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

This supplement consists of the following materials:

- Details of network architectures in Sec. 1.
- Optimization details in Sec. 2.
- Reconstruction inference in Sec. 3.
- Additional experimental results in Sec. 4.

1. Neural Network Architecture

We describe our model architecture in Tab. 1a, Tab. 1b, Tab. 1c, and Tab. 1d, for MNIST, SVHN, the Oxford-102 Flowers image, and caption respectively. $Z_{p,MNIST}$, $Z_{p,SVHN}$, $Z_{p,I}$, and $Z_{p,C}$ indicate the latent dimension of the private space of MNIST, SVHN, the Oxford-102 Flowers Image, and the Oxford-102 Flowers Caption respectively while Z_s is the latent dimension of the shared space.

2. Training Details

We use Adam optimizer for all datasets. For MNIST and SVHN modalities, we use batch size 100. 10 epochs are trained with learning rate $1e^{-3}$. The dimension of private latent space is 1 for MNIST and 4 for SVHN while the shared latent space has 10 dimension. λ_{MNIST} and λ_{SVHN} are set as 50 and 1 respectively to balance the contribution of the modalities to the PoE [2].

For the Oxford-102 Flowers [3], batch size 64 is used. After 10 epochs are trained with learning rate $2e^{-4}$, we fine-tune the model with learning rate $2e^{-5}$ for 20 more epochs. The dimension of private latent space is 3 for both image and caption while the shared latent space has 64 dimension. Both λ_{image} and $\lambda_{caption}$ are set as 1. In the caption modality, we use the BERT [1] tokenizer and the BERT base model, pre-trained on the uncased book corpus and English Wikipedia datasets in order to extract the sequence of the word embedding of 768d, which is fed to our caption network as an input. The maximum length of the sequence is 30 and sequence whose length is less than 30 is padded by zeros. In the image modality, we resize images into 224×224 and adopt image augmentation of horizontal flipping with probability 0.5.

3. Reconstruction Inference

In the main paper, Figure 4, we cross-synthesize MNIST and SVHN images from the opposite modality assuming the reconstruction inference setting (Sec. 4.2, Main paper), where the missing input image modality was replaced by sampling from $z_{p,MNIST} \sim \mathcal{N}(0,1)$ or $z_{p,SVHN} \sim \mathcal{N}(0,1)$. Here, we consider a more complete set of reconstruction experiments. Specifically, we consider the six reconstruction instances illustrated in Fig. 1.

The first instance, Fig. 1a corresponds to traditional "style transfer" experiments, where the "style" (private space) of x_1 is used to map onto the "content" determined by x_2. This can mean that x_1 could be an MNIST image of digit '1', while x_2 is the SVHN image of digit '2'. The task would be to create a synthetic MNIST image of digit '2' in the style of digit '1'. The second instance, in Fig. 1b, is that where there is no conditioning modality x_1, hence the "style" of x_1 is sampled from the prior. This corresponds to the task of synthesizing any MNIST image of SVHN class. Fig. 1c and Fig. 1d make the strong inference of z_s with both modalities using PoE [2]. The "style" is reflected from x_1 in Fig. 1c while Fig. 1d randomly samples the private latent factors from the prior. Fig. 1e corresponds to the instance of traditional reconstruction of a data point, where both style and digit information comes from x_1. The style of this reconstruction also can be varied by private latent codes from the prior distribution as in Fig. 1f.

4. Additional Experimental Results

In this section we present additional experimental results on MNIST, SVHN and the Flower dataset. Sec. 4.1 focuses on studying the two image modalities of MNIST and SVHN. Sec. 4.2 studies the image and text modalities with the Oxford-102 Flowers dataset.

Table 1: Neural Network of each dataset modality. $Z_{p,MNIST}$, $Z_{p,SVHN}$, $Z_{p,I}$, and $Z_{p,C}$ indicate the latent dimension of the private space of MNIST, SVHN, the Oxford-102 Flowers Image, and the Oxford-102 Flowers Caption respectively while Z_s is the latent dimension of the shared space.

Encoder	Decoder
Input: Image (1×28×28)	Input: Latents ($Z_{p,MNIST} + Z_s$)
Linear 784 × 256	Linear ($Z_{p,MNIST} + Z_s$) × 256
ReLU	ReLU
Linear 256 × ($Z_{p,MNIST} + Z_s$)	Linear 256 × 784
	Sigmoid

(a) MNIST network

Encoder	Decoder
Input: Image (3×32×32)	Input: Latents ($Z_{p,SVHN} + Z_s$)
32 Conv 4 × 4, stride 2, pad 1	Linear ($Z_{p,SVHN} + Z_s$) × (256×2×2)
ReLU	ReLU
64 Conv 4 × 4, stride 2, pad 1	256 Conv 4 × 4, stride 2, pad 1
ReLU	ReLU
128 Conv 4 × 4, stride 2, pad 1	128 Conv 4 × 4, stride 2, pad 1
ReLU	ReLU
256 Conv 4 × 4, stride 2, pad 1	64 Conv 4 × 4, stride 2, pad 1
ReLU	ReLU
Linear (256×2×2) × 512	32 Conv 4 × 4, stride 2, pad 1
ReLU	Sigmoid
Dropout 0.1	
Linear 512 × ($Z_{p,SVHN} + Z_s$)	

(b) SVHN network

Encoder	Decoder
Input: Image (3×224×224)	Input: Latents ($Z_{p,I} + Z_s$)
ResNet-101	Linear ($Z_{p,I} + Z_s$) × 2048
AvgPool2d 7 × 7	Sigmoid
Linear 2048 × ($Z_{p,I} + Z_s$)	

(c) The Oxford-102 Flowers Image network

Encoder	Decoder
Input: Caption embedding (768) from the BERT base model	Input: Latents ($Z_{p,C} + Z_s$)
Bi-LSTM w. one hidden layer of 512d	Linear ($Z_{p,C} + Z_s$) × 1024
Max pooling	
Linear 1024 × ($Z_{p,C} + Z_s$)	

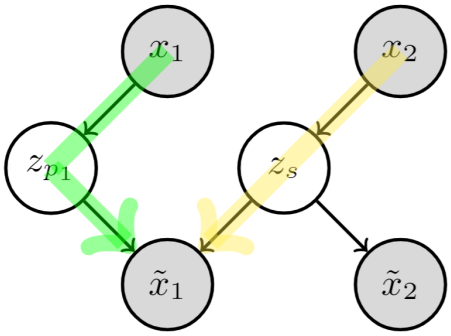
(d) The Oxford-102 Flowers Caption network

4.1. Image-Image modality

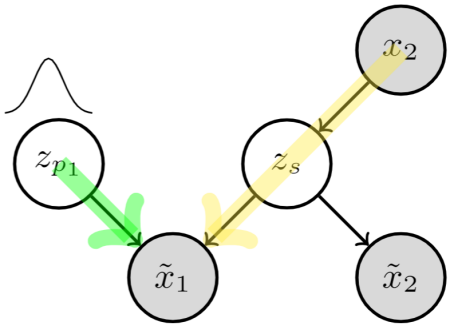
In Sec. 4.1.1, we conduct further qualitative evaluation on MNIST, SVHN based on the discussed reconditions in Sec. 3. We also provide a quantitative view of these synthesis results in Sec. 4.1.2. Finally, Sec. 4.1.3 demonstrates the separation of private and shared latent space by investigating the private latent space according to the digit identity and input image in 2-D space.

4.1.1 Qualitative Evaluation

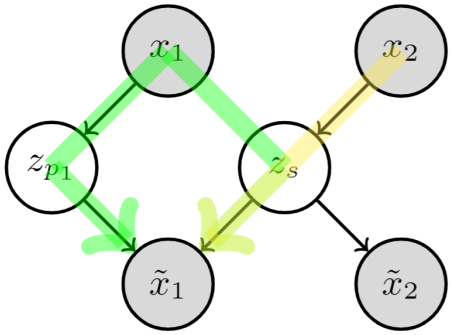
In Fig. 2 we depict the generated images of MNIST and SVHN under various reconstruction scenarios, defined in Fig. 1, beyond those in the main paper Figure 5. Fig. 2a is the ground truth images of MNIST and SVHN which are



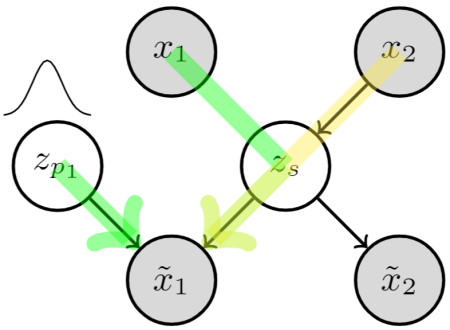
(a) Cross-reconstruction with conditioning



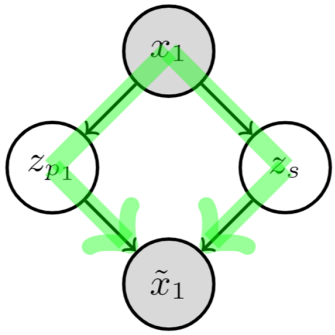
(b) Cross-reconstruction with missing modality



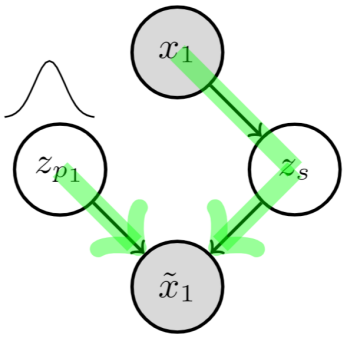
(c) PoE-reconstruction with conditioning



(d) PoE-reconstruction with prior distribution



(e) Reconstruction



(f) Reconstruction with prior distribution

Figure 1: ix instances of reconstruction of an image of MNIST or SVHN using the DMVAE model. **Cross-reconstruction with conditioning (a)**: both modalities x_1, x_2 are given; however, z_s is assumed to be inferred only from x_2, unlike the complete model. This corresponds to the case of "style transfer", where z_{p_1} (the style of x_1) is "injected" into the content z_s of x_2. **Cross-reconstruction with missing modality (b)**: depicts the reconstruction where the "style" of x_1 is not know, hence, z_{p_i} is sampled from its prior $p(z_p)$, the standard normal distribution. **PoE-reconstruction with conditioning (c)**: both modalities x_1, x_2 are given; while z_{p_1} is extracted only from x_1, z_s is assumed to be inferred from both x_1 and x_2 using PoE, which can enhance the shared latent code from both modalities. **PoE-reconstruction with prior distribution (d)**: both modalities x_1, x_2 are given; z_{p_1} is sampled from the prior distribution to have a random style and z_s is obtained from both x_1 and x_2. **Reconstruction (e)**: in this instance, we conduct traditional x_1 to $\tilde{x}_1$ reconstruction within a single modality. **Reconstruction with prior distribution (f)**: it gives variation on style of the traditional reconstruction by injecting the private factor from the prior distribution.

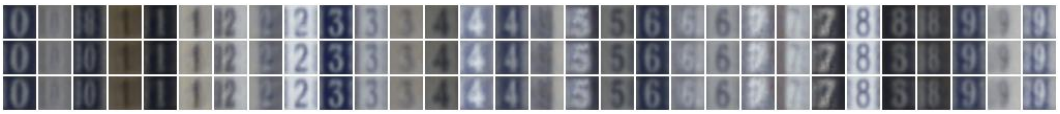
used as conditioning images for the shared or private factors in the following figures. Fig. 2b, Fig. 2c, Fig. 2d, and Fig. 2e show the generated images of MNIST and SVHN where each of three rows is reconstructed with the shared latent code coming from self-modality, PoE (both of the modalities), and cross-modality in order. In Fig. 2b and Fig. 2c,

(a) Ground truth images

(b) Generated MNIST where z_p inferred from (a)

(c) Generated SVHN where z_p inferred from (a)

(d) Generated MNIST where z_p sampled from $\mathcal{N}(0, I)$



(e) Generated SVHN where z_p sampled from $\mathcal{N}(0, I)$

Figure 2: (a) Ground truth images of MNIST and SVHN. (b) $\sim$ (e) reconstructed images of MNIST and SVHN. In each of (b) $\sim$ (e), the shared feature comes from self-modality, PoE (both of the modalities), and cross-modality in order. In (b) and (c), each row corresponds to Fig. 1e, Fig. 1c, and Fig. 1a in order which means the private latent factors comes from the given GT images in (a). In (d) and (e), each row corresponds to Fig. 1f, Fig. 1d, and Fig. 1b in order which means the private latent factors comes from the prior distribution $\mathcal{N}(0, I)$.

each row corresponds to Fig. 1e, Fig. 1c, and Fig. 1a in order, where the private latent code of each column is inferred from the given GT images in Fig. 2c in order. We note that the specific style from the conditioning image is well reflected in the generated image. In Fig. 2d and Fig. 2e, each row corresponds to Fig. 1f, Fig. 1d, and Fig. 1b in order, where the private latent factors are sampled from the prior distribution $\mathcal{N}(0, I)$. Different columns use different private factors from the prior distribution, however within one column, the same private factor is used for all rows. Even though the style is not specific, random sampling from the prior distribution enables synthesis of realistic reconstructions with identifiable digits.

4.1.2 Quantitative Evaluation of Synthesized Images

In Tab. 2, we assess the quality of reconstructions in Fig. 2. The quality of reconstruction is evaluated by computing the prediction accuracy of the digit identity using a separately trained CNN classifier, as described in the main paper. Tab. 2a shows the accuracy of reconstructions in Fig. 2b and Fig. 2c where private latent codes are conditioned on GT images in Fig. 2a. Tab. 2b accuracy corresponds to Fig. 2d

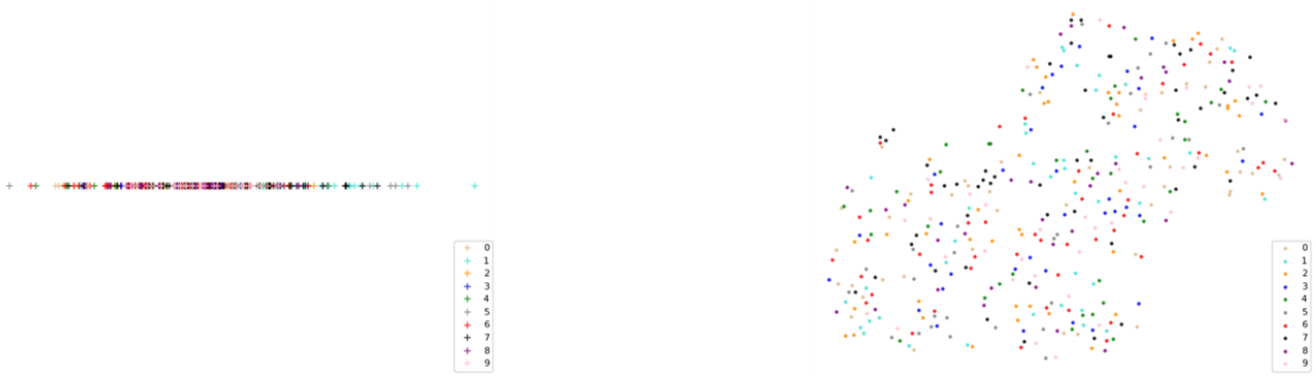
Table 2: Classification accuracy of the generated output conditioning on different shared latent code; self-modality, both modalities (PoE), or cross-modality. (a): accuracy corresponds to Fig. 2b and Fig. 2c where private latent codes are conditioned on GT images. (b): accuracy corresponds to Fig. 2d and Fig. 2e where private latent codes are sampled from the prior distribution $p(z) = \mathcal{N}(z|0, I)$.

Shared space conditioning	Generated output	
	MNIST	SVHN
Self-modality	95.37	81.49
Both modalities (PoE)	95.83	90.54
Cross-modality	84.16	90.04

(a) z_p inferred from GT image

Shared space conditioning	Generated output	
	MNIST	SVHN
Self-modality	91.87	79.12
Both modalities (PoE)	92.53	88.63
Cross-modality	83.73	88.13

(b) z_p sampled from $\mathcal{N}(0, I)$



(a) MNIST (b) SVHN

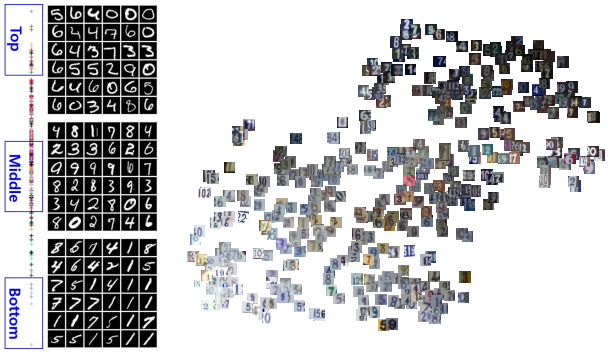
Figure 3: Visualization of the private latent features of (a) MNIST (b) SVHN using tSNE in 2-D space. In contrast to the Figure 5 in the main paper which is the tSNE embedding of the shared latent features, the private latent space does not learn any patterns according to the digit identity.

and Fig. 2e which use the private factor sampled from the prior distribution. As expected, the PoE can achieve the best accuracy in both Tab. 2a and Tab. 2b since it takes advantage of shared information from both modalities. One notable thing is that generated SVHN can be classified better when its shared space is from cross-modality (MNIST) than from its own modality (SVHN). This is because MNIST is simpler than SVHN to identify each digit, which makes it easier for the MNIST shared latent space to learn the digit identity. This can be examined in the shared space embedding plot in Figure 5 of the main paper. In the center of the plot, SVHN points are mixed while MNIST points are well-separated according to the digit identity. Therefore, cross-synthesized SVHN images with MNIST shared features can yield higher accuracy on digit id classification.

4.1.3 Separation of Private and Shared Latent Spaces

In addition to the tSNE embedding of the shared latent space in the Figure 5 of the main paper, we further investigate the separation of information across private and shared latent spaces by looking into the private latent space in 2-D space in Fig. 3 using the same 400 randomly selected samples as in the Figure 5 of the main paper. The 5 dimensional private latent space of SVHN is projected into a 2-D space with tSNE while the one dimensional private latent space of MNIST is plotted as it is. In both private latent spaces, any patterns cannot be found according to the digit identity, which indicates the private latent factors would be not involved in determining digit identity.

Furthermore, we check whether the style is learned in the private space. Since there are no labels for style, we illustrate the same private space embedding of style factors using the ground truth images, as in Fig. 4. For instance, Fig. 4a shows that one extreme of the private space (top) is aligned with wide digits, while the bottom characterizes



(a) MNIST (b) SVHN

Figure 4: Style analysis of the private space of MNIST / SVHN. Best interpreted magnified, together with Fig. 3.

slanted digits. Fig. 4b shows this for SVHN images. We can see that the color of digits and the background appears as the style factor associated with different potions of the private space. Bottom left are lighter digits, while the top-right are the darker ones. This affirms the ability of our model to isolate style factors into the private space, while coalescing the content (ID) into the shared space, as shown in Fig. 5 in the main paper.

4.2. Image-Text Modality

In this section, we provide additional qualitative results on the Oxford-102 Flowers dataset. Fig. 5 shows the top 5 retrieved captions when a flower image is given. In contrast, Fig. 6 illustrates the top 5 retrieved images given a text caption. Red colored text or red bounding box on a image indicates that the retrieved result has a different class label with the query. We note that even those incorrectly retrieved results are closely related to the given query, which causes the confusion.

Query	Rank1	Rank2	Rank3	Rank4	Rank5
	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 rounded white petals which are wide and very smooth	this orchid has five white petals, a yellow stigma, and two yellow stamen
	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	this flower has long orange petals with pointy tips and a long green and orange sepal	the petals are sharply pointed and orange, and there is one large, pointed sepal
	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	this flower has long orange petals with red dots, curling backwards	the flower is upside down that has orange petals with violet spots and also it has a lengthy stamen
	this flower has wide pale pink petals with rounded edges and thick veins	this pink flower has veined and rounded petals and pink stamen	this flower has overlapping pink petals with veins and rounded edges	this pink flower has rounded and veined petals and bright pink and white stamen	olive green, yellow and white pistil surrounded by large round lilac petals with light pink veins showing through
	this prickly flower has a plethora of thorns and a bright purple set of thin petals on top	this flower has a large, spiky pedicel and purple, spiky petals	this flower has long, purple, fiber like petals on a spiky sepal	this purple flower has long thin hair like petals and sits a top a green spiky bulb	this flower has a lot of very skinny and spiky purple petals
	this flower has a star-like configuration of bright blue petals with prominent veins	this flower has a star-like shape with bright blue petals that are veined	this flower is star shaped with five pointed blue petals which are veined	this flower has five pointed purple petals which are veined and star-shaped	this flower has light blue petals with dark blue veining arranged in a star-like shape
	this flower has closed yellow petals on a green pedicel	this flower has smooth yellow semi closed petals attached to a green sepal	a yellow flower with curved smooth petals and a green pedicel	this flower is yellow in color, with petals that are curled onward	this flower is yellow in color, with petals that are curled over on the center
	"the flower has a large white petal shaped like a bowl	this flower has petals that are white and closed together	a flower that has one petal that is white and wraps around	this white flower has one bowl shaped petal surrounding a cone shaped cluster of stamen	a white flower that wraps around the pedicel who's color bleeds into the petal that has a cone shape
	this strange looking flower has long thing red petals around its large center	this flower a lot of long petals and a large stamen	this flower has a large center of lightly colored stamen with sharply pointed petals	this wide blossom has many very thin pink stamen surrounded by splayed russet-red petals which are long and slightly tapered	the flower has a large brown center with long skinny petals that are pink in color

Figure 5: Given a query image, captions are retrieved. The red colored caption represents the incorrect retrieval.

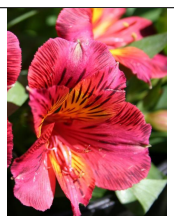
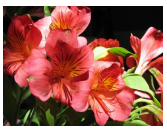
Query	Rank1	Rank2	Rank3	Rank4	Rank5
this unique flower has long thin pink petals with a big fussy stigma					
this flower has five elongated triangle shaped purple petals surrounding yellow stamen					
this flower has white petals with purple stripes and long pink stamen					
a pink flower made of thin, broad petals that sport dark pink veins and lightens in the center of the flower, where there are white pistils and yellow stamen					
this flower has a rounded ball of tiny pale purple blossoms					
the yellow-green petals are thin and form a flat circle around a cluster of yellow stamen					
the flower shown has green pedicel with a smooth white petal					
this flower has wide pink petals with accents of yellow and dark red					

Figure 6: Given a query caption, images are retrieved. The red bounding box on the image represents the incorrect retrieval.

References

[1] 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. 1

[2] 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] 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. 1