Cross-population Variational Autoencoders

Joe Davison[2,1], Kristen A. Severson[1], and Soumya Ghosh[1]

[1]MIT-IBM Watson AI Lab, IBM Research, Cambridge, MA
[2]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA
jddavison@g.harvard.edu, kristen.severson@ibm.com, ghoshso@us.ibm.com

1 Introduction

Unsupervised learning of latent representations is useful for a variety of tasks including dimensionality reduction, density estimation, and structure or sub-group discovery. Methods for recovering such representations typically rely on the assumption that the observed data is a manifestation of only a limited number of factors of variation [1, 2]. Variational autoencoders (VAE) [3], a combination of a non-linear latent variable model and an amortized inference scheme [4], is a popular method for recovering such latent structure. VAEs and its extensions have received considerable attention in recent years and have been shown useful for modeling text [5], images [6] , and other data exhibiting complex correlations [7]. However, barring a few notable exceptions [8], the vast majority of this line of work assumes that the data being modeled is independent and identically distributed.

In this work, we consider the task of modeling independent but not identically distributed data. In particular, we are interested in modeling data comprising two or more distinct but related sub-populations, with the intent of isolating latent representations that capture factors of variation common to all populations from those unique to particular populations. We show that by building models and inference procedures that are aware of the heterogeneity in data, we are able to learn representations which are both salient and disentangled across differing populations.

2 Cross-population Variational Autoencoder

We propose a model that generates data instance i of population k ($\mathbf{x}_{ki}$) given two latent variables $\mathbf{z}_{ki}$ and $\mathbf{t}_{ki}$. The latent variables are projected to the observed space using non-linear mappings, $f_{\theta_s}(\mathbf{z}_{ki})$ and $f_{\theta_k}(\mathbf{t}_{ki})$, each parameterized by a neural network. While we share the parameters θ_s among all populations, θ_k are only shared among instances belonging to the population k. This construction encourages the model to capture common latent structure in θ_s, while allowing θ_k to focus on factors of variation unique to the particular population k. We combine the contributions from the two mappings using an aggregation function g. The generative procedure can be summarized as,

$$\begin{aligned} &\mathbf{z}_{ki} \sim \mathcal{N}(0, \mathbf{I}), \qquad \mathbf{t}_{ki} \sim \mathcal{N}(0, \mathbf{I}), \\ &\mathbf{x}_{ki} \mid \mathbf{z}_{ki}, \mathbf{t}_{ki} \sim p(g(f_{\theta_s}(\mathbf{z}_i), f_{\theta_k}(\mathbf{t}_{ki}))), \quad \forall i \in \{1 \dots n_k\}, \forall k \in \{1, \dots, K\}, \end{aligned} \tag{1}$$

where p is an appropriately chosen distribution for modeling the observed data. In our experiments, we use an additive aggregation function $g(a, b) = a + b$, and select p to be the Gaussian distribution with mean parameterized by g and an isotropic diagonal covariance matrix, Ψ. We emphasize that other choices of aggregation function can easily be incorporated into the model. Exploring the space of aggregation functions is planned future work. Figure 1 presents the graphical model summarizing the conditional dependencies assumed by the model.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

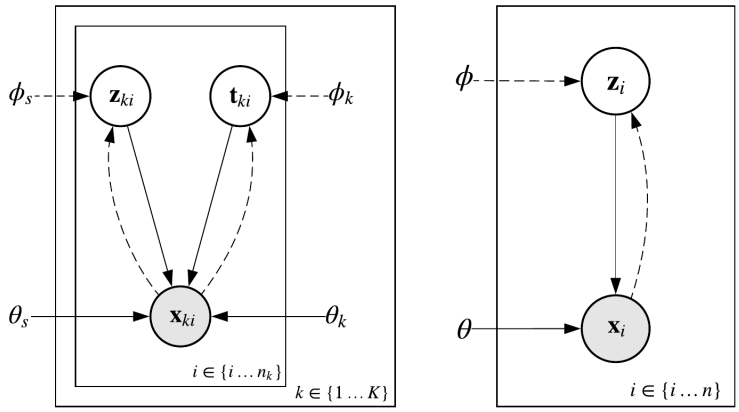

Figure 1: Graphical model of CPVAE (left) vs. standard VAE (right)

Amortized Variational Inference Amortized variational inference [4, 9] based on reparameterized gradients [10, 3] is straightforward to implement for the model described in Equation 1. We assume that the variational approximation factorizes conditioned on the observation,

$$q_\phi(\mathbf{z}_{ki}, \mathbf{t}_{ki} \mid \mathbf{x}_{ki}) = q_{\phi_s}(\mathbf{z}_{ki} \mid \mathbf{x}_{ki}) q_{\phi_k}(\mathbf{t}_{ki} \mid \mathbf{x}_{ki}). \tag{2}$$

We parameterize the variational distribution for each population with a single inference network with two distinct outputs, one for each latent variable. The model and variational parameters can then be jointly learned by optimizing the evidence lower bound (ELBO),

$$\begin{aligned}\mathcal{L}(\theta, \phi) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} & \mathbb{E}_{q_\phi(\mathbf{z}_{ki}, \mathbf{t}_{ki} | \mathbf{x}_{ki})}[\log p(\mathbf{x}_{ki} \mid \mathbf{z}_{ki}, \mathbf{t}_{ki}; \theta_s, \theta_k)] \\ & - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}_{ki}, \mathbf{t}_{ki} \mid \mathbf{x}_{ki}) || p(\mathbf{z}_{ki}, \mathbf{t}_{ki})\right)\end{aligned} \tag{3}$$

Importantly, in the above setup the per-population model and variational parameters θ_k and ϕ_k are learned only from the data in that population, encouraging them to learn representations for describing only that population, while the shared variational parameters θ_s and ϕ_s are learned for all data. We refer to this combination of the model described in equation 1 and the inference network as the cross population variational autoencoder (CPVAE), owing to its similarity with the VAE (see Figure 1) [3].

Mutual Information Regularized Inference The population specific mappings of CPVAE encourage latent factors of variation unique to the population to be modeled by the population specific latent variables $\mathbf{t}$. However, since every data instance is generated by a combination of shared and population specific representations, CPVAE does not explicitly prevent population specific representations from modeling variation shared across populations. In fact, when CPVAE is trained by maximizing the ELBO in equation 3 we find that the private representations of a population often exhibit features from the shared space and vice versa. This leads to "leakage" between populations where the private spaces of different populations end up capturing the same factors of variation that are shared across populations instead of modeling unique, population specific variation.

To alleviate such issues, we need further constraints. In this work, we propose to maximize the following augmented objective inplace of the ELBO,

$$\begin{aligned}J(\theta, \phi) = & \frac{1}{N} \sum_{k=1}^{K} \left[\sum_{i=1}^{n_k} \mathbb{E}_{q_\phi(\mathbf{z}_{ki}, \mathbf{t}_{ki} | \mathbf{x}_{ki})}[\log p(\mathbf{x}_{ki} \mid \mathbf{z}_{ki}, \mathbf{t}_{ki}; \theta_s, \theta_k)] - D_{\mathrm{KL}}\left(q_{\phi_s}(\mathbf{z}_{ki} | \mathbf{x}_{ki}) || p(\mathbf{z}_{ki})\right) \right] \\ & - \sum_{k=1}^{K} \frac{1}{N - n_k} \sum_{j \neq k}^{K} \sum_{i=1}^{n_j} D_{\mathrm{KL}}\left(q_{\phi_k}\left(\tilde{\mathbf{t}}_{ji} | \mathbf{x}_{ji}\right) || \mathcal{N}(\tilde{\mathbf{t}}_{ji} \mid \mathbf{0}, \mathbf{I})\right),\end{aligned} \tag{4}$$

where N is the total number of data points and n_k is the number of data points in population k. Let $\mathbf{x}_k = \{\mathbf{x}_{k1}, \dots, \mathbf{x}_{kn_k}\}$ denote the set of all data instances belonging to population k and $\mathbf{t}_k, \mathbf{z}_k$ denote the corresponding latent variables. We use $\mathbf{x}_{-k} = \{\mathbf{x}_j; \forall j \neq k \in K\}$ to denote the set of all

non-corresponding populations. We further introduce pseudo latent variables $\tilde{\mathbf{t}}_k = \{\{\tilde{\mathbf{t}}_{ji}\}_{i=1}^{n_j}\}_{j \neq k}$ each endowed with a zero-mean, unit variance Gaussian prior and let $q_{\phi_k}(\tilde{\mathbf{t}}_{ji} \mid \mathbf{x}_{ji})$ where $\mathbf{x}_{ji} \in \mathbf{x}_{-k}$ denote the corresponding variational approximations. Crucially these approximations are parameterized by ϕ_k — the inference network of population k and condition on only the members of non-corresponding populations, $\mathbf{x}_{-k}$.

This modification allows the model to learn useful population-specific features in the private representations while penalizing learning of features common to multiple populations. To see why, observe that the inference network ϕ_k for a population k is encouraged to encode members of populations $j \neq k$ to uninformative zero-mean, unit variance Gaussians through the last KL term in Equation 4. Moreover, unlike ELBO, no penalty is placed on members of population k encouraging informative posteriors, $q_{\phi_k}(\mathbf{t}_{ki} \mid \mathbf{x}_{ki})$.

We can arrive at Equation 4 from Equation 3 by observing that,

$$J(\theta, \phi) \leq \mathbb{E}_{p_D(\mathbf{x})}\big[\mathcal{L}(\theta, \phi)\big] + \sum_{k=1}^{K} \mathbb{E}_{p_D(\mathbf{x}_k)}\big[D_{\mathrm{KL}}\left(q_{\phi_k}(\mathbf{t}_k|\mathbf{x}_k)||p(\mathbf{t}_k)\right)\big] - \sum_{k}^{K} I_q(\mathbf{x}_{-k}; \tilde{\mathbf{t}}_k). \quad (5)$$

The above expression regularizes the ELBO by minimizing the mutual information between population specific pseudo latent variables and data from non-corresponding populations, while encouraging the divergence between the posterior over the population specific latent variables ($\mathbf{t}_k$) and the prior $\mathcal{N}(\mathbf{t}_k \mid \mathbf{0}, \mathbf{I})$ to be high, and hence informative. The mutual information term can be further upper bounded,

$$\begin{aligned} I_q(\mathbf{x}_{-k}; \tilde{\mathbf{t}}_k) &= \mathbb{E}_{p_D(\mathbf{x}_{-k})}\left[D_{\mathrm{KL}}(q_{\phi_k}(\tilde{\mathbf{t}}_k|\mathbf{x}_{-k})||p(\tilde{\mathbf{t}}_k))\right] - D_{\mathrm{KL}}(q_{\phi_k}(\tilde{\mathbf{t}}_k)||p(\tilde{\mathbf{t}}_k)) \\ &\leq \frac{1}{N - n_k} \sum_{j \neq k}^{K} \sum_{i}^{n_j} \left[D_{\mathrm{KL}}(q_{\phi_k}(\tilde{\mathbf{t}}_{ji}|\mathbf{x}_{ji})||p(\tilde{\mathbf{t}}_{ji}))\right], \end{aligned} \quad (6)$$

which leads to the objective in Equation 4. See the appendix A for additional information as well as a summary of the training procedure.

3 Related Work

Substantial recent work has explored methods for the unsupervised learning of *disentangled* representations. Though lacking a formal definition, the key idea behind a disentangled representation is that it should separate distinct informative factors of variation of the data [2, 1]. Several techniques have been proposed to encourage VAEs to learn disentangled latent representations. Some examples are β-VAE [11], AnnealedVAE [12], FactorVAE [13], β-TCVAE [14], and DIP-VAE [15], each of which proposes some variant of the VAE objective to encourage the variational distribution to be factorizble. Recently, it has been proposed that it is not possible to recover disentangled features without inductive bias or supervision [1]. CPVAE makes no assumptions about the composition of variational factors but imposes a form of weak supervision where population assignment is known and uses that information in choosing the model architecture and learning algorithm.

A few other techniques have been proposed to use weak supervision for learning improved representations. Multi-study factor analysis [16], [17] has similar aims and structure as compared to CPVAE but uses a linear model and focuses on applications related to high-throughput biological assay data. Contrastive latent variable models [18, 19] have non-linear variants but focus on the case where one dataset is the target to be compared/contrastive to another dataset. Multi-level variational autoencoders (ML-VAEs) have been proposed as a way to incorporate group-level data in unsupervised learning [8]. After dividing data into disjoint groups according to some factor of interest, the ML-VAE framework models latent structure both at the level of individual observations and of entire groups. This method effectively separates latent representations into semantically relevant parts, but differs from the CPVAE framework which models both private and shared structure at the observation level. Output-interpretable VAEs (oi-VAEs) have been proposed to leverage data that can be partitioned into within sample groups in an interpretable model [20]. The model structure is such that the components within each group are modeled with separate generative networks. Mappings from the latent representations to each of the groups are encouraged to be sparse via hierarchical Bayesian priors to improve interpretability. The primary difference between oi-VAE and CPVAE

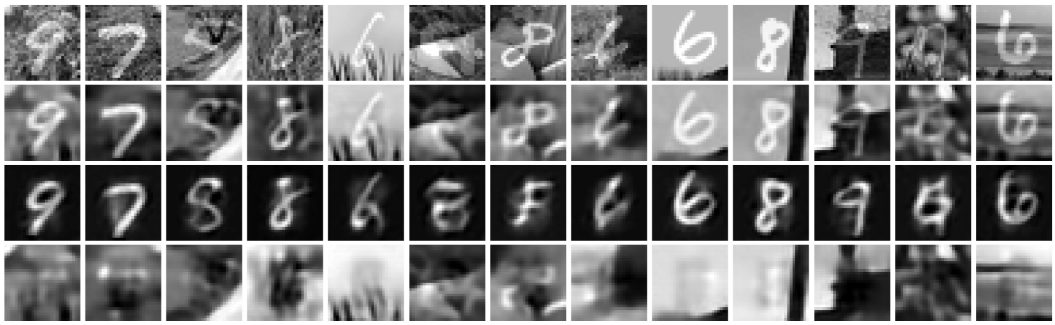

Figure 2: Grassy-MNIST reconstructions with two populations: digits 0-4 and digits 5-9. First row: original images. Second row: full reconstructions. Third row: private space (digit-only) reconstructions. Fourth row: shared space (background only) reconstructions. Note that this model has never seen the original digits or the original backgrounds.

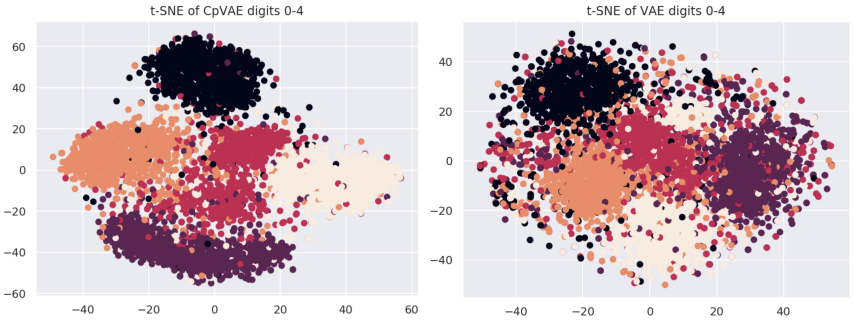

Figure 3: Visualization of subgroups within MNIST digit population after t-SNE projection of latent space for our method (left) vs a standard VAE (right). Each color represents one of the five digits within the population.

is that oi-VAE uses groupings over components, $\mathbf{x}_i \in \mathbb{R}^d$, $\mathbf{x}_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iK}]$, and $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ and CPVAE uses groupings over instances, $\mathbf{x}_{ki} \in \mathbb{R}^d$ and $\mathcal{D} = \{\{\mathbf{x}_{ki}\}_{i=1}^{n_k}\}_{k=1}^K$. There has also been recent work in learning disentangled representations from sequential data [21, 22]. These models share a representation across all elements of the sequence to learn global sequence dynamics while local aspects are modeled via time step specific representations. This problem is somewhere between the standard disentangled representation learning, where the goal is to learn disentangled features with no prior knowledge, and weak supervision as the sequential nature of the data inherently provides some structure. Our work presented here focuses on non-sequential data. Moreover, unlike us these works do not employ mutual information regularized inference, which we find to be crucial for recovering disentangled shared and private representations.

4 Experiments

In order to determine the effectiveness of our approach as well as demonstrate several possible applications, we evaluate it with a number of experiments on tasks including image denoising, sub-group discovery, and classification.

4.1 Denoised Generative Modeling Applied to Grassy-MNIST

We evaluate our model on a synthetic dataset of handwritten digits from the MNIST [23] superimposed on grassy backgrounds from ImageNet [24] (see Figure 2, top row for example images). In [25, 18, 19], the authors train contrastive models on this dataset along with the original grass images in order to learn more salient latent representations of the digits.

In our experiment, we split this synthetic data into two populations consisting of digits 0-4 and digits 5-9. We use a shared latent dimension of 100 and a population-specific latent size of 25. We show

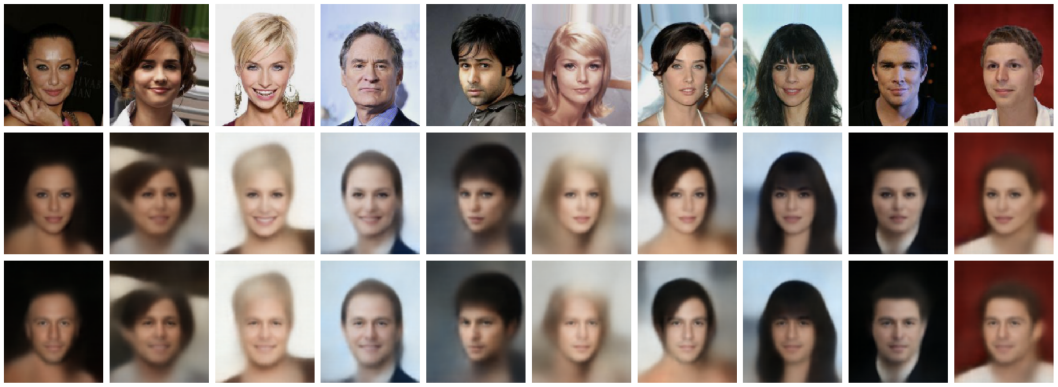

Figure 4: CPVAE reconstructions of CelebA dataset images. Top row: original images. Second row: reconstructions with `male` and shared space decoders. Third row: reconstructions with `non-male` and shared space decoders. Bottom row: shared decoder only.

Method	SHARED ARI	PRIVATE ARI
VAE	.1881	—
CPVAE, NO MI	.0306	.1926
CPVAE, $\alpha = 1$	.0031	.2281

Table 1: Adjusted rand index (ARI) for discovered clusters in the shared vs. private latent spaces.

that our model is able to effectively separate the complex background representations in the shared latent space from that of the digits in each private space. Crucially, the model is never shown either the original background grass images or the original non-noisy digits.

A sample of reconstructed images can be seen in Figure 2. The first and second rows show the original digits with noise and the full CPVAE reconstructions, respectively. The third and fourth rows show the reconstructions when the shared or the private latent spaces are ignored, respectively. These results qualitatively demonstrate CPVAE's ability to separate the shared and private feature representations.

As an evaluation of the salience of the private space representations, we also measure the ability of our model to do sub-group discovery within each population. K-means clustering is applied to the private space latent representations and compared to the true class labels using the adjusted rand index (ARI), which measures the correspondence of the cluster assignments with the true labels [26]. The results are in Table 1 and a TSNE visualization of one of the private latent spaces for this task compared with that of a standard VAE is in Figure 3. Our model outperforms the standard VAE, demonstrating the ability of our model to learn improved representations by incorporating cross-population structure. Utilizing the MI objective further improves ARI scores in the private space and worsens scores in the shared space, suggesting that the MI term effectively mitigates the leakage of population-specific features in the shared space.

4.2 Disentangling Labeled Attributes in Celebrity Images

CPVAE allows us to model data such that latent structure which corresponds to some labeled attribute of interest can be isolated. To demonstrate this, we perform an experiment on the Large-scale CelebFaces Attributes (CelebA) dataset [27]. This data consists of 202,599 aligned and cropped pictures of celebrities with 40 binary attributes labeled for each image. See Figure 4.

We train a CPVAE model on this dataset with two populations determined by the `male` attribute label. Under this setup, the model is incentivized to learn representations of gender-specific features in the private latent spaces while the shared space infers the remaining factors of variation. As a qualitative evaluation of this model, we autoencode a sample of images with each private space and examine the resulting reconstructions. By reconstructing an image with the non-corresponding population's

DATASET	ANNEAL α	$\alpha = 1$	NO MI	RESNET50
MNIST	99.5	98.2	97.3	99.6
MNIST-GRASS	77.5	75.1	70.5	85.7
CIFAR-10	76.2	47.9	42.2	84.1

Table 2: Maximum likelihood classification test accuracies (%) for our model with and without mutual information terms evaluated on different datasets. For reference, we also include accuracies from a ResNet50 classifier.

decoder, we get reconstructions that closely resemble the original image but with features that appear traditionally male when constructed from the male space and female when constructed from the non-male space. See Figure 4 for a sample of these results. This serves as additional evidence of our model's ability to separate population-specific features from shared latent structure.

4.3 Maximum Marginal Likelihood Classification

As an additional evaluation of our model's ability to learn disentangled population-specific representations, we test our model's ability to classify unseen data points into their corresponding populations. We assign an instance $\mathbf{x}_*$ to the population that maximizes its marginal likelihood,

$$\hat{k}_i = \arg\max_{k \in K} p(\mathbf{x}_* \mid \theta_k, \theta_s) = \arg\max_{k \in K} \mathbb{E}_{p(\mathbf{z}_{ki}, \mathbf{t}_{ki})} \left[p(\mathbf{x}_* \mid \mathbf{z}_{ki}, \mathbf{t}_{ki}; \theta_k, \theta_s) \right]. \tag{7}$$

We use importance sampling to compute the intractable expectation,

$$\mathbb{E}_{p(\mathbf{z}_{ki}, \mathbf{t}_{ki})} \left[p(\mathbf{x}_* \mid \mathbf{z}_{ki}, \mathbf{t}_{ki}; \theta_k, \theta_s) \right] = \mathbb{E}_{q_\phi(\mathbf{z}_{ki}, \mathbf{t}_{ki} \mid \mathbf{x}_i)} \left[p(\mathbf{x}_* \mid \mathbf{z}_{ki}, \mathbf{t}_{ki}; \theta_k, \theta_s) \frac{p(\mathbf{z}_{ki}, \mathbf{t}_{ki})}{q_\phi(\mathbf{z}_{ki}, \mathbf{t}_{ki} \mid \mathbf{x}_*)} \right]. \tag{8}$$

In order to achieve a high accuracy, the model must learn features in each population-specific latent space which are unique to its corresponding set. We therefore evaluate our model's classification performance on several labeled image datasets of varying difficulty — MNIST, CIFAR-10 [28], and the Grassy MNIST described in experiments above. In each case, we define a distinct population for each class and evaluate its performance on a held-out test set. We emphasize that our goal with this experiment is not to demonstrate state-of-the-art classification performance, but to provide a convenient, quantitative benchmark for evaluating the quality of the model's learned representations. For reference, we also provide classification accuracies from a ResNet-50 convolutional model [29, 30] trained by maximizing $p(k_* \mid \mathbf{x}_*)$ for five hundred epochs. Note that these CNN scores are not the state of the art for each task, but serve as a contextual reference point for understanding our model's performance. We compare this against our CPVAE models both with and without the mutual information regularization as well as a variant where the weight on the mutual information term, α, is gradually annealed over training.

The results can be seen in Table 2. Our model approaches the performance of the convolutional neural network despite not being trained to directly maximize classification performance. We find that performance improves by utilizing our mutual information regularized objective, particularly when α is annealed over training. This result provides compelling evidence that our model is able to effectively learn private space representations which are unique to each corresponding population.

5 Conclusion

In this work, we presented a framework for using a VAE-like architecture to model multiple sets of data which are independent but come from differing distributions. We developed an architecture which encourages the isolation of shared and private latent factors, and presented a mutual information regularized version of the evidence lower bound which discourages entanglement of the shared and population-specific latent vectors. Our experiments on the Grassy MNIST dataset demonstrated our model's ability to learn more salient representations and to effectively separate the shared and private latent factors on the task of image denoising. We also showed the effectiveness of our regularized objective in learning population-specific representations on several image classification tasks.

References

[1] F. Locatello, S. Bauer, M. Lucie, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *ICML*, 2019.

[2] Y. Bengio, A. Courvile, and P. Vincent, "Representation learning: A review and new perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, 2013.

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

[4] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The helmholtz machine," *Neural Computation*, vol. 5, pp. 889–904, 1995.

[5] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in *International conference on machine learning*, 2016, pp. 1727–1736.

[6] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville, "Pixelvae: A latent variable model for natural images," *arXiv preprint arXiv:1611.05013*, 2016.

[7] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS central science*, vol. 4, no. 2, pp. 268–276, 2018.

[8] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multi-level variational autoencoder: Learning disentangled representations from grouped observations," in *AAAI*, 2018.

[9] S. Gershman and N. Goodman, "Amoritized inference in probabilistic reasoning," in *Annual Meeting of the Cognitive Science Society*, 2014.

[10] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*, 2014.

[11] I. Higgins, I. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[12] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *arXiv preprint arXiv:1804.03599*, 2018.

[13] H. Kim and A. Mnih, "Disentangling by factorising," in *ICML*, 2018.

[14] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *NeurIPS*, 2018.

[15] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," in *ICLR*, 2017.

[16] R. De Vito, R. Bellio, L. Trippa, and G. Parmigiani, "Multi-study factor analysis," *arXiv:1611.06350v3*, 2018.

[17] ——, "Bayesian multi-study factor analysis for high-throughput biological data," *arXiv preprint arXiv:1806.09896*, 2018.

[18] K. A. Severson, S. Ghosh, and K. Ng, "Unsupervised learning with contrastive latent variable models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4862–4869.

[19] A. Abid and J. Y. Zou, "Contrastive variational autoencoder enhances salient features," *CoRR*, vol. abs/1902.04601, 2019. [Online]. Available: http://arxiv.org/abs/1902.04601

[20] S. K. Ainsworth, N. J. Foti, A. K. C. Lee, and E. B. Fox, "oi-vae: Output interpretable vaes for nonlinear group factor analysis," in *ICML*, 2018.

[21] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *NIPS*, 2017.

[22] Y. Li and S. Mandt, "Disentangled sequential autoencoder," in *ICML*, 2018.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *IEEE*, 1998, pp. 2278–2324.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[25] A. Abid and J. Zou, "Exploring patterns enriched in a dataset with contrastive principal component analysis," *Nature Communications*, vol. 9, p. 2134, 2018.

[26] T. Menezes and C. Roth, "Natural scales in geographical patterns," *CoRR*, vol. abs/1704.01036, 2017. [Online]. Available: http://arxiv.org/abs/1704.01036

[27] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[28] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

Appendices

A Details for the Mutual Information Based Objective

Let $I_q(\mathbf{x}_{-k}; \tilde{\mathbf{t}}_k)$ be the mutual information between a set $\mathbf{x}_{-k}$ and non-corresponding latent vector $\tilde{\mathbf{t}}_k$. Then we have,

$$
\begin{aligned}
I_q(\mathbf{x}_{-k}; \tilde{\mathbf{t}}_k) =& - \mathbb{E}_{q_{\phi_k}(\mathbf{x}_{-k}, \tilde{\mathbf{t}}_k)} \log \frac{q_{\phi_k}(\tilde{\mathbf{t}}_k)}{q_{\phi_k}(\tilde{\mathbf{t}}_k|\mathbf{x}_{-k})} \\
=& - \mathbb{E}_{q_{\phi_k}(\mathbf{x}_{-k}, \tilde{\mathbf{t}}_k)} \log \frac{q_{\phi_k}(\tilde{\mathbf{t}}_k) p(\tilde{\mathbf{t}}_k)}{q_{\phi_k}(\tilde{\mathbf{t}}_k|\mathbf{x}_{-k}) p(\tilde{\mathbf{t}}_k)} \\
=& - \mathbb{E}_{p_D(\mathbf{x}_{-k})} \mathbb{E}_{q_{\phi_k}(\tilde{\mathbf{t}}_k|\mathbf{x}_k)} \log \frac{p(\tilde{\mathbf{t}}_k)}{q_{\phi_k}(\tilde{\mathbf{t}}_k|\mathbf{x}_k)} - \mathbb{E}_{q_{\phi_k}(\tilde{\mathbf{t}}_k)} \log \frac{q_{\phi_k}(\tilde{\mathbf{t}}_k)}{p(\tilde{\mathbf{t}}_k)} \\
=& \mathbb{E}_{p_D(\mathbf{x}_{-k})} D_{\mathrm{KL}} \left(q_{\phi_k}(\tilde{\mathbf{t}}_k|\mathbf{x}_{-k}) || p(\tilde{\mathbf{t}}_k) \right) - D_{\mathrm{KL}} \left(q_{\phi_k}(\tilde{\mathbf{t}}_k) || p(\tilde{\mathbf{t}}_k) \right) \\
=& \frac{1}{N - n_k} \sum_{j \neq k}^{K} \sum_{i=1}^{n_j} \left[D_{\mathrm{KL}} \left(q_{\phi_k}(\tilde{\mathbf{t}}_{ji}|\mathbf{x}_{ji}) || p(\tilde{\mathbf{t}}_{ji}) \right) \right] - D_{\mathrm{KL}} \left(q_{\phi_k}(\tilde{\mathbf{t}}_k) || p(\tilde{\mathbf{t}}_k) \right) \\
\leq& \frac{1}{N - n_k} \sum_{j \neq k}^{K} \sum_{i=1}^{n_j} \left[D_{\mathrm{KL}} \left(q_{\phi_k}(\tilde{\mathbf{t}}_{ji}|\mathbf{x}_{ji}) || p(\tilde{\mathbf{t}}_{ji}) \right) \right]
\end{aligned}
\tag{9}
$$

In some scenarios, it may desirable to increase the importance of the mutual information term $I_q(\mathbf{x}_{-k}; \tilde{\mathbf{t}}_k)$ to our objective. To this end, we can simply add a scaling constant $\alpha \geq 1$ to the final KL term in (4) like so:

$$
\begin{aligned}
J(\theta, \phi) =& \frac{1}{N} \sum_{k=1}^{K} \left[\sum_{i=1}^{n_k} \mathbb{E}_{q_\phi(\mathbf{z}_{ki}, \mathbf{t}_{ki}|\mathbf{x}_{ki})} [\log p(\mathbf{x}_{ki} \mid \mathbf{z}_{ki}, \mathbf{t}_{ki}; \theta_s, \theta_k)] - D_{\mathrm{KL}} \left(q_{\phi_s}(\mathbf{z}_{ki}|\mathbf{x}_{ki}) || p(\mathbf{z}_{ki}) \right) \right] \\
& - \alpha \sum_{k=1}^{K} \frac{1}{N - n_k} \sum_{j \neq k}^{K} \sum_{i=1}^{n_j} D_{\mathrm{KL}} \left(q_{\phi_k} \left(\tilde{\mathbf{t}}_{ji}|\mathbf{x}_{ji} \right) || p(\mathbf{t}_{ji}) \right),
\end{aligned}
\tag{10}
$$

In our experiments, we often find substantially improved results by beginning with $\alpha = 1$ and gradually annealing its value over training.

B CPVAE Training Algorithm

Algorithm 1 Training procedure of CPVAE

Initialize conditional parameters $\theta = \{\theta_s\} \cup \{\theta_k; \forall k \in K\}$;
Initialize variational parameters $\phi = \{\phi_s\} \cup \{\phi_k; \forall k \in K\}$;
repeat
  Sample mini-batch from each population $\{\{\mathbf{x}_{ki}\}_{i=1}^{M}\}_{k=1}^{K}$
  for $k \in 1 \dots K$ **do**
    Sample shared codes $\mathbf{z}_k \sim q_{\phi_s}(\mathbf{z}_k|\mathbf{x}_k)$;
    Sample private codes $\mathbf{t}_k \sim q_{\phi_k}(\mathbf{t}_k|\mathbf{x}_k)$;
    for $j \neq k \in 1 \dots K$ **do**
      Sample fictitious codes $\tilde{\mathbf{t}}_j \sim q_{\phi_k}(\tilde{\mathbf{t}}_j|\mathbf{x}_j)$;
    end
  end
  Calculate $J(\theta, \phi)$ as in (4);
  Update $\theta^{t+1}, \phi^{t+1} \leftarrow \theta^t, \phi^t$ according to ascending gradient estimate of $J(\theta, \phi)$;
until *convergence*;