Variational Autoencoders and Information Bottleneck Models

Jose Roberto Tello Ayala
Harvard University
SEAS
Cambridge, MA
jtelloayala@g.harvard.edu

Sree Harsha Tanneru
Harvard University
SEAS
Cambridge, MA
sreeharshatanneru@g.harvard.edu

***Abstract*—In mathematics and its adjacent fields, it is customary to study a phenomena through equivalent or dual representations of the problem. Analyzing a question from the lens of different scientific disciplines can deepen the knowledge of the model while also shed light towards new research directions. The problem of encoding information and data efficiently has mesmerized scientist and engineers since the past century (Shannon, 1948). As data is becoming more valuable by the day it is imperative to infer underlying structures and representations from the data and to reliably transmit it.**

Variational Autoencoders (VAEs) (Kingma and Welling, 2014) and Information Bottleneck Models (IBs) (Tishby et al., 1999) are two compression schemes which share similarities in their optimization through equivalent variational bounds (Murphy, 2023). The paper will analyze the similarities of these two schemes as well as potentially lead the way into a research question regarding both of them.

Notation: each random variable is denoted with capital letters such as X. The probability distribution of the Random Variable X is denoted as P_X and the values it can take are in lower-case letters x. The parameters of the distribution appear as conditioned notation, so if a Random Variable X has parameters θ it is indicated as $P_X(x \mid \theta)$.

I. Introduction

Suppose that for a given event there are observations from a random variable X, and assume that the generative process of the data possesses hidden features from a random variable Z. The purpose of Latent Variable Models is to approximate the true data distribution of the observed variable X through the conditional distribution of X given Z (Gelman et al., 2013). The objective is to learn the joint distribution of X and Z by proposing know parametric distributions $P_Z(z|\theta)$ and $P_{X|Z}(x \mid z, \phi)$ where θ is a parameter to be estimated such that

$$P^{\star}_{X,Z}(x,z) \approx P_{X,Z}(x,z \mid \theta) = P_{X|Z}(x \mid z,\theta)P_Z(z \mid \theta), \quad (1)$$

where $P^{\star}_{X,Z}(x,z)$ is assumed to be the true unknown distribution. One application of learning latent variables is to obtain representations that summarize the main features of the observed variable X. As $P_{Z|X}(z|x,\theta)$ encodes the distribution of the hidden features given the observed data, if Z lies in a lower dimensional space than X, the unobserved variables can abstract high level features (Bengio et al., 2013). Ideally the latent features should conserve and summarize as much information from X as possible. A natural approach in statistics would be to estimate θ through maximum log-likelihood, however this results in a non-trivial problem as:

$$\log(P_X(x \mid \theta)) = \log\left(\int_{\Omega} P_{X,Z}(x,z \mid \theta)dz\right) \quad (2)$$

$$= \log\left(\int_{\Omega} P_{X|Z}(x \mid z,\theta)P_Z(z \mid \theta)dz\right). \quad (3)$$

The choice of the likelihood $P_{X|Z}(x \mid z,\theta)$ and the prior $P_Z(z \mid \theta)$ will only be conjugate in very specific situations, if the problem is high dimensional the procedure can become computationally expensive, and the integral might not have a close form (Gelman et al., 2013). As $P_X(x \mid \theta)$ is intractable, the posterior $P_{Z|X}(z|x,\theta) = \frac{P_{X,Z}(x,z|\theta)}{P_X(x|\theta)}$ will also be troublesome to evaluate. If the goal is to obtain inferences from the latent variables for a given example, the posterior distribution $P_{Z|X}(z|x,\theta)$ would be needed. To complicate matters further, as the goal is to capture complex structures from the data and given the success of deep learning to learn sophisticated patterns (Goodfellow et al., 2016 and Murphy, 2023), it is desirable to parameterize the latent variables through neural networks.

To overcome the intractable posterior and utilize the power of neural networks, Kingma and Welling, 2014 introduced the Variational Autoencoder (VAE). To overcome the estimation problem Kingma and Welling, 2014 utilized the framework of Bayesian Variational Inference (Blei et al., 2018), a family of methods consisting on approximating intractable integrals by optimizing a functional related to the model. A lower bound to the the log-likelihood $\log(P_X(x \mid \theta))$ named the Evidence Lower Bound (ELBO) is presented as the functional to be optimized over. The purpose of introducing the ELBO is to utilize a variational distribution, $Q_{Z|X}(z|x,\phi)$, that approximates the posterior $P_{Z|X}(z|x,\theta)$ making the problem tractable to compute and thus enable the optimization over the lower bound, *i.e.*

$$Q_{Z|X}(z \mid x,\phi) = Q_{Z|X}(e_\phi(z) \mid x) \quad (4)$$

$$\approx P_{Z|X}(z \mid x,\theta) \quad (5)$$

where e_ϕ is a neural network with parameters ϕ. Given that the latent representations can be understood as codes, the

distribution $Q_{Z|X}(e_\phi(z) \mid x)$ can be though as a probabilistic encoder and $P_{X|Z}(x \mid d_\theta(z))$ will act as a decoder where d_θ is a neural network with parameters θ (Kingma and Welling, 2014. The encoding and decoding are said to be probabilistic as for a given point, the posterior outputs a probability distribution given that data point. The ELBO is given by:

$$\log(P_X(x \mid \theta)) \geq \mathbb{E}_{Q_{Z|X}}\left[\log\left(\frac{P_{X,Z}(x,z \mid \theta)}{Q_{Z|X}(z \mid x,\phi)}\right)\right] \tag{6}$$

$$= \mathbb{E}_{Q_{Z|X}}\left[\log\left(P_X(x \mid z,\theta)\right)\right] \tag{7}$$

$$- \mathbb{D}(Q_{Z|X}(z \mid x,\phi) \mid\mid P_Z(z \mid \theta)) \tag{8}$$

$$\stackrel{\text{def}}{=} \mathcal{L}_{\text{VAE}}(\theta,\phi), \tag{9}$$

where $\mathbb{D}$ is the Kullback-Leibler Divergence (Cover and Thomas, 2006). The proof of this result can be found in Appendix A. The goal is to maximize the lower bound, which can be interpreted as maximizing the expected conditional log-likelihood while minimizing the difference between the approximate posterior and the true distribution of Z. This interpretation allows us to consider the Expectation (7) as a reconstruction term, in which the objective is to maximize how well the latent variables learn from the data while the Divergence (8) acts as a regularization factor that ensures that the variational distribution minimizes it distance to the posterior (Murphy, 2023). Through this manner, Z learns a latent lower dimensional representation of the data by acting as a bottleneck between the encoding and the reconstruction.

II. Disentanglement and the Information Bottleneck Perspective

Ideally, the latent variables Z should contain the relevant features of the generative process while being uniquely identifiable. In other words, the latent representations should be disentangled and factorized from each other to make them interpretable while allowing them to abstract the most information from the data as possible (Bengio et al., 2013). Tishby et al., 1999 introduced the Information Bottleneck Model to address the question of how to maximize the information you can obtain from a signal X through a compression $\hat{X}$. The goal, in the context of Section I, is to maximize the information we have from X and Z while limiting the amount of information from the reconstruction $\hat{X}$; in other words, we want maximize the amount of information that the latent representation can obtain while forcing the latent variable to act as a minimal sufficient static for the reconstruction (Tishby et al., 1999 and Alemi et al., 2016). Let the Markov Chain be $X \to Z \to \hat{X}$, the Information Bottleneck Procedure consist in the following optimization problem:

$$\max_{P_{\hat{X}|Z}} \ \mathbb{I}(Z;X) \text{ s.t. } \mathbb{I}(Z;\hat{X}) < \varepsilon, \tag{10}$$

in Lagrangian form:

$$\max_{P_{\hat{X}|Z}} \ \mathbb{I}(Z;X) - \beta\mathbb{I}(Z;\hat{X}), \tag{11}$$

where $\mathbb{I}$ represents the mutual information (Cover and Thomas, 2006). The Data Processing Inequality (Cover and Thomas, 2006) ensures that the information obtained from the compressed variable $\hat{X}$ can not be greater than that from the latent variable Z *i.e* $\mathbb{I}(Z;X) \geq \mathbb{I}(\hat{X};X)$. This implies a *trade-off between compressed representations and preservation of meaningful information* (Goldfeld and Polyanskiy, 2020). The β parameter modulates this trade-off between the amount of quantization $\hat{X}$ provides. As β increases the quantization level becomes greater making the representation more compressed and as β decreases the representation becomes more informative. Having a constraint over $\mathbb{I}(Z;\hat{X})$ might seem counterproductive as a sufficient statistic would ensure equality $\mathbb{I}(Z;X) = \mathbb{I}(\hat{X};X)$ by capturing all the information, however that would make the representation sufficient which implies that X would be independent of $Z \mid \hat{X}$ and would not full fill the goal of compressing X form Z (Achille and Soatto, 2018) The bottleneck is generated as the process of passing on the information that Z encodes about X goes through a constriction through the compressed $\hat{X}$ (Goldfeld and Polyanskiy, 2020).

III. β-VAE

Although the VAE model has proven to produce remarkable compression (Kingma and Welling, 2014), empirically it has proven to not produce disentangled representations in complex data sets (Higgins et al., 2017 and Voloshynovskiy et al., 2019). Expanding the VAE framework, Higgins et al., 2017 proposed the β-VAE model. The goal remains the same as of that of Section I. The objective is to learn the joint distribution $P_{X,Z}$ such that Z can conditionally generate the observed data X, while adding a constraint to how similar is the approximate posterior to the prior. As stated by Burgess et al., 2018 the goal of imposing a constraint on how similar can the approximate posterior be *constricts the effective encoding capacity of the latent bottleneck and encourages the latent representation to be more factorized*. The optimization problem proposed is:

$$\max_{\theta,\phi} \ \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(P_{X|Z}(x \mid z,\theta)\right)\right] \tag{12}$$

$$\text{s.t.} \ \ \mathbb{D}(Q_{Z|X}(z \mid x,\phi) \mid\mid P_Z(z \mid \theta)) < \varepsilon. \tag{13}$$

With the Lagrangian to the problem it follows that

$$\mathcal{L}(\theta,\phi,\beta) = \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(P_{X|Z}(x \mid z,\theta)\right)\right] \tag{14}$$

$$- \beta(\mathbb{D}(Q_{Z|X}(z \mid x,\phi) \mid\mid P_Z(z \mid \theta)) - \varepsilon) \tag{15}$$

$$\geq \mathbb{E}_{Q_\phi(z|x,\phi)}\left[\log\left(P_{X|Z}(x \mid z,\theta)\right)\right] \tag{16}$$

$$- \beta\mathbb{D}(Q_{Z|X}(z \mid x,\phi) \mid\mid P_Z(z \mid \theta)) \tag{17}$$

$$\stackrel{\text{def}}{=} \mathcal{L}_{\beta-\text{VAE}}(\theta,\phi,\beta) \tag{18}$$

where the inequality follows from the fact that $\beta\varepsilon \geq 0$ due to the complementary slackness of the Karsuh-Kunt-Tucker conditions (Boyd and Vandenberghe, 2004). The Lower Bound $\mathcal{L}_{\beta-\text{VAE}}(\theta,\phi,\beta)$ (18) is the optimization problem to be solved, reminiscent of the ELBO objective $\mathcal{L}_{\text{VAE}}(\theta,\phi)$ (9). β-VAE acts as a generalization of VAE, as for $\beta = 1$ the ELBO (9) is recovered from $\mathcal{L}_{\beta-\text{VAE}}(\theta,\phi,\beta)$ (18). The β parameter was empirically motivated in it inception by Higgins et al., 2017 to work in similar fashion as that of the Information Bottleneck

Model (11). It is argued that for $\beta > 1$ the latent variables will be disentangled if they are independent. Since then, work such as the one from Mathieu et al., 2019 has gave more intuition on why the disentanglement occurs. As the approximate posterior $Q_{Z|X}(z \mid x, \phi)$ overlaps with $P_Z(z \mid \theta)$, the data points in the latent space becomes unrecognizable from each other. Mathieu et al., 2019 argue that additional structure needs to be imposed over the aggregated posterior

$$Q_Z(z \mid \phi) = \frac{1}{n} \sum_{i=1}^{n} Q_Z(z \mid x_i, \phi). \tag{19}$$

If P_X^D is the empirical distribution of the data, and proposing $Q_{X,Z}(x, z \mid \phi) = Q_{Z|X}(z \mid x, \phi) P_X^D(x)$ it follows that

$$\mathbb{E}_{P_X^D}[\mathbb{D}(Q_{Z|X}(z \mid x, \phi) \mid\mid P_Z(z \mid \theta))] \tag{20}$$

$$= \mathbb{I}_\phi(X; Z) + \mathbb{D}(Q_Z(z \mid \phi) \mid\mid P_Z(z \mid \theta)), \tag{21}$$

where $\mathbb{I}_\phi(X; Z) = \mathbb{E}_{Q_{X,Z}(x,z|\phi)} \left[\log \left(\frac{Q_{X,Z}(x,z|\phi)}{P_X^D(x) Q_Z(z|\phi)} \right) \right]$. Notice how in expectation with the data, the β parameter in the β-VAE objective $\mathcal{L}_{\beta-\text{VAE}}(\theta, \phi, \beta)$ (18) regularizes the divergence term (21) by increasing or decreasing the mutual information between X and Z and the distance between the aggregated posterior and the prior distribution. The representations become untangled as β increases, similarly as in the Information Bottleneck, given that the maximization forces to extract as much information while controlling the average level of overlap between the aggregated posterior and the prior. This is specially true if the prior distribution is chosen to factorize over each latent variable, *i.e.* $P_{Z^n}(z^n \mid \theta) = \prod P_{Z_i}(z_i \mid \theta)$, as it would encourage the aggregated posterior to follow the same behavior as the prior.

IV. Deep Variational Information Bottleneck and β-VAE

The Information Bottleneck framework introduced in Section II is unfeasible to use in modern applications as computing the mutual information for arbitrary distributions is intractable (Tishby et al., 1999). In a similar vein as the VAE, Alemi et al., 2016 introduced the Deep Variational Information Bottleneck (DBI). The similarities between the DIB and the VAE are that it the DIB also uses a proposed approximate posterior distribution, the distributions are also parameterized by neural networks, and it is optimized through a variational lower bound.

The objective remains the same as in Section II, with a similar setting as that of the end of Section III. To reconstruct X from Z, $Q_{Z|X}(z|x, \phi)$ is proposed as an approximate posterior to $P_{Z|X}(z|x, \theta)$ and given the empirical data distribution $P_X^D(x)$ then $Q_{X,Z}(x, z \mid \phi) = Q_{Z|X}(z|x, \phi) P_X^D(X)$. The optimization objective becomes:

$$\max_{\phi} \quad \mathbb{I}_\phi(Z; X) - \beta \mathbb{I}_\phi(Z; \hat{X}), \tag{22}$$

where $\mathbb{I}_\phi$ is taken as in Section III, *i.e.*

$$\mathbb{I}_\phi(X; Z) = \mathbb{E}_{Q_{X,Z}(x,z|\phi)} \left[\log \left(\frac{Q_{X,Z}(x, z \mid \phi)}{P_X^D(x) Q_Z(z \mid \phi)} \right) \right] \tag{23}$$

as proposed by Voloshynovskiy et al., 2019. The Objective (22) still remains intractable due to how hard is to compute mutual information. The following lower bound is proposed to alleviate the problem and optimize through it. Coincidentally the lower bound shows a deeper relationship between β-VAE and DIB:

$$\mathbb{I}(Z; X) - \beta \mathbb{I}(Z; \hat{X}) = \mathbb{E}_{Q_{Z|X}(z|x,\phi)} \left[\frac{Q_{Z|X}(z|x, \phi)}{Q_\phi(z)} \right] \tag{24}$$

$$- \beta \mathbb{E}_{Q_\phi(\hat{y}, z)} \left[\frac{Q_{Z|X}(z|\hat{x}, \phi)}{Q_\phi(z)} \right] \tag{25}$$

$$\geq \mathbb{E}_{P_X^D(x)}[\mathcal{L}_{\beta-\text{VAE}}(\theta, \phi, \beta)] + \mathbb{H}(X) \tag{26}$$

The above shows that β-VAE is an instance of the DIB with the entropy of the data included as an irreducible term which doesn't affect the optimization. The proof of this results can be found in AppendixB.

V. Experiments and Practical Implementation

A. Data

β-VAE and Deep Variational Information Bottleneck were trained on a synthetic toy dataset sampled from four isotropic 2-D gaussians with means $[-1, -1], [-1, 1], [1, 1], [1, -1]$. We observe how the latent space changes with varying β, and also study how individual terms at convergence in objective function change with β.

B. The Reparametrization Trick

PyTorch was used to implement both the models, and leverage the reparametrization trick to backpropagate through the sampling function. The encoder network outputs parameters for a probability distribution over the latent variables, which are typically assumed to be continuous and normally distributed. Let's say the encoder network outputs two parameters, μ and σ, which represent the mean and standard deviation of the latent space distribution, respectively. The latent variables are then typically sampled from this distribution, and passed through the decoder to reconstruct the observed variable.

This sampling operation is not differentiable, which means we cannot use gradient-based methods to optimize the encoder network's parameters. To make the model differentiable, we can reparameterize the distribution in terms of a different set of parameters that are easy to sample from and are differentiable with respect to the original parameters. For example, we can reparameterize the distribution in terms of the standard deviation and a random noise variable as follows

$$z = \mu + \Sigma^{1/2} \cdot \epsilon$$

where μ is the mean vector, Σ the covarianve matrix, and ϵ is still a random noise variable that is sampled from a standard normal distribution. This reparameterization allows us to backpropagate gradients through the sampling operation

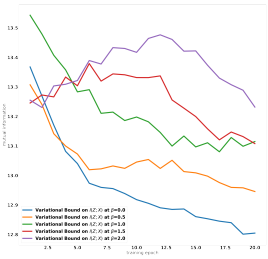
and optimize the encoder network's parameters. In practice, the reparameterization trick is implemented by sampling ϵ from a standard normal distribution and then passing it through a differentiable function, such as the inverse of the cumulative distribution function (CDF) of the desired distribution. This allows us to sample from any distribution that has a differentiable CDF, such as the normal, uniform, or logistic distributions.

C. Latent space plots for β-VAE and DVIB

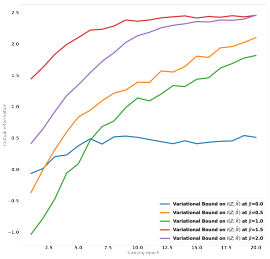
From the mathematical formulation, at lower β values, it is expected for the model to have more informative representations, and at higher β values, it is expected for to learn more compressed and disentangled representations. This is consistent with the results obtained from training latent space representations using DVIB and β-VAE. As β increases, the latent space is more distributed and has less correlation, an indicator of more disentanglement. The 2-D latent space representations of both models are shown in figures 1, and 2.

D. Loss terms

In Information Bottleneck method, β parameter signifies the trade-off between sufficiency and minimality of representations. As β increases, the theory would indicate for the representation to be more compressed, and as β decreases the representation becomes more informative. $\mathbb{I}(Z;X)$ is a good signal of how informative representations are, with high values of $\mathbb{I}(Z;X)$ indicating more informative representations. From the training results, it was noted that $\mathbb{I}(Z;X)$ increases with β, contrary to our hypothesis.



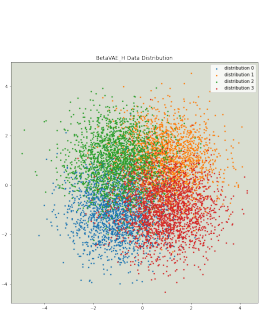
(a) $\mathbb{I}(Z;X)$ at varying β with training epoch



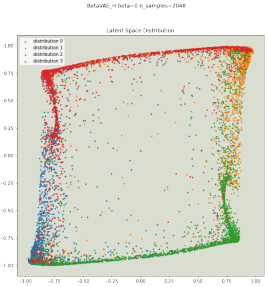
(b) $\mathbb{I}(Z;\hat{X})$ at varying β with training epoch

Fig. 3: Mutual Information terms in Deep Variational Information Bottleneck

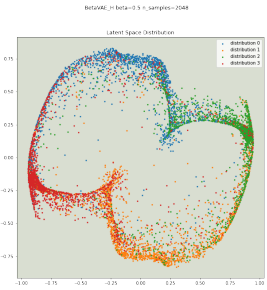
The β parameter in the β-VAE regularizes the divergence term by increasing or decreasing the mutual information between X and Z and the distance between the aggregated posterior and the prior distribution. As β increases, we expect the divergence term to decrease, and reconstruction loss to increase. We observe that the results in Figure [4] from training are consistent with the above hypothesis.



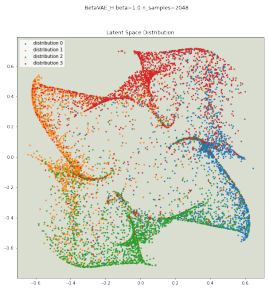
(a) Synthetic training data



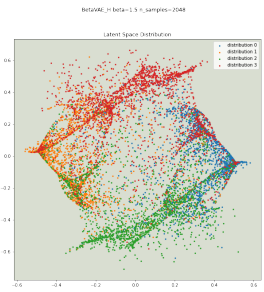
(b) Latent space of β-VAE at $\beta = 0$



(c) Latent space of β-VAE at $\beta = 0.5$



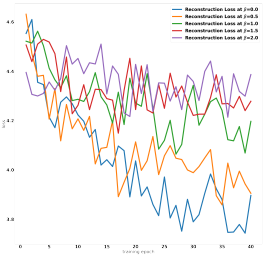
(d) Latent space of β-VAE at $\beta = 1.0$



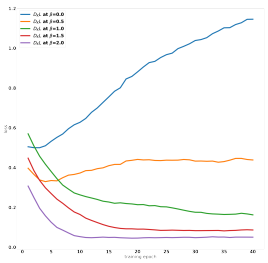
(e) Latent space of β-VAE at $\beta = 1.5$

(f) Latent space of β-VAE at $\beta = 2.0$

Fig. 1: β-VAE

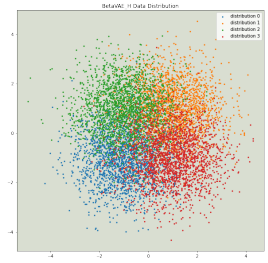


(a) Reconstruction loss at varying β with training epoch

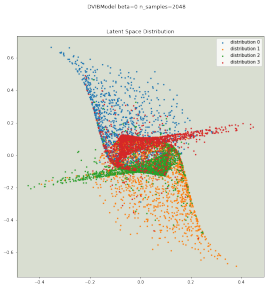


(b) $\mathbb{D}_{\mathbb{KL}}$ at varying β with training epoch

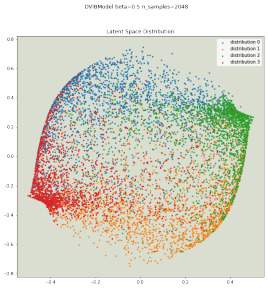
Fig. 4: Loss terms in β-VAE



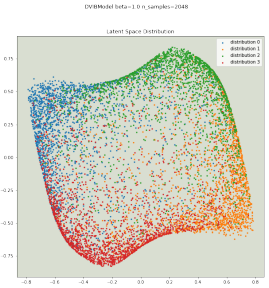
(a) Synthetic training data



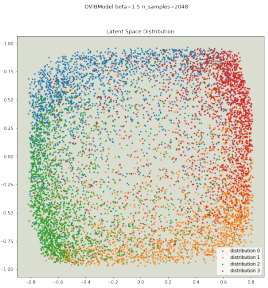
(b) Latent space of DIB at $\beta = 0$



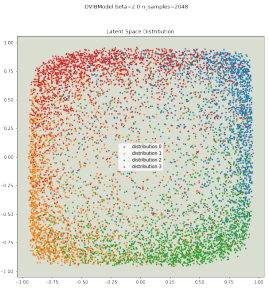
(c) Latent space of DIB at $\beta = 0.5$



(d) Latent space of DIB at $\beta = 1.0$



(e) Latent space of DIB at $\beta = 1.5$



(f) Latent space of DIB at $\beta = 2.0$

Fig. 2: Deep Variational Information Bottleneck

β	β-VAE	DVIB
0	0.5631	0.1267
0.5	0.72	0.6598
1.0	0.4628	0.7027
1.5	0.3502	0.7048
2.0	0.3566	0.7113

TABLE I: Classification Accuracy of β-VAE and Deep Variational Information Bottleneck on MNIST data

E. Classification Task Performace

Measuring downstream task performance is a good indicator of the amount of information in latent representations. We train β-VAE and DVIB at different β values on MNIST data, and train a linear ridge classifier on latent space representations ($Z_{dim} = 10$) to classify an image into one of the ten digits. The accuracy on test set is shown in Table I.

We notice that classification accuracy increases with β for Deep Variational Information Bottleneck, and decreases with β for β-VAE.

VI. Conclusion

As outlined in this work, the similarities between Deep Variational Information Bottlenecks and β-VAEs go beyond the resemblance of their optimization objective. In essence, both models attempt to learn the relevant features of the data by abstracting the most information possible while condensing it into a lower dimensional setting. Considerable progress has been done on understanding disentangling in representation learning but there is still work to be done to fully understand why current models find interpretable latent representations and in the design of more models that disentangle. Through finding the differences and similarities of the current existing techniques, the research community can strive to find the characteristic of what models need to fully abstract human known concepts.

References

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423. Retrieved April 22, 2003, from http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 368–377. https://arxiv.org/abs/physics/0004057

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory 2nd edition (wiley series in telecommunications and signal processing)*. Wiley-Interscience.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives [cite arxiv:1206.5538]. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828. http://arxiv.org/abs/1206.5538

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis*. CRC Press. https://books.google.com/books?id=eSHSBQAAQBAJ

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *CoRR*, *abs/1312.6114*.

Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2016). Deep variational information bottleneck [cite arxiv:1612.00410Comment: 19 pages, 8 figures, Accepted to ICLR17]. http://arxiv.org/abs/1612.00410

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., & Lerchner, A. (2017). Beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*.

Achille, A., & Soatto, S. (2018). Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Res.*, *19*(1), 1947–1980.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2018). Variational inference: A review for statisticians. https://arxiv.org/abs/1601.00670

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2018). Understanding disentangling in β-vae. *ArXiv*, *abs/1804.03599*.

Mathieu, E., Rainforth, T., Siddharth, N., & Teh, Y. W. (2019). Disentangling disentanglement in variational autoencoders. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Icml* (pp. 4402–4412). PMLR. http://dblp.uni-trier.de/db/conf/icml/icml2019.html#MathieuRST19

Voloshynovskiy, S., Kondah, M., Rezaeifar, S., Taran, O., Holotyak, T., & Rezende, D. J. (2019). Information bottleneck through variational glasses. *ArXiv*, *abs/1912.00830*.

Goldfeld, Z., & Polyanskiy, Y. (2020). The information bottleneck problem and its applications in machine learning. *IEEE J. Sel. Areas Inf. Theory*, *1*(1), 19–38. https://doi.org/10.1109/jsait.2020.2991561

Mushom, L. (2020). *Disentangled representations in variational autoencoders* (Master's thesis). Norwegian University of Science and Technology.

Murphy, K. P. (2023). *Probabilistic machine learning: Advanced topics*. MIT Press. http://probml.github.io/book2

APPENDIX

A. ELBO derivation

Consider the setting from Section I. The computation of the Evidence Lower Bound goes as follows:

$$\log(P_X(x \mid \theta)) = \log\left(\int_\Omega P_{X,Z}(x, z \mid \theta) dz\right) \tag{27}$$

$$= \log\left(\int P_{X,Z}(x, z \mid \theta) \frac{Q_{Z|X}(z \mid x, \phi)}{Q_{Z|X}(z \mid x, \phi)} dz\right) \tag{28}$$

$$= \log\left(\int Q_{Z|X}(z \mid x, \phi) \frac{P_{X,Z}(x, z \mid \theta)}{Q_{Z|X}(z \mid x, \phi)} dz\right) \tag{29}$$

$$= \log\left(\mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\frac{P_{X,Z}(x, z \mid \theta)}{Q_{Z|X}(z \mid x, \phi)}\right]\right) \tag{30}$$

$$\geq \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(\frac{P_{X,Z}(x, z \mid \theta)}{Q_{Z|X}(z \mid x, \phi)}\right)\right] \tag{31}$$

$$= \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(\frac{P_{X|Z}(x \mid z, \theta) P_Z(z \mid \theta)}{Q_{Z|X}(z \mid x, \phi)}\right)\right] \tag{32}$$

$$= \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(P_{X|Z}(x \mid z, \theta)\right) + \log\left(\frac{P_Z(z \mid \theta)}{Q_{Z|X}(z \mid x, \phi)}\right)\right] \tag{33}$$

$$= \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(P_{X|Z}(x \mid z, \theta)\right)\right] + \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(\frac{P_Z(z \mid \theta)}{Q_{Z|X}(z \mid x, \phi)}\right)\right] \tag{34}$$

$$= \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(P_{X|Z}(x \mid z, \theta)\right)\right] - \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(\frac{Q_{Z|X}(z \mid x, \phi)}{P_Z(z \mid \theta)}\right)\right] \tag{35}$$

$$= \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(P_{X|Z}(x \mid z, \theta)\right)\right] - \mathbb{D}(Q_{Z|X}(z \mid x, \phi) \mid\mid P_Z(z \mid \theta)), \tag{36}$$

where in Line (31) Jensen's inequality is used. To observe how tight the bound the Divergence between the approximate posterior is:

$$\mathbb{D}(Q_{Z|X}(z \mid x, \phi) \mid\mid P_{Z|X}(z \mid x, \theta)) = \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(\frac{Q_{Z|X}(z \mid x, \phi)}{P_{Z|X}(z \mid x, \theta)}\right)\right] \tag{37}$$

$$= \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(\frac{Q_{Z|X}(z \mid x, \phi)}{P_{Z|X}(z \mid x, \theta)} \frac{P_{X,Z}(x, z \mid \theta)}{P_{X,Z}(x, z \mid \theta)}\right)\right] \tag{38}$$

$$= \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(\frac{Q_{Z|X}(z \mid x, \phi)}{P_{Z|X}(z \mid x, \theta)} \frac{P_{Z|X}(z \mid x, \theta) P_X(x \mid \theta)}{P_{X,Z}(x, z \mid \theta)}\right)\right] \tag{39}$$

$$= \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(P_X(x \mid \theta) \frac{Q_{Z|X}(z \mid x, \phi)}{p_\theta(x, z)}\right)\right] \tag{40}$$

$$= \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(P_X(x \mid \theta)\right) - \log\left(\frac{P_{X,Z}(x, z \mid \theta)}{Q_{Z|X}(z \mid x, \phi))}\right)\right] \tag{41}$$

$$= \log\left(P_X(x \mid \theta)\right) - \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(\frac{P_{X,Z}(x, z \mid \theta)}{Q_{Z|X}(z \mid x, \phi)}\right)\right]. \tag{42}$$

And then,

$$\log(P_X(x \mid \theta)) = \mathbb{D}(Q_{Z|X}(z \mid x, \phi) \mid\mid P_{Z|X}(z \mid x, \theta)) + \mathbb{E}_{Q_{Z|X}(z|x,\phi)}\left[\log\left(\frac{P_{X,Z}(x, z \mid \theta)}{Q_{Z|X}(z \mid x, \phi)}\right)\right]$$

which allows us to know how tight the bound is. Whenever the approximate posterior is equals to the posterior, equality is attained.

B. Variational bound derivation for DIB

The proof was inspired by the work of Mushom, 2020. The details were full filled by the authors of this paper.

Let

$$\mathbb{I}_\phi(X; Z) = \mathbb{E}_{Q_{X,Z}(x,z|\phi)}\left[\log\left(\frac{Q_{X,Z}(x, z \mid \phi)}{P_X^D(x) Q_Z(z \mid \phi)}\right)\right] \tag{43}$$

and the objective is to find a lower bound to

$$\mathbb{I}_\phi(Z; X) - \beta \mathbb{I}_\phi(Z; \hat{X}). \tag{44}$$

As Equation (44) consists of two terms, the first quantity has to be upper bounded while the negative parameter lower bounded. The first upper bound is derived as:

$$\begin{aligned}
\mathbb{I}_{\phi}(Z;X) &= \int\int Q_{X,Z}(x,z \mid \phi) \log\left(\frac{Q_{X,Z}(x,z \mid \phi)}{P_X^D(x)Q_Z(z \mid \phi)}\right) dxdz \\
&= \int\int Q_{X,Z}(x,z \mid \phi) \log\left(\frac{Q_{X|Z}(x \mid z,\phi)Q_Z(z \mid \phi)}{P_X^D(x)Q_Z(z \mid \phi)}\right) dxdz \\
&= \int\int Q_{X,Z}(x,z \mid \phi) \log\left(\frac{Q_{X|Z}(x \mid z,\phi)}{P_X^D(x)}\right) dxdz \\
&= \int\int Q_{X|Z}(x \mid z,\phi)Q_Z(z \mid \phi) \log\left(\frac{Q_{X|Z}(x \mid z,\phi)}{P_X^D(x)}\right) dxdz \\
&= \int\int Q_Z(z \mid \phi)\left(Q_{X|Z}(x \mid z,\phi) \log\left(\frac{Q_{X|Z}(x \mid z,\phi)}{P_X^D(x)}\right)\right) dxdz \\
&= \int\int Q_Z(z \mid \phi)\left(Q_{X|Z}(x \mid z,\phi) \log\left(Q_{X|Z}(x \mid z,\phi)\right) - Q_{X|Z}(x \mid z,\phi) \log\left(P_X^D(x)\right)\right) dxdz \\
&= \int Q_Z(z \mid \phi)\left(\int Q_{X|Z}(x \mid z,\phi) \log\left(Q_{X|Z}(x \mid z,\phi)\right) dx - \int Q_{X|Z}(x \mid z,\phi) \log\left(P_X^D(x)\right) dx\right) dz,
\end{aligned}$$

as the Divergence is always positive by considering

$$\mathbb{D}(Q_{X|Z}(x \mid z,\phi) \mid\mid P_{X|Z}(x \mid z,\theta)) \geq 0 \tag{45}$$

it follow that

$$\int Q_{X|Z}(x \mid z,\phi) \log\left(Q_{X|Z}(x \mid z,\phi)\right) dx \geq \int Q_{X|Z}(x \mid z,\phi) \log\left(P_{X|Z}(x \mid z,\theta)\right) dx. \tag{46}$$

This in conjunction with the fact that $Q_{X,Z}(x,z \mid \phi) = Q_{Z|X}(z|x,\phi)P_X^D(X) = Q_Z(z \mid \phi)Q_{X|Z}(x \mid z,\phi)$ results in

$$\begin{aligned}
\mathbb{I}(Z;X)_{\phi} &= \int Q_Z(z \mid \phi)\left(\int Q_{X|Z}(x \mid z,\phi) \log\left(Q_{X|Z}(x \mid z,\phi)\right) dx - \int Q_{X|Z}(x \mid z,\phi) \log\left(P_X^D(x)\right) dx\right) dz && (47)\\
&\geq \int Q_Z(z \mid \phi)\left(\int Q_{X|Z}(x \mid z,\phi) \log\left(P_{X|Z}(x \mid z,\theta)\right) dx - \int Q_{X|Z}(x \mid z,\phi) \log\left(P_X^D(x)\right) dx\right) dz && (48)\\
&= \int\left(\int Q_Z(z \mid \phi)Q_{X|Z}(x \mid z,\phi) \log\left(P_{X|Z}(x \mid z,\theta)\right) dx - \int Q_Z(z \mid \phi)Q_{X|Z}(x \mid z,\phi) \log\left(P_X^D(x)\right) dx\right) dz && (49)\\
&= \int\left(\int P_X^D(x)Q_{Z|X}(z \mid x,\phi) \log\left(P_{X|Z}(x \mid z,\theta)\right) dx - \int P_X^D(x)Q_{Z|X}(z \mid x,\phi) \log\left(P_X^D(x)\right) dx\right) dz && (50)\\
&= \int\left(\int P_X^D(x)Q_{Z|X}(z \mid x,\phi) \log\left(P_{X|Z}(x \mid z,\theta)\right) dz - \int P_X^D(x)Q_{Z|X}(z \mid x,\phi) \log\left(P_X^D(x)\right) dz\right) dx && (51)\\
&= \int P_X^D(x)\left(\int Q_{Z|X}(z \mid x,\phi) \log\left(P_{X|Z}(x \mid z,\theta)\right) dz - \int Q_{Z|X}(z \mid x,\phi) \log\left(P_X^D(x)\right) dz\right) dx && (52)\\
&= \int P_X^D(x)\left(\int Q_{Z|X}(z \mid x,\phi) \log\left(P_{X|Z}(x \mid z,\theta)\right) dz - \log\left(P_X^D(x)\right)\right) dx && (53)\\
&= \int P_X^D(x) \int Q_{Z|X}(z \mid x,\phi) \log\left(P_{X|Z}(x \mid z,\theta)\right) dz - \int P_X^D(x) \log\left(P_X^D(x)\right) dx && (54)\\
&= \mathbb{E}_{P_X^D(x)}\left[\mathbb{E}_{Q_{Z|X}(z|x,\phi)}[\log\left(P_{X|Z}(x \mid z,\theta)\right)]\right] + \mathbb{H}(X). && (55)
\end{aligned}$$

It has been obtained that

$$\mathbb{I}_{\phi}(Z;X) \geq \mathbb{E}_{P_X^D(x)}\left[\mathbb{E}_{Q_{Z|X}(z|x,\phi)}[\log\left(P_{X|Z}(x \mid z,\theta)\right)]\right] + \mathbb{H}(X). \tag{56}$$

The upper bound for $\mathbb{I}_\phi(Z;\hat{X})$ is obtained similarly using the Inequality (46):

$$\mathbb{I}_\phi(Z;X) = \int\int Q_{X,Z}(x,z \mid \phi) \log\left(\frac{Q_{X,Z}(x,z \mid \phi)}{P_X^D(x)Q_Z(z \mid \phi)}\right) dxdz \tag{57}$$

$$= \int\int Q_{X,Z}(x,z \mid \phi) \log\left(\frac{Q_{Z|X}(z|x,\phi)P_X^D(X)}{P_X^D(x)Q_Z(z \mid \phi)}\right) dxdz \tag{58}$$

$$= \int\int Q_{X,Z}(x,z \mid \phi) \log\left(\frac{Q_{Z|X}(z|x,\phi)}{Q_Z(z \mid \phi)}\right) dxdz \tag{59}$$

$$= \int\int Q_{X,Z}(x,z \mid \phi) \log\left(\frac{Q_{Z|X}(z|x,\phi)}{Q_Z(z \mid \phi)}\right) dxdz \tag{60}$$

$$= \int\int Q_{X,Z}(x,z \mid \phi) \log\left(Q_{Z|X}(z|x,\phi)\right) - Q_{X,Z}(x,z \mid \phi) \log\left(Q_Z(z \mid \phi)\right) dxdz \tag{61}$$

$$= \int\int P_X^D(X)Q_{Z|X}(z|x,\phi) \log\left(Q_{Z|X}(z|x,\phi)\right) - Q_Z(z \mid \phi)Q_{X|Z}(x \mid z,\phi) \log\left(Q_Z(z \mid \phi)\right) dxdz \tag{62}$$

$$\leq \int\int P_X^D(X)Q_{Z|X}(z|x,\phi) \log\left(Q_{Z|X}(z|x,\phi)\right) - Q_Z(z \mid \phi)Q_{X|Z}(x \mid z,\phi) \log\left(P_Z(z \mid \theta)\right) dxdz \tag{63}$$

$$= \int\int P_X^D(X)Q_{Z|X}(z|x,\phi) \log\left(Q_{Z|X}(z|x,\phi)\right) - P_X^D(X)Q_{Z|X}(z|x,\phi) \log\left(P_Z(z \mid \theta)\right) dxdz \tag{64}$$

$$= \int\int P_X^D(X)Q_{Z|X}(z|x,\phi) \log\left(\frac{Q_{Z|X}(z|x,\phi)}{P_Z(z \mid \theta)}\right) dxdz \tag{65}$$

$$= \mathbb{E}_{P_X^D(y)}\left[\mathbb{D}(Q_{Z|X}(z|x,\phi) \mid\mid P_\theta(z))\right]. \tag{66}$$

Combining both both ir follows that

$$\mathbb{I}_\phi(Z;X) - \beta\mathbb{I}_\phi(Z;\hat{X}) \geq \mathbb{E}_{P_X^D(x)}\left[\mathbb{E}_{Q_{Z|X}(z|x,\phi)}[\log\left(P_{X|Z}(x \mid z,\theta)\right)]\right] + \mathbb{H}(X) - \beta\mathbb{E}_{P_X^D(y)}\left[\mathbb{D}(Q_{Z|X}(z|x,\phi) \mid\mid P_\theta(z))\right] \tag{67}$$

$$= \mathbb{E}_{P_X^D(x)}[\mathcal{L}_{\beta-\text{VAE}}(\theta,\phi,\beta)] + \mathbb{H}(X) \tag{68}$$