Dreaming Machines: On multimodal fusion and information retrieval using neural-symbolic cognitive agents

Leo de Penning[1], Artur d'Avila Garcez[2], and John-Jules C. Meyer[3]

1 **TNO Behaviour and Societal Sciences**
Soesterberg, The Netherlands
leo.depenning@tno.nl

2 **Department of Computing, City University**
London, UK
aag@soi.city.ac.uk

3 **Department of Information and Computing Sciences, Utrecht University**
Utrecht, The Netherlands
jj@cs.uu.nl

Abstract

Deep Boltzmann Machines (DBM) have been used as a computational cognitive model in various AI-related research and applications, notably in computational vision and multimodal fusion. Being regarded as a biological plausible model of the human brain, the DBM is also becoming a popular instrument to investigate various cortical processes in neuroscience. In this paper, we describe how a multimodal DBM is implemented as part of a Neural-Symbolic Cognitive Agent (NSCA) for real-time multimodal fusion and inference of streaming audio and video data. We describe how this agent can be used to simulate certain neurological mechanisms related to hallucinations and dreaming and how these mechanisms are beneficial to the integrity of the DBM. Finally, we will explain how the NSCA is used to extract multimodal information from the DBM and provide a compact and practical iconographic temporal logic formula for complex relations between visual and auditory patterns.

1998 ACM Subject Classification I.2.0 Cognitive simulation

Keywords and phrases Multimodal fusion, Deep Boltzmann Machine, Neural-Symbolic Cognitive Agent, Dreaming, Hallucinations

Digital Object Identifier 10.4230/OASIcs.ICCSW.2013.89

1 Introduction

The human brain has always inspired many of us to investigate and try to understand its complex processes. Ranging from neuroscientists that try to model the brain in terms of neurons, synapses and pathologies, to psychologists and cognitive scientists that try to model it in terms of human and social behaviour, to computer scientists that try to model it in terms of computational models that can perform intelligent tasks. A common tool in all these sciences is the use of abstract models of the human brain that help us to simulate, analyse and understand how it works. From a computer science perspective, computational models of the human brain are often based on models from neuroscience (e.g. neural networks) or models from cognitive and social sciences (e.g. cognitive models). These models have enabled computer scientists to build very complex systems that are able to perform tasks of human intelligence (e.g. visual recognition, speech recognition and driving a car). On the

2013 Imperial College Computing Student Workshop (ICCSW'13).
Editors: Andrew V. Jones, Nicholas Ng; pp. 89–94

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

other hand, these computational models have also been used in neural, cognitive and social sciences to investigate the human brain itself. For example, computational models have been used to investigate biological pathways in the visual cortex [5], neurological pathologies that cause hallucinations [10], the role of long-term memory in perception [8], and social dynamics of cognitive and affective processes [14]. Even in the investigation of more elusive and abstract processes related to the brain, computational models have been used. For example, to illustrate the role of mind and brain in cognitive psychology [12] and to explain the function of dreaming [1].

In this paper we will describe the use of a Deep Boltzmann Machine (DBM) as computational model to simulate certain neurological processes related to hallucination and dreaming and describe how these processes can be applied to multiple modalities, specifically streaming audio and video. Furthermore we will describe and illustrate how a Neural-Symbolic Cognitive Agent (NSCA) can be used to retrieve information from this model in a temporal logic formula that incorporates iconographic representations of the visual and auditory patterns.

2 Multimodal Deep Learning

Similar to the approach described in [9] we apply a Deep Boltzmann Machine (DBM) for multimodal fusion of visual and auditory information. A DBM can learn hierarchical representations of data, using several layers of Restricted Boltzmann Machines (RBMs) [11]. Each RBM represents a stochastic neural network with visible units $\mathbf{v}$, that represent input variables (or hidden-unit activations of lower-layer RBMs), and hidden units $\mathbf{h}$, that represent the likelihood of certain activation patterns in $\mathbf{v}$. There are symmetric weighted connections between the hidden and visible units with weights W, but no connections within the hidden units or visible units. The weights can be trained to model a joint probability distribution over $\mathbf{h}$ and $\mathbf{v}$ (Equation 1, where $\mathbf{b}$ and $\mathbf{c}$ denote the biases of the hidden and visible units and $\sigma(x)$ the logistic sigmoid function). This particular configuration makes it easy to compute the conditional probability distributions, when $\mathbf{v}$ or $\mathbf{h}$ is fixed (Equation 2), enabling the reconstruction of input data based on partial information in $\mathbf{v}$. This is done by sampling the conditional probability distribution in Equation 2, where $h'_j = 1$ with $p(h_j|\mathbf{v})$ (and $h'_j = 0$ otherwise), and calculating the reconstructed data $\mathbf{v'}$, where $v'_i = p(v_i|\mathbf{h'})$.

$$-\log P(\mathbf{v},\mathbf{h}) \propto E(\mathbf{v},\mathbf{h}) = -\mathbf{c}^T\mathbf{v} - \mathbf{b}^T\mathbf{h} - \mathbf{h}^T W\mathbf{v} \tag{1}$$

$$p(h_j|\mathbf{v}) = \sigma(b_j + w_j^T\mathbf{v}) \tag{2}$$

To train a DBM, each RBM layer is trained separately using Contrastive Divergence learning [5]. This learning algorithm tries to minimize the difference between $\mathbf{v}$ and $\mathbf{v'}$ by changing the weights using a Hebbian-like learning rule such that $\triangle W \cong \mathbf{v}\cdot\mathbf{h} - \mathbf{v'}\cdot\mathbf{h'}$, with the network in the long run learning to approximate the joint probability distribution $P(\mathbf{v},\mathbf{h})$.

Figure 1 depicts the RBMs used to model the auditory and visual patterns (1a, b) and two possible configurations for fusion of these patterns (1c, d). In this work we apply the same architecture as the bimodal DBN[1] (1d), to optimize the learning of relations across modalities (see [9]). We do not apply the deep autoencoders as proposed in [9] as we assume both modalities will be present during training and testing. Also, we will explain how

[1] Deep Boltzmann Machine are also referred to as Deep Belief Networks (DBN).

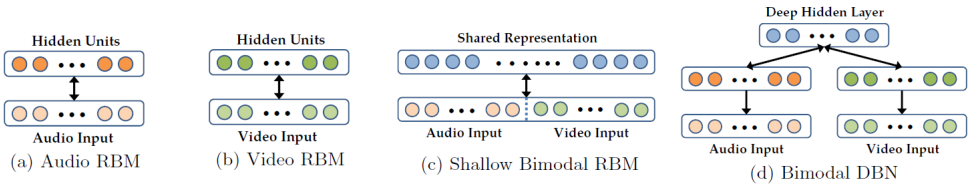

Figure 1 RBMs that model auditory (a) and visual patterns (b) and combine these patterns in higher-order multimodal representations (c, d).

certain neurological mechanisms can be used to overcome the multimodal inference problem addressed in [9]. To train the RBMs for audio (1a) and video (1b) we decode the audio stream as a spectrogram of 10 frames x 1024 frequencies using Discrete Cosine Transformation (DCT) on decoded audio samples, and the video stream as monochrome images that are reduced in scale, resulting in 160x120 pixels. Both transformations are fast and reversible allowing us to reconstruct video and audio from the DBM in real-time. As explained later, this approach will also enable us to do multimodal information retrieval and demonstrate the effect of neurological processes related to hallucinations and dreaming. As an extension to the DBM we also investigated the use of Recurrent Temporal RBMs in the top layer to model temporal sequences of audio and video patterns by taking into account the hidden unit activations in the previous time step [13].

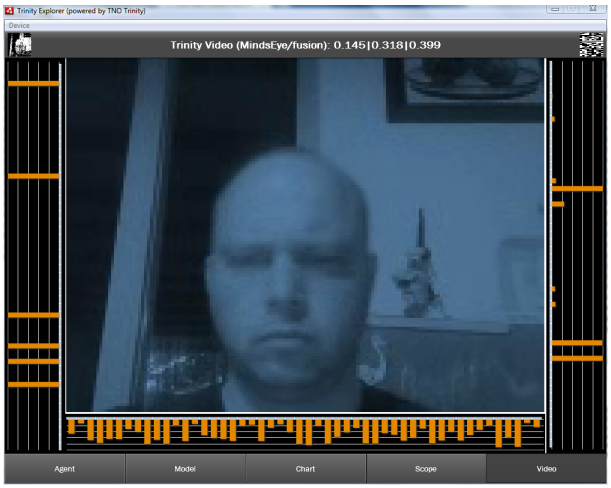

Figure 2 Adobe Flash based client that records from webcam and plays back reconstructed audio and video from a DBM. The real-time activations of all hidden units in the DBM is visualized as follows: on the left-side is depicted the hidden unit activations of the video, on the right-side the hidden unit activations of the audio, and on the bottom the hidden unit activations after multimodal fusion in the top layer of the DBM.

For demonstration purposes we implemented the DBM in a multi-agent platform, called Trinity[2], that supports real-time media streaming for Adobe Flash based clients that stream audio and video from a webcam. We implemented the DBM as part of a NSCA that enables the interpretation and reconstruction of audio and video information in a stream

[2] Trinity is a successor of the SimSCORM platform that has been developed for automated training and assessment [4].

and supports automated video indexing or assessment of observed human behaviour. As depicted in Figure 2, the client plays back the reconstructed audio and video and displays the real-time activation of the hidden units residing in each hidden layer. The use of a NSCA also allows to retrieve and investigate the contents of each visual, auditory and multimodal pattern that has learned by the DBM (see section 4). These patterns can be visualized in the tool by clicking on the bar of a related hidden unit in the activation graphs. We believe that this tool can help in future work on both multimodal fusion as well as the investigation of neurological processes.

3 Hallucinations and Dreaming

As described in [10], the DBM is a biologically plausible computational model for the investigation of neurological processes related to cortical learning, perception and diseases. It is able to simulate certain cortical processes that result in hallucinations due to loss of vision (i.e. Charles Bonnet Syndrome). Reichert describes how homoeostatic mechanisms in the cortex can stabilize neuronal activity to recover correct internal representations from degraded input. After some period this process can lead to complex vivid visual hallucinations. This mechanism can be implemented in the DBM as a regularization term for hidden unit biases (Equation 3) that is similar to mechanisms employed in other DBM-like models to enforce sparsity in the activations [9, 7].

$$\triangle b_i = \eta(p_i - a_i) \tag{3}$$

Using this model and regularization term, we have conducted several experiments that indeed demonstrate the forming of hallucinations when visual input is completely or partially blanked (mimicking loss of vision). These experiments also showed that when random noise is applied to the input, smaller overall bias shifts were necessary to restore original activity levels and produce hallucinations. This effect resembles another cognitive process, called reverse learning.

Reverse learning is a mechanism that is believed to be used in Rapid Eye Movement (REM) sleep to remove certain undesirable modes of interaction in networks of cells in the cerebral cortex. According to [1] the trace in the brain of the unconscious dream causes these modes to be weakened by applying random stimulation of the forebrain generated by the brain stem. This will tend to excite the inappropriate modes of brain activity, especially those which are too prone to be set off by random noise rather than by highly structured specific signals. Due to the random noise, overall neuron activity will drop, similar to the overall bias shift described before, automatically weakening the connections that encode these inappropriate modes.

Basically this means that reverse learning can also be regarded as a form of homoeostasis, which is beneficial to the integrity of the human brain and can be implemented in a DBM using the same stabilization mechanism as described before. We have implemented these mechanisms in all hidden layers of our DBM, resulting in a form of multimodal hallucination and dreaming. The effect of these mechanisms on the quality of the knowledge encoded in the DBM is still under investigation, but preliminary results have shown that the DBM indeed recovers from loss of audio or video input, producing hallucinations during stabilization of neuron activity, and that sparsity has improved the overall quality of the temporal relations encoded in the model.

4 Multimodal Information Retrieval

As described in [9], DBMs can produce good models for multimodal inference. For example, for the reconstruction of phonemes based on a visual representation of the mouth, and vice versa. But this approach will not explain the complex temporal relations encoded in a multimodal DBM. With a NSCA we are able to use an extraction mechanism that allows us to describe these multimodal relations, for example in terms of logic-based rules. As described in [3, 2], a NSCA uses the conditional probability distributions of a RBM to extract logic-based rules that describe the temporal relations between beliefs B encoded by the visible units and hypotheses H encoded by the hidden units. Typically, the temporal relations are represented by clauses of the form $H_1 \leftrightarrow B_1 \wedge B_3 \wedge \bullet H_1$ which denotes that hypothesis H_1 holds at time t if and only if beliefs B_1 and B_3 hold at time t and hypothesis H_1 holds at time $t-1$, where we use the previous time temporal logic operator $\bullet$ to denote $t-1$ [6]. If we extend this approach to a DBM we get clauses that describe hierarchical relations between hypotheses $H^{(l)}$ and lower-order hypotheses $H^{(l-1)}$. If we apply this notation to our multimodal DBM for audio and video we get clauses that describe higher-order temporal relations between auditory and visual patterns, such as $H_1^{fusion} \leftrightarrow H_1^{audio} \wedge H_4^{video} \wedge \bullet H_2^{fusion}$, and lower-order relations describing the most likely auditory and visual patterns in terms of pixels and frequencies, such as $H_4^{video} \leftrightarrow B_{10}^{video} \wedge B_{443}^{video} \wedge B_{753}^{video} \wedge ....$

Such textual descriptions would of course be very elaborate and impractical to understand at the level of individual pixels or frequencies. Therefore, we have implemented an iconographic representation for these visual and auditory patterns that enables us to present more compact and meaningful descriptions of H^{video} and H^{audio}. Similar to the approach suggested in [5], to investigate the weights of a RBM in terms of 2D images, we create icons from the pixel and frequency patterns that are extracted for each hypothesis H_j^{video} and H_k^{audio} and resample them in black and white to emphasize the most significant aspects of the patterns. An example, extracted during one of the experiments, of an iconographic temporal logic description of a multimodal relation is given in Equation 4. The first two icons show a person on the left side of the camera with a hand under his head. The other two icons visualize spectrograms of 10 frames x 1024 frequencies depicting the word "hel-lo" in phonemes.

$$H_{42}^{mind} \leftrightarrow \text{  } \wedge \text{  } \wedge \text{  } \wedge \text{  } \wedge \bullet H_7^{mind} \tag{4}$$

5 Conclusions and Future Work

Computational models used in AI research, such as the RBM and DBM, are becoming popular instruments in neural, cognitive and social sciences for the investigation of the human brain. In this paper, we discussed how these instruments can be used to model and simulate certain neurological processes, related to hallucinations and dreaming, such as homoeostasis and reverse learning. We have explained how such processes are beneficial to the recovery of appropriate and the reduction of inappropriate traces of the brain and implemented these mechanisms in a multimodal DBM for streaming audio and video. Early experiments with the DBM have shown similar effects as in homoeostasis and reverse learning (i.e. multimodal hallucinations and dreaming) and we expect these mechanisms will improve the integrity of the model, by stimulating sparsity, recovery of missing input, and unlearning inappropriate relations.

As part of future work we will investigate the actual improvements to the overall quality of the model, using benchmarks for comparison with other models, but also using the knowledge extracted from our model for expert analysis. In preparation of this, we already implemented the multimodal DBM as part of a NSCA that is able to extract temporal relations between auditory and visual patterns in the form of a iconographic temporal logic formula. Such a representation makes it practical to describe the visual and auditory patterns in terms of images and spectrograms. This will help us to understand and investigate the complex temporal relations encoded in multimodal DBMs and explain why certain neurological phenomenon occur, either in the DBM as a computational model or in the human brain that it tries to simulate.

References

1 F Crick and G Mitchison. The function of dream sleep. *Nature*, 304(5922):111–114, 1983.

2 Leo de Penning, Artur S. d'Avila Garcez, Luís C. Lamb, and John-Jules C. Meyer. A Neural-Symbolic Cognitive Agent for Online Learning and Reasoning. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, Barcelona, Spain, 2011.

3 Leo de Penning, R.J.M. den Hollander, H. Bouma, G.J. Burghouts, and A.S d'Avila Garcez. A Neural-Symbolic Cognitive Agent with a Mind's Eye. In *Workshop on Neural-Symbolic Learning and Reasoning at AAAI*, 2012.

4 Leo de Penning, Bart Kappé, and Eddy Boot. Automated Performance Assessment and Adaptive Training for Training Simulators with SimSCORM. In *Proc. of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, pages 1–7, Orlando, USA, 2009.

5 Geoffrey E Hinton. Learning to represent visual input. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1537):177–84, January 2010.

6 Luís C. Lamb, R.V. Borges, and Artur S. d'Avila Garcez. A connectionist cognitive model for temporal synchronisation and learning. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 827–832. AAAI Press, 2007.

7 Honglak Lee, C Ekanadham, and A Ng. Sparse deep belief net model for visual area V2. *Advances in Neural Information Processing Systems*, 20, 2008.

8 Martial Mermillod, Robert M. French, Paul C. Quinn, and Denis Mareschal. The importance of long-term memory in infant perceptual categorization. In *Proc. of the 25th Annual Conference of the Cognitive Science Society*, Boston, Massachusetts, 2003.

9 Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal Deep Learning. In *International Conference on Machine Learning (ICML)*, Bellevue, WA, USA, 2011.

10 D.P. Reichert, Peggy Series, and A.J. Storkey. Hallucinations in Charles Bonnet Syndrome Induced by Homeostasis: a Deep Boltzmann Machine Model. *Advances in Neural Information Processing Systems*, 23(23):2020–2028, 2010.

11 Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

12 E. Smith and S. Kosslyn. *Cognitive Psychology: Mind and Brain*. Prentice-Hall, 2006.

13 Ilya Sutskever. The recurrent temporal restricted boltzmann machine. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

14 Chantal Natalie van der Wal. *Social Agents: Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes*. PhD thesis, Vrije Universiteit Amsterdam, 2012.