Avoiding Latent Variable Collapse with Generative Skip Models

Adji B. Dieng
Columbia University

Yoon Kim
Harvard University

Alexander M. Rush
Harvard University

David M. Blei
Columbia University

Abstract

Variational autoencoders (VAEs) learn distributions of high-dimensional data. They model data with a deep latent-variable model and then fit the model by maximizing a lower bound of the log marginal likelihood. VAEs can capture complex distributions, but they can also suffer from an issue known as "latent variable collapse," especially if the likelihood model is powerful. Specifically, the lower bound involves an approximate posterior of the latent variables; this posterior "collapses" when it is set equal to the prior, i.e., when the approximate posterior is independent of the data. While VAEs learn good generative models, latent variable collapse prevents them from learning useful representations. In this paper, we propose a simple new way to avoid latent variable collapse by including skip connections in our generative model; these connections enforce strong links between the latent variables and the likelihood function. We study generative skip models both theoretically and empirically. Theoretically, we prove that skip models increase the mutual information between the observations and the inferred latent variables. Empirically, we study images (MNIST and Omniglot) and text (Yahoo). Compared to existing VAE architectures, we show that generative skip models maintain similar predictive performance but lead to less collapse and provide more meaningful representations of the data.

1 Introduction

Unsupervised representation learning aims to find good low-dimensional representations of high-dimensional data. One powerful method for representation learning is the variational autoencoder (VAE) [Kingma and Welling, 2013, Rezende et al., 2014]. VAEs have been studied for text analysis [Bowman et al., 2015, Miao et al., 2016, Dieng et al., 2016, Guu et al., 2017, Xu and Durrett, 2018], collaborative filtering [Liang et al., 2018], dialog modeling [Zhao et al., 2018], image analysis [Chen et al., 2016, van den Oord et al., 2017], and many other applications.

A VAE binds together modeling and inference. The model is a deep generative model, which defines a joint distribution of latent variables $\boldsymbol{z}$ and observations $\boldsymbol{x}$,

$$p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z}).$$

A typical VAE uses a spherical Gaussian prior $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and a likelihood parameterized by a deep neural network. Specifically, the likelihood of observation $\boldsymbol{x}_i$ is an exponential family whose natural parameter $\eta(\boldsymbol{z}_i; \theta)$ is a deep network with the latent representation $\boldsymbol{z}_i$ as input. Inference in VAEs is performed with variational methods.

VAEs are powerful, but they can suffer from a phenomenon known as *latent variable collapse* [Bowman et al., 2015, Hoffman and Johnson, 2016, Sønderby et al., 2016, Kingma et al., 2016, Chen et al., 2016, Zhao et al., 2017, Yeung et al., 2017, Alemi et al., 2018], in which the variational posterior collapses to the prior. When this phenomenon occurs, the VAE can learn a good generative model of the data but still fail to learn good representations of the individual data points. We propose a new way to avoid this issue.

Ideally the parameters of the deep generative model should be fit by maximizing the marginal likelihood of the observations,

$$\theta^* = \arg\max_\theta \sum_{i=1}^{N} \log \int p_\theta(\boldsymbol{x}_i, \boldsymbol{z}_i) d\boldsymbol{z}. \tag{1}$$

However each term of this objective contains an intractable integral. To this end, VAEs rely on amortized variational inference to approximate the posterior distribution. First posit a variational approximation $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i)$. This is an *amortized* family, a distribution over latent variables $\boldsymbol{z}_i$ that takes the observation $\boldsymbol{x}_i$ as input and uses a deep neural network to produce variational parameters. Using this family, the VAE

Primary contact: Adji B. Dieng abd2141@columbia.edu.

optimizes the evidence lower bound (ELBO),

$$\text{ELBO} = \sum_{i=1}^{N} E_{q_\phi(\boldsymbol{z}_i \mid \boldsymbol{x}_i)} \left[\log p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z}_i)\right] - \text{KL}(q_\phi(\boldsymbol{z}_i \mid \boldsymbol{x}_i) \,\|\, p(\boldsymbol{z}_i)). \quad (2)$$

The ELBO is a lower bound on the log marginal likelihood of Eq. 1. Thus the VAE optimizes Eq. 2 with respect to both the generative model parameters θ and the variational neural network parameters ϕ. Fitting θ finds a good model; fitting ϕ finds a neural network that produces good approximate posteriors.

This method is theoretically sound. Empirically, however, fitting Eq. 2 often leads to a degenerate solution where

$$q_\phi(\boldsymbol{z}_i \mid \boldsymbol{x}_i) \approx p(\boldsymbol{z}_i),$$

i.e. the variational "posterior" does not depend on the data; this is known as latent variable collapse. When the posterior collapses, $\boldsymbol{z}$ and $\boldsymbol{x}$ are essentially independent and consequently posterior estimates of the latent variable $\boldsymbol{z}$ do not represent faithful summaries of their data $\boldsymbol{x}$—the VAE has not learned good representations. This issue is especially a problem when the likelihood $p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$ has high capacity [Bowman et al., 2015, Sønderby et al., 2016, Kingma et al., 2016, Chen et al., 2016, Zhao et al., 2017, Yeung et al., 2017].

We propose a new method to alleviate latent variable collapse. The idea is to add skip connections in the deep generative model that parameterizes the likelihood function. Skip connections attach the latent input $\boldsymbol{z}_i$ to multiple layers in the model's neural network. The resulting *generative skip model* is at least as expressive as the original deep generative model, but it forces the likelihood to maintain a strong connection between the latent variables $\boldsymbol{z}_i$ and the observations $\boldsymbol{x}_i$. Consequently, as we show, posterior estimates of $\boldsymbol{z}_i$ provide good representations of the data.

VAEs with generative skip models—which we call Skip Variational Autoencoders (SKIP-VAEs)—produce both good generative models and good representations. Section 4 studies the traditional VAE [Kingma and Welling, 2013, Rezende et al., 2014] with PixelCNN/LSTM generative models analyzing both text and image datasets. For similar levels of model performance, as measured by the approximate likelihood, SKIP-VAEs promote more dependence between $\boldsymbol{x}$ and $\boldsymbol{z}$ as measured by mutual information and other metrics in Section 4. Moreover, the advantages of SKIP-VAEs increase as the generative model gets deeper.

Generative skip models can be used in concert with other techniques. For example Section 4 also studies generative skip models with the semi-amortized variational autoencoder (SA-VAE) [Kim et al., 2018][1], which have also been shown to mitigate posterior collapse. When used with the SA-VAE, generative skip models further improve the learned representations.

[1] resulting in the SKIP-SA-VAE

Related Work. Skip connections are widely used in deep learning, for example, in designing residual, highway, and attention networks [Fukushima, 1988, He et al., 2016b, Srivastava et al., 2015, Bahdanau et al., 2014]. They have not been studied for alleviating latent variable collapse in VAEs.

Many papers discuss latent variable collapse [Bowman et al., 2015, Hoffman and Johnson, 2016, Sønderby et al., 2016, Kingma et al., 2016, Chen et al., 2016, Zhao et al., 2017, Yeung et al., 2017, Alemi et al., 2018]. To address it, the most common heuristic is to anneal the KL term in the VAE objective [Bowman et al., 2015, Sønderby et al., 2016].

Several other solutions have also been proposed. One approach is to handicap the training of the generative model [Bowman et al., 2015] or weaken its capacity [Gulrajani et al., 2016, Yang et al., 2017], effectively encouraging better representations by limiting the generative model. Another approach replaces the simple spherical Gaussian prior with more sophisticated priors. For example van den Oord et al. [2017] and Tomczak and Welling [2017] propose parametric priors, which are learned along with the generative model. Still another approach uses richer variational distributions [Rezende and Mohamed, 2015]. In another thread of research, Makhzani et al. [2015] and Mescheder et al. [2017] replace the KL regularization term in the VAE objective with adversarial regularizers. Higgins et al. [2017] dampen the effect of the KL regularization term with Lagrange multipliers. Finally, one can appeal to new inference algorithms. For example Hoffman [2017] uses Markov chain Monte Carlo (MCMC) instead of variational inference and Kim et al. [2018] uses stochastic variational inference, initialized with the variational neural network parameters, to iteratively refine the variational distribution.

A very recent approach to address posterior collapse relies on ideas from directional statistics. Guu et al. [2017] and Xu and Durrett [2018] use the Von Mises-Fisher distribution for both the prior and the variational posterior and fixing the dispersion parameter of the Von Mises-Fisher distribution to make the KL term in the ELBO constant . We note this practice might result in less expressive approximate posteriors.

The generative skip models we propose differ from all of these strategies and can potentially complement them. They modify the generative model of the VAE using skip connections to enforce dependence between the observations and their latent variables. They can be used with any prior and variational distributions.

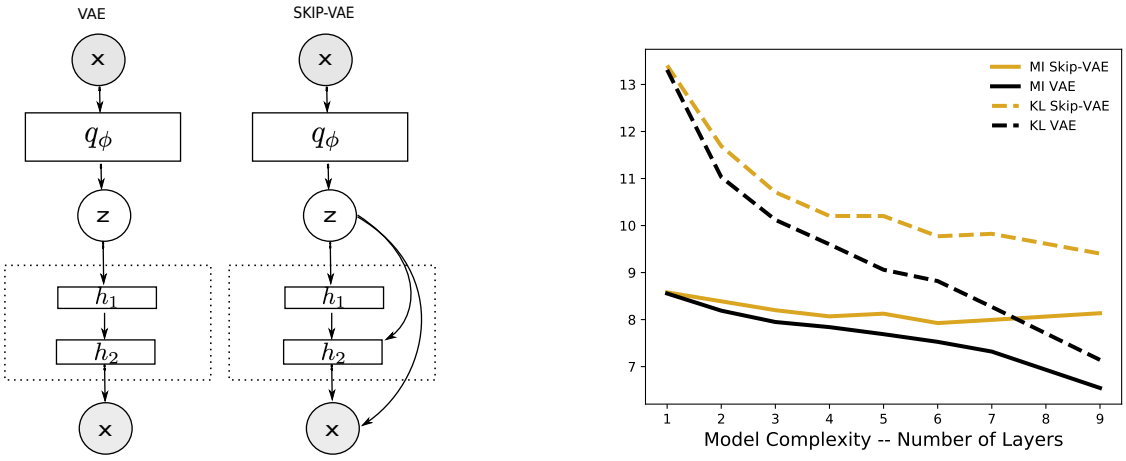

Figure 1: Left: The VAE and SKIP-VAE with a two-layer generative model. The function q_ϕ denotes the variational neural network (identical for VAE and SKIP-VAE). The difference is in the generative model class: the SKIP-VAE's generative model enforces residual paths to the latents at each layer. **Right:** The mutual information induced by the variational distribution and KL from the variational distribution to the prior for the VAE and the SKIP-VAE on MNIST as we vary the number of layers L. The SKIP-VAE leads to both higher KL and higher mutual information.

2 Latent variable collapse issue in VAEs

As we said, the VAE binds a deep generative model and amortized variational inference. The deep model generates data through the following process. First draw a latent variable $\boldsymbol{z}$ from a prior $p(\boldsymbol{z})$; then draw an observation $\boldsymbol{x}$ is $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$, the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{z}$. This likelihood is an exponential family distribution parameterized by a deep neural network,

$$\begin{aligned} p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) &= \text{ExpFam}\left(\boldsymbol{x}; \eta\left(\boldsymbol{z}; \theta\right)\right) \\ &= \nu(\boldsymbol{x}) \exp\left\{\eta\left(\boldsymbol{z}; \theta\right)^\top \boldsymbol{x} - A\left(\eta\left(\boldsymbol{z}; \theta\right)\right)\right\}, \end{aligned}$$

where $A(\cdot)$ is the log-normalizer of the exponential family. The exponential family provides a compact notation for many types of data, e.g., real-valued, count, binary, and categorical. We use this notation to highlight that VAEs can model many types of data.

The natural parameter $\eta(\boldsymbol{z}; \theta)$ is a hierarchical function of $\boldsymbol{z}$; see Figure 1 (left). Consider a function with L layers, where $\boldsymbol{h}^{(l)}$ is the hidden state in the l^{th} layer and $\boldsymbol{h}^{(1)}$ is the hidden state closest to $\boldsymbol{z}$. The natural parameter $\eta(\boldsymbol{z}; \theta)$ is computed as follows:

1. $\boldsymbol{h}^{(1)} = f_{\theta_0}(\boldsymbol{z})$
2. $\boldsymbol{h}^{(l+1)} = f_{\theta_l}\left(\boldsymbol{h}^{(l)}\right) \quad l = 1 \ldots L-1$
3. $\eta(\boldsymbol{z}; \theta) = f_{\theta_L}\left(\boldsymbol{h}^{(L)}\right)$.

The parameter θ is the collection $\{\theta_0, \ldots, \theta_L\}$. Given data, it should ideally be fit to maximize the log marginal likelihood; see Eq. 1.

However the integrals in Eq. 1 are intractable. To circumvent this issue, VAEs maximize a lower bound of the log marginal likelihood, also known as the ELBO; see Eq. 2. In the ELBO, $q(\boldsymbol{z} \mid \boldsymbol{x}; \phi)$ is an amortized variational distribution; its parameters are fit so that it approximates the intractable posterior $p(\boldsymbol{z} \mid \boldsymbol{x}; \theta)$. The ELBO is tight when $q(\boldsymbol{z} \mid \boldsymbol{x}; \phi) = p(\boldsymbol{z} \mid \boldsymbol{x}; \theta)$. The objective targets both a good likelihood and a good approximate posterior distribution.

Unfortunately, if the likelihood $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ is too flexible (e.g. a recurrent neural network that fully conditions on all previous tokens), it is difficult to achieve this balance. Consider an equivalent expression for the ELBO,

$$\begin{aligned} \text{ELBO} = &E_{p(\boldsymbol{x})} E_{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right] \\ &- \text{KL}\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z})\right), \end{aligned} \tag{3}$$

where $p(\boldsymbol{x})$ is the population distribution of $\boldsymbol{x}$. The flexible likelihood $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ in the first term allows the VAE to push the KL term to zero (i.e. setting $\text{KL}\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z})\right) \approx 0$) while still giving high probability to the data. This behavior results in a generative model that gives a meaningless approximate posteriors and thus poor latent representations. Chen et al. [2016] theoretically justify latent variable collapse via a "bits-back" argument: if the likelihood model is flexible enough to model the data distribution $p(\boldsymbol{x})$ without using any information from $\boldsymbol{z}$, then the global optimum is indeed obtained by setting $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) = p(\boldsymbol{x})$ and $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = p(\boldsymbol{z})$.

Let's understand this phenomenon from another angle, which will motivate our use of generative skip models. First we define the *variational joint* distribution.

Definition 1 *For any data $\boldsymbol{x}$ and variational posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$, the variational joint $q_\phi(\boldsymbol{x}, \boldsymbol{z})$ is the joint distribution of $\boldsymbol{x}$ and $\boldsymbol{z}$ induced by $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$. It induces a marginal*

$q_\phi(\boldsymbol{z})$ called the aggregated posterior [Makhzani et al., 2015, Mescheder et al., 2017]

$$q_\phi(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}) \cdot q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \quad \textit{and} \quad q_\phi(\boldsymbol{z}) = E_{p(\boldsymbol{x})} q_\phi(\boldsymbol{z} \mid \boldsymbol{x}).$$

With this definition in hand, consider a third form of the ELBO [Hoffman and Johnson, 2016],

$$\begin{aligned} \text{ELBO} = & E_{p(\boldsymbol{x})} \left\{ E_{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) \right] \right\} \\ & - \mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}) - \text{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z})) \end{aligned} \tag{4}$$

We expressed the $\text{KL}\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z})\right)$ of Eq. 3 as a function of a mutual information

$$\text{KL}\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z})\right)) = \mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}) + \text{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z})), \tag{5}$$

where $\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z})$ is defined as

$$\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}) = E_{p(\boldsymbol{x})} E_{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})} \log q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) - E_{q_\phi(\boldsymbol{z})} \log q_\phi(\boldsymbol{z}).$$

It is the mutual information between $\boldsymbol{x}$ and $\boldsymbol{z}$ induced by the variational joint and the aggregated posterior.

The ELBO in Eq. 4 reveals that setting the KL term to zero is equivalent to setting

$$\mathcal{I}_q(\boldsymbol{x}, \boldsymbol{z}) = \text{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z})) = 0.$$

This is true by non-negativity of KL and mutual information. If $\boldsymbol{z}$ is a good representation of $\boldsymbol{x}$, then the mutual information will be high and thus the KL term will be nonzero. But as can be seen from Eq. 4, the ELBO objective contains the negative of the mutual information between $\boldsymbol{z}$ and $\boldsymbol{x}$, and thus high mutual information is in contention with maximizing the ELBO. Of course, our goal is not merely to prevent the KL from being zero—a trivial way to prevent the KL from being zero is by only maximizing

$$\mathcal{L} = E_{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) \right].$$

which essentially corresponds to an auto-encoding objective. However maximizing this objective leads to poor generative models since the variational distribution is unregularized and distinct from the prior—the distribution used to generate samples once training is finished.

We next propose a method that still optimizes the ELBO but prevents the KL from collapsing to zero.

3 Generative skip models avoid latent variable collapse

We now describe SKIP-VAEs, a family of deep generative models that extend VAEs to promote high mutual information between observations and their associated latent variables.

A SKIP-VAE is a modified version of the exponential family model described in Section 2. The natural parameter $\eta(\boldsymbol{z}; \theta)$ is now computed as:

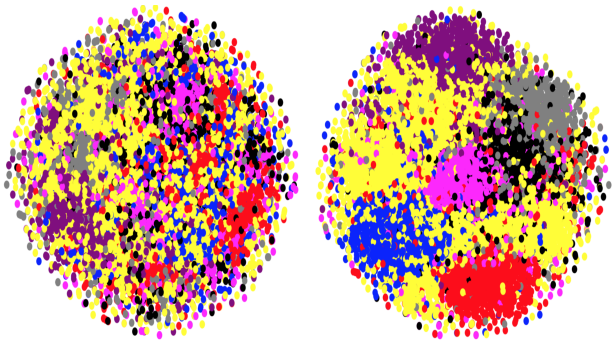

Figure 2: Clustering of the latent variables learned by fitting a VAE (left) and a SKIP-VAE (right) on MNIST and applying T-SNE on the test set. The model is a 9-layer PixelCNN and the variational neural network is a 3-layer ResNet. The colors represent digit labels. The SKIP-VAE clusters the latent variables better than the VAE; it discovers 7 digit classes. The remaining 3 classes are covered by the other classes. The latent variables learned by the VAE are not meaningful as they are spread out. The SKIP-VAE learns more useful latent representations.

1. $\boldsymbol{h}^{(1)} = f_{\theta_0}(\boldsymbol{z})$
2. $\boldsymbol{h}^{(l+1)} = g_{W_l}\left(f_{\theta_l}\left(\boldsymbol{h}^{(l)}\right), \boldsymbol{z}\right)$ for $l = 1 \ldots L-1$
3. $\eta(\boldsymbol{z}; \theta) = g_{W_L}\left(f_{\theta_L}\left(\boldsymbol{h}^{(L)}\right), \boldsymbol{z}\right)$.

At each layer l of the neural network producing $\eta(\boldsymbol{z}; \theta)$, the hidden state $\boldsymbol{h}^{(l)}$ is a function of the latent variable $\boldsymbol{z}$ as well as the previous hidden state. The functions f_{θ_l} are the same as in the non-skip generative model; the separately parameterized *skip* functions g_{W_l} are nonlinear and combine $\boldsymbol{z}$ with the previous hidden state. Figure 1 (left) illustrates this process.

Any type of VAE can be turned into its SKIP-VAE counterpart by adding skips/residual paths to its generative model.

Our main result is that SKIP-VAEs promote higher mutual information between $\boldsymbol{x}$ and $\boldsymbol{z}$ when trained with the ELBO objective in Eq. 2.

SKIP-VAEs are amenable to any type of skip function g; in this section we consider a simple subclass that empirically works well, specifically,

$$g_{W_l}\left(f_{\theta_l}(\boldsymbol{h}^{(l)}), \boldsymbol{z}\right) = \sigma\left(W_l^{(h)} f_{\theta_l}(\boldsymbol{h}^{(l)}) + W_l^{(z)} \boldsymbol{z}\right)$$

where σ is a nonlinear function such as sigmoid or ReLU (not used at the last layer), and $W_l^{(h)} \neq \mathbf{0}$ and $W_l^{(z)} \neq \mathbf{0}$ are learned deterministic weights. Similar to other uses of skip connections [Fukushima, 1988, He et al., 2016b, Srivastava et al., 2015, Bahdanau et al., 2014] we do not need to explicitly enforce the constraints $W_l^{(h)} \neq \mathbf{0}$ and $W_l^{(z)} \neq \mathbf{0}$ in practice.

Theorem 1 *Consider observation $\boldsymbol{x}$ from an L-layer deep generative model. Denote by $\mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z})$ the mutual information between $\boldsymbol{x}$ and $\boldsymbol{z}$ induced by the generative skip model. Similarly denote by $\mathcal{I}_p^{\text{VAE}}(\boldsymbol{x}, \boldsymbol{z})$ the mutual information between $\boldsymbol{x}$ and $\boldsymbol{z}$ induced by the counterpart generative model (the one constructed from the generative skip model by taking the skip connections out). Then*

$$\mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z}) \geq \mathcal{I}_p^{\text{VAE}}(\boldsymbol{x}, \boldsymbol{z})$$

Proof Sketch. The proof is based on several applications of the data processing inequality of information theory [Cover and Thomas, 2012] which states that the dependence of $\boldsymbol{x}$ to any hidden state in the hierarchy becomes weaker as one moves further away from $\boldsymbol{x}$ in that hierarchy. As a result $\boldsymbol{x}$ will depend less on the lower layers than the layers near it. Applying this inequality first in the generative skip model yields $\mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z}) \geq \mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{h}^{(l)})\ \forall l \in \{1, ..., L\}$. In particular $\mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z}) \geq \mathcal{I}_p^{\text{VAE}}(\boldsymbol{x}, \boldsymbol{h}^{(1)})$. Applying the inequality again—this time in the generative model of the VAE—we have $\mathcal{I}_p^{\text{VAE}}(\boldsymbol{x}, \boldsymbol{h}^{(1)}) \geq \mathcal{I}_p^{\text{VAE}}(\boldsymbol{x}, \boldsymbol{z})$. It then follows that $\mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z}) \geq \mathcal{I}_p^{\text{VAE}}(\boldsymbol{x}, \boldsymbol{z})$.

Theorem 1 says that for any VAE one can find a SKIP-VAE with higher mutual information. We now use this result to derive the implicit objective function optimized by a SKIP-VAE.

Consider a VAE with mutual information $\delta = \mathcal{I}_{\text{VAE}}(\boldsymbol{x}, \boldsymbol{z})$. We aim to learn the corresponding SKIP-VAE. Using the result of Theorem 1, rewrite the ELBO maximization problem under the generative skip model as a constrained maximization problem,

$$\theta^*, \phi^* = \arg\max_{\theta,\phi} \text{ELBO} \quad \text{s.t} \quad \mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z}) \geq \delta.$$

The equivalent Lagrange dual maximizes

$$\tilde{\mathcal{L}} = \text{ELBO} + \lambda \mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z})$$

where $\lambda > 0$ is the corresponding Lagrange multiplier. Using the expression of the ELBO in Eq. 2 and using the variational joint as defined in Definition 1, write the objective of SKIP-VAE as

$$\mathcal{L} = \mathcal{H}_p(\boldsymbol{x}) - \lambda \mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z}) + \text{KL}(q_\phi(\boldsymbol{x}, \boldsymbol{z}) \,\|\, p_\theta(\boldsymbol{x}, \boldsymbol{z})), \tag{6}$$

where $\mathcal{H}_p(\boldsymbol{x})$ is the entropy of the data distribution.

Minimizing Eq. 6 with respect to θ and ϕ is equivalent to joint distribution matching[2] under the constraint that the mutual information induced by the generative model $\mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z})$ is maximized. Minimizing Eq. 6 brings $p(\boldsymbol{x}, \boldsymbol{z}; \theta)$ closer to $q(\boldsymbol{x}, \boldsymbol{z}; \phi)$ thus also increasing $\mathcal{I}_q^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z})$—the mutual information under the variational joint. Note the SKIP-VAE increases $\mathcal{I}_q^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z})$ by acting on the generative model to increase $\mathcal{I}_p^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z})$. In doing so, it mitigates latent variable collapse. In experiments we see that the SKIP-VAE indeed increases $\mathcal{I}_q^{\text{SKIP-VAE}}(\boldsymbol{x}, \boldsymbol{z})$ relative to the VAE.

[2] Joint distribution matching in the context of VAEs means making the model joint $p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})$ close to the variational joint $q_\phi(\boldsymbol{x}, \boldsymbol{z}) = q_\phi(\boldsymbol{z} \mid \boldsymbol{x})p_{\text{data}}(\boldsymbol{x})$.

4 Empirical study

We assess SKIP-VAEs by applying skip connections to extend a standard VAE [Kingma and Welling, 2013, Rezende et al., 2014] and to the recently introduced SA-VAE [Kim et al., 2018]. We use benchmark datasets for images and text: MNIST, Omniglot, and the Yahoo corpus. Text datasets have been shown to be particularly sensitive to latent variable collapse when the likelihood is parameterized as a fully autoregressive model, such as a recurrent neural network [Bowman et al., 2015]. Note that we are interested in learning both a good generative model (as measured by the ELBO) and a good latent representation of the data (as measured by mutual information and other metrics). The prior for all studies is a spherical Gaussian, and the variational posterior is a diagonal Gaussian. We compare the performance of SKIP-VAE and the baselines when varying the dimensionality of the latent variable and the complexity of the generative model.

Evaluation We assess predictive performance—as given by a measure of held-out log-likelihood—and latent variable collapse. For image datasets we report the ELBO as a measure of log-likelihood; for text we report both the ELBO and perplexity (estimated using importance sampling).

Assessing latent variable collapse is more difficult. We employ three metrics: the KL-divergence, mutual information (MI), and number of active units (AU).

The first metric is the KL regularization term of the ELBO as written in Eq. 2.

The second measure is the mutual information induced by the variational joint $\mathcal{I}_q(\mathbf{x}, \mathbf{z})$. Using the expression of the KL in Eq. 5 we have

$$\mathcal{I}_q(\mathbf{x}, \mathbf{z}) = \text{KL}\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z})\right)) - \text{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z})).$$

We follow Hoffman and Johnson [2016] and approximate this mutual information using Monte Carlo estimates of the two KL terms. In particular,

$$\begin{aligned}\text{KL}(q_\phi(\mathbf{z}) \,\|\, p(\mathbf{z})) &= \mathbb{E}_{q_\phi(\mathbf{z})}\left[\log q_\phi(\mathbf{z}) - \log p(\mathbf{z})\right] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} \log q_\phi(\boldsymbol{z}^{(s)}) - \log p(\boldsymbol{z}^{(s)})\end{aligned}$$

where each aggregated posterior $q_\phi(\boldsymbol{z}^{(s)})$ is also approximated by Monte Carlo.

Table 1: Performance of SKIP-VAE vs VAE on MNIST as the dimensionality of the latent variable increases. SKIP-VAE outperforms VAE on all collapse metrics while achieving similar ELBO values.

	ELBO		KL		MI		AU	
Dim	VAE	SKIP-VAE	VAE	SKIP-VAE	VAE	SKIP-VAE	VAE	SKIP-VAE
2	**-84.27**	-84.30	3.13	**3.54**	3.09	**3.46**	2	2
10	-83.01	**-82.87**	8.29	**9.41**	7.35	**7.81**	9	**10**
20	-83.06	**-82.55**	7.14	**9.33**	6.55	**7.80**	8	**13**
50	-83.31	**-82.58**	6.22	**8.67**	5.81	**7.49**	8	**12**
100	-83.41	**-82.52**	5.82	**8.45**	5.53	**7.38**	5	**9**

Table 2: Performance of SKIP-VAE vs VAE on MNIST (Top) and Omniglot (Bottom) as the complexity of the generative model increases. The number of latent dimension is fixed at 20. Skip-VAE outperforms VAE on all collapse metrics while achieving similar ELBO values, and the difference widens as layers increase.

		ELBO		KL		MI		AU	
	Layers	VAE	SKIP-VAE	VAE	SKIP-VAE	VAE	SKIP-VAE	VAE	SKIP-VAE
MNIST	1	-89.64	**-89.22**	13.31	**13.40**	8.56	8.56	20	20
	3	-84.38	**-84.03**	10.12	**10.71**	7.95	**8.20**	16	16
	6	-83.19	**-82.81**	8.82	**9.77**	7.53	**7.93**	11	**13**
	9	-83.06	**-82.55**	7.14	**9.34**	6.55	**7.80**	8	**13**
Omniglot	1	-97.69	**-97.66**	**8.42**	8.37	**7.09**	7.08	20	20
	3	-93.95	**-93.75**	6.43	**6.58**	5.88	**5.97**	20	20
	6	-93.23	**-92.94**	5.24	**5.78**	4.94	**5.43**	20	20
	9	-92.79	**-92.61**	4.41	**6.12**	4.24	**5.65**	11	**20**

The third measure of latent variable collapse is the number of "active" units of the latent variable $\boldsymbol{z}$. This is defined in Burda et al. [2015] as

$$\text{AU} = \sum_{d=1}^{D} \mathbb{1}\left\{\text{Cov}_{p(\mathbf{x})}\left(\mathbb{E}_{q_\phi(\boldsymbol{z}\,|\,\mathbf{x})}[\boldsymbol{z}_d]\right) \geq \epsilon\right\},$$

where z_d is the d^{th} dimension of $\boldsymbol{z}$ and ϵ is a threshold. ($\mathbb{1}\{\cdot\}$ is an indicator giving 1 when its argument is true and 0 otherwise.) We follow Burda et al. [2015] and use a threshold of $\epsilon = 0.01$. We observe the same phenomenon: the histogram of the number of active dimensions of $\boldsymbol{z}$ is bi-modal, which means that it is not highly sensitive to the chosen threshold.

4.1 Images

Model We use a 3-layer ResNet [He et al., 2016a] (with 3×3 filters and 64 feature maps in each layer) as the variational neural network and a 9-layer Gated PixelCNN [van den Oord et al., 2016] (with 3×3 filters and 32 feature maps) as the likelihood. The baseline model uses a linear map from the sample (to project out to the image spatial resolution), concatenated with the original image, which is fed to the PixelCNN. This setup reflects the current state-of-the-art for image VAEs [Gulrajani et al., 2016, Chen et al., 2016].[3] The SKIP-VAE uses a linear map from the sample, concatenated with the output from each layer of the PixelCNN (before feeding it to the next layer). While this results in slightly more parameters for the SKIP-VAE model, we found that the baseline VAE's performance on the collapse metrics actually gets worse as the model size increases.

Results Table 1 shows the results on MNIST as we vary the size of the latent dimension. In all scenarios, the generative skip model yields higher KL between the variational posterior and the prior, higher mutual information (confirming the statement in Theorem 1), and uses more latent dimensions (as measured by AU).

Table 2 shows experiments on both MNIST and Omniglot as we vary the generative model's complexity by increasing its depth. We use a model with 20-dimensional latent variables. For VAE, as the generative model becomes more expressive the model becomes less reliant on $\mathbf{z}$. We see this in the poor performance on the collapse metrics. The SKIP-VAE mitigates this issue and performs better on all latent-variable collapse metrics. Note the ELBO is similar for both models. These results indicate that the family of generative skip models has a strong inductive bias to share more mutual information between the observation and the latent variable.

Similar results are observed when using weaker models. For example in Table 3 we used multilayer perceptrons (MLPS) for both the variational neural network and the generative model, and we set the dimensionality of the latent variables

[3] While our model capacity is similar to these works, our performance is slightly worse since we do not employ additional techniques such as data-dependent initialization [Chen et al., 2016].

Table 3: VAE and SKIP-VAE on MNIST using 50 latent dimensions with a simplified network. Here the encoder is a 2-layer MLP with 512 units in each layer and the decoder is also an MLP. The results below correspond to different number of layers for the decoder.

	ELBO		KL		MI		AU	
Layers	VAE	SKIP-VAE	VAE	SKIP-VAE	VAE	SKIP-VAE	VAE	SKIP-VAE
2	−94.88	**-94.80**	24.23	**26.35**	**9.21**	9.20	17	**24**
3	−95.38	**-94.17**	21.87	**26.15**	9.20	**9.21**	13	**21**
4	−97.09	**-93.79**	20.95	**25.63**	**9.21**	**9.21**	11	**21**

Table 4: SKIP-VAE and SKIP-SA-VAE perform better than their counterparts (VAE, SA-VAE) on the Yahoo corpus under all latent variable collapse metrics while achieving similar log-likelihoods. In particular, all latent dimensions are active when using SKIP-SA-VAE. Perplexity (PPL) for the variational models is estimated by importance sampling of the log marginal likelihood with 200 samples from $q_\phi(\mathbf{z} \mid \boldsymbol{x})$.

Model	Dim	PPL	ELBO	KL	MI	AU
LSTM LANGUAGE MODEL	-	61.60	-	-	-	-
VAE	32	62.38	−330.1	0.005	0.002	0
SKIP-VAE	32	61.71	−330.5	**0.34**	**0.31**	**1**
SA-VAE	32	59.85	−327.5	5.47	4.98	14
SKIP-SA-VAE	32	60.87	−330.3	**15.05**	**7.47**	**32**
SA-VAE	64	60.20	−327.3	3.09	2.95	10
SKIP-SA-VAE	64	60.55	−330.8	**22.54**	**9.15**	**64**

to 50. Even with this weaker setting the SKIP-VAE leads to less collapse than the VAE.

Latent Representations. We find qualitatively that the latent representations learned by the SKIP-VAE better capture the underlying structure. Figure 2 illustrates this. It shows a much clearer separation of the MNIST digits with the latent space learned by the SKIP-VAE compared to the latent space of the VAE. [4]

Quantitatively we performed a classification study on MNIST using the latent variables learned by the variational neural networks of VAE and SKIP-VAE as features. This study uses 50 latent dimensions, a 9-layer PixelCNN as the generative model, a 3-layer ResNet as the variational neural network, and a simple 2-layer MLP over the posterior means as the classifier. The MLP has 1024 hidden units, ReLU activations, and a dropout rate of 0.5. The classification accuracy of the VAE is 97.19% which is lower than the accuracy of the SKIP-VAE which is 98.10%. We also studied this classification performance on a weaker model. We replaced the 9-layer PixelCNN and the 3-layer ResNet above by two MLPS. The VAE achieved an accuracy of 97.70% whereas the SKIP-VAE achieved an accuracy of 98.25%.

[4]Note we did not fit a VAE and a SKIP-VAE with 2-dimensional latents for the visualization. Fitting 2-dimensional latents would have led to much better learned representations for both the VAE and the SKIP-VAE. However using 2-dimensional latents does not correspond to a realistic setting in practice. Instead we fit the VAE and the SKIP-VAE on 50-dimensional latents—as is usual in state-of-the-art image modeling with VAEs—and used t-SNE to project the learned latents on a 2-dimensional space.

4.2 Text

Model For text modeling, we use the training setup from Kim et al. [2018], a strong baseline that outperforms standard LSTM language models. The variational neural network is a 1-layer LSTM with 1024 hidden units, whose last hidden state is used to predict the mean vector and the (log) variance vector of the variational posterior. The generative model is also a 1-layer LSTM with 1024 hidden units. We found that the 1-layer model performed better than deeper models, potentially due to overfitting. In the VAEs the sample is used to predict the initial hidden state of the decoder and also fed as input at each time step. In the generative skip model we also concatenate the sample with the decoder's hidden state.

We also study the semi-amortized variational autoencoder, SA-VAE [Kim et al., 2018], which proposes a different optimization-based strategy for targeting the latent variable collapse issue when training VAEs for text. SA-VAE combines stochastic variational inference [Hoffman et al., 2013] with amortized variational inference by first using an inference network over $\mathbf{x}$ to predict the initial variational parameters and then subsequently running iterative inference on the ELBO to refine the initial variational parameters. We used 10 steps of iterative refinement for SA-VAE and SKIP-SA-VAE.

Results We analyze the Yahoo Answers dataset from Yang et al. [2017], a benchmark for deep generative models of text. Table 4 shows the results. We first note that successfully

training standard VAEs for text with flexible autoregressive likelihoods such as LSTMs remains a difficult problem. We see that VAE by itself experiences latent variables collapse. The SKIP-VAE is slightly better than VAE at avoiding latent variable collapse for similar log likelihoods, although the KL is only marginally above zero and the model only has one active unit.

When combining both approaches with the semi-amortized training, we see better use of latent variables in SA-VAE and SKIP-SA-VAE. While SA-VAE alone does mitigate collapse to an extent, skip connections learn generative models where the mutual information is even higher. Furthermore, the trend changes when adding more latent dimension. For the vanilla SA-VAE, the mutual information and active units are actually *lower* for a model trained with 64-dimensional latent variables than a model trained with 32-dimensional latent variables. This is a common issue in VAEs whereby simply increasing the dimensionality of the latent space actually results in a worse model. In contrast, models trained with skip connections make full use of the latent space and collapse metrics improve as we increase the number of dimensions. For example the SKIP-SA-VAE uses all the dimensions of the latent variables.

5 Conclusion

We have proposed a method for reducing latent variable collapse in VAEs. The approach uses skip connections to promote a stronger dependence between the observations and their associated latent variables. The resulting family of deep generative models (SKIP-VAEs) learn useful summaries of data. Theoretically we showed that SKIP-VAEs yield higher mutual information than their counterparts. We found that SKIP-VAEs—when used with more sophisticated VAEs such as the SA-VAE—lead to a significant improvement in terms of latent variable collapse.

References

A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168, 2018.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. In *Proceedings of International Conference on Learning Representations*, 2015.

X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

A. B. Dieng, C. Wang, J. Gao, and J. Paisley. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702*, 2016.

K. Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1 (2):119–130, 1988.

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

K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang. Generating sentences by editing prototypes. *arXiv preprint arXiv:1709.08878*, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2016a.

K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016b.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *Proceedings of International Conference on Learning Representations*, 2017.

M. D. Hoffman. Learning deep latent gaussian models with markov chain monte carlo. In *Proceedings of International Conference on Machine Learning*, pages 1510–1519, 2017.

M. D. Hoffman and M. J. Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference*, 2016.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 13:1303–1347, 2013.

Y. Kim, S. Wiseman, A. C. Miller, D. Sontag, and A. M. Rush. Semi-amortized variational autoencoders. *arXiv preprint arXiv:1802.02550*, 2018.

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

D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances*

in Neural Information Processing Systems, pages 4743–4751, 2016.

D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara. Variational autoencoders for collaborative filtering. *arXiv preprint arXiv:1802.05814*, 2018.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

Y. Miao, L. Yu, and P. Blunsom. Neural variational inference for text processing. In *International Conference on Machine Learning*, pages 1727–1736, 2016.

D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

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

C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. How to train deep variational autoencoders and probabilistic ladder networks. *arXiv preprint arXiv:1602.02282*, 2016.

R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

J. M. Tomczak and M. Welling. Vae with a vampprior. *arXiv preprint arXiv:1705.07120*, 2017.

A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, 2016.

A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6309–6318, 2017.

J. Xu and G. Durrett. Spherical latent spaces for stable variational autoencoders. *arXiv preprint arXiv:1808.10805*, 2018.

Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of International Conference on Machine Learning*, 2017.

S. Yeung, A. Kannan, Y. Dauphin, and L. Fei-Fei. Tackling over-pruning in variational autoencoders. *arXiv preprint arXiv:1706.03643*, 2017.

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

T. Zhao, K. Lee, and M. Eskenazi. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. *arXiv preprint arXiv:1804.08069*, 2018.