DOUBLY SEMI-SUPERVISED MULTIMODAL ADVERSARIAL LEARNING FOR CLASSIFICATION, GENERATION AND RETRIEVAL

Changde Du[1,2], *Changying Du*[3], *Huiguang He*[1,2,4,*]

[1]Research Center for Brain-Inspired Intelligence & NLPR, CASIA, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Huawei Noah's Ark Lab, Beijing, China
[4]Center for Excellence in Brain Science and Intelligence Technology, CAS, Beijing, China
{duchangde, ducyatict}@gmail.com, huiguang.he@ia.ac.cn

ABSTRACT

Learning over incomplete multi-modality data is a challenging problem with strong practical applications. Most existing multi-modal data imputation approaches have two limitations: (1) they are unable to accurately control the semantics of imputed modalities; and (2) without a shared low-dimensional latent space, they do not scale well with multiple modalities. To overcome the limitations, we propose a novel doubly semi-supervised multi-modal learning framework (DSML) with a modality-shared latent space and modality-specific generators, encoders and classifiers. We design novel softmax-based discriminators to train all modules adversarially. As a unified framework, DSML can be applied in multi-modal semi-supervised classification, missing modality imputation and fast cross-modality retrieval tasks simultaneously. Experiments on multiple datasets demonstrate its advantages.

Index Terms— Semi-supervised learning, multi-modal learning, incomplete data, adversarial learning

1. INTRODUCTION

Advances in sensor technologies lead to an increasing interest in acquiring and modeling data with multiple modalities. Since data from each modality can complement each other, combining the useful information from multiple modalities can significantly improve the prediction performance in various applications, such as emotion recognition [1], object recognition [2], etc. It is notable that most prior works on multi-modal learning make a common assumption that all training samples are with complete modalities and corresponding labels. Nevertheless, this assumption is excessive in practice, as (1) the data collection process may generate data points with missing modalities due to unforeseeable sensor malfunction or configuration issues; (2) in some applications (e.g., brain decoding [3]), the acquisition of multi-modality data is very expensive, while single-modality data is sufficient; and (3) the data labeling procedure requires lots of manual efforts, and hence only a small set of labeled samples is available in most cases.

To address the above incomplete data issues, some cross-modality data imputation/generation methods have been proposed recently [4, 5]. E.g., [4] developed a multi-view adversarially learned inference model, which formulates the modality imputation problem as a cross-view encoding-decoding task. However, these current solutions still have limitations. First, they are unable to accurately control the semantics of imputed modalities. Modality imputation is typically regarded as an unsupervised translation task [6], where we don't know what are the semantics of the translated result. How to make full use of the available semantic information (e.g., category labels) to guide the imputation process still remains challenging. Second, they do not scale well with multiple modalities. Most of the existing methods need to build bidirectional translators for any two modalities. Therefore, for n modalities, $n(n-1)$ translators would be needed. This quickly becomes unfeasible as the number of modalities increases.

In this paper, we model the statistical relationships between modalities by using modality-specific generators with a modality-shared latent space. Such a latent space can be beneficial in many ways, e.g., (1) translating data from one modality to another through the shared latent space is more scalable and efficient (only $2n$ low-dimensional mappings are needed) than direct translations between high-dimensional modalities; (2) once trained, we can perform efficient similarity computation in this low-dimensional space for fast cross-modality retrieval; and (3) fully-paired faking samples generated from the shared latent space may be used to augment the training set. To control the semantics of imputated/generated samples, we divide the shared latent representation into two parts $(\boldsymbol{c}, \boldsymbol{z})$, where $\boldsymbol{c}$ contains the designated semantic classes (category labels), and $\boldsymbol{z}$ encodes the styles. To disentangle $\boldsymbol{c}$ from $\boldsymbol{z}$, we further build the modality-specific classifiers and encoders to minimize the reconstruction error of $\boldsymbol{c}$ and $\boldsymbol{z}$ in the latent space, respectively. We finally design novel softmax-based discriminators to train all modules adversarially. In this way, the proposed DSML framework can not only utilize all available data flexibly, but also leverage the augmented information from the generated

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00011

data with high controllability. Experiments on multi-modal semi-supervised classification, missing modality generation and cross-modality retrieval tasks demonstrate its advantages.

2. RELATED WORK

Generation and inference. Adversarially learned inference (ALI) [7] and BiGAN [8] proposed an adversarial method for the joint distribution learning of data and latent code. They train a generation network and an inference network jointly, which can produce high-quality samples for both data and latent code. The ALICE model [9] adopted a similar idea, but with conditional entropy regularization, which was basically equivalent to the cycle-consistency principle in CycleGAN [6]. Recently, [4] proposed a multi-view ALI (MALI) model for cross-domain joint distribution matching. Our DSML model also draws inspirations from ALI for modality generation and latent code inference.

Semi-supervised cross-modality translation. Though ALI/BiGAN was originally designed for data generation and latent code inference, recent works [10, 11, 9, 12] extended it to cross-modality translations between two real data domains. Triangle GAN [11] is a semi-supervised framework that can be used to learn bi-directional mappings between domain $\boldsymbol{x}$ and domain $\boldsymbol{y}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ can be two data modalities, or data and the corresponding category label. Similar works includes TripleGAN [10] and JointGAN [12]. Though encouraging cross-modality translation results are reported, these models lack a mechanism to perform latent code inference, which is useful for many tasks.

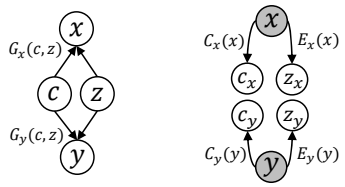
Semi-supervised multi-modal classification. Instead of using GANs, there also exists work that uses variational autoencoders (VAEs) [13] for semi-supervised multi-modal learning [1, 14]. E.g., [1] proposed a semi-supervised incomplete multi-view VAE model (SiMVAE) by treating the missing labels/modalities as latent variables and infer them for unlabeled/incomplete data. Our DSML model is distinct from SiMVAE in important ways, e.g., we separate the latent variables into two disentangled parts. Further, SiMVAE only focuses on the task of multi-modal classification, while DSML can be applied to a wide range of applications.

3. METHODOLOGY

We consider a doubly semi-supervised learning (SSL) setting, where both labels and modalities are incomplete. For a given instance, we assume $\boldsymbol{x}$ denotes one modality, $\boldsymbol{y}$ denotes the other modality, and $\boldsymbol{c}$ denotes its category label.

3.1. Overview

Fig. 1 illustrates the proposed DSML framework. Our key assumption is that the two modalities $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same instance can be generated via two modality-specific generators G_x and G_y with a shared low-dimensional latent space, respectively (cf. Fig. 1a). To accurately control the semantics of generated modalities, we separate the shared latent representation into two independent parts $(\boldsymbol{c}, \boldsymbol{z})$, where $\boldsymbol{c}$ contains the designated semantics, and $\boldsymbol{z}$ encodes other factors of variation. Note that, without specific constrains on the generative processes, $\boldsymbol{c}$ and $\boldsymbol{z}$ might be entangled in training phase. To overcome this, for each modality we build two separate inference networks, in which one acts as a classifier and the other as a common encoder (cf. Fig. 1b). Taking modality $\boldsymbol{x}$ as example, the classifier C_x and the encoder E_x define two conditional distributions $p_{c_x}(\boldsymbol{c}|\boldsymbol{x})$ and $p_{e_x}(\boldsymbol{z}|\boldsymbol{x})$ that are trained to approximate the true posteriors $p(\boldsymbol{c}|\boldsymbol{x})$ and $p(\boldsymbol{z}|\boldsymbol{x})$. In other words, C_x and E_x are trained to minimize the reconstruction error of $\boldsymbol{c}$ and $\boldsymbol{z}$ in the latent space, respectively. Similar strategy is adopted for $\boldsymbol{y}$. We elaborate the generative and inference processes in supplementary materials (SM) section A.



(a) G_x, G_y (b) C_x, C_y, E_x, E_y

Fig. 1. Overview of DSML framework: (a) the generators G_x, G_y; (b) the classifiers C_x, C_y and the encoders E_x, E_y. The grey and white units represent the observed and latent variables, respectively.

In practice, all the above modules are implemented as deep neural networks (DNNs), and their architectures depend on specific applications, such as deconvolutional neural network for image generation. Below, we show how to optimize all modules jointly by using adversarial learning.

3.2. Jointly adversarial learning

3.2.1. The pairs of modality and its latent code.

Following the ALI framework, we first construct an adversarial game to match the distributions of two different factorizations: $p_{g_x}(\boldsymbol{x}, \boldsymbol{z}) = \int p_{g_x}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{z})p(\boldsymbol{c})p(\boldsymbol{z})d\boldsymbol{c}$ and $p_{e_x}(\boldsymbol{x}, \boldsymbol{z}) = p_{e_x}(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})$. Specifically, the objective is to find the Nash equilibrium for the following minimax game:

$$\min_{G_x, E_x} \max_{D_{xz}} \mathcal{L}_{\boldsymbol{xz}} = \mathbb{E}_{(\tilde{\boldsymbol{x}}, \boldsymbol{z}) \sim p_{g_x}(\boldsymbol{x}, \boldsymbol{z})} \big[\log D_{xz}(\tilde{\boldsymbol{x}}, \boldsymbol{z}) \big] + \mathbb{E}_{(\boldsymbol{x}, \tilde{\boldsymbol{z}}) \sim p_{e_x}(\boldsymbol{x}, \boldsymbol{z})} \big[\log \big(1 - D_{xz}(\boldsymbol{x}, \tilde{\boldsymbol{z}}) \big) \big], \tag{1}$$

where the discriminator D_{xz} is trained to distinguish pairs $(\tilde{\boldsymbol{x}}, \boldsymbol{z}) \sim p_{g_x}(\boldsymbol{x}, \boldsymbol{z})$ from those come from $p_{e_x}(\boldsymbol{x}, \boldsymbol{z})$, which helps G_x and E_x to produce high-quality data samples and latent codes. Eq. (1) reaches optimum if and only if $p_{g_x}(\boldsymbol{x}, \boldsymbol{z}) = p_{e_x}(\boldsymbol{x}, \boldsymbol{z})$, and we have similar formulation for modality $\boldsymbol{y}$ (see SM section B)

To keep the style information (background, color, etc.) to be fully captured by $\boldsymbol{z}$, without being entangled with $\boldsymbol{c}$, here we introduce the reconstruction losses to $\boldsymbol{z}$ from the generated data $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$. Assume $\hat{\boldsymbol{z}}_x \sim p_{e_x}(\boldsymbol{z}|\tilde{\boldsymbol{x}})$ and $\hat{\boldsymbol{z}}_y \sim p_{e_y}(\boldsymbol{z}|\tilde{\boldsymbol{y}})$ denote the latent code reconstructions via $(\boldsymbol{c}, \boldsymbol{z}) \rightarrow \tilde{\boldsymbol{x}} \rightarrow \hat{\boldsymbol{z}}_x$ and $(\boldsymbol{c}, \boldsymbol{z}) \rightarrow \tilde{\boldsymbol{y}} \rightarrow \hat{\boldsymbol{z}}_y$, respectively. Then our idea can be formulated as

$$\min_{G_x, G_y, E_x, E_y} \mathcal{R}_{\boldsymbol{z}} = \mathbb{E}_{\boldsymbol{c}, \boldsymbol{z}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \hat{\boldsymbol{z}}_x, \hat{\boldsymbol{z}}_y} \big[\|\hat{\boldsymbol{z}}_x - \boldsymbol{z}\|_2 + \|\hat{\boldsymbol{z}}_y - \boldsymbol{z}\|_2 \big].$$

Intuitively, minimizing $\mathcal{R}_{\boldsymbol{z}}$ yields small $\|E_x(G_x(\boldsymbol{c}_1, \boldsymbol{z})) - E_x(G_x(\boldsymbol{c}_2, \boldsymbol{z}))\|_2$ and $\|E_y(G_y(\boldsymbol{c}_1, \boldsymbol{z})) - E_y(G_y(\boldsymbol{c}_2, \boldsymbol{z}))\|_2$, $\forall \boldsymbol{c}_1, \boldsymbol{c}_2 \sim p(\boldsymbol{c})$, which indicates $\boldsymbol{z}$ is disentangled from $\boldsymbol{c}$. Such a treatment can also be interpreted as applying the cycle-consistency principle [6] in the latent space. Besides, $\tilde{\boldsymbol{z}}_x \sim p_{e_x}(\boldsymbol{z}|\boldsymbol{x})$ and $\tilde{\boldsymbol{z}}_y \sim p_{e_y}(\boldsymbol{z}|\boldsymbol{y})$ inferred from the observable paired data $(\boldsymbol{x}, \boldsymbol{y})$ can be used to improve the training of encoders before G_x and G_y are able to generate high-quality data pairs $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$, i.e.,

$$\min_{G_x, G_y, E_x, E_y} \mathcal{R}^*_{\boldsymbol{z}} = \mathcal{R}_{\boldsymbol{z}} + \mathbb{E}_{\boldsymbol{x}, \boldsymbol{y}, \tilde{\boldsymbol{z}}_x, \tilde{\boldsymbol{z}}_y} \big[\|\tilde{\boldsymbol{z}}_x - \tilde{\boldsymbol{z}}_y\|_2 \big]. \quad (2)$$

It can be shown that minimizing $\mathcal{R}^*_{\boldsymbol{z}}$ w.r.t. the encoders E_x and E_y will not change the equilibriums of $\mathcal{L}_{\boldsymbol{xz}}$ and $\mathcal{L}_{\boldsymbol{yz}}$ [15].

3.2.2. The pairs of modality and its semantic label

In semi-supervised classification, the goal is to ensure that the joint distributions characterized by the generator and the classifier both converge to the empirical joint distribution. Taking modality $\boldsymbol{x}$ as example, it is required to match the distributions of joint pairs $(\boldsymbol{x}, \boldsymbol{c})$ drawn from $p_1(\boldsymbol{x}, \boldsymbol{c}) = \int_{\boldsymbol{z}} p_{g_x}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{z}) p(\boldsymbol{c}) p(\boldsymbol{z}) d\boldsymbol{z}$, $p_2(\boldsymbol{x}, \boldsymbol{c}) = p(\boldsymbol{x}) p_{c_x}(\boldsymbol{c}|\boldsymbol{x})$ and $p_3(\boldsymbol{x}, \boldsymbol{c}) = p(\boldsymbol{x}, \boldsymbol{c})$, respectively. Note that $p_3(\boldsymbol{x}, \boldsymbol{c})$ is simply the empirical joint distribution. Naively, one can employ two binary discriminators to distinguish these three kinds of joint pairs [11]. However, this may result in possibly conflicting (real vs. fake) assessments. To distinguish these three kinds of joint pairs consistently, here the discriminator D_{xc} is implemented as a neural network with 3-way softmax on the top layer, i.e., $\sum_{k=1}^{3} D_{xc}(\boldsymbol{x}, \boldsymbol{c})[k] = 1$ and $D_{xc}(\boldsymbol{x}, \boldsymbol{c})[k] \in (0, 1)$, where $D_{xc}(\boldsymbol{x}, \boldsymbol{c})[k]$ is an entry of $D_{xc}(\boldsymbol{x}, \boldsymbol{c})$. The minimax objective is given by

$$\min_{G_x, C_x} \max_{D_{xc}} \mathcal{L}_{\boldsymbol{xc}} = \sum_{k=1}^{3} \mathbb{E}_{p_k(\boldsymbol{x}, \boldsymbol{c})} \Big[\log D_{xc}(\boldsymbol{x}, \boldsymbol{c})[k] \Big]. \quad (3)$$

Compared with using two binary discriminators, our softmax-based discriminator can be considered as sharing the parameters between two binary discriminators except the top layer, thus reducing the number of parameters.

Proposition 1. *The equilibrium for the minimax objective $\mathcal{L}_{\boldsymbol{xc}}$ is achieved if and only if $p_1(\boldsymbol{x}, \boldsymbol{c}) = p_2(\boldsymbol{x}, \boldsymbol{c}) = p_3(\boldsymbol{x}, \boldsymbol{c})$ with the optimal discriminator $D^*_{xc}(\boldsymbol{x}, \boldsymbol{c})[k] = \frac{1}{3}$.*

The proof is provided in SM section C. This conclusion essentially motivates our design for semi-supervised classification. However, it turns out to be very difficult to achieve the desired convergence in practice, because there is little supervision to tell the generator $p_{g_x}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{z})$ what $\boldsymbol{c}$ essentially represents. As a result, G_x might generate low-quality samples that are not well aligned with their conditions. To address the issues, we force the classifier $p_{c_x}(\boldsymbol{c}|\boldsymbol{x})$ to reconstruct $\boldsymbol{c}$ in the latent space by introducing the following regularizer,

$$\min_{G_x, C_x} \mathcal{R}_{\boldsymbol{xc}} = \mathbb{E}_{p_3(\boldsymbol{x}, \boldsymbol{c})}[-\log p_{c_x}(\boldsymbol{c}|\boldsymbol{x})] + \mathbb{E}_{p_1(\boldsymbol{x}, \boldsymbol{c})}[-\log p_{c_x}(\boldsymbol{c}|\boldsymbol{x})]. \quad (4)$$

Intuitively, it aims to minimize the standard classification loss (i.e., cross-entropy loss) on both real and generated data, thus assigning the semantic labels to variable $\boldsymbol{c}$. Once G_x can generate high-quality samples that respect the label $\boldsymbol{c}$, the generated samples $(\boldsymbol{x}, \boldsymbol{c}) \sim p_1(\boldsymbol{x}, \boldsymbol{c})$ can be reused to augment the predictive power of C_x, which proves effective in SSL [10, 15]. Similar strategy is adopted for modality $\boldsymbol{y}$, and the adversarial game $\mathcal{L}_{\boldsymbol{yc}}$ and the regularizer $\mathcal{R}_{\boldsymbol{yc}}$ are provided in SM section D. Since we have a separate classifier for each modality, we use the classifier fusion strategy to deal with the classification of multi-modality data. E.g., the predicted label for two-modality data can be written as: label $=$ softmax$(O_x + O_y)$, where O_x and O_y are the outputs before softmax layer of each classifiers, respectively.

3.2.3. The pair of two modalities

Model performance can be improved by introducing an additional discriminator D_{xy} to drive $p_1(\boldsymbol{x}, \boldsymbol{y})$, $p_2(\boldsymbol{x}, \boldsymbol{y})$ and $p_3(\boldsymbol{x}, \boldsymbol{y})$ to concentrate on the empirical distribution $p_4(\boldsymbol{x}, \boldsymbol{y})$, where $p_1(\boldsymbol{x}, \boldsymbol{y})$, ..., $p_4(\boldsymbol{x}, \boldsymbol{y})$ denote the distributions of joint pairs $(\boldsymbol{x}_{\text{fake}}, \boldsymbol{y}_{\text{fake}})$, $(\boldsymbol{x}_{\text{fake}}, \boldsymbol{y}_{\text{real}})$, $(\boldsymbol{x}_{\text{real}}, \boldsymbol{y}_{\text{fake}})$ and $(\boldsymbol{x}_{\text{real}}, \boldsymbol{y}_{\text{real}})$, respectively (see SM section E). Similar to Eq. (3), here D_{xy} can be a 4-way softmax-based discriminator, and the minimax objective is given by

$$\min_{\boldsymbol{G}, \boldsymbol{E}, \boldsymbol{C}} \max_{D_{xy}} \mathcal{L}_{\boldsymbol{xy}} = \sum_{k=1}^{4} \mathbb{E}_{p_k(\boldsymbol{x}, \boldsymbol{y})} \Big[\log D_{xy}(\boldsymbol{x}, \boldsymbol{y})[k] \Big], \quad (5)$$

where $\boldsymbol{G} = \{G_x, G_y\}$, $\boldsymbol{E} = \{E_x, E_y\}$ and $\boldsymbol{C} = \{C_x, C_y\}$. Note that the empirical distribution $p_4(\boldsymbol{x}, \boldsymbol{y})$ might be biased because it is only characterized by the few paired samples. Fortunately, once G_x and G_y can generate high-quality samples, we can reuse the generated samples $(\boldsymbol{x}, \boldsymbol{y}) \sim p_1(\boldsymbol{x}, \boldsymbol{y})$ to augment the paired samples.

3.2.4. Full objective function

The overall objective of the proposed DSML framework is

$$\min_{\boldsymbol{G}, \boldsymbol{E}, \boldsymbol{C}} \max_{\boldsymbol{D}} \mathcal{L}_{\text{DSML}} = \mathcal{L}_{\boldsymbol{xz}} + \mathcal{L}_{\boldsymbol{yz}} + \mathcal{L}_{\boldsymbol{xc}} + \mathcal{L}_{\boldsymbol{yc}} + \mathcal{L}_{\boldsymbol{xy}} + \mathcal{R}_{\boldsymbol{xc}} + \mathcal{R}_{\boldsymbol{yc}} + \lambda \mathcal{R}^*_{\boldsymbol{z}}, \quad (6)$$

where $\boldsymbol{D} = \{D_{xz}, D_{yz}, D_{xc}, D_{yc}, D_{xy}\}$. Note that, various methods have been proposed to improve and stabilize the training of GAN, such as Wasserstein GAN [16], etc. Our framework is orthogonal to these methods, which could also be used to improve the training process.

3.3. Extensions to multiple modalities

The above formulation scales linearly with the number of modalities. When scaled to n-modalities, it is easy to see we need n generators/encoders/classifiers. As for D_{xy} in this case, we highlight the importance of our softmax-based design. Specifically, the input of this discriminator is the joint of all modalities, and we consider the following $(n+2)$ kinds of input: 1) all modalities are true, 2) all modalities are fake, and 3) one of the n modalities is fake and all other modalities are true. Finally the output layer of this discriminator is similar as a $(n+2)$-way classifier.

Diferences with StarGAN and RadialGAN. StarGAN [17] proposes a framework for multi-domain translation, where each domain represents a different attribute/class. However, without latent space, its single-generator design is not suitable when there are significant differences between domains. RadialGAN [18] performs data augmentation for multiple datasets simultaneously through a shared low-dimensional space. But its shared space is just used to align the distributions of different datasets rather than the multiple modalities of the same instance. In contrast, our DSML framework focuses on the semi-supervised classification of multi-modality data and the imputation of missing modality through the modality-shared latent space. Therefore, DSML significantly differs from StarGAN and RadialGAN w.r.t. architectures and applications.

4. EXPERIMENTS

For all experiments, the latent variables $\boldsymbol{z}$ are drawn from a $\mathcal{N}(0, \mathbf{I})$ distribution, with the dimension set to 100. We empirically set the regularization parameter $\lambda = 10$. The Adam optimizer [19] with learning rate 0.0002 is used for optimization. Many details including the network architectures and additional experiment results are given in the SM.

4.1. Multi-modal semi-supervised classification

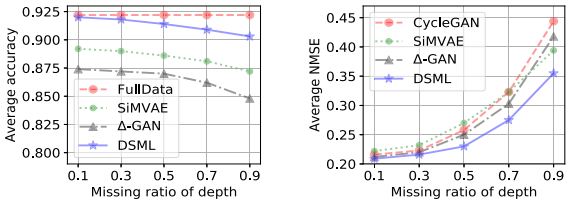
We deploy DSML for RGB-D object recognition on the RGB-D object dataset [20]. This dataset contains 41,877 RGB-D images capturing 300 objects from 51 categories. We regard the color and depth images as two different modalities, and resize them to 64×64 pixels. Next, we interpolate the missing values in the depth images with the mean of 5×5 nearest values. Furthermore, we extend the single channel depth images to three channels with surface normal processing, which is consistent with [21]. For each of the 10 random splits of training/test set provided by [20], we randomly labeled 5% samples (every class has equal number of labels) of the training set, and remain the rest unlabeled.

We first compared DSML with strong competitors in the case of complete modalities. All the experiments in Table 1 were repeated 10 times based on the given 10 different training/test splits, and the average results were reported for comparison. For the competitors, we considered the same setups (network structure, learning rate, etc.) as our DSML to keep comparisons fair. For unimodal algorithms, we evaluated their performance on each modality and the concatenation of two modalities, respectively. We note that the average accuracy of DSML significantly surpasses the competitors in multi-modality case. This is because our method can match the joint distribution of each modality and its labels adversarially, and the shared latent space can effectively capture the comment representation of both modality.

Table 1. Comparisons of semi-supervised classification accuracies (%) on partially labeled RGB-D dataset (without missing modality).

Methods	Algorithms	RGB	Depth	RGB-D
Unimodal baselines	M2 [22]	85.6±1.6	72.0±1.7	86.4±1.6
	SDGM [23]	85.8±1.5	75.4±1.7	86.7±1.5
	TripleGAN [10]	86.4±1.7	82.9±1.8	87.2±1.8
	Δ-GAN [11]	86.5±1.8	82.6±1.9	87.6±1.7
Multi-modal baselines	CT+SVM [21]	-	-	83.7±1.3
	DCNN [2]	-	-	89.2±1.3
	AMGL [24]	-	-	86.4±1.5
	SMVAE [1]	-	-	89.5±1.8
Proposed	DSML	-	-	**92.2**± 1.7

To simulate the doubly semi-supervised setting, we randomly selected a fraction of instances (from both labeled and unlabeled training data) to be unpaired examples, i.e., they are described by only one modality. We varied the missing ratio of depth modality from 0.1 to 0.9 with an interval of 0.2, while no missing modality in test sets. We compared DSML with SiMVAE [1], CycleGAN [6] and Δ-GAN in Fig. 2, where FullData means DSML with complete modalities. For Δ-GAN, we first estimated the missing modalities, and then conducted semi-supervised classification using Δ-GAN again. We measure the imputation errors (cf. Fig. 2b) using Normalized Mean Squared Error (NMSE), $\text{NMSE} = \frac{\|\mathbf{X}-\hat{\mathbf{X}}\|_F}{\|\mathbf{X}\|_F}$, where $\mathbf{X}$ and $\hat{\mathbf{X}}$ are the original and the recovered data matrices, respectively. $\|\cdot\|_F$ demotes the Frobenious norm.



(a) Recognition accuracies (b) Imputation errors

Fig. 2. Comparisons with different missing ratios.

From Fig. 2, we see that DSML has been successful even with a high missing ratio. With missing ration lower than 0.5, DSML roughly reaches FullData's performance. Moreover, the semi-supervised imputation methods DSML and SiMVAE outperforms the unsupervised imputation methods CycleGAN

and Δ-GAN when we have few paired data. This demonstrates that the label information also plays an important role in missing modality imputation. Except for utilizing label information, DSML can synthesize arbitrary number of paired data, which can also boost its performance.

4.2. Missing modality imputation

We evaluate the controllability of DSML in missing modality imputation tasks based on two publicly available datasets: (i) MNIST-to-MNIST-transpose [11], where two modalities are the MNIST images and their corresponding transposed ones. (ii) ImageNet-EEG [25], where two modalities are the ImageNet images and the evoked Electroencephalogram (EEG) signals (see SM for more data descriptions). Furthermore, we selected all images, belonging to the given image classes, from the ImageNet database as unpaired image data. All images are resized to 64×64 pixels. We randomly select 90% of the paired data for training, while the rest 10% for test (on which we impute the images using given EEG data).

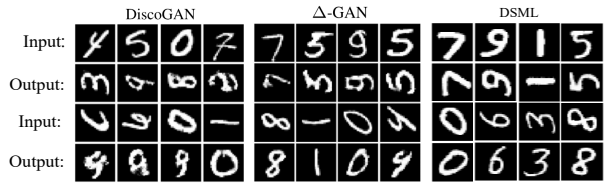


Fig. 3. Modality imputation results. DiscoGAN [26] does not need paired data, and we use 10% paired samples for Δ-GAN and DSML.

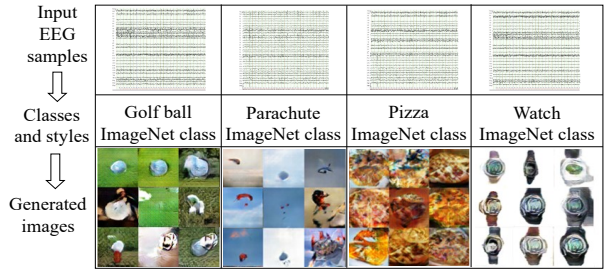


Fig. 4. Predicted images from EEG modality by using our DSML.

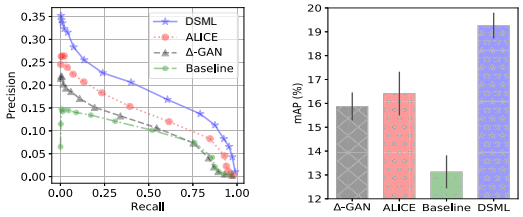
On both datasets, we assume only the paired samples have corresponding class labels in training phase. Results are shown in Figs. 3 and 4. We observe that DSML generally recovers missing modality with higher visual quality and strictly following the intrinsic semantics. For supporting quantitative evaluation, we use the pre-trained gold-standard classifier (Inception-v3 [27] for ImageNet-EEG) to classify the imputed images, and use the labels of test data as ground truth to calculate the accuracy. Results are displayed in Table 2, averaged over 5 runs with different random data splits. DSML achieves significantly better performance than Triple GAN and Δ-GAN, which indicates its effectiveness in controlling the semantics of imputed modalities.

Table 2. Classification accuracy (%) of the imputed images.

Algorithms	MNIST-to-MNIST-transpose			ImageNet-EEG
	100 paired	1000 paired	All paired	90% paired
DiscoGAN	-	-	15.00 ± 0.20	-
Triple GAN	63.79 ± 0.85	84.93 ± 1.63	86.70 ± 1.52	54.85 ± 1.21
Δ-GAN	83.20 ± 1.88	88.98 ± 1.50	93.34 ± 1.46	66.02 ± 1.09
DSML	**98.67** ± 1.43	**99.02** ± 1.41	**99.23** ± 1.30	**81.24** ± 0.98

4.3. Cross-modality retrieval

Another important feature of the proposed DSML is that its latent space can be used for cross-modality retrieval. Since the latent representation was separated into $\boldsymbol{c}$ (semantic class) and $\boldsymbol{z}$ (style) two parts, we conduct cross-modality retrieval on ImageNet-EEG dataset by considering these two aspects. Specifically, for each class, we randomly select 5 EEG samples from the test set as queries. For each selected EEG query, we find its $N \in \{2^0, 2^1, ..., 2^{15}\}$ nearest neighbors in the latent space of DSML, and return the corresponding images of these neighbors. In similarity search, we first match the label vector $\boldsymbol{c}$, and then perform ranking w.r.t. $\boldsymbol{z}$ based on Euclidean distance for samples with matched $\boldsymbol{c}$, and finally perform ranking w.r.t. $\boldsymbol{z}$ for samples with unmatched $\boldsymbol{c}$. Fig. 5 shows the results, where mean average precision (mAP) computes the area under the entire precision-recall curve and evaluates the overall retrieval performance. When evaluating these metrics, the number of ground truth neighbors is set to 100. The baseline method is that: given an EEG query, we first find its nearest neighbor in the paired EEG dataset, and then perform image retrieval based on the corresponding image representation of that nearest neighbor EEG instance.



(a) precision-recall curves (b) mean average precision

Fig. 5. Comparison of different cross-modal retrieval methods on the ImageNet-EEG dataset. Results were averaged over six subjects.

It is clear that three joint distribution matching methods (DSML, ALICE [9] and Δ-GAN) consistently outperform the baseline method. Further, DSML beats state-of-the-art methods ALICE and Δ-GAN. We ascribe this to the fact that DSML can synthesize arbitrary number of paired data based on the shared latent space, which contributes to the learning of modality mappings. Finally, the considered competitors can only perform similarity search in original high-dimensional space, which is inefficient. A natural advantage of DSML is its ability to conduct cross-modality retrieval in the latent space, which reduces the computational complexity effectively.

5. CONCLUSION

We focus on the issues of incomplete multi-modal learning in a doubly semi-supervised setting, where both labels and modalities are incomplete. The proposed DSML is a new framework for multi-modal jointly adversarial learning. The disentangled latent space allows DSML to accurately control the semantics of imputed modalities and synthesize arbitrary number of samples with complete labels and modalities to augment the training set. Experimental results demonstrated the superiorities of our framework over many state-of-the-art competitors.

6. ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (No. 91520202, 61602449), CAS Scientific Equipment Development Project under Grant YJKYYQ20170050, Beijing Municipal Science&Technology Commission (Z181100008918010), Youth Innovation Promotion Association CAS and Strategic Priority Research Program of CAS.

7. REFERENCES

[1] Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He, "Semi-supervised deep generative modelling of incomplete multi-modality emotional data," in *ACM MM*, 2018.

[2] Yanhua Cheng, Xin Zhao, Rui Cai, Zhiwei Li, Kaiqi Huang, and Yong Rui, "Semi-supervised multimodal deep learning for RGB-D object recognition," in *IJCAI*, 2016.

[3] Changde Du, Changying Du, Lijie Huang, and Huiguang He, "Reconstructing perceived images from human brain activities with Bayesian deep multiview learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[4] Changying Du, Changde Du, Xingyu Xie, Chen Zhang, and Hao Wang, "Multi-view adversarially learned inference for cross-domain joint distribution matching," in *SIGKDD*, 2018.

[5] Tran Luan, Xiaoming Liu, Jiayu Zhou, and Rong Jin, "Missing modalities imputation via cascaded residual autoencoder," in *CVPR*, 2017.

[6] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[7] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville, "Adversarially learned inference," in *ICLR*, 2017.

[8] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell, "Adversarial feature learning," in *ICLR*, 2017.

[9] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin, "Alice: Towards understanding adversarial learning for joint distribution matching," in *NIPS*, 2017.

[10] Chongxuan Li, Kun Xu, Jun Zhu, and Bo Zhang, "Triple generative adversarial nets," in *NIPS*, 2017.

[11] Zhe Gan, Liqun Chen, Weiyao Wang, Yunchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin, "Triangle generative adversarial networks," in *NIPS*, 2017.

[12] Yunchen Pu, Shuyang Dai, Zhe Gan, Weiyao Wang, Guoyin Wang, Yizhe Zhang, Ricardo Henao, and Lawrence Carin, "JointGAN: Multi-domain joint distribution learning with generative adversarial nets," in *ICML*, 2018.

[13] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[14] Masahiro Suzuki and Yutaka Matsuo, "Semi-supervised multimodal learning with deep generative models," 2018.

[15] Zhijie Deng, Hao Zhang, Xiaodan Liang, Luona Yang, Shizhen Xu, Jun Zhu, and Eric P. Xing, "Structured generative adversarial networks," in *NIPS*, 2017.

[16] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[17] Yunjey Choi, Minje Choi, and Munyoung Kim, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," .

[18] Jinsung Yoon, James Jordon, and Mihaela van der Schaar, "RadialGAN: Leveraging multiple datasets to improve target-specific predictive models using generative adversarial networks," in *ICML*, 2018.

[19] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *ICRA*, 2011, pp. 1817–1824.

[21] Yanhua Cheng, Xin Zhao, Kaiqi Huang, and Tieniu Tan, "Semi-supervised learning and feature evaluation for RGB-D object recognition," *Computer Vision and Image Understanding*, vol. 139, no. C, pp. 149–160, 2015.

[22] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, "Semi-supervised learning with deep generative models," in *NIPS*, 2014.

[23] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther, "Auxiliary deep generative models," in *ICML*, 2016.

[24] Feiping Nie, Jing Li, Xuelong Li, et al., "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *IJCAI*, 2016.

[25] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah, "Deep learning human mind for automated visual classification," in *CVPR*, 2017.

[26] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, 2017.

[27] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.