# Associative Variational Auto-Encoder with Distributed Latent Spaces and Associators

**Dae Ung Jo,[1] ByeongJu Lee,[1] Jongwon Choi,[2] Haanju Yoo,[3] Jin Young Choi[1]**
{mardaewoon, adolys, jychoi}@snu.ac.kr, {jw17.choi, haanju.yoo}@samsung.com
[1]Department of ECE, ASRI, Seoul National University, Korea
[2]Samsung SDS, Korea, [3]Samsung Research, Korea

## Abstract

In this paper, we propose a novel structure for a multi-modal data association referred to as Associative Variational Auto-Encoder (AVAE). In contrast to the existing models using a shared latent space among modalities, our structure adopts distributed latent spaces for multi-modalities which are connected through cross-modal associators. The proposed structure successfully associates even heterogeneous modality data and easily incorporates the additional modality to the entire network via the associator. Furthermore, in our structure, only a small amount of supervised (paired) data is enough to train associators after training auto-encoders in an unsupervised manner. Through experiments, the effectiveness of the proposed structure is validated on various datasets including visual and auditory data.

## Introduction

The brain combines multisensory information to understand the surrounding situation. Through various sensory experiences, humans learn the relationships between multisensory data and understand the experienced situation. This mechanism to learn the relationship among multiple stimuli is called associative learning (Bliss and Collingridge 1993; Buonomano and Merzenich 1998; Van Praag et al. 1999). Because of the associative learning mechanism, humans can robustly understand and perceive their surrounding situations even when only some of the modalities are available.

In the field of machine learning, utilizing multi-modality is also important issues because of its usefulness in a wide range of applications (Baltrušaitis, Ahuja, and Morency 2018; Bengio, Courville, and Vincent 2013). As a representative example, object recognition and scene understanding methods based on multi-modal data outperform the methods using only single-modal data (Hu, Li, and others 2016; Ngiam et al. 2011). Moreover, one can generate the synthesized data for a missing or desired modality (Cadena, Dick, and Reid 2016; Lim et al. 2018; Senocak et al. 2018; Spurr et al. 2018; Wang, van de Weijer, and Herranz 2018; Yoo et al. 2017). The multi-modal data association is one of the fundamental steps to understand the relationships among multi-modal data. Recently, along with the advances of deep learning, many studies have attempted to solve the multi-modal data association problem by deep learning algorithms (Baltrušaitis, Ahuja, and Morency 2018). The studies have adopted an approach that encodes multi-modal data into a shared latent space to memorize common features among multiple modalities (Cadena, Dick, and Reid 2016; Hu, Li, and others 2016; Ngiam et al. 2011; Spurr et al. 2018; Wu and Goodman 2018).

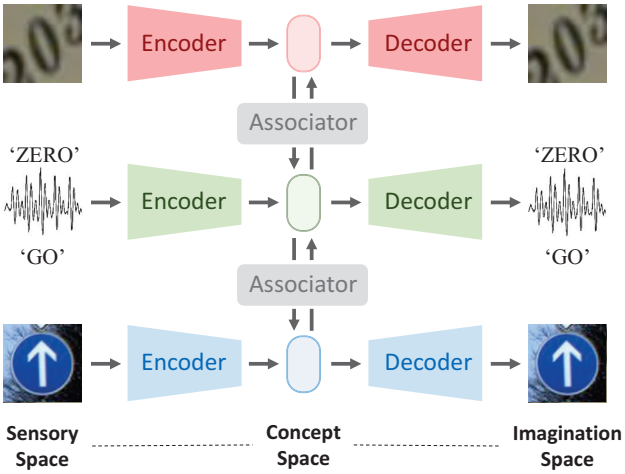


Figure 1: Conceptual illustration of the proposed AVAE. AVAE has modality-specific encoders and decoders for each modality (image, voice). Each modality has its own latent space, which is painted with a different color (red, green, blue). The latent spaces are connected via the proposed *associator* which associates two different modalities.

However, as pointed out by Chaudhury et al. (2017), most existing studies did not consider the case that the characteristic of each modality is very different from others. The encoding in the shared latent space is hard to represent all characteristics of the heterogeneous modalities or could be biased to a dominant modality. Furthermore, the capacity of the shared latent will be saturated as modalities increase and so encounters a scalability problem. To mitigate the limita-

tion of the shared latent space, we propose an approach that adopts distributed latent spaces. In our approach, as shown in Figure 1, each modality is encoded in each latent space separately by the variational auto-encoder (VAE) (Kingma and Welling 2013) and the distributed latent spaces are associated with the other modalities via associators.

The proposed structure is implemented with a deep neural network with multiple variational auto-encoders and variational associators. The loss function to train the network is derived by the variational inference framework. In experiments, the effectiveness and performance are evaluated through comparison with the existing methods and self-analysis using various datasets including voice and visual data. In addition, by self-experiments, the advantage of our structure is verified on generalization ability for semi-supervised learning, scalability of the network, and flexibility of distributed latent space dimensions.

## Related Works

### Multi-modality in Machine Learning

One of the major issues in machine learning is exploiting multi-modal data for various applications, such as data generation (Kingma and Welling 2013; Goodfellow et al. 2014; Spurr et al. 2018; Yoo et al. 2017), retrieval (Wang et al. 2016) and recognition (Hu, Li, and others 2016; Ngiam et al. 2011). There are a lot of studies that extract modality independent features by finding the shared representation of multi-modal data (Baltrušaitis, Ahuja, and Morency 2018). The shared representation is utilized in diverse applications such as handling a missing modality (Cadena, Dick, and Reid 2016; Spurr et al. 2018) or accomplishing better performance than models trained on single-modal data (Hu, Li, and others 2016; Ngiam et al. 2011). The research related to multi-modality can be categorized into two groups (Baltrušaitis, Ahuja, and Morency 2018).

One is a method mapping data from diverse modalities to the shared latent space. Wu and Goodman (2018) proposes the extended version of a Variational auto-encoder (Kingma and Welling 2013) which combines distribution parameters from encoders and calculate integrated distribution parameters. Spurr et al. (2018) also a variant of Variational auto-encoder for hand pose estimation with multi-modal data. The model proposed by Spurr et al. chooses the input modality and the output modality pair and train the corresponding encoder and decoder pair at every iteration. Cadena, Dick, and Reid (2016) trains an auto-encoder that takes RGB images, depth images and semantic images as its network input, then the trained model can a generate complete depth image and semantic image from an RGB image and partial depth and semantic image. Ngiam et al. (2011) builds a deep-belief network structure that maps audio data and lip images into the common hidden node for audio-visual speech recognition. Hu, Li, and others (2016) extends the RBM structure to reflect the sequential characteristic of a speech dataset.

The other group comprises methods that encode the corresponding data to the latent space of each modality but enforce similarity constraints to corresponded latent vectors. Wang, van de Weijer, and Herranz (2018) trains domain specific encoders and decoders, allowing encoders and decoders from different modality to be combined, then, the model is able to generate an unseen data pair by combining the encoders and decoders. Chaudhury et al. (2017) extracts low-level representation from original data first. Then they trains auto-encoders for each modality and enforces similarity constraints to embedding spaces of each auto-encoders for correlated data pair. In (Frome et al. 2013), a model is trained to maximize the similarity of an image feature and a vectorized label to infer a proper label for a given image.

### Associative Learning inspired by the Brain

The artificial neural network is an engineering model inspired by the biological mechanism of the brain. Parameters of those networks are usually updated by Hebbian learning rule where weight connections between firing nodes for input data are strengthened (Hebb 1961). The Hopfield network and Boltzmann machine are representative examples (Ackley, Hinton, and Sejnowski 1985). The Hopfield network models associative memory of human, thus network is trained to memorize specific patterns. Even if the input is incomplete, The Hopfield network can restore incomplete data through recurrent iteration. The Boltzmann machine is a stochastic version of the Hopfield network, which can learn a latent representation for input data through its hidden nodes.

There have been many studies that investigate associative learning from the perspective of neuroscience (Bliss and Collingridge 1993; Buonomano and Merzenich 1998; Van Praag et al. 1999). In the recent study which tried to analyze associative learning at the cellular substrate level (Wang and Cui 2018; 2017), they introduce the associative memory cells to describe brain neurons which are mainly involved in integration and storage of associated signals. A brain learns associated information by enhancing the strength of the synapses between co-activated associative memory cells activated by associated signals. In this paper, we realize the cross-modal association mechanism recently proposed by Wang and Cui (2018), which assumes the comprehensive diagram based on the associative memory cells.

## Method

### Problem Statements

According to recent studies (Wang and Cui 2017; 2018), associative learning process in the brain includes intra-modal and cross-modal association processes. The intra-modal association process is to make humans familiar with single-modal sensory information. On the other hand, the cross-modal association process is accomplished to enhance the strength of the synapses connecting multi-modal information to be associated. The goal of this paper is to establish the Bayesian formulation of these two association processes and to realize them in a variational auto-encoder framework.

### Graphical Model of Intra-Modal Association

Intra-modal association is the process of memorizing single-domain information. To efficiently memorize a vast amount of information, the model needs to extract the expressive

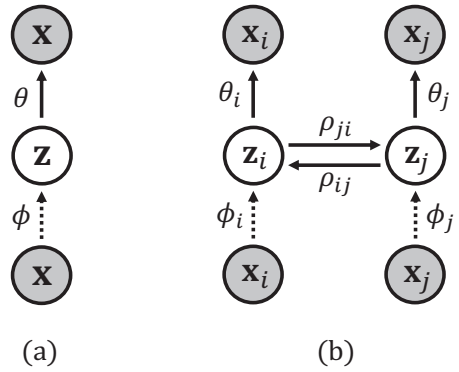


Figure 2: Graphical models for intra-modal and cross-modal association. Observable variables are illustrated as shadowed circles. $\theta, \phi, \rho$ are distribution parameters: $\theta$ for true distribution, $\phi$ for variational distribution, and $\rho$ for cross-modal association model. Subscripts denote modality. Dotted lines indicate variational approximation of true probability distribution. **(a) Intra-modal association** Latent variable $\mathbf{z}$ is obtained by $\mathbf{x}$ through $q_\phi(\mathbf{z}|\mathbf{x})$ and $\mathbf{x}$ is inferred from $\mathbf{z}$ through $p_\theta(\mathbf{x}|\mathbf{z})$ **(b) Cross-modal association between two modalities** The cross-modal association model has mutual connections between latent variables $\mathbf{z}_i$ and $\mathbf{z}_j$.

features of the data. One way to make the encoding model remember the features of the data in an unsupervised manner is to formulate a mathematical model reconstructing the original sensory data from the encoded information. Figure 2 (a) shows Bayesian graphical model to formulate the intra-modal association to memorize a distribution of the latent variable $\mathbf{z}_i$ associated with the input variable $\mathbf{x}_i$ for an observation in modality $i$. In the Bayesian framework, the objective is to infer model parameter $\theta_i$ of posterior distribution $p_{\theta_i}(\mathbf{z}_i|\mathbf{x}_i)$.

One of the most popular approaches to approximate an intractable posterior is the variational inference method. In this method, the variational distribution $q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)$ approximates the true posterior $p_{\theta_i}(\mathbf{z}_i|\mathbf{x}_i)$ by minimizing the Kullback-Leibler divergence, $D_{KL}\left(q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)\|p_{\theta_i}(\mathbf{z}_i|\mathbf{x}_i)\right)$. According to Kingma and Welling (2013), the minimization of $D_{KL}\left(q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)\|p_{\theta_i}(\mathbf{z}_i|\mathbf{x}_i)\right)$ can be replaced with the maximization of the evidence lower bound, given by

$$\begin{aligned}\mathcal{L}(q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)) = &- D_{KL}(q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)\|p_{\theta_i}(\mathbf{z}_i)) \\ &+ \mathbb{E}_{q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)}[\log p_{\theta_i}(\mathbf{x}_i|\mathbf{z}_i)],\end{aligned} \tag{1}$$

where $\mathbb{E}_{q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)}$ indicates expectation over distribution $q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)$.

## Graphical Model of Cross-Modal Association

In this section, we design a graphical model to represent the cross-modal association mechanism as in Figure 2 (b). Without loss of generality, we consider a path from modality $i$ to $j$. From observations of an associated variable pair $(\mathbf{x}_i, \mathbf{x}_j)$, the distribution parameter $\rho_{ji}$ is inferred to model the association between $\mathbf{z}_i$ and $\mathbf{z}_j$.

For a given observation pair $(\mathbf{x}_i, \mathbf{x}_j)$, the cross-posterior distribution $p_{\theta_i,\rho_{ji}}(\mathbf{z}_j|\mathbf{x}_i)$ is defined by marginalization for $\mathbf{z}_i$ as

$$p_{\theta_i,\rho_{ji}}(\mathbf{z}_j|\mathbf{x}_i) = \int p_{\rho_{ji}}(\mathbf{z}_j|\mathbf{z}_i)p_{\theta_i}(\mathbf{z}_i|\mathbf{x}_i)\, d\mathbf{z}_i. \tag{2}$$

To establish the cross-modal association model, we define a variational distribution for cross-posterior distribution $q_{\phi_i,\rho_{ji}}(\mathbf{z}_j|\mathbf{x}_i)$. Then, to infer the distribution parameters $(\phi_i, \rho_{ji})$, we minimize Kullback-Leibler divergence between $p_{\theta_j}(\mathbf{z}_j|\mathbf{x}_j)$ and $q_{\phi_i,\rho_{ji}}(\mathbf{z}_j|\mathbf{x}_i)$. To avoid clutter, subscripts for the distribution parameters are omitted in the remainders of this section. Kullback-Leibler divergence between $p(\mathbf{z}_j|\mathbf{x}_j)$ and $q(\mathbf{z}_j|\mathbf{x}_i)$ is given by

$$D_{KL}(q(\mathbf{z}_j|\mathbf{x}_i)\|p(\mathbf{z}_j|\mathbf{x}_j)) = \log p(\mathbf{x}_j) - \mathcal{L}(q(\mathbf{z}_j|\mathbf{x}_i)), \tag{3}$$

where

$$\mathcal{L}(q(\mathbf{z}_j|\mathbf{x}_i)) = \int q(\mathbf{z}_j|\mathbf{x}_i) \log \frac{p(\mathbf{x}_j)p(\mathbf{z}_j|\mathbf{x}_j)}{q(\mathbf{z}_j|\mathbf{x}_i)}\, d\mathbf{z}_j. \tag{4}$$

Since log-evidence $\log p(\mathbf{x}_j)$ is independent to the model parameter, the target problem is identical to maximizing the evidence lower bound $\mathcal{L}(q(\mathbf{z}_j|\mathbf{x}_i))$. With probabilistic tricks, $\mathcal{L}(q(\mathbf{z}_j|\mathbf{x}_i))$ can be decomposed as following.

$$\begin{aligned}\mathcal{L}(q(\mathbf{z}_j|\mathbf{x}_i)) = &- D_{KL}(q(\mathbf{z}_j|\mathbf{x}_i)\|p(\mathbf{z}_j)) \\ &+ \mathbb{E}_{q(\mathbf{z}_j|\mathbf{x}_i)}[\log p(\mathbf{x}_j|\mathbf{z}_j)].\end{aligned} \tag{5}$$

The detail derivation is given in Appendix A of supplementary document.In Eq. (5), the first term is a negative KL divergence term that leads $\mathbf{z}_j$ given by $\mathbf{x}_j$ to have similar distribution with a prior distribution of target modality. The expectation term in Eq. (5) minimizes the reconstruction error of decoded output from $\mathbf{z}_j$ fired from $\mathbf{x}_i$, which also promotes the inference for $\rho_{ji}$. By the similar steps, we can easily derive the opposite association from modality $j$ to modality $i$.

## Realization: Cross-Modal Association Network

We accomplish a realization of the aforementioned intra-modal and cross-modal association models by extending the Variational Auto-Encoder framework (VAE) (Kingma and Welling 2013). Figure 3 illustrates the proposed cross-modal association network for modality $i$ and $j$. Although only two modalities are considered in this paper, the proposed model can be applied to the association among three or more modalities also. In the proposed structure, the *encoder* produces the parameter of $q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i = x_i)$, and the *decoder* produces the parameter of $p_{\theta_i}(\mathbf{x}_i|\mathbf{z}_i = z_i)$. The encoder and decoder are realized by deep neural networks. Likewise, the latent space associating models $p_{\rho_{ji}}(\mathbf{z}_j|\mathbf{z}_i = z_i)$ and $p_{\rho_{ij}}(\mathbf{z}_i|\mathbf{z}_j = z_j)$ are also realized by deep neural networks, which are called by *associator*. Thus, the intra-modal association network contains several auto-encoders, each of which considers one of the multiple modalities only. The latent spaces of the auto-encoders are connected by *associators* in a pairwise manner, which configure the cross-modal association network.

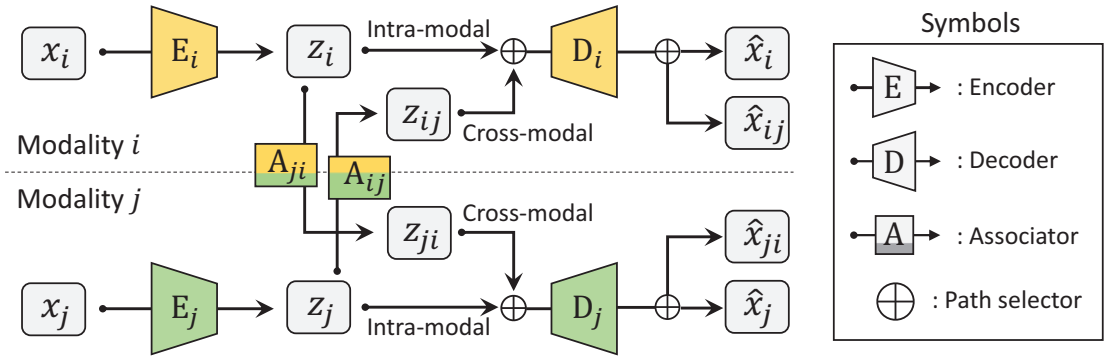


Figure 3: Overall structure of the proposed method for the modalities $i$ and $j$. For an observation sample $x_i$ for the variable $\mathbf{x}_i$, the intra-modal network of modality $i$ encodes $x_i$ into a latent vector $z_i$ for the latent variable $\mathbf{z}_i$ through the encoder $E_i$ and decodes $z_i$ to $\hat{x}_i$ through decoder $D_i$. In the case of association from modality $i$ to $j$, a sample $x_i$ is encoded to $z_{ji}$ through the encoder $E_i$ and the associator $A_{ji}$. Then, $z_{ji}$ is decoded to $\hat{x}_{ji}$ through $D_j$. The procedure for the opposite direction is performed in the same way.

The proposed network is trained in the two phases: intra-modal training phase and cross-modal training phase. In the intra-modal training phase, the auto-encoder in each modality is trained separately by minimizing the approximated version of the negative evidence lower bound in Eq. (1). As derived by Kingma and Welling (2013), variational distributions are assumed by the centered isotropic multivariate Gaussian distribution. For a given observation sample $x_i$, the encoder $E_i$ produces the mean $\mu_{\phi_i}$ and the variance $\sigma_{\phi_i}$ for a Gaussian distribution of $q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i = x_i)$. Then, the latent vector $z_i$ is sampled as $z_i = \mu_{\phi_i} + \sigma_{\phi_i} * \epsilon$ and $\epsilon \backsim N(0, I)$. Similarly, the decoder $D_i$ also produces the mean $\mu_{\theta_i}$ and the variance $\sigma_{\theta_i}$ for a Gaussian distribution of $p_{\theta_i}(\mathbf{x}_i|\mathbf{z}_i = z_i)$. Then, the reconstruction vector $\hat{x}_i$ is sampled as $\hat{x}_i = \mu_{\theta_i} + \sigma_{\theta_i} * \epsilon$ and $\epsilon \backsim N(0, I)$.

Using the samples, the empirical loss for auto-encoder can be derived as

$$\begin{aligned}\mathcal{L}_{int}(\theta_i, \phi_i; x_i) &= -\mathbb{E}_{q_{\phi_i}(\mathbf{z}_i|x_i)}[\log p_{\theta_i}(x_i|\mathbf{z}_i)] \\ &+ \lambda_{int}^{'} D_{KL}(q_{\phi_i}(\mathbf{z}_i|x_i)||p_{\theta_i}(\mathbf{z}_i)), \\ &= ||x_i - \hat{x}_i||_2^2 \\ &- \lambda_{int} \sum_{k}^{H} (1 + \log \sigma_{\phi_i(k)}^2 - \mu_{\phi_i(k)}^2 - \sigma_{\phi_i(k)}^2).\end{aligned} \tag{6}$$

where $\lambda_{int}$ is a user-defined parameter and $H$ is the dimension of the latent variable $\mathbf{z}_i$. $\mu_{\phi_i(k)}$ and $\sigma_{\phi_i(k)}^2$ denote the $k$-th element of $\mu_{\phi_i}$ and $\sigma_{\phi_i}^2$. The detail derivation presents in Appendix B of supplementary document.

After the convergence of the intra-modal training phase, the following cross-modal training phase proceeds to train the associators while freezing the weights of the auto-encoders. In the same way as in the intra-modal training phase, for a given observation pair $x_i$ and $x_j$, the encoders $E_i$ and $E_j$ produce the latent vectors $z_i$ and $z_j$, respectively. In addition, associators $A_{ji}$ and $A_{ij}$ produce the latent vectors $z_{ji}$ and $z_{ij}$ using inputs $z_i$ and $z_j$, respectively. Thereafter, the decoders $D_i$ and $D_j$ produce the reconstruction vectors $\hat{x}_{ij}$ and $\hat{x}_{ji}$ from $z_{ij}$ and $z_{ji}$, respectively.

Using the samples, the empirical loss for $A_{ji}$ is designed according to Eq. (5) as follows:

$$\begin{aligned}\mathcal{L}_{crs}(\rho_{ji}; x_i, x_j) &= -\mathbb{E}_{q_{\rho_{ji}}(\mathbf{z}_{ji}|x_i)}[\log p_{\theta_j}(x_j|\mathbf{z}_{ji})] \\ &+ \lambda_{crs}^{'} D_{KL}(q_{\phi_i,\rho_{ji}}(\mathbf{z}_{ji}|x_i)||p_{\theta_j}(\mathbf{z}_{ji})) \\ &= ||x_j - \hat{x}_{ji}||_2^2 \\ &- \lambda_{crs} \sum_{k}^{H} (1 + \log \sigma_{\rho_{ji}(k)}^2 - \mu_{\rho_{ji}(k)}^2 - \sigma_{\rho_{ji}(k)}^2).\end{aligned} \tag{7}$$

where $\lambda_{crs}$ is a user-defined parameter and $H$ is the dimension of the latent variable $\mathbf{z}_{ji}$. $(\mu_{\rho_{ji}}, \sigma_{\rho_{ji}}^2)$ are parameters for Gaussian distribution $q_{\phi_i,\rho_{ji}}(\mathbf{z}_{ji}|x_i)$ produced by $A_{ji}$. The detail derivation presents in Appendix B of supplementary document.

The loss $\mathcal{L}_{crs}(\rho_{ij}; x_i, x_j)$ for $A_{ij}$ is given in the same form of $A_{ji}$ except the index. Note that all $\mu$'s and $\sigma$'s in Eq. (6) and Eq. (7) are the functions of weights ($\mathbf{w}$) in encoders, decoders, or associators. Hence, the weights of the proposed network are trained by the negative direction of the gradient of the losses with respect to the weights ($\nabla_{\mathbf{w}}\mathcal{L}(\cdot)$).

## Advantages of Proposed Method

Owing to the newly introduced *associator*, the proposed model can associate heterogeneous modalities effectively. Reckless coalescence of heterogeneous data may have a fatal impact on associative learning such as the problem that shared latent vectors can be biased to the dominant modality. However, in our model, the associator acts as a translator between heterogeneous modalities and thus the characteristics of each latent space are preserved. Furthermore, in contrast to the existing models which adopt a shared latent space for the different modalities (Wu and Goodman 2018; Spurr et al. 2018; Cadena, Dick, and Reid 2016; Ngiam et al. 2011), our structure can provide a flexible dimensional

encoding in each latent space depending on the complexity of each modality. This provides better cross-modal data association results.

The proposed model easily incorporates additional modalities while maintaining the existing modalities. That is, a new modality can be added via training of only a new associator between an existing auto-encoder and a new auto-encoder. Though the associator only associates the new modality with one of the existing modalities, the model can associate the new modality with the rest of the modality by passing through multiple associators.

Finally, in contrast to existing models which always require paired data for cross-modal association, our structure can train the associator with the only small amount of paired data in a semi-supervised manner after learning each auto-encoder using unpaired data independently. Since obtaining paired data for cross-modal association is more expensive than obtaining unpaired data, our model is cost-effective. Furthermore, our model is plausible in that, when a person learns a cross-modal association, the paired examples are rarely given by a teacher after the person has become familiar with each modality via self-experience without a teacher.

The aforementioned advantages of the proposed structure are validated in the following experiment section.

## Experiment

The implementation details for network architectures are provided in Appendix C of the supplementary document.

### Datasets

**Google Speech Commands (GSC) (Warden 2018):** As the data for the auditory modality, we used the GSC dataset, which consists of 105,829 audio samples containing utterances of 35 short words. Each audio sample is one-second-long and encoded with a sampling rate of 16KHz. Among 35 words, we chose 14 words, including words for each digit ('ZERO' to 'NINE') and four traffic commands ('GO,' 'STOP,' 'LEFT,' 'RIGHT'). The chosen set has 54,239 samples. We extracted the Mel-Frequency Cepstral Coefficient (MFCC) from each audio clip to generate an audio feature. MFCC has been widely used in the processing of voice data because it reflects the human auditory perception mechanism well (Muda, Begam, and Elamvazuthi 2010; Logan and others 2000; Rubin et al. 2016). The resulting features are $40 \times 101$ matrices. We randomly divided the original dataset into training, validation and test sets at the ratio of 8:1:1.

**German Traffic Sign Recognition Benchmark (GTSRB) (Stallkamp et al. 2011):** For the visual data that correspond to the traffic commands in GSC, we used the GTSRB dataset, which consists of 51,839 RGB color images illustrating 42 kinds of traffic signals. In particular, to evaluate the performance on pairs of traffic sign images and voice commands in GSC, we chose four pair sets, where each pair set has similar semantic meaning, i.e., ('Ahead only,' 'GO'), ('No entry for vehicle,' 'STOP'), ('Turn left and ahead,' 'LEFT'), and ('Turn right and ahead,' 'RIGHT'). The first and the second element are taken from GTSRB and GSC dataset, respectively. Then, to prevent the four signs from occupying the entire latent space, we chose additional sign images in GTSRB such as 'No overtaking,' 'Entry to 30kph zone,' 'Prohibit overweighted vehicle,' 'No-waiting zone,' and 'Roundabout'. The chosen set includes 10,709 samples. All of the chosen signs have a circular backboard. The size of each image varies from $15 \times 15$ to $250 \times 250$ pixels for each RGB channel in the original dataset. In our experiments, we resized all images into $52 \times 52$.

**MNIST (LeCun et al. 1998):** We used the MNIST dataset as the corresponding visual data to the GSC for each digit. The MNIST consists of center-aligned $28 \times 28$ gray-scale images for handwritten digits from 0 to 9. The dataset contains 60k and 10k samples for the training set and testing set, respectively.

**SVHN (Netzer et al. 2011):** We used the SVHN dataset as another visual modality. Even though SVHN and MNIST are equally categorized by digits, their capturing environments are very different from each other. The SVHN consists of $32 \times 32$ RGB images for digits from 0 to 9. The dataset contains 73257 and 26032 samples for the training set and testing set, respectively.

**Fashion-MNIST (F-MNIST) (Xiao, Rasul, and Vollgraf 2017):** To validate that the proposed model can associate even not semantically related datasets, we used the F-MNIST dataset. We associated F-MNIST with the MNIST and the GSC dataset. After this association learning, we can imagine the clothing items (F-MNIST) from their numberings (MNIST). The F-MNIST consists of center-aligned $28 \times 28$ gray-scale images assigned with a label from 10 kinds of clothing such as T-shirt, Trouser and Sneaker. The dataset contains 60k and 10k samples for the training set and testing set, respectively.

### Evaluation Metric

Since aforementioned datasets have no direct matching relationships, we cannot measure cross-likelihood $p(\mathbf{x}_1|\mathbf{x}_2)$ for paired sample $(\mathbf{x}_1, \mathbf{x}_2)$ used in recent works (Wu and Goodman 2018; Suzuki, Nakayama, and Matsuo 2016). In our work, we used the classification accuracy for the reconstructed results as the evaluation metric of the association models. The quality of results reconstructed by an association model can be a valid measure to evaluate the association model since the quality of the reconstructed results is acceptable to both the human and the classifier. Table 2 shows the performance of the classifiers trained with the each dataset, which shows sufficient performance for evaluating the reconstructed results of the compared encoders. For the GSC dataset, we get performance comparable to the 88.2% (Warden 2018).

### Intra-Modal Association

As mentioned in the problem statements section, it is also essential for the proposed model to learn the intra-modal association that encodes single modal input data into the latent space. For a fair comparison to existing works, we trained

Table 1: Evaluation of cross-modal association models. Accuracy is measured by the classifier for the reconstructed data of the target modality from the input data of the other modality. Bold font denotes the best performance for each case.

| | Classification Accuracy (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | SVHN → MNIST | F-MNIST → MNIST | MNIST → F-MNIST | MNIST → GSC | GSC → MNIST | F-MNIST → GSC | GSC → F-MNIST | GTSRB → GSC | GSC → GTSRB |
| VAE | 38.73 | 47.36 | 41.86 | 10.34 | 28.61 | 12.83 | 22.18 | 35.93 | 19.44 |
| VAE-CG | 66.05 | 82.41 | 83.84 | 32.46 | 66.62 | 29.87 | 63.79 | 28.43 | 55.00 |
| JMVAE | 64.93 | **83.49** | 88.14 | 28.15 | 62.31 | **47.58** | 51.23 | 41.02 | 65.18 |
| CVA | 57.01 | 76.51 | 85.88 | 24.61 | 65.04 | 18.70 | 59.73 | 31.02 | 77.78 |
| MVAE | 31.18 | 62.65 | 77.62 | 23.04 | 46.70 | 13.52 | 33.24 | 28.06 | 69.17 |
| **ours** | **74.20** | 82.02 | **94.26** | **59.47** | **88.66** | 43.95 | **77.84** | **58.89** | **77.87** |
| **ours-flex** | - | - | - | - | - | - | - | **61.39** | **80.56** |

Table 2: Performance of the classifier trained with the each dataset and reconstruction performance of VAE. dim($\mathbf{z}$) denotes the dimension of latent space of VAE.

| Dataset | Acc (%) | VAE (%) | dim($\mathbf{z}$) |
|---|---|---|---|
| MNIST | 97.97 | 96.12 | 64 |
| F-MNIST | 89.22 | 80.54 | 64 |
| SVHN | 93.73 | 78.22 | 64 |
| GSC | 88.65 | 81.93 | 64 |
| GTSRB | 98.53 | 95.70 | 64 |
| GTSRB | - | 95.50 | 256 |

encoders and decoders for each dataset with the fixed dimension of latent space ($\dim(\mathbf{z}) = 64$). In addition, to show the advantages of the proposed model where the dimension of the latent space can be flexibly designed according to the complexity of target modality, we trained additional auto-encoder whose latent space dimension is $256$ for GTSRB dataset.

Table 2 shows the performance of the classifiers and the intra-modal association network implemented by VAE. Performance of VAE is also measured by the classifier on the results reconstructed by VAE. As shown in the table 2, the voice data in the GSC dataset shows much degraded accuracy, which means that the voice data are hard to be reconstructed than other modalities. Since F-MNIST has confused classes such as pullover, coat, and shirt, performance on F-MNIST dataset is also degraded.

## Cross-Modal Association

Cross-modal problem is defined to develop a model that can generate the sample of target modality from a given sample of source modality, where samples are semantically associated. We evaluated the proposed model on five scenarios: (1) Association between MNIST and GSC, (2) GTSRB and GSC, (3) F-MNIST and GSC, (4) F-MNIST and MNIST, (5) SVHN and MNIST. Scenario (1), (2) and (3) are for association between heterogeneous datasets, i.e. voice and image datasets. Scenario (3) and (4) is for association between datasets which have no semantic relations between classes. Scenario (5) is for association between semantically related datasets, through two datasets have different dataset characteristics like image size. In order to train cross-modal association, we used randomly paired training samples from each dataset belonging to the correlated class. For example, we paired a randomly chosen sample in '0' class of MNIST dataset with a randomly chosen sample in 'ZERO' class of GSC dataset.

To evaluate the proposed associator, the following methods were compared: **VAE** and **VAE-CG** are variants of the standard VAE. When training VAE, we constructed the training data with vectors concatenated with data from two associated modalities, whereas one modality in the concatenated vector for 50% of training data was set to zero-vector to learn the case of the missing modality. **VAE-CG** is trained to generate the target modality sample from a given input sample of other modality. VAE-CG has to be trained only by supervised data with input and output pairs. Joint Multimodal Variational Auto-Encoder (**JMVAE**) (Suzuki, Nakayama, and Matsuo 2016) has two kinds of latent spaces: one is for each modality and the other is for jointly encoding of two modalities. The joint latent space is shared for association between two modalities. The training for encoding in the joint latent space is done to minimize Kullback-Leibler divergence between the latent vector of each encoder and the joint latent vector of the joint encoder. In comparison, the hyper-parameter $\alpha$ was set to 0.01 for whole scenarios. Cross-modal Variational Auto-Encoder (**CVA**) (Spurr et al. 2018) is an extension of VAE for cross-modal data. In CVA, the latent space is shared between two modalities. In the training process, the selected sample pair are trained alternately throughout iteration. Multimodal Variational Auto-Encoder (**MVAE**) (Wu and Goodman 2018) is also a variant of VAE for cross-modal data. MVAE uses the standard VAE for each modality, but each latent space is associated via a shared latent space expressing the unified distribution of the association modalities. We trained MVAE by using the sub-sampled training paradigm presented in their paper.

To evaluate the flexibility of encoding dimension in our model, we have conducted an experiment where each modality is encoded in a different dimensional space from the other. **ours-flex** has large dimension of latent space for GT-

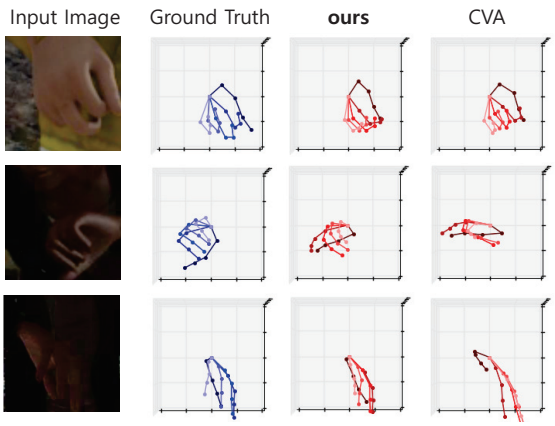


Figure 4: Qualitative results for 3D hand pose estimation on RHD dataset. Each column corresponds to input images, ground truth 3D keypoints, estimated 3D keypoints in order from left to right.

SRB dataset ($dim(\mathbf{z}) = 256$). Except for ours-flex, all compared models use the same VAE of which the latent space dimension is 64.

Table 1 shows the evaluation result of the proposed model and the compared models for the cross-modal association. The proposed model accomplishes significant enhancement from the compared algorithms for most of the scenarios. Interestingly, in the challenging scenarios such as the association between heterogeneous modalities, for instance, between voice (GSC) and image data (MNIST, GTSRB), the proposed model achieves a remarkable improvement compared to the existing methods. The qualitative results of our model are presented in Appendix D of supplementary material.

### Application: 3D Hand pose estimation

We have conducted additional experiments for 3D hand pose estimation on Rendered Hand pose Dataset (RHD) (Zimmermann and Brox 2017). RHD dataset provides $320 \times 320$ RGB image, depth map, segmentation map and 21 keypoints for each hand. The dataset contains 41258 training and 2728 testing samples. The association target is to generate 3D keypoints from the RGB image. The evaluation metric is the average End-Point-Error (EPE), which measures Euclidean distance between ground truth keypoints and estimated keypoints. We used the same encoders and decoders structure to CVA and added our associator. The proposed model achieves **13.15**, which outperforms recent 3D hand pose estimation algorithms such as CVA (19.73) and HPS (30.42) (Zimmermann and Brox 2017). Figure 4 shows qualitative results for 3D hand pose estimation.

### Semi-supervised Learning

We conducted an additional experiment to verify the effectiveness of the proposed associator in semi-supervised learning. Figure 5 illustrates a trend of performance variation depending on the proportion of paired data, from 100% to 1% in the GSC $\rightarrow$ MNIST scenario. The result shows that the proposed associator can achieve eminent performance with only a small proportion of paired data (5%) in a semi-supervised manner.

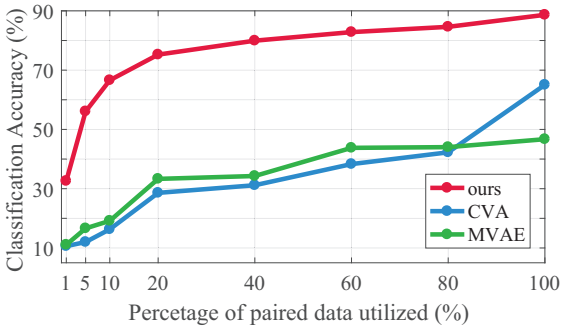


Figure 5: Semi-supervised learning. Performance variation while reducing the proportion of paired data from 100% to 1% in the GSC $\rightarrow$ MNIST. Our method can achieve much better performance with only 5% paired data than the existing methods with 80% paired data.

Table 3: Performance of the proposed model in the case of cascading association and direct association.

| GSC $\rightarrow$ MNIST | GSC $\rightarrow$ F-MNIST $\rightarrow$ MNIST |
|---|---|
| 88.66 | 76.99 |

### Scalability

The proposed structure can easily expand a new modality while maintaining the existing modalities. That is, a new modality can be added via training of only a new associator between an existing auto-encoder and the new auto-encoder. Since the associator connect only two latent spaces, if the existing network associates $N$ modality, $N$ associators need to be trained newly. In our model, this inefficiency can be mitigated by cascading association through multiple associators. Table 3 compares the results of cascading association and direct association for the example of MNIST, F-MNIST and GSC dataset. The F-MNIST is utilized as a medium between MNIST and GSC. Although the cascading association has some performance degradation, it still has good performance compared to other algorithms presented in Table 1.

## Conclusion

We proposed a novel multi-modal association network structure that consists of multiple modal-specific auto-encoders and associators for cross-modal association. By adopting the associators, the proposed multi-modal network can incorporate new modalities easily and efficiently while preserving the encoded information in the latent space of each modality. In addition, the proposed network can effectively associate even heterogeneous modalities by designing each latent space independently and can be trained by a small amount of paired data in a semi-supervised manner. Based

on the validation of our structure in experiments, future work can attempt to implement a large-scale multi-modal association network for practical use.

## Acknowledgement

This work was supported by Next-Generation ICD Program through NRF funded by Ministry of S&ICT [2017M3C4A7077582] and ICT R&D program of MSIP/IITP [No.B0101-15-0552, Predictive Visual Intelligence Technology].

## References

Ackley, D. H.; Hinton, G. E.; and Sejnowski, T. J. 1985. A learning algorithm for boltzmann machines. *Cognitive science* 9(1):147–169.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35(8):1798–1828.

Bliss, T. V., and Collingridge, G. L. 1993. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* 361(6407):31.

Buonomano, D. V., and Merzenich, M. M. 1998. Cortical plasticity: from synapses to maps. *Annual review of neuroscience* 21(1):149–186.

Cadena, C.; Dick, A. R.; and Reid, I. D. 2016. Multi-modal auto-encoders as joint estimators for robotics scene understanding. In *Robotics: Science and Systems*.

Chaudhury, S.; Dasgupta, S.; Munawar, A.; Khan, M. A. S.; and Tachibana, R. 2017. Conditional generation of multi-modal data using constrained embedding space mapping. *arXiv preprint arXiv:1707.00860*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2121–2129.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Hebb, D. O. 1961. *The organization of behavior*. na.

Hu, D.; Li, X.; et al. 2016. Temporal multimodal learning in audio-visual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3574–3582.

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

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Lim, J.; Yoo, Y.; Heo, B.; and Choi, J. Y. 2018. Pose transforming network: Learning to disentangle human posture in variational auto-encoded latent space. *Pattern Recognition Letters*.

Logan, B., et al. 2000. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, volume 270, 1–11.

Muda, L.; Begam, M.; and Elamvazuthi, I. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.

Rubin, J.; Abreu, R.; Ganguli, A.; Nelaturi, S.; Matei, I.; and Sricharan, K. 2016. Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients. In *Computing in Cardiology Conference (CinC), 2016*, 813–816. IEEE.

Senocak, A.; Oh, T.-H.; Kim, J.; Yang, M.-H.; and Kweon, I. S. 2018. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4358–4366.

Spurr, A.; Song, J.; Park, S.; and Hilliges, O. 2018. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 89–98.

Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2011. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *IEEE International Joint Conference on Neural Networks*, 1453–1460.

Suzuki, M.; Nakayama, K.; and Matsuo, Y. 2016. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Van Praag, H.; Christie, B. R.; Sejnowski, T. J.; and Gage, F. H. 1999. Running enhances neurogenesis, learning, and long-term potentiation in mice. *Proceedings of the National Academy of Sciences* 96(23):13427–13431.

Wang, J.-H., and Cui, S. 2017. Associative memory cells: formation, function and perspective. *F1000Research* 6.

Wang, J.-H., and Cui, S. 2018. Associative memory cells and their working principle in the brain. *F1000Research* 7.

Wang, K.; Yin, Q.; Wang, W.; Wu, S.; and Wang, L. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*.

Wang, Y.; van de Weijer, J.; and Herranz, L. 2018. Mix and match networks: encoder-decoder alignment for zero-pair image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5467–5476.

Warden, P. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.

Wu, M., and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, 5580–5590.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yoo, Y.; Yun, S.; Chang, H. J.; Demiris, Y.; and Choi, J. Y. 2017. Variational autoencoded regression: high dimensional regression of visual data on complex manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3674–3683.

Zimmermann, C., and Brox, T. 2017. Learning to estimate 3d hand pose from single rgb images. Technical report, arXiv:1705.01389. https://arxiv.org/abs/1705.01389.