Variational Inference via a Joint Latent Variable Model with Common Information Extraction

J. Jon Ryu* **Young-Han Kim**
Dept. of ECE, UCSD
La Jolla, CA 92093, USA
{jongharyu,yhk}@ucsd.edu

Yoojin Choi **Mostafa El-Khamy** **Jungwon Lee**
SoC R&D, Samsung Semiconductor Inc.
San Diego, CA 92121, USA
{yoojin.c,mostafa.e,jungwon2.lee}@samsung.com

Abstract

Inspired by the distributed simulation problem in network information theory, this paper proposes a new joint latent variable model to learn a succinct common information and presents a simple training scheme for this model, where the succinctness is measured by Wyner's common information. The proposed model includes two groups of latent variables—first, the common latent variable that captures the common information (e.g., a shared concept) of the two data variables, and second, the local latent variables that capture the remaining randomness (e.g., texture and style) in respective data variables. As an example to illustrate the efficiency of the proposed model and accompanied training techniques, conditional generation (given an image of $N \in \{0, \dots, 9\}$, generate an image of $(N+1) \mod 10$) and style transfer (replicate a similar style of a given digit image in generating a new image) experiments using the MNIST dataset are exhibited.

1 Introduction

Suppose that given a sample from an underlying joint distribution $q(\mathbf{x}, \mathbf{y})$, we wish to simulate, or more precisely, to draw samples from $q(\mathbf{x}, \mathbf{y})$. Then, what is the minimum description rate we would need to make the simulation as exact as possible? In network information theory, A. Wyner formualted this problem as *distributed simulation* of correlated sources, and proved that the minimum description rate of $\mathbf{Z}$ is characterized by the so-called Wyner's *common information* (CI)

$$J(\mathbf{X}; \mathbf{Y}) := \min_{\mathbf{X}-\mathbf{Z}-\mathbf{Y}} I(\mathbf{X}, \mathbf{Y}; \mathbf{Z}), \tag{1}$$

where the minimum is over all conditional distributions $q(\mathbf{z}|\mathbf{x}, \mathbf{y})$ such that $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$ form a Markov chain [1, 2]. Note that the minimization over all possible joint encoders $q(\mathbf{z}|\mathbf{x}, \mathbf{y})$ in (1) is equivalent to the minimization over all possible *decoders* $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})$ that are consistent with $q(\mathbf{x}, \mathbf{y})$, i.e., $\int p(\mathbf{z})p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})\, \mathrm{d}\mathbf{z} = q(\mathbf{x}, \mathbf{y})$. Wyner's common information also appears in distributed channel synthesis [3] and the recent study of finding universal features from weakly correlated Gaussian vectors [4].

Based on this observation, we define the random variable $\mathbf{Z}$ corresponding to the minimizer in (1) as the most succinct "common representation" of correlated sources $q(\mathbf{x}, \mathbf{y})$, and we will refer it as *Wyner's common latent variable*. In this view, the compactness of an arbitrary random variable $\mathbf{Z}$ in learning $q(\mathbf{x}, \mathbf{y})$ is naturally quantified by the mutual information $I(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$. Interestingly, the optimization problem in (1) can be relaxed and optimized efficiently using the standard variational learning techniques developed in Bayesian statistics and variational autoencoders as will be shown shortly. After learning the decoder model, we can also perform conditional inference from $\mathbf{X}$ to $\mathbf{Y}$ by learning the posterior distribution $p(\mathbf{z}|\mathbf{x})$ using the same variational technique. The conditional

*This work was performed in part during the author's internship at Samsung Semiconductor Inc.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

inference will then be performed following the Markov chain $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$. Once we establish the variational learning of Wyner's common latent variable, we propose a new latent variable model with auxiliary random variables that explicitly represent the randomness in the stochastic decoders.

2 Main Results

2.1 Variational Learning of Wyner's Common Latent Variable

Variational relaxation of Wyner's CI. Let us use the shorthand notation $\mathbf{W}$ for $(\mathbf{X}, \mathbf{Y})$. The following derivation in this section holds for any random vector $\mathbf{W}$, unless specified otherwise. With $\mathbf{W} = (\mathbf{X}, \mathbf{Y})$, the Markovity $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$ implies that $p(\mathbf{w}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})$.

To ease the optimization problem in (1), we first assume that the joint distribution $p_\theta(\mathbf{z})p_\theta(\mathbf{w}|\mathbf{z})$ belongs to a parametric family parameterized by θ. Then eq. (1) can be rewritten as

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & I(\mathbf{W}_\theta; \mathbf{Z}_\theta) \\ \text{subject to} \quad & p_\theta(\mathbf{w}) = q(\mathbf{w}), \end{aligned} \tag{2}$$

where $p_\theta(\mathbf{w}) := \int p_\theta(\mathbf{z})p_\theta(\mathbf{w}|\mathbf{z})\,\mathrm{d}\mathbf{z}$, and $(\mathbf{W}_\theta, \mathbf{Z}_\theta) \sim p_\theta(\mathbf{z})p_\theta(\mathbf{w}|\mathbf{z})$. We introduce two approximations to relax the optimization problem to a tractable form. First, we replace the equality constraint in (2) for consistency with an inequality constraint introducing a slackness in terms of a KL divergence, i.e., for some small $\Delta > 0$, $D(q(\mathbf{w}) \parallel p_\theta(\mathbf{w})) \leq \Delta$. Secondly, we introduce an approximate posterior distribution $q_\phi(\mathbf{z}|\mathbf{w})$ as a proxy to the intractable model posterior $p_\theta(\mathbf{z}|\mathbf{w})$[2], which leads us an upper bound on the KL divergence term as

$$\begin{aligned} D(q(\mathbf{w}) \parallel p_\theta(\mathbf{w})) &\leq D(q(\mathbf{w})q_\phi(\mathbf{z}|\mathbf{w}) \parallel p_\theta(\mathbf{w})p_\theta(\mathbf{z}|\mathbf{w})) \\ &= D(q(\mathbf{w})q_\phi(\mathbf{z}|\mathbf{w}) \parallel p_\theta(\mathbf{z})p_\theta(\mathbf{w}|\mathbf{z})). \end{aligned} \tag{3}$$

Also, we replace the mutual information term $I(\mathbf{W}_\theta; \mathbf{Z}_\theta) = D(p_\theta(\mathbf{z})p_\theta(\mathbf{w}|\mathbf{z}) \parallel p_\theta(\mathbf{w})p_\theta(\mathbf{z}))$ with $D(q(\mathbf{w})q_\phi(\mathbf{z}|\mathbf{w}) \parallel q(\mathbf{w})p_\theta(\mathbf{z}))$ to make it trainable. All the approximation steps become exact when $q(\mathbf{w})q_\phi(\mathbf{z}|\mathbf{w}) \equiv p_\theta(\mathbf{z})p_\theta(\mathbf{w}|\mathbf{z})$, i.e., the latent variable model is consistent with $q(\mathbf{w})$ and the approximate posterior $q_\phi(\mathbf{z}|\mathbf{w})$ matches to the true posterior $p_\theta(\mathbf{z}|\mathbf{w})$.

By introducing a Lagrange multiplier $\lambda > 0$, then the optimization problem can be converted into an unconstrained form

$$\underset{\theta,\phi}{\text{minimize}} \; D(q(\mathbf{w})q_\phi(\mathbf{z}|\mathbf{w}) \parallel p_\theta(\mathbf{z})p_\theta(\mathbf{w}|\mathbf{z})) + \lambda D(q(\mathbf{w})q_\phi(\mathbf{z}|\mathbf{w}) \parallel q(\mathbf{w})p_\theta(\mathbf{z})), \tag{4}$$

or equivalently,

$$\underset{\theta,\phi}{\text{minimize}} \; \mathsf{E}_{q(\mathbf{w})}\left[(1+\lambda)D(q_\phi(\mathbf{z}|\mathbf{W}) \parallel p_\theta(\mathbf{z})) + \int q_\phi(\mathbf{z}|\mathbf{W}) \log \frac{1}{p_\theta(\mathbf{W}|\mathbf{z})}\,\mathrm{d}\mathbf{z}\right]. \tag{5}$$

If $\lambda = 0$ in (5), it becomes equivalent to the expected value of the evidence lower bound (ELBO) in variational Bayesian methods (see, e.g., [5]), and thus it appears as the loss function for the variational autoencoder (VAE) [6, 7]. The first and the second terms are called the *regularization* loss and the *reconstruction* loss, respectively. With $\lambda = 0$ in (4), it can be viewed as *matching two joint distributions* by minimizing the KL divergence. We also remark that the weighted version of the objective with $\lambda > 0$ also appears in the recent β-VAE model [8].

Practical optimization procedure and deep neural networks. In practice, we have access to the underlying distribution $q(\mathbf{w})$ only in the form of the empirical distribution $q_{\mathrm{emp}}(\mathbf{w})$ given a sample $\{\mathbf{w}_i\}_{i=1}^N$, and thus the expectation over $q(\mathbf{w})$ is replaced with the summation over the sample.

With a certain parametric family, the regularization loss $D(q_\phi(\mathbf{z}|\mathbf{w}) \parallel p_\theta(\mathbf{z}))$ can be given as an analytic form of a function of the parameters of the distributions, which are then differentiable with respect to θ and ϕ. For example, the most common choice is to take $p_\theta(\mathbf{z}) = p(\mathbf{z})$ as a standard Gaussian, and both the encoders p_θ and the decoders q_ϕ as diagonal Gaussians. The gradient of the reconstruction loss involving the integral can be estimated efficiently by Monte Carlo approximation

[2] This is equivalent to the standard technique of introducing an approximate posterior in variational Bayesian methods [5] to detour the often intractable marginal distribution $p_\theta(\mathbf{w})$.

of the integral with the reparameterization trick [6]. Deep neural networks that provide rich parametric families for $q_\phi(\mathbf{z}|\mathbf{w})$ and $p_\theta(\mathbf{w}|\mathbf{z})$ can be simply plugged in, and then the model parameters can be trained efficiently with backpropagation. We note that, throughout the paper, every distribution is parameterized with a separate neural network although it is indexed by the same parameter θ or ϕ.

Back to Wyner's common latent variable. We note that in (5) the regularization loss corresponds to the mutual information $I(\mathbf{Z};\mathbf{W}) = I(\mathbf{Z};\mathbf{X},\mathbf{Y})$, while the reconstruction loss corresponds to the conditional entropy $h(\mathbf{W}|\mathbf{Z}) = h(\mathbf{X}|\mathbf{Z}) + h(\mathbf{Y}|\mathbf{Z})$, which quantifies the expected uncertainty in guessing $\mathbf{W} = (\mathbf{X},\mathbf{Y})$ given $\mathbf{Z}$. Therefore, variational optimization of Wyner's common latent variable tries to extract the most succinct common structure from the data pair $\mathbf{X}$ and $\mathbf{Y}$ by minimizing the mutual information, while it also tries to decrease uncertainty in guessing $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{Z}$ by minimizing the conditional entropy.

We remark that the joint objective in (5) with $\lambda = 0$ still minimizes $I(\mathbf{Z};\mathbf{X},\mathbf{Y})$ in disguise, and thus finds a succinct factor implicitly. As a simple showcase of the idea, we demonstrate how the succinctness of the common latent variable $\mathbf{Z}$ can be controlled by manipulating the size of $\mathbf{Z}$ when $\lambda = 0$ is taken; see Section 3.

2.2 The Proposed W-model

Building upon the variational learning of Wyner's common latent variable, we propose a new latent variable model with common information extraction by refining the decoders as follows. We introduce an auxiliary random variable $\mathbf{T}$, and reparameterize $p_\theta(\mathbf{w}|\mathbf{z})$ by $p_\theta(\mathbf{w})\delta(\mathbf{w} - \mathbf{w}_\theta(\mathbf{t},\mathbf{z}))$, where $\delta(\cdot)$ denotes the Dirac delta function, and $\mathbf{w}_\theta(\mathbf{t},\mathbf{z})$ is a deterministic mapping parameterized by θ. The new latent variable model then becomes $p_\theta(\mathbf{z})p_\theta(\mathbf{t})\delta(\mathbf{w} - \mathbf{w}_\theta(\mathbf{t},\mathbf{z}))$. This technique is similar to the functional representation lemma in network information theory (see, e.g., [2, Appendix B]) and the reparameterization trick proposed in [6].

The corresponding graphical model for $\mathbf{W} = (\mathbf{X},\mathbf{Y})$ is illustrated in Fig. 1(b). Here, we denote $\mathbf{U}$ and $\mathbf{V}$ in place of $\mathbf{T}$ for the implicit randomness in the stochastic decoders $p_\theta(\mathbf{x}|\mathbf{z})$ and $p_\theta(\mathbf{y}|\mathbf{z})$, respectively, and refer as the *local latent variable* for each $\mathbf{X}$ and $\mathbf{Y}$. In this case, the motivation for the reparameterization of the decoders is clear; each $\mathbf{U}$ and $\mathbf{V}$ corresponds to the local randomness used to generate $\mathbf{X}$ and $\mathbf{Y}$, respectively, in the distributed simulation. For simplicity, we further assume the additional conditional independence for the joint posterior distribution $q_\phi(\mathbf{u},\mathbf{v},\mathbf{z}|\mathbf{x},\mathbf{y}) = q_\phi(\mathbf{u}|\mathbf{x},\mathbf{y})q_\phi(\mathbf{v}|\mathbf{x},\mathbf{y})q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ in this paper. (See Figure 1(c).)

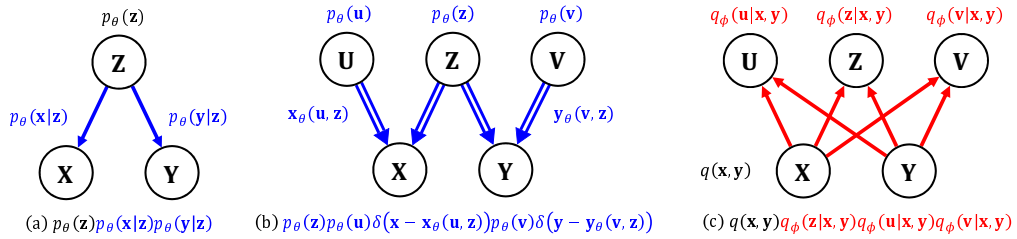


Figure 1: (a) The original joint latent variable model, (b) the new decoder model with refined decoders, and (c) the corresponding encoder model for variational learning. In the graphical models above, solid and double arrows denote stochastic and deterministic mapping, respectively.

The advantage in the refinement is clearly twofold. First, it improves the expressivity of the decoders by far than the parameterization by product distributions. Moreover, it leads us to learning the local latent variables $\mathbf{U}$ and $\mathbf{V}$ explicitly via the variational learning technique, so that we can use them in the subsequent inference tasks; see Section 2.3.

The new decoder model can be trained via the same procedure as previously derived in Section 2.1, with the proper joint encoder $q_\phi(\tilde{\mathbf{z}}|\mathbf{w})$, where $\tilde{\mathbf{Z}} := (\mathbf{Z},\mathbf{U},\mathbf{V})$. For the ill-defined terms in the objective function due to the Dirac delta function, we replace it with an isotropic Gaussian with small fixed variance, i.e., $\delta(\mathbf{w} - \mathbf{w}_\theta(\mathbf{t},\tilde{\mathbf{z}})) \approx \mathcal{N}(\mathbf{w}|\mathbf{w}_\theta(\mathbf{t},\tilde{\mathbf{z}}),\epsilon^2 I)$ for some small $\epsilon > 0$. Then from (5), the regularization loss becomes

$$(1+\lambda)D(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y}) \parallel p_\theta(\mathbf{z})) + D(q_\phi(\mathbf{u}|\mathbf{x},\mathbf{y}) \parallel p_\theta(\mathbf{u})) + D(q_\phi(\mathbf{v}|\mathbf{x},\mathbf{y}) \parallel p_\theta(\mathbf{v})),$$

and the reconstruction loss becomes

$$\frac{1}{2\epsilon^2}\int q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})\left(\int q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x})\|\mathbf{x}-\mathbf{x}_\theta(\mathbf{u},\mathbf{z})\|_2^2\,\mathrm{d}\mathbf{u}+\int q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})\|\mathbf{y}-\mathbf{y}_\theta(\mathbf{v},\mathbf{z})\|_2^2\,\mathrm{d}\mathbf{v}\right)\mathrm{d}\mathbf{z}.$$

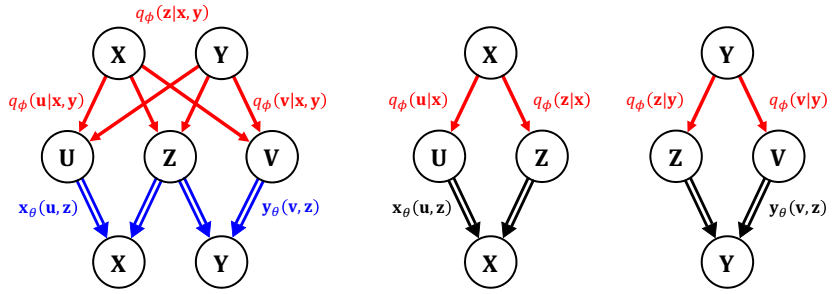
Here, the Lagrange multiplier λ for Wyner's CI is put only on $D(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})\parallel p_\theta(\mathbf{z}))$, which corresponds to the mutual information term $I(\mathbf{X},\mathbf{Y};\mathbf{Z})$.

For conditional inference via the common latent variable, we need to learn the approximate posteriors $q_\phi(\mathbf{z},\mathbf{u}|\mathbf{x})=q_\phi(\mathbf{z}|\mathbf{x})q_\phi(\mathbf{u}|\mathbf{x})$ and $q_\phi(\mathbf{z},\mathbf{v}|\mathbf{y})=q_\phi(\mathbf{z}|\mathbf{y})q_\phi(\mathbf{v}|\mathbf{y})$ as proxy to the model posteriors $p_\theta(\mathbf{z},\mathbf{u}|\mathbf{x})$ and $p_\theta(\mathbf{z},\mathbf{v}|\mathbf{y})$. All encoders are then to be trained with the distribution matching framework as described for the joint model, while the overall training scheme is readily discussed in the next paragraph. After Wyner, we name the entire probabilistic model including the new joint model (Fig. 1(b)) and all the approximate posteriors (Fig. 1(c)) as the W-model. Once the encoders and the decoders are parameterized by deep neural networks, we name the corresponding neural network architecture as the W-VAE.

Training. There are many possible training schemes for the W-model, but we present here a simple two-step algorithm to illustrate the idea. In the first step, we train all θ and ϕ in matching joint distributions q_ϕ and p_θ over $(\mathbf{X},\mathbf{Y},\mathbf{Z},\mathbf{U},\mathbf{V})$. (See Figure 2(a).) In the second step, we then train the marginal encoder $q_\phi(\mathbf{z}|\mathbf{x})$ by matching the joint distributions q_ϕ and p_θ over $(\mathbf{X},\mathbf{Z},\mathbf{U})$, i.e.,

$$\underset{\phi}{\text{minimize}}\, D(q_{\text{emp}}(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x})q_\phi(\mathbf{u}|\mathbf{x},\mathbf{z})\parallel p_\theta(\mathbf{z})p_\theta(\mathbf{u})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{u},\mathbf{z}))).$$

Here, the marginal encoder is to be fit only based on the decoders that is already trained in the first step.[3] The encoders $q_\phi(\mathbf{z}|\mathbf{y})q_\phi(\mathbf{v}|\mathbf{y})$ in the other direction of the Markov chain can be trained similarly by symmetry. (See Figure 2(b).)



(a) Training the joint model (b) Training the marginal models

Figure 2: The proposed two-step training algorithm.

Comparison to existing VAE architectures. The W-model improves upon the existing VAE architectures, namely, the conditional VAE (CVAE) [10, 11] and the joint VAE (JVAE) [12, 13, 9], by taking the best of the two architectures. First, the CVAE tends to overfit the realization of the conditioning variable, and suffer from poor generation performance for high-dimensional data. Since the W-model trains the joint distribution first, however, it does not suffer such problems. Second, by specifying and learning the local latent variables $\mathbf{U}$ and $\mathbf{V}$ explicitly, it can perform new inference tasks utilizing them as described below.

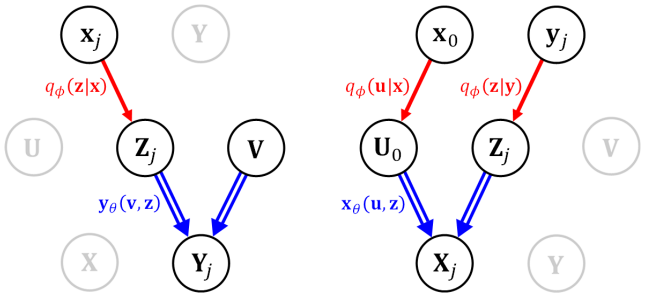
2.3 Statistical Inference via the W model

We describe how to perform joint generation, conditional generation, and style transfer via the proposed W-model.

- **Joint generation.** We sample $(\mathbf{U},\mathbf{V},\mathbf{Z})\sim p_\theta(\mathbf{u})p_\theta(\mathbf{v})p_\theta(\mathbf{z})$, and then take $\mathbf{X}=\mathbf{x}_\theta(\mathbf{U},\mathbf{Z})$ and $\mathbf{Y}=\mathbf{y}_\theta(\mathbf{V},\mathbf{Z})$.
- **Conditional generation.** Sampling $\mathbf{Y}$ given $\mathbf{X}=\mathbf{x}$ can be performed in three steps; see Figure 3(a). First sample $\mathbf{Z}\sim q_\phi(\mathbf{z}|\mathbf{x})$, then sample $\mathbf{U}\sim p_\theta(\mathbf{u})$, and finally take $\mathbf{Y}=\mathbf{y}_\theta(\mathbf{V},\mathbf{Z})$.

[3] We note that this training scheme is similar to one proposed in [9] as an alternative.

- **Style transfer.** Suppose that the common information $\mathbf{Z}$ is well-extracted, and the local randomness $\mathbf{U}$ and $\mathbf{V}$ capture the style and/or texture of each random variable. Then, we can utilize the local latent variables for inference tasks such as *style transfer*; see Figure 3(b). For example, let $(\mathbf{X}, \mathbf{Y})$ be a pair of correlated images generated from the common concept but with different styles. Suppose that a reference image $\mathbf{x}_0$ for style and a set of reference images $\mathbf{y}_1, \ldots, \mathbf{y}_M$ for concept are given. Given encoders $q_\phi(\mathbf{z}, \mathbf{u}|\mathbf{x})$ and $q_\phi(\mathbf{z}, \mathbf{v}|\mathbf{y})$ and a decoder $\mathbf{x}_\theta(\mathbf{z}, \mathbf{u})$, we generate samples $\mathbf{X}$ with the same style of the reference sample $\mathbf{x}_0$ by first sample $\mathbf{U}_0 \sim q_\phi(\mathbf{u}|\mathbf{x}_0)$ (guessing style of $\mathbf{x}_0$), next sample $\mathbf{Z}_j \sim q_\phi(\mathbf{z}|\mathbf{y}_j)$ (guessing concept from $\mathbf{y}_i$), and finally take $\mathbf{X}_j = \mathbf{x}_\theta(\mathbf{U}_0, \mathbf{Z}_j)$ (generating new image $\mathbf{X}_j$ by combining $\mathbf{Z}_j$ and $\mathbf{U}_0$).



(a) Conditional generation (b) Style transfer

Figure 3: Schematic illustrations for conditional generation and style transfer.

3 Experiment

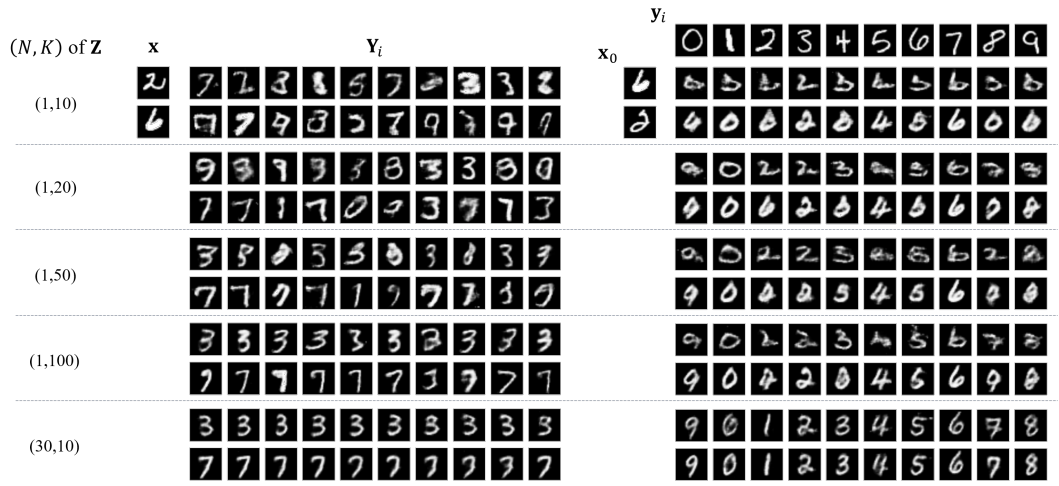
We generated a set of pairs of digit images $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^N$ from the MNIST dataset [14] such that $\mathrm{label}(\mathbf{Y}_i) = \mathrm{label}(\mathbf{X}_i) + 1 \mod 10$. There are 10 different classes in this dataset, where each class consists of 5,000 samples. Note that $\mathbf{X}$ and $\mathbf{Y}$ are correlated only through the label information.

In this experiment, we took $\lambda = 0$ and demonstrated the performance of conditional generation and style transfer. To control the succinctness of the common latent variable and the expressivity of the decoders by the sizes of the latent variables, we took all the latent vectors as categorical random vectors, and thus the training was done with the reparameterization with the Gumbel–softmax parameterization [15, 16]. For simplicity, we call the dimension N and cardinality K of a categorical random vector $\mathbf{Z}$ as the *size* of the random vector, and denote it as $|\mathbf{Z}|$.

All encoder and decoder networks have the same architecture of three fully connected layers with 512 hidden units each. For better visual quality, we trained a convolutional autoencoder with the original MNIST dataset, and used the trained autoencoder as a wrapper outside the stochastic encoders and decoders. For each case, the maximum number of epochs was taken to be 50, and the learning rate and the temperature for the Gumbel-softmax was annealed along the training epochs. For the objective function, the hyperparameter ϵ was taken such that $2\epsilon^2 = 10^{-4}$.

3.1 Varying $|\mathbf{Z}|$ with fixed $|\mathbf{U}|, |\mathbf{V}|$

We fixed the size of the local latent vectors $\mathbf{U}$ and $\mathbf{V}$ by $(N, K) = (20, 10)$, and increased the size of the common latent vector $\mathbf{Z}$ to show how the common information extraction and the corresponding generation performance are affected by the size of $\mathbf{Z}$ (See Figure 4). In conditional generation (Figure 4(a)), we wish to achieve both high accuracy of label and high variability in style, but there exists a tradeoff. When the size of $\mathbf{Z}$ was too small to contain the common information (e.g., $(N, K) = (1, 10)$), it resulted in a poor accuracy. On the other hand, when the size was too large (e.g., $(N, K) = (30, 10)$), $\mathbf{Z}$ tended to contain all the information of $\mathbf{X}$ and $\mathbf{Y}$, while $\mathbf{U}$ and $\mathbf{V}$ contained no information, which resulted in a poor variability. For style transfer (Figure 4), similar observations can be made. The consistency of the style along the row implies that $\mathbf{U}$ contained the style information, and generated images with correct labels along the column imply that $\mathbf{Z}$ contained the label information.



(a) Conditional generation (b) Style transfer

Figure 4: Sample results with varying dimension (N) and cardinality (K) of $\mathbf{Z}$ with fixed size of local latent variables $\mathbf{U}, \mathbf{V}$ with $(N, K) = (2, 10)$. In (a), the leftmost column shows the inputs $\mathbf{x}$, and the rest columns are the generated samples $\mathbf{Y}_i$. In (b), the leftmost column shows the reference images for style, and the topmost row shows the reference images for label. In the current example, $(N, K) = (1, 100)$ shows the most visually pleasing result in terms of accuracy and variability, for both tasks.

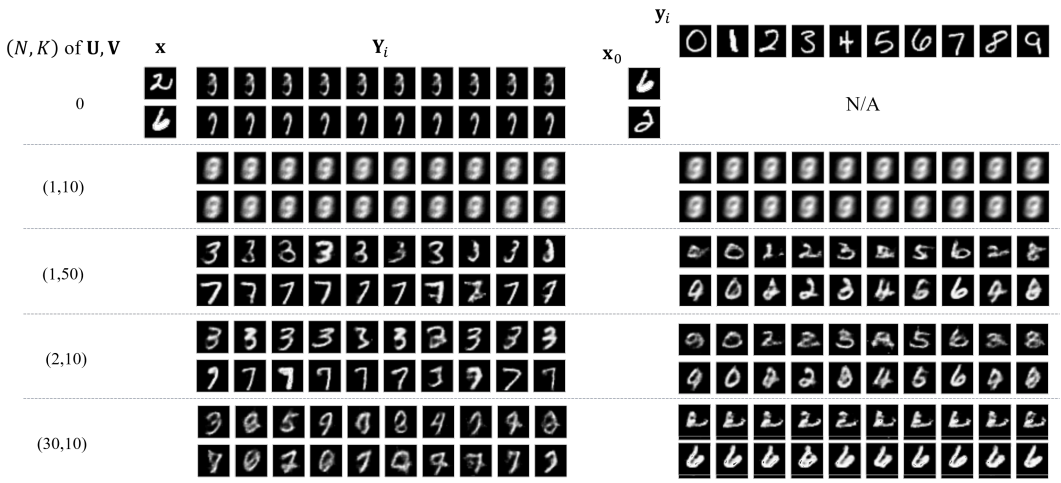
3.2 Varying $|\mathbf{U}|, |\mathbf{V}|$ with fixed $|\mathbf{Z}|$

We also empirically show that the expressivity of the decoders can be controlled through the dimensions of local latent vectors $\mathbf{U}$ and $\mathbf{V}$. As the baseline, we include a degenerate model without any local randomness, where the mean squared error (MSE) is used instead of the log-loss in the reconstruction. (Note that this case corresponds to $2\epsilon^2 = 1$ in our modification.) Since the diagonal Gaussian has a limited expressivity by assuming independent pixels, taking actual samples from the Gaussian distributions may perform worse. Hence, we took the output of the neural network (e.g., $\mathbf{x}_\theta(\mathbf{z})$) as a sample. Intuitively, increasing the number of dimensions of the local randomness gave better generation results. However, it can be also seen that too large local latent variables (e.g., $(N, K) = (30, 10)$) also inhibited the proper learning of the model, since then $\mathbf{U}$ and $\mathbf{V}$ may contain all the information about both $\mathbf{X}$ and $\mathbf{Y}$ while $\mathbf{Z}$ learns nothing.

4 Concluding Remarks

In this preliminary work, based on the notion of common information in network information theory, we developed a new latent variable model, described its variational learning method, and illustrated how to perform various inference tasks using the proposed model. The preliminary experimental results demonstrated the potential of the proposed model as a new way of learning the succinct common latent variable that can be further developed and refined for more complex dataset.

Here we remark several future directions to be explored. First, we plan to properly justify Wyner's common information as a measure of succinctness of a shared randomness of correlated sources. For example, is there any optimality of the Wyner's common latent variable in predicting $\mathbf{X}$ and/or $\mathbf{Y}$? Further, the effect of the Lagrange multiplier λ on the tradeoff between learning succinct common information and predicting $\mathbf{X}$ and $\mathbf{Y}$ will be carefully studied. Provided that these points can be addressed properly, we believe that the variational learning of Wyner's common information can be a new information theoretic principle in representation learning as the information bottleneck principle [17].



(a) Conditional generation (b) Style transfer

Figure 5: Sample results with varying dimension (N) and cardinality (K) of $\mathbf{U}$ and $\mathbf{V}$ with fixed size of common latent vector $\mathbf{Z}$ with $(N, K) = (1, 100)$. The presentation format is same to Figure 4.

References

[1] Aaron Wyner. The common information of two dependent random variables. *IEEE Trans. Inf. Theory*, 21(2):163–179, 1975.

[2] Abbas El Gamal and Young-Han Kim. *Network Information Theory*. Cambridge University Press, Cambridge, 2011.

[3] Paul Cuff. Distributed channel synthesis. *IEEE Trans. Inf. Theory*, 59(11):7071–7096, November 2013.

[4] Shao-Lun Huang, Gregory W. Wornell, and Lizhong Zheng. Gaussian universal features, canonical correlations, and common information. In *Proc. IEEE Inf. Theory Workshop*, pages 440–444, 2018.

[5] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.*, 112(518):859–877, 2017.

[6] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Int. Conf. Learn. Repr.*, 2014.

[7] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. Int. Conf. Mach. Learn.*, pages 1278–1286, 2014.

[8] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Int. Conf. Learn. Repr.*, 2017.

[9] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. In *Int. Conf. Learn. Repr.*, 2018.

[10] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Adv. Neural Info. Proc. Syst.*, pages 3581–3589, 2014.

[11] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Adv. Neural Info. Proc. Syst.*, pages 3483–3491, 2015.

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

[13] Rui Shu, Hung H. Bui, and Mohammad Ghavamzadeh. Bottleneck conditional density estimation. In *Proc. Int. Conf. Mach. Learn.*, 2017.

[14] Yann LeCun. The MNIST database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`, 1998.

[15] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *Int. Conf. Learn. Repr.*, 2016.

[16] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Int. Conf. Learn. Repr.*, 2016.

[17] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proc. 37th Ann. Allerton Conf. Comm. Control Comput.*, pages 368–377, 1999.