Multi-View Variational Autoencoder for Robust Classification against Irrelevant Data

Daichi Nishikawa, Ryosuke Harakawa and Masahiro Iwahashi
Department of Electrical, Electronics and Information Engineering, Nagaoka University of Technology, Niigata, Japan
E-mail: s183155@stn.nagaokaut.ac.jp, {harakawa, iwahashi}@vos.nagaokaut.ac.jp

***Abstract*—Multi-view variational autoencoders (MVAEs) can extract latent variables with high discriminative power for classification by considering correlations among multi-view data, i.e., multiple kinds of data. However, if we input irrelevant data, i.e., multiple kinds of data that capture the different object, to MVAEs, discriminative power is reduced. To solve this problem, we propose an MVAE including a novel objective function. Our proposed MVAE reconstructs multi-view data without the negative effect of irrelevant data. Specifically, we derive an objective function that focuses on latent variables of relevant data, i.e., multiple kinds of data that capture the same object. Experimental results show that the proposed method improved the discriminative power of latent variables even if irrelevant data are input.**

I. Introduction

Multi-view variational autoencoders (MVAEs) [1]–[9] can extract latent variables with high discriminative power for classification by considering correlations among multi-view data. In this paper, we define multiple kinds of data (a color image that captures a digit and a grayscale image that capture a digit) as multi-view data; discriminative power is defined as the accuracy of the classifier constructed by using latent variables via MVAEs. Methods presented in previous work [1], [6] can extract latent variables by learning the bi-directional generation process between multi-view data. Methods in previous work [4], [5] enabled the extraction of latent variables even if some data are missing. Schonfeld et al. [8] introduced distribution alignment and cross alignment objective functions into an MVAE for zero-shot learning. Hwang et al. [7] proposed an objective function for decomposing latent variables into two types, shared and exclusive representations. Thus, we can extract disentangled latent variables for domain-invariant and domain-specific representations. Methods in previous work [2], [9] introduced labels into an MVAE. This improves the discriminative power of latent variables. Huang et al. [3] introduced graph embedding into an MVAE. This idea enabled graph embedding considering the correlation among multi-view data for social media contents.

Multi-view data can be divided into *relevant data* and *irrelevant data*. In this paper, we define multiple kinds of data that capture the same object as relevant data, and those that capture the different object as irrelevant data. Conventional methods [1]–[9] do not assume that irrelevant data are input. Therefore, in the training phase, MVAEs learn to reconstruct only relevant data. (see Fig. 1 (a)). As a result, if irrelevant data are input to MVAEs in the test phase, latent variables to reconstruct irrelevant data are calculated. This reduces the discriminative power of the latent variables. Even if we assume that irrelevant data are input to MVAEs (see Fig. 1 (b)), MVAEs extract latent variables that reconstruct irrelevant data, and discriminative power of latent variables is reduced in the test phase.

In this paper, we propose an MVAE including a novel objective function that extracts latent variables with high discriminative power even if irrelevant data are input. Specifically, we focus on conventional methods to reconstruct the original input data in the training phase, and derive an objective function that reconstructs relevant data from irrelevant data in the training phase (see Fig. 1 (c)). As a result, even if irrelevant data are input in the test phase, latent variables focused on relevant data are extracted and, the negative effect of irrelevant data on the latent variable is reduced. Experimental results for MNIST [10] and SVHN [11] datasets show the effectiveness of our MVAE.

II. Problem of Conventional Objective Function

In this section, we describe the problem of conventional objective functions [1]–[9]. Original variational autoencoder (VAE) [12] is a probabilistic variant of the traditional autoencoder. The important characteristic is that it assumes that latent variables follow a prior distribution. The VAE assumes that high-dimensional data can be represented by low-dimensional latent variables $\boldsymbol{z}$ that follow a certain prior distribution. The dimension of the input data is compressed by an encoder, and then the compressed data is reconstructed to the original input data by a decoder. This process allows the VAE to extract the latent variables $\boldsymbol{z}$ with high discriminative power. MVAEs are based on the VAE and take multi-view data as input. MVAEs extract latent variables that capture the characteristics shared by multi-view data. The objective function of MVAEs is a variational lower bound $\mathcal{L}(\theta, \phi; \boldsymbol{x}_{1:M})$ on the log-likelihood $\log p_\theta(\boldsymbol{x}_{1:M})$:

$$\begin{aligned}
\log p_\theta(\boldsymbol{x}_{1:M}) &= \log \int p_\theta(\boldsymbol{x}_{1:M}, \boldsymbol{z}) d\boldsymbol{z} \\
&= \log \int q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M}) \frac{p_\theta(\boldsymbol{x}_{1:M}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})} d\boldsymbol{z} \\
&\geq \int q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M}) \log \frac{p_\theta(\boldsymbol{x}_{1:M}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})} d\boldsymbol{z} \\
&= \mathbb{E}_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})} \left[\log \frac{p_\theta(\boldsymbol{x}_{1:M}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})} \right] \\
&= \mathcal{L}(\theta, \phi; \boldsymbol{x}_{1:M}), \qquad (1)
\end{aligned}$$

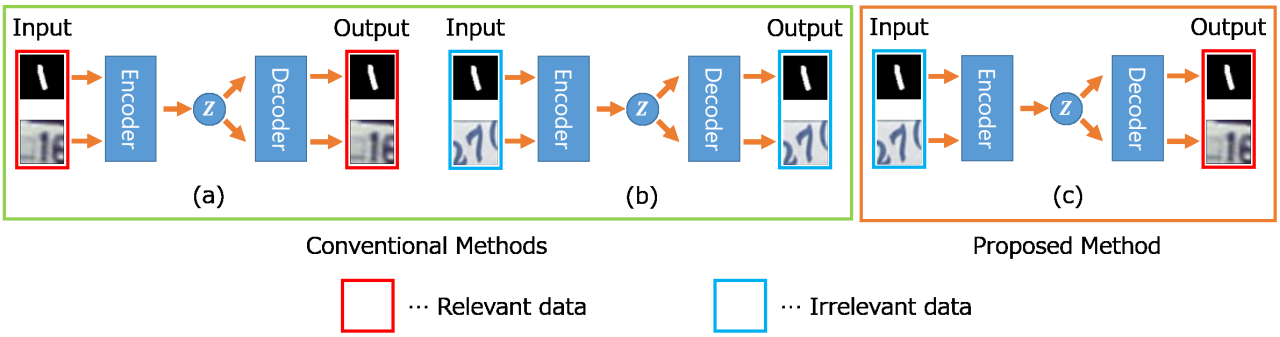


Fig. 1. Reconstruction process of the conventional method and the proposed method in the training phase. (a) Conventional methods with only relevant data. (b) Conventional methods with irrelevant data. (c) Our MVAE that reconstructs relevant data from irrelevant data.

where θ is generative model parameter, $\boldsymbol{x}_{1:M}$ are input multi-view data and $\boldsymbol{z}$ is a latent variable. $q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})$ is the posterior distribution approximated by an encoder that takes the true posterior distribution $p_\theta(\boldsymbol{z}|\boldsymbol{x}_{1:M})$ as its parameter ϕ. Conventional methods [1]–[9] derive original objective functions on the basis of $\mathcal{L}(\theta, \phi; \boldsymbol{x}_{1:M})$.

However, the objective function $\mathcal{L}(\theta, \phi; \boldsymbol{x}_{1:M})$ has a problem. Specifically, it assumes that only relevant data are input to the MVAEs [1]–[9]. If irrelevant data are input to the MVAEs, latent variables are extracted to reconstruct irrelevant data. As a result, the discriminative power is reduced.

III. Novel Objective Function Robust against Irrelevant Data

We propose a novel objective function that assumes that irrelevant data are input to an MVAE in the training phase. Similar to the conventional method, the proposed method considers the log-likelihood $\log p_\theta(\boldsymbol{x}^R_{1:M})$ of relevant data $\boldsymbol{x}^R_{1:M}$, which consist of M data, to be marginalized by a latent variable $\boldsymbol{z}$.

$$\log p_\theta(\boldsymbol{x}^R_{1:M}) = \log \int p_\theta(\boldsymbol{x}^R_{1:M}, \boldsymbol{z}) d\boldsymbol{z}. \tag{2}$$

In the conventional objective function, the true posterior probability $p_\theta(\boldsymbol{z}|\boldsymbol{x}^R_{1:M})$ for relevant data is approximated by the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x}^R_{1:M})$. Therefore, if irrelevant data are input to MVAEs, latent variables that reconstruct irrelevant data are extracted in the training phase. In the proposed method, we assume that the true posterior probability $p_\theta(\boldsymbol{z}|\boldsymbol{x}^R_{1:M})$ for relevant data is approximated by the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})$ for irrelevant data $\boldsymbol{x}_{1:M}$ as follows.

$$\begin{aligned}
\log p_\theta(\boldsymbol{x}^R_{1:M}) &= \log \int q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M}) \frac{p_\theta(\boldsymbol{x}^R_{1:M}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})} d\boldsymbol{z} \\
&\geq \int q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M}) \log \frac{p_\theta(\boldsymbol{x}^R_{1:M}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})} d\boldsymbol{z} \\
&= \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})} \left[\log \frac{p_\theta(\boldsymbol{x}^R_{1:M}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x}_{1:M})} \right]. \quad (3)
\end{aligned}$$

The conventional objective function [1] improves the discriminative power of latent variables by using Importance Weighted Autoencoder (IWAE) [13] and mixture-of-experts [14]. Motivated by this fact, we propose a novel objective function $\mathcal{L}_p(\Theta, \Phi; \boldsymbol{x}^R_{1:M}, \boldsymbol{x}_{1:M})$ based on Eq. (3).

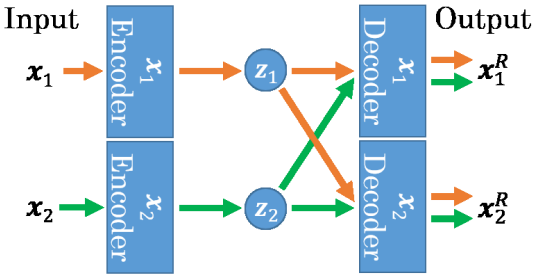


Fig. 2. Architecture of the proposed method ($M = 2$). Orange arrows represent the reconstruction process from $\boldsymbol{x}_1$. Green arrows represent the reconstruction process from $\boldsymbol{x}_2$.

$$\begin{aligned}
&\mathcal{L}_p(\Theta, \Phi; \boldsymbol{x}^R_{1:M}, \boldsymbol{x}_{1:M}) \\
&= \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\boldsymbol{z}^{1:K}_m \sim q_{\phi_m}(\boldsymbol{z}|\boldsymbol{x}_m)} \left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{p_\Theta(\boldsymbol{x}^R_{1:M}, \boldsymbol{z}^k_m)}{q_\Phi(\boldsymbol{z}^k_m|\boldsymbol{x}_{1:M})} \right], \quad (4)
\end{aligned}$$

where Θ is all generative model parameters, Φ is all encoder parameters, ϕ_m is m-th encoder parameter, $\boldsymbol{x}_m$ is each of input multi-view data, and K is the number of samples for the Monte Carlo method in IWAE. This objective function allows us to learn latent variables focusing on relevant data rather than irrelevant data. The proposed objective function is optimized using AMSGrad [15]. The proposed method utilizes the same architecture as MMVAE [1] (see Fig. 2), which can extract highly discriminative latent variables. In the proposed method, the architecture is defined depending on the input data (see Sec. IV-A).

IV. Experimental Results

In this experiment, we show that the proposed MVAE (**Ours**) has high discriminative power even if irrelevant data are input. Specifically, we constructed Support Vector Machine (SVM) [16] using the latent variables calculated by **Ours** and verify their discrimination accuracy. We assumed that irrelevant data is caused by pairing with mislabeled data,

referring to previous studies [17]–[21]. In other words, when we paired with mislabeled data, we evaluate the discrimination accuracy of SVM.

A. Conditions

The following two methods are used as comparative methods.

CM1: MMVAE [1] using only relevant data in the training phase.

CM2: MMVAE using irrelevant data in the training phase.

In this experiment, we used the pair of the Modified National Institute of Standards and Technology database (MNIST) [10] (consisting of grayscale handwritten numeric images) and the Street View House Numbers database (SVHN) [11] (consisting of color images of house numbers obtained from Google Street View[1]) as multi-view data. First, we prepared a training dataset without irrelevant data. Specifically, we randomly selected images with the same digit class from 60,000 MNIST training images and 73,257 SVHN training images and prepared multi-view data by pairing with these images. By this process, we prepared 135,000 multi-view data for each digit class (the digits 0-9) (1,350,000 in total).

Next, we prepared a test dataset including both relevant data and irrelevant data as follows:

Step 1 Similar to the training dataset, we prepared 25,500 multi-view data for each class (255,000 in total) by randomly selecting and pairing with the same digit classes from 10,000 MNIST test images and 26,032 SVHN test images.

Step 2 We randomly selected data from the digit classes according to the transition matrices shown in Fig. 3. For example, we randomly selected 30% from MNIST images of digit classes 2, 3, 5, and 8 from the test dataset.

Step 3 According to the transition matrices shown in Fig. 3, we randomly rewrote the image selected in Step 2 For example, we rewrote the MNIST images of "class 2" selected in Step 2 to images of "class 3".

Finally, we used the irrelevant data for training of **CM2** and **Ours**. Therefore, we prepared training dataset $\boldsymbol{x}_{1:M}$ and $\boldsymbol{x}^R_{1:M}$ of **CM2** and **Ours** in the same manner as Step 2 and Step 3. Note that **CM1** is trained to reconstruct only relevant data, as shown in Fig. 1 (a). **CM2** is trained to reconstruct both relevant data and irrelevant data, as shown in Fig. 1 (b). **Ours** is trained to reconstruct relevant data from irrelevant data, as shown in Fig. 1 (c).

In this experiment, the architecture of MMVAE [1] was adopted for **CM1**, **CM2**, and **Ours**. The details of the architecture are shown in Tables I and II. For all methods, the batch size, the number of training sessions, K, and the learning rate were set to 128, 30 epochs, 30, and 0.001, respectively.

We used the linear SVM [16] classifier to evaluate the performance of all methods. Specifically, we prepared two SVMs for **CM1**, **CM2**, and **Ours** because the MVAEs of these methods extracts two latent variables from multi-view data. The first SVM was trained by 27,000 latent variables extracted from MNIST. The second SVM was trained by 27,000 latent variables extracted from SVHN. We calculated the average of the prediction probabilities [24] from the two SVMs for the test dataset and defined the class with the highest probability as the discrimination result.

[1]https://www.google.co.jp/maps

TABLE I
ARCHITECTURE OF THE ENCODER AND DECODER FOR MNIST. FC. IS A FULLY CONNECTED LAYER. L IS THE NUMBER OF DIMENSIONS OF LATENT VARIABLES.

Encoder	**Decoder**
Input$\in \mathbb{R}^{1\times28\times28}$	Input $\in \mathbb{R}^{L}$
FC. 400 ReLU [22]	FC. 400 ReLU
FC. L, FC. L	FC. 1×28×28 Sigmoid [23]

TABLE II
ARCHITECTURE OF THE ENCODER AND DECODER FOR SVHN. *conv.* IS A CONVOLUTION LAYER. *pad.* IS A PADDING PROCESSING. *upconv.* IS A DECONVOLUTION LAYER.

Encoder
Input$\in \mathbb{R}^{3\times32\times32}$
4×4 conv. 32 stride 2 pad. 1 & ReLU
4×4 conv. 64 stride 2 pad. 1 & ReLU
4×4 conv. 128 stride 2 pad. 1 & ReLU
4×4 conv. L stride 1 pad. 0, 4×4 conv. L stride 1 pad. 0
Decoder
Input$\in \mathbb{R}^{L}$
4×4 upconv. 128 stride 1 pad. 0 & ReLU
4×4 upconv. 64 stride 2 pad. 1 & ReLU
4×4 upconv. 32 stride 2 pad. 1 & ReLU
4×4 upconv. 3 stride 2 pad. 1 & Sigmoid

B. Results

Table III shows the accuracy of classification by SVM. This table confirms that **Ours** is 8% more accurate than **CM1** and **CM2** on average. For the classes containing irrelevant data (classes 0, 1, 2, 3, 5, 7, 8, and 9), we can confirm that the accuracy is improved by 10% compared with **CM1** and **CM2**. These results show that **CM2** is robust against SVHN mislabeled images, but vulnerable against MNIST mislabeled images. This means that conventional MVAE is not robust against irrelevant data, because it learns latent variables to reconstruct irrelevant data.

Figures 4 and 5 show the reconstruction results of irrelevant data. Figure 4 shows that **Ours** can reconstruct the SVHN image to MNIST image more accurately than **CM1** and **CM2**. Figure 5 shows that **Ours** does not reconstruct the SVHN image as well as **CM1** and **CM2**. This confirms that **Ours** does not learn latent variables to reconstruct the mislabeled SVHN.

Figure 6 shows the confusion matrix that represents the relationship between ground truths and prediction results for each digit class. **CM1** is accurate for classes 4 and 6, which do not contain irrelevant data. On the contrary, for classes 0, 1, 2, 3, 5, 7, 8, and 9, which contain irrelevant data, the discriminative power of latent variables is reduced due to the negative effect of irrelevant data. **CM2** extracted the

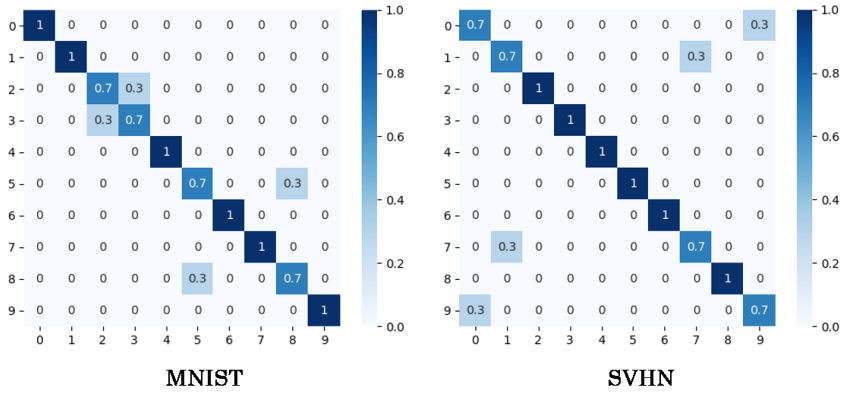


Fig. 3. Transition matrices of MNIST and SVHN.

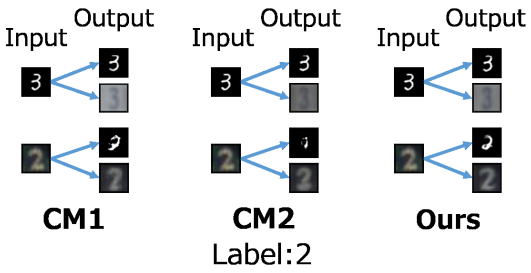


Fig. 4. Reconstruction of irrelevant data including MNIST mislabeled image of **CM1**, **CM2** and **Ours**.

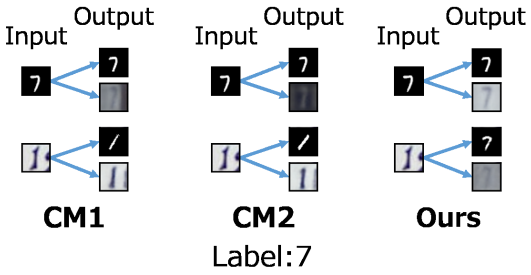


Fig. 5. Reconstruction of irrelevant data including SVHN mislabeled image of **CM1**, **CM2** and **Ours**.

discriminative power of latent variables in classes 0, 1, 7, and 9 compared with **CM1**. However, for classes 2, 3, 5, and 8, the discriminative power of latent variables is reduced in **CM2** compared with **CM1**. On the other hand, **Ours** extracts latent variables that are more discriminative than **CM1** and **CM2** in each digit class containing irrelevant data.

V. Conclusion

Conventional MVAE is designed based on the assumption that only the relevant data are input. This assumption causes the problem that the discriminative power of latent variables is reduced when irrelevant data are input. To solve this problem, we proposed MVAE including a novel objective function. Specifically, we derived an objective function that reconstructs relevant data from irrelevant data in the training phase. Experimental results show that our proposed method improved the discriminative power of latent variables even if irrelevant data are input.

TABLE III
Accuracy of **CM1**, **CM2** and **Ours**. Classes 0, 1, 2, 3, 5, 7, 8 and 9 include irrelevant data. Classes 4 and 6 do not include irrelevant data.

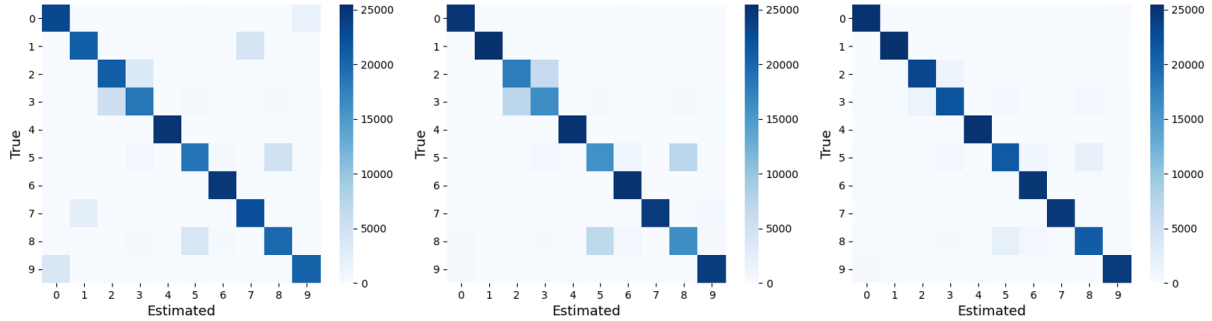
Method	CM1	CM2	Ours
Classes 0, 1, 2, 3, 5, 7, 8 and 9	0.810	0.811	**0.916**
Classes 4 and 6	0.972	**0.982**	0.978
Average	0.843	0.846	**0.928**

Acknowledgement

This work was partly supported by JSPS KAKENHI Grant Number JP21K11934. We wish to thank Tokyo Electric Power Company Holdings, Incorporated for supporting our research.

References

[1] Y. Shi, N. Siddharth, B. Paige, and P.H.S. Torr, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," in *Proc. Advances in Neural Information Processing Systems*, 2019, pp. 15718–15729.
[2] D. Khattar, J.S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conference*, 2019, pp. 2915–2921.
[3] F. Huang, X. Zhang, J. Xu, C. Li, and Z. Li, "Network embedding by fusing multimodal contents and links," *Knowledge-Based Systems*, vol. 171, pp. 44–55, May 2019.
[4] M.Wu and N.Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *Proc. Advances in Neural Information Processing Systems*, 2018, pp. 5575–5585.
[5] M. Zambelli, A. Cully, and Y. Demiris, "Multimodal representation models for prediction and control from partial information," *Robotics and Autonomous Systems*, vol. 123, pp. 103312, 2020.
[6] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," in *Proc. International Conference on Learning Representations Workshops*, 2016.



(a) **CM1** (b) **CM2** (c) **Ours**

Fig. 6. Confusion matrix of **CM1**, **CM2** and ours.

[7] H.J. Hwang, G.H. Kim, S. Hong, and K.E. Kim, "Variational interaction information maximization for cross-domain disentanglement," in *Proc. Advances in Neural Information Processing Systems*, 2020, pp. 22479–22491.

[8] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero-shot learning via aligned variational autoencoders," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[9] K. Lin, X. Xu, L. Gao, Z. Wang, and H.T. Shen, "Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval," in *Proc. the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 2020, pp. 11515–11522.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A.Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. Advances in Neural Information Processing Systems Workshops*, 2011.

[12] D.P. Kingma, P. Diederik, and M. Welling, "Auto-encoding variational bayes," in *Proc. International Conference on Learning Representations*, 2014.

[13] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," in *Proc. International Conference on Learning Representations*, 2016.

[14] S.J. Nowlan and G.E. Hinton, "Evaluation of adaptive mixtures of competing experts," in *Proc. Advances in Neural Information Processing System*, 1990, pp. 774–780.

[15] S.J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *Proc. International Conference on Learning Representations*, 2018.

[16] V. Vapnik, *The nature of statistical learning theory*, Springer science & business media, 2013.

[17] D. Hendrycks, M. Mazeika, D. Wilson, and K. Gimpel, "Using trusted data to train deep networks on labels corrupted by severe noise," in *Proc. Advances in Neural Information Processing Systems*, 2018, pp. 10477–10486.

[18] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.J. Li, "Learning from noisy labels with distillation," in *Proc. the IEEE International Conference on Computer Vision*, 2017, pp. 1910–1918.

[19] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Advances in Neural Information Processing Systems*, 2018, pp. 8536–8546.

[20] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?," in *Proc. International Conference on Machine Learning*, 2019, pp. 7164–7173.

[21] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13726–13735.

[22] A. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[23] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018.

[24] T.F. Wu, C.J. Lin, and R.C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, no. Aug, pp. 975–1005, 2004.