2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–15

Towards Hierarchical Discrete Variational Autoencoders

Valentin Liévin VALV@DTU.DK
Technical University of Denmark, Copenhagen, Denmark

Andrea Dittadi ADIT@DTU.DK
Technical University of Denmark, Copenhagen, Denmark

Lars Maaløe LM@CORTI.AI
Corti, Copenhagen, Denmark

Ole Winther OLWI@DTU.DK
Technical University of Denmark, Copenhagen, Denmark

Abstract

Variational Autoencoders (VAEs) have proven to be powerful latent variable models. However, the form of the approximate posterior can limit the expressiveness of the model. Categorical distributions are flexible and useful building blocks for example in neural memory layers. We introduce the Hierarchical Discrete Variational Autoencoder (HD-VAE): a hierarchy of variational memory layers. The Concrete/Gumbel-Softmax relaxation allows maximizing a surrogate of the Evidence Lower Bound by stochastic gradient ascent. We show that, when using a limited number of latent variables, HD-VAE outperforms the Gaussian baseline on modelling multiple binary image datasets. Training very deep HD-VAE remains a challenge due to the *relaxation bias* that is induced by the use of a surrogate objective. We introduce a formal definition and conduct a preliminary theoretical and empirical study of the bias.

1. Introduction

Unsupervised learning has proven powerful at leveraging vast amounts of raw unstructured data (Kingma et al., 2014; Radford et al., 2017; Peters et al., 2018; Devlin et al., 2018). Through unsupervised learning, latent variable models learn the explicit likelihood over an unlabeled dataset with an aim to discover hidden factors of variation as well as a generative process. An example hereof, is the Variational Autoencoder (VAE) (Kingma and Welling, 2013; Rezende et al., 2014) that exploits neural networks to perform amortized approximate inference over the latent variables. This approximation comes with limitations, both in terms of the latent prior and the amortized inference network (Burda et al., 2015; Hoffman and Johnson, 2016). It has been proposed to go beyond Gaussian priors and approximate posterior using, for instance, autoregressive flows (Chen et al., 2016; Kingma et al., 2016), a hierarchy of latent variables (Sønderby et al., 2016; Maaløe et al., 2016, 2019), a mixture of priors (Tomczak and Welling, 2017) or discrete distributions (van den Oord et al., 2017; Razavi et al., 2019; Rolfe, 2016; Vahdat et al., 2018b,a; Sadeghi et al., 2019).

Current state-of-the-art deep learning models are trained on web-scaled datasets and increasing the number of parameters has proven to be a way to yield remarkable results

(Radford et al., 2019). Nonetheless, time complexity and GPU memory are scarce resources, and the need for both resources increases linearly with the depth of neural network. Li et al. (2016) and Lample et al. (2019) showed that large memory layers are an effective way to increase the capacity of a model while reducing the computation time.

Bornschein et al. (2017) showed that discrete variational distributions are analogous to neural memory (Graves et al., 2016), which can be used to improve generative models (Li et al., 2016; Lample et al., 2019). Also, memory values are yet another way to embed data, allowing for applications such as one-shot transfer learning (Rezende et al., 2016) and semi-supervised learning that scales (Jang et al., 2016).

Depth promises to bring VAEs to the next frontier (Maaløe et al., 2019). However, the available computing resources may shorten that course. Motivated by the versatility and the scalability of discrete distributions, we introduce the Hierarchical Discrete Variational Autoencoder. HD-VAE is a VAE with a hierarchy of factorized categorical latent variables. In contrast to the existing discrete latent variable methods, our model (a) is hierarchical, (b) trained using Concrete/Gumbel-Softmax, (c) relies on a conditional prior that is learned end-to-end and (d) uses a variational distribution that is parameterized as a large stochastic memory layer. Despite being optimized for a biased surrogate objective we show that a shallow HD-VAE outperforms the baseline Gaussian-based models on multiple binary images datasets in terms of test log-likelihood. This motivates us to introduce a definition of the relaxation bias and to measure how it is affected by the configuration of latent variables.

2. Hierarchical Discrete VAE

Hierarchical VAE Hierarchical VAEs define a model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$ where $\mathbf{x}$ is an observed variable and $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_L\}$ is a hierarchy of latent variables so that $p_\theta(\mathbf{z})$ is factorized into L layers. The inference model $q_\phi(\mathbf{z}|\mathbf{x})$ usually exploits the inverse dependency structure. A vanilla hierarchical VAE results in the following model:

$$p_\theta(\mathbf{z}) = p_\theta\left(\mathbf{z}_L\right) \prod_{i=1}^{L-1} p_\theta\left(\mathbf{z}_i|\mathbf{z}_{i+1}\right) \qquad q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi\left(\mathbf{z}_1|\mathbf{x}\right) \prod_{i=2}^{L} q_\phi\left(\mathbf{z}_i|\mathbf{z}_{i-1}\right) \ . \tag{1}$$

The choice of the VAE architecture is independent of the choice of the variational family and deeper models can easily be defined (see appendix F).

Variational Neural Memory Each stochastic layer consists of N categorical random variables with K class probabilities $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$ and can be parametrized as a memory layer. Lample et al. (2019) recently proposed a scalable approach to attention-based memory layers that can be directly translated to the stochastic setting: Each categorical distribution is parametrized by factored keys $\{\mathbf{k}_1, \dots, \mathbf{k}_K\}, \mathbf{k}_i \in \mathbb{R}^{d_1}$ and a parametric query model $Q(\mathbf{h})$. If $\{\mathbf{v}_1, ..., \mathbf{v}_K\}, \mathbf{v}_i \in \mathbb{R}^{d_2}$ are the memory values, for $c \in \mathbb{R}$ and $i = 1, \dots, K$, then the output of the memory layer for one variable is

$$\mathbf{y} = \sum_{i=1}^{K} z_i \mathbf{v}_i, \qquad \mathbf{z} \sim \mathrm{Cat}(\boldsymbol{\pi}), \qquad \log \pi_i = Q(\mathbf{h})^T \mathbf{k}_i + c \ . \tag{2}$$

Optimization We wish to maximize the Evidence Lower Bound (ELBO):

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[f_{\theta,\phi}(\mathbf{x},\mathbf{z})\right] \equiv \mathcal{L}_1(\theta,\phi) \qquad f_{\theta,\phi}(\mathbf{x},\mathbf{z}) = \log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \,. \tag{3}$$

The subscript of $\mathcal{L}$ denotes the number of importance weighted samples.

Guided by the analysis of Sønderby et al. (2017), we chose to use the Concrete/Gumbel-Softmax relaxation (Jang et al., 2016; Maddison et al., 2016) for differentiable, approximate sampling of categorical variables. A relaxed categorical sample can be obtained as

$$\tilde{z}_i = \frac{\exp\left(\left(\log \pi_i + g_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\left(\log \pi_j + g_j\right)/\tau\right)} \qquad \text{for } i = 1,\ldots,K \,, \tag{4}$$

where $\{g_i\}$ are i.i.d. samples drawn from Gumbel(0,1), and $\tau \in \mathbb{R}^{*+}$ is a temperature parameter. As in the categorical case, the output of the memory layer is a convex combination of the memory values weighted by the entries of $\tilde{\mathbf{z}}$: $\mathbf{y} = \sum_{i=1}^{K} \tilde{z}_i \mathbf{v_i}$, for $i = 1,\ldots,K$. The relaxed samples $\tilde{\mathbf{z}}$ follow a Concrete/Gumbel-Softmax distribution q_ϕ^τ which depends on τ and converges to the categorical distribution $q_\phi^{\tau=0} = q_\phi$ as $\tau \to 0$ which is equivalent to applying the Gumbel-Max trick to soft samples, meaning $\mathbf{z} = H(\tilde{\mathbf{z}})$, $H = \text{one hot} \circ \arg\max$.

When we extend the definition of $f_{\theta,\phi}$ to the domain of the relaxed samples, as in appendix D, the surrogate objective that is maximized becomes

$$\mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|\mathbf{x})}\left[f_{\theta,\phi}(\mathbf{x},\tilde{\mathbf{z}})\right] \equiv \mathcal{L}_1^{\tau>0}(\theta,\phi) \tag{5}$$

which is not guaranteed to be a lower bound of $\log p_\theta(\mathbf{x})$. Hence, we are interested in the relaxation bias that we define as:

$$\delta^\tau(\theta,\phi) \equiv \left|\mathcal{L}_1^{\tau>0}(\theta,\phi) - \mathcal{L}_1^{\tau=0}(\theta,\phi)\right| \tag{6}$$

where $\mathcal{L}_1^{\tau=0}(\theta,\phi) = \mathcal{L}_1(\theta,\phi)$ is the original ELBO.

If $f_{\theta,\phi}$ is a κ-Lipschitz for $\mathbf{z}$, we can derive an upper bound for the relaxation bias as well as a new log-likelihood bound (*relaxed ELBO*) by adding a corrective term to the surrogate objective (derivation in appendix C). For a one layer Ladder Variational Autoencoder (LVAE), it results in the following bounds:

$$\delta^\tau(\theta,\phi) \leq \kappa\ \mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|\mathbf{x})}\left[\|\tilde{\mathbf{z}} - H(\tilde{\mathbf{z}}))\|_2\right], \tag{7}$$

$$\log p(\mathbf{x}) \geq \mathcal{L}_1^{\tau>0}(\theta,\phi) - \kappa\ \mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|\mathbf{x})}\left[\|\tilde{\mathbf{z}} - H(\tilde{\mathbf{z}}))\|_2\right] \,. \tag{8}$$

This new bound shows that, if the model is unconstrained, the relaxation bias is free to grow and that it grows with the number of discrete variables. In section 4.2, we provide empirical results supporting the monotonically increasing property of the relaxation bias with regards to the number of stochastic units.

Table 1: Sample estimates of the $KL(q_{\phi,\theta}(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))$ and the ELBO for 1000 importance weighted hard samples ($\tau = 0$) using the same LVAE architecture and hyperparameters across all datasets.

	L = 1				L = 2				L = 3			
	$\mathcal{L}_{1000}^{\tau=0}$		$KL^{\tau=0}$		$\mathcal{L}_{1000}^{\tau=0}$		$KL^{\tau=0}$		$\mathcal{L}_{1000}^{\tau=0}$		$KL^{\tau=0}$	
	DISCRETE	NORMAL	DISCRETE	NORMAL	DISCRETE	NORMAL	DISCRETE	NORMAL	DISCRETE	NORMAL	DISCRETE	NORMAL
BINMNIST	-87.34	-90.61	26.46	26.75	-80.52	-81.46	24.95	25.62	-79.39	-80.07	25.58	25.94
CALTECH	-103.29	-107.77	35.37	32.70	-88.64	-94.60	36.17	34.57	-85.31	-90.84	35.84	37.36
FASHION	-105.16	-111.49	30.43	28.81	-95.85	-100.09	28.98	28.29	-94.01	-95.31	29.30	29.72
OMNIGLOT	-100.38	-107.65	31.55	26.37	-95.55	-98.30	31.82	31.14	-94.19	-96.40	32.49	32.68

3. Related work

Li et al. (2016) used a deterministic memory layer as building blocks for VAEs, Bornschein et al. (2017) introduced memory as a stochastic layer. Razavi et al. (2019) introduced a hierarchy of discrete variables trained using vector-quantitization and with an autoregressive prior trained offline. Rolfe (2016); Vahdat et al. (2018b); Sadeghi et al. (2019) used Bernouilli random variables with RBM priors trained using an alternative relaxation and coupled with continuous latent variables. Alternatively, one may optimize a categorical variable model using unbiased gradient estimators such as Mnih and Rezende (2016); Tucker et al. (2017); Grathwohl et al. (2017). To the best of our knowledge, HD-VAE is the only work that attempts to transform memory layers into a general purpose variational distribution.

4. Experiments

4.1. Modelling Binary Images

We trained HD-VAE for different number of layers of latent variables using the surrogate objective defined in the equation 5. In this experiment, we observe that HD-VAE consistently outperforms the baseline Gaussian model for multiple datasets and different number of latent layers (table 1). This shows that using variational memory layers yields a more flexible model than for the VAE with a Gaussian prior and the same number of latent variables. Furthermore, optimizing latent variable models is challenging (Sønderby et al., 2016; Chen et al., 2016). In this experiment, the measured KL is higher for the discrete model, suggesting a well-tempered optimization behavior. Finally, we observe that increasing the depth of HD-VAE consistently improves on the log-likelihood, with a limit of three layers latent layers.

4.2. Relaxation Bias

The relaxation bias (section 2) may increase with the number of discrete latent variables. We trained HD-VAE for different numbers of stochastic units and different depths on Binarized MNIST using the surrogate objective defined in the equation 5. We measured the relaxation bias $\delta^{\tau=0.1}$ on the test set (figure 1, table 4). The relaxation bias monotonically increases with the total number of discrete latent variables for different numbers of latent variables.

This may explain why we found that HD-VAE with a large number of latent variables is not yet competitive with the Gaussian counterparts.

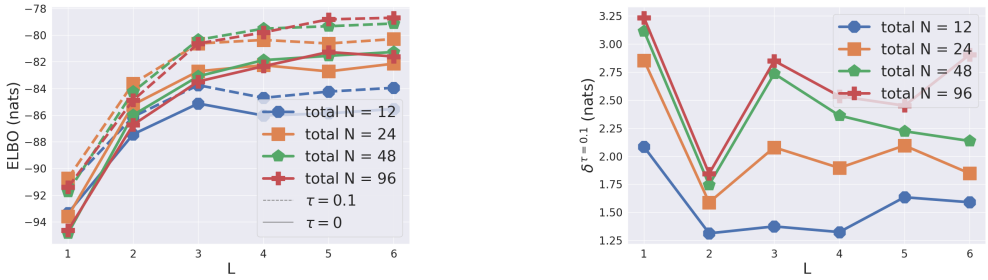


Figure 1: (left) ELBO $\mathcal{L}_1^{\tau=0}$ and relaxed objective $\mathcal{L}_1^{\tau=0.1}$ and (right) the relaxation bias $\delta^{\tau=0.1} = \left|\mathcal{L}_1^{\tau=0.1} - \mathcal{L}_1^{\tau=0}\right|$ for different HD-VAE models trained with different total number of latent variables (N), different depths (L) and K = 256 evaluated on the binarized MNIST test set. The relaxation bias grows monotonically with N. The hyperparameters search was performed for L = 2.

5. Conclusion

In this preliminary research, we have introduced a design for variational memory layers and shown that it can be exploited to build hierarchical discrete VAEs, that outperform Gaussian prior VAEs. However, without explicitly constraining the model, the relaxation bias grows with the number of latent layers, which prevents us from building deep hierarchical models that are competitive with state-of-the-art methods. In future work we will attempt to harness the *relaxed-ELBO* to improve the performance of the HD-VAE further.

References

Jörg Bornschein, Andriy Mnih, Daniel Zoran, and Danilo J Rezende. Variational memory addressing in generative models. September 2017.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. September 2015.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. November 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. October 2018.

Adji B Dieng, Yoon Kim, Alexander M Rush, and David M Blei. Avoiding latent variable collapse with generative skip models. July 2018.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. October 2017.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, October 2016.

Matthew D Hoffman and Matthew J Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound. 2016.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. November 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

Diederik P Kingma and Max Welling. Auto-Encoding variational bayes. December 2013.

Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Semi-Supervised learning with deep generative models. June 2014.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. June 2016.

Brenden M Lake, Ruslan R Salakhutdinov, and Josh Tenenbaum. One-shot learning by inverting a compositional causal process. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2526–2534. Curran Associates, Inc., 2013.

Guillaume Lample, Alexandre Sablayrolles, Marc'aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Large memory layers with product keys. July 2019.

Chongxuan Li, Jun Zhu, and Bo Zhang. Learning to generate with memory. February 2016.

Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. February 2016.

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. BIVA: A very deep hierarchy of latent variables for generative modeling. February 2019.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. November 2016.

B Marlin, K Swersky, B Chen, and N Freitas. Inductive principles for restricted boltzmann machine learning. *on Artificial Intelligence and ...*, 2010.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. February 2018.

Andriy Mnih and Danilo J Rezende. Variational inference for monte carlo objectives. February 2016.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. February 2018.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. April 2017.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse High-Fidelity images with VQ-VAE-2. June 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. January 2014.

Danilo Jimenez Rezende, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra. One-Shot generalization in deep generative models. March 2016.

Jason Tyler Rolfe. Discrete variational autoencoders. September 2016.

Hossein Sadeghi, Evgeny Andriyash, Walter Vinci, Lorenzo Buffoni, and Mohammad H Amin. PixelVAE++: Improved PixelVAE with discrete prior. August 2019.

Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks.

Tim Salimans and Diederik P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. February 2016.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. February 2016.

Casper Kaae Sønderby, Ben Poole, and Andriy Mnih. Continuous relaxation training of discrete latent variable image models. August 2017.

Jakub M Tomczak and Max Welling. VAE with a VampPrior. May 2017.

George Tucker, Andriy Mnih, Chris J Maddison, Dieterich Lawson, and Jascha Sohl-Dickstein. REBAR: Low-variance, unbiased gradient estimates for discrete latent variable models. March 2017.

Arash Vahdat, Evgeny Andriyash, and William G Macready. DVAE#: Discrete variational autoencoders with relaxed boltzmann priors. May 2018a.

Arash Vahdat, William G Macready, Zhengbing Bian, Amir Khoshaman, and Evgeny Andriyash. DVAE++: Discrete variational autoencoders with overlapping transformations. February 2018b.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. November 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. August 2017.

Appendix A. Experimental Setup

Datasets We evaluate the test log-likelihood on the HD-VAE on statically binarized MNIST (Salakhutdinov and Murray), statically binarized Omniglot (Lake et al., 2013), statically binarized Fashion MNIST (Xiao et al., 2017) and Caltech 101 Silhouettes (Marlin et al., 2010).

Gaussian Baseline and Number of Latent Variables The relaxation bias grows with the number of latent variables (section 2), hence we use a small number of stochastic units (maximum 32). Since the prior is learned, we simply use a large enough K value for each layer (architecture detailed in the table 2). The choice of the Gaussian baseline is not trivial as each discrete latent variable with K classes describes a $K-1$-simplex in the relaxed case and a set of K distinct values in the zero limit of τ. Because a categorical variable can represent at minimum a discretization of a continuous variable defined on the real line, we chose to use the same number of latent variables N for the continuous and the discrete models. One should keep in mind that the performances of the Gaussian model may increase with N while the performances of HD-VAE may reach a plateau due to the relaxation bias.

Architecture We use a Ladder Variational Autoencoder (LVAE) (Sønderby et al., 2016) with skip connections with one, two and three layers of latent variables. Each intermediate connection is parametrized by a sequence of 3 gated residual convolutional blocks with skip connection, 64 filters and weight normalization (Salimans and Kingma, 2016) similarly to Kingma et al. (2016); Maaløe et al. (2019). We denote L the number of stochastic layers, N the total number of stochastic units and K the number of class for a given categorical variable. We use factored keys (Lample et al., 2019), which results in a set of $2\sqrt{K}$ values for K effective keys. We use values and keys of size $d_1 = d_2 = 8$. Using factored keys led to substantially improved performances. For reference, HD-VAE and the Gaussian LVAE have respectively 43.1M and 16.6M parameters.

Optimization During training, we mitigate the posterior collapse using the *freebits* (Kingma et al., 2016) strategy with $\lambda = 2$ for each stochastic layer. A dropout of 0.5 is used to avoid overfitting. We linearly decrease the temperature τ from 0.8 to 0.3 during the first $2 \cdot 10^5$ steps and from 0.3 to 0.1 during the next $2 \cdot 10^5$ steps. We use the Adamax optimizer (Kingma and Ba, 2014) with initial learning rate of $2 \cdot 10^{-3}$ for all parameters except for the memory values that are trained using a learning rate of $2 \cdot 10^{-2}$ to compensate for sparsity. We use a batch size of 128. All models are trained until they overfit and we evaluate the log-likelihood using 1000 importance weighted samples (Burda et al., 2015). Despite its large number of parameters, HD-VAE seems to be more robust to overfitting, which may be explained by the sparse update of the memory values.

Runtime Sparse CUDA operations are currently not used, which means there is room to make HD-VAE more memory efficient. Even during training, one may truncate the relaxed samples to benefit from the sparse optimizations. The table 3 shows the average elapsed time training iteration as well as the memory usage for a 6 layers LVAE with 6×16 stochastic units and $K = 16^2$ and batch size of 128.

Table 2: Architectures of the HD-VAE and the baseline used in the binary image modelling experiment.

L	N	K
2	16 + 8	$16^2 + 8^2$
3	16 + 8 + 4	$16^2 + 8^2 + 4^2$
3	16 + 8 + 4 + 4	$16^2 + 8^2 + 4^2 + 4^2$
3	16 + 8 + 4 + 4 + 4	$16^2 + 8^2 + 4^2 + 4^2 + 4^2$

Table 3: Runtime performances for a 6 layer HD-VAE.

	elapsed time (seconds / iteration)	GPU memory usage (MB)
Gaussian	0.25	4723
Discrete	0.25	5245

Appendix B. Tabular Results for the Relaxation Bias Experiment

Table 4: Measured one-importance-weighted ELBO on binarized MNIST for a LVAE model with different number of layers and different numbers of stochastic units using relaxed ($\tau = 0.1$) and hard samples ($\tau = 0$). We report $N = \sum_{l=1}^{L} n_l$, where n_l relates to the number of latent variables at the layer l and we set K = 256 for all the variables.

	$N = 12$			$N = 24$			$N = 48$			$N = 98$		
L	$L_1^{\tau=0.1}$	$L_1^{\tau=0}$	$\delta^{\tau=0.1}$	$L_1^{\tau=0.1}$	$L_1^{\tau=0.1}$	$\delta^{\tau=0.1}$	$L_1^{\tau=0.1}$	$L_1^{\tau=0}$	$\delta^{\tau=0}$	$L_1^{\tau=0.1}$	$L_1^{\tau=0}$	$\delta^{\tau=0.1}$
1	-91.27	-93.35	**2.08**	-90.74	-93.59	**2.85**	-91.76	-94.87	**3.11**	-91.41	-94.65	**3.23**
2	-86.13	-87.44	**1.31**	-83.63	-85.22	**1.59**	-84.25	-85.99	**1.75**	-84.88	-86.73	**1.84**
3	-83.76	-85.13	**1.37**	-80.63	-82.71	**2.08**	-80.35	-83.09	**2.73**	-80.63	-83.48	**2.85**
4	-84.71	-86.03	**1.32**	-80.36	-82.26	**1.90**	-79.52	-81.88	**2.36**	-79.80	-82.34	**2.54**
5	-84.24	-85.87	**1.63**	-80.63	-82.72	**2.09**	-79.33	-81.56	**2.22**	-78.82	-81.27	**2.45**
6	-83.95	-85.54	**1.59**	-80.30	-82.14	**1.85**	-79.13	-81.27	**2.14**	-78.70	-81.60	**2.90**

Appendix C. Adjusted Evidence Lower Bound for relaxed categorical variables (relaxed-ELBO)

Let $\mathbf{x}$ be an observed variable, and consider a VAE model with one layer of N categorical latent variables $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ each with K classes. The generative model is $p_\theta(\mathbf{x}, \mathbf{z})$ and the inference model is $q_\phi(\mathbf{z}|\mathbf{x})$.

For a temperature parameter $\tau > 0$, the equivalent relaxed concrete variables are denoted $\hat{\mathbf{z}} = \{\hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_N\}, \hat{\mathbf{z}}_i \in [0,1]^K$. We define $H = \text{one hot} \circ \arg\max$ and

$$f_{\theta,\phi,x} : \mathbf{z} \in [0,1]^{N\times K} \to \mathbb{R} \text{ such that } f_{\theta,\phi,x}(\mathbf{z}) = f_{\theta,\phi}(\mathbf{x}, \mathbf{z}) = \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \ . \tag{9}$$

Following Tucker et al. (2017), using the Gumbel-Max trick, one can notice that

$$\mathbb{E}_{q_\phi^{\tau=0}(\mathbf{z}|\mathbf{x})}[f_{\theta,\phi,x}(\mathbf{z})] = \mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|\mathbf{x})}[f_{\theta,\phi,x}(H(\tilde{\mathbf{z}}))] \ .$$

We now assume that $f_{\theta,\phi,x}$ is κ-Lipschitz for L^2. Then, by definition,

$$\forall(\mathbf{a}, \mathbf{b}) \in ([0,1]^{N\times K})^2, \ |f_{\theta,\phi,x}(\mathbf{b}) - f_{\theta,\phi,x}(\mathbf{a})| \leq \kappa \left\|\mathbf{b} - \mathbf{a}\right\|_2 \tag{10}$$

The relaxation bias can therefore be bounded as follows:

$$\begin{aligned}
\delta^\tau(\theta,\phi) &\equiv \left|\mathcal{L}_1^{\tau>0}(\theta,\phi) - \mathcal{L}_1^{\tau=0}(\theta,\phi)\right| \\
&= \left|\mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|\mathbf{x})}\left[f_{\theta,\phi,x}(\tilde{\mathbf{z}})\right] - \mathbb{E}_{q_\phi^{\tau=0}(\mathbf{z}|\mathbf{x})}\left[f_{\theta,\phi,x}(\mathbf{z})\right]\right| \\
&= \left|\mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|\mathbf{x})}\left[f_{\theta,\phi,x}(\tilde{\mathbf{z}}) - f_{\theta,\phi,x}(H(\tilde{\mathbf{z}}))\right]\right| \\
&\leq \mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|\mathbf{x})}\left[|f_{\theta,\phi,x}(\tilde{\mathbf{z}}) - f_{\theta,\phi,x}(H(\tilde{\mathbf{z}}))|\right] \quad \text{(Jensen's Inequality)} \\
&\leq \kappa \ \mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|\mathbf{x})}\left[\|\tilde{\mathbf{z}} - H(\tilde{\mathbf{z}}))\|_2\right] \qquad (\kappa\text{-Lipschitz}).
\end{aligned} \tag{11}$$

Furthermore, we can define the adjusted Evidence Lower Bound for relaxed categorical variables (*relaxed-ELBO*):

$$\begin{aligned}
\log p(\mathbf{x}) &\geq \mathcal{L}_1^{\tau=0}(\theta,\phi) \\
&\geq \mathcal{L}_1^{\tau=0}(\theta,\phi) - \left|\mathcal{L}_1^{\tau>0}(\theta,\phi) - \mathcal{L}_1^{\tau=0}(\theta,\phi)\right| \\
&\geq \mathcal{L}_1^{\tau>0}(\theta,\phi) - \kappa \ \mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|\mathbf{x})}\left[\|\tilde{\mathbf{z}} - H(\tilde{\mathbf{z}}))\|_2\right] \ .
\end{aligned} \tag{12}$$

As shown by the experiment presented in the section 4.2, the quantity $\mathcal{L}_1^{\tau>0}(\theta,\phi) - \mathcal{L}_1^{\tau=0}(\theta,\phi)$ appears to be a positive quantity. Furthermore, as the model attempts to exploit the relaxation of $\mathbf{z}$ to maximize the surrogate objective, one may consider that $\mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|x)}\left[|f_{\theta,\phi,x}(\tilde{\mathbf{z}}) - f_{\theta,\phi,x}(H(\tilde{\mathbf{z}}))|\right]$ is a tight bound of $\delta^\tau(\theta,\phi)$, meaning that the relaxed-ELBO is a tight lower bound of the ELBO.

The relaxed-ELBO is differentiable and may enable automatic control of the temperature as left and right terms of the relaxed-ELBO seek respectively seek for high and low temperature.

κ-Lipschitz neural networks can be designed using Weight Normalization (Salimans and Kingma, 2016) or Spectral Normalization (Miyato et al., 2018). Nevertheless handling

residual connections and multiple layers of latent variables is not trivial. We note however that in the case of a one layer VAE, one only needs to constrain the VAE decoder to be κ-Lispchitz as the surrogate objective is computed as

$$\mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|x)}\left[f_{\theta,\phi,x}(\tilde{\mathbf{z}})\right] = \mathbb{E}_{q_\phi^{\tau>0}(\tilde{\mathbf{z}}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\tilde{\mathbf{z}}) + \log p_\theta(H(\tilde{\mathbf{z}})) - \log q_\phi(H(\tilde{\mathbf{z}})|\mathbf{x})\right] \ . \tag{13}$$

In the appendix E, we show how the relaxed-ELBO can be extended to multiple layers of latent variables in the LVAE setting.

Appendix D. Defining $f_{\theta,\phi}$ on the domain of the relaxed Categorical Variables $\tilde{\mathbf{z}}$

$f_{\theta,\phi}$ is only defined for categorical samples. For relaxed samples $\tilde{\mathbf{z}}$, we define $f_{\theta,\phi}$ as:

$$f_{\theta,\phi}(\mathbf{x},\tilde{\mathbf{z}}) = \underbrace{\log p_\theta(\mathbf{x}|\tilde{\mathbf{z}})}_{\textbf{(a)}} + \underbrace{\log p_\theta(H(\tilde{\mathbf{z}}))}_{\textbf{(b)}} - \underbrace{\log q_\phi(H(\tilde{\mathbf{z}})|\mathbf{x})}_{\textbf{(c)}} \ . \tag{14}$$

The introduction of the function H is necessary as the terms **(b)** and **(c)** are only defined for categorical samples. This expression remains valid for hard samples $\tilde{\mathbf{z}}$.

During training, relaxing the expressions **(b)** and **(c)** can potentially yield gradients of lower variance. In the case of a single categorical variable $\mathbf{z}$ described by the set of K class probabilities $\boldsymbol{\pi} = \{\pi_1, ...\pi_K\}$. One can define:

$$\log \tilde{p}(\tilde{\mathbf{z}} \mid \boldsymbol{\pi}) \equiv \sum_{i=1}^{K} \tilde{z}_i \log \pi_i \ . \tag{15}$$

Alternatively, asides from being a relaxed Categorical distribution, the Concrete/Gumbel-Softmax also defines a proper continuous distribution. When treated as such, this results in a proper probabilistic model with continuous latent variables, and the objective is unbiased. In that case, the density is given by

$$p(\tilde{\mathbf{z}}|\boldsymbol{\pi}) = (K-1)!\tau^{K-1}\prod_{i=1}^{K}\left(\frac{\pi_i(\tilde{z}_i)^{-\tau-1}}{\sum_{j=1}^{K}\pi_j(\tilde{z}_j)^{-\tau}}\right) \ . \tag{16}$$

Appendix E. The relaxed-ELBO for Ladder Variational Autoencoders

We consider now an LVAE model:

$$p_\theta(\mathbf{z}) = p_\theta\left(\mathbf{z}_L\right) \prod_{i=1}^{L-1} p_\theta\left(\mathbf{z}_i|\mathbf{z}_{i+1}\right) \qquad q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi\left(\mathbf{z}_L|\mathbf{x}\right) \prod_{i=1}^{L-1} q_{\phi,\theta}\left(\mathbf{z}_i|\mathbf{z}_{i+1},\mathbf{x}\right)$$

In the following, we will leave the conditioning on $\mathbf{x}$ implicit for convenience. The ELBO estimated with relaxed samples (relaxed-ELBO) is:

$$\mathcal{L}_1^{\tau>0}(\theta,\phi) = \mathbb{E}_{q^{\tau>0}(\tilde{\mathbf{z}})}\left[\log\frac{p(H(\tilde{\mathbf{z}}_L))}{q(H(\tilde{\mathbf{z}}_L))} + \sum_{i=1}^{L-1}\log\frac{p(H(\tilde{\mathbf{z}}_i)|\tilde{\mathbf{z}}_{i+1})}{q(H(\tilde{\mathbf{z}}_i)|\tilde{\mathbf{z}}_{i+1})} + \log p(\mathbf{x}|\tilde{\mathbf{z}}_1)\right]$$

The correct ELBO can be rewritten as follows:

$$\begin{aligned}\mathcal{L}_1^{\tau=0}(\theta,\phi) &= \mathbb{E}_{q^{\tau=0}(\mathbf{z})}\left[\log\frac{p(\mathbf{z}_L)}{q(\mathbf{z}_L)} + \sum_{i=1}^{L-1}\log\frac{p(\mathbf{z}_i|\mathbf{z}_{i+1})}{q(\mathbf{z}_i|\mathbf{z}_{i+1})} + \log p(\mathbf{x}|\mathbf{z})\right] \\ &= \mathbb{E}_{q^{\tau>0}(\tilde{\mathbf{z}})}\left[\log\frac{p(H(\tilde{\mathbf{z}}_L))}{q(H(\tilde{\mathbf{z}}_L))} + \sum_{i=1}^{L-1}\log\frac{p(H(\tilde{\mathbf{z}}_i)|H(\tilde{\mathbf{z}}_{i+1}))}{q(H(\tilde{\mathbf{z}}_i)|H(\tilde{\mathbf{z}}_{i+1}))} + \log p(\mathbf{x}|H(\tilde{\mathbf{z}}_1))\right]\end{aligned}$$

Let us define a shorthand notation for the functions involved here:

$$\begin{aligned}f_0(\mathbf{t}) &= f_0(\mathbf{t};\mathbf{x}) = \log p(\mathbf{x}|\mathbf{t}) \\ f_i^p(\mathbf{t}) &= f_i^p(\mathbf{t};\tilde{\mathbf{z}}_i) = \log p(H(\tilde{\mathbf{z}}_i)|\mathbf{t}) \\ f_i^q(\mathbf{t}) &= f_i^q(\mathbf{t};\tilde{\mathbf{z}}_i) = \log q(H(\tilde{\mathbf{z}}_i)|\mathbf{t})\end{aligned}$$

for $i = 1,\dots,L-1$. We assume that all these functions are Lipschitz functions with constants κ_0, κ_i^p, κ_i^q. The relaxation bias for an LVAE can be bounded as follows:

$$\begin{aligned}\delta^\tau(\theta,\phi) &= \left|\mathbb{E}_{q^{\tau>0}(\tilde{\mathbf{z}})}\left[\sum_{i=1}^{L-1}\left(f_i^p(\tilde{\mathbf{z}}_{i+1}) - f_i^p(H(\tilde{\mathbf{z}}_{i+1})) + f_i^q(H(\tilde{\mathbf{z}}_{i+1})) - f_i^q(\tilde{\mathbf{z}}_{i+1})\right)\right.\right. \\ &\qquad \left.\left. + f_0(\tilde{\mathbf{z}}_1) - f_0(H(\tilde{\mathbf{z}}_1))\right]\right| \\ &\le \mathbb{E}_{q^{\tau>0}(\tilde{\mathbf{z}})}\left[\sum_{i=1}^{L-1}\left(\left|f_i^p(\tilde{\mathbf{z}}_{i+1}) - f_i^p(H(\tilde{\mathbf{z}}_{i+1}))\right| + \left|f_i^q(\tilde{\mathbf{z}}_{i+1}) - f_i^q(H(\tilde{\mathbf{z}}_{i+1}))\right|\right)\right. \\ &\qquad \left. + \left|f_0(\tilde{\mathbf{z}}_1) - f_0(H(\tilde{\mathbf{z}}_1))\right|\right] \\ &\le \kappa_0\ \mathbb{E}_{q^{\tau>0}(\tilde{\mathbf{z}})}\Big[\left\|\tilde{\mathbf{z}}_1 - H(\tilde{\mathbf{z}}_1)\right\|_2\Big] + \sum_{i=1}^{L-1}(\kappa_i^p + \kappa_i^q)\ \mathbb{E}_{q^{\tau>0}(\tilde{\mathbf{z}})}\Big[\left\|\tilde{\mathbf{z}}_{i+1} - H(\tilde{\mathbf{z}}_{i+1})\right\|_2\Big]\end{aligned}$$

Note that the terms for $\mathbf{z}_L$ cancel out when taking the difference, because both in the original and relaxed-ELBO we evaluate the log-density ratio of the categorical distributions at the hard samples.

Appendix F. Hierarchical Variational Autoencoders

In this section we define the VAE (Rezende et al., 2014; Kingma et al., 2016; Dieng et al., 2018), the LVAE (Sønderby et al., 2016) and BIVA (Maaløe et al., 2019). All models are characterized by a generative model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$ and can be coupled with any variational distribution.

Variational Autoencoder (VAE)

$$p_\theta(\mathbf{z}) = p_\theta\left(\mathbf{z}_L\right) \prod_{i=1}^{L-1} p_\theta\left(\mathbf{z}_i|\mathbf{z}_{i+1}\right) \qquad q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi\left(\mathbf{z}_1|\mathbf{x}\right) \prod_{i=2}^{L} q_\phi\left(\mathbf{z}_i|\mathbf{z}_{i-1}\right) \ . \tag{17}$$

Variational Autoencoder with Skip-Connections (Skip-VAE)

$$p_\theta(\mathbf{z}) = p_\theta\left(\mathbf{z}_L\right) \prod_{i=1}^{L-1} p_\theta\left(\mathbf{z}_i|\mathbf{z}_{>i}\right) \qquad q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi\left(\mathbf{z}_1|\mathbf{x}\right) \prod_{i=2}^{L} q_\phi\left(\mathbf{z}_i|\mathbf{z}_{<i},\ \mathbf{x}\right) \ . \tag{18}$$

Ladder Variational Autoencoder (LVAE)

$$p_\theta(\mathbf{z}) = p_\theta\left(\mathbf{z}_L\right) \prod_{i=1}^{L-1} p_\theta\left(\mathbf{z}_i|\mathbf{z}_{i+1}\right) \qquad q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi\left(\mathbf{z}_L|\mathbf{x}\right) \prod_{i=1}^{L-1} q_{\phi,\theta}\left(\mathbf{z}_i|\mathbf{z}_{i+1}, \mathbf{x}\right) \tag{19}$$

Ladder Variational Autoencoder with Skip-Connections (Skip-LVAE)

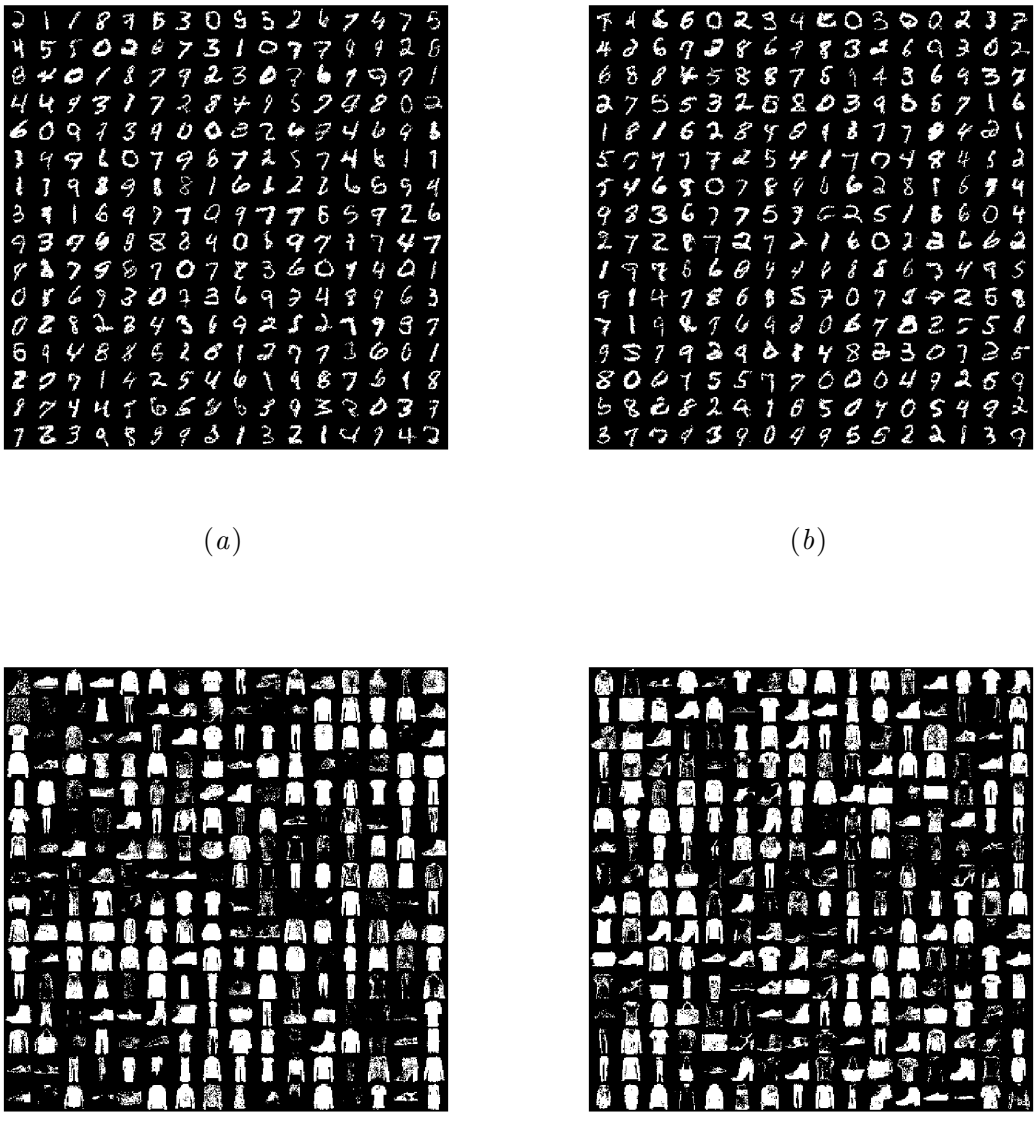
$$p_\theta(\mathbf{z}) = p_\theta\left(\mathbf{z}_L\right) \prod_{i=1}^{L-1} p_\theta\left(\mathbf{z}_i|\mathbf{z}_{>i}\right) \qquad q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi\left(\mathbf{z}_L|\mathbf{x}\right) \prod_{i=1}^{L-1} q_{\phi,\theta}\left(\mathbf{z}_i|\mathbf{z}_{>i}, \mathbf{x}\right) \tag{20}$$

Bidirectional Variational Autoencoder (BIVA)

$$p_\theta(\mathbf{z}) = p_\theta\left(\mathbf{z_L}\right) \prod_{i=1}^{L-1} p_\theta\left(\mathbf{z}_\mathbf{i}^{\mathrm{BU}}|\mathbf{z}_{>\mathbf{i}}\right) p_\theta\left(\mathbf{z}_\mathbf{i}^{\mathrm{TD}}|\mathbf{z}_{>\mathbf{i}}\right) \tag{21}$$

$$q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi\left(\mathbf{z_L}|\mathbf{x}, \mathbf{z}_{<\mathbf{L}}^{\mathrm{BU}}\right) \prod_{i=1}^{L-1} q_\phi\left(\mathbf{z}_\mathbf{i}^{\mathrm{BU}}|\mathbf{x}, \mathbf{z}_{<\mathbf{i}}^{\mathrm{BU}}\right) q_{\phi,\theta}\left(\mathbf{z}_\mathbf{i}^{\mathrm{TD}}|\mathbf{x}, \mathbf{z}_{<\mathbf{i}}^{\mathrm{BU}}, \mathbf{z}_{>\mathbf{i}}^{\mathrm{BU}}, \mathbf{z}_\mathbf{i}^{\mathrm{TD}}\right) \tag{22}$$

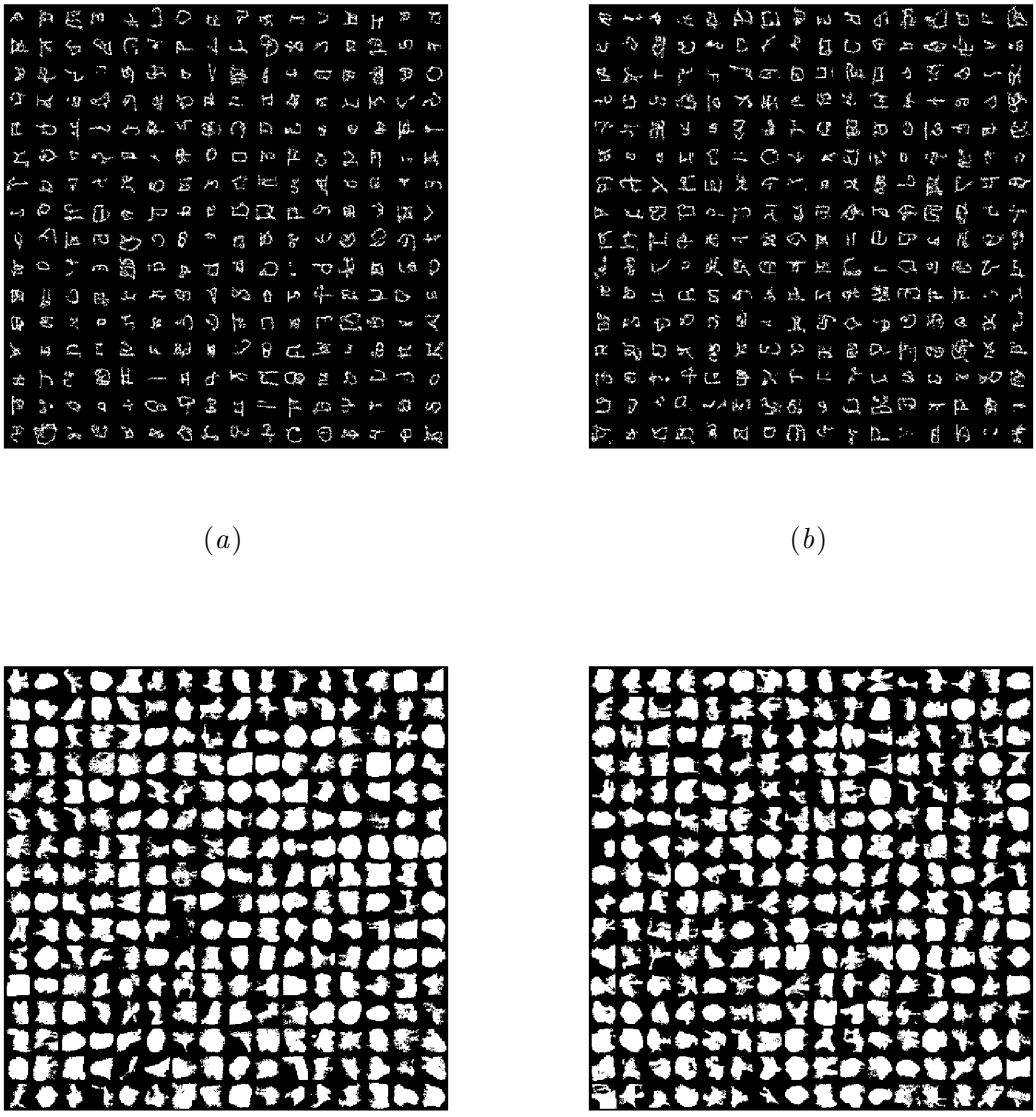
Appendix G. Samples



(a) (b)

(c) (d)

Figure 2: Prior Samples generated using the Gaussian LVAE (a, c) and HD-VAE (b, d) with L = 3 and $\tau = 0$ for statically Binarized MNIST and binarized Fashion MNIST.



(a) (b)

(c) (d)

Figure 3: Prior Samples generated using the Gaussian LVAE (a, c) and HD-VAE (b, d) with L = 3 and $\tau = 0$ for Omniglot and Caltech 101 Silouhettes.