Proposal of a Generative Model of Event-based Representations for Grounded Language Understanding

Simon Brodeur[§], *Luca Celotti*[§] , *Jean Rouat*

NECOTIS, Département génie électrique et génie informatique,
Université de Sherbrooke, Sherbrooke, Canada
Simon.Brodeur@USherbrooke.ca, Luca.Celotti@USherbrooke.ca, Jean.Rouat@USherbrooke.ca

Abstract

Grounding is the problem of correspondence between the symbolic concepts of language and the physical environment. The research direction that we propose to tackle language acquisition and grounding is based on multimodal event-based representations and probabilistic generative modeling. First, we establish a new multimodal dataset recorded from a mobile robot and describe how such multimodal signals can be efficiently encoded into compact, event-based representations using sparse coding. We highlight how they could be better suited to ground concepts. We then describe a generative probabilistic model based on those event-based representations. We discuss possible applications of this probabilistic framework in the context of a cognitive agent, such as detecting novelty at the inputs or reasoning by building internal simulations of the environment. While this work is still in progress, this could open new perspectives on how representational learning can play a key role in the ability to map structures of the multimodal scene to language.

Index Terms: probabilistic modeling, sparse coding, event-based representation, generative modeling, cognitive agent.

1. Introduction

Human language relies on the ability to map the physical nature of concepts to words during real-life interactions. This correspondence problem is known as symbol grounding [1]. Since words can have meanings related to objects, actions or even emotions, grounding must involve the very same sensors that we use to perceive objects, plan motor actions or express emotions. Here we will use *modalities* to refer to sensors, and *multimodal grounding* to mean to associate words to multiple sensory inputs. More specifically, grounding represents the link between symbolic constituents of language and the internal representation of the multimodal scene. This allows to make sense of the communicated message in ambiguous situations. Ultimately, conceptual representations are learned and used independently of the presence of the objects concerned [2]. In this paper, we will, however, focus on the functional representations linked with the sensorimotor experiences. Abstract concepts that have no realization in the external world are not discussed in this research contribution.

Grounding depends on representational learning to extract relevant structures from the multimodal scene that is perceived by the interactants. Symbolic knowledge can be acquired from the multimodal scene by associating particular words to the related objects that are perceived (e.g. [3]). This requires learning in a situational context, with an embodied interactants sharing the same external environment. The main idea of this paper is that the internal representation may be learned in such way to facilitate future grounding and acquisition of symbols if constraints such as sparsity are used to obtain object and event-based representations. We can summarize the contribution of this work as follows:

- we establish a new multimodal dataset recorded from a mobile robot interacting with the environment. We discuss how this dataset can be used for offline evaluation of algorithms on learning multimodal representation and grounding action-related concepts (Section 2.1).
- we propose a method to learn internal representations of temporal inputs using sparse coding. Sparse coding is used to generate a part and event-based representation that could be better suited to ground concepts (Section 2.2).
- we describe a generative probabilistic model of multimodal scenes based on part and event-based representations, in the perspective that internal simulations of the environment based on sampling such model would be part of the grounding process (Section 2.3).
- we discuss the role of probabilistic and generative modeling in the context of a cognitive agent acquiring language (Section 3).

2. Proposed Object-based Representation and Generative Modeling of Sensory Inputs

2.1. Dataset for Grounding Action-related Concepts

Sensory and symbolic knowledge can be acquired from recorded multimodal data to facilitate experiments and ensure reproducibility of the results. We thus designed a dataset [1] in the context of a mobile robot that can learn multimodal representation of its environment. The ability of the robot to perform actions and navigate in the environment is used to learn the dependency and relationship between modalities (e.g. vision, audio, proprioception). An example of the multimodal signals from a subset of the 15 sensors is shown in Figure 1. A little over 9 hours of multimodal data have been recorded in various scenarios (e.g. exploration only, or with human interaction). Such database is of interest for language acquisition and grounding since it is possible to work directly on the raw input signals and represent them in a vector space to obtain the environmental representations. Low-level knowledge of the environment can be manipulated and studied with that database to ground the actions of the robot within the environment with a very simple and limited semantic that characterizes its actions. The multimodal dataset can be used in several ways within the goal of sensorimotor knowledge acquisition, such as multimodal unsupervised object learning or multimodal prediction.

[§] Both authors contributed equally to this work

[1] Dataset specifications and tools available online: https://github.com/sbrodeur/ros-icreate-bbb

10.21437/GLU.2017-14

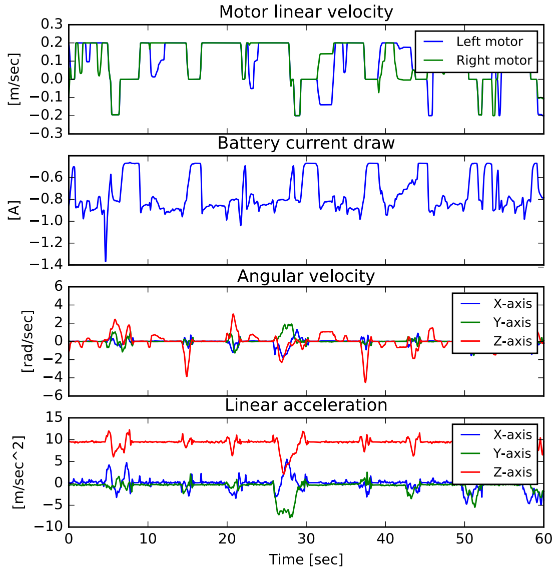


Figure 1: *Example of multimodal signals from various sensors of a mobile robot navigating in a room. An internal representation of the multimodal scene characterized by the various sensors allows to ground action-related concepts into the sensorimotor state of the robot. For instance, the robot could acquire the meaning of moving forward by learning how the motors impact the battery current draw, linear acceleration and angular velocity. It could also acquire the meaning of being manipulated (e.g. as occurring during the time interval of [25,30] seconds) when such interaction event is linked with large variations in linear accelerations.*

The dataset has to be annotated with a vocabulary of about 20 words. Examples of language annotations linked to the sensorimotor state of the robot are: "sharp turn left", "lifted up", "wandering", "accelerating forward", "hitting an obstacle". Aside from simple words such as "left", "right" or "forward", the annotations could also include a variety of adjectives and modifiers such as "sharp". This provides a way to acquire sensorimotor and symbolic knowledge related to diverse concepts such as actions, directions and speeds when interacting with the environment.

2.2. Sparse coding of time series

Assume a simple linear decomposition based on the discrete convolution [2] as $\mathbf{S} = \mathbf{X} * \mathbf{D}$. The variable $\mathbf{S}$ defines a temporal sequence of length N_t, with feature vectors of dimensionality N_s, i.e. $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_{N_t}] \in \Re^{N_t \times N_s}$. The dictionary $\mathbf{D}$ is a multidimensional array (tensor) of rank 3, i.e. $\mathbf{D} \in \Re^{K \times W \times N_s}$. The dictionary defines a set of K bases (or convolution filters) $\mathbf{d}_k$ of dimensionality $W \times N_s$, where W is the temporal length of the bases and N_s is the number of features from time samples of the input signal $\mathbf{S}$. The set of coefficients $\mathbf{X}$ is a multidimensional array of rank 3, i.e. $\mathbf{X} \in \Re^{N_t \times K \times N_s}$ that encode the input $\mathbf{S}$ by the means of sparse feature maps.

Sparse coding (for review, see [4]) aims at providing a set of sparse coefficients by the penalization of the L^0 norm of $\mathbf{X}$, as shown in Equation 1. The constant α controls the trade-off between the reconstruction error and the L^0 norm sparsity of the set of coefficients $\mathbf{X}$. Sparse coding allows to compact the energy of the representation into a few high-valued coefficients while letting the remaining coefficients sit at exactly zero. The linear decomposition based on the discrete convolution adds time shift (translation) invariance properties to the dictionary $\mathbf{D}$, and $\mathbf{D} \in \Re^{K \times W \times N_s}$ can now have temporal length $W < N_t$ to represent local features in the input signal $\mathbf{S}$.

$$(\mathbf{X}^*, \mathbf{D}^*) = \arg\min_{\mathbf{X},\mathbf{D}} \frac{1}{2}\|\mathbf{S} - \mathbf{X} * \mathbf{D}\|_2^2 + \alpha\|\mathbf{X}\|_0 \qquad (1)$$

$$\text{with constraint} \quad \|\mathbf{d}_k\|_2 = 1 \quad \text{for} \quad k \in \{1, 2, \ldots, K\}$$

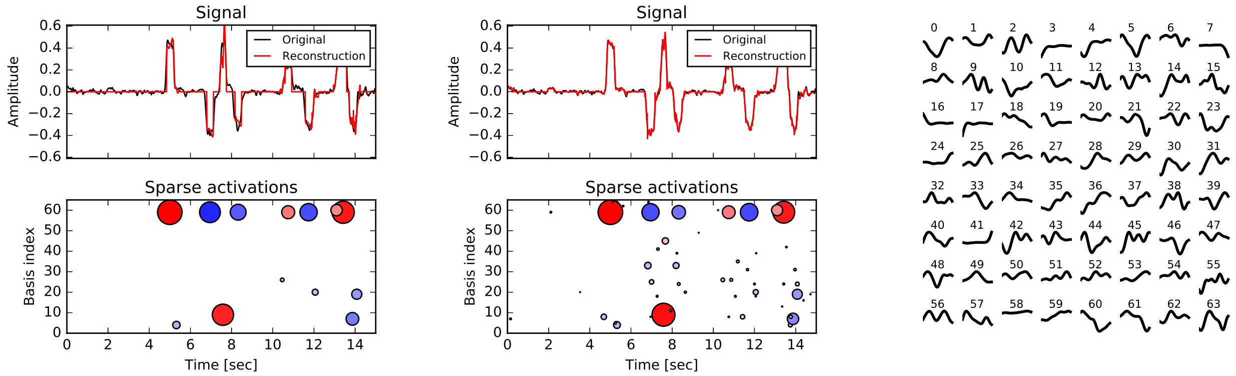
The optimization problem described in Equation 1 is NP-hard because of the constraint on the L^0 norm [5]. Matching pursuit [6] is one possible iterative algorithm to find an approximative solution to the optimal set of coefficients $\mathbf{X}$ with minimal non-zero support, given a dictionary $\mathbf{D}$. The approximative solution is good enough to remove redundancy and encode relevant structures from the input signal. The dictionary $\mathbf{D}$ may even be learned using other methods such as principal component analysis (PCA) or clustering using K-means. An example of real-life learned dictionary on gyroscope signals on a mobile robot is shown in Figure 2c. The real strength of sparse coding is the inference of a minimal set of coefficients to describe the input signal, given a learned dictionary $\mathbf{D}$. This is illustrated in Figure 2a and 2b, where complex signals are represented with an arbitrary accuracy with only a few non-zero coefficients. This process gives an object-based representation of the signal that we hypothesize to ease grounding.

If the projection on a set of bases is overcomplete (i.e. there are more bases that input features), the use of a sparsity constraint introduces a non-linearity that is beneficial to compactly represent the input [9]. Overcomplete representations are also better at approximating the statistical density of the data [10]. This is why sparse coding is well adapted to feature extraction and denoising (e.g. [11]). Researchers have applied sparseness constraints to convolutional deep belief networks [12], energy-based models [13], rectifier neural networks [14] and non-negative matrix factorization methods [15] as a means to learn part-based or object-based representations. However none of these specific algorithms use a constraint on the L^0 norm [3] as in the convolutional sparse coding framework presented above. Those approaches cannot be easily interpreted as event-based representations.

In the brain, neuronal spiking can also be seen as an event-based representation where the timing of the spikes carry information across neural assemblies and brain regions. Unlike discrete signals with fixed sampling rate, it is an asynchronous representation and thus can be very sparse in the time dimension. Using a L^0 norm constraint achieves true sparsity (i.e. negligible coefficients are exactly zero) in time and bases domain and allows to convert a time series into an asynchronous representation. In fact, every non-zero sparse activation in the set of coefficients $\mathbf{X}$ can be thought of as an event occurring at a particular discrete time t_n, with a specific basis b_n and a given amplitude a_n. So like in the brain where raw sensory inputs (i.e. analog signals) are encoded to structured sequences of spikes by neural populations, discrete signals can also be converted into events

[2] The definition of convolution from machine learning is used here, where the filter is not time-reversed. This convolution is similar to a cross-correlation.

[3] Sparsity regularization is most commonly done by penalizing the L^1 norm of the output coefficients or activations, mainly due to the fact that L^0 norm is non-differentiable and non-convex.



(a) *10 dB SNR*

(b) *20 dB SNR*

(c) *Dictionary*

Figure 2: *Example of sparse encoding of a segment of z-axis angular velocity signal from our dataset, using the low-complexity orthogonal matching pursuit (LoCOMP) algorithm [7]. Sparse coding is used to convert the signal into an event-based representation with discrete time. The encoding is shown for different levels of signal-to-noise ratios (SNR). The signal (shown on top) is encoded as a linear combination of sparse activations distributed through time and basis indices of the dictionary (shown below). The dictionary learned using convolutional K-means [8] is composed of 64 bases with a temporal span of 320 ms, as shown on right. The size of circles (in (a) and (b))) is proportional to the absolute amplitude of the sparse activation, with blue and red colours respectively representing negative and positive amplitude. Encoding the coarse dynamic of the signal only requires a few sparse coefficients. One advantage of sparse coding is that redundancy and noise can be removed relatively easily. The learned bases might be linked with action-related words, such as turning left or right, and provide a means to ground these concepts into the sensorimotor state of the robot.*

that encode part or object-based features through time. For such representation, the set of coefficients is simply interpreted as a sequence of events $\mathbf{X} \rightarrow \{e_1, \ldots, e_{N_{nz}}\}$, where $N_{nz} = \|\mathbf{X}\|_0$ is the number of non-zero coefficients in the representation. Each event is described by the triple $e_n = \{t_n, b_n, a_n\}$. Event-driven algorithms can be very energy-efficient since they only need to update their internal state or output when a new event occurs, regardless of the time interval between two successive events (e.g. [16, 17, 18]).

2.3. Probabilistic Modelling of Time Series

Probabilistic modeling of time series has long been possible with hidden Markov models [20] and use commonly in speech recognition [21] and speech synthesis [22]. However, with the recent advances in machine learning, artificial neural networks have become more popular and proved to push further and further the state of the art in modeling and generation quality. For instance, conditional restricted Boltzmann machines has been used to model and generate motion style from human skeleton data [23]. Autoregressive neural networks have been able to generate very realistic speech in text-to-speech applications by working directly in time domain on raw speech signals [24]. Most if not all of these approaches have had great success in generating realistic and high quality signals by doing sample-by-sample prediction, i.e. predicting the next time domain samples sequentially one at a time.

In this section, we would like to introduce a very general framework for probabilistic modeling of multimodal event-based time series, which is illustrated in Figure 3. The architecture we propose is based on a long short-term memory (LSTM) [25] and a mixture density network (MDN) [26]. We have discovered recent parallel work which uses those algorithms to generate realistic accelerometer data [27]. The main difference is that our framework considers event-based inputs and outputs as described in Section 2.2. Instead of doing sample-by-sample prediction, we propose to explore a similar mechanism but that can operate on a larger temporal context at the level of events. Recall that events here describe objects or part-of-objects in the signal. We propose to predict the probability of the next event $e_{next} = (t_{next}, b_{next}, a_{next})$ given a past history of N_h events $h = \{e_1, e_2, \ldots, e_{N_h}\}$, as shown in Equation 2. In the multimodal case, those events h are related to different modalities (e.g. visual features, sound) and processed as parallel streams as illustrated in Figure 3.

$$P(e_{next}|e_1, e_2, \ldots, e_{N_h}; \theta) = P(e_{next}|h; \theta) \qquad (2)$$

The variable θ corresponds to parameters of the density function. To simplify the modeling of this density function, we consider that all three predicted variables of the next event (i.e. t_{next}, b_{next} and a_{next}) are statistically independent from each other, leading to the factorization in Equation 3.

$$\begin{aligned} P(e_{next}|h) &= P(t_{next}, b_{next}, a_{next}|h; \theta) \\ &= P(t_{next}|h; \theta_t) \cdot P(b_{next}|h; \theta_b) \cdot P(a_{next}|h; \theta_a) \end{aligned} \qquad (3)$$

The proposed framework is suitable to model the densities of both discrete and continuous variables. Time will be considered as discrete in our event-based representation because the input signal is a discrete signal. The basis information is already discrete since it refers to a finite set of bases in the sparse coding dictionary $\mathbf{D}$. To simplify the density functions, we can map the time and amplitude of the predicted event to discrete symbols. The Lloyd algorithm [28] can be used for optimal quantization of the time and amplitude, based on the statistical distribution of the data. With this quantization procedure, each density function can now be modeled by a multinomial distribution. This gives the ability to model any kind of discrete

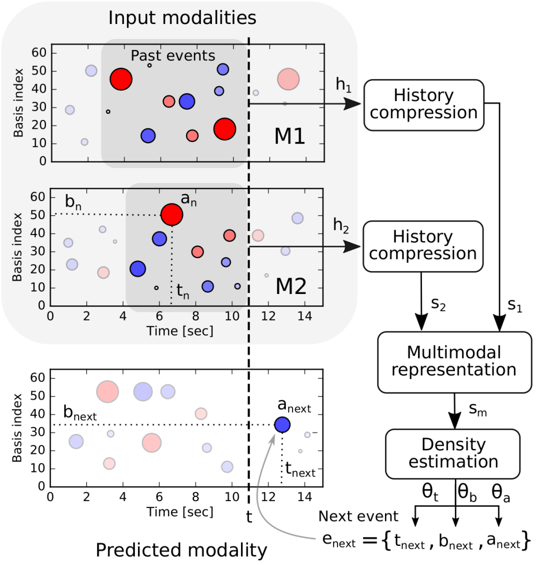


Figure 3: *Proposed architecture for probabilistic modeling of time series with multimodality and event-based representation. The inputs are sequences of events describing up to time t the past history $h_i = \{e_1, e_2, \ldots, e_{N_h}\}_i$ for each input modalities i. For each modality i, the variable-length event sequence is compressed into a fixed-dimensional vector s_i. A multimodal representation of the past history is built by aggregating those states into a single fixed-dimensional vector $s_m = \{s_1, s_2, \ldots, s_{N_m}\}$, where N_m is the number of input modalities. Using probability density estimation, the sets of parameters θ_t, θ_b and θ_a are estimated. Sampling the density functions $P(t_{next}|h;\theta_t)$, $P(b_{next}|h;\theta_b)$ and $P(a_{next}|h;\theta_a)$ allows to generate the next event e_{next}. Alternatively, the density functions can be used to evaluate the likelihood of the next event e_{next} given the set of histories $\{h_1, h2, \ldots, h_{N_m}\}$. History compression can be implemented with a Phased-LSTM [19]. Density estimation can be implemented for multinomial distributions by multilayer perceptrons with softmax outputs.*

distribution, without any assumption. The framework described in Figure 3 can be trained end-to-end using stochastic gradient descent to maximize the probability of guessing the next event correctly. This can be achieved by using loss functions such as the negative log-likelihood.

We have been performing preliminary experiments and evaluations of the proposed framework. A variation of the LSTM called Phased-LSTM [19] is currently tested for history compression due to its ability to deal with asynchronous data streams. We are using multilayer perceptrons for multimodal representation learning and probability density estimation. While this implementation currently achieves preliminary prediction performances greater than chance on our multimodal dataset, a deeper analysis remains to be done and results should provide greater understanding of the potential of the event-based probabilistic model compared to modeling the raw samples.

3. Discussion for a cognitive agent

A probabilistic model learned on multimodal time series inputs can have multiple applications in the context of a cognitive agent that has to interact with the environment and with humans and learn from its actions and sensory experiences. One important ability that such learning agent must possess is that it should be able *to know what it doesn't know*. This is the problem of novelty detection (for review, see [29]), where the agent itself is able to tell if a new multimodal scene has never been experienced before. The strength of probabilistic modeling is that by modeling the density of the input data, the rarity and novelty of sensory experiences can be estimated given the model. Novelty detection could be used in a proactive way to guide questions about the multimodal scene to interactively learn new concepts and their sensorimotor correlates. This would be useful in the context of the CHIST-ERA IGLU project, where we collaborate with several groups to build an agent that is able to learn via interactive dialogue with a human.

A probabilistic model can also be used as a generative model, since it describes a probability density function that can be sampled. This would be useful for a cognitive agent in a reinforcement learning architecture where imagined sensorimotor experiences are used to explore possible alternative solutions to maximize a reward function (e.g. [30]). Such reward function could be based on curiosity and desire to acquire new sensory and symbolic knowledge by interacting with the environment (e.g. [31]). The internal simulations would implement a basic conceptual system [32], and thus provide a much more powerful theory of grounded cognition. In this perspective, generating sensorimotor experiences could be used to ground the meaning of sentences that can be simulated, and assess if they make sense. For instance, the agent could interpret the sentence "move forward and you will hit a wall" by simulating the actions with the generative model and assessing of the consequences are similar, depending on the multimodal scene. If the agent is in front of a wall, it should be predictable that a collision will occur in those circumstances. However, if the agent is in a large open space, the sentence is ambiguous. Grounding language and understanding the meaning of sentences could thus be helped by sampling from a generative probabilistic model of sensorimotor experiences to assess the likelihood of action and object-related words given the context.

4. Conclusion

In this paper, we proposed an approach and a potential solution to build the sensorimotor representations in the process of grounding language in the real world. It will be validated and evaluated in the near future. We have designed and recorded a multimodal dataset from a mobile robot to be able to experiment with computational models of grounding action-related concepts into the sensorimotor state of the robot. We proposed an object and event-based representation along with a probabilistic model that can be learned on those representations. They integrate as potential components of the grounding process. We discussed probabilistic generative modeling capabilities in a cognitive agent, and the usefulness of internal simulations of the environment and novelty detection to acquire language. Future work will test the validity of our propositions.

5. Acknowledgements

The authors would like to thank the ERA-NET (CHIST-ERA) and FRQNT organizations for funding this research as part of the European IGLU project. NVIDIA Corporation supported this research with the donation of a Tesla K40. Many thanks to the reviewers for their constructive comments.

6. References

[1] S. Harnad, “The symbol grounding problem,” *Physica D*, vol. 42, pp. 335–346, 1990.

[2] P. Perniss and G. Vigliocco, “The bridge of iconicity: from a world of experience to the experience of language,” *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 369, no. 1651, 2014.

[3] C. Yu and D. H. Ballard, “On the integration of grounding language and learning objects,” in *Proceedings of the 19th National Conference on Artifical Intelligence*, ser. AAAI’04. AAAI Press, 2004, pp. 488–493.

[4] R. Rubinstein, A. M. Bruckstein, and M. Elad, “Dictionaries for sparse representation modeling.” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[5] B. K. Natarajan, “Sparse approximate solutions to linear systems,” *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, Apr. 1995.

[6] S. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *Trans. Sig. Proc.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[7] B. Mailhe, R. Gribonval, F. Bimbot, and P. Vandergheynst, “A low complexity orthogonal matching pursuit for sparse signal approximation with shift-invariant dictionaries,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3445–3448.

[8] A. Coates and A. Y. Ng, *Learning Feature Representations with K-Means*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 561–580.

[9] B. A. Olshausen and D. J. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?” *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[10] M. S. Lewicki and T. J. Sejnowski, “Learning overcomplete representations,” *Neural Comput.*, vol. 12, no. 2, pp. 337–365, Feb. 2000.

[11] A. Hyvarinen, P. Hoyer, and E. Oja, “Sparse code shrinkage for image denoising,” in *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, vol. 2, May 1998, pp. 859–864 vol.2.

[12] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML ’09. New York, NY, USA: ACM, 2009, pp. 609–616.

[13] M. aurelio Ranzato, C. Poultney, S. Chopra, and Y. L. Cun, “Efficient learning of sparse representations with an energy-based model,” in *Advances in Neural Information Processing Systems 19*, P. B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 1137–1144.

[14] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15. Fort Lauderdale, FL, USA: PMLR, 11–13 Apr 2011, pp. 315–323.

[15] D. D. Lee and H. S. Seung, “Learning the parts of objects by nonnegative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[16] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura *et al.*, “A million spiking-neuron integrated circuit with a scalable communication network and interface,” *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[17] G. Séguin-Godin, F. Mailhot, and J. Rouat, “Efficient event-driven approach using synchrony processing for hardware spiking neural networks,” in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2015, pp. 2696–2699.

[18] S. K. Esser, P. A. Merolla, J. V. Arthur, A. S. Cassidy, R. Appuswamy, A. Andreopoulos, D. J. Berg, J. L. McKinstry, T. Melano, D. R. Barch, C. di Nolfo, P. Datta, A. Amir, B. Taba, M. D. Flickner, and D. S. Modha, “Convolutional networks for fast, energy-efficient neuromorphic computing,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 41, pp. 11 441–11 446, 2016.

[19] D. Neil, M. Pfeiffer, and S.-C. Liu, “Phased lstm: Accelerating recurrent network training for long or event-based sequences,” in *Advances In Neural Information Processing Systems*, 2016, pp. 3882–3890.

[20] L. E. Baum and T. Petrie, “Statistical inference for probabilistic functions of finite state markov chains,” *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554–1563, 12 1966.

[21] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.

[22] A. Ljolje and F. Fallside, “Synthesis of natural sounding pitch contours in isolated utterances using hidden markov models,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1074–1080, Oct 1986.

[23] G. W. Taylor and G. E. Hinton, “Factored conditional restricted boltzmann machines for modeling motion style,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML ’09. New York, NY, USA: ACM, 2009, pp. 1025–1032.

[24] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio.” *CoRR*, vol. abs/1609.03499, 2016.

[25] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] C. M. Bishop, “Mixture density networks,” Dept. of Computer Science and Applied Mathematics, Aston University, Tech. Rep. NCRG/94/004, 1994.

[27] M. Alzantot, S. Chakraborty, and M. Srivastava, “Sensegen: A deep learning architecture for synthetic sensor data generation,” in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, March 2017, pp. 188–193.

[28] S. Lloyd, “Least squares quantization in PCM,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, March 1982.

[29] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, “A review of novelty detection,” *Signal Processing*, vol. 99, pp. 215–249, 2014.

[30] R. S. Sutton, “Dyna, an integrated architecture for learning, planning, and reacting,” *SIGART Bull.*, vol. 2, no. 4, pp. 160–163, Jul. 1991.

[31] J. Gottlieb, P. Y. Oudeyer, M. Lopes, and A. Baranes, “Information-seeking, curiosity, and attention: Computational and neural mechanisms,” *Trends in Cognitive Sciences*, vol. 17, no. 11, pp. 585–593, 2013.

[32] L. Barsalou, “Perceptual symbol systems,” *Behavioral and Brain Sciences*, vol. 22, no. 4, pp. 577–609, 1999.