ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 444-447
doi: 10.18178/wcse.2017.06.076

Multimodal Learning with Deep Associative Model

Dongwei Guo[+], Zhihua Zeng

College of Computer Science and Technology, Jilin University, JLU

Abstract. An associative-generated model based on Deep Belief Network (DBN) and Bi-directional Associative Memory (BAM) neural network is presented. The model is capable of extracting features from multiple input modalities and forming an associative memory. By such associative memorization, the model can regenerate one channel from the other, and perform classification with missing inputs.

Keywords: associative memory, generation, multimodal learning, deep belief network

1. Introduction

Humans learn a given concept through multiple channels and in different forms. As human being, we associate information received in multiple channels by associative memorization. Hence we can reproduce information from a channel using given information from another. In this paper, we are demonstrating a novel associative memorization model which was built based on DBN [1] and BAM [2] to simulate this associated learning. The model we developed is capable of learning multi-modal data and forming an associative memory. Thus it can reconstruct the missing information. For instance, given an audio signal, it can reconstruct an associated image. The features extracted from multiple channels are combined as a common representation, and such a representation are then transferred over to a classifier. Owing to the associative memorization, our system can perform well on classification tasks even when some signals are missing.

2. Background

There are two technologies related to our associative memorization system introduced in this chapter.

2.1. Deep Belief Network (DBN)

Containing stacked layers of Restricted Boltzmann Machine (RBM) [3], Deep Belief Network (DBN) shown in Fig. 1 is competent to learn multiple levels of abstraction so that the feature extraction process can be intellectualized. [4] A RBM has a two-layer structure: a visible layer (v) and a hidden layer (h), which are fully connected by bidirectional weighted edge, denoted by w. Layer v has bias b while layer h has bias c. The energy function of RBM is shown in formula 1.

$$E(v,h) = -\sum_{i \in v} b_i v_i - \sum_{j \in h} c_j h_j - \sum_{i,j} v_i h_j w_{ij} \tag{1}$$

By adjusting the weights and biases to lower the energy function (1), RBM can model the distribution of its input [5], and therefore DBN offer an unsupervised manner to learn multi-layer representations of data by stacking many layers of them.

[+] Corresponding author. 13009006800
E-mail address: guodw@jlu.edu.cn

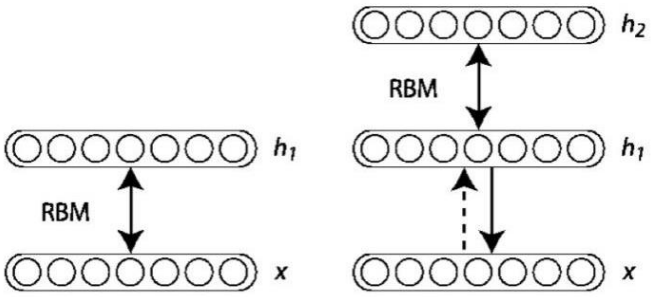

Fig. 1: Deep belief network.

2.2. Exponential Bidirectional Associative Memory Network (E-BAM)

Proposed by Kosko [6], BAM network is a two-layer hetero-associative neural network that is capable of storing and recalling bipolar pairs $(A_1, B_1), \cdots, (A_p, B_p)$, where $A \in \{-1,1\}^n, B \in \{-1,1\}^m$. Assume that the pairs needed to be storied are $\left(\mathrm{A_i}, \mathrm{B_i}\right)(\mathrm{i} = 1, 2, \cdots, \mathrm{m})$, and the evolutionary behaviour of the BAM is shown below:

$$\mathrm{A} \to \mathrm{W} \to \mathrm{B}$$

$$\mathrm{A}' \leftarrow \mathrm{W}^{\mathrm{T}} \leftarrow \mathrm{B}$$

where $\mathrm{A_i} = \left(\mathrm{a_{i1}}, \mathrm{a_{i2}}, \cdots \mathrm{a_{in}}\right)^{\mathrm{T}}$, $\mathrm{B_i} = \left(\mathrm{b_{i1}}, \mathrm{b_{i2}}, \cdots \mathrm{b_{ip}}\right)^{\mathrm{T}}$, $\mathrm{W} = \sum_{\mathrm{i}}^{\mathrm{m}} \mathrm{A_i^T B_i}$. Given an initial pair $\left(\mathrm{A^{(0)}}, \mathrm{B^{(0)}}\right)$, the evolutionary rule is shown in formula 2.

$$\begin{aligned} \mathrm{B^{(t+1)}} &= \mathrm{sgn}\left(\mathrm{WA^{(t)}}\right) \\ \mathrm{A^{(t+1)}} &= \mathrm{sgn}\left(\mathrm{W^T B^{(t)}}\right) \end{aligned} \tag{2}$$

Based on this hetero-associative network, Jin Cong proposed an exponential bidirectional associative memory network [7], which provided a solution to the problem that the capacity of pattern pair of Kosko's network is lower than the dimensions of the pattern spaces. And the evolutionary rule of E-BAM is shown in formula 3.

$$\begin{aligned} \mathrm{b_k^{(t+1)}} &= \mathrm{sgn}\left(\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{b_{ik}} \theta^{\mathrm{A_i^T \cdot A^{(t)}}}\right), 1 \le \mathrm{k} \le \mathrm{p} \\ \mathrm{a_j^{(t+1)}} &= \mathrm{sgn}\left(\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{a_{ij}} \theta^{\mathrm{B_i^T \cdot B^{(t+1)}}}\right), 1 \le \mathrm{j} \le \mathrm{n} \end{aligned} \tag{3}$$

3. Theory

A deep belief network can perform feature extraction task. Initially, two deep belief networks independently receive their input signals which are handwritten images and audio signals. Each DBN contains several stacked layers of RBMs. We use contrastive divergence (CD) algorithm [8-9] to extract features layer by layer in DBNs. After the two DBNs have been trained, the learned representations are used as inputs for our E-BAM network, in this manner, our system can associatively memorize two features extracted by two different DBNs. Therefore, we can regenerate one channel given the other. The image channel of the system learns handwritten images of Arabic numeral 0 to 9 while the audio channel learns audio representation of spoken digits zero to nine. After extracting the features of audio and image, and associating them together, a classifier, which consists of 10 softmax units, is trained. The input of this classifier is the combination of the features extracted by DBNs. The associated generative model is shown in Fig. 2.

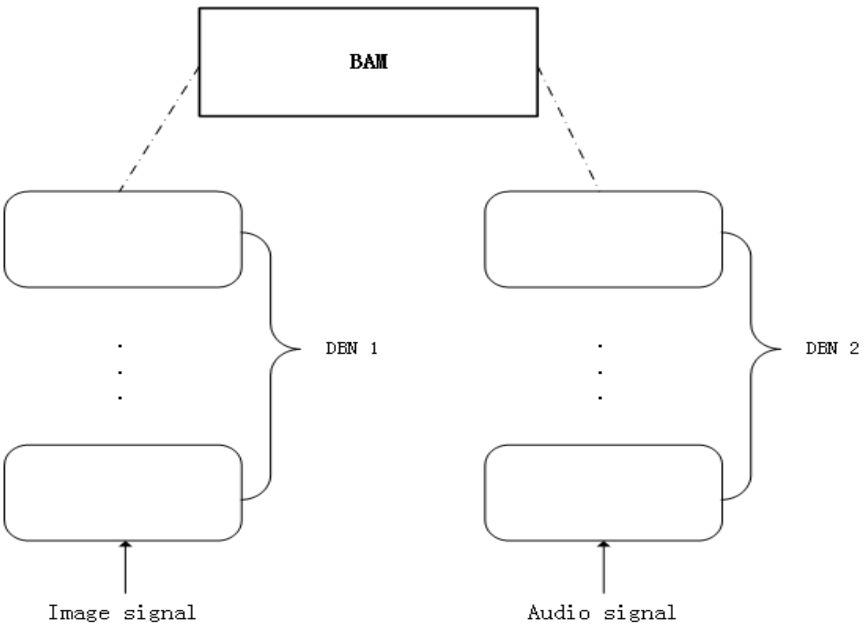

Fig. 2: The associated generative model.

4. Experiments

The main objective is to test the ability of our system to learn muti-channel and reconstruct one channel given the other. We use the MNIST digital image dataset [10] as image input with each image containing 28*28 pixels. The dataset of audio recordings is collected from 10 students, 5 males and 5 females. Each individual was asked to speak digits 0 through 9 for 10 times. The audio, at 48000Hz, is transformed to a frequency spectrum using short-time Fourier transform (STFT) [11], and we use 12000 length hamming window function and 50% overlap. We keep only the absolute values of the STFT and ends up with inputs of dimension 42007. We train the "audio DBN" with data from eight students and test it with the remaining two students' data. Correspondingly, we train the 'image DBN' with 800 images randomly chosen from MNIST dataset. After the layer-wise training of two DBN, we train our BAM network with the learned representations.

4.1. Regenerating One Channel from the Other

As is shown in Fig. 3, our model, as an associated generative network, is capable of reconstructing the input on one channel given another. Because the STFT transformation of audio data is irreversible, we cannot produce an audio signal. As a result, we verify the feasibility of the associated generative model by reconstructing images given audio signal and evaluating the results using MAP (mean average precision). Figure x shows some images reconstructed from the audio test set. Our model achieves a MAP over 10 classes of 71.3%. Table 1 shows the precision of each digit over 10 runs.

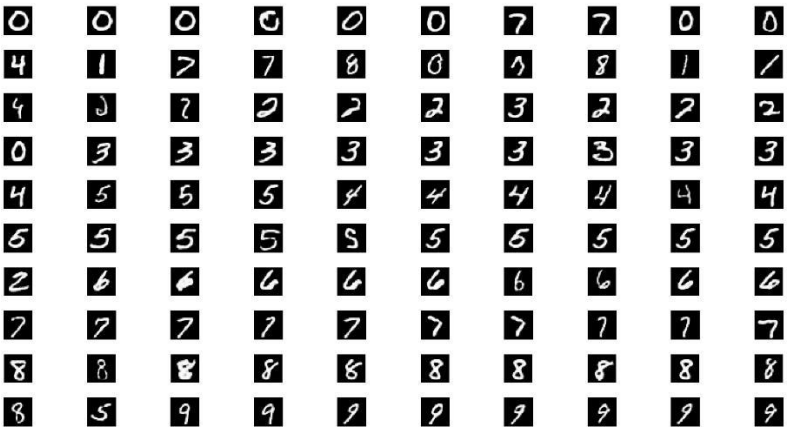

Fig. 3: Example of generated images.

Table 1: Regeneration Precision

Digit	0	1	2	3	4	5	6	7	8	9	Average
Precision(%)	43.57	75.20	50.62	87.45	57.79	77.51	91.03	66.37	92.10	71.36	71.30

4.2. Classification with Missing Inputs

As argued above, our BAM network has memorized two input channels' features, which are the top-level representations of DBNs, and can be used as input for classification. Hence we can naturally infer missing values through our associative memorization model. In order to verify the validity of the method to complete missing data, we train discriminative model which consists of 10 softmax units indicating 10 categories of digits using the training set as given above, and at test time, missing data is filled in by associated memorization. Table 2 show the averaged results of the classification of uni-modal inputs by filling in another modality, and the results of multi-modal inputs.

Table 2: Results of the classification of Uni-modal Inputs

Input	0	1	2	3	4	5	6	7	8	9	Average
Image	96.3	95.9	89.4	92.0	96.6	92.7	97.7	91.3	93.9	90.1	93.6
Audio	86.1	96.5	91.6	82.9	82.2	86.9	98.7	91.1	91.0	79.4	88.7
Both	97.6	98.6	96.2	96.5	97.4	97.0	99.7	96.4	95.2	95.5	97.0

5. Conclusion

We introduced an associative generated model that is capable of learning deep representation from image channel and audio channel, regenerating one channel from the other. It also works nicely when performing classification with missing modality. In the future, we hope that the data modalities can be more diverse, for instance, including video modality and text modality.

6. References

[1] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[2] Kosko B. Adaptive bidirectional associative memories[J]. Applied optics, 1987, 26(23): 4947-4960.

[3] Hinton G. A practical guide to training restricted Boltzmann machines[J]. Momentum, 2010, 9(1): 926.

[4] Hinton G E. Learning multiple layers of representation[J]. Trends in cognitive sciences, 2007, 11(10): 428-434.

[5] Tieleman T. Training restricted Boltzmann machines using approximations to the likelihood gradient[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 1064-1071.

[6] Kosko B. Bidirectional associative memories[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1988, 18(1): 49-60.

[7] Cong Jin. On memory capacity of the discrete exponential bidirectional associative memory network[J]. Pattern recognition and artificial intelligence, 2000, 13(1): 12-15. (in Chinese)

[8] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[9] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends® in Machine Learning, 2009, 2(1): 1-127.

[10] LeCun Y, Cortes C, Burges C J C. The MNIST database of handwritten digits[J]. 1998.

[11] Allen J. Short-term spectral analysis, and modification by discrete Fourier transform[J]. IEEE Transactions on Acoustics Speech and Signal Processing, 1977, 25(3): 235-238.