Adversarial Learning of a Variational Generative Model with Succinct Bottleneck Representation

J. Jon Ryu[1], Yoojin Choi[2], Young-Han Kim[1], Mostafa El-Khamy[2], Jungwon Lee[1,2]
[1]Dept. of ECE, University of California, San Diego, [2]SoC R&D, Samsung Semiconductor Inc.
[1]{jongharyu,yhk}@ucsd.edu, [2]{yoojin.c,mostafa.e,jungwon2.lee}@samsung.com

Abstract

A new bimodal generative model is proposed for generating conditional and joint samples, accompanied with a training method with learning a succinct bottleneck representation. The proposed model, dubbed as the variational Wyner model, is designed based on two classical problems in network information theory—distributed simulation and channel synthesis—in which Wyner's common information arises as the fundamental limit on the succinctness of the common representation. The model is trained by minimizing the symmetric Kullback–Leibler divergence between variational and model distributions with regularization terms for common information, reconstruction consistency, and latent space matching terms, which is carried out via an adversarial density ratio estimation technique.

1 Introduction

This paper studies how to learn a *good* common representation $\mathbf{Z}$ of a pair of an arbitrarily correlated random vectors $(\mathbf{X}, \mathbf{Y})$ from data, which is often referred to as the *cross-domain disentanglement* problem [4] in the representation learning literature [9]. This is a fundamental problem in machine learning with numerous applications including joint and conditional generative tasks (also known as domain transfer or image-to-image translation) and cross-domain retrieval tasks [25, 4, 10, 12, 5, 24, 15]. The main difficulty of this problem lies with the lack of a notion of a "good" common representation $\mathbf{Z}$ that captures the commonality of $(\mathbf{X}, \mathbf{Y})$. While there have been several information theoretic proposals on learning a good bottleneck representation including the famous information bottleneck principle [20] and a recent proposal [6], to name a few, there is no definitive answer in the literature. This work proposes a new information-theoretic representation learning principle with a recipe for training its deep-generative-model manifestation.

To motivate our perspective, consider the following game between Alice ("encoder") and Bob ("decoder") that captures the problem setting of conditional generation. Given an image of a child's photo $\mathbf{X}$, Alice is asked to encode $\mathbf{X}$ and send its description $\mathbf{Z}$ to Bob who draws a portrait $\mathbf{Y}$ of how the child will grow up based on it. In this game, we wish Bob to draw nice adulthood portraits, as various as possible, given a child's photo. In this cooperative game, Alice needs to help Bob in the process by providing a *good* description $\mathbf{Z}$ of the child's photo $\mathbf{X}$. Intuitively, seeking the *most succinct description* $\mathbf{Z}$ that contains information *common in* $\mathbf{X}$ *and* $\mathbf{Y}$ may be beneficial in their guessing process, since Alice need not describe any extra information beyond that is contained in $\mathbf{X}$ and Bob is not required to filter out any redundant information from $\mathbf{Z}$ for generating $\mathbf{Y}$.

P. Cuff (2013) formulated this game of conditional generation as the *channel synthesis* problem in network information theory and characterized the minimum description rate for such conditional generation by Wyner's *common information* [22, 3] denoted by $J(\mathbf{X}; \mathbf{Y})$ and defined as the optimal value of the optimization problem

$$\text{minimize} \quad I(\mathbf{X}, \mathbf{Y}; \mathbf{Z}) \qquad \text{subject to} \quad \mathbf{X} \leftrightarrow \mathbf{Z} \leftrightarrow \mathbf{Y}. \tag{1}$$

Submitted to the 6th Bayesian Deep Learning (BDL) Workshop at 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Here, $I(\mathbf{X},\mathbf{Y};\mathbf{Z})$ is the mutual information between $(\mathbf{X},\mathbf{Y})$ and $\mathbf{Z}$ and $\mathbf{X} \leftrightarrow \mathbf{Z} \leftrightarrow \mathbf{Y}$ denotes that $\mathbf{X},\mathbf{Z},\mathbf{Y}$ form a Markov chain, or equivalently, $\mathbf{X}$ is independent of $\mathbf{Y}$ given $\mathbf{Z}$ [1]. Furthermore, notably, the same quantity $J(\mathbf{X};\mathbf{Y})$ arises as the fundamental limit of the *distributed simulation* of correlated sources studied originally by A. Wyner (1975) in which two distributed agents wish to simulate a target distribution $q(\mathbf{x},\mathbf{y})$ (i.e., joint generation of $(\mathbf{X},\mathbf{Y})$) based on the least possible amount of shared common randomness. In this sense, the joint distribution $q(\mathbf{x},\mathbf{y})$ and the conditional distributions $q(\mathbf{y}|\mathbf{x})$, $q(\mathbf{x}|\mathbf{y})$ have the same common information structure characterized by the optimization problem (1).

2 The Proposed Method

(1) The variational Wyner model Thus motivated, we propose a probabilistic model that finds a common representation of the joint and conditional distributions and its training objectives based on the Wyner's optimization problem (1). For modeling the joint distribution $q(\mathbf{x},\mathbf{y})$, we consider the latent variable model $p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{y}|\mathbf{z})$, where $\mathbf{Z} \sim p_\theta(\mathbf{z})$ signifies the common randomness fed into the *probabilistic* decoders $p_\theta(\mathbf{x}|\mathbf{z})$ and $p_\theta(\mathbf{y}|\mathbf{z})$. We further parameterize the probabilistic decoders $p_\theta(\mathbf{x}|\mathbf{z})$ and $p_\theta(\mathbf{y}|\mathbf{z})$ by (deterministic) functions $\mathbf{x}_\theta(\mathbf{z},\mathbf{u})$ and $y_\theta(\mathbf{z},\mathbf{v})$ with independent *local randomness* $\mathbf{U} \sim p_\theta(\mathbf{u})$ and $\mathbf{V} \sim p_\theta(\mathbf{v})$. To model the conditional distribution $q(\mathbf{y}|\mathbf{x})$, we consider the bottleneck conditional model $q_\theta(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{y}|\mathbf{z})$ that follows $\mathbf{X} \leftrightarrow \mathbf{Z} \leftrightarrow \mathbf{Y}$; note that the decoder $p_\theta(\mathbf{y}|\mathbf{z})$ is shared by the joint model. The other direction for modeling $q(\mathbf{x}|\mathbf{y})$ is symmetric. Lastly, we introduce three additional *variational* encoders: (1) a joint encoder $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ that plays a key role of an anchor for tying the joint and conditional models, and (2) two local encoders $q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x})$ and $q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})$, which can be viewed as *style extractors* for each modality $\mathbf{x}$ and $\mathbf{y}$. Note that the local encoders $q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x})$ and $q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})$ are designed to satisfy the conditional independence structure $q_\phi(\mathbf{z},\mathbf{u},\mathbf{v}|\mathbf{x},\mathbf{y}) = q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x})q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})$ implied by the joint model $p_\theta(\mathbf{z})p_\theta(\mathbf{u})p_\theta(\mathbf{v})\mathbf{x}_\theta(\mathbf{z},\mathbf{u})\mathbf{y}_\theta(\mathbf{z},\mathbf{v})$, differing from existing works, e.g., [4, 6, 21].

We call the entire model with all the components introduced above as the (bimodal) *variational Wyner model*. See Table 1 for a summary of the four distributions over $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v})$ defined under the variational Wyner model and their shorthand notations.

Table 1: Induced distributions and their shorthand notation.

Type	Distribution over $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v})$	Notation
joint ($\to$ xy)	$p_\theta(\mathbf{z})p_\theta(\mathbf{u})p_\theta(\mathbf{v})\mathbf{x}_\theta(\mathbf{z},\mathbf{u})\mathbf{y}_\theta(\mathbf{z},\mathbf{v})$	$p_{\to\mathsf{xy}}$
cond. (x $\to$ y)	$q(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{v})\mathbf{y}_\theta(\mathbf{z},\mathbf{v})q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x})$	$p_{\mathsf{x}\to\mathsf{y}}$
cond. (y $\to$ x)	$q(\mathbf{y})q_\theta(\mathbf{z}|\mathbf{y})p_\theta(\mathbf{u})\mathbf{x}_\theta(\mathbf{z},\mathbf{u})q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})$	$p_{\mathsf{y}\to\mathsf{x}}$
variational (xy $\to$)	$q(\mathbf{x},\mathbf{y})q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x})q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})$	$q_{\mathsf{xy}\to}$

(2) Training objectives The main components of our training objectives are derived from the Wyner's optimization problem (1). For each model distribution $p_{\mathsf{model}} \in \{p_{\to\mathsf{xy}}, p_{\mathsf{x}\to\mathsf{y}}, p_{\mathsf{y}\to\mathsf{x}}\}$, we can convert (1) with variational relaxation to an unconstrained Lagrangian form $\mathrm{minimize}\ \mathcal{D}^{\mathsf{xyzuv}}_{\mathsf{model}} + \lambda^{\mathsf{CI}}_{\mathsf{model}} I_{\mathsf{model}}$. Here we define the *distribution matching term* $\mathcal{D}^{\mathsf{xyzuv}}_{\mathsf{model}} := D_{\mathsf{sym}}(q_{\mathsf{xy}\to}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}), p_{\mathsf{model}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}))$ for the symmetric KL divergence $D_{\mathsf{sym}}(p(\mathbf{s}),q(\mathbf{s})) := D_{\mathsf{KL}}(p(\mathbf{s}) \parallel q(\mathbf{s})) + D_{\mathsf{KL}}(q(\mathbf{s}) \parallel p(\mathbf{s}))$ [7] following Pu et al. [16] and define the *common information (CI) regularization term* $I_{\mathsf{model}} := I_{\mathsf{model}}(\mathbf{X},\mathbf{Y};\mathbf{Z}) := D_{\mathsf{KL}}(p_{\mathsf{model}}(\mathbf{x},\mathbf{y},\mathbf{z}) \parallel p_{\mathsf{model}}(\mathbf{x},\mathbf{y})p_{\mathsf{model}}(\mathbf{z}))$. On top of the base objectives, we further propose to add several terms that can guide training, including (1) reconstruction consistency terms (joint $\mathcal{R}_{\mathsf{xy}\to\mathsf{xy}}$ (data), $\mathcal{R}_{\mathsf{zuv}\to\mathsf{zuv}}$ (latent); conditional $\mathcal{R}_{\mathsf{x}\to\mathsf{y}}, \mathcal{R}_{\mathsf{y}\to\mathsf{x}}$; marginal $\mathcal{R}_{\mathsf{x}\to\mathsf{x}}, \mathcal{R}_{\mathsf{y}\to\mathsf{y}}$, all definitions omitted), (2) latent matching terms ($\mathcal{M}_{\mathsf{x}\to\mathsf{y}} := D_{\mathsf{sym}}(p_{\mathsf{x}\to\mathsf{y}}(\mathbf{z}), p_\theta(\mathbf{z}))$ and $\mathcal{M}_{\mathsf{y}\to\mathsf{x}}$) that enforce the joint and conditional models to share the same latent space, and (3) the *cross-matching* term $\mathcal{D}^{\mathsf{xyzuv}}_{\mathsf{x}\leftrightarrow\mathsf{y}} := D_{\mathsf{sym}}(p_{\mathsf{x}\to\mathsf{y}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}), p_{\mathsf{y}\to\mathsf{x}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}))$ that may improve the quality of representation for cross-domain tasks. When training, we optimize a weighted combination of the proposed objectives, where the weights are hyperparameters to be tuned.

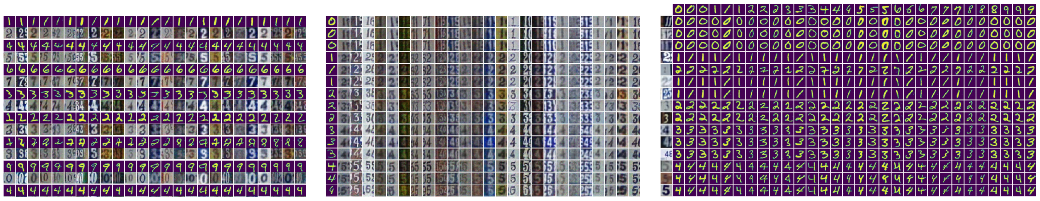
(3) Approximate training with adversarial density ratio estimation To minimize the proposed objective without explicitly assuming densities of the model distributions, we adopt the adversarial density ratio estimation technique proposed by [16]. The idea is to estimate the density ratio $p(\mathbf{s})/q(\mathbf{s})$ by a solution $r(\mathbf{s})$ of the following variational characterization of the Jensen–Shannon divergence $D_{\mathsf{JS}}(p(\mathbf{s}),q(\mathbf{s})) = \max_{r(\mathbf{s})}\{\mathbb{E}_{p(\mathbf{s})}[\log\sigma(\log r(\mathbf{s}))] + \mathbb{E}_{q(\mathbf{s})}[\log\sigma(-\log r(\mathbf{s}))]\}$, since the maximum is attained if and only if $r^*(\mathbf{s}) \equiv p(\mathbf{s})/q(\mathbf{s})$. Here, $\sigma(x) = 1/(1+e^{-x})$ denotes the sigmoid function. As in a typical GAN training procedure, we alternate between training the variational Wyner model components and training the discriminators batch-by-batch, freezing one while training the other. When training the variational Wyner model, we freeze the density ratio estimators and estimate $D_{\mathsf{sym}}(p(\mathbf{s}), q(\mathbf{s})) \approx \mathbb{E}_{p(\mathbf{s})}[\log r(\mathbf{s})] - \mathbb{E}_{q(\mathbf{s})}[\log r(\mathbf{s})]$ assuming that $r(\mathbf{s}) \approx p(\mathbf{s})/q(\mathbf{s})$. A mutual information $I_{\mathsf{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$ term can be handled by the same technique. We use the tilde notation to denote a corresponding discriminator objective of a generator objective requiring a discriminator, e.g., e.g., $\tilde{\mathcal{D}}^{\mathsf{xyzuv}}_{\to\mathsf{xy}}$ for $\mathcal{D}^{\mathsf{xyzuv}}_{\to\mathsf{xy}}$. Lastly, our density ratio estimation network has several important design choices, including (1) a shared joint data feature map with (2) deterministic parameterization of encoders using the instance noise trick [19], for computational efficiency and stable training.

3 Experiment

(1) MNIST–SVHN add-one dataset To validate the efficacy of the proposed approach and illustrate the effect of information decomposition in our model, we considered a synthetic image-image pair dataset constructed from MNIST [8] and SVHN [14] datasets, similar to Shi et al. [18] as a toy example. Here, we randomly picked an MNIST image $\mathbf{X}_i$ of label $\ell_i \in \{0, \ldots, 9\}$ and paired with $m = 4$ randomly picked SVHN images of label $(\ell_i + 1) \mod 10$; we call the resulting dataset a MNIST-SVHN add-one dataset. Note that the images are paired only through their labels, and clearly the common information structure we seek is the underlying label of a pair.

We trained the variational Wyner model with all joint and conditional models, with the objective $\mathcal{D}^{\mathsf{xyzuv}}_{\to\mathsf{xy}} + \mathcal{D}^{\mathsf{xyzuv}}_{\mathsf{x}\to\mathsf{y}} + \mathcal{D}^{\mathsf{xyzuv}}_{\mathsf{y}\to\mathsf{x}} + \lambda^{\mathsf{CI}}(I_{\to\mathsf{xy}} + I_{\mathsf{x}\to\mathsf{y}} + I_{\mathsf{y}\to\mathsf{x}}) + \mathcal{R}_{\mathsf{xy}\to\mathsf{x}} + \mathcal{R}_{\mathsf{xy}\to\mathsf{y}} + \mathcal{R}_{\mathsf{x}\to\mathsf{y}} + \mathcal{R}_{\mathsf{y}\to\mathsf{x}}$ for training the variational Wyner model and $\tilde{\mathcal{D}}^{\mathsf{xyzuv}}_{\to\mathsf{xy}} + \tilde{\mathcal{D}}^{\mathsf{xyzuv}}_{\mathsf{x}\to\mathsf{y}} + \tilde{\mathcal{D}}^{\mathsf{xyzuv}}_{\mathsf{y}\to\mathsf{x}} + \tilde{I}_{\to\mathsf{xy}} + \tilde{I}_{\mathsf{x}\to\mathsf{y}} + \tilde{I}_{\mathsf{y}\to\mathsf{x}}$ for training the discriminator with the dimension of the latent space $(\mathbf{Z}, \mathbf{U}, \mathbf{V}) = (16, 8, 8)$. We tried four different CI regularization weight $\lambda^{\mathsf{CI}} \in \{0, 0.1, 0.5, 1\}$ to demonstrate the effect of the regularization for 50 epochs and the averaged ℓ_1 distance over dimensions was used for the reconstruction loss functions.

In Figure 1, we present a few joint and conditional samples generated from the trained model with $\lambda^{\mathsf{CI}} = 1$ at the end of training. In the figure, $\mathbf{z}$ is shared across the row, and $\mathbf{u}$ and/or $\mathbf{v}$ are shared across the column. In particular, the top row of the last panel (c) shows the reference samples whose style are transferred downward along each column. The samples clearly indicate that the learned model successfully disentangles the common and local representations. For example, in Figure 1(b), in the first three rows, regardless of the specifics of the input MNIST images independent to their label 0, the generated samples coherently present the correct label 1 as well as sharing the same style fixed along each column. Figure 1(c) illustrates that using the local variational encoder $q_\phi(\mathbf{u}|\mathbf{z}, \mathbf{x})$ we can generate conditional samples given a fixed style extracted from a reference image.



(a) Joint generation

(b) Conditional generation

(c) Conditional generation with style transfer

Figure 1: MNIST–SVHN add-one samples from the variational Wyner model.

In order to numerically examine the effect of CI regularization in the model, we computed two metrics (1) custom Frechet distance (FD) scores and (2) classification accuracy of generated samples; the details for computing these metrics will be reported in a full paper. The results are summarized in Figure 2. The dashed lines are the evaluated metrics for a baseline model whose discriminator was trained only with distribution matching terms, i.e., without any CI related term. As shown in the figure, increasing λ^{CI} improves the quality of generated samples in terms of the smaller FD scores and improved the digit accuracy. Note that the high digit accuracy with large FD score under no or small CI regularization indicates that such models generate coherent samples yet with less variation.

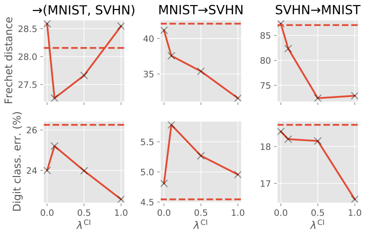


Figure 2: A summary of numerical evaluations for MNIST–SVHN add-one dataset.

(2) Zero-shot sketch based image retrieval To demonstrate the utility of learned representations beyond generative modeling, we consider the *zero-shot sketch based image retrieval* (ZS-SBIR) task proposed by Yelamarthi et al. [23], where the goal is to construct a good retrieval model that retrieves relevant photos from a sketch, with a training set of no overlapping classes with a test set. For this experiment, we borrowed the the same setting from Hwang et al. [6]. We trained and evaluated our model with the Sketchy Extended dataset [17, 13], which consists of total 75,479 sketches ($\mathbf{X}$) and 73,002 photos ($\mathbf{Y}$) from 125 different classes. To perform the retrieval task after training, we first find and keep the common representations $\{\mathbf{z}_i\}$ of test photos $\{\mathbf{y}_i\}$ using the model encoder $q_\theta(\mathbf{z}|\mathbf{y})$. Then, given a query sketch $\mathbf{x}_o$, we find the common representation $\mathbf{z}_o \sim q_\theta(\mathbf{z}|\mathbf{x}_o)$ to retrieve the K-nearest neighbors of $\mathbf{z}_o$ from $\{\mathbf{z}_i\}$ with respect to the cosine similarity.

For this task, we trained our model only with conditional model components, as we only need to learn good model encoders $q_\theta(\mathbf{z}|\mathbf{x})$ and $q_\theta(\mathbf{z}|\mathbf{y})$. Specifically, we trained with the objective $\mathcal{D}^{\mathsf{xyzuv}}_{\mathsf{x}\to\mathsf{y}} + \mathcal{D}^{\mathsf{xyzuv}}_{\mathsf{y}\to\mathsf{x}} + \mathcal{D}^{\mathsf{xyzuv}}_{\mathsf{x}\leftrightarrow\mathsf{y}} + \lambda^{\mathsf{CI}}(I_{\mathsf{x}\to\mathsf{y}} + I_{\mathsf{y}\to\mathsf{x}}) + \lambda^{\mathsf{rec}}(\mathcal{R}_{\mathsf{x}\to\mathsf{y}} + \mathcal{R}_{\mathsf{y}\to\mathsf{x}} + \mathcal{R}_{\mathsf{x}\to\mathsf{x}} + \mathcal{R}_{\mathsf{y}\to\mathsf{y}})$ for training the variational Wyner model and $\tilde{\mathcal{D}}^{\mathsf{xyzuv}}_{\to\mathsf{xy}} + \tilde{\mathcal{D}}^{\mathsf{xyzuv}}_{\mathsf{x}\to\mathsf{y}} + \tilde{\mathcal{D}}^{\mathsf{xyzuv}}_{\mathsf{y}\to\mathsf{x}} + \tilde{I}_{\mathsf{x}\to\mathsf{y}} + \tilde{I}_{\mathsf{y}\to\mathsf{x}}$ for training the discriminator. We used the ℓ_2^2-distance averaged over dimensions for the reconstruction loss functions. The dimension of the latent space $(\mathbf{Z}, \mathbf{U}, \mathbf{V})$ was $(64, 64, 64)$.

Table 2: Evaluation of the ZS-SBIR task with the Sketchy Extended dataset.

Models	P@100	mAP
LCALE [11]	0.583	0.476
IIAE [6]	0.659	0.573
Variational Wyner	**0.703**	**0.629**

As a quantitative evaluation, we computed the Precision@100 (P@100) and mean average precision (mAP) scores for the test split; see Table 2. The reported scores for the adversarially learned Wyner model was obtained with $\lambda^{\mathsf{CI}} = 0.1$ and $\lambda^{\mathsf{rec}} = 8$. We outperform the scores reported by Hwang et al. [6], who already demonstrated that their scores significantly improved upon the existing work tailored to extra information; for example, LCALE [11] incorporated word embedding during training. The improvement corroborates the power of our approach in learning disentangled representations. For an ablation study, we trained our model with degenerate local encoders $q_\phi(\mathbf{u}|\mathbf{x})$ and $q_\phi(\mathbf{u}|\mathbf{y})$, i.e., without conditioning with $\mathbf{z}$, and achieved suboptimal scores (0.670,0.591); it justifies the design of our local encoders $q_\phi(\mathbf{u}|\mathbf{z}, \mathbf{x})$ and $q_\phi(\mathbf{v}|\mathbf{z}, \mathbf{y})$.

Some examples of retrieved photos are show in Figure 3. Note that most of the falsely retrieved photos share visual similarity with the query sketches.

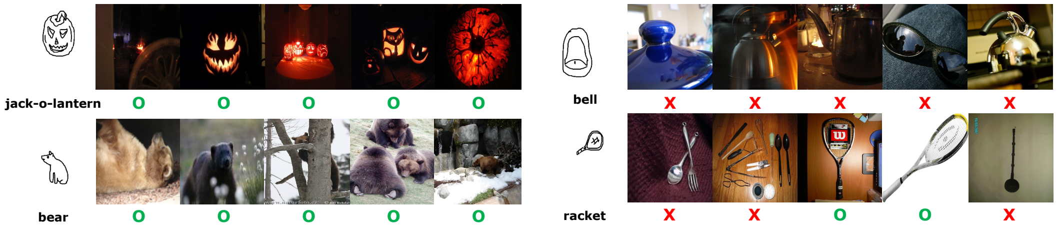


Figure 3: A few examples of retrieved samples from the Sketchy Extended dataset. For each query sketch, the top-5 retrieved images are shown, where the top-1 is in the leftmost. The O/X's indicate whether the retrievals belong to the same class of the query.

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, New York, second edition, 2006.

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

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

[4] Abel Gonzalez-Garcia, Joost Van De Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Adv. Neural Info. Proc. Syst.*, volume 31, 2018.

[5] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proc. Eur. Conf. Comput. Vis.*, pages 172–189, 2018.

[6] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Variational interaction information maximization for cross-domain disentanglement. *Adv. Neural Info. Proc. Syst.*, 33, 2020.

[7] Harold Jeffreys. *The Theory of Probability*. OUP Oxford, 1998.

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

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[10] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proc. Eur. Conf. Comput. Vis.*, pages 35–51, 2018.

[11] Kaiyi Lin, Xing Xu, Lianli Gao, Zheng Wang, and Heng Tao Shen. Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval. In *Proc. AAAI Conf. Artif. Int.*, volume 34, pages 11515–11522, 2020.

[12] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *Adv. Neural Info. Proc. Syst.*, volume 31, 2018.

[13] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 2862–2871, 2017.

[14] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[15] Ori Press, Tomer Galanti, Sagie Benaim, and Lior Wolf. Emerging disentanglement in auto-encoder based unsupervised image content transfer. In *Int. Conf. Learn. Repr.*, 2019.

[16] Yuchen Pu, Weiyao Wang, Ricardo Henao, Liqun Chen, Zhe Gan, Chunyuan Li, and Lawrence Carin. Adversarial symmetric variational autoencoder. In *Advances in Neural Information Processing Systems*, pages 4330–4339, 2017.

[17] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 2016.

[18] Yuge Shi, Narayanaswamy Siddharth, Brooks Paige, and Philip HS Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Adv. Neural Info. Proc. Syst.*, volume 32, 2019.

[19] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. In *Int. Conf. Learn. Repr.*, 2017. URL `https://openreview.net/forum?id=S1RP6GLle`.

[20] 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.

[21] Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

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

[23] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proc. Eur. Conf. Comput. Vis.*, pages 300–317, 2018.

[24] Xiaoming Yu, Yuanqi Chen, Thomas Li, Shan Liu, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *Adv. Neural Info. Proc. Syst.*, volume 32, 2019.

[25] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Adv. Neural Info. Proc. Syst.*, volume 30, 2017.