Private-Shared Disentangled Multimodal VAE for Learning of Latent Representations

Mihee Lee
Rutgers University
Piscataway, NJ, USA
ml1323@rutgers.edu

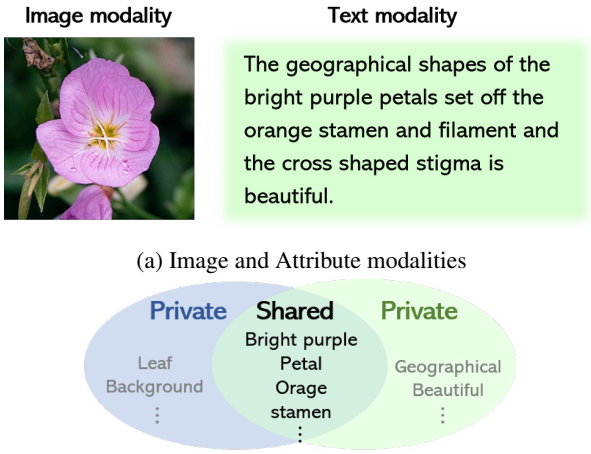
Vladimir Pavlovic
Rutgers University
Piscataway, NJ, USA
vladimir@cs.rutgers.edu

Abstract

Multi-modal generative models represent an important family of deep models, whose goal is to facilitate representation learning on data with multiple views or modalities. However, current deep multi-modal models focus on the inference of shared representations, while neglecting the important private aspects of data within individual modalities. In this paper, we introduce a disentangled multi-modal variational autoencoder (DMVAE) that utilizes disentangled VAE strategy to separate the private and shared latent spaces of multiple modalities. We demonstrate the utility of DMVAE two image modalities of MNIST and Google Street View House Number (SVHN) datasets as well as image and text modalities from the Oxford-102 Flowers dataset. Our experiments indicate the essence of retaining the private representation as well as the private-shared disentanglement to effectively direct the information across multiple analysis-synthesis conduits.

1. Introduction

Representation learning is a key step in the process of data understanding, where the goal is to distill interpretable factors associated with the data. Representation learning approaches typically focus on data observed in a single modality, such as text, images, or video. Nevertheless, most real world data comes from processes that manifest itself in multiple views or modalities. In computer vision, image-based data is typically accompanied with text description to promote understanding of its latent factors. For example, in Fig. 1a, an image of a flower is augmented with captions describing the detailed characteristics of the flower. To study about the flower, the background of the image is unnecessary but the additional information of text description is helpful. Therefore, accurate modeling of the underlying data representation has to consider both the **private** aspects of individual modalities as well as what those modalities **share**, as illustrated in Fig. 1b.



(a) Image and Attribute modalities

(b) Private and shared factors

Figure 1: (a) Example of bimodal data, where one modality, `I`, is an image of a flower and the other, `T`, represents a textual caption describing the flower. (b) Only some of the factors, here aligned with the caption for simplicity, are shared by both modalities in `Shared` $\mathtt{I} \cap \mathtt{T}$. Other factors are private to individual modalities, grouped in separate `Private` spaces. By definition, the three spaces are **disentangled** from each other.

In this paper, we propose a generative variational model that can learn both the private and the shared latent space of each modality, with each latent variable attributed to a disentangled representational factor. The model extends the well-known family of Variational AutoEncoders (VAEs) [10] by introducing separate shared and private spaces, whose representations are induced using pairs of individual modality encoders and decoders. To create the shared representation, we impose consistency of representations using a product-of-experts (PoE) [6] inference network.

While the shared latent representation can be used to model the compatibility of the two modalities, the representation can also enable cross-reconstruction of one modality from another. We demonstrate that this essential task can and has to be effectively combined in an end-to-end learning framework with the private-shared disentangled VAE, resulting in our novel disentangled multi-modal variational autoencoder (DMVAE).

We apply the DMVAE to two multi-modal representation learning problems. In the first setting, we consider the problem of learning the shared/private generative representations of digit images from two datasets of different styles, where the shared property becomes the digit class and the private property becomes the style of each dataset. In the second setting, we generalize the modality types further into images and text, aiming to model the joint representation of flower appearance and the corresponding captions, which describe the visual characteristics of the flower. We show that DMVAE excels both as an analysis tool as well as the (cross) synthesis generative model.

Our main contributions are as follows.

- We segregate the latent representation space into the union of the private and the shared spaces. We show that the private latent space is critical for modeling the disjoint properties of each modality while the shared latents enable linking and cross-synthesis across domains, as signified in the experiments in Sec. 5.1 and Sec. 5.3.
- We improve the compatibility between modalities by introducing the cross-VAE task (loss), whose aim is the cross-modal reconstruction through the shared latent space. The impact of the cross-VAE direction, induced by the properties of the linked datasets is examined in the ablation study in Sec. 5.3.
- By applying our model to (image, text) as well as (image, image) representation modeling problems, we demonstrate the universal applicability and effectiveness of the the DMVAE framework, across different data types.

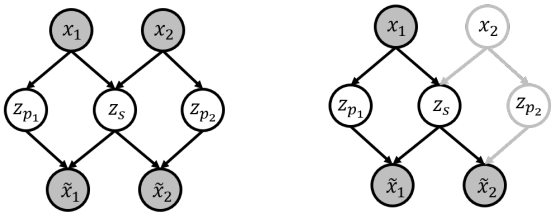
2. Related Work

Several lines of related work can be linked to our proposed DMVAE. Our modeling task is intimately related to image-to-image translations problems, the task of translating between different representations of one image, such as the sketch-photographic, summer-winter, *etc*., views of a visual object or a scene. We begin by reviewing relevant prior work in this area, subsequently extending it to multi-modal learning that can take any input data type.

Image-to-image translation. Multiple research efforts have attempted to solve the image-to-image translation problem by framing it as a two-modality matching setup. [23] utilize GAN [5] framework, which takes the image from one modality as the fake sample against another modality. They combine VAE into GAN so that the latent space encodes information about the ground truth outputs, rectifying the mode collapse problem. The diversity of the output that the latent factor can provide is enhanced with the latent regressor GAN, which tries to generate output from randomly drawn latent factors and then attempts to recover the latent code again.

[4] disentangles the latent representation into two parts; the shared between the two modalities and the exclusive within each modality. Using only the shared part of the representation in the image translation, the domain-specific variation is reduced. Furthermore, adding noise with the shared latent factors for generation improves the diversity in translation between images. However, the paired input images are necessary to train these models. [12, 8] show that cross-domain mapping and cross-cycle consistency enable an effective style transfer using unpaired data. They separate a domain-invariant content and a domain-specific attribute (style) latent space using an adversarial loss. [2] separates private and shared networks in each domain utilizing DANN [16] to make it possible for the unlabeled target domain to learn the transferred information from the labeled source more effectively with only the latent codes from the shared network. Although these methods are able to achieve realistic and diverse image translation, they make use of the strong within-image-modality conditioning, which may fail when the modalities exhibit vastly different properties (*e.g*., text and image).

Multi-modal Learning. Several prior works have considered the problem of modeling multi-modal data using gen-



(a) Full-modality (b) Missing-modality

Figure 2: Unrolled graphical model representation of DMVAE. The gray circles illustrate observed variables. z_{p_1}, z_{p_2} denote the private latents of modalities x_1, x_2. z_s denotes the shared latents between two modalities. $\tilde{x}_1, \tilde{x}_2$ denote reconstructed views, which should match the observed data in this unrolled generative model. (b) illustrates the missing modality instance network, which is critical for test-time inference of x_2 from x_1. We elaborate on the inference in the missing modality in Sec. 4.2.

erative VAE-inspired models. JMVAE [19] exploits the joint inference network $q(z|x_1, x_2)$ to learn the interaction of two modalities, x_1 and x_2. To address the missing modality problem, where some of the data samples are not paired (i.e., do not have both views present), they train inference networks $q(z|x_1)$, $q(z|x_2)$ in addition to the bimodal inference model $q(z|x_1, x_2)$, and then minimize the distance between uni- and multi-modal based latent distribution. JVAE [20] adopts a product-of-expert (PoE) [6] for the joint posterior $q(z|x_1, x_2)$ of multi modalities in the inference network. The approach leverages the unimodal inference networks, whose predictions are constrained and made consistent through the PoE. JVAE trains the model with two-stage process to handle both paired and missing modality data. Due to this fact, the number of required inference networks increases exponentially for more than two modalities. To alleviate the inefficiency of JVAE, MVAE [21] considers only partial combination of observed modalities. This helps reduce the number of parameters and increase the computational efficiency of learning. [18] applies Mixture-of-Expert (MoE) to jointly learn the shared factors across multi-modalities. Though they introduces the concept of the private and shared information of multi-modalities, it is implicitly conceived. Moreover, the use IWAE for the approximation makes the contribution of MoE vague.

However, the aforementioned prior works based on VAE use a single latent space to represent the multi-modal data. Although [2] attempt to separate the private and shared networks, their method uses deterministic latent features. Moreover, they require target label information to train their model. Within one common latent space under the VAE framework, modality-specific factors could be entangled with the shared factors across all moralities, reducing the ability of these generative models to represent the data and infer the "true" latent factors.

In this paper, we address these challenges by explicitly separating the shared from the disjoint private spaces, using individual inference networks to achieve this goal. This is illustrated in Fig. 2. In subsequent sections, we review the core VAE framework, followed by the details of our DMVAE modeling approach, and the experimental evaluation.

3. Background

Our DMVAE framework builds upon the VAE model of [10]. We first highlight the relevant aspects of VAE-based models, which we then leverage to construct the DMVAE in Sec. 4.

Variational Autoencoder. A variational autoencoder (VAE) [10] implements variational inference for the latent variable via autoencoder structure. The objective of the VAE is to maximize the marginal distribution $p(x) = \int p_\theta(x|z)p(z)dx$ which is, however, intractable. Thus, VAE introduces the evidence lower bound (ELBO) which uses an approximated recognition model $q_\phi(z|x)$ instead of the intractable true posterior. It maximizes $\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]$ while minimizing $\mathrm{KL}(q_\phi(z|x), p(z))$.

$$\begin{aligned}\log p(x) &\geq \mathbb{E}_{p(x)}\left[ELBO(x;\theta,\phi)\right] \\ &= \mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - KL(q_\phi(z|x)||p(z))\right] \quad (1)\end{aligned}$$

$q_\phi(z|x)$ and $p_\theta(x|z)$ are represented as encoder and decoder in the network with the learning parameter ϕ and θ, respectively. The first term of ELBO (Eq. (1)) is the reconstruction error and the second term plays a role of regularizer not to be far from the prior distribution $p(z)$. We next discuss in more detail the effect of the second term on disentanglement.

4. DMVAE Framework

In this section, we introduce the new DMVAE model. Sec. 4.1 describes the architecture of private and shared latent spaces within the disentangled representation. In Sec. 4.2 and Sec. 4.3, we define the DMVAE inference models, accompanied with the learning objective in Sec. 4.4.

4.1. Private / Shared-Disentangled Multi-Modal VAE

The assumption in this paper is that under the multi-modal description of a concept, the latent space of the concept is divided into a private space of each modality and one shared space across all modalities[1]. Our goal is to obtain well-separated private and shared spaces. This separation is critical; the shared latent space can only transfer the information common across modalities, but it will fail to model the individual aspects of the modalities. In a generative model, such as the VAE, modeling the private factors is critical as those factors enable both the high fidelity of the data reconstruction as well as the improved separation (disentanglement) of the latent factors across modalities.

Our model is illustrated in Fig. 3 for the case of two modalities. Given paired *i.i.d.* data $\{(x_1, x_2)\}$, we infer the latents $z_1 \sim q_{\phi_1}(z|x_1), z_2 \sim q_{\phi_2}(z|x_2)$, where ϕ_1, ϕ_2 are the parameters of each individual modal inference network. We assume the latents can be factorized into $z_1 = [z_{p_1}, z_{s_1}]$ and $z_2 = [z_{p_2}, z_{s_2}]$, where z_{p_1}, z_{p_2} represent the private latents of modalities x_1, x_2, respectively, and z_{s_1}, z_{s_2} represent the shared latents, which are to model the commonality between the two modalities.

For the desired shared representation in z_{s_1}, z_{s_2} we seek to effectively make $z_{s_1} = z_{s_2}$. We describe how to accomplish this using a PoE-based consistency model in Sec. 4.2, which approximates the shared inference network $p(z_s|x_1, x_2)$.

[1] Other more intricate representations of private and shared spaces may arise in the presence of more than two modalities. However, we do not

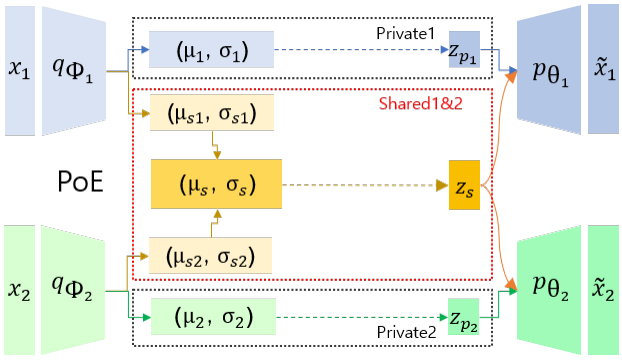


Figure 3: Model architecture of DMVAE. Each modality is used to infer the shared latent representation of that modality alone (subscripts $s1$ and $s2$) which are then aligned by the product-of-expert (subscript s). Private spaces (numeric-only/"p" subscripts) are left unaligned. The dashed lines indicate sampling from respective distributions.

4.2. Latent Space Inference

First, we define the latent space inference in our model. Given N modalities $\boldsymbol{x} = (x_1, \ldots, x_N)$, each modality has the posterior distribution $p(z_i|x_i)$, approximated by inference networks $q(z_i|x_i) = q(z_{p_i}, z_{s_i}|x_i)$. Since the shared latent space should reflect the information shared across all modalities, we require that the representation be consistent, i.e., $z_{s_i} = z_s$ w.p.1, $\forall i$. Consequently, we separate the *private inference* $q(z_{p_i}|x_i)$ from the *shared inference* network $q(z_s|\boldsymbol{x})$, defined using the product-of-experts (PoE) model [6], adopted in [20, 21]:

$$q(z_s|\boldsymbol{x}) \propto p(z_s)\prod_{i=1}^{N} q(z_s|x_i). \tag{2}$$

In the case where all inference networks and priors assume conditional Gaussian forms, $p(z) = \mathcal{N}(z|0, I)$ and $q(z|x_i) = \mathcal{N}(z|\mu_i, C_i)$ of i-th Gaussian expert with the covariance C_i, the PoE shared inference network will have the closed form of $q(z|\boldsymbol{x})$ as $\mathcal{N}(z|\mu, C)$ where $C^{-1} = \sum_i C_i^{-1}$ and $\mu = C\sum_i C_i^{-1}\mu_i$.

Missing-mode Inference. An important benefit of the PoE-induced shared inference is that the individual modality shared networks can also be used for inference in instances when one (or more) of the modalities is missing. Specifically, as illustrated in the bi-modal case of Fig. 2b, under the x_2 missing, the shared latent space would be simply inferred using the remaining modality shared inference network $q(z_s|x_1)$; and vice-versa for missing x_1.

consider this setting in our current work.

4.3. Reconstruction Inference

In addition to inferring the latent factors, a key enabler in VAE is the reconstruction inference, or encoding-decoding. Specifically, we seek to infer $p(\tilde{x}|x) = \int p(\tilde{x}, z|x)dz = \int p(\tilde{x}|z)p(z|x)dz = \mathbb{E}_{p(z|x)}[p(\tilde{x}|z)] \approx \mathbb{E}_{q(z|x)}[p(\tilde{x}|z)]$.

The reconstruction inference in multi-modal settings, much like the latent space inference, has to consider the cases of complete and missing modality data. We assume bi-modality without loss of generality. The first case is the self-reconstruction within a single modality, $\mathbb{E}_{q(z_{p_i}|x_i)q(z_s|x_i)}[p(\tilde{x}_i|z_{p_i}, z_s)]$ for $i = 1, 2$. The second form is the joint multi-modal reconstruction, $\mathbb{E}_{q(z_{p_i}|x_i)q(z_s|x_1,x_2)}[p(\tilde{x}_i|z_{p_i}, z_s)]$ for $i = 1, 2$. It is also possible to consider the cross-modal reconstruction, e.g., $p(\tilde{x}_2|x_1) = \mathbb{E}_{p(z_{p_2})q(z_s|x_1)}[p(\tilde{x}_2|z_{p_2}, z_s)]$, illustrated in Fig. 2b. This instance, where x_2 is missing, is facilitated using the prior on the private space of x_2, $p(z_{p_2})$.

The different reconstruction inference modes are essential for model learning but also valuable for understanding the model performance. For instance, one may seek to see how successful the multi-modal DMVAE is in learning the shared and private representations in the context of synthesizing one modality from another. We highlight these cross-synthesis experiments in Sec. 5.1 and Sec. 5.2.

4.4. Learning Objective

In general, for each data point $\boldsymbol{x} = (x_1, x_2, ..., x_N)$ and N modalities, the learning objective assumes the form:

$$\begin{aligned}\sum_i \mathbb{E}_{p(x_i)}\Big[&\lambda_i \mathbb{E}_{q_\phi(z_{p_i},|x_i),q_\phi(z_s,|\boldsymbol{x})}\left[\log p_\theta(x_i|z_{p_i}, z_s)\right] \\ &- KL(q_\phi(z_{p_i}|x_i)||p(z_{p_i})) - KL(q_\phi(z_s|\boldsymbol{x})||p(z_s)) \\ &+ \sum_j \Big(\lambda_i \mathbb{E}_{q_\phi(z_{p_i},|x_i),q_\phi(z_s,|x_j)}\left[\log p_\theta(x_i|z_{p_i}, z_s)\right] \\ &- KL(q_\phi(z_{p_i}|x_i)||p(z_{p_i})) - KL(q_\phi(z_s|x_j)||p(z_s))\Big)\Big],\end{aligned} \tag{3}$$

where λ_i balances the reconstruction across different modalities. The first term models the accuracy of reconstruction with the jointly learned shared latent factor, compensated by the KL-divergence from the prior. The second set of terms assesses the accuracy of the cross-modal reconstruction, $x_i \leftarrow x_j$ for i $\neq j$ and the accuracy of self-reconstruction for $i = j$, again compensated by the divergence.

5. Experiments

We demonstrate the effectiveness of our proposed DMVAE framework on two experimental setup. In Sec. 5.1, we

Table 1: Classification accuracy for cross-synthesized MNIST and SVHN and joint matching accuracy

Model	Cross(M$\rightarrow S$)	Cross(S$\rightarrow M$)	Joint
MVAE [21]	9.5	9.3	12.7
MMVAE [18]	86.4	69.1	42.1
DMVAE	**88.1**	**83.7**	**44.7**

show how DMVAE can learn the common latent representation given two image modalities of MNIST and street-view house number (SVHN) datasets. We evaluate our model both quantitatively and qualitatively by cross-synthesizing images from one to another modality. Sec. 5.2 further investigates DMVAE on the image and text modalities using the Oxford-102 Flowers dataset. We examine how well the flower image and its descriptions are retrieved between two modalities. Sec. 5.3 evaluates the effectiveness of different model components in an ablation study. The code for our DMVAE model and the experiments in this section is available at https://github.com/seqam-lab/DMVAE.

5.1. Image-Image Modality

As in MMVAE [18], we ground the bi-modal setup by giving one modality as MNIST images and another modality as SVHN images. By assuming that the pair of (MNIST, SVHN) images is constructed according to the digit identity $\{0, \ldots, 9\}$, we expect the shared information between the MNIST and SVHN modalities to be the digit identity and each private latent space includes styles of of the digits, such as width, tilt, background etc. We follow MMVAE to create the paired data [2].

For MNIST, we assume one dimensional private space while SVHN which has more diverse style requires four dimension for its private latent space. To model the ten class factors as the shared latent representation, we set 10 dimensional shared latent space. The details of the model (encoder/decoder) architectures and the optimization are described in the Supplement.

Quantitative Evaluation. In order to assess whether the desired shared latent representations are learned, we generate two kinds of images at test time. The first one is prior-synthesized images. After sampling the shared latent code from the prior distribution, we generate the MNIST and SVHN images based on the same shared space sample. Secondly, given an image of one modality, the shared latent code is extracted and transferred to another modality in order to synthesize an image. For both of the cases, we feed the private latent factors sampled from the prior distribution $\mathcal{N}(0, 1)$, which promotes the diversity of the generated image. We use the same protocol as in MMVAE to evaluate the cross-synthesized images. To predict the digit class, the

[2]The code for paired data generation is available at https://github.com/iffsid/mmvae.

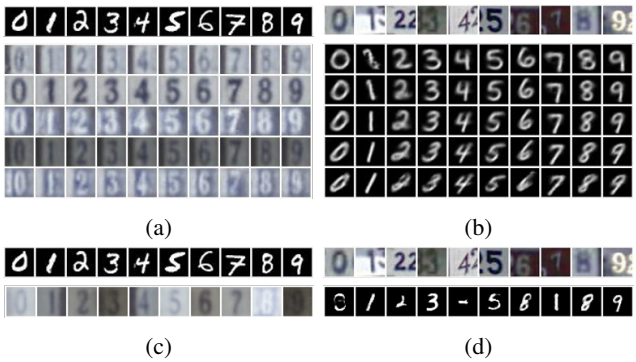
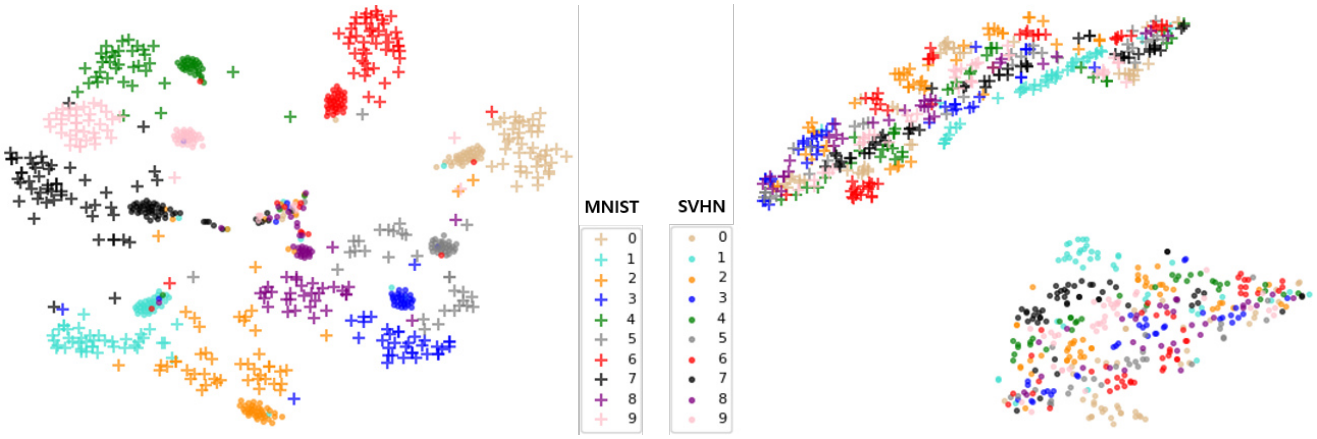


Figure 4: Cross-synthesis images from the opposite modality. (a) and (b) are results with DMVAE, (c) and (d) are results with MVAE. In each image, the first row is the ground truth images from which the share latent code comes. The following rows are the cross-synthesized images.

separate CNN classifiers are trained for MNIST and SVHN using the code from MMVAE for the fair comparison. For the prior-synthesized images, we compute the joint matching frequency which means how often the prior-synthesized images of MNIST and SVHN generated from the same prior distribution sample correspond to each other. For cross-synthesized images, the predicted labels based on the cross-synthesized images are compared to the ground-truth labels. Thus, the cross-modal inference in our generative model becomes the process of classification. We compare our results against MVAE [21] and MMVAE in Tab. 1. We outperform baselines in both direction of cross-synthesis, suggesting the desired common latent code, which is the digit class, is learned and transferred through the shared latent space. Moreover, when a random shared latent code is fed to each of MNIST and SVHN decoders, our model is able to generate the same class images of MNIST and SVHN with higher probability than baselines models. These results underline the ability of DMVAE to disentangle latent factors and distil them into the shared factor representing the digit label and the private factors surmising the image style. We investigate further why DMVAE can achieve significant improvement from SVHN to MNIST generation in Sec. 5.3 through the ablation study.

Qualitative Evaluation. Fig. 4 shows the cross-synthesized images conditioned on the opposite modality images. For one modality to be inferred by the generative model, both its private and shared latent factors are necessary. The shared factor is determined by the opposite conditioning modality. The private image factor is sampled from its prior distribution $\mathcal{N}(0, 1)$. Fig. 4 shows the results of our model and MMVAE. In each of Fig. 4a, Fig. 4b, Fig. 4c, and Fig. 4d, the first row is the ground-truth conditioning image to transfer its shared latent code to another modality and



(a) (b)

Figure 5: Visualization of 2-D embeddings of latent features of MNIST and SVHN, using tSNE. '+' and 'o' represent the MNIST and SVHN test data points respectively and each color associates with one of ten digit classes, $\{0, 1, ..., 9\}$. (a) DMVAE embedding result. (b) MMVAE embedding result.

the following rows are the cross-synthesized images. For DMVAE results Fig. 4a and Fig. 4b, cross-synthesized images require private latent code for the style. We generate 5 different rows of the synthesized images with 5 different private factors which are depicted from the second row. Though the private latent values can be sampled from the Gaussian prior distribution, we pick those latent values that can express "extreme" styles to assess whether the visually distinct style is kept. When SVHN image is synthesized from MNIST image as in Fig. 4b, styles of SVHN such as (dark / bright) (letter / background) color, or overall shadow, or width of digit are reflected in the cross-synthesized images as well as the conditioning MNIST classes are kept. In Fig. 4b, MNIST needs one dimensional private latent factor to be generated using the SVHN shared latent factor. We vary the private latent value from -1 to 3 to generate 5 different rows of the synthesized images. From top to bottom, the synthesized MNIST images show the varied width and slanted styles. Fig. 4c and Fig. 4d illustrate the results of MMVAE with the same ground truth images as in those of DMVAE. As well as the synthesized images cannot reflect the digit identity from the ground truth images clearly, there is no freedom of feeding diverse styles since MMVAE does not separate private latent space aside from the shared latent space.

In order to investigate how the shared latent space encodes each modality, we project the latent features inferred by the encoders, on the test set, into a 2-D space with tSNE. We use 400 randomly selected samples to plot the embedded features. In Fig. 5a, even though our model is trained without class label information, MNIST shared feature and SVHN shared features are gathered nearby according to the digit identity, which indicates DMVAE learns the digit identity only with the paired data. In contrast to DMVAE where MNIST and SVHN data points with the same class are heading for the same direction, MMVAE embeddings are clustered separately per dataset. This represents that DMVAE is able to amplify the role of the shared features by placing the styles aside into the private space, compared to MMVAE baseline.

5.2. Image-Text Modality

We further examine DMVAE on the Oxford-102 Flowers dataset [15], where each flower image is paired with ten captions that describe the visual characteristics of the flower. This dataset consists of 102 classes of 8,189 flower images, split into 62 training, 20 validation, and 20 test classes. Since the categories of the test set are disjoint from those of the training and validation sets, the problem falls within the scope of zero-shot test-time tasks. For the fair comparison with the prior works [1, 13, 22, 17], which utilize the class label of the Flowers dataset during training, we apply the following additional matching loss suggested in [22, 17].

$$L_M^I = \frac{1}{B}\sum_{m=1}^{B}\sum_{n=1}^{B} p_{m,n} \log \frac{p_{m,n}}{q_{m,n}+\epsilon}, \tag{4}$$

where $p_{m,n} = \frac{\exp(z_{SI_m}^\top \tilde{z}_{ST_n})}{\sum_{k=1}^{B}\exp(z_{SI_m}^\top \tilde{z}_{ST_k})}$ is the probability of matching the m-th image shared feature to the normalized n-th text shared feature for $m, n \in [1, B]$, with B the batch size; z_{SI_m} is the m-th image shared features, and $\tilde{z}_{ST_n}$ is the normalized n-th text shared feature, $\tilde{z}_{ST_n} = \frac{z_{ST_n}}{||z_{ST_n}||}$. $q_{m,n} = \frac{Y_{m,n}}{\sum_{k=1}^{B} Y_{m,k}}$ is the normalized true matching proba-

bility where $Y_{m,n} = 1$ if the pair is matched and 0 otherwise. L_M^T, the matching loss from text shared feature to the normalized image shared feature is computed in a similar manner. These losses take the advantage of the class label information to construct matched and unmatched pairs within a batch, in order to minimize the compatibility with unmatched pairs.

Given a $224 \times 224 \times 3$ dimensional input image, we first apply pre-trained ResNet-101 to generate $7 \times 7 \times 2048$ dimensional features. After global-average pooling layer, a FC layer is used to extract 64d shared latent feature and 3d private latent feature. For the simplicity of the network, our generative model decodes the 2048d feature, the reconstruction of the feature produced from the global-average pooling layer in the encoder, instead of the ambient image.

For the caption, we first extract the sequence of the word embedding using the BERT [3] tokenizer and the BERT base model, pre-trained on the uncased book corpus and English Wikipedia datasets, where the maximum length of the sequence is 30. A sequence whose length is less than 30 is padded by zeros. Given the 768d caption embedding in the BERT base model, we construct our text inference network using a bidirectional LSTM [7] of hidden dimension 512, followed by a max pooling layer and a FC layer to create the final 64d shared feature and the 3d private feature. For the same reasons as the image modality, the text modality decoder produces 1024d features corresponding to the output of max pooling in the text encoder. We use Adam optimizer [9] with batch size 64.

Quantitative Evaluation. As suggested in [1, 13, 22, 17], we evaluate the compatibility of image and text modalities in terms of recognition and retrieval on the shared latent space. The shared features of the text modality are averaged per class for the evaluation in both directions. For image-to-text cross generation, recognition is assessed with the Rank1 score and for text-to-image cross generation, retrieval is measured with AP@50. To compute Rank1, after ranking the cosine similarity between a given query image feature and all per-class-averaged text features, we assert whether the closest text feature shares same label with the query image. To compute AP@50, images are first ranked according to their cosine similarity with a given query text feature, averaged per class, assessing the fraction of the closest 50 images with the same class label as the query text, finally averaged over all classes. Tab. 2 shows the recognition and retrieval evaluation results. DMVAE outperforms competition on the image-to-text cross-recognition while achieving identical performance on text-to-image retrieval task.

Qualitative Evaluation. In Fig. 6, each row depicts the top3 retrieved captions given a query image according to the cosine similarity in the shared latent space. All the retrieved captions have the same class label as the query image except for the red colored caption in the last row, where the class is that of the image in the second row. In spite of the incorrectly retrieved caption, we can observe that the description

Table 2: Recognition and retrieval results on the Oxford-102 Flowers dataset.

Model	Img2Txt (Rank1)	Txt2Img (AP@50)
Word2Vec [14]	54.2	52.1
GMM+HGLMM [11]	54.8	52.8
Word CNN [1]	60.7	56.3
Word CNN-RNN [1]	65.6	59.6
Triplet [13]	64.3	64.9
IATV [13]	68.9	69.7
CMPM+CMPC [22]	68.4	70.1
TIMAM [17]	70.6	**73.7**
DMVAE	**73.3**	**73.6**

Query	Rank1	Rank2	Rank3
	this flower has smooth white petals, two which are rounded and three which are oblong	a flower with five white petals and a yellow pistil	this flower has five very smooth white petals with rounded edges
	this flower has orange upright petals that have pointed tips	this flower has a lightly multicolored pedicel that holds the upright sharply pointed orange petals	this flower has knife like orange petals that stick up vertically
	the petals on this flower are long and droopy with an orange color to them	the petals are curled, orange, and covered with dark red spots	a bird shaped flower with shiny orange petals that sprout out of it's pedicel

Figure 6: Given a query image, captions are retrieved. The red colored caption represents the incorrect retrieval.

Query	Rank1	Rank2	Rank3
this unique flower has long thin pink petals with a big fussy stigma			
this flower has five elongated triangle shaped purple petals surrounding yellow stamen			
this flower has white petals with purple stripes and long pink stamen			

Figure 7: Given a query caption, images are retrieved. The red bounding box on the image represents the incorrect retrieval.

Original query	this unique flower has long thin pink petals with a big fussy stigma.		
Synthesized query	Rank1	Rank2	Rank3
this unique flower has long thin yellow petals with a big fussy stigma			
this unique flower has short pink petals with a long white stamen			
this unique flower has round blue petals with red spots			

Figure 8: Given an original caption, captions are synthesized by swapping some words with the blue colored words that can represent different visual characteristics. Images are retrieved for each synthesized query. The retrieved images for the original query can be found in Fig. 7.

is closely related to the query image. Fig. 7 illustrates the reverse retrieval. Given a query caption, top3 images are retrieved. We provide the incorrect retrieval case at the rank3 of the first row, indicated by the red bounding box. It has the fussy stigma and the color similar with flowers of the correct class (Rank1 and Rank2 images), however it shows a different petal shape and a larger center.

We further examine what images the synthesized text retrieves in Fig. 8. From the original text query at the first row of the Fig. 7, we synthesize a new caption by swapping some words that can represent different visual characteristics. We mark those words with blue color in Fig. 8. In the first row, we observe that the retrieved images are yellow colored and with the characteristic of a fussy stigma. In the third row, no image exists in the test set that the synthesized caption describes. Thus, the similar images with round shaped and blue colored but no red spots are retrieved. This suggests that DMVAE learns a sufficiently meaningful shared latent space that allows retrieving between the image and text modalities.

5.3. Ablation Study

We examine the effectiveness of each component of DMVAE on MNIST and SVHN paired datasets in Tab. 3. pM and pS represent the private space of MNIST and that of SVHN respectively. Thus, pM and pS columns indicate the absence (x) or presence (o) of the private space for MNIST and SVHN, respectively. crVAE indicates the cross-VAE loss where M2S means the cross-reconstruction is conducted by transferring MNIST shared latent code to SVHN shared latent code; and vice-versa for crVAE S2M. For example, in the third row, the **"x"** of crVAE (M2S) implies no cross VAE loss in Eq. 3 for transfer from MNIST to SVHN.

In the second row of Tab. 3, without PoE, the alignment between two modalities becomes weaker, leading to lower performance in both cross-generation tasks. In terms of cross-VAE from the 3rd to the 5th rows, we can observe that the role of S2M cross-VAE is critical in order to achieve meaningful accuracy on SVHN to MNIST cross generation. SVHN alone is not able to learn the latent representation sufficient enough to classify each digit identity because the SVHN images are challenging for digit identity analysis. While trying to learn to generate MNIST, a simpler domain to recognize the digit classes, the SVHN shared latent space can obtain the knowledge about the digit identity. On the other hand, the performance on MNIST-to-SVHN cross-generation is improved significantly by the presence of the private space. This is because the variety of the SVHN styles cannot be expressed within the shared latent code from MNIST. These results provide the evidence of the ability of each component in DMVAE to disentangle latent factors into the shared factor and the private factor components.

Table 3: Ablation study on MNIST and SVHN modalities to analyze each component of DMVAE. pM and pS represent private space of MNIST and private space of SVHN respectively. crVAE indicates the cross-VAE loss.

Components					Accuracy	
pM	pS	crVAE (M2S)	crVAE (S2M)	PoE	M $\rightarrow S$	S $\rightarrow M$
o	o	o	o	o	88.13	83.73
o	o	o	o	x	87.33	77.33
o	o	x	o	o	87.85	82.83
o	o	o	x	o	82.7	17.03
o	o	x	x	o	83.19	12.75
x	x	o	o	o	12.38	82.72

6. Conclusion

In this paper, we introduce a novel multi-modal VAE model with separated private and shared spaces. We verify that having a private space per modality as well as the common shared space can significantly impact the representational performance of multimodal VAE models. We also demonstrate that VAE with the cross-reconstruction is important for separation of factors across the two sets of spaces. Application to image-to-image and image-to-text modeling tasks demonstrates the universal properties and effectiveness of DMVAE across different data types.

Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant No. 1955404.

References

[1] Learning deep representations of fine-grained visual descriptions, booktitle = IEEE Computer Vision and Pattern Recognition, year = 2016, author = Scott Reed and Zeynep Akata and Bernt Schiele and Honglak Lee,. 6, 7

[2] Konstantinos Bousmalis, George Trigeorgis, N. Silberman, Dilip Krishnan, and D. Erhan. Domain separation networks. In *NIPS*, 2016. 2, 3

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 7

[4] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1294–1305, Red Hook, NY, USA, 2018. Curran Associates Inc. 2

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 2

[6] G. E. Hinton. Products of experts. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, volume 1, pages 1–6 vol.1, Sep. 1999. 1, 3, 4

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. 9(8):1735–1780, Nov. 1997. 7

[8] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 7

[10] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1, 3

[11] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. pages 4437–4446, 06 2015. 7

[12] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018. 2

[13] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. pages 1908–1917, 10 2017. 6, 7

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 10 2013. 7

[15] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 6

[16] Daniel Quang, Yifei Chen, and Xiaohui Xie. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763, 10 2014. 2

[17] N. Sarafianos, X. Xu, and I. Kakadiaris. Adversarial representation learning for text-to-image matching. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5813–5823, 2019. 6, 7

[18] Yuge Shi, N. Siddharth, Brooks Paige, and P. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *ArXiv*, abs/1911.03393, 2019. 3, 5

[19] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016. 3

[20] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 3, 4

[21] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5575–5585. Curran Associates, Inc., 2018. 3, 4, 5

[22] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, 2018. 6, 7

[23] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. NIPS'17, page 465–476, Red Hook, NY, USA, 2017. Curran Associates Inc. 2