JOINT MULTIMODAL LEARNING WITH DEEP GENERATIVE MODELS

Masahiro Suzuki, Kotaro Nakayama, Yutaka Matsuo
The University of Tokyo
Bunkyo-ku, Tokyo, Japan
{masa,k-nakayama,matsuo}@weblab.t.u-tokyo.ac.jp

ABSTRACT

We investigate deep generative models that can exchange multiple modalities bi-directionally, e.g., generating images from corresponding texts and vice versa. Recently, some studies handle multiple modalities on deep generative models. However, these models typically assume that modalities are forced to have a conditioned relation, i.e., we can only generate modalities in one direction. To achieve our objective, we should extract a joint representation that captures high-level concepts among all modalities and through which we can exchange them bi-directionally. As described herein, we propose a joint multimodal variational autoencoder (JMVAE), in which all modalities are independently conditioned on joint representation. In other words, it models a joint distribution of modalities. Furthermore, to be able to generate missing modalities from the remaining modalities properly, we develop an additional method, JMVAE-kl, that is trained by reducing the divergence between JMVAE's encoder and prepared networks of respective modalities. Our experiments show that JMVAE can generate multiple modalities bi-directionally.

1 INTRODUCTION

In our world, information is represented through various modalities and people often exchange such information bi-directionally. To do so, it is important to extract a joint representation that captures high-level concepts among all modalities. Deep neural network architectures have been used widely for multimodal learning by sharing the top of hidden layers in modality specific networks as joint representations (Ngiam et al., 2011; Srivastava & Salakhutdinov, 2012). Among them, generative approaches using deep Boltzmann machines (DBMs) (Srivastava & Salakhutdinov, 2012; Sohn et al., 2014) offer the important advantage that these can generate modalities bi-directionally.

Recently, variational autoencoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014) have been proposed to estimate flexible deep generative models by variational inference methods. These models can be trained on large-scale and high-dimensional dataset compared with DBMs with MCMC training. However, some previous studies (Kingma et al., 2014; Sohn et al., 2015; Pandey & Dukkipati, 2016) are forced to model conditional distribution. Therefore, it can only generate modalities in one direction.

We develop a novel multimodal learning model with VAEs, which we call a joint multimodal variational autoencoder (JMVAE). The most significant feature of our model is that all modalities, $\mathbf{x}$ and $\mathbf{w}$ (e.g., images and texts), are conditioned independently on a latent variable $\mathbf{z}$ corresponding to joint representation, i.e., the JMVAE models a joint distribution of all modalities, $p(\mathbf{x}, \mathbf{w})$. Therefore, we can extract a high-level representation that contains all information of modalities. Moreover, since it models a joint distribution, we can draw samples from both $p(\mathbf{x}|\mathbf{w})$ and $p(\mathbf{w}|\mathbf{x})$. Because, at this time, modalities that we want to generate are usually missing, the inferred latent variable becomes incomplete and generated samples might be collapsed in the testing time when missing modalities are high-dimensional and complicated. To prevent this issue, we propose a method of preparing the new encoders for each modality, $p(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z}|\mathbf{w})$, and reducing the divergence between the multimodal encoder $p(\mathbf{z}|\mathbf{x}, \mathbf{w})$, which we call JMVAE-kl. This contributes

to more effective bi-directional generation of modalities, e.g., from face images to texts (attributes) and vice versa.

2 PROPOSED METHOD

We consider $i.i.d.$ dataset $(\mathbf{X}, \mathbf{W}) = \{(\mathbf{x}_1, \mathbf{w}_1), ..., (\mathbf{x}_N, \mathbf{w}_N)\}$, where two modalities $\mathbf{x}$ and $\mathbf{w}$ have different kinds of dimensions and structures. Our objective is to generate two modalities bi-directionally. For that reason, we assume that these are conditioned independently on the same latent concept $\mathbf{z}$: joint representation. Therefore, we assume their generating processes as $\mathbf{z} \sim p(\mathbf{z})$ and $\mathbf{x}, \mathbf{w} \sim p(\mathbf{x}, \mathbf{w}|\mathbf{z}) = p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})p_{\theta_\mathbf{w}}(\mathbf{w}|\mathbf{z})$, where $\theta_\mathbf{x}$ and $\theta_\mathbf{w}$ represent the model parameters of each independent p. One can see that this models joint distribution of all modalities, $p(\mathbf{x}, \mathbf{w})$. Therefore, we designate this model as a *joint multimodal variational autoencoder* (JMVAE).

Considering an approximate posterior distribution as $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})$, we can estimate a lower bound of the log-likelihood $\log p(\mathbf{x}, \mathbf{w})$ as follows:

$$\mathcal{L}_{JM}(\mathbf{x}, \mathbf{w}) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})||p(\mathbf{z})) + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})] + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log p_{\theta_\mathbf{w}}(\mathbf{w}|\mathbf{z})]. \quad (1)$$

To optimize the lower bound $\mathcal{L}(\mathbf{x})$ with respect to parameters, we estimate gradients of Equation 1 using stochastic gradient variational Bayes (SGVB).

In the JMVAE, we can extract joint latent features by sampling from the encoder $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})$ at testing time. Our objective is to exchange modalities bi-directionally, e.g., images to texts and vice versa. In this setting, modalities that we want to sample are missing, so that inputs of such modalities are set to zero. However, if missing modalities are high-dimensional and complicated such as natural images, then the inferred latent variable becomes incomplete and generated samples might collapse.

We propose a method to solve this issue, which we designate as JMVAE-kl. Suppose that we have encoders with a single input, $q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})$ and $q_{\phi_\mathbf{w}}(\mathbf{z}|\mathbf{w})$, where $\phi_\mathbf{x}$ and $\phi_\mathbf{w}$ are parameters. We would like to train them by bringing their encoders close to an encoder $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})$. Therefore, the object function of JMVAE-kl becomes

$$\mathcal{L}_{JM_{kl}(\alpha)}(\mathbf{x}, \mathbf{w}) = \mathcal{L}_{JM}(\mathbf{x}, \mathbf{w}) - \alpha \cdot [D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})||q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})) + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})||q_{\phi_\mathbf{w}}(\mathbf{z}|\mathbf{w}))], \quad (2)$$

where α is a factor that regulates the KL divergence terms.

3 EXPERIMENTS

In this experiment, we used CelebA (Liu et al., 2015) dataset. CelebA consists of 202,599 color face images and corresponding 40 binary attributes such as male, eyeglasses, and mustache. In this work, we regard them as two modalities. Beforehand, we cropped the images to squares and resized to 64×64 and normalized. In this experiment, we trained the JMVAE-kl ($\alpha = 0.1$) lower bound as JMVAE. In order to generate clearer images, we combined JMVAE with generative adversarial networks (GANs) (Goodfellow et al., 2014) in the same way as a VAE-GAN model (Larsen et al., 2015). The models were implemented using Theano (Team et al., 2016), Lasagne (Dieleman et al., 2015) and Tars[1]. Our code is available online[2].

Table 1 presents the evaluations of marginal and conditional log-likelihood. We compare the test marginal log-likelihood against VAEs (Kingma & Welling, 2013; Rezende et al., 2014) and the test conditional log-likelihood against CVAEs (Kingma et al., 2014; Sohn et al., 2015) and CMMAs (Pandey & Dukkipati, 2016). Just like the JMVAE setting, we combine all competitive models with GAN. From this table, it is apparent that values of both marginal and conditional log-likelihood with JMVAEs are larger than those with other competitive methods.

Next, we confirm that JMVAE can generate images from attributes. Figure 1(a) portrays generated faces conditioned on various attributes. We find that we can generate an average face of each attribute and various random faces conditioned on a certain attributes. Figure 1(b) shows that samples are gathered for each attribute and that locations of each variation are the same irrespective of attributes. From these results, we find that manifold learning of joint representation with images and attributes works well.

[1] https://github.com/masa-su/Tars

[2] https://github.com/masa-su/jmvae

Table 1: Evaluation of log-likelihood : *left*, marginal log-likelihood; *right*, conditional log-likelihood.

	$\leq \log p(\mathbf{x})$			$\leq \log p(\mathbf{x}\|\mathbf{w})$
	multiple	single	CVAE	-4152
VAE		-4439	CMMA	-4147
JMVAE	**-4141**	-4144	JMVAE	**-4130**

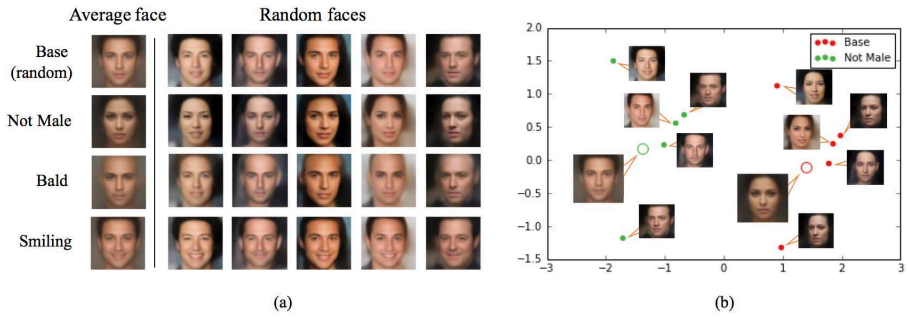

Figure 1: (a) Generation of average faces and corresponding random faces. We first set all values of attributes $\{-1, 1\}$ randomly and designate them as Base. Then, we choose an attribute that we want to set (e.g., Male, Bald, Smiling) and change this value in Base to 2 (or -2 if we want to set "Not"). Each column corresponds to same attribute according to legend. Average faces are generated from $p(\mathbf{x}|\mathbf{z}_{mean})$, where $\mathbf{z}_{mean}$ is a mean of $q(\mathbf{z}|\mathbf{w})$. Moreover, we can obtain various images conditioned on the same values of attributes such as $\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})$, where $\mathbf{z} = \mathbf{z}_{mean} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\zeta})$, and $\boldsymbol{\zeta}$ is the parameter which determines the range of variance. Each row in random faces has the same $\boldsymbol{\epsilon}$. (b) PCA visualizations of latent representation. Colors indicate which attribute each sample is conditioned on.

Finally, we demonstrate that JMVAE can generate bi-directionally between faces and attributes. Figure 2 shows that JMVAE can generate both attributes and changed images conditioned on various attributes from images which had no attribute information. This way of generating an image by varying attributes is similar to the way of the CMMA (Pandey & Dukkipati, 2016). However, the CMMA cannot generate attributes from an image because it only generates images from attributes in one direction.

4 Conclusion

In this paper, we introduced a novel multimodal learning model with VAEs, the joint multimodal variational autoencoders (JMVAE). In this model, modalities are conditioned independently on joint representation, i.e., it models a joint distribution of all modalities. We further proposed the method (JMVAE-kl) of reducing the divergence between JMVAE's encoder and a prepared encoder of each modality to prevent generated samples from collapsing when modalities are missing. We confirmed that the JMVAE can obtain appropriate joint representations and high log-likelihoods on CelebA datasets. Moreover, we demonstrated that the JMVAE can generate multiple modalities bi-directionally.

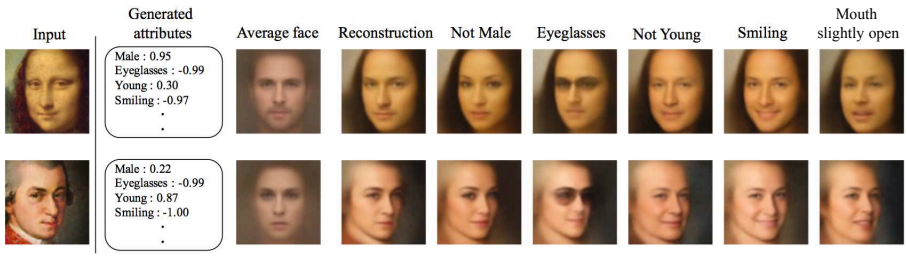

Figure 2: Portraits of the Mona Lisa (upper) and Mozart (lower), generated their attributes, and reconstructed images conditioned on varied attributes, according to the legend. We cropped and resized it in the same way as CelebA. The procedure is as follows: generate the corresponding attributes $\mathbf{w}$ from an unlabeled image $\mathbf{x}$; generate an average face $\mathbf{x}_{mean}$ from the attributes $\mathbf{w}$; select attributes which we want to vary and change the values of these attributes; generate the changed average face $\mathbf{x}'_{mean}$ from the changed attributes; and obtain a changed reconstruction image $\mathbf{x}'$ by $\mathbf{x} + \mathbf{x}'_{mean} - \mathbf{x}_{mean}$.

REFERENCES

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Sander Dieleman, Jan Schlter, Colin Raffel, Eben Olson, Sren Kaae Snderby, Daniel Nouri, Daniel Maturana, Martin Thoma, Eric Battenberg, Jack Kelly, Jeffrey De Fauw, Michael Heilman, Diogo Moitinho de Almeida, Brian McFee, Hendrik Weideman, Gbor Takcs, Peter de Rivaz, Jon Crall, Gregory Sanders, Kashif Rasul, Cong Liu, Geoffrey French, and Jonas Degrave. Lasagne: First release., August 2015. URL `http://dx.doi.org/10.5281/zenodo.27878`.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696, 2011.

Gaurav Pandey and Ambedkar Dukkipati. Variational methods for conditional multimodal learning: Generating human faces from attributes. *arXiv preprint arXiv:1603.01801*, 2016.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Kihyuk Sohn, Wenling Shang, and Honglak Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pp. 2141–2149, 2014.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pp. 3483–3491, 2015.

Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pp. 2222–2230, 2012.

The Theano Development Team, Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, et al. Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*, 2016.

A TEST LOWER BOUNDS

A lower bound used to estimate test marginal log-likelihood $\log p(\mathbf{x})$ of the JMVAE are as follows:

$$\mathcal{L}(\mathbf{x}) = E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log \frac{p_{\theta_x}(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}]. \tag{3}$$

We can also estimate test conditional log-likelihood $\log p(\mathbf{x}|\mathbf{w})$ from this lower bound as

$$\mathcal{L}(\mathbf{x}|\mathbf{w}) = E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log \frac{p_{\theta_x}(\mathbf{x}|\mathbf{z})p_{\theta_w}(\mathbf{w}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}] - \log p(\mathbf{w}), \tag{4}$$

where $\log p(\mathbf{w}) = \log E_{p(\mathbf{z})}[p_{\theta_w}(\mathbf{w}|\mathbf{z})] = \log \frac{1}{N_w} \sum_i^{N_w} p_{\theta_w}(\mathbf{w}|\mathbf{z}^{(i)})$ and $\mathbf{z}^{(i)} \sim p(\mathbf{z})$. In this paper, we set $N_w = 10$.

We can obtain a tighter bound on the log-likelihood by k-fold importance weighted sampling (Burda et al., 2015). For example, we obtain an importance weighted bound on $\log p(\mathbf{x})$ from Equation 3 as follows:

$$\log p(\mathbf{x}) \geq E_{\mathbf{z}_1,\ldots,\mathbf{z}_k \sim q_{\phi_x}(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log \frac{1}{k} \sum_{i=1}^{k} \frac{p_{\theta_x}(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_{\phi_x}(\mathbf{z}|\mathbf{x},\mathbf{w})}] = \mathcal{L}^k(\mathbf{x}). \tag{5}$$