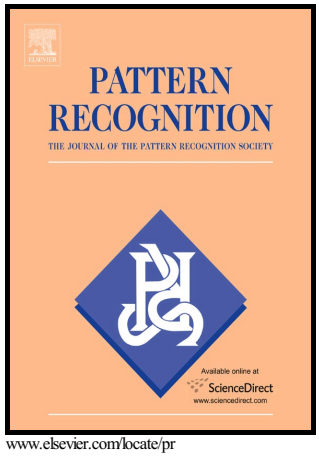
Multimodal deep learning for solar radio burst classification

Lin Ma, Zhuo Chen, Long Xu, Yihua Yan



www.elsevier.com/locate/pr

PII: S0031-3203(16)30059-0
DOI: http://dx.doi.org/10.1016/j.patcog.2016.04.013
Reference: PR5709

To appear in: *Pattern Recognition*

Received date: 31 January 2016
Revised date: 20 April 2016
Accepted date: 22 April 2016

Cite this article as: Lin Ma, Zhuo Chen, Long Xu and Yihua Yan, Multimoda deep learning for solar radio burst classification, *Pattern Recognitio* http://dx.doi.org/10.1016/j.patcog.2016.04.013

Multimodal Deep Learning for Solar Radio Burst Classification

Lin Ma[a], Zhuo Chen[b], Long Xu[b,*], Yihua Yan[b]

[a]*Huawei Noah's Ark Lab, Hong Kong*
[b]*Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China*

Abstract

In this paper, multimodal deep learning for solar radio burst classification is proposed. We make the first attempt to build multimodal learning network to learn the joint representation of the solar radio spectrums captured from different frequency channels, which are treated as different modalities. In order to learn the representation of each modality and the correlation and interaction between different modalities, autoencoder together with the structured regularization is used to enforce and learn the modality-specific sparsity and density of each modality, respectively. Fully-connected layers are further employed to exploit the relationships between different modalities for the joint representation generation of the solar radio spectrums. Based on the learned joint representation, solar radio burst classification is performed. With the validation on the constructed solar radio spectrum database, experimental results have demonstrated that the proposed multimodal learning network can effectively learn the representation of the solar radio spectrum, and improve the classification accuracy.

Keywords: Multimodal learning, solar radio spectrum, classification

*Corresponding author
Email addresses: forest.linma@gmail.com (Lin Ma), chenzhuo@nao.cas.cn (Zhuo Chen), lxu@nao.cas.cn (Long Xu), yyh@nao.cas.cn (Yihua Yan)

1. Introduction

Solar radio astronomy is an emerging interdisciplinary field of radio astronomy and solar physics. The discovery of radio waves from the Sun provides a new window to exploit and investigate the solar atmosphere, as new information about the Sun can be obtained. With proper devices, the properties of the solar corona are much more easily depicted with the captured signals at radio wavelengths. As solar radio telescopes have improved a lot in recent years, fine structures in solar radio bursts can thus be easily and accurately detected. In this study, in order to analyze the solar burst behavior, we use the data obtained by solar broadband radio spectrometer (SBRS) of China [1] which is a solar dedicated radio spectrometer for capturing solar radio strength along time over multiple frequency channels in the microwave region. Its functionality is to monitor the 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 (specifically the 0.7-1.5, 1.0-2.0, 2.6-3.8, 4.5-7.5, and 5.2-7.6 GHz wave bands). As SBRS monitors the solar radio bursts in daytime, it produces massive data about the solar radio information. However, the solar activity researchers are only interested in the data reflecting the burst activity of the Sun in the massive data. However, the data reflecting the Sun burst activity is very rare (1% of the captured data). Moreover, the data is always accompanied with the interference during the capturing process. As such, it is of heavy labor for human to identify whether the data contains burst information or not timely. To the end, analyzing the captured data automatically (burst or not) are highly demanded and beneficial to the solar radio astronomy study.

Nowadays, with the available massive data, especially visual data including images and videos, many algorithms have been developed to learn the representation with unsupervised and supervised methods for the tasks of visual, classification [18], localization [24], and so on. Recent progresses on deep learning [2] have demonstrated state-of-the-art performances in a wide variety of tasks, including visual recognition [3, 4], audio recognition [5, 6], and natu-

ral language processing [7], cross modality relationship [17, 15, 19], and so on. These techniques are super powerful because they are capable of learning useful features directly from both unlabeled and labeled data to avoid the need of hand-engineering. For solar radio spectrums, we also have massive data. Firstly, a large amount of the captured solar radio spectrums are unlabeled, most of which do not contain the burst information of Sun. Secondly, the professional experts who have the knowledge of solar physics are employed to label some of our captured data, which is time intensive and labor intensive. In this paper, we apply the deep learning, specifically the multimodal deep learning on the massive data of solar radio spectrums. The powerful ability of the deep network is expected to learn the inherent structural information of the solar radio spectrums for an effective and automatic classification of solar radio spectrums.

There are several kinds of disciplines for realizing unsupervised learning of deep neural networks for the massive data, such as Boltzmann machine, autoencoder (AE). AE is an unsupervised learning algorithm that applies backpropagation by setting the target values to be equal to the inputs. AE tries to learn a function which makes the input similar to the output of the function. In other words, it is trying to learn an approximation to the identity function, so as to output of the network that is similar to the input. The identity function seems a particularly trivial function to be trying to learn. But by placing constraints on the network, such as by limiting the number of hidden units, interesting structure about the data can be learnt. Therefore, AE is very helpful for representation learning of data, including visual data. The variances of AE, such as denoising AE [8], stacked AE (SAE) [9] were also developed widely. In [10], the authors proposed an automatic dimensionality reduction to facilitate the classification, visualization, communication, and storage of high-dimensional data through 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. Using 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 is also named as deep belief network (DBN). With the achievements of these learning methods, we can learn the representation of the solar radio spectrum even better than [11], which will be employed for further solar radio spectrum analysis, such as clustering, classification, and so on. However, Both AE and DBN, as well as their variants, treat the input signals equally. If the network takes different signals as the input, the characteristics between different input modalities cannot be distinguished. Thus the interaction and contributions between different modality inputs cannot be well exploited and captured.

For solar radio spectrums, the signals are captured from different frequency channels, which depict the Sun's activity from different perspectives. In this paper, we firstly employ the multimodal learning method, specifically the AE with the structured regularization, to learn the representation of the solar radio spectrum by distinguishing the contribution of each modality. By further stacking more fully-connected (FC) layers, the joint representation of the solar radio spectrums are generated, which is input to the softmax layer for classification. By evaluating the constructed multimodal network on the solar radio spectrum database, the experimental results demonstrate that the multimodal learning method can effectively analyze the solar radio spectrum.

The rest of the paper is organized as following. In Section 2, a multimodal learning architecture is introduced. In Section 3, a deep neural network based on the multimodal learning architecture is proposed to classify the solar radio spectrum. Section 4 gives the experimental results on representation learning and classification. And the final section concludes the paper.

2. Multimodal Learning Architecture

We propose a multimodal learning architecture for the purpose of solar radio burst classification, which is illustrated in Figure 1. The proposed multimodal learning architecture takes different numbers and types of modalities as the input and generates their joint representation for the targeted task, such as

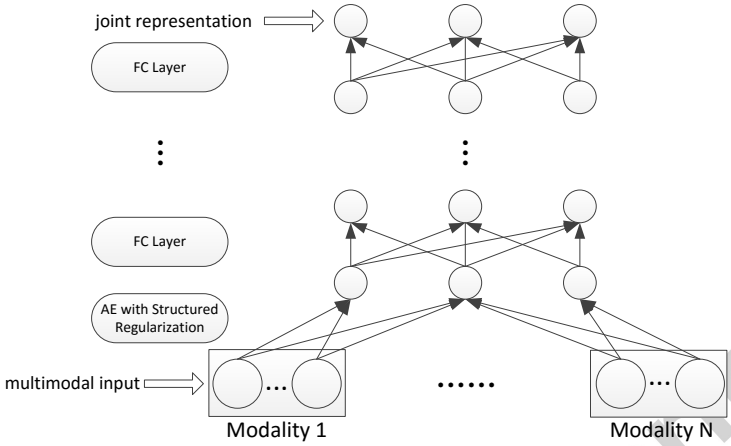


Figure 1: The framework of the multimodal learning architecture.

classification. The proposed multimodal learning model needs to adequately learn the representation of each individual modality. Most importantly, the inter-modality relationships and interactions need to be accurately captured to generate the joint representation. As illustrated in Figure 1, our proposed multimodal learning model relies on AE with the structured regularization to model each modality individually and jointly capture their interactions. Afterwards, several FC layers are stacked and employed to nonlinearly transform the intermediate representation to the final joint representation for the specifical target task. The benefits of the introduced FC layers are twofold. Firstly, the multiple FC layers with the nonlinear activation function will increase the nonlinearity of our proposed multimodal learning architecture, which will further make the final decision function (such as classification) more discriminative [18]. Secondly, the multiple modalities can interact more closely with each other through layers of FC nonlinear transformation. As such, the proposed multimodal learning model learns the multimodal abstractive representations from the detailed information contained in each modality.

We formulate the proposed multimodal learning architecture as:

$$\nu_{jr} = f_t^n\big(\cdots\big(f_t^1\big(f_{SR}(x_1, x_2, \cdots, x_m)\big)\big)\big) \tag{1}$$

where ν_{jr} is the final learned joint representation from the input with m different modalities $x_1, x_2, \cdots, x_m$. f_{SR} takes the input different multiple modalities and learns their intermediate representations. $f_t^1, \cdots, f_t^n$ are the following n FC layers, which are stacked together and transform the intermediate representation learned from f_{SR} to the final joint representation ν_{SR}. As illustrated in Figure 1, f_{SR} is realized by the AE together with the structured regularization in this paper. AE aims at transforming the input signal into output signal with the smallest distortions. AE treats each node of the input signal equally by performing the mapping process from the input to the output. As such, the different contributions of different modalities to the nodes of the output signal cannot be well learned and captured. However, different modalities may contribute differently to the specific task. In order to overcome this limitation and fully exploit the contributions and interactions between different modalities, the structured regularization is introduced to AE, which makes the proposed multimodal learning network distinguish different modalities with individual treatments for the intermediate representation. Moreover, our proposed multimodal learning network can be trained greedily layer by layer, as such stacking architecture ensures the scalability of the learning ability. On one hand, as aforementioned more nonlinear transformation layers can help improve the nonlinearity representation ability of the neural network, thus make the proposed network more discriminative. On the other hand, more parameters will be inevitably introduced. More parameters require more training data for adequately training and avoid overfitting. As such, the depth of our proposed multimodal learning architecture needs to be determined by the specific task as well as the number of available training samples.

2.1. Autoencoder (AE)

An AE with the simplest form is a feedforward neural net, presenting similarly to the multilayer perceptron (MLP), which consists of one input layer, one hidden layer, and one output layer. Compared with MLP, the main difference is that the node number of the output layer needs to be identical with that

of the input layer. AE is regarded to consist of two components, namely the encoder and decoder. The encoder encodes the input $x \in R^d$ to some hidden representation $y \in R^{d_h}$, while a decoder decodes the hidden representation y back to the reconstructed signal $\bar{x}$. AE is trained to make the reconstructed signal $\bar{x}$ to be as close as possible to the input signal x. The encoder process can be viewed as a mapping function f_e with nonlinear activation:

$$y = f_e(x) = \sigma(\omega x + b), \tag{2}$$

where x is the input signal of the encoder and y is the generated hidden representation given x as the input signal. ω and b are the weighting and bias parameters of the encoder function f_e, respectively. σ is an element-wise nonlinear activation function, which can employ sigmoid, tanh, and rectified linear unit (ReLU) [16] functions.

Afterwards, the generated hidden representation y from the encoder is mapped onto the reconstruction signal $\bar{x}$ with the same shape as x:

$$\bar{x} = f_d(y) = \sigma(\omega^T y + \bar{b}), \tag{3}$$

where f_d is the decoder function, which can be viewed as the inverse process of f_e. ω^T and $\bar{b}$ are the weighting and bias parameters of the decoder function f_d, where ω^T is obtained by transposing ω in Eq. (2). σ is the nonlinear activation function, same as the one in Eq. (2).

As aforementioned, AE is trained to make the reconstructed signal $\bar{x}$ as close as possible to the input signal x. The reconstruction error (with the squared error as defined) is minimized:

$$\begin{aligned} \mathscr{L}(\bar{x}, x) &= \| \bar{x} - x \|_2^2 = \| f_d(y) - x \|_2^2 = \| f_d(f_e(x)) - x \|_2^2 \\ &= \| \sigma(\omega^T(\sigma(\omega x + b)) + \bar{b}) - x \|_2^2 . \end{aligned} \tag{4}$$

Normally, the dimension R^{d_h} of the hidden representation y is smaller than the dimension R^d of the input signal x. As such, the encoder function $f_e(x)$ can be regarded as a compact and compressed representation of the input signal x. If R^{d_h} is larger than R^d, AE tends to learn the identical function, which may still learn some useful features [2].

2.2. Structured Regularization

In order to learn the representations and exploit different behaviors of different modalities, we employ the structured regularization (SR) to regularize the connections of AE between the hidden nodes and the multimodal input nodes. The connections between the input and hidden nodes as well as the corresponding weights are learnt with a data-driven manner, which are expected to distinguish and learn the representation from different multimodal inputs to generate the final joint representation.

SR for handling the multimodal inputs is inspired by [12, 13]. Suppose M as a $P \times Q$ binary matrix, where P indicates the total number of modalities and Q denotes the total number of the input units. The element $M_{k,i}$ indicates the membership of the input unit x_i in the specific modality k. If the input unit x_i belongs to the modality k, $M_{k,i}$ is 1 and 0 otherwise. As such, for each node in the hidden layer, each modality will be treated as a regularization group separately. Such process presents similar behavior with the group regularization, which treats each input modality differently and thus learns the complicated relationship and correlation between different modalities. And the weights $\omega_{i,j}$ of our constructed multimodal network are real valued. SR is defined as:

$$SR(\omega) = \sum_{j=1}^{N} \sum_{k=1}^{P} \left(\sum_{i=1}^{Q} \left(M_{k,i} \mid (\omega_{i,j})^{\lambda} \mid \right) \right)^{\frac{1}{\lambda}}, \tag{5}$$

where N denotes the total number of the hidden nodes. $SR(\omega)$ regularizes on the weighting parameters ω by summing the the Minkowski distance of ω. Such regularization term penalizes the the summation value of the weights ω in the form of Minkowski distance. With $\lambda \to \infty$, the regularization term $SR(\omega)$ is reformulated as:

$$SR(\omega) = \sum_{j=1}^{N} \sum_{k=1}^{P} \left(\max_{i} (M_{k,i} \mid \omega_{i,j} \mid) \right), \tag{6}$$

which directly penalizes the maximum weight value from each input node to the hidden node. Furthermore, in order to prevent over-constraining, we further modified the regularization term by penalizing the nonzero weight maxima

from each modality to each hidden node without additional penalty of the larger values of these maxima. The can be further viewed as group sparse regularization. The regularization term in Eq. (6) is further expressed as:

$$SR(\omega) = \sum_{j=1}^{N} \sum_{k=1}^{P} \left(f_B(\max_i(M_{k,i} \mid \omega_{i,j} \mid) > 0) \right), \tag{7}$$

where $f_B(\cdot)$ indicates a Boolean function that takes a value of 1 if its variable is true, and 0 otherwise. It can be observed that $SR(\omega)$ performs the direct penalty on the number of modalities connected to each hidden node.

It can be observed that each modality in our multimodal network is treated as a regularization group separately. And such process presents similar behavior with the group regularization, compared with the fully dense and modality-specific models. The fully dense model simply concatenates the multimodal inputs as a vector and treats each modality equally. The modality-specific model [12] assumes that the ideal low-level features for each modality are purely unimodal, while higher layer features are purely multimodal. Compared with the fully dense and modality-specific models, our proposed multimodal network not only learns correlated features between multiple input modalities, but also regularizes the number of modalities used for each hidden unit and thus discourages learning weak correlations between different modalities. With SR, the multimodal network can enforce and learn the modality-specific sparsity as well as the density of each modality.

2.3. Integrating AE with Structured Regularization

By integrating AE with the structured regularization, we can ensure the group sparsity on the multimodal inputs, which directly regularizes the number of the modalities connected to each hidden node as shown in Figure 2. The reconstruction error as defined in Eq. (4) is further modified by integrating the regularization term on the weighting parameters ω. As such, the objective function for training the multimodal network as shown in Figure 1 is formulated

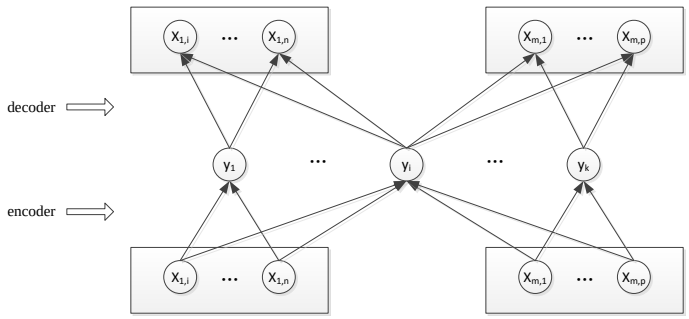


Figure 2: AE with structured regularization.

as:

$$\begin{aligned}\omega^* &= \mathscr{L}(\bar{x}, x) + \alpha \cdot SR(\omega) \\ &= \arg\min_{\omega} \parallel \bar{x} - x \parallel_2^2 + \alpha \cdot SR(\omega).\end{aligned} \tag{8}$$

where $\bar{x}$ is the signal reconstructed by the decoder of AE by Eq. (3). α is the parameter to balance the error and the regularization terms. ω^* is the learned parameters for the AE with the structured regularization.

By integrating SR into AE, the obtained representation y only connects to partial nodes in the hidden layer. As shown in Eq. (7), in order to minimize $SR(\omega)$, the zero number of ω should be as large as possible, leading to some nodes in the hidden layer connected to only part of the nodes in the input layer. As such, the contributions of each modality to each hidden node can thus be optimized with the constraints on the connections. AE with structured regularization demonstrates that the multimodal network could distinguish different modalities and learn the correlations between them automatically. Furthermore, the effective weight parameters of the paramter ω can be greatly reduced compared with the fully connected AE. It can further help prevent overfitting of the multimodal network with a limited number of training samples.

3. Multimodal Learning for Classification of Solar Radio Spectrum

In this section, we first introduce how to pre-process the solar radio spectrum in order to be fed into our proposed multimodal learning network as introduced

in Section. 2. Afterwards, the framework of the proposed network for the solar radio spectrum are introduced in details.

3.1. Pre-processing of Solar Radio Spectrum

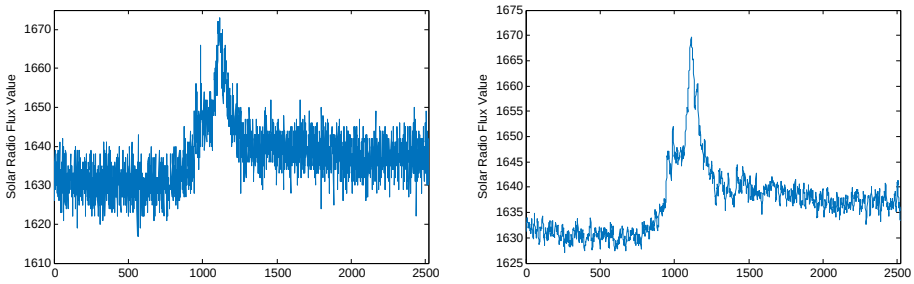
As aforementioned, the solar radio spectrum is captured by SBRS of China [1], which is a solar dedicated radio spectrometer for capturing solar radio strength along time over multiple frequency channels in the microwave domain. Each channel is responsible for the designed frequency range to capture the solar radio strength. Compared with the entire solar radio spectrum, the radio strength variation of each channel can present more detailed characteristics of the Sun's activities. Therefore, in this paper, the captured solar radio signals from different channels are regarded as different modalities to be fed into our proposed multimodal learning architectures, which not only learn the representation of each channel but also capture the interactions and relations between different channels for the joint representation.

3.1.1. Solar Radio Spectrum

The solar radio signal sensed from each channel is treated individually. In total, there are 120 channels working toward capturing the solar radio information at the same time. Moreover, each captured file contains both the left and right circular polarization parts, which should be separated and processed individually for further processing. We extract the captured data from each channel as a row vector, which is organized according to its sensing time. Afterwards, all row vectors from the 120 channels are assembled together to form an image, which can be further processed for visualization and processing. As there are 120 channels and 2520 sensing time points in 8 ms recorded file, the resolution of the generated image is 120×2520.

3.1.2. Channel Denoising and Normalization

For SBRS of China sensing the solar radio signal with radio antenna, the noise is thus inevitably introduced, which presents to be of strong white noise presenting dramatic fluctuations as shown in Figure 3 (a). The noise will be



(a) (b)

Figure 3: The solar radio signal before (a) and after (b) Gaussian filtering.

very annoying and seriously affect the following analyze and process of the radio spectrums. As such, in order to make the radio signals more expressive and representative, we employ the Gaussian filter to suppress the noise:

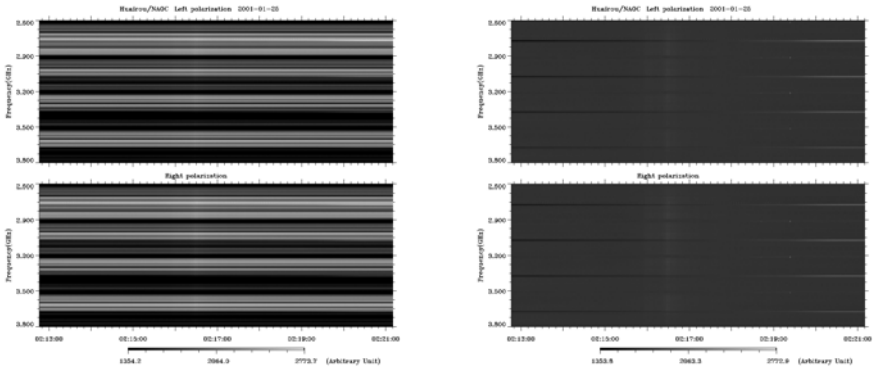
$$S_i = G \otimes S_i, \tag{9}$$

where S_i denotes the radio signal captured in each channel. $\otimes$ denotes the convolution process. G is the Gaussian kernel defined as:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}. \tag{10}$$

σ is the standard deviation of the Gaussian distribution, which is set as 5 in this paper. As the noise gain for sensing the solar radio signal is different for each channel, the one-dimension convolution process is performed on each channel individually. The denoised radio spectrum after the Gaussian filtering is illustrated in Figure 3 (b). It can be observed that the dramatic fluctuations has been alleviated while the global shape of the solar radio flux signals has been well preserved, which will be helpful for further processes.

It can be observed that there are horizontal-stripes-like interference signals in Figure 4 (a). It is named as the channel effect in solar radio observation, which is caused by different gains of different channels. The channel effect may disturb the presentation of bursts. In order to eliminate such channel effect, we propose one method for channel normalization, which is formulated as following:



(a) (b)

Figure 4: The solar radio spectrum before (a) and after (b) channel normalization.

$$\bar{S} = S - S_{LM} + S_{GM} \tag{11}$$

where S the whole solar radio spectrum after denoising, $\bar{S}$ is the obtained solar radio spectrum after performing the channel normalization, S_{LM} and S_{GM} denote the local mean and global mean values of the solar radio spectrum, respectively. The local mean S_{LM} is calculated by the mean of each channel. S_{GM} accounts for the mean of whole radio spectrum. As shown in Eq. (11), S_{LM} is to alleviate the effect of uneven channel gain, while S_{GM} compensates each channel by adding a global background. The solar radio after performing the channel normalization is illustrated in Figure 4 (b). Compared with Figure 4 (a), the horizontal-stripes-like interference signals are alleviated in Figure 4 (b). The solar radio signal variations along the time can be more clearly detected after normalization, which can help learning the final joint representation for classification.

3.1.3. Down-sampling via Channel Competition

After the channel denoising and normalization, a noise free and channel effect free solar radio spectrum is obtained, which can be fed into the proposed multimodal learning network for further classification task. However, the reso-

lution of the entire solar radio spectrum is very large. In order to reduce the number of the input nodes as well as the parameters of the multimodal learning network, we proposed one channel competition method to further down-sample the entire solar radio spectrum.

As discussed in [1], the radio emission will be greatly enhanced when a solar burst occurs, like a solar flare or coronal mass ejection (CME), which results from a local release of energy in the Sun's low corona. Such process would produce numerous radio spectral structures observed with radio spectrometer. As solar radio bursts occur, the flux values of solar radio spectrum will increase in certain channels. In order to more accurately depict such behavior and property of each channel, we defined an activity term to discriminate the difference between channels as:

$$D_i = \max(\bar{S}_i) - mean(\bar{S}_i) \tag{12}$$

where D_i represents the maximum flux value of the i-th channel $\bar{S}_i$ minus the mean of the corresponding channel. D_i is assumed to reflect the flux activity of each channel in the solar radio spectrum. As such, the inner structural information of each channel is exploited for the entire solar radio spectrum .

By ranking D_i with descending order, we select the top k channels from the total 120 channels as the most representative ones of the entire solar radio spectrum. In this paper, k is set as 10. After selecting the top k channels, each channel is further down-sampled with the bicubic filter. The original spectrum with the resolution as 120×2520 is finally down-sampled as 10×200. As such, the dimension of input data has been greatly reduced. Moreover, with the channel competition scheme, the most representative channels are kept in the final down-sampled version.

3.2. Multimodal Network for Solar Radio Spectrum Classification

In order to further perform the classification of the solar radio spectrum, we design one multimodal network as shown in Figure 5. The input of the

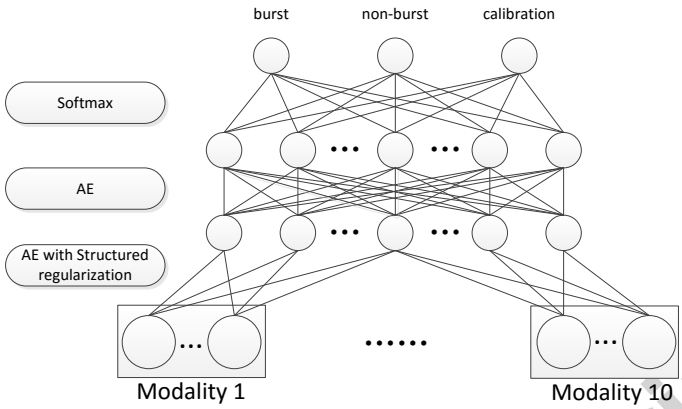


Figure 5: The framework of multimodal network for solar radio spectrum classification.

proposed multimodal network is the pre-processed solar radio spectrum. As introduced in Section 3.1, the raw captured solar radio spectrum is processed with denoising, normalization, and downsampling. The final obtained spectrum is of 10×200. As each channel senses the solar radio flux at each specific channel, we treat these 10 channel radio spectrum as 10 different modalities as the input to our proposed multimodal network. As such, each modality contains 200 input nodes representing the radio spectrum in each channel. AE with SR takes the 10 modalities as input and generate the intermediate representation with 200 nodes, which is expected to capture both the intra channel and inter channel relations and interactions. As discussed in Section 2.3, each node of the first hidden layer is obtained by regularizing the number of modalities connected to each hidden node. As such, the multimodal network could distinguish different modalities and learn the correlations between them automatically. On top of the intermediate representation, AE as the FC layer performs an additional nonlinear mapping to generate the final joint representation. AE is employed here to increase nonlinearity of the proposed multimodal network, which could make the decision function more discriminative [18]. Based on the joint representation, the softmax is used to perform the corresponding classification.

The "I-H-H-O" network structure is shown in Figure 5. "I" indicates the

input data from all the multimodal inputs, the total dimension of which is $10 \times 200 = 2000$. "H" denotes the hidden nodes of the two FC layers, which consists of 200 nodes for the first hidden layer and 100 nodes for the second layer. "O" is output nodes for the classification, which give the probabilities of each input sample belonging to the pre-defined classes. In this paper, 3 main observations about the solar radio spectrum, namely "burst", "non-burst", and "calibration", are employed as the 3 classes in the classification layer.

The object of the proposed multimodal learning network is to realize the non-linear mapping relations between the input solar radio spectrum and the pre-defined labels. The inference can be realized by the following function:

$$\bar{L} = \arg\max_{L} p(L|S;\Theta), \tag{13}$$

where S is set of multimodal input of $S_1, \ldots, S_{10}$, with S_i is the captured radio signal at the i-th channel. Θ is the parameters, consisting of the weight parameters of the first layer AE with structured regularization, the second layer of AE, and the softmax layer.

In order to make reliable classification of the solar radio spectrum, the parameters of the constructed multimodal network, specifically the parameters of the three different layers, needs to be learned. For the parameters of AE with structured regularization, we obtain the initialized parameters with pre-training of AE according to Eq. (8). Specifically, the pre-training process simply learns features from unlabeled data automatically aiming to transform the input spectrums into outputs with the least amount of distortion. With such pre-training process, the constructed network can effectively avoid the risk of trapping in poor local optima. Afterwards, the fine-tuning process is further performed for the classification of solar radio spectrum. A log-likelihood function with the structured regularization as the constraint is employed as the object function to train the whole multimodal network:

$$\bar{\Theta} = \arg\max_{\Theta} \sum log(p(\bar{L} = L|S;\Theta)) - \alpha SR(\omega_{SR}), \tag{14}$$

where L is the true label of the input radio spectrum, $\bar{L}$ is the output of the

network. Backprorogation (BP) [23] is employed to fine-tune parameters of the constructed network. BP is proposed 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 [12]. Simultaneously, it can improve generalization performance of the learning system. Furthermore, in order to prevent over-fitting in training neural network, dropout 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. Dropout is applied on all the layers and the probability is set as $p = 0.2$.

Observing the input layer, the 10 channel radio spectrums are regarded as the multimodal inputs. In general applications [19] [20] [21] [22], each modality may has the different form of the data, e.g., audio, image, text and so on, which represent the similar semantic meanings. For our application on solar radio spectrum classification, each frequency channel captures the information of the solar burst from one specific perspective. These different frequency channels are treated as different modalities, whose interactions and relations can be further learned with the proposed multimodal network. There are three possible models for multimodal learning. One naive and straightforward way of applying feature learning to multimodal data is to simply take the whole data vector as the input to the model. This approach name as fully dense model, may fail to learn associations between modalities with very different underlying statistics. Additionally, it learns features prematurely, which can easily tend to be overfitting. Instead of the fully dense model, modality-specific sparse model trains a first layer representation for each modality separately. This approach assumes that the ideal low-level features for each modality are purely unimodal without any relations and interactions, while the higher-layer features are purely multimodal. This approach may work better for some problems where the modalities have very different basic representations, such as the video and audio data as

Table 1: The number of solar radio bursts observed with each component spectrometer of SBRS by the end of 2001.

Frequency Range	0.5-1.5	1.0-2.0	2.6-3.8	4.5-7.5	5.2-7.6
Number of Bursts	108	526	921	233	550

shown in [19]. However, in our application, frequency channels of spectrums are treated as the modalities. These modalities have strong correlations and similar behaviors between each other, which means that the learning of low-level correlations and interactions may lead to better features. Therefore, the proposed AE with the structured regularization method is employed ,which can be viewed as the group sparse model by constraints on the connection between the hidden nodes to the modalities.

4. Experimental Results

In order to evaluate the proposed method, the experiments are performed on a solar radio spectrum database. First, we will briefly introduce the built database of solar radio spectrum. Afterwards, we will demonstrate and discuss the classification results of solar radio spectrums on the database with the comparisons with other approaches.

4.1. Solar Radio Spectrum Database

As mentioned before, the SBRS of China [1] is designed to acquire dynamic spectrograms of solar radio bursts with the combination of wide frequency coverage from 0.7GHz to 7.6 GHz. It is of the high temporal resolution, high spectral resolution, and high sensitivity. It consists of five "component spectrometers" operating at five different wavelength bands. All the five "component spectrometers" work simultaneously to make a full observation of the solar radio bursts from the perspective of sensing frequencies. More detailed information about SBRS can be referred to [1].

In total, SBRS captured about millions of solar radio spectrums during the period from 1995 to 2001. However, there are only a small portion of solar radio bursts in these captured data which are meaningful for solar physics research, as the experts are mostly interested in the burst activities of the Sun. By the end of 2001, about 2000 burst solar radio spectrums are observed. Compared with millions of captured radio spectrums, the burst solar radio spectrums are very rare. Therefore, it brings the experts a huge labor to distinguish the burst radio spectrums from the non-burst ones. That is also the reason why we resort to multimodal learning architecture to automatically perform the solar radio spectrum classification. Detailed information about the burst radio spectrum at each frequency range can be found in Table 1. It can be observed that the burst behaviors captured in the 2.6-3.8 GHz frequency range are more easily detected by the human experts. It means that the spectrums captured in this frequency range can be represented as one image such as Figure 4 for the experts to easily indicate whether the solar radio bursts exists or not. As such, we select the most representative solar radio spectrums in this frequency range to construct the database of solar radio spectrum.

We select 4408 captured data files of 2.6-3.8 GHz frequency band. After performing the processes as introduced in Section 3.1.1, each data file provides two spectrums with the size of 120×2520, i.e., the left and right spectrums. By treating the generated left and right spectrums separately, we obtain 8816 solar radio spectrums in total. Each row of the spectrum denotes the frequency for capturing the solar radio wave, while the column indicates the sensing time of the solar radio wave. These generated spectrums are labeled by the professional experts with five classes (0: no burst or hard to identify the burst activity, 1: weak burst, 2: moderate burst, 3: large burst, 4: calibration). The calibration signals are generated by the antenna to align different frequency channels and make sure that the signal captured by the solar radio telescopes is effective. For most cases of calibration signals, the flux values vary non-continuously within each channel. Detailed information about the labeled spectrums in the dataset is illustrated in Table 2. As mentioned before, the solar researchers are only

Table 2: The number of solar radio spectrums with different activities in the database. (0 indicates non-burst or hard to identify; 1 indicates weak burst; 2 denotes moderate burst; 3 denotes strong burst; 4 indicates calibration.)

Burst Strength	0	1	2	3	4	total
Spectrum Number	6670	618	268	272	988	8816

interested in the burst spectrums for further study of the Sun's activity. Therefore, it is significantly meaningful to distinguish the burst spectrums from the other ones. As such, in this paper we focus on the classification of the three coarse categories, specifically the "burst", "non-burst", and "calibrations". The spectrums of the "burst" category denotes the spectrums containing weak, moderate, and strong burst information. Finally, the total solar radio spectrums labeled as "bursts", "non-burst", and "calibrations" are 6670, 2146, and 988, respectively.

4.2. Performance Comparisons

As discussed before, the researchers in solar activity are mostly interested in the burst radio spectrums other than non-burst and calibration spectrums. As such, we examine the classification ability of the proposed multimodal network on the solar radio spectrums of different behaviors. As such, we employ the true positive rate (TPR) and false positive rate (FPR) from the binary classification to evaluate the corresponding performances. TPR measures the proportion of positives that are correctly identified as such. FPR, on the other hand, measures the proportion of negatives that are misclassified as positive.

As listed in Table 1, there are 8816 solar radio spectrums in total. We randomly select 900 "burst", 800 "non-burst", 800 "calibration" from the dataset for training the proposed multimodal network as well as the competitor models. And the rest solar radio samples are employed for testing. The trained model achieves good performances when the category with highest possibility output by the model matches the labeled category of the input spectrum. To ensure that the proposed model is robust across the content of the solar radio spectrum

Table 3: Performance comparisons between the proposed multimodal model, DBN, and PCA+SVM.

	Mutimodal		DBN		PCA+SVM	
	TPR	FPR	TPR	FPR	TPR	FPR
Burst	82.2%	22.5%	67.4%	13.2%	52.7%	26.6%
Non-Burst	83.3%	9.6%	86.4%	14.1%	0.1%	16.6%
Calibration	92.5%	1.7%	95.7%	0.4%	38.3%	72.2%

and is not biased by the specific train-test split, random processes with the same splitting percentage is repeated 20 times.

We compare our proposed mutlimodal network model with DBN as well as the PCA+SVM approach. PCA+SVM employs the PCA to perform the dimension reduction on the solar radio spectrum and the SVM to classify the processed solar radio spectrum. DBN takes the raw data of the solar radio spectrum as the input to perform the classification. Due to the constraint of the training data number, only one hidden layer is used in DBN.

The average TPR and FPR of proposed network and competitor models are reported in Table 3. It can be observed that the proposed multimodal network is better than DBN with respect to the classification accuracy of "burst". More specifically, the TPR for the burst solar radio is over 82% for the proposed multimodal network, which is higher than DBN and PCA+SVM. The performance gain can be attributed to that the proposed multimodal network not only well represent each modality (solar radio signals from each sensing channel) but also exploit the differences between different modalities (solar radio signals from different sensing channels) and learn their relations and interactions to generate the final joint representation for the final classification. However, such performance is not as good as that for general image classification, as the Sun's activity in solar radio spectrum is very hard to detect even for the solar activity researcher. In some cases, the solar activity researchers cannot identify "burst" for "non-burst" for some spectrums with strong noise and weak activ-

ity. Also it can be observed that PCA+SVM perform the worst. The features from the sola radio spectrum is extracted by PCA. With SVM as the classifier, the complicated relations and behaviors of the solar radio spectrum cannot be well exploited. That is the main reason why PCA+SVM perform the worst. Compared PCA+SVM, DBN ensemble the feature learning and classification together as an end-to-end learning strategy. As such, the complicated behaviors within the solar radio spectrum can be well discovered, which produces a better performance compared with PCA+SVM. However, DBN treats the captured signals from different frequency channels equally, which cannot distinguish the differences and contributions between them. Moreover, the correlations and interactions between the signals of different modalities cannot be well captured. That is the main reason why DBN performs inferiorly to our proposed multimodal network. For "non-burst", the proposed multimodal network presents a slight worse performance. However, for solar radio researches, we mostly focus on discovering the "burst" information from the massive data. Therefore, the degradation for "non-burst" is acceptable, compared with improvement on the "burst" solar radio spectrum classification. Moreover, the performance on "calibration" is much better than those on "burst" and "non-burst" for our proposed multimodal network and DBN. Such "calibration" spectrums present very simple feature pattern, which can be easily learnt with an end-to-end learning approach.

4.2.1. The Number of Hidden Layers

In this subsection, we examine the effect of different numbers of hidden layers for our proposed multimodal network. We compared different multimodal with different hidden layers. Specifically, three multimodal networks are of one hidden layer (I(2000)-H(200)-O(3)), two hidden layers (I(2000)-H(200)-H(100)-O(3)) and three hidden layers (I(2000)-H(200)-H(200)-H(50)-O(3)), respectively, where the node number of each layer is also illustrated. As shown in Table 4, it can be observed that the proposed multimodal network with two hidden layers yields the best performances. The first layer employs AE with

Table 4: Performance comparisons between the multimodal learning networks with different hidden layers.

	I(2000)-H(200) -H(100)-O(3)		I(2000)-H(200) -O(3)		I(2000)-H(200) -H(200)-H(50)-O(3)	
	TPR	FPR	TPR	FPR	TPR	FPR
Burst	82.2%	22.5%	72.1%	15.8%	73.6%	17.4%
Non-Burst	83.3%	9.6%	82.4%	15.7%	79.5%	14.4%
Calibration	92.5%	1.7%	93.6%	0.02%	94.2%	0.03%

the structured regularization to distinguish the differences and contributions between the captured signals from different channels. Compared with the network with only one hidden layer, the multimodal network with two hidden layers introduce another FC layer to nonlinearly map the intermediate representation to the final joint representation, which can further increase the nonlinearity of the system and make the decision function more discriminative [18]. However, by stacking an additional hidden layer, the performance of the network with three hidden layers is inferior to the one with two hidden layers. Also, with more hidden layers, the nonlinearity of the system can be enhanced and the discriminative of the decision function is ensured. However, with more hidden layers, the FC layer is inevitable introduce a even larger number of parameters for tuning the whole network. Thus, it will be more prune to overfitting. That is also the main reason why the network with three hidden layers performs inferiorly to that with two hidden layers.

4.2.2. The Number of Hidden Nodes

In this section, we examine the effect of the hidden node number on the proposed multimodal network with two hidden layers. The two layer node numbers are set as 220-100, 200-50, 1000-100, respectively. The experimental results are illustrated in Table 5. With the same node number of the first hidden layer, the network with the second layer node as 100 outperforms the one with the

Table 5: Performance comparisons between the multimodal learning networks with different hidden node numbers.

	I(2000)-H(200) -H(100)-O(3)		I(2000)-H(200) -H(50)-O(3)		I(2000)-H(1000) -H(100)-O(3)	
	TPR	FPR	TPR	FPR	TPR	FPR
Burst	82.2%	22.5%	77.9%	19.7%	69.8%	14.9%
Non-Burst	83.3%	9.6%	77.3%	12.1%	83.3%	16.4%
Calibration	92.5%	1.7%	93.6%	0.03%	92.0%	0.02%

seconde layer node as 50. Therefore, the second FC layer with larger number of nodes can map the learned intermediate representation to the joint representation for better classification, which can help to more accurately model the complicated relationships and behaviours with the solar radio spectrum. With the fix number of second layer node, the network with the first layer node as 1000 performs inferiorly to the one with the seconde layer node as 200. Also the layer with 1000 nodes can more accurately learn the relationships between different modalities with a larger space for the intermediate representation. However, the larger number of nodes will also introduce more parameters for the constructed multimodal network, which is also more prune to overfitting.

5. Conclusion

In this paper, we proposed a novel multimodal learning network for the solar radio spectrum classification. By integrating autoencoder with structured regularization, the proposed multimodal learning network regularizes the number of modalities connected to each hidden node, which can simultaneously learn the density of each modality and enforce the modality-specific sparsity. As such, the multimodal network could distinguish different modalities and learn the interactions and correlations between them automatically. The experimental results demonstrated the superiority of the proposed multimodal network on the solar radio spectrum classification task.

6. Acknowledgements

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).

References

[1] Q. Fu, H. Ji, Z. Qin, Z. Xu, Z. Xia, H. Wu, Y. Liu, Y. Yan, G. Huang, Z. Chen, Z. Jin, Q. Yao, C. Cheng, F. Xu, M. Wang, L. Pei, S. Chen, G. Yang, C. Tan, and S. Shi "A new solar broadband radio spectrometer (SBRS) in China," solar Physics, 2004.

[2] Y. Bengio, "Learnin deep architectures for AI," Foundations and Trends in Machine Learning, 2009.

[3] Q. V. Le, M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng "Building high-level features using large scale unsupervised learning," International Conference on Machine Learning, 2012.

[4] K. Sohn, D. Y. Jung, H. Lee, and A. O. Hero III, "Efficient learning of sparse, distributed, convolutional feature representations for object recognition," International Conference on Computer Vision, 2011.

[5] H. Lee, and Y. Largman, P. Pham, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," Annual Conference on Neural Information Processing Systems, 2011.

[6] A. Mohamed, G. E. Dahl, and G. Hinton, "Unsupervised feature learning for audio classification using convolutional deep belief networks," IEEE Transactions on Audio, Speech, and Language Proessing,2012.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa "Natural language processing (almost) from scratch", The Journal of Machine Learning Research, 2011.

[8] M. Chen, K. Weinberger, F. Sha, and Y. Bengio "Marginalized denoising auto-encoders for nonlinear representations," International Conference on Machine Learning, 2014.

[9] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized Stacked Denoising Autoencoders for Domain Adaptation," Internaltional Conference on Maching Learning, 2012.

[10] G. Hinton, R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, 2006.

[11] Z. Chen, L. Ma, L. Xu, C. Tan, and Y. Yan, "Imaging and Representation Learning of Solar Radio Spectrums for Classification," Multimedia Tools and Applications, 2015.

[12] L. Ian, H. Lee, and A. Saxena, "Deep Learning for Detecting Robotic Grasps," International Journal on Robotics Research, 2015.

[13] A. Jalali, P. Ravikumar, S. Aanghavi, and C. Ruan, "A dirty model for multi-task learning," Advances in Neural Information Processing Systems, 2010.

[14] T. Fawcett, "An introduction to ROC analysis", Pattern Recognition Letters, 2006.

[15] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," The 30th AAAI Conference on Artificial Intelligence, 2016.

[16] G. E. Dahl, T. N. Sainath, and G. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[17] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal Convolutional Neural Networks for Matching Image and Sentence," International Conference on Computer Vision, 2015.

[18] K. Simonyan, A. Zisserman "Very Deep Convolutional Networks for Large-Scale Image Recognition," International Conference on Learning Representation, 2015.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," International Conference on Machine Learning, 2011.

[20] N. Srivastava, R. Salakhutdinov, "Multimodal Learning with Deep Boltzmann machines,", Neural Information Processing Systems, 2012.

[21] R. Kiros, R. Salakhutdinov, and R. Zemel, "Unifying Visual-Semantic Embedding with Multimodal Neural Language Models," Transactions of the Association for Computational Linguistics, 2015.

[22] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal Neural Language Models," International Conference on Machine Learning, 2014.

[23] L. Deng, "Three classes of deep learning architectures and their applications: a tutorial survey," APSIPA Transactions on Signal and Information Processing, 2012

[24] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C. Loy, and X. Tang, "DeepID-net: Deformable deep convolutional neural netowkrs for object detection," IEEE Conference on Computer Vision and Pattern Recognition, 2015.