Data Fusion with Deep Boltzmann Machines for High-level Image Perception

Bingjie Xu, Ming Jiang, Shaojing Fan, Qi Zhao
National University of Singapore
bingjiexu, mjiang@u.nus.edu idmfs, eleqiz@nus.edu.sg

Abstract

Datasets have been collected independently for various studies in visual perception. Data fusion is crucial because of the high costs for data collection yet infrequent use in connection with each other - the power of big data in many tasks such as recognition and retrieval suggests effective fusion of them. In this paper, we propose to use Deep Boltzmann Machines (DBMs) to fuse various datasets. Our approach is advantageous in providing novel insights into visual perception study by incorporating multiple modality ground truth and large-scale datasets.

1. Introduction

Data fusion is crucial to save collection cost and provide novel insights across various problems. By fusing a number of partially overlapping datasets, Fan et al. [2] proposed a statistical model for image perception analysis. In this paper, we propose a deep learning model that addresses the following problems to make it more generalizable. First, the previous approach was based on linear statistical modeling but human perception is not a linear mechanism. Second, it only took into account the human perceptual ratings without utilizing any other ground truth. Third, the previous statistical modeling might not be able to deal with high-dimensional multimodal data, which cannot be represented by simple structures [3].

Deep Boltzmann Machines (DBM) [7] is advantageous by using unlabelled data to extract a good generative model and fill in missing data. In this work, we feed DBM with perceptual ratings plus image features, and we train the model in an unsupervised way that is necessary to fuse unlabelled datasets. The method is shown to be effective to fuse high-dimensional datasets. Our contributions are as follows. First, we take advantage of unsupervised learning ability of DBM to study non-linear relations among perceptual attributes. Second, the deep net is more scalable to incorporate multiple modalities [8] and handle high-dimensional fused modalities.

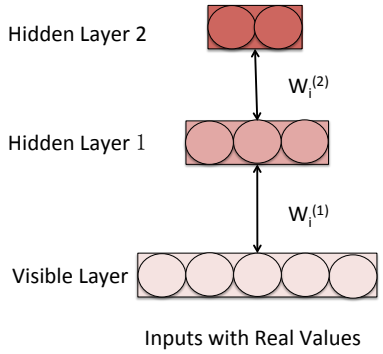

Figure 1. Visualization of the DBM used. Gaussian RBM is used to model the pathway with real-valued inputs. Inputs can either be perceptual ratings or ratings plus computational features of scenes.

2. Method

We propose a deep-learning approach to fuse multiple datasets using DBM. Figure 1 displays the representation of the model. Inputs are based on various experimental purposes: real-valued ratings for building the perceptual structure; real-valued ratings combined with image object bank features for missing data imputation and classification. Moreover, unit numbers of three layers serve for different training purposes: unit numbers in visible, the first hidden and the second hidden layer are 26, 6, 1 respectively for building the perceptual structure; 5602, 1024, 1024 for imputation and classification. We use Gaussian Restricted Boltzmann Machine (RBM) [5] to model visible layer with real values while the other layers with binary units.

Through unsupervised training, the model is able to extract the latent structure of image perceptions and imputed attributes for further classification tasks. We use Persistent Contrastive Divergence (PCD) for layer-wise training [4].

3. Experimental Results

We test on the visual realism dataset [1] fused with the memorability dataset [6]. After pruning out subjective attributes, 26 perceptual attributes are used in our experiments. There are 4000 images in total with 2000 in each

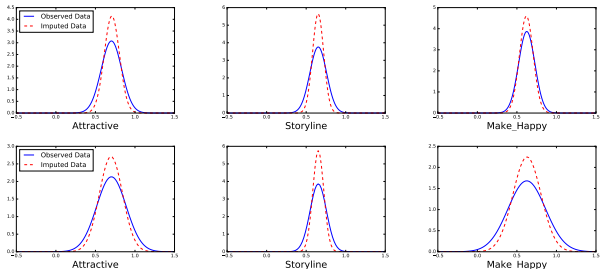

Figure 2. Sample frequency distribution of only the observed ratings and ratings after imputation of various attributes in the fused dataset. Distributions are from normal fitting. The first row indicates visual realism dataset and the second row indicates memorability dataset.

dataset. Furthermore, object bank features are of 5576 dimensions. We show the feasibility of applying DBM in modeling perception in the following experiments. First, we visualize the structure connecting basic to high-level image perceptual attributes. We train and test the model three times with different training parameters to see the stability of the structure. Second, we investigate its ability to impute missing data. Following the guide [4], we randomly initialize missing ratings and feed all ratings to the model, together with computational features . Third, we test our model on binary classification tasks, validating multimodality and data fusion power. SVMs are trained on all ratings after imputation and on only the observed ratings.

Our empirical results are as follows. First, three heavily weighted attributes connected to each hidden unit are shown stable through the three different runs. The structure indicates that the basic-level perceptions share common high-level properties. This consolidates previous visual perception studies and gives more insights into human perceptual judgments of natural scene images. Figure 3 displays the visualization of the structure learned from the fused dataset. Second, RBM can impute missing data through unsupervised training. Sample results are in Figure 2. It is shown that the imputed data share similar frequency distributions with the observed data. Finally, the imputation and multimodality improve classification performance. Classification results are shown in the Table 1. The deep net is also shown scalable to handle high-dimensional dataset fusion.

Method	Visual Realism		Memorability	
	AUC	Accuracy	AUC	Accuracy
All Ratings	0.79	0.82	0.51	0.81
Observed Ratings	0.77	0.80	0.50	0.81

Table 1. Classification results for visual realism and memorability of two datasets, respectively. Area under ROC curve (AUC) and accuracy are used as evaluation metrics.

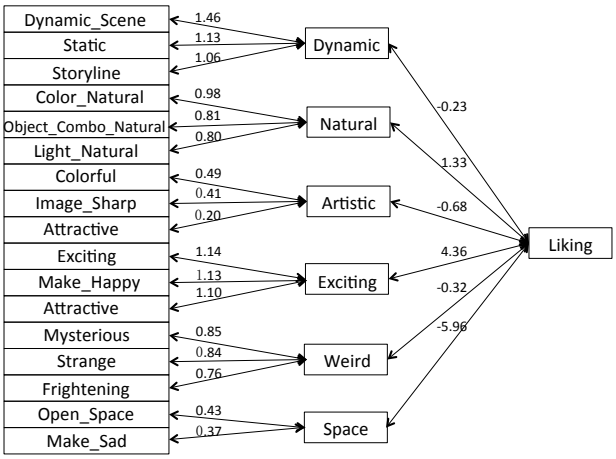

Figure 3. Human perception model of the fused dataset based on weights of DBM.

4. Conclusions

We propose to use DBM to fuse datasets and provide insights of human visual perception. The multimodal dataset fusion is shown to be effective in improving visual realism and memorability recognition performance. In future work, separate modalities with different representations, for instance sparse word count vectors, can be added to our model to fill in missing data by retrieval. This may further strengthen the imputation power of our model.

References

[1] S. Fan, T.-T. Ng, J. S. Herberg, B. L. Koenig, C. Y.-C. Tan, and R. Wang. An automated estimator of image visual realism based on human cognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4201–4208, 2014. 1

[2] S. Fan, T.-T. Ng, B. L. Koenig, M. Jiang, and Q. Zhao. A paradigm for building generalized models of human image perception through data fusion. *CVPR*, 2016. 1

[3] G. Hinton. Ucl tutorial on: Deep belief nets. 2009. 1

[4] G. Hinton. A practical guide to training restricted boltzmann machines. *Momentum*, 9(1):926, 2010. 1, 2

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

[6] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, pages 2429–2437, 2011. 1

[7] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *International conference on artificial intelligence and statistics*, pages 448–455, 2009. 1

[8] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012. 1