Solar Radio Astronomical Big Data Classification

Long Xu[*], Ying Weng[**], and Zhuo Chen

Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China
School of Computer Science, Bangor University, Bangor, UK
lxu@nao.cas.cn;y.weng@bangor.ac.uk;chenzhuo@nao.cas.cn

Abstract. *The Solar Broadband Radio Spectrometer (SBRS) monitors the solar radio busts all day long and produces solar radio astronomical big data for analysis every day, which usually have been accumulated in mass images for scientific study over decades. In the observed mass data, burst events are rare and always along with interference, so it seems impossible to identify whether the mass data contain bursts or not and figure out which type of burst it is by manual operation timely. Therefore, we take advantage of high performance computing and machine learning techniques to classify the huge volume astronomical imaging data automatically. The professional line of multiple NVIDIA GPUs has been exploited to deliver 78x faster parallel processing power for high performance computing of the astronomical big data, and neural networks have been utilized to learn the representations of the solar radio spectrum. Experimental results have demonstrated that the employed network can effectively classify the solar radio image into the labeled categories. Moreover, the processing time is dramatically reduced by exploring GPU parallel computing.* environment.

Keywords: Solar radio, big data, deep learning, classification

1 Introduction

Solar radio astronomy is an interdisciplinary subject of radio astronomy and solar physics. The discovery of radio waves from the Sun provided a new window to investigate the solar atmosphere. For example, the properties of the solar corona were much more easily determined at radio wavelengths. Solar radio telescopes have been improved a lot recently, so that fine structures in solar radio bursts can be detected. In this study, we use data obtained by Solar Broadband Radio Spectrometer (SBRS) of China [1].The SBRS is with characteristics of high time resolution, high-frequency resolution, high sensitivity, and wide frequency

[*] This work was partially supported by a grant from the National Natural Science Foundation of China under Grant 61202242, 100-Talents Program of Chinese Academy of Sciences (No. Y434061V01).

[**] Corresponding author

coverage in the microwave region. Its functionality is to monitor solar radio bursts in the frequency range of 0.7-7.6 GHz with time resolution of 1-10 ms. It consists of five 'component spectrometers' which work in five different wave bands (0.7-1.5 GHz, 1.0-2.0 GHz, 2.6-3.8 GHz, 4.5-7.5 GHz, and 5.2-7.6 GHz, respectively). The SBRS monitors the solar radio bursts all day long producing mass of data for researchers to analyze. In the observed data, burst events are rare and always with interference in the meantime. So it seems impossible to identify whether the data containing bursts or not and figure out which type of burst it is by manual operation timely. Thus, classifying the observed data automatically will be quite helpful for solar radio astronomical study.

Nowadays, for mass of data, many algorithms have been developed to learn the representation with unsupervised and supervised methods, especially the deep learning methods. Current methods based on deep learning [2] have demonstrated competitive performance in a wide variety of tasks, including visual recognition [3][4], audio recognition [5][6], and natural language processing [7]. These techniques are especially powerful because they are capable of learning useful features directly from both unlabeled and labeled data, avoiding the need for hand-engineering, which will be much helpful to the automatic analysis of the solar radio spectrum. Autoencoder (AE) can also be employed to learn the representation from the available mass data. AE is an unsupervised learning algorithm that applies backpropagation, setting the target values equal to the inputs. The AE tries to learn a function to make the input similar to the output of the function. There are many other variations of the AE, such as denoising AE [8], stacked AE (SAE) [9]. In [10], the authors proposed the automatic dimensionality reduction to facilitate the classification, visualization, communication, and storage of high-dimensional data. An adaptive, multilayer "encoder" network to transform the high-dimensional data into a low-dimensional code and a similar "decoder" network to recover the data from the code. With the random weights as the initialization in the two networks, they can be trained together by minimizing the discrepancy between the original data and its reconstruction. Then the representation can be learned in an unsupervised manner. The network can be further named as deep belief network (DBN). With the achievements of these learning methods, we can learn the representations of the solar radio spectrums, which will be employed for further solar radio image analysis, such as clustering, classification, and so on. In this paper, we make the first attempt to employ the deep learning method, specifically the DBN, to learn the representation of the solar radio spectrum. Based on the representation, we can further classify the solar radio spectrums into different categories automatically.

The rest of the paper is organized as following. In Section 2, the learning architecture is introduced to learn the representation of the solar radio image. Section 3 gives the experimental results on representation learning and classification.

2 Representation Learning and Classification for Solar Radio Images

SBRS contains several channels to monitor the solar burst in different frequencies. Therefore, the signal sensed from each channel will be treated individually. In total, there are 120 channels working toward the solar radio information captured at the same time. Moreover, each captured file contains both left and right circular polarization parts, which should be separated and processed individually. We extract the captured data from each channel as a row vector, which is stored according the sensing time. Afterwards, all the vectors from the 120 channels will be assembled together according the frequency values to form a solar radio spectrum, which is used for visualization and further processing. To reduce computational complexity, the solar radio spectrum is down-sampled into 75×30 image with the nearest neighbor sampling method.

We employ DBN to learn solar radio image representation. DBN is a multilayer, stochastic generative model which is created by stacking multiple restricted Boltzmann machines (RBMs). Each RBM is trained by taking the hidden activities of the previous RBM as its input data. Each time a new RBM is added to the stack, the new DBN has a better variational lower bound on the log probability of the data than the previous DBN, provided the new RBM is learned in the appropriate way [11].

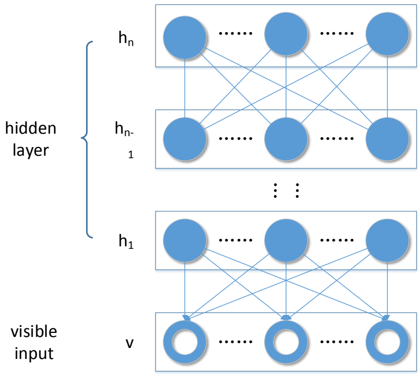

Fig. 1. DBN learning structure

2.1 RBM

RBM is a type of graphical model in which the nodes are divided into two sets, specifically, the visible and hidden. Each visible node is only connected to the hidden nodes. It means that there are no intra-visible or intra-hidden connections, which can be illustrated in each layer of Fig. 1. The energy function of an RBM with V visible units and H hidden units is defined in the following.

$$E(v,h) = -\sum_{i=1}^{V}\sum_{j=1}^{H} v_i h_j \omega_{ij} - \sum_{i=1}^{V} v_i b_i^v - \sum_{j=1}^{H} h_i b_i^h \tag{1}$$

where v is the binary state vector of the visible nodes, h is the binary state vector of the hidden nodes, v_i is the state of visible node i, h_j is the state of the hidden node j, ω_{ij} is the real-valued weight between the visible node i, the hidden node j. b_i^v is the real-valued bias into visible node i, and b_i^h is the real-valued bias into hidden node j. The joint distribution of the visible and hidden nodes is defined in the following:

$$p(v,h) = \frac{e^{-E(v,h)}}{\sum_u \sum_g e^{-E(u,g)}} \tag{2}$$

It can be observed that low energy results in high probability and high energy brings is assigned low probability. Also the probability of a visible node turning on is independent from the states of other visible nodes, given the states of the hidden nodes. Likewise the hidden states are independent from each other given the visible states. The property of RBM makes sampling extremely efficient, as one can sample all the hidden nodes simultaneously and then all the visible nodes simultaneously.

2.2 DBN

As mentioned before, each layer of DBN is composed of an RBM, where the weights in layer l are trained by keeping all the weights in the lower layers constant and taking as data the activities of the hidden units at layer $l+1$. Therefore, the DBN training algorithm trains the layers greedily and sequentially. Layer l is trained after layer $l-1$. If the size of the second hidden layer is the same as the size of the first hidden layer and the weights of the second is borrowed from the weights of the first, it can be proven that training the second hidden layer while keeping the first hidden layer's weights constant improves the log likelihood of the data under the model [12]. Fig. 1 illustrates the multilayer DBN. The probability of the DBN assigns to a visible vector is defined as:

$$p(v) = \sum_{h_1,\ldots,h_n} p(h_{n-1},h_n) \prod_{k=2}^{n-1} p(h_{k-1} h_k) p(v|h_1) \tag{3}$$

where n defines the number of hidden layers. In this study, we employ the DBN to learn the representation and perform the classification of the solar radio images.

2.3 Neural network for solar radio image classification

Based on the learning architecture in previous section, we propose a simple network for representation learning and classification of solar radio images. A classification layer with three output nodes is added on top of one RBM layer, which takes learned representation as input and outputs the classification results for each type of the solar radio image. For each type, the classification layer will determine the possibility about how the inputs will result in the specific type.

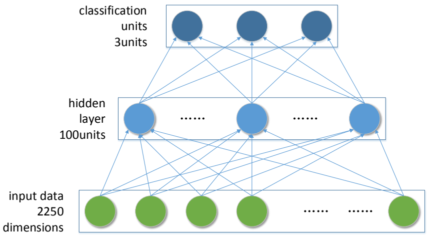

Fig. 2. DBN learning structure

The depth of the neural network depends on the problem and the size of the training set. Overfitting will occur with high probabilities if the training samples are insufficient, as the network requires a larger number of parameters. In this case, due to the limit number of solar radio images, only one hidden layer is employed. Then we propose the $I-H-C$ structure network for the experiment, as illustrated in Fig. 2. C , standing for the classification, is defined to give the prediction which most possible type the input is. I, indicating the number nodes of the input layer, is set as 2250 which is the number of dimensions of preprocessed data. H, standing for hidden, is defined as 100 nodes of hidden layer. The bottom layer of the employed network is the RBM and the top layer is a softmax layer for classification. In order to realize the non-linear mapping function for the classification, the object of the learning network is defined as following:

$$\hat{o} = \arg\min p(o|x;\Theta), \tag{4}$$

where Θ include all the parameters in RBM and softmax layers. In order to make the inference, we need to obtain the parameters of the constructed network, i.e., the parameters of RBM and softmax layer, respectively. For the parameters in the RBM layer, the standard contrastive divergence learning procedure is employed for pre-training. Detailed information about the pre-training method can be found in [13]. With the process of pre-training, the constructed network can effectively avoid the risk of trapping in poor local optima. After the pre-training process, the fine-tuning process needs to be further performed to make the network more suitable for solar radio spectrum classification. Thereby, a log-likelihood function is employed as the object function for further training the parameters in the softmax layers and fine-tuning the parameters in the RBM layer:

$$\Theta^* = \arg\max \sum_{t=1}^{k} \log P(\hat{L} = L|x; \Theta), \tag{5}$$

where k indicates the number of categories for determination, L represents the label of the inputs, and $\hat{L}$ represents the outputs of the network. For the parameter training, the traditional back-prorogation (BP) [14] is employed to fine-tune parameters of the constructed deep network. This algorithm is firstly proposed by Rumelhart and McCelland, the essence of which is to minimize the mean squared error between actual output and desired output based on gradient descent. BP algorithm is especially powerful because it can extract regular knowledge from input data and memory on the weights in the network automatically. Furthermore, in order to prevent over-fitting in training neural network, drop-out is introduced. Typically the outputs of neurons are set to zero with a probability of p in the training stage and multiplied with 1-p in the test stage. By randomly masking out the neurons, dropout is an efficient approximation of training many different networks with shared weights. In our experiments, we apply the dropout to all the layers and the probability is set as $p = 0.2$.

3 Experimental results

To evaluate the proposed representation learning and classification of solar radio spectrums, a solar radio spectrum database is established firstly. Then, the representation learning and classification of solar radio spectrums are tested on this database. For GPU acceleration, a high performance computing server with 4 GeForce GTX 780 GPU for computing and one GeForce 210 GPU for display is used in our simulation.

In this database, 4408 observational data files are labeled by the experts into six categories (0=no burst or hard to identify, 1= weak burst, 2=moderate burst, 3= large burst, 4=data with interference, 5=calibration). Since the objective of our experiment is to distinguish the bursts from others, the solar radio image in the database has been selected and relabeled to form a new database for the experiment. Three coarse categories, i.e., 'bursts', 'non-burst', and 'calibrations' are included in the database.

Table 1. The details of the database. 0=no burst or hard to identify, 1= weak burst, 2=moderate burst, 3= large burst, 4=data with interference, 5=calibration

Algorithm	0	1	2	3	4	5	total
Image Number	6670	618	268	272	570	988	8816

After preprocessing, we input the training set data to the network as batches. The hidden layer is firstly pre-trained to initialize the parameters in an unsupervised way. Then both the hidden layer and the classification layer are fine-tuned with labeled data. After that, the preprocessed testing set data will be input sequentially and the network will output the classification results in possibilities how likely the input data belongs to each category respectively. The model classifies a solar radio image successfully when the category with highest possibility output by the algorithm matches the labeled category of the file input. The classification results can be found in Table 2.

We also exploit the professional line of multiple NVIDIA GPUs to accelerate the computing of neural network. Since the computing of neural network concerns the same processing for the nodes, it is benefited greatly from GPU computing. In our simulation, the computing time by using only CPU (2 Inter(R) Xeon (R) CPU E5-2620 v2 @ 2.10GHz) is about 1716.64 minutes, and can be dramatically reduced to 21.90 minutes by employing GPU acceleration (4 GeForce GTX 780 GPU). Therefore, the GPU acceleration can deliver 78x faster parallel processing power for high performance computing of solar radio spectrum classification.

Table 2. Performance of DBN

	TPR	FPR
Burst	67.4%	13.2%
Non-burst	86.4%	14.1%
Calibration	95.7%	0.4%

References

1. Fu Q, Ji H, Qin Z, et al, A new solar broadband radio spectrometer (SBRS) in China, Solar Physics, 2004, 222(1): 167-173.
2. Y. Bengio, Learning deep architectures for AI, FTML, 2(1):1-127, 2009.
3. Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen,G. Corrado, J. Dean, and A. Ng, Building high-level features using large scale unsupervised learning, ICML, 2012.
4. K. Sohn, D. Y. Jung, H. Lee, and A. Hero, Efficient learning of sparse, distributed, convolutional feature representations for object recognition, ICCV, 2011.
5. H. Lee, Y. Largman, P. Pham, and A. Y. Ng, Unsupervised feature learning for audio classification using convolutional deep belief networks, NIPS, 2009.
6. A. R. Mohamed, G. Dahl, and G. E. Hinton, Acoustic modeling using deep belief networks, IEEE Trans Audio, Speech, and Language Processing, 20 (1):14-22, 2012.
7. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, Natural language processing (almost) from scratch, JMLR, 12:2493-2537, 2011.
8. M. Chen, K. Weinberger, F. Sha, Y. Bengio, Marginalized Denoising Autoencoders for Nonlinear Representation, ICML, 2014.
9. M. Chen, Z. Xu, K. Weinberger, F. Sha, Marginalized Stacked Denoising Autoencoders for Domain Adaptation 29th International Conference on Machine Learning (ICML), 2012.
10. Hinton, G. E. and Salakhutdinov, R. R, Reducing the dimensionality of data with neural networks. Science, Vol. 313. no. 5786, pp. 504 - 507, 28 July 2006.

11. G. E. Hinton, S. Osindero, and Y. the, A fast learning algorithm for deep belief nets, Neural Computation, 18:1527-1554, 2006.
12. Salakhutdinov, R., Murray, I., 2008, On the Quantitative Analysis of Deep Belief Networks, ICML 2008.
13. G. E. Hinton. A practical guide to training restricted Boltzmann machines. Technical Report, University of Toronto, 2010.
14. L. Deng, Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey, APSIPA Transactions on Signal and Information Processing, 2012.