Consistency Regularization for Variational Auto-Encoders

Samarth Sinha [1] **Augustus Odena** [2] **Adji B. Dieng** [3]

Abstract

Variational auto-encoders (VAEs) are a powerful approach to unsupervised learning. They enable scalable approximate posterior inference in latent-variable models using variational inference (VI). A VAE posits a variational family parameterized by a deep neural network—called an *encoder*—that takes data as input. This encoder is shared across all the observations, which amortizes the cost of inference. However the encoder of a VAE has the undesirable property that it maps a given observation and a semantic-preserving transformation of it to different latent representations. This "inconsistency" in the representations induced by the encoder negatively affects generalization. In this paper, we propose a regularization method to enforce consistency in VAEs. The idea is to minimize the Kullback-Leibler (KL) divergence between the variational distribution when conditioning on the observation and the variational distribution when conditioning on a random semantic-preserving transformation of this observation. This regularization is applicable to any VAE. In our experiments we apply it to three different VAE variants on several benchmark datasets and found it always improves generalization but also yields more interpretable latent variables as measured by mutual information.

1. Introduction

VAEs have significantly impacted research on unsupervised learning. They have been used in several areas, including density estimation for images (Kingma & Welling, 2013; Rezende et al., 2014), image generation (Gregor et al., 2015), text generation (Bowman et al., 2015; Fang et al., 2019), language modeling (Dieng et al., 2016), music generation (Roberts et al., 2018), topic modeling (Miao et al., 2016; Dieng et al., 2019), and recommendation systems (Liang et al., 2018).

VAEs extend deterministic auto-encoders to probabilistic generative modeling. The encoder of a VAE parameterizes an approximate posterior distribution over latent variables of a generative model. The encoder is shared between all observations, which amortizes the cost of posterior inference. Once fitted, the encoder of a VAE can be used to obtain low-dimensional representations of data, (e.g. for downstream tasks.) The quality of these representations is therefore very important to a successful application of VAEs.

Researchers have looked at ways to improve the quality of the latent representations of VAEs, often tackling the so-called *latent variable collapse* problem—in which the approximate posterior distribution induced by the encoder collapses to the prior over the latent variables (Bowman et al., 2015; Kim et al., 2018; Dieng et al., 2018; He et al., 2019; Fu et al., 2019).

In this paper, we focus on a different problem pertaining to the latent representations of VAEs for image data. Indeed, the encoder of a fitted VAE, tends to map a semantics-preserving transformation of an image to a different latent representation than the original image. This "inconsistency" negatively affects generalization. We propose a simple idea to enforce consistency in VAEs. The idea is to maximize the likelihood of both the images and their semantics-preserving transformations and minimize the KL divergence between the approximate posterior distribution induced by the encoder when conditioning on the image, on one hand, and its transformation, on the other hand. This regularization technique can be applied to any VAE variant to improve generalization. We call a VAE with this form of regularization, a consistency-regularized variational auto-encoder (CR-VAE).

Section 1 illustrates the inconsistency problem of VAEs and how CR-VAEs address this problem. The figure shows the representations learned by a VAE fitted on MNIST (Figure 1a.) The red dots represent the representations of few images from the test set and the blue dots represent the representations of their transformations. We applied

[1]University of Toronto, Toronto, Canada [2]Google Brain, Mountain View, CA, USA [3]Columbia University, New York, NY, USA. Correspondence to: Samarth Sinha <samarth.sinha@mail.utoronto.ca>.

Second workshop on *Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models* (ICML 2020), Virtual Conference

semantics-preserving transformations: rotation, translation, and scaling. The VAE maps each image and its transformation to different parts in the latent space as evidenced by the long arrows connecting each pair. Even when the VAE is fitted using MNIST augmented with semantics-preserving image transformations, this inconsistency problem still occurs (Figure 1b.) The CR-VAE does not suffer from the inconsistency problem; it maps each image and its transformation to nearby areas in the latent space, as evidenced by the short arrows connecting each pair.

In Section 4 we apply the proposed technique to three VAE variants, the original VAE (Kingma & Welling, 2013), the importance-weighted auto-encoder (IWAE) (Burda et al., 2015), and the β-VAE (Higgins et al., 2017). We found, on three different benchmark datasets, that CR-VAEs always generalize better than their base VAEs. We also found CR-VAEs learn more meaningful latent representations as measured by mutual information and number of active units in the learned representations. An ablation study reveals the imposed KL constraint further improves predictive and qualitative performance.

2. Method

We consider a latent-variable model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}) \cdot p(\mathbf{z})$, where $\mathbf{x}$ denotes an observation and $\mathbf{z}$ is its associated latent variable. The marginal $p(\mathbf{z})$ is a prior over the latent variable and $p_\theta(\mathbf{x}|\mathbf{z})$ is an exponential family distribution whose natural parameter is a function of $\mathbf{z}$ parameterized by θ, e.g. through a neural network. Our goal is to learn the parameters θ and a posterior distribution over the latent variables. The approach of VAEs is to maximize the evidence lower bound (ELBO), a lower bound on the log marginal likelihood of the data,

$$\mathcal{L}_{\text{VAE}} = \text{ELBO} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right)\right] \tag{1}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is an approximate posterior distribution over the latent variables. The idea of a VAE is to let the parameters of the distribution $q_\phi(\mathbf{z}|\mathbf{x})$ be given by the output of a neural network, with parameters ϕ, that takes $\mathbf{x}$ as input. The parameters θ and ϕ are then jointly optimized by maximizing a Monte Carlo approximation of the ELBO using the reparameterization trick (Kingma & Welling, 2013).

Consider a semantic-preserving transformation $t(\tilde{\mathbf{x}}|\mathbf{x})$ of data $\mathbf{x}$ (e.g. rotation or translation for images.) A good representation learning algorithm should provide similar latent representations for $\mathbf{x}$ and $\tilde{\mathbf{x}}$. This is not the case for the VAE that maximizes Equation 1 and its variants. Once fit to data, the encoder of a VAE is unable to yield similar latent representations for a data $\mathbf{x}$ and its tranformation $\tilde{\mathbf{x}}$. This is because there is nothing in Equation 1 that forces this desideratum.

We now propose a regularization method that ensures *consistency* of the encoder of a VAE. We call a VAE with such a regularization a CR-VAE. The regularization proposed is applicable to many variants of the VAE such as the IWAE (Burda et al., 2015) and the β-VAE (Higgins et al., 2017). In what follows, we use the standard VAE, the one that maximizes Equation 1, as the base VAE to regularize.

A CR-VAE maximizes the following objective,

$$\begin{aligned}\mathcal{L}_{\text{CR-VAE}} = \mathcal{L}_{\text{VAE}} + \mathbb{E}_{t(\tilde{\mathbf{x}}|\mathbf{x})}\mathbb{E}_{q_\phi(\tilde{\mathbf{z}}|\tilde{\mathbf{x}})}\left[\log\left(\frac{p_\theta(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})}{q_\phi(\tilde{\mathbf{z}}|\tilde{\mathbf{x}})}\right)\right] \\ - \eta \cdot \mathbb{E}_{t(\tilde{\mathbf{x}}|\mathbf{x})}\left[\text{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||q_\phi(\tilde{\mathbf{z}}|\tilde{\mathbf{x}})\right)\right].\end{aligned} \tag{2}$$

Here $\mathcal{L}_{\text{VAE}}$ is the base VAE being regularized. The second term of Equation 2 is the regularizer; it is an expectation over the transformations. This regularization has two components. The first component renders transformed data $\tilde{\mathbf{x}}$ likely under the model; it is a data augmentation term. The second component of the regularizer ensures $\mathbf{x}$ and its transformation $\tilde{\mathbf{x}}$ have close representations in the latent space. Although we defined closeness using KL. Here λ and η are hyperparameters that determine the strength of the regularization.

We draw samples from $t(\tilde{\mathbf{x}}|\mathbf{x})$ by applying random semantic-preserving transformations to $\mathbf{x}$. More concretely, for image data, we apply translations with randomly sampled length or rotations with randomly sampled angle to get $\tilde{\mathbf{x}}$. Consider an image $\mathbf{x}$. We draw $\tilde{\mathbf{x}}$ from $t(\tilde{\mathbf{x}}|\mathbf{x})$ as follows:

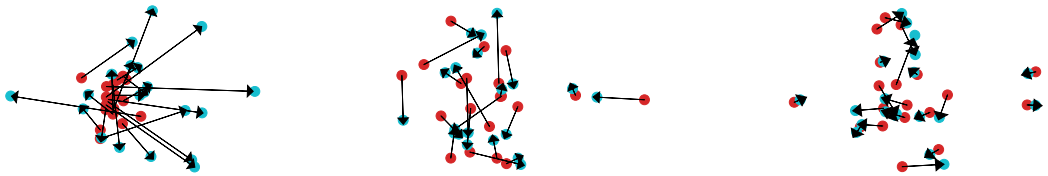
$$\tilde{\mathbf{x}} \sim t(\tilde{\mathbf{x}}|\mathbf{x}) \iff \epsilon \sim p(\epsilon), \tilde{\mathbf{x}} = g(\mathbf{x}, \epsilon). \tag{3}$$

Here $g(\mathbf{x}, \epsilon)$ is a random semantics-preserving transformation of the image $\mathbf{x}$, e.g. translation with random length ϵ drawn from $p(\epsilon) = \mathcal{U}[-\delta, \delta]$ for some threshold δ.

Note Equation 2 is intractable because the expectations are intractable. We approximate $\mathcal{L}_{\text{CR-VAE}}$ using Monte Carlo with the reparameterization trick. That is, we approximate Equation 2 by drawing one-sample from $t(\tilde{\mathbf{x}}|\mathbf{x})$ following Equation 3 and one sample from $q_\phi(\mathbf{z}|\mathbf{x})$ and $q_\phi(\tilde{\mathbf{z}}|\tilde{\mathbf{x}})$ using the reparameterization trick.

3. Related Work

Applying consistency regularization to VAEs, as we do in this paper, has not been previously explored. Consistency regularization is a widely used technique for semi-supervised learning (Bachman et al., 2014; Sajjadi et al., 2016; Laine & Aila, 2016; Miyato et al., 2018; Xie et al., 2019). The core idea behind consistency regularization for semi-supervised learning is to force classifiers to learn representations that are insensitive to semantics-preserving changes to images, so as to improve classification of unlabeled images. Examples of semantics-preserving changes



(a) VAE **(b)** VAE+Aug **(c)** CR-VAE

Figure 1. Illustration on MNIST of the inconsistency issue of VAEs and how CR-VAEs address this problem. The **red** dots correspond to the representations of few images from the test set. The **blue** dots correspond to the representations of the transformed images. The transformations used here are rotations, translations, and scaling; they are semantics-preserving. **(a)**: the VAE maps the two sets of images to different areas in the latent space. **(b)**: even when trained with the original dataset augmented with the transformed images, the VAE still maps the two sets of images to different parts in the latent space. **(c)**: the CR-VAE maps each pair in the two sets of images to nearby areas in the latent space.

used in the literature include rotation, zoom, translation, crop, or adversarial attacks. Consistency is often enforced by minimizing the $\mathbb{L}_2$ distance between a classifier's logit output for an image and the logit output for its semantics-preserving transformation (Sajjadi et al., 2016; Laine & Aila, 2016), or by minimizing the KL divergence between the classifier's label distribution induced by the image and that of its tranformation (Miyato et al., 2018; Xie et al., 2019). Prior work in denoising and contraactive auto-encoders have also explored similar problems (Rifai et al., 2011; Vincent et al., 2008; 2010).

More recently, consistency regularization has been applied to generative adversarial networks (GANs) (Goodfellow et al., 2014). Indeed (Wei et al., 2018; Zhang et al., 2020) show that applying consistency regularization on the discriminator of a GAN—also a classifier—can substantially improve its performance. The work differs from the works above in two ways. First, it applies consistency regularization to VAEs for image data. Second, it leverages consistency regularization, not in the label or logit space, as done in the works mentioned above, but in the latent space which is lower-dimensional.

4. Empirical Study

In this section we show CR-VAE significantly improves the generalization performance of its base VAE and yields more interpretable latent representations. We also show that the proposed regularization method is amenable to different VAE variants by applying it to the IWAE and the β-VAE. Finally, we conduct an ablation study to assess the effect of λ and η in Equation 2. We found that only regularizing with data augmentation ($\lambda > 0$) significantly improves performance but that accounting for the KL term ($\eta > 0$) further improves both quantitative and qualitative performance.

4.1. Experimental Protocol

We study three benchmark datasets: MNIST, OMNIGLOT, and CELEBA.

We consider three transformations $t(\tilde{\mathbf{x}}|\mathbf{x})$. The first randomly translates an image $[-2, 2]$ pixels in any direction. The second transformation randomly rotates an image uniformly in $[-15, 15]$ degrees clockwise. Finally the third transformation randomly scales an image by a factor uniformly sampled from $[0.9, 1.1]$.

To test the effectiveness of the proposed method to generalize to changes in distribution, we create a *transformed*-test set (t-Test set) from the original test set. To obtain the *transformed*-test set we perform the same transformations as above but increase their magnitude. Specifically, we randomly translate each image in the test set in $[-4, 4]$ pixels in any direction, randomly rotate each image in $[-30, 30]$ degrees clockwise, and randomly scale each image by a factor uniformly sampled from $[0.75, 1.25]$. By increasing the magnitude of the transformations, we are able to test the ability to generalize to distribution shifts. We measure held-out log-likelihood on the original test set and the t-Test set.

We consider mutual information (MI) and number of active units (AU) as measures of the quality of the learned latent variables (Dieng et al., 2018).

The VAEs are built on the same architecture as (Tolstikhin et al., 2017). The networks are trained with the Adam optimizer with a learning rate of 10^{-4} (Kingma & Ba, 2014) and trained for 100 epochs with a batch size of 64. We set the dimensionality of the latent variables to 50, therefore the maximum number of active units in the latent space is 50. We set $\lambda = 1$ for all the experiments. We found $\eta = 0.1$ to be best according to cross-validation using held-out log-likelihood and exploring the range $[1e^{-4}, 1.0]$ datasets. In an ablation study we explore $\eta = 0$. For the β-VAE we set

Table 1. CR-VAEs generalize better than their base VAEs on MNIST as evidenced by lower negative held-out log-likelihood on both the original test set and the t-Test set. CR-VAEs also learn more meaningful latent representations as evidenced by higher mutual information and active units.

	NLL		MI		AU	
Method	Test Set	t-Test Set	Test Set	t-Test Set	Test Set	t-Test Set
VAE	83.7 ± 0.3	338.7 ± 1.7	124.5 ± 1.1	171.6 ± 5.4	36 ± 0.8	36 ± 0.8
CR-VAE	**81.2** ± 0.2	**129.5** ± 0.4	**126.3** ± 0.9	**216.4** ± 1.2	**47** ± 0.5	**47** ± 0.5
IWAE	81.7 ± 0.3	318.0 ± 0.8	127.1 ± 0.7	180.9 ± 3.9	39 ± 0.5	39 ± 0.5
CR-IWAE	**79.7** ± 0.3	**124.8** ± 0.8	**129.7** ± 1.0	**223.9** ± 0.8	**50** ± 0	**50** ± 0
β-VAE ($\beta = 0.5$)	92.6 ± 0.3	285.3 ± 0.8	284.3 ± 1.1	359.9 ± 2.2	50 ± 0	50 ± 0
β-CR-VAE ($\beta = 0.5$)	**85.7** ± 0.6	**123.4** ± 0.6	**291.9** ± 0.7	**490.2** ± 1.0	**50** ± 0	**50** ± 0
β-VAE ($\beta = 10$)	**126.1** ± 1.8	361.3 ± 1.5	6.3 ± 0.6	7.6 ± 0.9	8 ± 1.7	8 ± 1.7
β-CR-VAE ($\beta = 10$)	126.2 ± 0.5	**225.8** ± 0.9	**6.9** ± 0.6	**10.6** ± 0.4	**10** ± 0.5	**10** ± 0.5

Table 2. CR-VAEs generalize better than their base VAEs on OMNIGLOT as evidenced by lower negative held-out log-likelihood on both the original test set and the t-Test set. CR-VAEs also learn more meaningful latent representations as evidenced by higher mutual information and active units.

	NLL		MI		AU	
Method	Test Set	t-Test Set	Test Set	t-Test Set	Test Set	t-Test Set
VAE	128.2 ± 0.8	863.1 ± 4.2	105.4 ± 1.2	115.4 ± 2.5	50 ± 0	50 ± 0
CR-VAE	**124.1** ± 0.1	**381.2** ± 1.6	**107.8** ± 1.1	**215.6** ± 0.6	**50** ± 0	**50** ± 0
IWAE	127.5 ± 0.5	861.1 ± 2.9	110.3 ± 1.1	112.4 ± 1.4	50 ± 0	50 ± 0
CR-IWAE	**123.6** ± 0.5	**381.2** ± 1.2	**115.3** ± 0.8	**224.5** ± 0.6	**50** ± 0	**50** ± 0
β-VAE ($\beta = 0.5$)	137.1 ± 0.2	947.7 ± 5.2	143.4 ± 1.0	148.3 ± 0.9	50 ± 0	50 ± 0
β-CR-VAE ($\beta = 0.5$)	**132.5** ± 0.3	**388.2** ± 1.0	**169.5** ± 0.5	**378.3** ± 2.0	**50** ± 0	**50** ± 0
β-VAE ($\beta = 10$)	**157.5** ± 1.1	937.0 ± 3.4	1.4 ± 0.2	8.3 ± 0.6	4 ± 0.9	4 ± 0.9
β-CR-VAE ($\beta = 10$)	157.6 ± 0.6	**562.0** ± 2.1	**1.6** ± 0.1	**22.5** ± 0.5	**4** ± 0.5	**4** ± 0.5

Table 3. CR-VAEs generalize better than their base VAEs on CELEBA as evidenced by lower negative held-out log-likelihood on both the original test set and the t-Test set. CR-VAEs also learn more meaningful latent representations as evidenced by higher mutual information and active units.

	NLL		MI		AU	
Method	Test Set	t-Test Set	Test Set	t-Test Set	Test Set	t-Test Set
VAE	66.1 ± 0.2	108.3 ± 0.4	33.8 ± 0.2	55.1 ± 1.0	32 ± 0.9	32 ± 0.9
CR-VAE	**65.9** ± 0.2	**98.3** ± 0.8	**34.9** ± 0.5	**67.1** ± 0.9	**33** ± 1.2	**33** ± 1.2
IWAE	65.3 ± 0.1	105.1 ± 0.6	36.9 ± 0.5	61.2 ± 1.0	36 ± 1.6	36 ± 1.6
CR-IWAE	**65.0** ± 0.2	**97.6** ± 0.8	**38.4** ± 0.5	**74.1** ± 1.1	**36** ± 1.9	**36** ± 1.9
β-VAE ($\beta = 0.5$)	68.7 ± 0.2	110.1 ± 0.6	75.8 ± 0.5	124.3 ± 1.8	49 ± 0.5	49 ± 0.5
β-CR-VAE ($\beta = 0.5$)	**68.2** ± 0.1	**96.7** ± 0.2	**77.1** ± 0.1	**141.6** ± 1.0	**50** ± 0	**50** ± 0
β-VAE ($\beta = 10$)	92.7 ± 0.5	172.4 ± 1.5	3.6 ± 0.3	3.7 ± 0.3	7 ± 0.8	7 ± 0.8
β-CR-VAE ($\beta = 10$)	**92.6** ± 0.1	**147.6** ± 0.6	**3.7** ± 0.4	**4.1** ± 0.3	**9** ± 0.9	**9** ± 0.9

$\eta = 0.1 \cdot \beta$ and study both $\beta = 0.1$ and $\beta = 10$, two regimes under which the β-VAE performs qualitatively very differently (Higgins et al., 2017).

4.2. Results

Table 1, Table 2, and Table 3 show generalization performance and the quality of the learned latent representations for different VAE variants and their CR-VAEs counterparts. These results show that applying consistency regularization always improves generalization performance. More interestingly, CR-VAEs improve out-of-distribution generalization of their base VAEs, as evidenced by improved negative log-likelihood scores on the transformed test set.

Although our original focus is to improve generalization by solving the inconsistency issue of VAEs, the results in Table 1, Table 2, and Table 3 show that CR-VAEs also yield more meaningful learned latent representations than their

Table 4. Ablation study on the impact of the KL term in Equation 2. On all datasets, the KL term of the CR-VAE objective helps improve both generalization and qualitative performance.

Method	NLL		MI		AU	
	Test Set	t-Test Set	Test Set	t-Test Set	Test Set	t-Test Set
VAE+Aug+MNIST	82.8 ± 0.4	131.3 ± 0.5	125.9 ± 0.2	208.4 ± 0.4	42 ± 0.5	42 ± 0.5
CR-VAE+MNIST	**81.2** ± 0.2	**129.5** ± 0.4	**126.3** ± 0.9	**216.4** ± 1.2	**47** ± 0.5	**47** ± 0.5
VAE+Aug+Omniglot	125.7 ± 0.2	402.5 ± 1.4	105.9 ± 0.7	208.3 ± 0.8	**50** ± 0	**50** ± 0
CR-VAE+Omniglot	**124.1** ± 0.1	**381.2** ± 1.6	**107.8** ± 1.1	**215.6** ± 0.6	**50** ± 0	**50** ± 0
VAE+Aug+CelebA	66.0 ± 0.2	98.4 ± 0.3	34.1 ± 0.8	63.1 ± 0.7	**33** ± 0.9	**33** ± 0.9
CR-VAE+CelebA	**65.9** ± 0.2	**98.3** ± 0.8	**34.9** ± 0.5	**67.1** ± 0.9	**33** ± 1.2	**33** ± 1.2

base VAEs as evidenced by higher mutual information and active units.

Table 4 shows the results of an ablation study to assess the importance of the KL term imposed on the approximated posterior induced by the encoder when conditioning on the data $\mathbf{x}$ ($q_\phi(\mathbf{z}|\mathbf{x})$) and when conditioning on its transformation $\tilde{\mathbf{x}}$ ($q_\phi(\tilde{\mathbf{z}}|\tilde{\mathbf{x}})$.) From Table 4, we conclude this constraint further helps improve performance, both in terms of generalization and in terms of the quality of the learned latents.

5. Conclusion

We proposed a simple idea to constrain encoders of VAEs to learn similar latent representations for an image and a semantics-preserving transformation of the image. The idea consists in maximizing the likelihood of the images and their semantics-preserving transformations and to minimize the KL divergence between the variational distribution induced by the encoder when conditioning on the image and its transformation. We applied this technique to three VAE variants on three benchmark datasets. We found it always leads to better generalization as measured by held-out log-likelihood and more meaningful latent representations as measured by mutual information and number of active units. An ablation study revealed the KL constrain further improves performance.

References

Bachman, P., Alsharif, O., and Precup, D. Learning with pseudo-ensembles. In *Advances in neural information processing systems*, pp. 3365–3373, 2014.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., 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*, 2015.

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

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.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*, 2019.

Fang, L., Li, C., Gao, J., Dong, W., and Chen, C. Implicit deep latent variable models for text generation. *arXiv preprint arXiv:1908.11527*, 2019.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.

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*, pp. 2672–2680, 2014.

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., and Wierstra, D. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

He, J., Spokoyny, D., Neubig, G., and Berg-Kirkpatrick, T. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*, 2019.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Liang, D., Krishnan, R. G., Hoffman, M. D., and Jebara, T. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference*, pp. 689–698, 2018.

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

Miyato, T., Maeda, S.-i., Ishii, S., and Koyama, M. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

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

Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. Contractive auto-encoders: Explicit invariance during feature extraction. 2011.

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., and Eck, D. A hierarchical latent vector model for learning long-term structure in music. *arXiv preprint arXiv:1803.05428*, 2018.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, 2016.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

Wei, X., Gong, B., Liu, Z., Lu, W., and Wang, L. Improving the improved training of wasserstein gans: A consistency term and its dual effect. *arXiv preprint arXiv:1803.01541*, 2018.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., and Le, Q. V. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

Zhang, H., Zhang, Z., Odena, A., and Lee, H. Consistency regularization for generative adversarial networks. 2020.