Deep Modality Invariant Adversarial Network for Shared Representation Learning

Kuniaki Saito, Yusuke Mukuta, Yoshitaka Ushiku
The University of Tokyo
{k-saito,mukuta,ushiku}@mi.t.u-tokyo.ac.jp

Tatsuya Harada
The University of Tokyo, RIKEN
harada@mi.t.u-tokyo.ac.jp

Abstract

In this work, we propose a novel method to learn the mapping to the common space wherein different modalities have the same information for shared representation learning. Our goal is to correctly classify the target modality with a classifier trained on source modality samples and their labels in common representations. We call these representations modality-invariant representations. Our proposed method has the major advantage of not needing any labels for the target samples in order to learn representations. For example, we obtain modality-invariant representations from pairs of images and texts. Then, we train the text classifier on the modality-invariant space. Although we do not give any explicit relationship between images and labels, we can expect that images can be classified correctly in that space. Our method draws upon the theory of domain adaptation and we propose to learn modality-invariant representations by utilizing adversarial training. We call our method the Deep Modality Invariant Adversarial Network (DeMIAN). We demonstrate the effectiveness of our method in experiments.

1. Introduction

Significant improvements have been made in classifying various modalities including images, texts, and videos, which use large-scale labeled datasets [28, 13, 23]. However, high labor costs are involved in collecting such a large amount of labeled samples.

Shared representation learning (SRL) is based on two modalities of information, namely, the source modality and target modality. During the training time, we are given paired source and target modality samples. Also, we are provided with labeled source modality samples although we do not have access to labeled target ones. In the training phase, we learn the mapping to the common space by using the paired samples, and then, under the common space, we train a classifier by using the labeled source samples. The goal is to classify the target samples using the learned classifier. For SRL, we have to consider learning mapping to the common space, where different modality samples have the same information, and a classifier is trained using source modality samples and their labels. If the target samples are correctly classified in the space, we do not need any labels for the target samples.

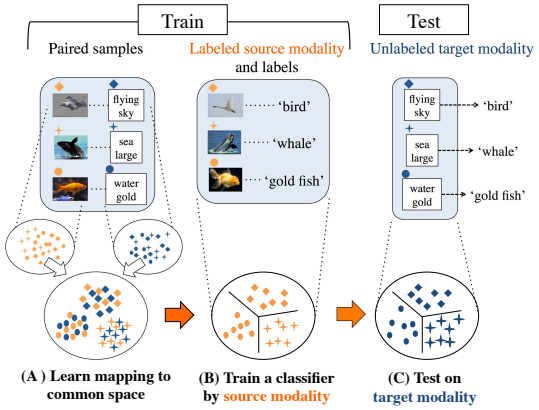


Figure 1. Illustration of our proposed method. We proposed a method that learns modality-invariant representations for shared representation learning. We can utilize the labeled source modality information to classify unlabeled target modality in this method. (A) We aim to learn modality-invariant representations by utilizing the relationship between paired samples and making the distributions similar. (B) We obtain decision boundaries from labeled source modality. (C) We can classify unlabeled target modality by the learned boundaries through modality-invariant representations.

Then, we propose a novel method that aims to learn representations from two modalities, which are interchangeable in a classification problem. We call such representations modality-invariant representations. We show the overview of our method in Fig. 1.

We define modality-invariant representations as representations that perform two functions. The first is that the representations include discriminative information. Learned representations must contain discriminative information to categorize them correctly. The second is that, under modality-invariant representations, the classifier trained on one modality can be transferred to the other modality.

We have to incorporate the following three properties in the common space to realize the functions above.

1. The paired samples are placed close together.
2. The distributions of different modalities match.
3. The projected samples are made dissimilar with each other.

Here, we provide a detailed explanation on why unlabeled modalities can be categorized correctly using modality-invariant representations.

The first and second properties are based on the domain adaptation (DA) theory. DA deals with samples generated from various domains, which are related but different from each other. DA mainly assumes the existence of two domains, the source and target domains. One aims to transfer the knowledge gained from labeled source samples to target samples, and classify the target samples accurately. David *et al*. [4] demonstrates that the expected error in target samples is bound by the summation of three terms:(i) the expected error on source domain; (ii) the divergence between source and target domains; and (iii) the minimum of the summation of the error in the source and target domains. The third term is considered to be very low when the representations are sufficiently discriminative. In our problem setting, we regard the source and target domains as the labeled and unlabeled modality, respectively.

The first property ensures that the first and third terms will be low. We can extract co-occurring information by utilizing paired samples. We think that the co-occurring information between modalities includes high-level information; thus, we can obtain discriminative representations.

The second property decreases the second term in the theory. Previous studies have shown that matching the distributions between different domains is effective in reducing the second term [4, 17, 9].

The third property is for avoiding bad representations. When considering only the first and second properties, the learned representations can be made very similar. If the learned representations are reduced to similar points, the first and second properties are satisfied, but contain little information. Therefore, we can state that the property that prevents the reduction is required.

One can think that projecting paired samples into exactly the same point will achieve distribution matching. However, the constraints of distribution matching are easier to satisfy than projecting paired samples into the sample point. Furthermore, considering the fact that our goal is to classify learned representations, we do not need to completely project into the same point.

In this paper, we propose a novel method for learning modality-invariant representations from paired modalities, which satisfy the three important properties above. We incorporate the idea of matching paired samples and making the distributions of two modalities similar in order to obtain the representations. For this purpose, we utilize adversarial training, namely a minimax two-player game involving a generator and discriminator. We call our proposed method the Deep Modality Invariant Adversarial Network (DeMIAN). Our contributions are as follows:

- We propose a novel method to learn interchangeable representations between different modalities for zero-shot learning.
- We demonstrate the effectiveness of our method through experiments.

2. Related Works

2.1. Multimodal Learning

Here, we focus on obtaining interchangeable representations between modalities. Ngiam *et al*. [21] used canonical correlation analysis (CCA) to learn the latent space between audio and video features, trained a classifier using only one modality, and tested it on the other modality. Sharma *et al*. [30] proposed a supervised extension of CCA, which utilized the label information along with the paired relationship. Our method focuses on the case where we have no access to complete training sets, namely, paired samples and the corresponding labels. In our algorithm, we propose to utilize the relationship between paired samples by minimizing the distance used in the CCA formulation. Moreover, we added a term that makes the distribution of different modalities similar. Thus, our model can be regarded as one that efficiently incorporates a modality-invariant factor with multimodal learning.

2.2. Domain Adaptation

In unsupervised DA, since one is provided with labeled source samples and unlabeled target samples, minimizing errors on source samples and domain divergence is required to obtain a good classifier in the target domain. For this purpose, many previous methods [9, 17, 1, 33] proposed the reduction of the divergence between the distributions of the source and target representations. Ganin *et al*. [9] introduced the idea of generative adversarial networks [10] for DA. They used adversarial training for domain-invariant feature extraction that identified the domain from which hidden features were generated. They trained two models to obtain domain-invariant representations: one was the main network classifier, a convolutional neural network (CNN); the other was a domain-classifier network, which distinguished the domain labels of hidden features extracted from the main network. To obtain domain-invariant discriminative representations, they trained the two networks simultaneously: the main CNN was trained to classify source samples correctly and to deceive the domain classifier, while the domain network classifier was trained to identify the

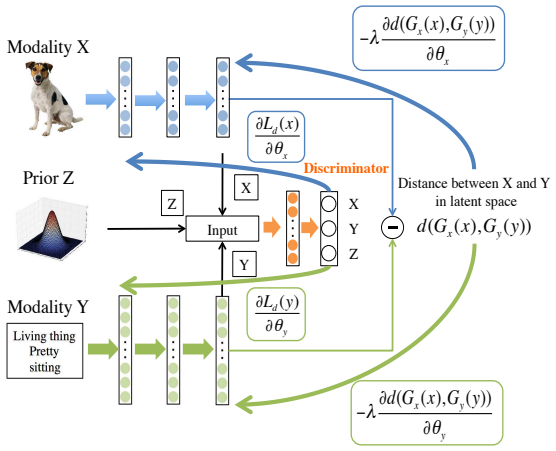


Figure 2. The proposed model: Deep Modality Invariant Adversarial Network. We aim to learn modality-invariant representations using a minimax two-player game involving a generator and discriminator. The discriminator tries to minimize the loss of predicting the modality of the input. The generator tries to minimize the distance between paired samples in the latent space and to maximize the error of the discriminator.

domain from which the hidden features were generated. Through this training, they observed that the distribution of different domains matched their hidden feature space.

3. Proposed Method

In this section, we describe the details of the proposed model. We first explain the problem setting and requirements for our model. Then, we discuss the components of our model for satisfying the requirements. Finally, we explain the learning procedure. We present an overview of our model in Fig. 2.

We are given two kinds of training sets: paired modal samples $\{(x_i, y_i)\}_{i=1}^n \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, and one modality with the corresponding labels $\{(y_i, t_i)\}_{i=1}^m \in \mathbb{R}^{d_y} \times \mathcal{C}$. Our goal is to learn mapping to a common space $G_x : \mathbb{R}^{d_x} \to \mathbb{R}^{d_z}$, $G_y : \mathbb{R}^{d_y} \to \mathbb{R}^{d_z}$ from $\{(x_i, y_i)\}_{i=1}^n$ and a classifier $h : \mathbb{R}^{d_z} \to \mathcal{C}$ from $\{(G_y(y_i), t_i)\}_{i=1}^m$, such that both classifiers $h \circ G_x$ and $h \circ G_y$ work well. We first learn mapping using paired samples $\{(x_i, y_i)\}_{i=1}^n$, then train a classifier by learned representations. We call the function G_x, G_y as generators. We denote the parameters of the generators as θ_x, θ_y, and denote the representations obtained from each generator as $f_x = G_x(x; \theta_x)$ and $f_y = G_y(y; \theta_y)$.

The following three properties are required to obtain modality-invariant representations as discussed in the Sec. 1,

1. We have to project samples into a common space by considering the relationship between paired samples.
2. We have to make the distribution of f_x, f_y similar. (We utilize the idea of DA)
3. We have to make each of f_x, f_y dissimilar.

For the first requirement, we obtain discriminative information using the relationship between paired samples. With regard to the second requirement, from the perspective of DA, we assume that we can train a classifier that works well for both modalities by making the distribution similar. To make the projected points informative, we need the third requirement. We have to train the generators that simultaneously satisfy the three requirements.

3.1. Learning Relationship between Modalities

For the first requirement, we define the objective for this matching as

$$J(\theta_x, \theta_y) = \sum_{i=1}^{n} d(G_x(x_i), G_y(y_i)) \tag{1}$$

where we use the l2-squared distance or cosine distance for $d(G_x(x_i), G_y(y_i))$, which can consider the matching of paired modality as used in CCA.

3.2. Modality-Discriminator to Match Distributions

For the second requirement, we have to measure the divergence between distributions f_x and f_y. However, divergence measurement is non-trivial, given that f_x and f_y are high dimensional, and that distributions change constantly as learning progresses. Hence, we utilize the modality-discriminator, D_d with the parameters θ_d. We can estimate divergence by looking at the loss of D_d, provided that D_d has been trained to discriminate between f_x and f_y. If the trained discriminators are deceived by the generated representations, then the distributions will match correctly.

Therefore, we seek the parameters θ_x and θ_y that maximize the loss of D_d, while simultaneously seeking the parameters θ_d that minimize the loss of D_d. This is the adversarial training method for our model. In other words, we seek to minimize the loss of $J(\theta_x, \theta_y)$, as well as the loss for adversarial training.

3.3. Gaussian Prior

The learned generator can always produce the similar points with the two training objectives mentioned above. The objectives do not include any terms that will make the generated points more dissimilar to each other. If the generator produces the similar points, the demands for the paired relationship will be met and the discriminator can be deceived easily. However, such a point does not have any discriminative information. A similar problem is reported in training generative adversarial networks [29].

In our work, we try to solve this problem by utilizing samples generated from Gaussian distributions. We input the samples generated from Gaussian distributions that have

Dataset	loss	activation	λ	image units	language units	batch size	learning rate
MNIST	L2-squared	ReLU	5.0	[392,1000,50]	[392,1000,50]	500	2.0×10^{-4}
Mir Flickr	L2-squared	ReLU	5.0	[3857,1000,200]	[2000,1000,200]	500	2.0×10^{-3}
SUN (Zero-Shot)	cosine distance	ReLU	10.0	[4096,2000,1000]	[102,2000,1000]	1000	2.0×10^{-4}
CUB-200-201	cosine distance	ELU	10.0	[4096,1000,1000]	[312,1000,1000]	1000	2.0×10^{-4}

Table 1. The architectures used for DeMIAN. Note that *image units* and *language units* mean the structure used for the network. In the experiment on MNIST, we input image features for both units.

the same dimensions as f_x and f_y, to the discriminator as well as f_x and f_y. Then, we train the discriminator to classify samples from the Gaussian distributions as modality z. That is, we expect f_x and f_y to follow Gaussian distributions. With this objective, the projected features will not contract into similar points, because the randomly generated samples from Gaussian distributions are dispersed. Our final objective for the generator consists of two terms: the loss of the deceiving discriminator in a three-class classification problem, and the loss of the paired relationship. We can avoid the contraction of representations by effectively utilizing the adversarial network structure.

3.4. Formulation of DeMIAN and its Optimization

Given the discussion above, the objective for our model can be written as follows:

$$\begin{aligned} z_i &\sim N(0, I_{z_d}) \\ L(\theta_x, \theta_y, \theta_d) &= \log\left(Pr(D_d(G_x(x_i)) = 1)\right) \\ &\quad - \log\left(Pr(D_d(G_y(y_i)) = 1)\right) \\ &\quad - \log\left(Pr(D_d(z_i) = 1)\right) \\ &\quad + \log\left(Pr(D_d(G_y(y_i)) = 2)\right) \\ &\quad + \log\left(Pr(D_d(z_i) = 3)\right) \\ E(\theta_x, \theta_y, \theta_d) &= J(\theta_x, \theta_y) - \lambda L(\theta_x, \theta_y, \theta_d). \end{aligned} \tag{2}$$

where λ is a parameter for balancing the loss of multimodal matching and adversarial training. We chose this parameter by validation split of each dataset. We assign the domain labels $1, 2, 3$ respectively for modality x,y,z. $Pr(D_d(G_x(x_i)) = 1)$ denotes the probability that generated features came from modality x, and likewise for other modalities. We seek network parameters by playing the following two-player minimax game,

$$\min_{\theta_x, \theta_y} \max_{\theta_d} E(\theta_x, \theta_y, \theta_d). \tag{3}$$

At the saddle point, the parameters θ_x, θ_y minimize both the modality classification loss and the loss for matching paired samples.

For the activation function, we used ReLU or ELU [8] and BatchNormalization (BN) [12] after the activation. BN is known to be highly effective for optimizing generative adversarial networks [26], and we confirmed that BN can also stabilize and improve the performance of our model. For the discriminator, we used ReLU for activation in all the experiments.

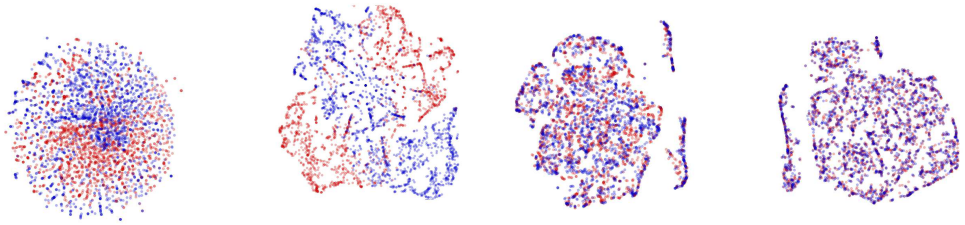
4. Experiment

We tested our model by classification for SRL and zero-shot learning. The difference between SRL and zero-shot learning is that we completely omitted the unseen target samples for zero-shot learning, while we used samples from all classes to train the model for SRL. For SRL, we used MNIST and Mir Flickr [11]. For zero-shot learning, we used SUN attribute [24] and Caltech-UCSD Birds-200-2011 (CUB-200-2011) [34], which are the benchmark image datasets for zero-shot learning. In all the experiments, we used Adam [14] for optimization of our model.We evaluated both deep and shallow models in SRL experiments. The effect of distribution matching will be seen in the shallow models. Note that the notation DeMIAN in our results refers to our proposed model with three layers, while MIAN refers to our model with two layers, namely the shallow model. MIAN includes linear and non-linear models, which we will describe clearly. We trained the logistic regression of learned representations for SRL and the multilayer-perceptron for zero-shot learning experiments. We show the architectures used to train DeMIAN in Table 1. The architectures include the activation function for the generator, loss function of $d(G_x(x_i), G_y(y_i))$, the value of λ, the structure of networks, batch size, and learning rate. We show the hyperparameters of MIAN in our supplementary material. In the SRL experiment, we compared our method to deep CCA (DCCA) [3] in addition to CCA. We used the optimization method proposed in [35] and the same structure as our proposed method and added the BN layer for a fair comparison.

4.1. MNIST

We divided a digital image into a left half and a right half as in [3] to input in our model separately. We normalized the raw pixel values to 0-1 before splitting. We regarded the left half and right half of images with 392 dimensions as different modalities. There exists a clear relationship between the paired half images and we evaluated our model's ability to extract modality-invariant information from these paired samples. We used 60,000 paired samples for training and 10,000 for testing. We followed Andrew *et al.*, [3], wherein 6,000 samples of the training dataset were used in validation split. We tested the non-linear models of MIAN and DeMIAN in this experiment.

Table 2 shows the recognition experiment results learned



(a) CCA (b) DCCA (c) MIAN (d) DeMIAN

Figure 3. Comparison of the embedding in the MNIST experiments. Red points are left half samples and blue points are right half ones.

	Source modality →Target modality			
Method	Left → Right	Right → Left	Left → Left	Right → Right
CCA	0.703	0.675	**0.857**	**0.843**
DCCA	0.147	0.139	0.618	0.553
MIAN (non-linear) w/o z	0.702	0.712	0.724	0.719
MIAN (non-linear)	0.716	0.731	0.762	0.751
DeMIAN w/o z	0.523	0.520	0.529	0.516
DeMIAN	**0.810**	**0.804**	0.820	0.794

Table 2. Result of the MNIST recognition experiment. **Source modality** means the input modality for the supervised training, whereas **Target modality** means the input modality when testing.

	Source modality →Target modality				
Method	Tag → Image	Image → Tag	Tag →Tag	Image →Image	Tag and Image→Tag and Image
DBM	–	–	–	–	0.528
CCA	0.312	0.404	0.428	0.381	0.496
DCCA	0.438	0.455	0.463	0.464	0.570
MIAN (linear) w/o z	0.451	0.417	0.419	0.485	0.486
MIAN (linear)	0.458	0.438	**0.528**	0.548	0.598
DeMIAN w/o z	0.367	0.361	0.366	0.367	0.398
DeMIAN	**0.544**	**0.487**	0.512	**0.567**	**0.599**

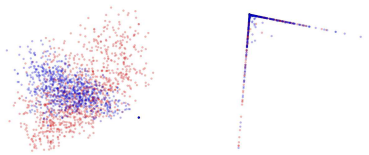
Table 3. Result of Mir Flickr recognition experiment based on MAP. **Source modality** means the input modality when training a linear classifier. **Target modality** means the input for the testing. Note that we show the result where we had access to 25,000 labeled samples. The result of DBM is from [32].

by each model. Our proposed model achieved better performance compared to CCA and DCCA. The model showed little change in accuracy despite the difference in training modality; for example, Left → Right and Right → Right in DeMIAN. This indicates that making distributions similar enabled the adaptation of different modalities in this classification task. On the other hand, while CCA showed the best performance in a standard classification setting, such as Right → Right, its performance substantially declined in Right → Left. This is because CCA did not include a mechanism for learning modality-invariant representations. Our model showed significantly better performance than our model without $z \sim N(0, I_z)$. Thus, merely introducing BN is not sufficient to obtain good representations. This suggests that the introduction of prior z can lead to much better representations, especially in training a deep model. We visualized the learned representation using t-SNE [18] in Fig. 3. From this figure, we can observe that the distribution between the left half and right half digits in the learned common space densely matched compared to the embedding of CCA.

4.2. Mir Flickr

This dataset consists of 1 million images from the social photography website Flickr, along with their user-assigned tags. Twenty-five thousand images were annotated for 38 classes, where each image may belong to several classes [11]. We used 15,000 images for training and 10,000 for testing within labeled samples. Five thousand images of the training split were used for validation. We used the mean average precision (MAP) for the evaluation based on an existing work [32]. Each tag input was represented using the vocabulary of the 2,000 most frequently used tags. The images were represented by 3,857-dimensional features extracted by concatenating the pyramid histogram of word features [5], gist features [22], and MPEG-7 descriptors [19](EHD, HTD, CSD, CLD, SCD).

We show the results of Mir Flickr in Table 3. In this table, we show the modality used for training and testing in supervised training. Image → Tag means that we used



(a) Training *with* z (b) Training *without* z

Figure 4. Visualization of representations embedded by PCA in Mir Flickr experiment. We show the top two principal components. The red points are image features and the blues ones are tag features. **Left**: Representations learned by DeMIAN with prior z. **Right**: Representations learned by DeMIAN without prior z.

the image features and its labels for supervised training, and tested Tag features. Image and Tag $\rightarrow$ Image and Tag means that we used both the image features and Tag features for supervised training, and tested both features. We averaged the output of the classifier for image features and tag features in this setting. Our model achieved better performance compared to other existing methods for both Image $\rightarrow$ Tag and Tag $\rightarrow$ Image. Comparing CCA and MIAN (linear), we can see the direct effect of our modality adaptation method. Our model learned rich representations that were useful for both the source modality and target modality. From the results of Image and Tag $\rightarrow$ Image and Tag, our model performed better than DBM [32], which is one of the most successful models for multimodal learning. The results also show that representations from both modalities were effective for training a linear classifier. In this sense, our proposed model learned modality-invariant rich representation. Furthermore, the clear effects of samples from Gaussian distributions in training DeMIAN were evident in the results. We also visualized the learned representations in Figure. 4. It shows that when training without modality z, the projected points can be very similar. On the other hand, training with modality z makes each projected point dissimilar.

4.3. SUN Attribute

The SUN attribute database contains 717 classes of images and annotations.

For zero-shot learning, unlike in the experiment with SRL, we did not use unseen image features for training. We completely omitted the unseen image features in both the unsupervised and supervised training phases. We followed the protocol of [38] for image features and splitting datasets. We used the VGGNet [31] features and selected 10 classes from the unseen classes following their setting. We tuned the parameters of our model using 10 classes of the seen classes. Then, we reported the the 10-times average of the best scores during supervised training for zero-shot learning.

We show the results of the zero-shot recognition experiment with SUN in Table 4. Our model improved the state-of-the-art accuracy by approximately 2%. Notably, our model achieved state-of-the-art accuracy using a single method, whereas the state-of-the-art accuracy was previously achieved by an ensemble of methods [27] + SP-ZSR [39].

Method	CUB-200-2011	SUN Attribute
Akata *et al.* [2]	40.3	–
Kodirov *et al.* [15]	47.9	–
Peng *et al.* [25]	49.87	
Lampert *et al.* [16]	–	72.00
Paredes *et al.* [27]	–	82.10 ± 0.32
SSE-ReLU[37]	30.41 ± 0.20	82.50 ± 1.32
JLSE [38]	42.11 ± 0.55	83.83 ± 0.29
Bucher *et al.* [7]	43.29 ± 0.38	84.41 ± 0.71
Hard Negative [6]	45.87 ± 0.34	86.21 ± 0.88
Morgado *et al.* [20]	**59.5**	–
Xu *et al.* [36]	53.6	84.5
DeMIAN w/o z	12.0 ± 0.82	43.5 ± 1.2
DeMIAN	57.5 ± 0.56	$\mathbf{87.6 \pm 1.3}$
[27] + SP-ZSR[39]	–	89.5
JLSE + SP-ZSR[39]	55.34 ± 0.77	86.12 ± 0.99

Table 4. Result of the zero-shot learning. Our model achieved the highest accuracy for the SUN and CUB-200-2011 dataset. In particular, SUN's score was the best one including the ensemble method.

4.4. CUB-200-2011

We used the VGGNet [31] features and attributes features following [38]. We used 150 bird species as the seen classes for training and the remaining 50 species as the unseen classes [38] for testing. We selected 50 seen classes for validation as in SUN.

Then, we reported the 10 times average of the best scores during the supervised training for zero-shot learning.

We show the results in Table 4. Our model improved state-of-the-art accuracy by approximately 3%. The effect of using prior is also clear in this experiment.

5. Conclusion

In this paper, we proposed a novel method to learn modality-invariant representations for shared representation learning, called the Deep Modal Invariant Adversarial Network (DeMIAN). Our network incorporated the idea of domain adaptation and multimodal learning. We learned modality-invariant representations through adversarial training and observed the effect of our network in embedding learned representations. Our proposed algorithm showed excellent performance in experiments on SRL and zero-shot learning.

6. Acknowledgement

This work was partially funded by the ImPACT Program of the Council for Science, Technology, and Innovation (Cabinet Office, Government of Japan), and was partially supported by CREST, JST.

References

[1] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand. Domain-adversarial neural networks. *arXiv:1412.4446*, 2014. 2

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 6

[3] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, 2013. 4

[4] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 2

[5] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *ICCV*. IEEE, 2007. 5

[6] M. Bucher, S. Herbin, and F. Jurie. Hard negative mining for metric learning based zero-shot classification. In *ECCV*, 2016. 6

[7] M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *ECCV*, 2016. 6

[8] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *ICLR*, 2016. 4

[9] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2014. 2

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[11] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *ICMR*, 2008. 4, 5

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015. 4

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1

[14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 4

[15] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015. 6

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *PAMI*, 36(3):453–465, 2014. 6

[17] M. Long and J. Wang. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[18] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9:2579–2605, 2008. 5

[19] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *TCSVT*, 11(6):703–715, 2001. 5

[20] P. Morgado and N. Vasconcelos. Semantically consistent regularization for zero-shot recognition. *arXiv:1704.03039*, 2017. 6

[21] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011. 2

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 5

[23] I. Partalas, A. Kosmopoulos, N. Baskiotis, T. Artieres, G. Paliouras, E. Gaussier, I. Androutsopoulos, M.-R. Amini, and P. Galinari. Lshtc: A benchmark for large-scale text classification. *arXiv:1503.08581*, 2015. 1

[24] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *IJCV*, 108(1-2):59–81, 2014. 4

[25] P. Peng, Y. Tian, T. Xiang, Y. Wang, and T. Huang. Joint learning of semantic and latent attributes. In *ECCV*, 2016. 6

[26] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015. 4

[27] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. 6

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1

[29] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016. 3

[30] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012. 2

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 6

[32] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012. 5, 6

[33] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, 2016. 2

[34] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 4

[35] W. Wang, R. Arora, K. Livescu, and N. Srebro. Stochastic optimization for deep cca via nonlinear orthogonal iterations. In *Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2015. 4

[36] X. Xu, F. Shen, Y. Yang, D. Zhang, H. T. Shen, and J. Song. Matrix tri-factorization with manifold regularizations for zero-shot learning. In *Proc. of CVPR*, 2017. 6

[37] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *CVPR*, 2015. 6

[38] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016. 6

[39] Z. Zhang and V. Saligrama. Zero-shot recognition via structured prediction. In *ECCV*, 2016. 6