

Lars Maaløe¹ Marco Fraccaro² Valentin Liévin³ Ole Winther^{3,4}

Abstract

With the introduction of the variational autoencoder (VAE), probabilistic latent variable models have received renewed attention as powerful generative models. However, their performance in terms of test likelihood and quality of generated samples has been surpassed by autoregressive models without stochastic units. Furthermore, flow-based models have recently been shown to be an attractive alternative that scales well to high-dimensional data. In this paper we close the performance gap by constructing VAE models that can effectively utilize a deep hierarchy of stochastic variables and model complex covariance structures. We introduce the Bidirectional-Inference Variational Autoencoder (BIVA), characterized by a skip-connected generative model and an inference network formed by a bidirectional stochastic inference path. We show that BIVA reaches state-of-the-art test likelihoods, generates sharp and coherent natural images, and uses the hierarchy of latent variables to capture different aspects of the data distribution. We observe that BIVA, in contrast to recent results, can be used for anomaly detection. We attribute this to the hierarchy of latent variables which is able to extract high-level semantic features. Finally, we extend BIVA to semi-supervised classification tasks and show that it performs comparably to state-of-the-art results by generative adversarial networks.

1. Introduction

One of the key aspirations in recent machine learning research is to build models that *understand the world* (Kingma, 2013; Rezende et al., 2014; Goodfellow et al., 2014; van den Oord et al., 2016c). Generative models are providing the means to learn from a plethora of unlabeled data in order to model a complex data distribution, e.g. natural images, text, and audio. These models are evaluated by their ability to *generate* data that is similar to the input data distribution from which they were trained on. The range of applications that come with generative models are vast, where audio synthesis (van den Oord et al., 2016a) and semi-supervised



Figure 1. 64x64 CelebA samples generated from a BIVA with increasing levels of stochasticity in the model (going from close to the mode to the full distribution). In each column the latent variances are scaled with factors 0.1, 0.3, 0.5, 0.7, 0.9, 1.0. Images in a row look similar because they use the same Gaussian random noise ϵ to generate the latent variables. BIVA has $L = 20$ stochastic latent layers connected by three layer ResNet blocks.

classification (Rasmus et al., 2015; Maaløe et al., 2016; Salimans et al., 2016) are examples hereof. Generative models can be broadly divided into explicit and implicit density models. The generative adversarial network (GAN) (Goodfellow et al., 2014) is an example of an implicit model, since it is not possible to procure a likelihood estimation from this model framework. The focus of this research is instead within explicit density models, for which a tractable or approximate likelihood estimation can be performed.

The three main classes of powerful explicit density models are autoregressive models (Larochelle & Murray, 2011; van den Oord et al., 2016c), flow-based models (Dinh et al., 2014; 2016; Kingma & Dhariwal, 2018), and probabilistic latent variable models (Kingma, 2013; Rezende et al., 2014; Mnih & Gregor, 2014). In recent years autoregressive models, such as the PixelRNN and the PixelCNN (van den Oord et al., 2016c; Salimans et al., 2017), have achieved superior likelihood performance and flow-based models have proven efficacy on large-scale natural image generation tasks (Kingma & Dhariwal, 2018). However, in the

¹Corti, Copenhagen Denmark ²Unumed, Copenhagen, Denmark

³Technical University of Denmark, Lyngby, Denmark

⁴University of Copenhagen, Copenhagen, Denmark. Correspondence to: Lars Maaløe <lm@corti.ai>

autoregressive models the runtime performance of generation is scaling poorly with the complexity of the input distribution. The flow-based models do not possess this restriction and do indeed generate visually compelling natural images when sampling close to the mode of the distribution. However, generation from the actual learned distribution is less compelling and the likelihood is far from the one of autoregressive models (Kingma & Dhariwal, 2018).

Probabilistic latent variable models possess intriguing properties that are different from the other classes of explicit density models. They are characterized by a posterior distribution over the latent variables of the model, derived from Bayes' theorem, which is typically intractable and needs to be approximated. This distribution most commonly lies on a low-dimensional manifold that can provide insights into the internal representation of the data (Bengio et al., 2013). However, the latent variable models have largely been disregarded as powerful generative models due to *blurry* generations and poor likelihood performances on natural image tasks. Larsen et al. (2016); Dosovitskiy & Brox (2016), amongst others, attribute this tendency to the usage of a similarity metric in pixel space. Contrarily, we attribute it to the lack of overall model expressiveness for accurately modeling complex input distributions, as discussed in Zhao et al. (2017); Rezende & Viola (2018).

There has been much research into explicitly defining and learning more expressive latent variable models. Here, the complementary research into learning a covariance structure through a framework of normalizing flows (Rezende & Mohamed, 2015; Tomczak & Welling, 2016; Kingma et al., 2016) and the stacking of a hierarchy of latent variables (Burda et al., 2015a; Ranganath et al., 2016; Maaløe et al., 2016; Sønderby et al., 2016) have shown promising results. However, despite significant improvements, the reported performance of these models has still been inferior to their autoregressive counterparts. This has spawned a new class of explicit density models that adds an autoregressive component to the generative process of a latent variable model (Gulrajani et al., 2016; Chen et al., 2017). In this combination of model paradigms the latent variables can be viewed as merely a *lossy* representation of the input data and the model still suffers from the same issues as autoregressive models.

Contributions. In this research we argue that latent variable models that are defined in a sufficiently expressive way can compete with autoregressive and flow-based models in terms of test log-likelihood and quality of the generated samples. We introduce the Bidirectional-Inference Variational Autoencoder (BIVA), a model formed by a deep hierarchy of stochastic variables that uses skip-connections to enhance the flow of information and avoid inactive units. To define a flexible posterior approximation, we construct a

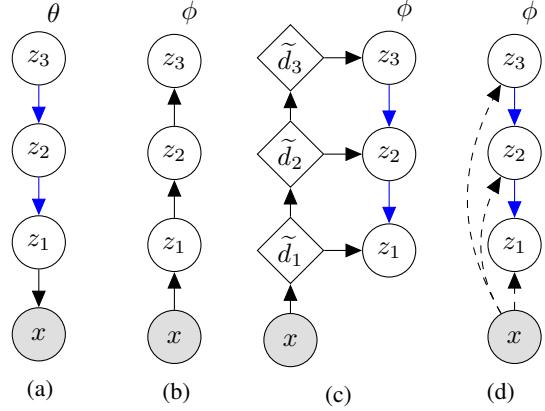


Figure 2. (a) Generative model of a VAE/LVAE with $L = 3$ stochastic variables, (b) VAE inference model, (c) LVAE inference model, and (d) skip connections among stochastic variables in the LVAE where dashed lines denote a skip-connection. Blue arrows indicate that shared parameters between the inference and generative model in LVAE.

bidirectional inference network using stochastic variables in a bottom-up and a top-down inference path. We perform an in-depth analysis of BIVA and show (i) an ablation study that analyses the contributions of the individual novel components, (ii) that the model is able to improve on state-of-the-art results on benchmark image datasets, (iii) that a small extension of the model can be used for semi-supervised classification and performs comparably to current state-of-the-art models, and (iv) that the model, contrarily to other state-of-the-art explicit density models (Nalisnick et al., 2018), can be utilized for anomaly detection on complex data distributions.

2. Deep Learning and Variational Inference

The introduction of stochastic backpropagation (Paisley et al., 2012; Hoffman et al., 2013) and the variational auto-encoder (VAE) (Kingma, 2013; Rezende et al., 2014) has made approximate Bayesian inference and probabilistic latent variable models applicable to machine learning problems considering complex data distributions, e.g. natural images, audio, and text. The VAE is a generative model parameterized by a neural network θ and is defined by an observed variable x that depends on a hierarchy of stochastic latent variables $\mathbf{z} = z_1, \dots, z_L$ so that: $p_\theta(x, \mathbf{z}) = p_\theta(x|z_1)p_\theta(z_L) \prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})$. This is illustrated in Figure 2a.

The distributions $p_\theta(z_i|z_{i+1})$ over the latent variables of the VAE are normally defined as Gaussians with diagonal covariance, whose parameters depend on the previous latent variable in the hierarchy (with the top latent variable $p_\theta(z_L) = \mathcal{N}(z_L; 0, I)$). The likelihood $p_\theta(x|z_1)$ is typically a Gaussian distribution for continuous data, or a Bernoulli distribution for binary data.

In order to learn the parameters θ we seek to maximize the log marginal likelihood over a training set: $\sum_i \log p_\theta(x_i) = \sum_i \log \int p_\theta(x_i, \mathbf{z}_i) d\mathbf{z}_i$. However, complex data distributions require an expressive model, which makes the above integral intractable. In order to circumvent this we use Variational Inference (Jordan et al., 1999), and introduce a posterior approximation $q_\phi(\mathbf{z}|x)$, known as *inference network* or *encoder*, that is parameterized by a neural network ϕ . Using Jensen’s inequality we can derive the *evidence lower bound* (ELBO), a lower bound to the integral in the marginal likelihood which is a function of the variational approximation $q_\phi(\mathbf{z}|x)$ and the generative model $p_\theta(x, \mathbf{z})$:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(\mathbf{z}|x)} \left[\log \frac{p_\theta(x, \mathbf{z})}{q_\phi(\mathbf{z}|x)} \right] \equiv \mathcal{L}(\theta, \phi). \quad (1)$$

The parameters θ and ϕ can be optimized by maximizing the ELBO with stochastic backpropagation and the reparameterization trick, which allows to use gradient ascent algorithms with low variance gradient estimators (Kingma, 2013; Rezende et al., 2014). As illustrated in Figure 2b, in a VAE the variational approximation is factorized with a bottom-up structure, $q_\phi(\mathbf{z}|x) = q_\phi(z_1|x) \prod_{i=1}^{L-1} q_\phi(z_{i+1}|z_i)$, so that each latent variable is conditioned on the variable below in the hierarchy. For ease of computation, all the factors in the variational approximation are typically assumed to be Gaussians whose mean and diagonal covariance are parameterized by neural networks.

Latent variable collapse in VAEs. A deep hierarchy of latent stochastic variables will result in a more expressive model. However, the additional variables come at a price. As shown in Chen et al. (2017); Maaløe et al. (2017), we can rewrite the ELBO (eq. (1)):

$$\begin{aligned} \mathcal{L}(\theta, \phi) = & \mathbb{E}_{q_\phi(\mathbf{z}|x)} \left[\log \frac{p_\theta(x, z_{<L}|z_L)}{q_\phi(z_{<L}|x)} \right] - \\ & \mathbb{E}_{q_\phi(z_{<L}|x)} [KL[q_\phi(z_L|z_{<L})||p_\theta(z_L)]] . \end{aligned} \quad (2)$$

From the above it becomes obvious that, during the optimization of the VAE, the top stochastic latent variables may have a tendency to *collapse* into the prior, i.e. $q_\phi(z_L|z_{<L}) = p_\theta(z_L) = \mathcal{N}(z_L; 0, I)$, if the model $p_\theta(x, z_{<L}|z_L)$ is powerful enough. This is supported by empirical results in Sønderby et al. (2016); Bowman et al. (2015) amongst others. The tendency has limited the applicability of deep VAEs in problems with complex data distributions, and has pushed VAE research towards the extension of shallow VAEs with autoregressive models, that allow to capture a *lossy* representation in the latent space while achieving strong generative performances (Gulrajani et al., 2016; Chen et al., 2017). Another research direction has focused on learning more complex prior distributions through normalizing flows (Rezende & Mohamed, 2015; Tomczak & Welling, 2016; Kingma et al., 2016). Our research considers instead

the original goal of building expressive models that can exploit a deeper hierarchy of stochastic latent variables while avoiding variable collapse.

The Ladder VAE (LVAE) (Sønderby et al., 2016) is amongst the first attempts towards VAEs that can effectively leverage multiple layers of stochastic variables. This is achieved by parameterizing the variational approximation with a *bottom-up* deterministic path followed by a *top-down* inference path that shares parameters with the top-down structure of the generative model: $q_{\phi,\theta}(\mathbf{z}|x) = q_\phi(z_L|x) \prod_{i=1}^{L-1} q_{\phi,\theta}(z_i|z_{i+1}, x)$. A graphical representation of the LVAE inference network can be found in Figure 2c. Thanks to the bottom-up path, all the latent variables in the hierarchy have a deterministic dependency on the observed variable x , which allows data-dependent information to skip all the stochastic variables lower in the hierarchy (Figure 2d). The stochastic latent variables that are higher in the hierarchy will therefore receive less noisy inputs, and will be empirically less likely to collapse. Despite the improvements obtained thanks to the more flexible inference network, in practice LVAEs with a very deep hierarchy of stochastic latent variables will still experience variable collapse. In the next section we will introduce the Bidirectional-Inference Variational Autoencoder, that manages to avoid these issues by extending the LVAE in 2 ways: (i) adding a deterministic top-down path in the generative model and (ii) defining a factorization of the latent variables z_i at each level of the hierarchy that allows to construct a bottom-up *stochastic* inference path.

3. Bidirectional-Inference Variational Autoencoder

In this section we will first describe the architecture of the Bidirectional-Inference Variational Autoencoder (Figure 3), and then provide the motivation behind the main ideas of the model as well as some intuitions on the role of each of its novel components. Finally, we will show how this model can be used for a novel approach to detecting anomalous data.

3.1. Model architecture

Generative model. In BIVA, at each layer $1, \dots, L-1$ of the hierarchy we split the latent variable in two components, $z_i = (z_i^{\text{BU}}, z_i^{\text{TD}})$, which belong to a bottom-up (BU) and top-down (TD) inference path, respectively. More details on this will be given when introducing the inference network. The generative model of BIVA is illustrated in Figure 3a. We introduce a deterministic top-down path d_{L-1}, \dots, d_1 that is parameterized with neural networks and receives as input at each layer i of the hierarchy the latent variable z_{i+1} . This is done by defining a fully convolutional model

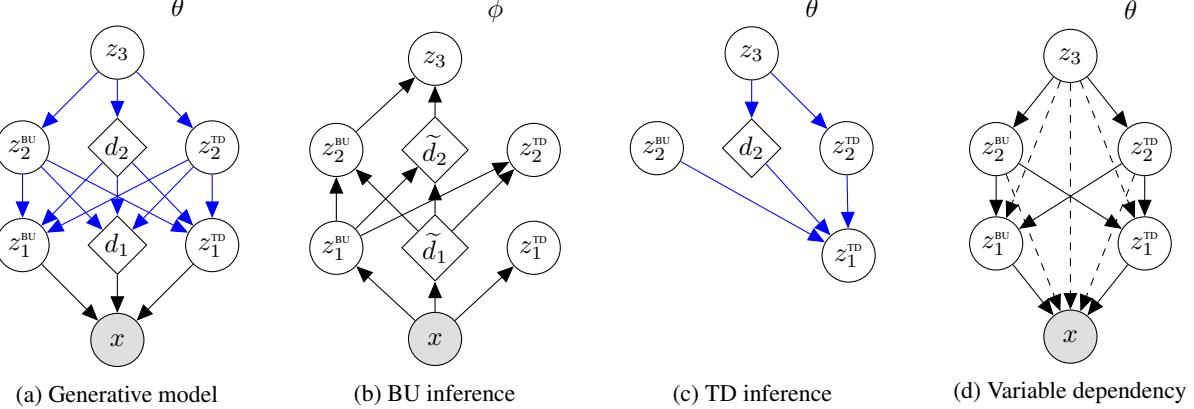


Figure 3. A $L = 3$ layered BIVA with (a) the generative model, (b) bottom-up (BU) inference path, (c) top-down (TD) inference path, and (d) variable dependency of the generative models where dashed lines denote a skip-connection. Blue arrows indicate that the deterministic parameters are shared within the generative model or between the generative and inference model.

and concatenating $(z_{i+1}^{\text{BU}}, z_{i+1}^{\text{TD}})$ and d_{i+1} along the features' dimension. d_i can therefore be seen as a deterministic variable that summarizes all the relevant information coming from the stochastic variables higher in the hierarchy, $z_{>i}$. The latent variables z_i^{BU} and z_i^{TD} are conditioned on all the information in the higher layers, and are conditionally independent given $z_{>i}$. The joint distribution of the model is then given by:

$$p_\theta(x, \mathbf{z}) = p_\theta(x|\mathbf{z})p_\theta(z_L) \prod_{i=1}^{L-1} p_\theta(z_i^{\text{BU}}|z_{>i})p_\theta(z_i^{\text{TD}}|z_{>i}),$$

where θ are the parameters of the generative model. Each stochastic latent variable $1, \dots, L$ is parameterized by a Gaussian distribution with diagonal covariance, with one neural network $\mu(\cdot)$ for the mean and another neural network $\sigma(\cdot)$ for the variance. Since the z_{i+1}^{BU} and z_{i+1}^{TD} variables are on the same level in the generative model and of the same dimensionality, we share all the deterministic parameters going to the layer below. See Figure 3a and Appendix A for details.

Bidirectional inference network. Due to the nonlinearities in the neural networks that parameterize the generative model, the exact posterior distribution $p_\theta(\mathbf{z}|x)$ is intractable and needs to be approximated. As for VAEs, we therefore define a variational distribution $q_\phi(\mathbf{z}|x)$, that needs to be flexible enough to approximate as closely as possible the true posterior distribution. We define a bottom-up (BU) and a top-down (TD) inference path, which are computed sequentially when constructing the posterior approximation for each data point x , see Figures 3b and 3c. The variational distribution over the BU latent variables depends on the data x and on all BU variables lower in the hierarchy, i.e. $q_\phi(z_i^{\text{BU}}|x, z_{<i}^{\text{BU}})$, where ϕ denotes all the parameters of the BU path. As seen in Figure 3b, z_i^{BU} has a direct dependency only on the BU variable below,

z_{i-1}^{BU} . The dependency on $z_{<i-1}^{\text{BU}}$ is achieved, similarly to the generative model, through a deterministic bottom-up path $\tilde{d}_1, \dots, \tilde{d}_{L-1}$. The TD variables as well depend on the data and the BU variables lower in the hierarchy through the BU inference path, but also on all variables above in the hierarchy through the TD inference path in Figure 3c. The variational approximation over the TD variables is thereby $q_{\phi, \theta}(z_i^{\text{TD}}|x, z_{<i}^{\text{BU}}, z_{>i}^{\text{BU}}, z_{>i}^{\text{TD}})$. Importantly, all the parameters of the TD path are shared with the generative model, and are therefore denoted as θ . The overall inference network can be factorized as follows:

$$q_\phi(\mathbf{z}|x) = q_\phi(z_L|x, z_{<L}^{\text{BU}}) \prod_{i=1}^{L-1} q_\phi(z_i^{\text{BU}}|x, z_{<i}^{\text{BU}}) q_{\phi, \theta}(z_i^{\text{TD}}|x, z_{<i}^{\text{BU}}, z_{>i}^{\text{BU}}, z_{>i}^{\text{TD}}),$$

where the variational distributions over the BU and TD latent variables are Gaussians whose mean and diagonal covariance are parameterized with neural networks that take as input the concatenation over the feature dimension of the conditioning variables (BIVA is fully convolutional). Training of BIVA is performed, as for VAEs, by maximizing the ELBO in eq. (1) with stochastic backpropagation and the reparameterization trick.

3.2. Motivation

BIVA can be seen as an extension of the LVAE in which we (i) add a deterministic top-down path and (ii) apply a bidirectional inference network. We will now provide the motivation and some intuitions on the role of these two novel components, that will then be validated with the experiments of Section 4.

Deterministic top-down path. Skip-connections represent one of the simplest yet most powerful advancements of deep learning in recent years. They allow to construct



Figure 4. (left) images from the CelebA dataset preprocessed to 64x64 following (Larsen et al., 2016). (right) $\mathcal{N}(0, I)$ generations of BIVA with $L = 20$ layers that achieves a $\mathcal{L}_1 = 2.48$ bits/dim on the test set.

very deep neural networks, by better propagating the information throughout the model and reducing the issue of vanishing gradients. Skip connections form for example the backbone of deep neural networks such as ResNets (He et al., 2015), which have shown impressive performances on a wide range of classification tasks. Our goal in this paper is to build very deep latent variable models that are able to learn an expressive latent hierarchical representation of the data. In our experiments we however found that the LVAE still had difficulties in activating the top latent variables for deeper hierarchies. To limit this issue, we add skip connections among the latent variables in the generative model by adding the deterministic top-down path, that makes each variable depend on all the variables above in the hierarchy (see Figure 3d for a graphical representation). This allows a better flow of information in the model and thereby avoids the collapse of latent variables. A related idea was recently proposed by Dieng et al. (2018), that add skip connections among the neural network layers parameterizing a shallow VAE with a single latent variable.

Bidirectional inference. The inspiration for the bidirectional inference network of BIVA comes from the work on Auxiliary VAEs (AVAE) by Ranganath et al. (2016); Maaløe et al. (2016). An AVAE can be viewed as a shallow VAE with a single latent variable z and an auxiliary variable a that increases the expressiveness of the variational approximation $q_\phi(z|x) = \int q_\phi(z|a, x)q_\phi(a|x)da$. By making the inference network $q_\phi(z|a, x)$ depend on the stochastic variable a , the AVAE adds covariance structure to the posterior approximation over the stochastic unit z , since it no longer factorizes over its components $z^{(k)}$, i.e. $q_\phi(z|x) \neq \prod_k q_\phi(z^{(k)}|x)$. As discussed in the following, by factorizing the latent variables at each level of the hierarchy of BIVA we are able to achieve similar results without introducing additional auxiliary variables in the model. To see this, we can focus for example on the highest latent variable z_L . In BIVA, the presence of the z_i^{BU} variables makes the bottom-up inference path *stochastic*, as opposed to the deterministic BU path of the LVAE. While the conditional distribution $q_\phi(z_L|x, z_{<L}^{\text{BU}})$ still factorizes over the components of z_L , due to the stochastic BU

variables the marginal distribution over z_L no longer factorizes, i.e. $q_\phi(z_L|x) = \int q_\phi(z_L|x, z_{<L}^{\text{BU}})q_\phi(z_{<L}^{\text{BU}}|x)dz_{<L}^{\text{BU}} \neq \prod_{k=1}^K q(z_L^{(k)}|x)$. Therefore, the BU inference path enables the learning of a complex covariance structure in the higher TD stochastic latent variables, which is fundamental in the model to extract *good* high-level semantic features from the data distribution. Notice that in BIVA only z_1^{BU} will have a marginally factorizing inference network.

3.3. Anomaly detection with BIVA

Anomaly detection is considered to be one of the most important applications of explicit density models. However, recent empirical results suggest that these models are not able to distinguish between two clearly distinctive data distributions, as they can assign a higher likelihood to data points from a data distribution that is very different from the one the model was trained on (Nalisnick et al., 2018). Based on a thorough study, Nalisnick et al. (2018) state that the main issue is the fact that explicit density models have a tendency to capture low-level statistics, as opposed to the high-level semantics that are preferable when doing anomaly detection. We hypothesize that the latent representations in the higher layers of BIVA can capture the high-level semantics of the data and that these can be used for improved anomaly detection.

In the standard ELBO from eq. (1), the main contribution to the expected log-likelihood term is coming from averaging over the variational distribution of the lower level latent variables. This will thus emphasize low-level statistics. So in order to perform anomaly detection with BIVA we instead need to emphasize the contribution from the higher layers. We can achieve this with an alternative log-likelihood lower bound that partly replaces the inference network with the generative model. It will be a weaker bound than the ELBO, but it has the advantage that it explicitly uses the generative hierarchy of the stochastic variables. In the following we define the hierarchy of stochastic latent variables as $\mathbf{z} = z_1, z_2, z_3, \dots, z_L$ with $z_i = (z_i^{\text{BU}}, z_i^{\text{TD}})$. Instead of using the variational approximation $q_\phi(\mathbf{z}|x)$ over all stochastic variables in the model, we use the prior distribution for the

first k layers and the variational approximation for the others, i.e. $p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x)$.

The new ELBO becomes:

$$\mathcal{L}^{>k} = \mathbb{E}_{p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x)} \left[\log \frac{p_\theta(x|\mathbf{z})p_\theta(z_{>k})}{q_\phi(z_{>k}|x)} \right]. \quad (3)$$

$\mathcal{L}^{>0} = \mathcal{L}$ is the ELBO in eq. (1). As for the ELBO, we approximate the computation of $\mathcal{L}^{>k}$ with Monte Carlo integration. Sampling from $p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x)$ can be easily performed by obtaining samples $\hat{z}_{>k}$ from the inference network, that are then used to sample $\hat{z}_{\leq k}$ from the conditional prior $p_\theta(z_{\leq k}|\hat{z}_{>k})$.

Due to the sampling from the prior, eq. (3) will generally return a worse likelihood approximation than the ELBO. Despite this, $\mathcal{L}^{>k}$ with high values of k represents a useful metric for anomaly detection. By only sampling the top $L - k$ variables from the variational approximation, in fact, we are forcing the model to only rely on the high-level semantics encoded in the highest variables of the hierarchy when evaluating this metric, and not on the low-level statistics encoded in the lower variables.

4. Experiments

BIVA is empirically evaluated by (i) an ablation study analyzing each novel component, (ii) likelihood and semi-supervised classification results on binary images, (iii) likelihood results on natural images, and (iv) an analysis of anomaly detection in complex data distributions. We employ a *free bits* strategy with $\lambda = 2$ (Kingma et al., 2016) for all experiments to avoid latent variable collapse during the initial training epochs. Trained models are reported with 1 importance weighted sample, \mathcal{L}_1 , and 1000 importance weighted samples, \mathcal{L}_{1e3} (Burda et al., 2015b). We evaluate the natural image experiments by bits per dimension (bits/dim), $\mathcal{L}/(hwc \log(2))$, where h, w, c denotes the height, width, and channels respectively. In the Appendix we give further details and samples for the image tasks and initial results for the model applied to text. For a detailed description of the experimental setup see Appendix C and the source code¹.

4.1. Ablation Study

BIVA can be viewed as an extension of the LVAE from Sønderby et al. (2016) where we add (i) extra dependences in the generative model ($p_\theta(x|z_1) \rightarrow p_\theta(x|\mathbf{z})$ and $p_\theta(z_i|z_{i+1}) \rightarrow p_\theta(z_i|z_{>i})$) through the skip connections obtained with the deterministic top-down path and (ii) a bottom-up (BU) path of stochastic latent variables to the inference model. In order to evaluate the effects of each added

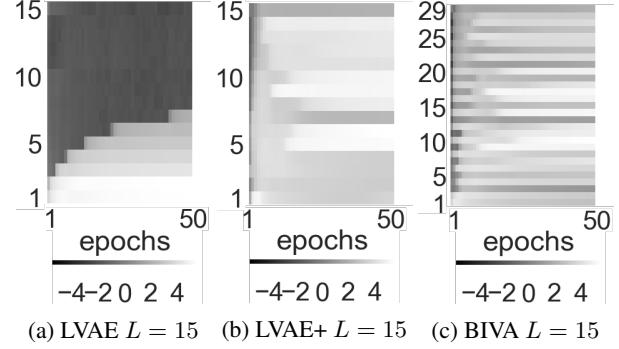


Figure 5. The $\log KL(q||p)$ for each stochastic latent variable as a function of the training epochs on CIFAR-10. (a) is a $L = N = 15$ stochastic latent layer LVAE with no skip-connections and no bottom-up inference. (b) is a $L = N = 15$ LVAE+ with skip-connections and no bottom-up inference. (c) is a $L = 15$ stochastic latent layer ($N = 29$ latent variables) BIVA for which $1, 2, \dots, N$ denotes the stochastic latent variables following the order $z_1^{\text{BU}}, z_1^{\text{TD}}, z_2^{\text{BU}}, z_2^{\text{TD}}, \dots, z_L$.

	PARAMETERS	BITS/DIM
LVAE L=15, \mathcal{L}_1	10.85M	≤ 3.60
LVAE+ L=15, \mathcal{L}_1	11.36M	≤ 3.41
LVAE+ L=29, \mathcal{L}_1	21.99M	≤ 3.45
BIVA L=15, \mathcal{L}_1	18.48M	≤ 3.12

Table 1. A comparison of the LVAE with no skip-connections and no bottom-up inference, the LVAE+ with skip-connections and no bottom-up inference, and BIVA. All models are trained on the CIFAR-10 dataset.

component we define an LVAE with the exact same architecture as BIVA, but without the BU variables and the deterministic top-down path. Next, we define the LVAE+, where we add to the LVAE’s generative model the deterministic top-down path. It is therefore the same model as in Figure 3 but without the BU variables. Finally, we investigate a LVAE+ model with $2L - 1$ stochastic layers. This corresponds to the depth of the hierarchy of the BIVA inference model $x \rightarrow z_1^{\text{BU}} \rightarrow \dots \rightarrow z_{L-1}^{\text{BU}} \rightarrow z_L \rightarrow z_{L-1}^{\text{TD}} \rightarrow \dots \rightarrow z_1^{\text{TD}}$. If this model is competitive with BIVA then it is an indication that it is the depth that determines the performance. The ablation study is conducted on the CIFAR-10 dataset against the best reported BIVA with $L = 15$ layers (Section 4.3), which means $2L - 1 = 29$ stochastic latent layers in the deep LVAE+.

Table 1 presents a comparison of the different model architectures. The positive effect of adding the skip connections in the generative models can be evaluated from the difference between the LVAE $L = 15$ and LVAE+ $L = 15$ results, for which there is close to a 0.2 bits/dim difference in the ELBO. Thanks to the more expressive posterior approximation obtained using its bidirectional inference network, BIVA improves the ELBO significantly w.r.t the LVAE+, by more than 0.3 bits/dim. Notice that a deeper hierarchy of stochastic latent variables in the LVAE+ will not necessarily

¹The source code will be released on [Github](#).

provide a better likelihood performance, since the LVAE+ $L = 29$ performs worse than the LVAE+ $L = 15$ despite having significantly more parameters.

In Figure 5 we plot for LVAE, LVAE+ and BIVA the KL divergence between the variational approximation over each latent variable and its prior distribution, $KL(q||p)$. This KL divergence is 0 when the two distributions match, in which case we say that the variable has collapsed, since its posterior approximation is not using any data-dependent information. We can see that while the LVAE is only able to utilize its lowest 7 stochastic variables, all variables in both LVAE+ and BIVA are active. We attribute this tendency to the deterministic top-down path that is present in both models, which creates skip-connections between all latent variables that allow to better propagate the information throughout the model.

4.2. Binary Images

We evaluate BIVA $L = 6$ in terms of test log-likelihood on statically binarized MNIST (Salakhutdinov & Murray, 2008), dynamically binarized MNIST (LeCun et al., 1998) and dynamically binarized OMNIGLOT (Lake et al., 2013). The model parameterization and optimization parameters have been kept identical for all binary image experiments (see Appendix C). For each experiment on binary image datasets, we *finetune* each model by setting the free bits to $\lambda = 0$ until convergence in order to test the tightness of the \mathcal{L}_1 ELBO.

To the best of our knowledge, BIVA achieves state-of-the-art results on statically binarized MNIST, outperforming other latent variable models, autoregressive models, and flow-based models (see Table 2). Finetuning the model with $\lambda = 0$ improves the \mathcal{L}_1 ELBO significantly and achieves slightly better performance for the 1000 importance weighted samples. For dynamically binarized MNIST and OMNIGLOT, BIVA achieves similar improvements with $\mathcal{L}_{1e3} = 78.41$ (state-of-the-art) and $\mathcal{L}_{1e3} = 91.34$ respectively, see Appendix E for additional results and details.

Semi-supervised learning. BIVA can be easily extended for semi-supervised classification by adding a categorical variable y to represent the class, as done in Kingma et al. (2014). We add a classification model $q_\phi(y|x, z_{<L}^{BU})$ to the inference network, and a class-conditional distribution $p_\theta(x|\mathbf{z}, y)$ to the generative model (see Appendix B for a detailed description). We train 5 different semi-supervised models on MNIST, each using a different set of just 100 randomly chosen and evenly distributed MNIST labels. Table 3 presents the classification results on the test set (mean and standard deviation over the 5 runs), that shows that BIVA achieves comparable performance to recent state-of-the-art results by generative adversarial networks.

	− log $p(x)$
<i>Results with autoregressive components</i>	
PIXELCNN (van den Oord et al., 2016c)	= 81.30
DRAW (Gregor et al., 2015)	< 80.97
IAFVAE (Kingma et al., 2016)	≤ 79.88
PIXELVAE (Gulrajani et al., 2016)	≤ 79.66
PIXELRNN (van den Oord et al., 2016c)	= 79.20
VLAЕ (Chen et al., 2017)	≤ 79.03
<i>Results without autoregressive components</i>	
DISCRETE VAE (Rolfe, 2017)	≤ 81.01
BIVA, \mathcal{L}_1	≤ 81.22
BIVA, \mathcal{L}_{1e3}	≤ 78.67
BIVA FINETUNED, \mathcal{L}_1	≤ 80.47
BIVA FINETUNED, \mathcal{L}_{1e3}	≤ 78.59

Table 2. Test log-likelihood on statically binarized MNIST for different number of importance weighted samples. The finetuned models are trained for an additional number of epochs with no *free bits*, $\lambda = 0$.

	ERROR %
M1+M2 (Kingma et al., 2014)	3.33% (± 0.14)
VAT (Miyato et al., 2015)	2.12%
CATGAN (Springenberg, 2015)	1.91% (± 0.10)
SDGM (Maaløe et al., 2016)	1.32% (± 0.07)
LADDERNET (Rasmus et al., 2015)	1.06% (± 0.37)
ADGM (Maaløe et al., 2016)	0.96% (± 0.02)
IMPGAN (Salimans et al., 2016)	0.93% (± 0.07)
TRIPLEGAN (Li et al., 2017)	0.91% (± 0.58)
SSLGAN (Dai et al., 2017)	0.80% (± 0.10)
BIVA	0.83% (± 0.02)

Table 3. Semi-supervised test error for BIVA on MNIST for 100 randomly chosen and evenly distributed labelled samples.

4.3. Natural Images

We trained and evaluated BIVA $L = 15$ on 32x32 CIFAR-10, 32x32 ImageNet (van den Oord et al., 2016c), and another BIVA $L = 20$ on 64x64 CelebA (Larsen et al., 2016). For the output decoding we employ the discretized logistic mixture likelihood from Salimans et al. (2017) (see Appendix C for more details). In Table 4 we see that for the CIFAR-10 dataset BIVA outperforms other state-of-the-art non-autoregressive models and performs slightly worse than state-of-the-art autoregressive models. Notice that BIVA has fewer parameters (18.48M) than PixelCNN++ (28.57M parameters, Salimans et al. (2017)). For the 32x32 ImageNet dataset BIVA achieves better performance than flow-based models, but the performance gap to the autoregressive models remains large (Table 5). This may be due to the added complexity (more categories) of the 32x32 ImageNet dataset, requiring an even more flexible model. More research should be invested in defining an improved architecture for BIVA that holds more parameters and thereby achieves better performances.

Figure 4 shows generated samples from the $\mathcal{N}(0, I)$ prior of a BIVA $L = 20$ trained on the CelebA dataset. From a visual inspection, the samples are far superior to previous natural

	BITS/DIM
<i>Results with autoregressive components</i>	
CONVDRAW (GREGOR ET AL., 2016)	< 3.58
IAFVAE \mathcal{L}_1 (KINGMA ET AL., 2016)	≤ 3.15
IAFVAE \mathcal{L}_{1e3} (KINGMA ET AL., 2016)	≤ 3.11
GATEDPIXELCNN (VAN DEN OORD ET AL., 2016B)	$= 3.03$
PIXELRNN (VAN DEN OORD ET AL., 2016C)	$= 3.00$
VLAЕ (CHEN ET AL., 2017)	≤ 2.95
PIXELCNN++ (SALIMANS ET AL., 2017)	$= 2.92$
<i>Results without autoregressive components</i>	
NICE (DINH ET AL., 2014)	$= 4.48$
DEEPGMMS (VAN DEN OORD & SCHRAUWEN, 2014)	$= 4.00$
REALNVP (DINH ET AL., 2016)	$= 3.49$
DISCRETEVAE++ (VAHDAT ET AL., 2018)	≤ 3.38
GLOW (KINGMA & DHARIWAL, 2018)	$= 3.35$
BIVA L=10, \mathcal{L}_1	≤ 3.17
BIVA L=15, \mathcal{L}_1	≤ 3.12
BIVA L=15, \mathcal{L}_{1e3}	≤ 3.08

Table 4. Test log-likelihood on CIFAR-10 for different number of importance weighted samples. We evaluated two different BIVA with various number of layers (L).

	BITS/DIM
<i>Results with autoregressive components</i>	
CONVDRAW (GREGOR ET AL., 2016)	< 4.10
PIXELRNN (VAN DEN OORD ET AL., 2016C)	$= 3.63$
GATEDPIXELCNN (VAN DEN OORD ET AL., 2016B)	$= 3.57$
<i>Results without autoregressive components</i>	
REALNVP (DINH ET AL., 2016)	$= 4.28$
GLOW (KINGMA & DHARIWAL, 2018)	$= 4.09$
BIVA, \mathcal{L}_1	≤ 3.98
BIVA, \mathcal{L}_{1e3}	≤ 3.96

Table 5. Test log-likelihood on ImageNet 32x32 for different number of importance weighted samples.

image generations by latent variable models. We believe that previous claims stating that this type of model can only generate *blurry* images should be disregarded (Larsen et al., 2016). Rather the limited expressiveness/flexibility of previous models should be blamed. Additional samples from BIVA can be found in Figure 1 and Appendix E.

4.4. Does BIVA know what it doesn't know?

We test the anomaly detection capabilities of BIVA replicating the most challenging experiment of Nalisnick et al. (2018). We train BIVA $L = 15$ on the CIFAR-10 dataset, and evaluate eq. (3) for various values of k both on the CIFAR-10 test set and on the SVHN dataset (Netzer et al., 2011). The results can be found in Table 6 and Figure 6, and are reported in terms of bits per dimension (lower is better). We see that for $k = 0$, corresponding to the standard ELBO, BIVA wrongly assigns lower values to data points from SVHN. This is in line with the results obtained with other explicit density models in Nalisnick et al. (2018), and shows that by using the standard ELBO the low-level image statistics prevail and the model is not able to correctly detect out-of-distribution samples. However, for higher values of k , the situation is reversed. We take this as an indication that

	CIFAR-10 TEST (BITS/DIM)	SVHN TEST (BITS/DIM)
BIVA $\mathcal{L}_1^{>13}$	79.36	121.04
BIVA $\mathcal{L}_{100}^{>13}$	54.91	71.38
BIVA $\mathcal{L}_{100}^{>11}$	35.34	58.82
BIVA $\mathcal{L}_1^{>9}$	20.93	26.76
BIVA $\mathcal{L}_1^{>0}$	3.12	2.28

Table 6. The $\mathcal{L}^{>k}$ for $k = 13, 11, 9, 0$ evaluated in bits/dim by a model trained on the CIFAR-10 train set and evaluated on the CIFAR-10 test set and the SVHN test set. It is possible to use BIVA for anomaly detection of high-level semantics in complex datasets when using a high k .

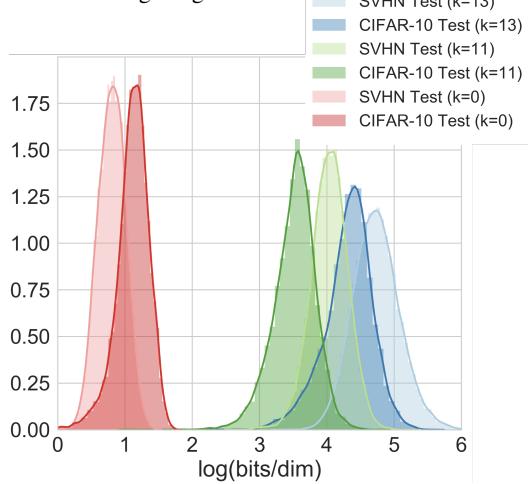


Figure 6. Histograms and kernel density estimation of the $\mathcal{L}^{>k}$ for $k = 13, 11, 0$ evaluated in bits/dim by a model trained on the CIFAR-10 train dataset and evaluated on the CIFAR-10 and the SVHN test set.

BIVA uses the high-level semantics inferred from the data to better differentiate between the CIFAR-10 and the SVHN distributions.

5. Conclusion

In this paper we have introduced BIVA, that significantly improves performances over previously introduced probabilistic latent variable models and flow-based models. BIVA is able to generate natural images that are both sharp and coherent, to improve on semi-supervised classification benchmarks and, contrarily to other models, allows for anomaly detection using the extracted high-level semantics of the data.

On natural image generation tasks, the autoregressive models are still superior to BIVA. Future directions should include experiments on BIVA with more parameters and testing for better architectures. Furthermore, the model flexibility can be increased by defining more expressive stochastic latent variables, such as *inverse autoregressive flows* (Kingma et al., 2016). Finally, BIVA can easily be complemented by state-of-the-art autoregressive models.

References

- Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 2013.
- Bowman, S., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.
- Burda, Y., Grosse, R., and Salakhutdinov, R. Importance Weighted Autoencoders. *arXiv preprint arXiv:1509.00519*, 2015a.
- Burda, Y., Grosse, R., and Salakhutdinov, R. Accurate and conservative estimates of mrf log-likelihood using reverse annealing. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2015b.
- Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. Variational Lossy Autoencoder. In *International Conference on Learning Representations*, 2017.
- Dai, Z., Yang, Z., Yang, F., Cohen, W. W., and Salakhutdinov, R. Good semi-supervised learning that requires a bad GAN. In *Advances in Neural Information Processing Systems*, 2017.
- Dieng, A. B., Kim, Y., Rush, A. M., and Blei, D. M. Avoiding latent variable collapse with generative skip models. *arXiv preprint arXiv:1807.04863*, 2018.
- Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
- Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.
- Dosovitskiy, A. and Brox, T. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, 2016.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2014.
- Gregor, K., Danihelka, I., Graves, A., and Wierstra, D. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.
- Gregor, K., Besse, Fredric, R. D. J., Danihelka, I., and Wierstra, D. Towards conceptual compression. *arXiv preprint arXiv:1604.08772*, 2016.
- Gulrajani, I., Kumar, K., Ahmed, F., Ali Taiga, A., Visin, F., Vazquez, D., and Courville, A. PixelVAE: A latent variable model for natural images. *arXiv e-prints*, 1611.05013, November 2016.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
- Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, November 1999.
- Kingma, D. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 12 2014.
- Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 2018.
- Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. Semi-Supervised Learning with Deep Generative Models. In *Proceedings of the International Conference on Machine Learning*, 2014.
- Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*. 2016.
- Kingma, Diederik P; Welling, M. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 12 2013.
- Lake, B. M., Salakhutdinov, R. R., and Tenenbaum, J. One-shot learning by inverting a compositional causal process. In *Advances in Neural Information Processing Systems*. 2013.
- Larochelle, H. and Murray, I. The neural autoregressive distribution estimator. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011.
- Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the International Conference on Machine Learning*, 2016.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2278–2324, 1998.

- Li, C., Xu, K., Zhu, J., and Zhang, B. Triple generative adversarial nets. *arXiv preprint arXiv:1703.02291*, 2017.
- Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. Auxiliary Deep Generative Models. In *Proceedings of the International Conference on Machine Learning*, 2016.
- Maaløe, L., Fraccaro, M., and Winther, O. Semi-supervised generation with cluster-aware generative models. *arXiv preprint arXiv:1704.00637*, 2017.
- Miyato, T., Maeda, S.-i., Koyama, M., Nakae, K., and Ishii, S. Distributional Smoothing with Virtual Adversarial Training. *arXiv preprint arXiv:1507.00677*, 7 2015.
- Mnih, A. and Gregor, K. Neural variational inference and learning in belief networks. In *Proceedings of the International Conference on Machine Learning*, pp. 1791–1799, 2014.
- Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Do deep generative models know what they don’t know? *arXiv preprint arXiv:1810.09136*, 2018.
- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *Deep Learning and Unsupervised Feature Learning, workshop at Neural Information Processing Systems 2011*, 2011.
- Paisley, J., Blei, D. M., and Jordan, M. I. Variational bayesian inference with stochastic search. In *Proceedings of the International Conference on Machine Learning*, pp. 1363–1370, 2012.
- Ranganath, R., Tran, D., and Blei, D. M. Hierarchical variational models. In *Proceedings of the International Conference on Machine Learning*, 2016.
- Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, 2015.
- Rezende, D. J. and Mohamed, S. Variational Inference with Normalizing Flows. In *Proceedings of the International Conference on Machine Learning*, 2015.
- Rezende, D. J. and Viola, F. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.
- Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv preprint arXiv:1401.4082*, 04 2014.
- Rolfe, J. T. Discrete variational autoencoders. In *Proceedings of the International Conference on Learning Representations*, 2017.
- Salakhutdinov, R. and Murray, I. On the quantitative analysis of deep belief networks. In *Proceedings of the International Conference on Machine Learning*, 2008.
- Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2016.
- Salimans, T., Kingma, D. P., and Welling, M. Markov chain monte carlo and variational inference: Bridging the gap. In *Proceedings of the International Conference on Machine Learning*, 2015.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.
- Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. Pixelenn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint:1701.05517*, 2017, 2017.
- Semeniuta, S., Severyn, A., and Barth, E. A hybrid convolutional variational autoencoder for text generation. *arXiv preprint arXiv:1702.02390*, 2017.
- Shah, H., Zheng, B., and Barber, D. Generating sentences using a dynamic canvas, 2018.
- Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems 29*, 2016.
- Springenberg, J. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.
- Tomczak, J. M. and Welling, M. Improving variational auto-encoders using householder flow. *arXiv preprint arXiv:1611.09630*, 2016.
- Tran, D., Ranganath, R., and Blei, D. M. Variational Gaussian process. In *Proceedings of the International Conference on Learning Representations*, 2016.
- Vahdat, A., Macready, W. G., Bian, Z., Khoshaman, A., and Andriyash, E. DVAE++: discrete variational autoencoders with overlapping transformations. In *Proceedings of the International Conference on Machine Learning*, 2018.

van den Oord, A. and Schrauwen, B. Factoring variations in natural images with deep gaussian mixture models. In *Advances in Neural Information Processing Systems*, 2014.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016a.

van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., and Kavukcuoglu, K. Conditional image generation with pixelcnn decoders. *arXiv preprint arXiv:1606.05328*, 2016b.

van den Oord, A., Nal, K., and Kavukcuoglu, K. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 01 2016c.

Zhao, S., Song, J., and Ermon, S. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.

A. Detailed Model Description

Generative model. The generative model has a top-down path going from z_L through the intermediary stochastic latent variables to x . Between each stochastic layer there is a ResNet block with M layers set up similarly to (Salimans et al., 2017). Weight normalization (Salimans & Kingma, 2016) is applied in all neural network layers. In the generative model the BU and TD units are not distinguished so we write $z_i = (z_i^{\text{BU}}, z_i^{\text{TD}})$. We use $f_{i,j}$ to denote the neural network function (a function of generative model parameters θ) of ResNet layer j associated with stochastic layer i . The feature maps are written as $d_{i,j}$. The generative process can then be iterated as $z_L \sim \mathcal{N}(0, I)$ and $i = L-1, L-2, \dots, 1$:

$$d_{i,0} = z_{i+1} \quad (4)$$

$$d_{i,j} = \langle f_{\theta_{i,j}}(d_{i,j-1}); d_{i+1,j} \rangle \text{ for } j = 1, \dots, M \quad (5)$$

$$z_i = \mu_{\theta,i}(d_{i,M}) + \sigma_{\theta,i}(d_{i,M}) \otimes \epsilon_i, \quad (6)$$

where $d_{L,j} = 0$, $\langle \cdot \rangle$ denotes concatenation of feature maps in the convolutional network and hidden units in the fully connected network, $\epsilon \sim \mathcal{N}(0, I)$ and $\mu(\cdot)$ and $\sigma(\cdot)$ are parameterized by neural networks. To complete the generative model $p(x|\mathbf{z})$ is written in terms of z_1 and d_1 through a ResNet block f_0 .

Inference model. The inference model consists of a bottom-up (BU) and top-down (TD) paths such that bottom-up stochastic units only receive bottom-up information whereas the top-down units receive both bottom-up and top-down information. The top-down path shares parameters with the generative model. For each stochastic latent variable z_i in $i = 1, \dots, L$ we use a ResNet block with M layers and there are associated neural network functions $g_{i,j}$, $j = 1, \dots, M$ with parameters collectively denoted by ϕ . The deterministic feature map of layer i, j is denoted by $\tilde{d}_{i,j}$:

$$\tilde{d}_{i,0} = \begin{cases} x & i = 1 \\ \langle z_{i-1}; \tilde{d}_{i-1,M} \rangle & \text{otherwise} \end{cases} \quad (7)$$

$$\tilde{d}_{i,j} = \langle g_{i,j}(\tilde{d}_{i,j-1}); \tilde{d}_{i-1,j} \rangle \text{ for } j = 1, \dots, M, \quad (8)$$

$$z_i^{\text{BU}} = \mu_i^{\text{BU}}(\tilde{d}_{i,M}) + \sigma_i^{\text{BU}}(\tilde{d}_{i,M}) \otimes \epsilon_i^{\text{BU}} \quad (9)$$

where $\epsilon \sim \mathcal{N}(0, I)$. Finally, to infer the top-down latent we use the bottom-up latent z_i^{TD} inferred in eq. (9) and pass them through the generative path eq. (5) for $i = L-1, L-2, \dots, 2$ to determine $d_{i,M}$ and

$$z_i^{\text{TD}} = \mu_i^{\text{TD}}(\langle \tilde{d}_{i,M}; d_{i,M} \rangle) + \sigma_i^{\text{TD}}(\langle \tilde{d}_{i,M}; d_{i,M} \rangle) \otimes \epsilon_i^{\text{TD}}. \quad (10)$$

B. Semi-Supervised Learning

When defining BIVA for semi-supervised classification tasks we follow the approach described for the M2 model in Kingma et al. (2014). In addition to BIVA, described in detail in Appendix A, we introduce a classification model $q_\phi(y|x, z_{<L}^{\text{BU}})$ in the inference model, where y is the class variable, and a Categorical latent variable dependency in the generative model.

Inference model. For the classification model we introduce another deterministic hierarchy with an equivalent parameterization as $\tilde{d}_{i,1}, \dots, \tilde{d}_{i,M}$. We denote the hierarchy $\tilde{d}_{i,1}^c, \dots, \tilde{d}_{i,M}^c$. The forward-pass is performed by:

$$\tilde{d}_{i,0}^c = \begin{cases} x & i = 1 \\ \tilde{d}_{i-1,M}^c & \text{otherwise} \end{cases} \quad (11)$$

$$\tilde{d}_{i,j}^c = \langle g_{\phi_{i,j}}^c(\tilde{d}_{i,j-1}^c); z_i^{\text{BU}} \rangle \text{ for } j = 1, \dots, M \quad (12)$$

$$y = g_{\phi_{i,M+1}}^c(\tilde{d}_{i,M}^c), \quad (13)$$

where $g_{\phi_{i,M+1}}^c$ is a final densely connected neural network layer, of the same dimension as the number of categories, and a Softmax activation function. The inference model is thereby factorized by:

$$q_\phi(\mathbf{z}, y|x) = q_\phi(z_L|x, y, z_{<L}^{\text{BU}}) q_\phi(y|x, z_{<L}^{\text{BU}}) \prod_{i=1}^{L-1} q_\phi(z_i^{\text{BU}}|x, z_{<i}^{\text{BU}}) q_{\phi,\theta}(z_i^{\text{TD}}|x, y, z_{<i}^{\text{BU}}, z_{>i}^{\text{BU}}, z_{>i}^{\text{TD}}). \quad (14)$$

Generative model. For each stochastic latent variable, \mathbf{z} , and the observed variable x in the generative model, as well as the TD path of the inference model, we add a conditional dependency on a categorical variable y :

$$p_{\theta}(x, y, \mathbf{z}) = p_{\theta}(x|\mathbf{z}, y)p_{\theta}(z_L)p_{\theta}(y) \prod_{i=1}^{L-1} p_{\theta}(z_i|z_{>i}, y). \quad (15)$$

Evidence lower bound. In a semi-supervised learning problem, we have labeled data and unlabeled data which results in two formulations of the ELBO. The ELBO for labeled data points is given by:

$$\log p_{\theta}(x, y) \geq \mathbb{E}_{q_{\phi}(\mathbf{z}|x, y)} \left[\log \frac{p_{\theta}(x, y, \mathbf{z})}{q_{\phi, \theta}(\mathbf{z}|x, y)} \right] \equiv -\mathcal{F}(\theta, \phi). \quad (16)$$

Since the classification model is not included in the above definition of the ELBO we add a classification loss term (a categorical cross-entropy), equivalent to the approach in Kingma et al. (2014):

$$\bar{\mathcal{F}}(\theta, \phi) = \mathcal{F}(\theta, \phi) - \alpha \cdot \mathbb{E}_{q(z < L|x)} [\log q_{\phi}(y|x, z_{< L}^{\text{BU}})], \quad (17)$$

where α is a hyperparameter that we define as in Maaløe et al. (2016). For the unlabeled data points, we marginalize over the labels:

$$\log p_{\theta}(x) \geq \mathbb{E}_{q_{\phi}(\mathbf{z}, y|x)} \left[\log \frac{p_{\theta}(x, y, \mathbf{z})}{q_{\phi, \theta}(\mathbf{z}, y|x)} \right] \equiv -\mathcal{U}(\theta, \phi). \quad (18)$$

The combined objective function over the labeled, (x_l, y_l) , and unlabeled data points, (x_u) , are thereby given by:

$$\mathcal{J}(\theta, \phi) = \sum_{x_l, y_l} \bar{\mathcal{F}}(\theta, \phi; x_l, y_l) + \sum_{x_u} \mathcal{U}(\theta, \phi; x_u). \quad (19)$$

C. Experimental Setup

Throughout all experiments we follow the BIVA model description that is described in detail in Appendix A and B.

Optimization. All models are optimized using Adamax (Kingma & Ba, 2014) with a hyperparameter setting similar to the one used in (Kingma et al., 2016). They are trained with a batch-size of 48 where the binary image experiments are trained on a single GPU and the natural image experiments are trained on two GPUs (by splitting the batch in 2 and then taking the mean over the gradients). For evaluation we use exponential moving averages of the parameters space, similar to Kingma et al. (2016); Salimans et al. (2017).

Binary image architecture. BIVA has $L = 6$ layers. The g_{ϕ_1} neural networks are defined by $M = 3, 64 \times 5 \times 5$ (number of kernels x kernel width x kernel height) convolutional layers and an overall stride of 2. Neural networks $i = 2, \dots, 6$ are defined by four $M = 3, 64 \times 3 \times 3$ convolutional layers. The final neural network, $i = 6$, applies a stride of 2. All stochastic latent variables are densely connected layers of dimension 48, 40, 32, 24, 16, 8 for $1, \dots, L$ respectively. We apply a dropout rate of 0.5 for both the deterministic layers in the generative as well as the inference model.

Natural image architecture (32x32). BIVA has $L = 15$ layers. The g_{ϕ_1} neural networks are defined by $M = 3, 96 \times 5 \times 5$ convolutional layers and an overall stride of 2. Neural networks $i = 2, \dots, 15$ are defined by $M = 3, 96 \times 3 \times 3$ convolutional layers. Neural networks 11 and 15 are defined with a stride of 2. All stochastic latent variables are parameterized by convolutional layers with 38, 36, 34, ..., 10 feature maps for $1, 2, 3, \dots, L$ respectively. The kernel width and height of the stochastic latent variables are defined similarly to the dimension of the subsequent output after striding. We apply a dropout rate of 0.2 in the deterministic layers of the inference model.

Natural image architecture (64x64). BIVA has $L = 20$ layers. The g_{ϕ_1} and g_{ϕ_2} neural networks are defined by $M = 3, 64 \times 7 \times 7$ and $64 \times 5 \times 5$ convolutional layers respectively with a stride of 2 in each. Neural networks $i = 3, \dots, 11$ are defined by $M = 3, 64 \times 3 \times 3$ convolutional layers. Neural network 11 is defined with a stride of 2. Neural networks $i = 12, \dots, 20$ are defined by $M = 3, 128 \times 3 \times 3$ convolutional layers and network 20 has a stride of 2. All stochastic latent variables are parameterized by convolutional layers with 20, 19, 18, ..., 1 feature maps for $1, 2, 3, \dots, L$ respectively. The kernel width and height of the stochastic latent variables are defined similarly to the dimension of the subsequent output after striding. We apply a dropout rate of 0.2 in the deterministic layers of the inference model.

D. Initial Results on Text Generation Tasks

	PARAMETERS	$-\log p(x)$	KL	PPL
<i>Results with autoregressive components, no dropout</i>				
LSTM	15.0M	= 41.49	—	36.28
RNN-VAE (Bowman et al., 2015), \mathcal{L}_1 , WARMUP	23.7M	≤ 42.09	1.61	38.21
RNN-VAE (Bowman et al., 2015), \mathcal{L}_1 , FINETUNED	23.7M	≤ 42.41	5.13	39.26
HYBRID VAE (Semeniuta et al., 2017), \mathcal{L}_1 , FINETUNED	23.7M	≤ 42.24	4.67	38.70
BIVA L=7, \mathcal{L}_1, FINETUNED	23.0M	≤ 42.34	10.15	39.04
<i>Results without autoregressive components, no dropout</i>				
HYBRID VAE (Semeniuta et al., 2017), \mathcal{L}_1 , FINETUNED	15.0M	≤ 54.53	14.10	112.1
BIVA L=7 FINETUNED, \mathcal{L}_1	14.0M	≤ 54.13	15.33	108.3

Table 7. Test performances on the BookCorpus with 1 importance weighted sample (sentences limited to 40 words). The RNN-VAE and Hybrid VAE are trained and evaluated from our own implementation.

Optimizing generative models coupled with autoregressive models is a difficult task. Such coupling causes the posterior to collapse, and the latent variables are ignored. Nonetheless, autoregressive components remain a cornerstone of the generative models for text (Bowman et al., 2015; Semeniuta et al., 2017; Shah et al., 2018). In order to enforce the model to use the latent variable, previous efforts aim at weakening the decoder using powerful regularizing *tricks*, such as word dropout (Bowman et al., 2015). We investigate the use of BIVA in the context of sentence modelling without weakening the decoder. We show that it allows optimizing the latent variables more effectively, resulting in a higher measured KL when compared to the RNN-VAE (Bowman et al., 2015) and the Hybrid VAE (Semeniuta et al., 2017).

Dataset. We use the Bookcorpus dataset (Zhu et al., 2015) of sentences of maximum 40 words, no preprocessing is performed and sentences are tokenized using the white spaces. We defined a vocabulary of 20000 words and filtered out the sentences that contain non-indexed tokens. We randomly sampled 10000 sentences for testing and used the remaining 56M sentences for training.

Models. We couple BIVA with an LSTM decoder, using the output of the convolutional model as an input sequence for the auto-regressive model. We compare our model against a LSTM language model (Hochreiter & Schmidhuber, 1997), the RNN-VAE (Bowman et al., 2015), and the Hybrid VAE (Semeniuta et al., 2017), which couples a convolutional architecture with an LSTM decoder. We also perform experiments without using autoregressive components.

All LSTM models are parameterized by 1024 units and we use embeddings of dimension 512. This results in an RNN-VAE model with 23.7M parameters and we limit the other models to use the same total number of parameters. This results in using a limited number of stochastic layers for the BIVA and small a small number of kernels of 128.

Training. We trained the models for 5 epochs with an initial learning rate of $2e^{-3}$ using the Adamax optimizer. We used batches of size 512 and used only one stochastic sample. We train all latent variable models using the free bits method from (Kingma et al., 2016) with an initial KL budget of 30 nats distributed equally over the stochastic variables and we incrementally decrease the free bits value *on plateau*. We also train the RNN-VAE baseline using the deterministic warmup method (Bowman et al., 2015; Sønderby et al., 2016) for comparison.

Likelihood and latent variables usage. We report the test set results in table 7 and test samples in 8 and reconstructions in table 9. While BIVA without the autoregressive decoder is not competitive with an LSTM language model, we observe that replacing the LSTM inference model by a BIVA model allows exploiting the latent space more actively, which results in a higher measured KL than the RNN-VAE and Hybrid VAE baselines.

BIVA+LSTM	RNN-VAE
<p>he said .</p> <p>i tried to think of something to say to him , but he was already on his way back to the house . it sounded as if he was going to say something .</p> <p>“ and that ‘ s why you ‘ re coming . ”</p> <p>“ what ? ”</p> <p>she swallowed .</p> <p>“ i want you . ”</p> <p>glancing up , i saw the way he was staring at me with a look of pure hatred .</p> <p>i need a favor . ”</p> <p>he did n’t .</p> <p>you ‘ re not dead .</p> <p>i stood , and he followed .</p> <p>“ can i sit on the couch and talk ? ”</p> <p>“ it was n’t until i was fifteen , i was n’t in the mood to be around .</p> <p>i looked down at my lap .</p> <p>the smile disappeared .</p> <p>it was hard to tell which one was more of a rock .</p> <p>i ‘ m not sure it ‘ s a good idea .</p> <p>the first two .</p> <p>he was there .</p> <p>“ all of you , ” joe said .</p> <p>he did n’t care if he was n’t a vampire .</p> <p>her mouth curved up , then she nodded .</p> <p>just tell me what you want in the end .</p> <p>and again .</p> <p>the other man ‘ s voice was hoarse and ragged .</p> <p>i had n’t known that was a bad idea , but i had n’t been able to get it out of my head .</p> <p>your brother is the most important thing to me .</p> <p>you dont need to go to the police , right ?</p> <p>there was a long silence .</p> <p>i looked up .</p> <p>he nodded , and he looked at me , and i could tell he was thinking about it .</p> <p>“ hang on , baby .</p> <p>we had to be close to the city , and we could n’t afford to be here .</p> <p>you know , it would be better if you were n’t so stupid . ”</p> <p>excuse me ?</p> <p>you know how much i love you , too .</p> <p>a woman ‘ s voice , a voice that was familiar .</p> <p>i have a very important business to attend to , and i ‘ m going to have to make a decision .</p> <p>they sat on the small wooden table in the center of the room .</p> <p>“ it ‘ s fine . ”</p> <p>she felt a rush of relief .</p> <p>maria , he says .</p> <p>what ?</p> <p>“ it does n’t seem like a lot to me , ” he said .</p> <p>he ‘ d told her everything .</p> <p>“ she ‘ s in shock . ”</p> <p>“ after all , ” he murmured , “ i ‘ m going to go get the rest of the stuff . ”</p> <p>and then , finally , she ‘ d done it .</p> <p>her words were a whisper , but it was n’t enough .</p>	<p>“ two .</p> <p>“ you do n’t have to do this . ”</p> <p>the light from the lamp was dim , but the light was dim and the room was dark . or a nuclear bomb , or something .</p> <p>“ the baby ? ”</p> <p>“ you ‘ re not going to kill me . ”</p> <p>she was n’t going to .</p> <p>“ i guess we could have been more careful , ” he said .</p> <p>there are some things that are not good .</p> <p>“ you ‘ re a good man . ”</p> <p>i had n’t been able to get it out .</p> <p>“ you ‘ re going to have to be careful , ” he said .</p> <p>it ‘ s not a bad idea .</p> <p>he asked .</p> <p>“ this is a bad idea , ” he said , his voice a little hoarse .</p> <p>“ i ‘ m sure he ‘ s in love with you . ”</p> <p>as he stepped out of the car , he saw the man standing in the doorway , his eyes wide and his face pale .</p> <p>“ no . ”</p> <p>“ in the meantime , i need to get some sleep , ” i said .</p> <p>i was n’t .</p> <p>did i want to talk to you ?</p> <p>“ i want to hear you say it . ”</p> <p>the train was already in the driveway .</p> <p>“ good . ”</p> <p>i just needed to get out of here , and i needed to get out of here .</p> <p>“ this is a good idea . ”</p> <p>“ hey . ”</p> <p>she took a deep breath and let it out .</p> <p>then he kissed her .</p> <p>i felt a warm hand on my shoulder and a warm smile spread across my face .</p> <p>“ he ‘ s dead . ”</p> <p>at the time , i was going to have to get out of the house .</p> <p>he was so close to the edge of the bed .</p> <p>“ i do n’t know . ”</p> <p>“ i do n’t have a choice . ”</p> <p>i know i ‘ m not going to let him touch me , but i do .</p> <p>i could n’t see the face of the man who ‘ d just been in the doorway .</p> <p>in the end , we all know that we are not going to be able to get out of this .</p> <p>“ yes . ”</p> <p>“ what are you doing here ? ”</p> <p>so the only thing that mattered was that he was here .</p> <p>neither of them spoke .</p> <p>from now on , you will be able to get out of here .</p> <p>the thought of having to kill him made him want to kill her .</p> <p>the other two were staring at me , their eyes wide .</p> <p>i did n’t want to be a part of it , but i was n’t going to let it go .</p> <p>“ i do n’t want to talk about it . ”</p> <p>she looked at him , her eyes wide .</p> <p>“ that ‘ s what you ‘ re going to do . ”</p>

Table 8. Samples decoded from the prior of the BIVA with LSTM decoder and baseline RNN-VAE.

Input	BIVA-LSTM	RNN-VAE
" a sad song , being sung alone in the basement . "	" it sounds like you 've been through a lot . "	" you 're going to be a great father . "
more often , though , wherever she sank , beck was there .	more than anything , she wanted to be with him .	in the end , we all knew what was going on .
he looked just about as pale as i had ever seen him .	he 's still a lot more than a friend .	he was n't going to let her go .
caleb turned and shoved him back as he took his true form .	he lifted me up , his arms still wrapped around my waist .	he was standing in the doorway , his hands folded in front of him .
i gasped , tried to pull away , squeezed my legs together .	i gasped , and he wasn't able to stop himself .	i felt my body tense , and i could n't help but smile .
agreed as i adjusted myself and sat heavily in my chair .	i tried to ignore it , but my eyes were still closed .	i did n't want to be the one to tell him .
you bind me , UNK in darkness , though , in light .	he decided to take her home , to make her feel safe .	he was more than willing to let her go .
they promise me things , ask me questions , whisper and plead .	they might be able to do something about it , but they don't .	we need to talk , " he said , his voice low .
i glowered as i held the bear , almost bigger than me .	i started to close my eyes , but he was too strong .	i could n't help but smile at the sight of her .
i wondered how much he paid them to be his guard dogs .	i had to admit that it was n't a good idea .	i don't want to be a part of this .
" hummm , " richard muttered , and headed up the path .	" jeesus , " he said , his voice barely audible .	" but you 're going to be a father .
he was happy that he had found it in the UNK hall .	he was n't going to be the one to go .	he was it 's sure if he was going to make it .
at the shack , at the condo , at the hangar .	at the moment , the only thing that mattered was that he was n't alone .	he was staring at the floor , his eyes wide .
" i 'd pop to go to the dance with you . "	" i 'd prefer to go to the hospital .	i 'm going to go to the bathroom .
someday . i 'll share them with the rest of the world .	and now i have a lot of my own .	we 're going to have to do something about it .
" maybe i 'm not the right person for this one . "	" maybe we can get a little more of a ride . "	i don't think you 're going to be able to do that .
" gin is my sister , and she 's coming with me .	" there 's a chance i can get a little more sleep . "	" if you want to , i 'll be there . "
thick desire stormed her ... along with a bittersweet curl of emotion .	the tension was gone , and he was n't looking at me .	the air smelled of stale cigarette smoke .
they caused him to stagger back and drove him to the ground .	they had to be at the top of the hill .	he 'd found a way to get her to safety .
you 're not much of a friar , friar , he says .	you 're not supposed to be around here , are you ? "	you 're not going to be able to do that , are you ? "

Table 9: Reconstruction of samples from the test set using the BIVA with LSTM decoder and the RNN-VAE baseline. The samples are decoded from the posterior distribution by using greedy decoding, without teacher forcing.

E. Additional Results

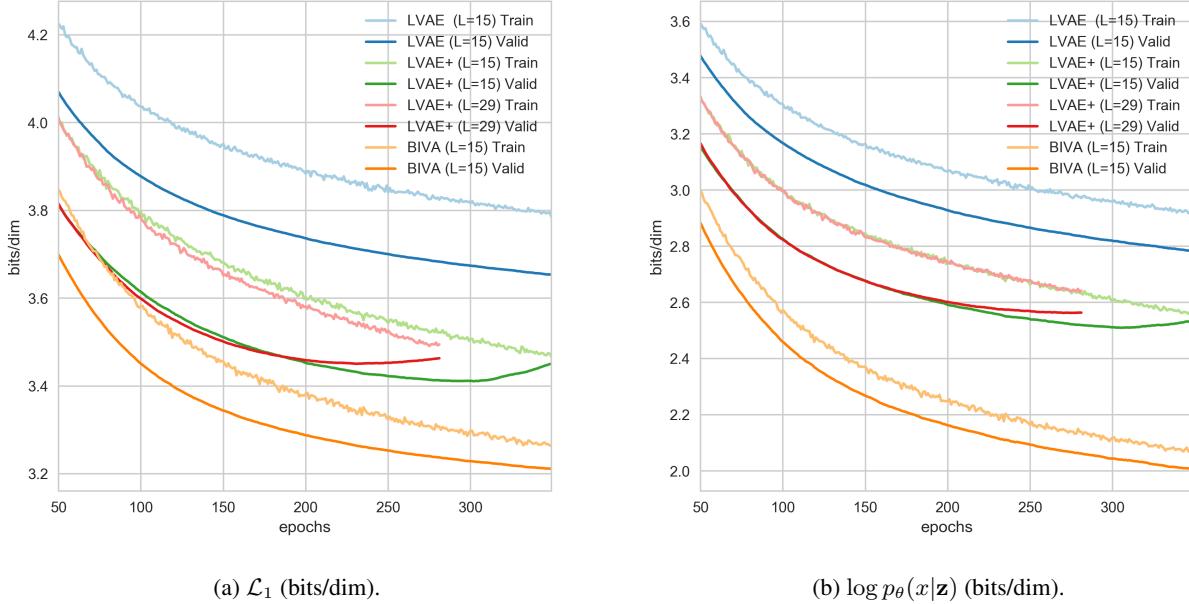


Figure 7. Convergence plot on CIFAR-10 training for the LVAE with $L = 15$, the LVAE+ with $L = 15$, the LVAE+ with $L = 29$, and BIVA with $L = 15$. (a) shows the convergence of the 1 importance weighted ELBO, \mathcal{L}_1 , calculated in bits/dim. (b) shows the convergence of the reconstruction loss. The discrepancy between (a) and (b) is explained by the added cost from the stochastic latent variables, the Kullback-Leibler divergence $KL[p(\mathbf{z})||q(\mathbf{z}|x)]$.

	$-\log p(x)$
<i>Results with autoregressive components</i>	
DRAW+VGP (TRAN ET AL., 2016)	< 79.88
IAFVAE (KINGMA ET AL., 2016)	≤ 79.10
VLAЕ (CHEN ET AL., 2017)	≤ 78.53
<i>Results without autoregressive components</i>	
IWAE (BURDA ET AL., 2015A)	≤ 82.90
CONVAE+HVI (SALIMANS ET AL., 2015)	≤ 81.94
LVAE (SØNDERBY ET AL., 2016)	≤ 81.74
DISCRETE VAE (ROLFE, 2017)	≤ 80.04
BIVA, \mathcal{L}_1	≤ 80.60
BIVA, \mathcal{L}_{1e3}	≤ 78.49
BIVA FINETUNED, \mathcal{L}_1	≤ 80.06
BIVA FINETUNED, \mathcal{L}_{1e3}	≤ 78.41

Table 10. Test log-likelihood on dynamically binarized MNIST for different number of importance weighted samples. The finetuned models are trained for an additional number of epochs with no *free bits*, $\lambda = 0$.

	$-\log p(x)$
<i>Results with autoregressive components</i>	
DRAW GREGOR ET AL. (2015)	< 96.50
CONVDRAW (GREGOR ET AL., 2016)	≤ 91.00
VLAЕ (CHEN ET AL., 2017)	≤ 89.83
<i>Results without autoregressive components</i>	
IWAE (BURDA ET AL., 2015A)	≤ 103.38
LVAE (SØNDERBY ET AL., 2016)	≤ 102.11
DVAE (ROLFE, 2017)	≤ 97.43
BIVA, \mathcal{L}_1	≤ 95.90
BIVA FINETUNED, \mathcal{L}_1	≤ 93.54
BIVA FINETUNED, \mathcal{L}_{1e3}	≤ 91.34

Table 11. Test log-likelihood on dynamically binarized OMNIGLOT for different number of importance weighted samples. The finetuned models are trained for an additional number of epochs with no *free bits*, $\lambda = 0$.

	$-\log p(x)$
BIVA, \mathcal{L}_1	≤ 94.05
BIVA FINETUNED, \mathcal{L}_1	≤ 93.54
BIVA FINETUNED, \mathcal{L}_{1e3}	≤ 87.98

Figure 7. Table 1: Test log-likelihood on statically binarized Fashion MNIST for different number of importance weighted samples. The finetuned models are trained for an additional number of epochs with no *free bits*, $\lambda = 0$.



Figure 8. BIVA $\mathcal{N}(0, \sigma^2)$ generations with varying $\sigma^2 = 0.01, 0.1, 0.5, 1.0$ for (a), (b), (c) and (d) respectively. We follow the same generating procedure of Figure 1. BIVA has $L = 20$ stochastic latent variables and is trained on the CelebA dataset, preprocessed to 64x64 images following Larsen et al. (2016). BIVA achieves a $\mathcal{L}_1 = 2.48$ bits/dim on the test set. Close to the mode of the latent distribution there is very little variance in generated natural images. When we *loosen* the samples towards the full distribution, $\sigma^2 = 1$, we can see how the generated images are adopting different styles and contexts.

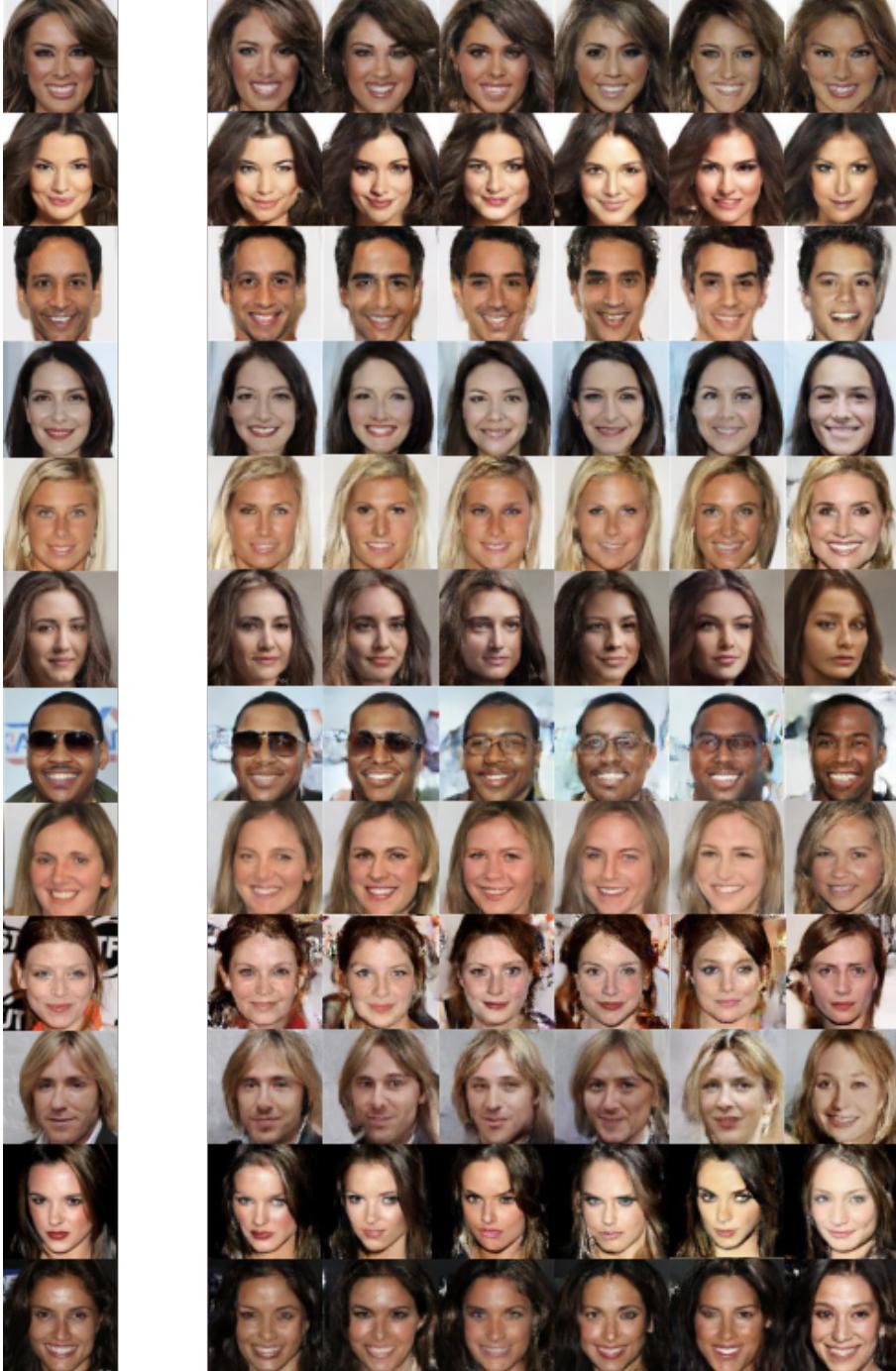


Figure 9. BIVA $L = 20$ generations (right) from fixed $z_{>i}$ given an input image (left), for different layers throughout the stochastic variable hierarchy (from left to right $i = 12, 14, 16, 17, 18, 19$). The model is trained on CelebA, preprocessed to 64x64 images following Larsen et al. (2016). $z_{>i}$ are fixed by passing the original image through the encoder, after which $z_{\leq i}$ are sampled from the prior. When generating from a higher z_i (columns) it is shown how the model has more *freedom* to augment the input images. BIVA achieves a $\mathcal{L}_1 = 2.48$ bits/dim on the test set.



Figure 10. BIVA $\mathcal{N}(0, I)$ generations on a model trained on CIFAR-10. BIVA has $L = 15$ stochastic latent variables and achieves a 3.08 bits/dim on the test set. The images are still not as sharp and coherent as the PicelCNN++ (Salimans et al., 2017) (3.08 vs. 2.92), however, it does achieve to find coherent structure resembling the categories of the CIFAR-10 dataset.