as possible. If we push this error to zero, in an infinite chain sampling operation, the model would produce the exact same data with our training distribution. Thus the model of the data and the learnt model would be identical.

2.2 Conditional Restricted Boltzmann Machine

The RBM could model in our case static frames of data. We can model temporal dependencies by treating the observa-

Figure 2: Some example frames of our data. The two left columns come from the labelled streams, while the rest come from the test streams (left to right, speakers 1, 2, 3). The top row contains a typical frame of the stream while the bottom one some extreme cases.

tions of past time slices as additional inputs [8]. Thus we now model: $p\left(x^t, h^t | x^{t-1} \dots x^{t-n}\right)$, creating a cRBM which is visible in figure 1. We can add connections from the previous time slice observations to the current one in the form of a dynamic bias and to the hidden layer in the form of undirected incoming connections without complicating inference. In our problem we set $n = 4$, taking into account the previous four observation vectors. Let d_{ij}^{t-q} denote the weight connecting node i of observation x^{t-q} to node j of hidden layer h^t, and a_{ki}^{t-q} substitute a_i with weights connecting node k of observation x^{t-q} to node i of observation x^t. The CD learning rules for these weights now become

$$\begin{aligned} \Delta d_{ij}^{t-q} &\propto x_i^{t-q} \langle h_j^t \rangle_{data} - \langle h_j^t \rangle_{recon} \\ \Delta a_k^{t-q} i &\propto x_k^{t-q} \langle x_i^t \rangle_{data} \langle x_i^t \rangle_{recon} \end{aligned} \tag{5}$$

2.3 Building a Deep Architecture

Once we have learnt the parameters of the cRBM, we can add layers as in a Deep Belief Network (DBN) [5]. The previously learnt cRBM is used to get a sequence of hidden layer vectors from the training data. This sequence can now be used as visible data for a new hidden layer. We expect higher levels of the model to capture higher concepts appearing in the data. This greedy learning can be justified using a variatonal bound [5]. At this point, the final structure can be seen as an autoregressor with Gaussian noise, which takes into account the specific temporal relations of the data.

2.4 Inference using the Deep Architecture

There is a different number of ways to infer the quantity in question from the deep architecture. For example, we can train one deep architecture per class, and classify a novel example as the class that creates a more accurate reconstruction. Alternatively we can use any classification method on the informative, low dimensional hidden layer(s). Different methods are discussed thoroughly in [4].

In this work we tune the final architecture as a sigmoid belief net. In this way, we expect to alter slightly only the weights that will perform optimal classification in the subspace learnt during CD training [1].

3. FEATURE EXTRACTION

The audio and video streams generated by a person are complex highly varying signals, making it very hard to detect correlations in the raw pixel and audio data. There are many different descriptors for such data, each one capturing the most informative part of the stream for a given task. Since we want to model the correlation between a person's face movement and the corresponding audio stream, we use the features that produced the state-of-the-art result in lip-movement generation [3].

From the video stream we create an active appearance model of the person [9] , and then detect the points of interest on each frame of the stream. Some example frames are visible in figure 2. There are 52 points tracked on each face throughout the stream. Some of these points correspond to less informative parts of the face in terms of speaking activity. However we expect the cRBM to automatically learn the informative subspace of the high-dimensional input data, and utilise only information *relative* to the inference objective.

In the audio stream we extract the 13 first MFCCs. In speaker diarization or recognition tasks, it is a common technique to concatenate the first and second order differences of these vectors. However, in the cRBM setting, consequent audio descriptors will be connected to common hidden variables. Thus, their weights will *compete* during the training procedure to explain the data, and therefore useful information regarding the descriptor differences can automatically be detected.

4. EXPERIMENTS

In this section we first describe the experimental setting and the objectives we wanted to meet. In the second subsection we present our experimental results along with the results acquired on the same data using the algorithms of [6] and [2].

4.1 Experimental Setup

We trained our cRBM using 3000 sequences from a discussion recording. Example frames of this data set are visible on the left side in Figure 2. We used the weights acquired during the cRBM training to initialise a sigmoid belief net

Accuracy %	Speaker 1	Speaker 2	Speaker 3
Proposed Approach	67	**70**	**76**
MI Based	45	57	43
D.B.N.	**88**	62	61

Table 1: The accuracy results in speaker diarization on different speakers of the meeting.

with the same structure, and performed speaker diarization on novel recording. The objective of our task, is to infer whether each person is speaking or not, on each frame of the novel recording.

The novel recording comes from a meeting setting with three participants. The first participant is dominant, speaking 70% of the time, while the two other participants vocalise significantly less. The second and third participant are moving their lips, or in different ways occludes their face, in order to make audio-visual synchronisation detection harder. We note here, that none of the participants in the training recording exhibits such extreme behaviour.

We evaluate our model in terms of the percentage of frames where the speaker status was correctly classified. In pattern recognition terms the total of true positive and true negative examples divided by the total number of classifications. Furthermore we compare our results to those acquired using two models proposed in different works. The first one, labelled **MI based** estimates the mutual information between the pixel value variations of a person and the corresponding audio script [2]. The speaker selected is the person most correlated to the stream. The second method, labelled as **D.B.N**, uses this information as well as information about the voice and appearance of the speakers. All this information is fused in a dynamic Bayesian network whose parameters are acquired directly from the test data [6].

4.2 Results & Discussion

In table 1 we can see the results acquired by our approach and the methods presented in [6] and [2]. Our model achieved an overall classification accuracy of 68%, which is close to the 76% achieved by the dynamic Bayesian network presented of [6].

Furthermore we observe that the acquired results are much more robust when it comes to speakers 2 and 3 which vocalise less. The *D.B.N.*, outperforms our proposed method when it comes to speaker 1, because he provides enough data to acquire a reliable model of his voice. Furthermore, our cRBM is outperforming both models when it comes to a speaker with behaviour very different from that of the training set.

This data set, although limited, can help drawn some important conclusions. Firstly, the cRBM manages to generalise well, using a limited training set to classify a novel test set in a very sparse domain. Secondly, adaptation to the new data set is really important. However, the first training phase of a cRBM is computationally intensive and novel data is not easily incorporated. Finally, a possible incorporation of RBM-like observational models to a dynamic probabilistic framework can account well for complex multimodal patterns, as in this example the audio and video streams generated by a person.

5. DISCUSSION OF CORE CONTRIBUTION

This paper, to our best knowledge is the first time that a cRBM is applied on audio visual data. Similar deep probabilistic architectures gain increased attention in modern machine learning research, and since their theoretical advantages are well fitted to multimodal fusion we expect them to be widely used in coming years.

The first advantage of these models regards their representational power. cRBMs belong to the family of product of experts. Therefore they are able to capture complex, high-varying probability distribution functions, and thus express the complex relationships appearing in multimodal streams. Furthermore, cRBMs have a distributed hidden state with a representational capacity which is linear to the number of components used. This compares favourably to the widely used HMMs that require 2^N hidden states to model N bits of information from the past.

The second advantage regards the two phases training procedure of cRBMs. In the first phase, the domain data is used without any labels, and cRBMs extract the structure of the dataset. In the second phase a few training examples can be used to tune the parameters for the required task. This can allow context specific training for the social agents without a time consuming procedure. Furthermore a large data set can be used in the first phase ,and achieve good high-concept classification results with just a few labelled examples of the required concept group.

The third advantage regards the application possibilities offered by the complete model. We can use it to perform accurate, on-line inference on a multiple range of concept groups, since the final model has neural-network like inference possibilities. Furthermore we can use the model to generate data, like in [8], and thus produce human like behaviour on the social agents side.

6. CONCLUSION

In this work, we describe the application of a cRBM on audio-visual data for HCI. We present the theoretical intuition behind choosing a cRBM to model this kind of data, and exhibit our experimental results on speaker detection. We show that a cRBM models naturally the complex patterns appearing in this kind of data without the need of any manual hard-coding of the process dynamics. Furthermore the ability to add more layers in the observation model can map sensor observations to high-level concepts without any user interference.

7. REFERENCES

[1] Yoshua Bengio. Learning deep architectures for ai. Technical Report 1312, Dept. IRO, Universite de Montreal, 2007.

[2] Trevor Darrell, John W. Fisher Iii, Paul Viola, and Mit Ai Lab. Audiovisual segmentation and the cocktail party effect. In *in International Conference on Multimodal Interfaces*, pages 32–40, 2000.

[3] Gwenn Englebienne, Tim Cootes, and Magnus Rattray. A probabilistic model for generating realistic lip movements from speech. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 401–408. MIT Press, Cambridge, MA, 2008.

[4] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, August 2002.

[5] Geoffrey E. Hinton, S. Onsidero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006.

[6] Athanasios K. Noulas and Ben J. A. Kröse. On-line multimodal speaker diarization. In *International Conference on Multimodal Interfaces*, pages 351–358, 2007.

[7] Ilya Sutskever and Geoffrey Hinton. Learning multilevel distributed representations for high-dimensional sequences. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

[8] Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. Modeling human motion using binary latent variables. In Bernhard Scho"lkopf, John Platt, and Thomas Hoffman, editors, *NIPS*, pages 1345–1352. MIT Press, 2006.

[9] T.F.Cootes, G.J. Edwards, and C.J.Taylor. Active appearance models. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[10] Max Welling, Michal Rosen-Zvi, and Geoffrey Hinton. Exponential family harmoniums with an application to information retrieval. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1481–1488. MIT Press, Cambridge, MA, 2005.