Pattern Recognition 100 (2019) 107085

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Semi-supervised cross-modal image generation with generative adversarial networks

Dan Li[a,b], Changde Du[a,b], Huiguang He[a,b,c,*]

[a] Research Center for Brain-inspired Intelligence and National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China
[b] University of Chinese Academy of Sciences, Beijing 100190, China
[c] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 5 May 2019
Revised 8 September 2019
Accepted 15 October 2019
Available online 12 November 2019

Keywords:
Multi-modality
Semi-supervised learning
Semantic networks
Generative adversarial networks
Multi-label learning

ABSTRACT

Cross-modal image generation is an important aspect of the multi-modal learning. Existing methods usually use the semantic feature to reduce the modality gap. Although these methods have achieved notable progress, there are still some limitations: (1) they usually use single modality information to learn the semantic feature; (2) they require the training data to be paired. To overcome these problems, we propose a novel semi-supervised cross-modal image generation method, which consists of two semantic networks and one image generation network. Specifically, in the semantic networks, we use image modality to assist non-image modality for semantic feature learning by using a deep mutual learning strategy. In the image generation network, we introduce an additional discriminator to reduce the image reconstruction loss. By leveraging large amounts of unpaired data, our method can be trained in a semi-supervised manner. Extensive experiments demonstrate the effectiveness of the proposed method.

1. Introduction

With the development of sensor technology, we can obtain multiple modalities dataset using different acquisition equipment. For example, natural image and its text description, visual stimulus and its evoked brain signal (as is shown in Fig. 1). Different modalities of the dataset can provide complementary information. In recent years, there are more and more applications of multimodal learning [1–4]. Current multimodal learning mainly focuses on multimodal fusion task [5–8]. There are also some researchers have explored the cross-modal prediction task, i.e., predict one modality from another. Cross-modal prediction includes cross-modal retrieval [9–11], image caption [12–14], cross-modal image generation [15–17], etc.

In this paper, we focus on cross-modal image generation: generating images from non-image modalities. As is known, one of the challenges in the cross-modal image generation task is how to build a bridge between two modalities to reduce the modality gap. In other words, how to learn semantic feature from non-image modality is very important. To address this problem, some cross-modal image generation methods have been proposed recently [16,17]. These methods learn the semantic feature of non-image modality before image generation. Kavasidis et al. used LSTM network [18] to learn a compact and noise-free representation of Electroencephalograph (EEG) data [17]. Tirupattur et al. proposed a deep neural network architecture in which a 1D convolutional neural network (CNN) is followed by another 2D CNN to encode the EEG signals [16]. In the feature learning process, they used the semantic labels as supervision information [16,17].

Although the above methods have achieved notable progress, there are still some limitations. First of all, in the process of semantic feature learning, they only use single modality information. Most of them treated the image modality as the generated target, but they did not consider using the image information to assist non-image modality in semantic feature learning. In fact, image modality has obvious feature information which can be used to supervise the non-image modality feature learning. Furthermore, different modalities may provide complementary information which can improve the model performance in the prediction task [6]. Scott et al. pre-train a deep convolutional recurrent network on the structured joint embedding of text captions with 1024-dimensional GoogLeNet [19] image embedding to obtain a semantic text feature for image generation [15]. Nevertheless, this method requires training data to be paired.

Aforementioned methods are all under an assumption that the training data are paired. In reality, the number of the training pairs

* Corresponding author.
E-mail addresses: lidan2017@ia.ac.cn (D. Li), duchangde2016@ia.ac.cn (C. Du), huiguang.he@ia.ac.cn (H. He).

https://doi.org/10.1016/j.patcog.2019.107085
0031-3203/

(a) Image and tag (b) Image and brain signal

Fig. 1. Multi-modality data.

is always limited as non-image modality is not easy to obtain, for example, the brain signals need to take a lot of time to collect or the text may require the efforts of experienced human annotators. The image modality is usually sufficient, models should have the ability to leverage large amounts of unpaired image data. In multimodal learning, unpaired data can be viewed as suffering missing modality problem. A series of methods are proposed [20–22] to solve this problem, and these methods need to generate the missing modality for unpaired data. For some cross-modal image generation tasks, there is no need to perform modal completion on unpaired data. In fact, unpaired data itself contains a lot of information, and it is important to be able to use the information directly.

The image generation is a very natural application for generative adversarial networks [23]. In the process of adversarial image generation, some models often used the Mean Squared Error (MSE) loss as the image reconstruction loss [24,25]. However, the MSE loss makes the adversarial learning to be unstable and the generated images to be blurry [20,26].

In this paper, we propose a novel semi-supervised cross-modal image generation model, which consists of two semantic networks and one image generation network. In the semantic networks which contain image network and non-image network, we use a deep mutual learning strategy for semantic feature learning. In the image generation network which is conditioned on semantic features and labels that encoded by non-image modality, we train a deep convolutional conditional generative adversarial network for image generation. The image generation network consists of one generator and two discriminators. For the two discriminators, one is used to ensure that the distribution of generated images is close to that of real images, and the other is used to reduce the image reconstruction loss. Meanwhile, by leveraging large amounts of unpaired data, our method can be trained in a semi-supervised manner. Extensive experiments on three datasets including single-label (fMRI-HC, EEG-ImageNet) and multi-label (MIRFLICKR-25K) demonstrate that the proposed method outperforms state-of-the-art approaches. It also shows that our method can be applied to various fields, such as neuroscience field and computer vision field.

The main contributions of our method are summarized as follows:

- By leveraging a deep mutual learning strategy in the training process, our method can use the image modality information to assist the semantic feature learning of non-image modality.
- Our method does not require all training data to be paired. It can utilize a large amount of unpaired data during the training process, which achieves semi-supervised characteristics.
- For the image generation network, we add an additional discriminator which is able to reduce the image reconstruction loss. Comparing with the MSE loss, the adversarial learning is more stable in this way.

The rest of this paper is organized as follows: Section 2 introduces the most recent related work. We present our method in Section 3, including the model formulation, semi-supervised learning and optimizing strategy. Experiments are shown in Section 4. Finally, we conclude in Section 5.

2. Related work

The proposed method contains two parts: semantic feature learning and image generation. We use the image modality to assist non-image modality in semantic feature learning, which can be viewed as the process of multimodal fusion. And the image generation is an application of generative adversarial networks. In this section, we introduce recent related work from two aspects: multimodal fusion and generative adversarial networks.

Multimodal fusion is the concept of integrating information from multiple modalities to improve the performance of some task [27–29]. Ngiam et al. proposed a deep auto-encoder model to learn the shared representation between the video and audio for audio-visual speech classification [7]. Srivastava et al. used a deep Boltzmann machine [30] to extract a unified representation that fuses modalities together for classification and information retrieval [8]. Unlike the above methods, which learn one common representation from multiple modalities and ignoring the dependencies between target labels and multiple features, Chen et al. proposed a supervised multimodal deep learning model that used a joint deep network to learn one representation from each modality. The dependencies among multiple modalities are captured by both shared feature layer and label layer of the network [6].

Generative Adversarial Networks (GANs) has emerged as a powerful method for generating images when given a noise [23,31,32]. The basic strategy of GANs is to train a generative model and a discriminative model simultaneously via the adversarial process: the goal of the generator is to capture the data distribution whereas the discriminator tries to distinguish the generated samples from the real samples. Comparing with other generative models, the adversarial process of GANs no longer requires a hypothetical data distribution. The disadvantage of GANs is no control on modes of the data being generated because the input of the generator is only noise without any additional information. Thus, the conditional generative adversarial networks (CGAN) came into being [33]. The conditional variable is introduced in the modeling process of the generator and the discriminator, which is used to guide the data generation process. Most methods utilized simple conditioning variables such as attributes or class labels of images [33–36]. There are also some other methods conditioned on images to generate images, including photo editing [37,38], domain transfer [39–42], etc. Phillip et al. used conditional adversarial networks as a general purpose solution to image-to-image translation problems [39]. Similarly, some cross-modal image generation methods are proposed. Kavasidis et al. proposed a method to convert brain signals into images [17]. Reed et al. proposed a method that generated the image from the text. [15]. These are applications of CGAN. Although the cross-modal image generation method [15] also used multimodal information in the process of feature learning, it usually applies when training data is paired.

3. Proposed method

Motivated by supervised multimodal deep learning and conditional generative adversarial networks (CGAN), we propose a novel semi-supervised cross-modal image generation method.

As illustrated in Fig. 2, the whole model is an end-to-end framework by seamlessly integrating two parts: the semantic learning part and the image generation part. For the semantic learning part, it contains semantic feature learning and semantic label prediction. We use a deep mutual learning strategy for semantic feature learning. In the image generation part, we use a CGAN that contains one generator and two discriminators for image generation.

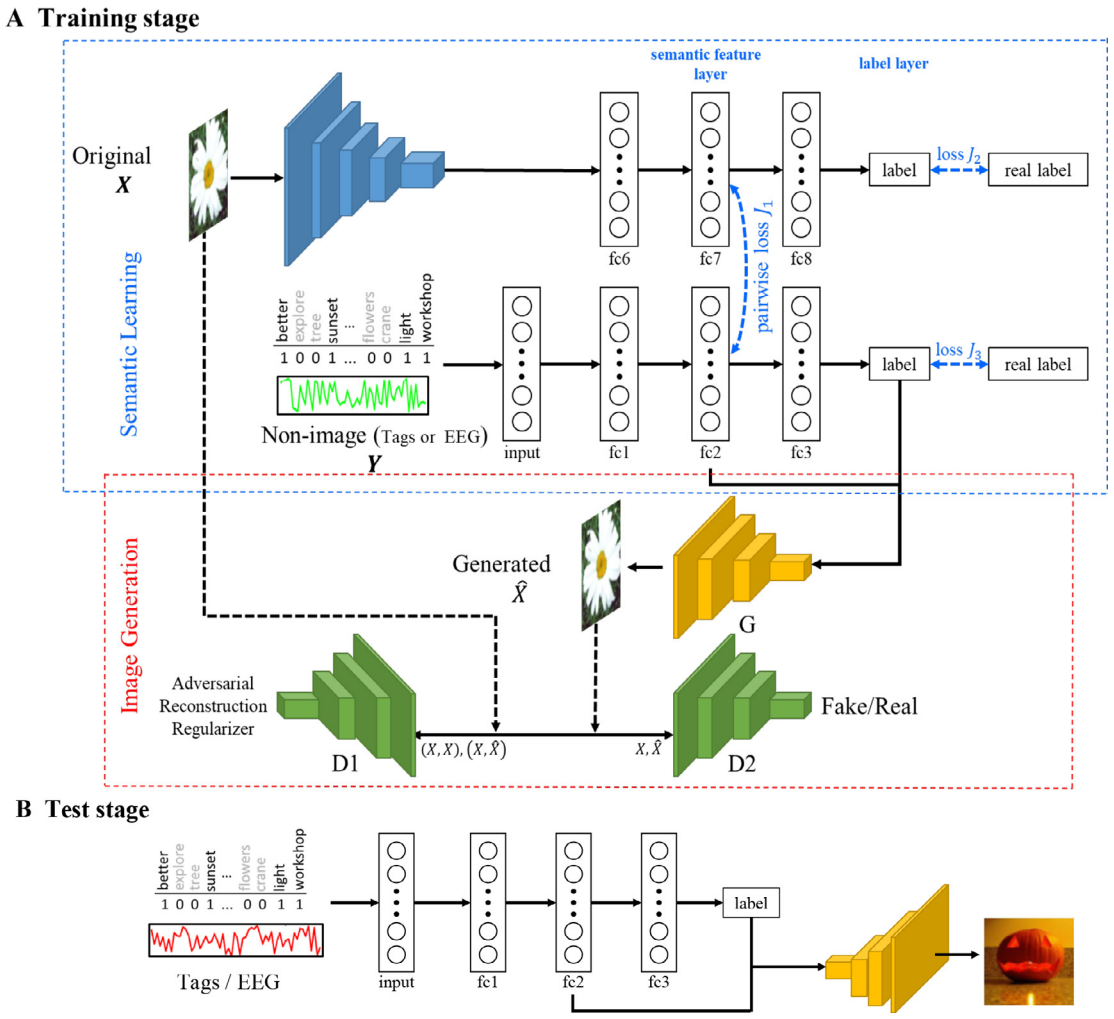


Fig. 2. The framework of our proposed cross-modal generation method. It contains two parts: the semantic learning part and the image generation part. We use the image modality and the non-image modality in the training stage. In the test stage, we only use non-image modality to generate image modality.

3.1. Supervised learning

We first introduce the proposed method under supervised learning, that is, the dataset without missing modality problem. Although the proposed framework is not limited to two modalities, we assume that the training set has n instances, and each instance has two modalities. We use $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ to denote the image modality, and $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^n$ to denote the non-image modality that matches with images. The semantic features of $\mathbf{X}$ and $\mathbf{Y}$ are denoted by $\boldsymbol{f}^{(X)} = \{\boldsymbol{f}_i^{(X)}\}_{i=1}^n$ and $\boldsymbol{f}^{(Y)} = \{\boldsymbol{f}_i^{(Y)}\}_{i=1}^n$, respectively. $\boldsymbol{l} = \{\boldsymbol{l}_i\}_{i=1}^n (\boldsymbol{l}_i \in \{0,1\}^K)$ are the real labels with dimension K, $\boldsymbol{l}^{(X)}$ and $\boldsymbol{l}^{(Y)}$ are the predicted labels from the $\mathbf{X}$ and $\mathbf{Y}$. The pairwise label similarity matrix denotes $\boldsymbol{S} \in R^{K \times K}$, where we define $S_{ij} = 1$ if x_i and y_j share as least one label, otherwise $S_{ij} = 0$. $\mathbf{z} \in \mathbb{R}^Z \sim p_z(\boldsymbol{z})$ is the prior distribution of the latent variables. The generator network is denoted by $G : \mathbb{R}^Z \times \mathbb{R}^F \times \mathbb{R}^K \longrightarrow \mathbb{R}^D$, the first discriminator is $D_1 : \mathbb{R}^D \times \mathbb{R}^F \times \mathbb{R}^K \longrightarrow \{0,1\}$, the second discriminator is $D_2 : \mathbb{R}^D \times \mathbb{R}^D \longrightarrow \{0,1\}$, where F is the dimension of semantic features, D is the dimension of the images, and Z is the dimension of the noise.

3.1.1. Semantic learning network

Since different modalities usually provide complementary information, combining the useful information from multiple modalities can improve model performance [6]. Therefore, the semantic learning part contains two deep neural networks, one for image modality and the other for non-image modality. The two networks adopt a deep mutual learning strategy for semantic feature learning. The deep neural network (DNN) for image modality is a convolutional neural network (CNN) adapted from AlexNet [43], and DNN for non-image modality is a three fully-connected network. Both of these networks contain a semantic feature layer and a softmax layer. The main tasks of this part are semantic feature learning and classification. Real labels for training data not only used for classification, but also can be viewed as supervised information for semantic feature learning.

Because of the different modalities belonging to an instance are semantically relevant [10], the semantic feature for two modalities which belonging to the same category should be similar. This can be formulated by minimizing the following loss function [9,10,44]:

$$\mathcal{J}_1 = -\sum_{i,j=1}^{n} (s_{ij}\delta_{ij} - log(1 + e^{\delta_{ij}})) \tag{1}$$

$$s_{ij} = \begin{cases} 1, & \boldsymbol{l}_i^T * \boldsymbol{l}_j > 0, \\ 0, & \boldsymbol{l}_i^T * \boldsymbol{l}_j = 0. \end{cases} \tag{2}$$

where $\delta_{ij} = \boldsymbol{f}_i^{(X)} * \boldsymbol{f}_j^{(Y)}$. From $\mathcal{J}_1$ it can be seen, in the condition $S_{ij} = 1$, the larger δ_{ij} is, the smaller $\mathcal{J}_1$ is, the greater the similarity of $\boldsymbol{f}_i^{(X)}$ and $\boldsymbol{f}_j^{(Y)}$ is, and vice versa. This ensures that the two

modalities have the same semantic space in which we use image modality to supervise non-image modality for semantic feature learning.

Furthermore, in order to maintain the semantic information during the deep mutual learning process, we use the cross-entropy loss to minimize the distance between the predicted labels and the real labels:

$$\mathcal{J}_2 = -\sum_{i=1}^{n}[\boldsymbol{l}_i \log(\boldsymbol{l}_i^{(X)})] + (1-\boldsymbol{l}_i)\log(1-\boldsymbol{l}_i^{(X)})] \tag{3}$$

$$\mathcal{J}_3 = -\sum_{i=1}^{n}[\boldsymbol{l}_i \log(\boldsymbol{l}_i^{(Y)})] + (1-\boldsymbol{l}_i)\log(1-\boldsymbol{l}_i^{(Y)})] \tag{4}$$

Thus, the overall objective function of semantic learning network for different modalities can be written as:

$$\min_{f^{X,Y},l^{X,Y},\theta^{X,Y}} \mathcal{J} = \mathcal{J}_1 + \alpha_1\mathcal{J}_2 + \alpha_2\mathcal{J}_3 \tag{5}$$

where α_1 and α_2 are hyper-parameters, $\theta^{X,Y}$ is the parameter of image network and non-image network. $\mathcal{J}_1$ is used to preserve the semantic feature similarity between two modalities. $\mathcal{J}_2$ and $\mathcal{J}_3$ are the classification loss of image modality and non-image modality, respectively

3.1.2. Image generation network

Inspired by CGAN and deep convolutional GAN [23,26,32,33], our approach trains a deep convolutional CGAN as image generation network. This network conditioned on semantic features and labels that encoded by non-image modality for image generation.

In the generator G, inference proceeds as in a deconvolutional network: we feed-forward semantic features and labels through the generator G. The generator G is parameterized by β. And a synthetic image $\widehat{\boldsymbol{x}}$ is generated via $\widehat{\boldsymbol{x}} \leftarrow G(\boldsymbol{z}, \boldsymbol{f}^{(Y)}, \boldsymbol{l}^{(Y)})$ ($\widehat{\boldsymbol{x}} \sim p_\beta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{f}^{(Y)}, \boldsymbol{l}^{(Y)})$, $\boldsymbol{x} \sim q(\boldsymbol{x})$). For the discriminator D_1 and D_2, we perform several layers of stride2 convolution with spatial batch normalization followed by leaky RELU. The discriminator D_1 is parameterized by η and the discriminator D_2 is parameterized by γ.

To ensure $\widehat{\boldsymbol{x}}$ is completely match $\boldsymbol{x}$, the most straightforward way is to minimize the euclidean distance between them: $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2^2$. But as a consequence of a ℓ_2-based pixel-wise loss, the generated images tend to be blurry [20,26]. The minimized element-wise loss for images does not consider any global similarity, like structural similarity. Therefore, inspired by the model [26], we add the discriminator D_1 and use fully adversarial training enforces $\widehat{\boldsymbol{x}}$ to be close to $\boldsymbol{x}$. Specifically, given the real image $\boldsymbol{x}$, D_1 can distinguish between two joints: the joint of $\boldsymbol{x}$ and itself, and the joint of $\boldsymbol{x}$ and its reconstruction $\widehat{\boldsymbol{x}}$. The objective function of adversarial learning for discriminator D_1 can be expressed as follows:

$$\min_{\beta}\max_{\eta}\mathcal{L}_1 = \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log(D_1(\boldsymbol{x},\boldsymbol{x}))] + \mathbb{E}_{\widehat{\boldsymbol{x}}^p\sim p_\beta(\boldsymbol{x}^p|\boldsymbol{z},\boldsymbol{f}^{(Y)},\boldsymbol{l}^{(Y)}),\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1-D_1(\boldsymbol{x},\widehat{\boldsymbol{x}}))] \tag{6}$$

where $q(\boldsymbol{x})$ denotes the true distribution of the images, $p_\beta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{f}^{(Y)}, \boldsymbol{l}^{(Y)})$ denotes the distribution of generated images. Our method replaces the MSE loss with $\mathcal{L}_1$ and encourages the generated images to be close to the original images in a global view.

The role of discriminator D_2 is the same as the standard GAN. Its goal is to match the marginal $p_\beta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{f}^{(Y)}, \boldsymbol{l}^{(Y)})$ to $q(\boldsymbol{x})$. The minmax objective of discriminator D_2 is given by the following expression:

$$\min_{\beta}\max_{\gamma}\mathcal{L}_2 = \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}\left[\log\left(D_2\left(\boldsymbol{x} \mid \boldsymbol{f}^{(Y)}, \boldsymbol{l}^{(Y)}\right)\right)\right] + \mathbb{E}_{\widehat{\boldsymbol{x}}^p\sim p_\beta\left(\boldsymbol{x}^p|\boldsymbol{z},\boldsymbol{f}^{(Y)},\boldsymbol{l}^{(Y)}\right),\ \boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\log\left(1-D_2\left(\widehat{\boldsymbol{x}} \mid \boldsymbol{f}^{(Y)}, \boldsymbol{l}^{(Y)}\right)\right)\right] \tag{7}$$

Therefore, the whole objective function of image generation network is defined as follow:

$$\min_{\beta}\max_{\eta,\gamma}\mathcal{L} = \lambda_1\mathcal{L}_1 + \mathcal{L}_2 \tag{8}$$

where λ_1 is the hyper-parameter that aims to balance the two loss function terms.

3.2. Semi-supervised learning

Many multi-modality datasets in reality often suffer from missing modality problem, which has been a challenge to address in multi-modality data analysis. For this problem, the commonly used method is to remove the samples with missing modality. As a result, it dramatically reduces the sample size and diminishes the statistical power of subsequent analysis.

Our method can solve this problem and does not need to generate the missing modality for incomplete data. In other words, our method can utilize a large amount of missing modality data during the training process. More specifically, we denote the data without missing modality as paired data and other as unpaired data. Unlike the typical semi-supervised learning which aims to utilize a large amount of unlabeled data and some labeled data to build a better classifier [45,46], in our paper, the semi-supervised learning refers to we can use some paired data and a large amount of unpaired data to build a better generation model. The paired data is denoted by $\boldsymbol{D}^p = \{(\boldsymbol{x}_i^p, \boldsymbol{y}_i^p)\}\,|_{i=1}^{n_p}$. The unpaired data of image and non-image are represented with $\boldsymbol{D}_X^u = \{(\boldsymbol{x}_i^u)\}|_{i=1}^{n_X^u}$, $\boldsymbol{D}_Y^u = \{(\boldsymbol{y}_i^u)\}\,|_{i=1}^{n_Y^u}$, respectively. Similarly, corresponding semantic features and predicted labels for paired data and unpaired data denotes $\boldsymbol{f}_p^{(X)}$, $\boldsymbol{l}_p^{(X)}$, $\boldsymbol{f}_p^{(Y)}$, $\boldsymbol{l}_p^{(Y)}$ and $\boldsymbol{f}_u^{(X)}$, $\boldsymbol{l}_u^{(X)}$, $\boldsymbol{f}_u^{(Y)}$, $\boldsymbol{l}_u^{(X)}$, respectively. The real labels of paired data and unpaired data are l^p and l^u, respectively.

Due to a large number of image resources on the internet, image modality missing is infrequent. Instead, the non-image modality that matches the image, such as brain signal is often difficult to obtain. So in our semi-supervised learning process, we only consider the problem of non-image modality missing. The unpaired images are used to train the image network from which we can get the semantic information of unpaired images. At the same time, we use this semantic information to train CGAN. Benefited by a large amount of unpaired image data, the deep network can be trained very well and the image network performs better in the semantic features learning process. Considering the image network and non-image network use the mutual learning strategy in the training process, the unpaired image is also conducive to non-image network. Simultaneously, the image generation network trained by a large number of semantic features $\boldsymbol{f}_u^{(X)}$ and labels $\boldsymbol{l}_u^{(X)}$ can generate images more realistic. The objective function of the semantic networks in semi-supervised learning can be expressed as follows:

$$\min_{f_p^{X,Y},l_p^{X,Y},\theta^{X,Y}} \mathcal{J} = \mathcal{J}_1 + \alpha_1\mathcal{J}_2 + \alpha_2\mathcal{J}_3 + \mathcal{J}_4 + \alpha_3\mathcal{J}_5 \tag{9}$$

$$\mathcal{J}_4 = -\sum_{i,j=1}^{n}\left(\widetilde{s_{ij}}\widetilde{\delta_{ij}} - log\left(1+e^{\widetilde{\delta_{ij}}}\right)\right) \tag{10}$$

$$\widetilde{s_{ij}} = \begin{cases} 1, & (\boldsymbol{l}_i^p)^T * \boldsymbol{l}_j^u > 0, \\ 0, & (\boldsymbol{l}_i^p)^T * \boldsymbol{l}_j^u = 0. \end{cases},\quad \widetilde{\delta_{ij}} = (\boldsymbol{f}_p^{(Y)})_i * (\boldsymbol{f}_u^{(X)})_j \tag{11}$$

$$\mathcal{J}_5 = \sum_{i=1}^{K}[\boldsymbol{l}_i^u \log(\boldsymbol{l}_{ui}^{(X)})] + (1-\boldsymbol{l}_i^u)\log(1-\boldsymbol{l}_{ui}^{(X)})]. \tag{12}$$

where $\mathcal{J}_1$, $\mathcal{J}_2$ and $\mathcal{J}_3$ are the same as above, α_1, α_2 and α_3 are the hyper-parameters.

The objective function for image generation network in semi-supervised learning is denoted as follow:

$$\begin{aligned}\min_{\beta}\max_{\eta,\gamma}\mathcal{L} &= \lambda_1\mathcal{L}_1 + \mathcal{L}_2 + \lambda_2\mathcal{L}_3 + \mathcal{L}_4 \\ &= \lambda_1\mathbb{E}_{\boldsymbol{x}^p\sim q(\boldsymbol{x}^p)}[\log(D_1(\boldsymbol{x}^p,\boldsymbol{x}^p))] \\ &\quad +\lambda_1\mathbb{E}_{\widehat{\boldsymbol{x}}^p\sim p_\beta(\boldsymbol{x}^p|\boldsymbol{z},\boldsymbol{f}^{(Y)},\boldsymbol{l}^{(Y)}),\ \boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1-D_1(\boldsymbol{x}^p,\widehat{\boldsymbol{x}}^p))] \\ &\quad +\mathbb{E}_{\boldsymbol{x}^p\sim q(\boldsymbol{x}^p)}\left[\log\left(D_2\left(\boldsymbol{x}^p \mid \boldsymbol{f}_p^{(Y)},\boldsymbol{l}_p^{(Y)}\right)\right)\right] \\ &\quad +\mathbb{E}_{\widehat{\boldsymbol{x}}^p\sim p_\beta(\boldsymbol{x}^p|\boldsymbol{z},\boldsymbol{f}^{(Y)},\boldsymbol{l}^{(Y)}),\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\log\left(1-D_2\left(\widehat{\boldsymbol{x}}^p \mid \boldsymbol{f}_p^{(Y)},\boldsymbol{l}_p^{(Y)}\right)\right)\right] \\ &\quad +\lambda_2\mathbb{E}_{\boldsymbol{x}^u\sim q(\boldsymbol{x}^u)}[\log(D_1(\boldsymbol{x}^u,\boldsymbol{x}^u))] \\ &\quad +\lambda_2\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\log\left(1-D_1\left(\boldsymbol{x}^u,G\left(\boldsymbol{z},\boldsymbol{f}_u^{(X)},\boldsymbol{l}_u^{(X)}\right)\right)\right)\right] \\ &\quad +\mathbb{E}_{\boldsymbol{x}^u\sim q(\boldsymbol{x}^u)}\left[\log\left(D_2\left(\boldsymbol{x}^u \mid \boldsymbol{f}_u^{(X)},\boldsymbol{l}_u^{(X)}\right)\right)\right] \\ &\quad +\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\log\left(1-D_2\left(G\left(\boldsymbol{z},\boldsymbol{f}_u^{(X)},\boldsymbol{l}_u^{(X)}\right) \mid \boldsymbol{f}_u^{(X)},\boldsymbol{l}_u^{(X)}\right)\right)\right] \quad (13)\end{aligned}$$

where λ_1 and λ_2 are the hyper-parameters.

3.3. Optimizing strategy

The proposed method can integrate semantic learning and image generation into a unified framework. However, the semantic networks may not match well with the image generation network in the training process. This is because the convergence speed between the networks is quite different in the process of optimization. Therefore, we should exploit an alternative strategy to optimize the parameters of these networks.

Specifically, in the optimizing process, we first set different learning rate for the two semantic networks and the one image generation network. The convolution layers of image network come from AlexNet that have been pre-trained on ImageNet dataset. The reason for pre-training the convolution layers was to increase the speed of training the networks for faster experimentation. We randomly initialize the fully-connected layers of image network and non-image network and train it. Secondly, we utilize the obtained semantic features and labels in semantic networks for image generation network training. Finally, repeating such training procedure for semantic networks and image generation network until convergence. The entire optimization process is end-to-end.

Owing to the optimization strategy for supervised learning is the same as semi-supervised learning, here we only show the optimization details of semi-supervised learning. Algorithm 1 outlines the whole learning algorithm. All network parameters are learned by utilizing the stochastic gradient descent (SGD) with the back-propagation (BP) algorithm.

Algorithm 1 The semi-supervised leaning algorithm for our method.

Input:
Paired image set X^p, Paired non-image set Y^p, Paired Label set l_p, Unpaired image set X^u, Unpaired label set l_u
Output: Semantic label l_u, l_p, Generated image $\widehat{x}_p$, $\widehat{x}_u$
Intitialization
Initialize network parameters $\theta^{(X)}$, $\theta^{(Y)}$, β, γ, η
hyper-parameters: $\alpha_{1\sim3}$, $\lambda_{1\sim2}$
learning rate: μ_1, μ_2
mini-batch size: N = 64
maximum iteration number: T_{max}
repeat
  for T_{max} iteration **do**
    Update $\theta^{(X)}, \theta^{(Y)}$ by SGD algorithm
    $\theta^{(X,Y)} \leftarrow \theta^{(X,Y)} - \mu_1\partial\mathcal{J}/\partial\theta^{(X,Y)}$
    Update γ, η by SGD algorithm
    $(\gamma, \eta) \leftarrow (\gamma, \eta) - \mu_2\partial\mathcal{L}/\partial(\gamma, \eta)$ {Update discrimators}
    Update β by SGD algorithm
    $\beta \leftarrow \beta - \mu_2\partial\mathcal{L}/\partial\beta$ {Update generator}
  end for
until convergence

4. Experiments

4.1. Datasets

We briefly introduce the three datasets: fMRI-HC, EEG-ImageNet and MIRFLICKR-25K, which are used in the following experiments.

For the fMRI (functional magnetic resonance imaging)-HC (Handwritten Characters) dataset, each instance is an fMRI-image pair. The fMRI data acquisition experiments have subjects viewing grey-scale handwritten character images from 6 semantic categories. Furthermore, we selected all the images from the handwritten character image dataset that the categories are the same as the given stimulus to verify the semi-supervised function of our method.

In EEG-ImageNet dataset, each instance is an image-EEG pair [47]. The EEG data are recorded when people were shown visual stimuli of objects. The dataset used for visual stimuli is a subset of ImageNet, containing 40 classes of easily recognizable objects. During the EEG data acquisition experiment, 2000 images (50 from each class) are shown in bursts for 0.5 s each. And they conducted the experiments using a 128-channel cap with active. Furthermore, we selected all the images from the ImageNet dataset that the categories are the same as the given stimulus to verify the semi-supervised function of our method.

The MIRFLICKR-25K dataset contains 25,000 images collected from Flickr website [48]. Each image is associated with several textual tags. Each instance is an image-tags pair. This dataset can be seen as multimodal data. Following the paper [9], we also select the instances with at least 20 tags for our experiment. The text for each instance is represented as a 1386 dimensional BOW vectors, and each instance is manually annotated with at least one of 24 unique labels.

For paired dataset, we randomly select 80% of the dataset as the training data, and the remaining 20% as the test data. All unpaired data are used as part of the training data. In the cross-modal generation task, we only use the non-image modality in the testing process.

4.2. Compared methods

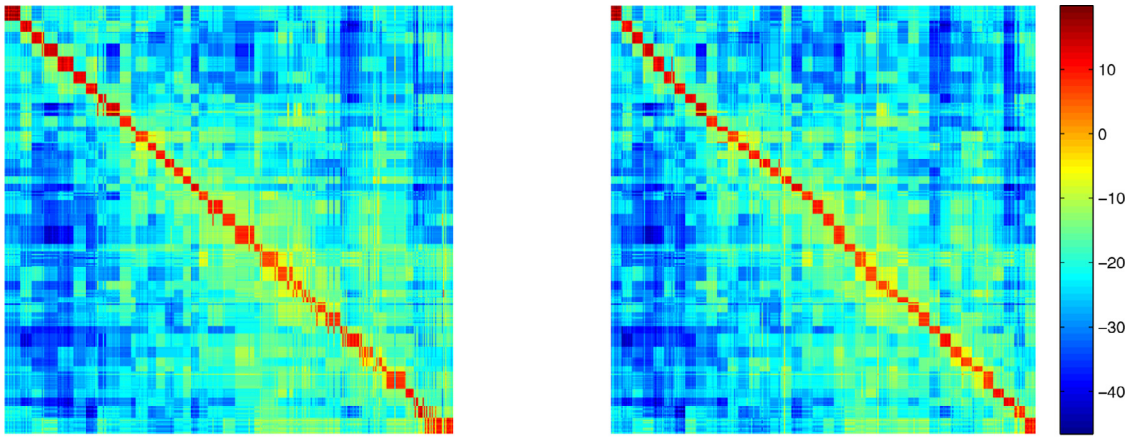
The following state-of-the-art methods are used as baselines:

Multimodal Variational Autoencoder (**MVAE**) [49] is a method that contains two parts: encoder and decoder. For the fMRI-HC dataset, the input for the encoder is the combination of the semantic features of fMRI and the semantic features of the image. And in the test stage, we only use the semantic feature of fMRI. For the EEG-ImageNet dataset, the setting of VAE [50] is the same as paper [17].

Brain2image [17] is a method that generates the images from EEG signals with GAN. For the EEG-ImageNet dataset of our paper, the training set and test set are divided in the same way as Brain2image method.

Text-to-image [15] is a method that generates the image from text with GAN.

CGAN [33] provides an algorithm for learning to generate samples conditioned on a discrete label. Conditioning on the semantic features encoded by non-image modalities, we use CGAN method to generate images in our paper.



(a) EEG and Images (b) EEG and Unpaired Images

Fig. 3. The visualization of similarity matrix.

ACGAN [35] is a conditional image synthesis method with auxiliary classifier. The discriminator of the ACGAN can give the semantic labels of the input data. In our paper, the input for the generator is the combination of the noise and semantic features.

TripleGAN [51] is a method that contains three parts: generator G, discriminator D and classifier C. The input for the generator G is the combination of the noise and semantic labels, and the input for the classifier C is the real images or generated images.

4.3. Implementation details

4.3.1. Image network

The DNN for image modality is CNN. The training image size was set to $224 \times 224 \times 3$. In order to apply CNN to our model, we first pre-train the AlexNet on ImageNet. And we reserve the convolution layers of AlexNet for our image network. Then we built a three-layer fully-connected neural network ($4096 \rightarrow 200 \rightarrow K$), and the parameters for the fully-connected network are randomly initialized. The nodes of output layer K are the number of labels.

4.3.2. Non-image network

We built the non-image network by using a three-layer feed-forward neural network, which is used to project an EEG or tags into semantic features and labels ($Y \rightarrow 4096 \rightarrow 200 \rightarrow K$).

4.3.3. Generative network for images

We build the generator network by using a deconvolutional network, and the network contains four convolution layers. The number of convolution kernels is 256, 128, 64, 3. The size of the convolution kernels is 5×5. The network architecture of the discriminator is almost symmetrical with the generator. The generated images size is $64 \times 64 \times 3$.

4.4. Experimental results and discussions

4.4.1. Semantic feature analysis

In order to demonstrate the effectiveness of our proposed deep mutual learning strategy, we perform a visual analysis of the generated semantic features on EEG-ImageNet. In this analysis, the non-image modality comes from the test data and the images match the non-image modality in the test data. Unpaired images are from the training data.

For EEG-ImageNet dataset, we visualize the semantic features similarity between the image modality and the EEG signal modality. The similarity metric is the inner product. Visualization results are shown in Fig. 3. Elements in Fig. 3(a) indicate the semantic features similarity between the EEG and the paired image. Elements in Fig. 3(b) indicate the semantic features similarity between the EEG and the unpaired image. The block diagonal structures of Fig. 3(a) and (b) is the similarity of semantic features which belong to the same category.

From Fig. 3, it can be seen that the semantic features of non-image network are similar to those of image network. It is manifesting that non-image modality is well assisted by image modality in semantic feature learning.

Table 1
Multi-label classification results on tags. ↑ means the larger the better, ↓ means the smaller the better.

Method	*MicroF1*↑	*MacroF1*↑	*Hamming loss*↓	*Ranking loss*↓
DNN-tags	0.590	0.550	0.117	0.108
TRAM [53]	0.572	–	0.132	0.112
CA2E [52]	0.590	0.544	0.133	0.120
Our Method	**0.602**	**0.560**	**0.108**	**0.098**

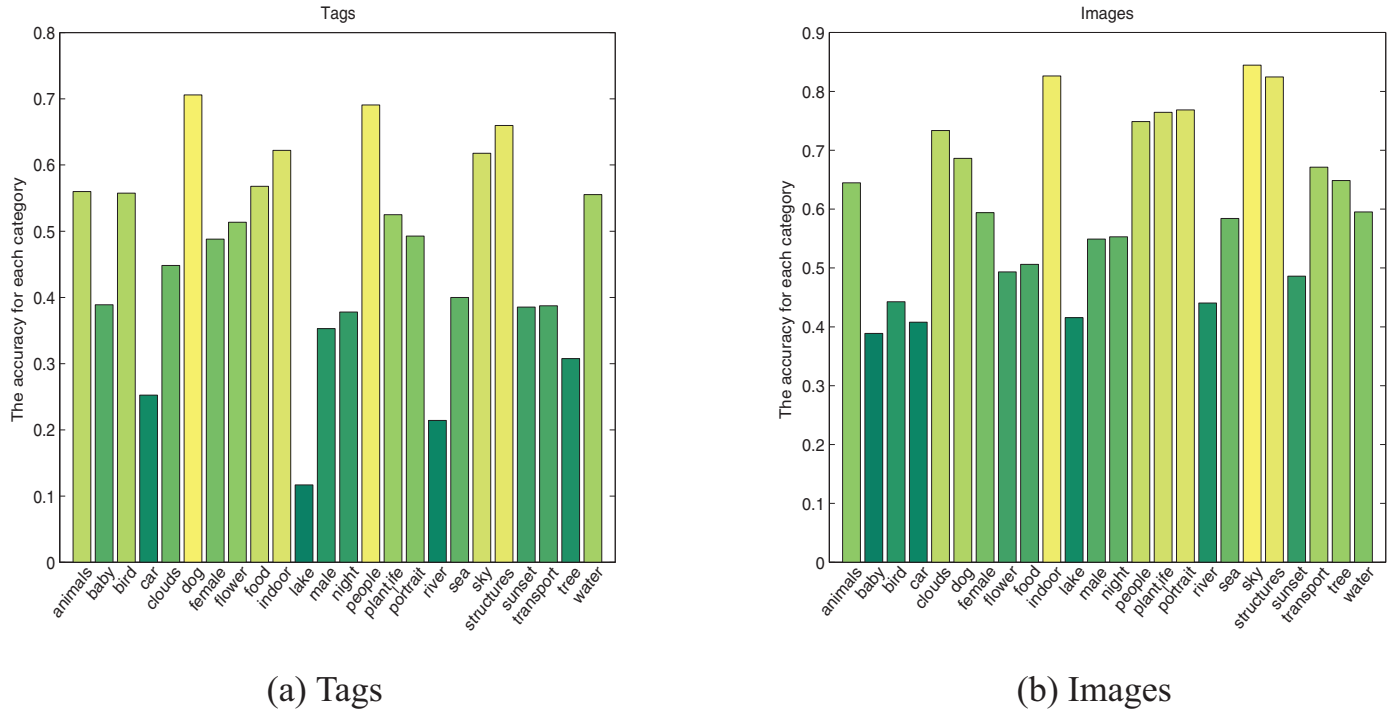
4.4.2. Multi-label classification on MIRFLICKR-25k

We also conduct experiments on MIRFLICKR-25K dataset to validate the effectiveness of the proposed deep mutual learning strategy. In previous papers, this dataset was mainly used for multi-label classification and retrieval tasks [9,10]. For our experiments, we first carry out the multi-label classification on non-image modality (tags) for this dataset. The activation function used for the output layer is sigmoid. For the hyper-parameters in semantic networks, we find that the optimal result can be obtained when $\alpha_1 = \alpha_2 = 1$. The learning rate is chosen to 2^{-4}.

For the semantic networks, the training data is images and tags, and the test data can be tags.

We first compare our method with DNN-tags method which only uses tags network in the training process. In addition to a baseline method of three layers DNN, we use **CA2E** [52], a method that joint feature and label embedding by a deep latent space. And CA2E is composed of 2 layers of fully connected neural network and both single fully connected layer structures. We also compare our method with **TRAM** which is a transductive multi-label learning method [53]. Table 1 reports the experimental results of our method and other multi-label learning algorithms on the MIRFLICKR-25K data.

MicroF1, MacroF1, Hamming loss and Ranking loss are multi-label classification evaluation criteria [53]. MicroF1 can evaluate both microaverage of precision and micro average of recall with



(a) Tags (b) Images

Fig. 4. The accuracy for each category on tags and images.

equal importance:

$$MicroF1 = \frac{2 \times \sum_{i \in \mu} |l_i^{(Y)} \cap l_i|}{\sum_{i \in \mu} |l_i^{(X)}| + \sum_{i \in \mu} |l_i|}$$

MacroF1 can calculate metrics for each label, and find their unweighted mean:

$$MacroF1 = \frac{1}{K} \sum_{j=1}^{K} \frac{2 p_j r_j}{p_j + r_j},$$

where p_j and r_j are precision and recall of ith label. Hamming loss can evaluate how many times an instance-label pair is misclassified:

$$Hamming\ loss = \frac{1}{|Y_\mu|} \sum_{i \in \mu} \frac{1}{m} |l_i^{(Y)} \triangle l_i|,$$

where $\triangle$ stands for the symmetric difference of two sets, Y_μ is a test data. Ranking loss can evaluate the average fraction of label pairs that are not correctly ordered:

$$Ranking\ loss = \frac{1}{|Y_\mu|} \sum_{i \in \mu} \frac{1}{|l_i||\bar{l}_i|} |\mathcal{S}_{rank}^i|,$$

$$\mathcal{S}_{rank}^i = \{(u, v) | l_{iu}^{(Y)} < l_{iv}^{(Y)}, l_{iu} = 1, l_{iv} = 0\}$$

where the $\bar{l}_i$ denotes the complementary set of l_i in entire label space. For each evaluation criteria, $\uparrow$ means the larger the better, $\downarrow$ means the smaller the better. We select the labels that predicted probability value greater than 0.5 as the final predicted label. From Table 1, it can be seen the method that used multimodal information on classification is superior to other single-modal classification methods.

The accuracy for each category on tags and images is shown in Fig. 4. Tags and images are matched in the test dataset. It can be seen from Fig. 4 that the accuracy of 14 categories is more than 40%. The category with the lowest classification accuracy is “lake”, and the “water” with the high classification accuracy. I think the reason for this is that there are a few samples corresponding to the “lake”, and the samples corresponding to “lake” also contains the label of “water”. “Water” is a broad word with a large number of samples. The probability that the sample is predicted to be the label “water” is much larger than the “lake”. Furthermore, we can find that the accuracy for each category on tags has the same trend with images.

The above results of multi-label classification show that deep mutual learning improves tags classification accuracy. It also demonstrates that the non-image network is well supervised by image network.

Table 2
Comparison of experimental results on the fMRI-HC dataset.

Method	PCC	ACC-SVM
MVAE [49,50]	0.517	0.633
CGAN [33]	0.534	0.700
ACGAN [35]	0.556	0.733
TripleGAN [51]	0.561	0.733
Our Method	**0.570**	**0.767**

4.4.3. Cross-modal image generation

In this section, we study the performance of our method for cross-modal image generation on three datasets. For fair comparison, we use the same network architecture in all compared methods.

Firstly, we conduct experiments on the fMRI-HC dataset to generate images from fMRI data. Performance comparisons are list in Table 2. The evaluation criteria are PCC and Accuracy-SVM (ACC-SVM). PCC is Pearsons Correlation Coefficient. For the ACC-SVM, we train the linear SVM on the presented visual images. The pre-trained model is regarded as a gold-standard classifier to classify the generated images. The labels of test data are used as ground truth to calculate the classification accuracy. From Table 2, we can see that our method consistently outperforms the baselines. The

Fig. 5. Generated images from fMRI on the fMRI-HC dataset.

Table 3
Ablation study results on the EEG-ImageNet datatset.

Method	Inception scores
One-hot Conditioning	6.41
Supervised	5.89
MSE	6.27
Two-stage	6.01
Without-image-modality	5.01
Our Method	**6.76**

Table 4
Comparison of experimental results on the EEG-ImageNet dataset.

Method	Inception scores
VAE [50]	4.49
Brain2image [17]	5.07
CGAN [33]	5.06
ACGAN [35]	5.10
TripleGAN [51]	5.49
Our Method	**6.76**

MVAE method achieves the worst performance. Comparing with CGAN, the results of ACGAN and TripleGAN are slightly higher because the two methods add a semantic classifier for generated images and real images. Fig. 5 shows some images generated from our method and the compared methods. As can be seen from Fig. 5, the contour, texture and color of generated handwritten characters in our method are very similar to the original images. And the generated images in our method is very clear. One reason for this is that our method can utilize a large amount of unpaired data in the case of the number of paired data limited.

Then we conduct experiments on the EEG-ImageNet dataset. Generating the image from the EEG signal is a challengeable problem since the brain signal always contains noise and the number of brain signal is small. In the training process, we use paired data and unpaired image data. We find that when α_1, α_2 and $\alpha_3 \in \{0.005, 0.01, 1\}$ and $\lambda_1 = \lambda_2 = 1$ the optimal result can be obtained. Performance comparisons are listed in Tables 3 and 4. We choose a recently proposed numerical assessment approach Inception Score for quantitative evaluation [54]:

$$I = exp(\mathbb{E}_x D_{kl}(p(l|x) \| p(l)))$$

where x denotes a generated image, and l is the label predicted by the Inception models. The intuition behind this metric is that good models should generate diverse but meaningful images.

To validate that the semantic feature learned from the semantic networks can be viewed as a bridge between the different modalities, we compare our method with the method (**One-hot Conditioning**) only conditioned on one-hot labels for image generation. To validate the advantages of semi-supervised learning, we compare our method with the method (**Supervised**) that only uses paired data in the training process. To validate our discriminator D_1 is more conducive to adversarial learning than MSE loss, we compare our method with the method (**MSE**) that use MSE loss and discriminator D_2 in the image generation process. The MSE loss is the Euclidean distance between generated images and real images, and the dimension of the vector that used to calculate the Euclidean distance is 64*64*3. To validate the effectiveness of our proposed self-optimization strategy, we compare our method with the method (**Two-stage**) that semantic feature learning and image generation are divided into two stages in the training process. To validate the effectiveness of image modality in the image generation task, we compare our method with the method (**Without-image-modality**) that no image modality in the training process. From Table 3, it can be seen we validate the contributions mentioned above on five aspects. Our method outperforms all of its variants, which shows the effectiveness of each module. The **Without-image-modality** method achieves the worst performance, we find that image modality plays an important role in the proposed framework.

The results of the comparison experiment on the EEG-ImageNet dataset are shown in Table 4. From Table 4, we can see our method achieves the best performance and significantly outperforms other methods. As with the fMRI-HC dataset, the VAE method achieves the worst performance. The Brain2image, CGAN, and ACGAN are the variant of GAN. In the cross-modal image generation task, these methods can only use paired data (without modality missing) in the training process. However, our method can utilize a large amount of unpaired image data, so the results of these methods are significantly lower than our method. Although the TripleGAN method can also use a large amount of unpaired data in the training process, the result of our method is still higher than TripleGAN. Fig. 6 shows some of the images generated from our method. From Fig. 6, we find that the images generated by our method are clear and have obvious semantic features.

Finally, we conduct experiments on the MIRFLICKR-25K dataset for cross-modal image generation. The MIRFLICKR-25K is a very complex dataset that contains multiple objects and variable backgrounds. We use the same semantic networks architecture- and CGAN architecture as the EEG-ImageNet. In the training of semantic networks, the difference between EEG-ImageNet and

(a) Panda

(b) Airliner

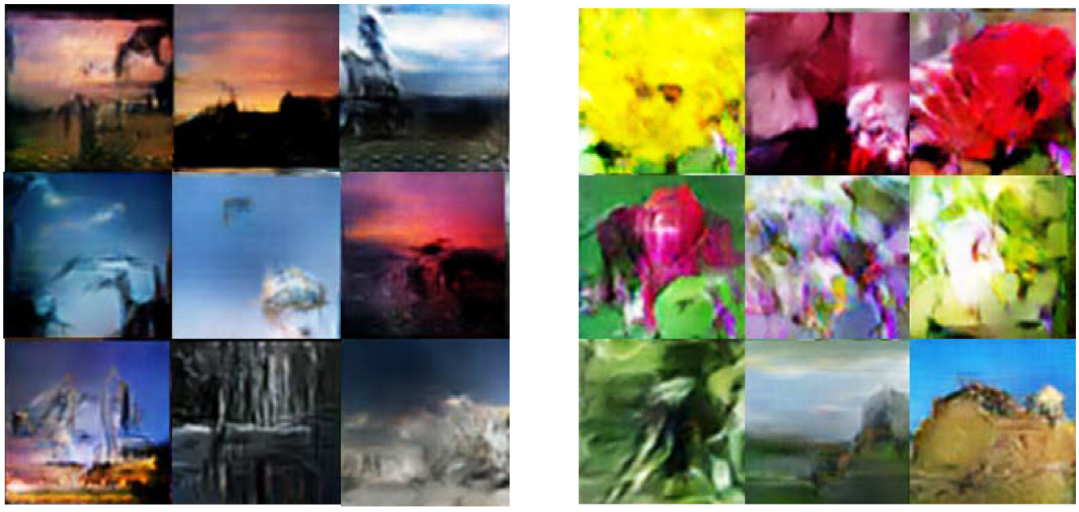
(c) Pizza

(d) Daisy

(e) Bolete

(f) Reflex camera

Fig. 6. Generated images from EEG on the EEG-ImageNet dataset.



(a) sky, clouds, sunset, night

(b) flower, plant-life, tree

(c) food, indoor

(d) water, sea, lake

Fig. 7. Generated images from tags on the MIRFLICKR-25K dataset.

Table 5
Ablation study results on the MIRFLICKR-25K datatset. O.F1, C.F1 denote microF1, macroF1 respectively.

Method	MicroF1	MacroF1	Inception scores
One-hot Conditioning	0.31	0.14	4.95
Two-stage	0.33	0.17	5.01
MSE	0.31	0.11	4.80
Without-image-modality	0.33	0.16	5.13
Our Method	**0.35**	**0.20**	**5.31**

Table 6
Comparison of experimental results on the MIRFLICKR-25K datatset. O.F1, C.F1 denote microF1, macroF1 respectively.

Method	MicroF1	MacroF1	Inception scores
Text-to-image [15]	0.29	0.12	4.90
CGAN [33]	0.31	0.15	5.03
ACGAN [35]	0.32	0.16	5.20
TripleGAN [51]	0.31	0.16	5.25
Our Method	**0.35**	**0.20**	**5.31**

MIRFLICKR-25K is that there is no single object category for each class of MIRFLICKR-25K. We use Inception scores and microF1 / macroF1 as the evaluation metrics in this part. The microF1 / macroF1 is the same as above, which is multi-label classification evaluation criteria. Here, we use the pre-trained AlexNet to classify the generated images, and the microF1 / macroF1 are used to evaluate multi-label classification results. In these experiments, we do not use unpaired data. The results of ablation study are shown in Table 5. The same as the above, we validate our method from four aspects. The results of the comparison experiment on the MIRFLICKR-25K dataset are shown in Table 6. From Table 6, we can see our method achieves better performance on the MIRFLICKR-25K dataset. The generated images are shown in Fig. 7. Because MIRFLICKR-25K is a very complex dataset, there are still a lot of images that can't be generated very well.

4.5. Sensitivity analysis

In this section, we conduct experiments to analyze the sensitivity of hyper-parameters and the sensitivity of the dimension of semantic features.

In our method, we have hyper-parameters $\alpha_{1\sim3}$ and $\lambda_{1\sim2}$. The hyper-parameters α_1, α_2 and α_3 control the classification loss, and the hyper-parameters λ_1 and λ_2 control the adversarial loss of the additional discriminator. To facilitate the study of the sensitivity of our method to different hyper-parameter values, we plot the Inception Scores of generated images on EEG-ImageNet dataset by tuning the values of $\alpha(\alpha_1 = \alpha_2 = \alpha_3)$ and $\lambda(\lambda_1 = \lambda_2)$. The results are shown in Fig. 8. For the different α, the performance of our method slightly fluctuates. For the different λ, our method also

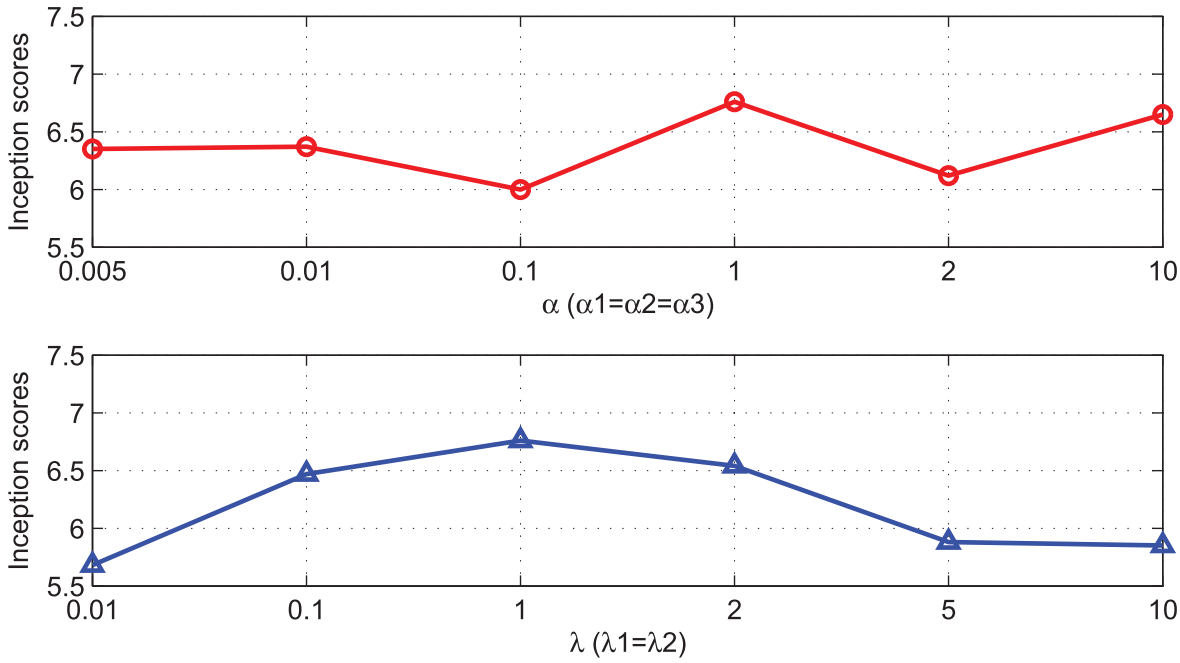


Fig. 8. A sensitivity analysis of the hyper-parameters.

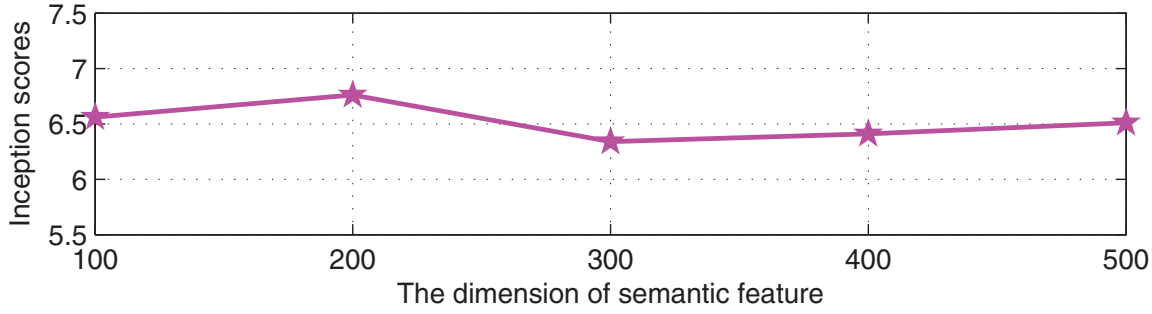


Fig. 9. A sensitivity analysis of the dimension of semantic feature.

performs stably in a fixed range of λ. This result also verifies the importance of the additional discriminator.

For the sake of studying the sensitivity of our method with respect to different dimensions of semantic features, we plot the Inception Scores of generated images on EEG-ImageNet dataset by tuning the dimensions of semantic features. The results are shown in Fig. 9. From Fig. 9, we can see the Inception Scores of our method changed from 6.34 to 6.76, which demonstrates that our method performs stably with different dimensions of semantic features.

5. Conclusion

In this paper, we propose a novel cross-modal semi-supervised image generation method. Our method learns the semantic feature from non-image modality to solve the modality gap problem. In the semantic feature learning, we use image modality to assist non-image modality and adopt a deep mutual learning strategy in the training process. For the image generation, our method adds an additional discriminator to ensure the generated images of completely match the real images. Our method can easily utilize a large amount of unpaired data to solve the problem of lack of training pairs data. And we adopt an alternative training strategy to jointly optimize the parameters of all networks. Extensive experiments on three datasets clearly validate our method is effective for two tasks. It also shows that our method can be applied to many fields.

In the process of semantic feature learning, we only consider the semantic feature of the middle layer of the semantic networks. In the future work, we will investigate the method for determining which layer of the semantic network should or should not be used for deep mutual learning.

Acknowledgment

This work was supported by National Natural Science Foundation of China (81701785, 61906188), CAS Scientific Research Equipment Development Project (YJKYYQ20170050), Strategic Priority Research Program of CAS (XDB32040200) and the Beijing Municipal Science Technology Commission (Z181100008918010).

References

[1] Y. Li, Z. Zhang, Y. Cheng, L. Wang, T. Tan, Mapnet: multi-modal attentive pooling network for RGB-D indoor scene classification, Pattern Recognit. 90 (2019) 436–449.

[2] C. Du, C. Du, H. Wang, J. Li, W.-L. Zheng, B.-L. Lu, H. He, Semi-supervised deep generative modelling of incomplete multi-modality emotional data, in: Proceedings of the ACM international conference on Multimedia, ACM, 2018, pp. 108–116.

[3] W. Zheng, W. Liu, Y. Lu, B. Lu, A. Cichocki, Emotionmeter: a multimodal framework for recognizing human emotions., IEEE Trans. Cybern. 49 (3) (2019) 1110–1122.

[4] A. Shahroudy, T.-T. Ng, Y. Gong, G. Wang, Deep multimodal feature analysis for action recognition in RGB-D videos, IEEE Trans. Pattern Anal. Mach.Intell. 40 (5) (2018) 1045–1058.
[5] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi–path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit. 86 (2019) 376–385.
[6] T. Chen, S. Wang, S. Chen, Deep multimodal network for multi-label classification, in: International Conference on Multimedia and Expo (ICME), IEEE, 2017, pp. 955–960.
[7] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of ICML, 2011, pp. 689–696.
[8] N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, in: Advances in Neural Information Processing Systems, 2012, pp. 2222–2230.
[9] Q.-Y. Jiang, W.-J. Li, Deep cross-modal hashing, in: Proceedings of CVPR, 2017, pp. 3232–3240.
[10] C. Li, C. Deng, N. Li, W. Liu, X. Gao, D. Tao, Self-supervised adversarial hashing networks for cross-modal retrieval, in: Proceedings of CVPR, 2018, pp. 4242–4251.
[11] V.E. Liong, J. Lu, Y.-P. Tan, Cross-modal discrete hashing, Pattern Recognit. 79 (2018) 114–129.
[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of ICML, 2015, pp. 2048–2057.
[13] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, B. Plank, Automatic description generation from images: a survey of models, datasets, and evaluation measures, J. Artif. Intell. Res. 55 (2016) 409–442.
[14] X. Xiao, L. Wang, K. Ding, S. Xiang, C. Pan, Dense semantic embedding network for image captioning, Pattern Recognit. 90 (2019) 285–296.
[15] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: Proceedings of ICML, 2016, pp. 1060–1069.
[16] P. Tirupattur, Y.S. Rawat, C. Spampinato, M. Shah, Thoughtviz: visualizing human thoughts using generative adversarial network, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2018, pp. 950–958.
[17] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, M. Shah, Brain2image: Converting brain signals into images, in: Proceedings of the ACM international conference on Multimedia, ACM, 2017, pp. 1809–1817.
[18] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of CVPR, 2015, pp. 1–9.
[20] L. Cai, Z. Wang, H. Gao, D. Shen, S. Ji, Deep adversarial learning for multi-modality missing data completion, in: Proceedings of KDD, ACM, 2018, pp. 1158–1166.
[21] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, S. Ji, Deep learning based imaging data completion for improved brain disease diagnosis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2014, pp. 305–312.
[22] L. Tran, X. Liu, J. Zhou, R. Jin, Missing modalities imputation via cascaded residual autoencoder, in: Proceedings of CVPR, 2017, pp. 1405–1414.
[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
[24] M. Fabbri, G. Borghi, F. Lanzi, R. Vezzani, S. Calderara, R. Cucchiara, Domain translation with conditional GANs: from depth to RGB face-to-face, in: International Conference on Pattern Recognition (ICPR), 2018.
[25] K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, M. van Gerven, Generative adversarial networks for reconstructing natural images from brain activity, NeuroImage 181 (2018) 775–785.
[26] C. Li, H. Liu, C. Chen, Y. Pu, L. Chen, R. Henao, L. Carin, Alice: Towards understanding adversarial learning for joint distribution matching, in: Advances in Neural Information Processing Systems, 2017, pp. 5495–5503.
[27] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: a survey and taxonomy, IEEE Trans. Pattern Anal. Mach.Intell. 41 (2) (2019) 423–443.
[28] K. Li, C. Zou, S. Bu, Y. Liang, J. Zhang, M. Gong, Multi-modal feature fusion for geographic image annotation, Pattern Recognit. 73 (2018) 1–14.
[29] S.H. Amiri, M. Jamzad, Efficient multi-modal fusion on supergraph for scalable image annotation, Pattern Recognit. 48 (7) (2015) 2241–2253.
[30] R. Salakhutdinov, G. Hinton, Deep Boltzmann machines, in: Artificial Intelligence and Statistics, 2009, pp. 448–455.
[31] E.L. Denton, S. Chintala, R. Fergus, et al., Deep generative image models using a- Laplacian pyramid of adversarial networks, in: Advances in Neural Information Processing Systems, 2015, pp. 1486–1494.
[32] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv:1511.06434 (2015).
[33] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv:1411.1784 (2014).
[34] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: interpretable representation learning by information maximizing generative adversarial nets, in: Advances in Neural Information Processing Systems, 2016, pp. 2172–2180.
[35] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, in: Proceedings of ICML, 2017, pp. 2642–2651.
[36] X. Yan, J. Yang, K. Sohn, H. Lee, Attribute2image: Conditional Image Generation From Visual Attributes, in: Proceedings of ECCV, Springer, 2016, pp. 776–791.
[37] A. Brock, T. Lim, J.M. Ritchie, N. Weston, Neural photo editing with introspective adversarial networks, arXiv:1609.07093 (2016).
[38] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, A.A. Efros, Generative visual manipulation on the natural image manifold, in: Proceedings of ECCV, Springer, 2016, pp. 597–613.
[39] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of CVPR, 2017, pp. 1125–1134.
[40] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems, 2017, pp. 700–708.
[41] J. Lim, Y. Yoo, B. Heo, J.Y. Choi, Pose transforming network: learning to disentangle human posture in variational auto-encoded latent space, Pattern Recognit. Lett. 112 (2018) 91–97.
[42] Y. Li, L. Song, X. Wu, R. He, T. Tan, Learning a bi-level adversarial network with global and local perception for makeup-invariant face verification, Pattern Recognit. 90 (2019) 99–108.
[43] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[44] N. Li, C. Li, C. Deng, X. Liu, X. Gao, Deep joint semantic-embedding hashing., in: Proceedings of IJCAI, 2018, pp. 2397–2403.
[45] X.J. Zhu, Semi-supervised learning literature survey, Technical Report, University of Wisconsin-Madison Department of Computer Sciences, 2005.
[46] G. Yu, G. Zhang, C. Domeniconi, Z. Yu, J. You, Semi-supervised classification based on random subspace dimensionality reduction, Pattern Recognit. 45 (3) (2012) 1119–1135.
[47] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, M. Shah, Deep learning human mind for automated visual classification, in: Proceedings of CVPR, 2017, pp. 6809–6817.
[48] M.J. Huiskes, M.S. Lew, The mir flickr retrieval evaluation, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, ACM, 2008, pp. 39–43.
[49] C. Du, C. Du, L. Huang, H. He, Reconstructing perceived images from human brain activities with Bayesian deep multiview learning, IEEE Trans. Neural Netw. Learning Syst. (2018).
[50] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv:1312.6114 (2013).
[51] L. Chongxuan, T. Xu, J. Zhu, B. Zhang, Triple generative adversarial nets, in: Advances in Neural Information Processing Systems, 2017, pp. 4088–4098.
[52] C.-K. Yeh, W.-C. Wu, W.-J. Ko, Y.-C.F. Wang, Learning deep latent space for multi-label classification., in: Proceedings of AAAI, 2017, pp. 2838–2844.
[53] X. Kong, M.K. Ng, Z.-H. Zhou, Transductive multilabel learning via label set propagation, IEEE Trans. Knowl. Data Eng. 25 (3) (2013) 704–719.
[54] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks, in: Proceedings of ICCV, 2017, pp. 5907–5915.

Dan Li received the M.S. degree in mathematics from the ShangHai University, Shanghai, China, in 2017. She is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Sciences. Her current research interests include machine learning and computational neuroscience.

Changde Du received the M.S. degree in data mining from the University of Chinese Academy of Sciences, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Sciences. His current research interests include machine learning, Bayesian statistics and brain-inspired intelligence.

Huiguang He received the Ph.D. degree (with honor) in pattern recognition and intelligent systems from Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He is currently a full professor with CASIA. His research interests include pattern recognition, medical image processing, and brain computer interface (BCI).