The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

SHOT-VAE: Semi-supervised Deep Generative Models With Label-aware ELBO Approximations

Hao-Zhe Feng,[1] Kezhi Kong,[2] Minghao Chen[1]
Tianye Zhang,[1] Minfeng Zhu,[1] Wei Chen[1*]
[1] Zhejiang University
[2] University of Maryland, College Park
{fenghz, chenminghao, zhangtianye1026, minfeng_zhu, chenvis}@zju.edu.cn, kong@cs.umd.edu

Abstract

Semi-supervised variational autoencoders (VAEs) have obtained strong results, but have also encountered the challenge that **good ELBO values do not always imply accurate inference results**. In this paper, we investigate and propose two causes of this problem: (1) The ELBO objective cannot utilize the label information directly. (2) A bottleneck value exists, and continuing to optimize ELBO after this value will not improve inference accuracy. On the basis of the experiment results, we propose SHOT-VAE to address these problems without introducing additional prior knowledge. The SHOT-VAE offers two contributions: (1) A new ELBO approximation named *smooth-ELBO* that integrates the label predictive loss into ELBO. (2) An approximation based on *optimal interpolation* that breaks the ELBO value bottleneck by reducing the margin between ELBO and the data likelihood. The SHOT-VAE achieves good performance with 25.30% error rate on CIFAR-100 with 10k labels and reduces the error rate to 6.11% on CIFAR-10 with 4k labels.

Introduction

Most deep learning models are trained with large labeled datasets via supervised learning. However, in many scenarios, although acquiring a large amount of original data is easy, obtaining corresponding labels is often costly or even infeasible (Wei et al. 2020). Thus, semi-supervised variational autoencoder (VAE) (Kingma et al. 2014) is proposed to address this problem by training classifiers with multiple unlabeled data and a small fraction of labeled data.

Based on the latent variable assumption (Doersch 2016), semi-supervised VAE models combine the evidence lower bound (ELBO) and the classification loss as objective, so that it can not only learn the required classification representations from labeled data, but also capture the disentangled factors which could be used for data generation. Although semi-supervised VAE models have obtained strong empirical results on many benchmark datasets (e.g. MNIST, SVHN, Yale B) (Narayanaswamy et al. 2017), it still encounters one common problem that **good ELBO values do not always imply accurate inference results** (Zhao et al. 2017). To address this problem, existing works introduce prior knowledge that needs to be set manually, e.g., the stacked VAE structure (M1+M2, Kingma et al. 2014; Davidson et al. 2018), the prior domain knowledge (Louizos et al. 2016; Ilse et al. 2019) and mutual information bounds (Dupont 2018).

In this study, we investigate the training process of semi-supervised VAE with extensive experiments and propose two possible causes of the problem. (1) First, the ELBO cannot utilize label information directly. In the semi-supervised VAE framework (Kingma et al. 2014), the classification loss and ELBO learn from the labels and unlabeled data separately, making it difficult to improve the inference accuracy with ELBO. (2) Second, an *"ELBO bottleneck"* exists, and continuing to optimize the ELBO after a certain bottleneck value will not improve inference accuracy. Thus, we propose **S**moot**H**-ELBO **O**ptimal In**T**erpolation VAE (SHOT-VAE) to solve the *"good ELBO, bad performance"* problem without requiring additional prior knowledge, which offers the following contributions:

- **The smooth-ELBO objective that integrates the classification loss into ELBO**.
  We derive a new ELBO approximation named *smooth-ELBO* with the label-smoothing technique (Müller et al. 2019). Theoretically, we prove that the *smooth-ELBO* integrates the classification loss into ELBO. Then, we empirically show that a better inference accuracy can be achieved with *smooth-ELBO*.
- **The margin approximation that breaks the ELBO bottleneck.**
  We propose an approximation of the margin between the real data distribution and the one from ELBO. The approximation is based on the *optimal interpolation* in data space and latent space. In practice, we show this optimal interpolation approximation (*OT-approximation*) can break the *"ELBO bottleneck"* and achieve a better inference accuracy.
- **Good semi-supervised performance.**
  We evaluate SHOT-VAE on 4 benchmark datasets and the results show that our model achieves good performance with 25.30% error rate on CIFAR-100 with 10k labels and reduces the error rate to 6.11% on CIFAR-10 with 4k labels. Moreover, we find it can get strong results even with

*Corresponding author

fewer labels and smaller models, for example obtaining a 14.27% error rate on CIFAR-10 with 500 labels and 1.5M parameters.

Background

Semi-supervised VAE

In supervised learning, we are facing with training data that appears as input-label pairs $(\mathbf{X}, \mathbf{y})$ sampled from the labeled dataset $\mathbb{D}_L$. While in semi-supervised learning, we can obtain an extra collection of unlabeled data $\mathbf{X}$ denoted by $\mathbb{D}_U$. We hope to leverage the data from both $\mathbb{D}_L$ and $\mathbb{D}_U$ to achieve a more accurate model than only using $\mathbb{D}_L$.

Semi-supervised VAEs (Kingma et al. 2014) solve the problem by constructing a probabilistic model to disentangle the data into continuous variables $\mathbf{z}$ and label variables $\mathbf{y}$. It consists of a generation process and an inference process parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ respectively. The generation process assumes the posterior distribution of $\mathbf{X}$ given the latent variables $\mathbf{z}$ and $\mathbf{y}$ as

$$p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{z}, \mathbf{y}) = \mathcal{N}(\mathbf{X}; f_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{y}), \boldsymbol{\sigma}^2). \tag{1}$$

The inference process assumes the posterior distribution of $\mathbf{z}$ and $\mathbf{y}$ given $\mathbf{X}$ as

$$\begin{aligned} q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{X})); \\ q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X}) &= \mathrm{Cat}(\mathbf{y}|\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X})). \end{aligned} \tag{2}$$

where $\mathrm{Cat}(\mathbf{y}|\boldsymbol{\pi})$ is the multinomial distribution of the label, $\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X})$ is a probability vector, and the functions $f_{\boldsymbol{\theta}}$, $\boldsymbol{\mu}_{\boldsymbol{\phi}}$, $\boldsymbol{\sigma}_{\boldsymbol{\phi}}$ and $\boldsymbol{\pi}_{\boldsymbol{\phi}}$ are represented as deep neural networks.

To make the model learn disentangled representations, the independent assumptions (Kingma et al. 2014; Dupont 2018) are also widely used as

$$\begin{aligned} p(\mathbf{z}, \mathbf{y}) &= p(\mathbf{z})p(\mathbf{y}); \\ q_{\boldsymbol{\phi}}(\mathbf{z}, \mathbf{y}|\mathbf{X}) &= q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X}). \end{aligned} \tag{3}$$

For the unlabeled dataset $\mathbb{D}_U$, VAE models want to learn the disentangled representation of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})$ and $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ by maximizing the evidence lower bound of $\log p(\mathbf{X})$ as

$$\begin{aligned} \mathrm{ELBO}_{\mathbb{D}_U}(\mathbf{X}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{y}|\mathbf{X})}[\log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{z}, \mathbf{y})] \\ &- D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}) \| p(\mathbf{z})) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X}) \| p(\mathbf{y})). \end{aligned} \tag{4}$$

For the labeled dataset $\mathbb{D}_L$, the labels $\mathbf{y}$ are treated as latent variables and the related ELBO becomes

$$\begin{aligned} \mathrm{ELBO}_{\mathbb{D}_L}(\mathbf{X}, \mathbf{y}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X},\mathbf{y})}[\log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{z}, \mathbf{y})] \\ &- D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}, \mathbf{y}) \| p(\mathbf{z})). \end{aligned} \tag{5}$$

Considering the label prediction $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ contributes only to the unlabeled data in (4), which is an undesirable property as we wish the semi-supervised model can also learn from the given labels, Kingma et al. (2014) proposes to add a *cross-entropy (CE)* loss as a solution and the extended target is as follows:

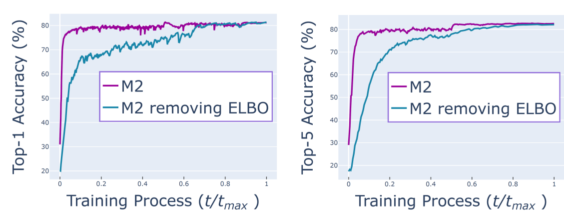
$$\begin{aligned} &\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{\mathbf{X}\sim\mathbb{D}_U}[-\mathrm{ELBO}_{\mathbb{D}_U}(\mathbf{X})] + \\ &\mathbb{E}_{(\mathbf{X},\mathbf{y})\sim\mathbb{D}_L}[-\mathrm{ELBO}_{\mathbb{D}_L}(\mathbf{X}, \mathbf{y}) + \alpha \cdot \mathrm{CE}(q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X}), \mathbf{y})] \end{aligned} \tag{6}$$

where α is a hyper-parameter controlling the loss weight.

Good ELBO, Bad Inference

However, a frequent phenomenon is that good ELBO values do not always imply accurate inference results (Zhao et al. 2017), which often occurs on realistic datasets with high variance, such as CIFAR-10 and CIFAR-100. In this paper, we investigate the training process of semi-supervised VAE models on the above two datasets and propose two possible causes of the *"good ELBO, bad inference"* problem.

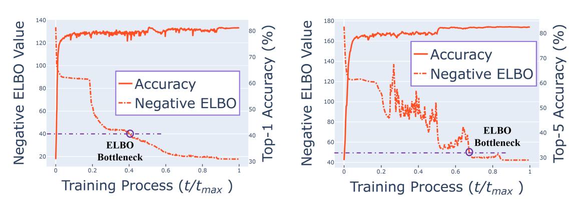
The ELBO cannot utilize the label information. As mentioned in equation (6), the label prediction $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ only contributes to the unlabeled loss $-\mathrm{ELBO}_{\mathbb{D}_U}(\mathbf{X})$, which indicates that the labeled loss $-\mathrm{ELBO}_{\mathbb{D}_L}(\mathbf{X}, \mathbf{y})$ can not utilize the label information directly. We assume this problem will make the ELBO value irrelevant to the final inference accuracy. To evaluate our assumption, we compare the semi-supervised VAE (M2) models (Kingma et al. 2014) with the same model but removing the $\mathrm{ELBO}_{\mathbb{D}_L}$ in (6). As shown in Figure 1, the results indicate that $\mathrm{ELBO}_{\mathbb{D}_L}$ can accelerate the learning process of $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$, but it fails to achieve a better inference accuracy than the one removing $\mathrm{ELBO}_{\mathbb{D}_L}$.



(a) CIFAR-10 (b) CIFAR-100

Figure 1: Test accuracy of semi-supervised VAE (M2) model with and w/o $\mathrm{ELBO}_{\mathbb{D}_L}$. Results indicates that the $\mathrm{ELBO}_{\mathbb{D}_L}$ fails to achieve a better inference accuracy.

The *"ELBO bottleneck"* effect. Another possible cause is the *"ELBO bottleneck"*, that is, continuing to optimize ELBO after a certain bottleneck value will not improve the inference accuracy. Figure 2 shows that the inference accuracy raises rapidly and peaks at the bottleneck value. After that, the optimization of ELBO value does not affect the inference accuracy.



(a) CIFAR-10 (b) CIFAR-100

Figure 2: Comparison between the negative ELBO value and accuracy for semi-supervised VAE. Results indicate that a ELBO bottleneck exists, and continuing to optimize ELBO after this bottleneck will not improve the inference accuracy.

Existing works introduce prior knowledge and specific structures to address these problems. Kingma et al. (2014), Maaløe et al. (2016) and Davidson et al. (2018) propose the stacked VAE structure (M1+M2), which forces the model to utilize the representations learned from $\text{ELBO}_{\mathbb{D}_L}$ to inference the $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$. Louizos et al. (2016) and Ilse et al. (2019) incorporate domain knowledge into models, making the ELBO representations relevant to the label prediction. Zhao et al. (2017) and Dupont (2018) utilize the ELBO decomposition technique (Hoffman and Johnson 2016), setting the mutual information bounds to perform feature selection. These methods have achieved great success on many benchmark datasets (e.g. MNIST, SVHN, Yale B). However, the related prior knowledge and structures need to be selected manually. Moreover, for some standard datasets with high variance such as CIFAR-10 and CIFAR-100, the semi-supervised performance of VAE is not satisfactory.

Instead of introducing additional prior knowledge, we propose a novel solution based on the ELBO approximations, **S**moot**H**-ELBO **O**ptimal in**T**erpolation VAE.

SHOT-VAE

In this section, we derive the SHOT-VAE model by introducing its two improvements. First, we derive a new ELBO approximation named *smooth-ELBO* that unifies the ELBO and the label predictive loss. Then, we create the differentiable *OT-approximation* to break the ELBO value bottleneck. Due to space limitations, the proofs are provided in the Appendix, which is available at Arxiv[1].

Smooth-ELBO: Integrating the Classification Loss into ELBO

To overcome the problem that the $\text{ELBO}_{\mathbb{D}_L}$ cannot utilize the label information directly, we first perform an *"ELBO surgery"*. Following previous works (Doersch 2016), the $\text{ELBO}_{\mathbb{D}_L}$ can be derived with *Jensen-Inequality* as:

$$\log p(\mathbf{X},\mathbf{y}) = \log \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X},\mathbf{y})} \frac{p(\mathbf{X},\mathbf{y},\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X},\mathbf{y})} \\ \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X},\mathbf{y})} \log \frac{p(\mathbf{X},\mathbf{y},\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X},\mathbf{y})} = \text{ELBO}_{\mathbb{D}_L}(\mathbf{X},\mathbf{y}) \tag{7}$$

Utilizing the independent assumptions in equation (3), we have $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X},\mathbf{y}) = q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})$. In addition, the labels $\mathbf{y}$ are treated as latent variables directly in $\text{ELBO}_{\mathbb{D}_L}$, which equals to obey the empirical degenerate distribution, i.e. $\hat{p}(\mathbf{y}_i|\mathbf{X}_i) = 1, \forall(\mathbf{X}_i,\mathbf{y}_i) \in \mathbb{D}_L$. Substituting the above two conditions into (7), we have

$$\text{ELBO}_{\mathbb{D}_L}(\mathbf{X},\mathbf{y}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}),\hat{p}(\mathbf{y}|\mathbf{X})} \log \frac{p(\mathbf{X},\mathbf{y},\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})\hat{p}(\mathbf{y}|\mathbf{X})} \\ = \mathbb{E}_{q_{\boldsymbol{\phi}},\hat{p}} \log p(\mathbf{X}|\mathbf{z},\mathbf{y}) - D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})\|p(\mathbf{z})) \\ - D_{\text{KL}}(\hat{p}(\mathbf{y}|\mathbf{X})\|p(\mathbf{y})). \tag{8}$$

In equation (8), the last objective $D_{\text{KL}}(\hat{p}(\mathbf{y}|\mathbf{X})\|p(\mathbf{y}))$ is irrelevant to the label prediction $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$, which causes the *"good ELBO, bad inference"* problem. Inspired by this, we derive a new ELBO approximation named *smooth-ELBO*.

[1]https://arxiv.org/abs/2011.10684

The *smooth-ELBO* provides two improvements. First, we propose a more flexible assumption of the empirical distribution $\hat{p}(\mathbf{y}|\mathbf{X})$. Instead of treating $\hat{p}(\mathbf{y}|\mathbf{X})$ as the degenerate distribution, we use the label smoothing technique (Müller et al. 2019) and view the one-hot label $\mathbf{1_y}$ as the parameters of the empirical distribution $\hat{p}(\mathbf{y}|\mathbf{X})$, that is, $\forall(\mathbf{X},\mathbf{y}) \in \mathbb{D}_L$

$$\hat{p}(\mathbf{y}|\mathbf{X}) = \text{Cat}(\mathbf{y}|\text{smooth}(\mathbf{1_y})); \\ \text{smooth}(\mathbf{1_y})_i = \begin{cases} 1-\epsilon & \text{if } \mathbf{1}_{\mathbf{y},i} = 1, \\ \frac{\epsilon}{K-1} & \text{if } \mathbf{1}_{\mathbf{y},i} = 0. \end{cases}, \tag{9}$$

where K is the number of classes and ϵ controls the smooth level. We use $\epsilon = 0.001$ in all experiments.

Then, we derive the following convergent approximation with the smoothed $\hat{p}(\mathbf{y}|\mathbf{X})$:

$$D_{\text{KL}}(\hat{p}(\mathbf{y}|\mathbf{X})\|q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})) + D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})\|p(\mathbf{y})) \\ \to D_{\text{KL}}(\hat{p}(\mathbf{y}|\mathbf{X})\|p(\mathbf{y})) \\ \text{when } q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X}) \to \hat{p}(\mathbf{y}|\mathbf{X}). \tag{10}$$

The proof can be found in Appendix A. Combining equations (8)-(10), we propose the *smooth-ELBO* for $\mathbb{D}_L$:

$$\text{smooth-ELBO}_{\mathbb{D}_L}(\mathbf{X},\mathbf{y}) \\ = \mathbb{E}_{q_{\boldsymbol{\phi}},\hat{p}} \log p(\mathbf{X}|\mathbf{z},\mathbf{y}) - D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})\|p(\mathbf{z})) \\ - D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})\|p(\mathbf{y})) - D_{\text{KL}}(\hat{p}(\mathbf{y}|\mathbf{X})\|q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})). \tag{11}$$

Theoretically, we demonstrate the following properties.

Smooth-ELBO integrates the classification loss into ELBO. Compared with the original $\text{ELBO}_{\mathbb{D}_L}$, *smooth-ELBO* derives two extra components, $D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})\|p(\mathbf{y}))$ and $D_{\text{KL}}(\hat{p}(\mathbf{y}|\mathbf{X})\|q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X}))$. Utilizing the decomposition in (Hoffman and Johnson 2016), we can rewrite the first component into

$$\mathbb{E}_{\mathbb{D}_L} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})\|p(\mathbf{y})) = \mathbf{I}_{q_{\boldsymbol{\phi}}}(\mathbf{X};\mathbf{y}) + D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{y})\|p(\mathbf{y}))$$

where $\mathbf{I}_{q_{\boldsymbol{\phi}}}(\mathbf{X};\mathbf{y})$ is the constant of mutual information between $\mathbf{X}$ and $\mathbf{y}$, $p(\mathbf{y})$ is the true marginal distribution for $\mathbf{y}$ which can be estimated with $\hat{p}(\mathbf{y}|\mathbf{X})$ in $\mathbb{D}_L$ and $q_{\boldsymbol{\phi}}(\mathbf{y}) = \frac{1}{|\mathbb{D}_L|}\sum_{(\mathbf{X},\mathbf{y})\in\mathbb{D}_L} q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ is the estimation of marginal distribution. By optimizing $D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{y})\|p(\mathbf{y}))$, the first component can learn the marginal distribution $p(\mathbf{y})$ from labels.

For the second component, with Pinsker's inequality, it's easy to prove that for all $i = 1, 2, \ldots, K$

$$|\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X}) - \text{smooth}(\mathbf{1_y})|_i \leq \sqrt{\frac{1}{2} D_{\text{KL}}(\hat{p}(\mathbf{y}|\mathbf{X})\|q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X}))}$$

The proof can be found in Appendix B, which indicates that $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ converges to $\hat{p}(\mathbf{y}|\mathbf{X})$ in training process.

Convergence analysis. As mentioned above, $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ converge to the smoothed $\hat{p}(\mathbf{y}|\mathbf{X})$ in training. Based on this property, we can assert that the *smooth-ELBO* converges to $\text{ELBO}_{\mathbb{D}_L}$ with the following equation:

$$|\text{smooth-ELBO}_{\mathbb{D}_L}(\mathbf{X},\mathbf{y}) - \text{ELBO}_{\mathbb{D}_L}(\mathbf{X},\mathbf{y})| \leq C_1\delta + C_2\frac{\delta^2}{\epsilon}$$

The proof can be found in Appendix C. C_1, C_2 are the constants related to class number K and $\delta = \sup_i |\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X})_i - \text{smooth}(\mathbf{1_y})_i|$ is the distance between $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ and $\hat{p}(\mathbf{y}|\mathbf{X})$.

To summarize the above, **smooth-ELBO can utilize the label information directly**. Compared with the original $-\text{ELBO}_{\mathbb{D}_L} + \alpha\text{CE}$ loss in (6), *smooth-ELBO* has three advantages. First, it not only learns from single labels, but also learns from the marginal distribution $p(\mathbf{y})$. Second, we do not need to manually set the loss weight α. Moreover, it also takes advantages of the $\text{ELBO}_{\mathbb{D}_L}$, such as disentangled representations and convergence assurance. The extensive experiments will also show that a better model performance can be achieved with smooth-ELBO.

OT-approximation: Breaking the ELBO Bottleneck

To overcome the *ELBO bottleneck* problem, we first analyze what the semi-supervised VAE model does after reaching the bottleneck, then we create the differentiable *OT-approximation* to break it, which is based on the *optimal interpolation* in latent space.

As mentioned in equation (4), VAE aims to learn disentangled representations $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})$ and $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ by maximizing the lower bound of the likelihood of data as $\log p(\mathbf{X}) \geq \text{ELBO}(\mathbf{X})$, while the margin between $\log p(\mathbf{X})$ and $\text{ELBO}(\mathbf{X})$ has the following closed form:

$$\log p(\mathbf{X}) - \text{ELBO}(\mathbf{X}) = D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})\|p(\mathbf{z},\mathbf{y}|\mathbf{X})) \tag{12}$$

Ideally, the optimization process of ELBO will make the representation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})$ and $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ converge to their ground truth $p(\mathbf{z}|\mathbf{X})$ and $p(\mathbf{y}|\mathbf{X})$. However, the unimproved inference accuracy of $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ indicates that **optimizing ELBO after the bottleneck will only contribute to the continuous representation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})$, while the $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$ seems to get stuck in the local minimum**. Since the ground truth $p(\mathbf{y}|\mathbf{X})$ is not available for the unlabeled dataset $\mathbb{D}_U$, it is hard for the model to jump out by itself. Inspired by this, we create a differentiable approximation of $D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})\|p(\mathbf{y}|\mathbf{X}))$ for $\mathbb{D}_U$ to break the bottleneck.

Following previous works (Lee 2013), the approximation is usually constructed with two steps: creating the pseudo input $\tilde{\mathbf{X}}$ with data augmentations and creating the pseudo distribution $\tilde{p}(\mathbf{y}|\tilde{\mathbf{X}})$ of $\tilde{\mathbf{X}}$. Recent advanced works use autoaugment (Cubuk et al. 2019) and random mixmatch (Berthelot et al. 2019) to perform data augmentations. However, these strategies will greatly change the representation of continuous variable $\mathbf{z}$, e.g., changing the image style and background. To overcome this, we propose the *optimal interpolation* based approximation.

The optimal interpolation consists of two steps. First, for each input $\mathbf{X}_0$ in $\mathbb{D}_U$, we find the optimal match $\mathbf{X}_1$ with the most similar continuous variable $\mathbf{z}$, that is, $\arg_{\mathbf{X}_1\in\mathbb{D}_U} \min D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}_0)\|q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}_1))$. Then, on purpose of jumping out the stuck point $q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})$, we take the widely-used mixup strategy (Zhang et al. 2018) to create pseudo input $\tilde{\mathbf{X}}$ as follows:

$$\tilde{\mathbf{X}} = (1-\lambda)\mathbf{X}_0 + \lambda\mathbf{X}_1, \tag{13}$$

where λ is sampled from the uniform distribution $\mathbf{U}(0,1)$.

The mixup strategy can be understood as calculating the optimal interpolation between two points $\mathbf{X}_0, \mathbf{X}_1$ in input space with the maximum likelihood:

$$\max_{\tilde{X}} (1-\lambda)\cdot\log(p_{\boldsymbol{\theta}}(\tilde{\mathbf{X}}|\mathbf{z}_0,\mathbf{y}_0)) + \lambda\cdot\log(p_{\boldsymbol{\theta}}(\tilde{\mathbf{X}}|\mathbf{z}_1,\mathbf{y}_1)),$$

where $\{\mathbf{z}_i,\mathbf{y}_i\}_{i=0,1}$ is the latent variables for the data points $\mathbf{X}_0, \mathbf{X}_1$, and the proof can be found in Appendix D.

To create the pseudo distribution $\tilde{p}(\mathbf{y}|\tilde{\mathbf{X}})$ of $\tilde{\mathbf{X}}$, it is a natural thought that **the optimal interpolation in data space could associate with the same in latent space** with D_{KL} distance used in ELBO. Inspired by this, we propose the optimal interpolation method to calculate $\tilde{p}(\mathbf{y}|\tilde{\mathbf{X}})$ as

Proposition 1 *The optimal interpolation derived from D_{KL} distance between $q_{\boldsymbol{\phi}}(\mathbf{y}|\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X}_0))$ and $q_{\boldsymbol{\phi}}(\mathbf{y}|\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X}_1))$ with $\lambda\in[0,1]$ can be written as*

$$\min_{\tilde{\boldsymbol{\pi}}} (1-\lambda)\cdot D_{\text{KL}}(\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X}_0)\|\tilde{\boldsymbol{\pi}}) + \lambda\cdot D_{\text{KL}}(\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X}_1)\|\tilde{\boldsymbol{\pi}})$$

and the solution $\tilde{\boldsymbol{\pi}}$ satisfying

$$\tilde{\boldsymbol{\pi}} = (1-\lambda)\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X}_0) + \lambda\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X}_1). \tag{14}$$

The proof can be found in Appendix E.

Combining the optimal interpolation in data space and latent space, we derive the **o**ptimal in**t**erpolation approximation (OT-approximation) for $\mathbb{D}_U$ as

$$\text{OT}_{\mathbb{D}_U}(\mathbf{X}_0,\mathbf{X}_1,\lambda) = D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{y}|\tilde{\mathbf{X}})\|\tilde{p}(\mathbf{y}|\tilde{\mathbf{X}}))$$
$$\textbf{s.t.} \begin{cases} \tilde{\mathbf{X}} = (1-\lambda)\mathbf{X}_0 + \lambda\mathbf{X}_1 \\ \tilde{p}(\mathbf{y}|\tilde{\mathbf{X}}) = \text{Cat}(\mathbf{y}|\tilde{\boldsymbol{\pi}}) \\ \tilde{\boldsymbol{\pi}} = (1-\lambda)\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X}_0) + \lambda\boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X}_1) \end{cases}. \tag{15}$$

Notice that the *OT-approximation* does not require additional prior knowledge and is easy to implement. Moreover, although *OT-approximation* utilizes the mixup strategy to create pseudo input $\tilde{\mathbf{X}}$, our work has two main advantages over mixup-based methods (Zhang et al. 2018; Verma et al. 2019). First, mixup methods directly assume the pseudo label $\tilde{\mathbf{y}}$ behaves linear in latent space without explanations. Instead, we derive the $\mathcal{D}_{KL}$ from ELBO as the metric and utilize the optimal interpolation (15) to construct $\tilde{\mathbf{y}}$. Second, mixup methods use $\|\cdot\|_2^2$ loss between $q_{\boldsymbol{\phi}}(\mathbf{y}|\tilde{\mathbf{X}})$ and $\tilde{p}(\mathbf{y}|\tilde{\mathbf{X}}))$, while we use the $\mathcal{D}_{KL}$ loss and achieve better semi-supervised learning performance.

Algorithm 1 SHOT-VAE training process with epoch t.

Input:
Batch of labeled data $(\mathbf{X}_L, \mathbf{y}_L)\in\mathbb{D}_L$;
Batch of unlabeled data $\mathbf{X}_U\in\mathbb{D}_U$;
Mixup $\lambda\sim\mathbf{U}(0,1)$;
Loss weight w_t;
Model parameters: $\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\phi}^{(t-1)}$;
Model optimizer: SGD

Output:
Updated parameters: $\boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}^{(t)}$

1: $L_{\mathbb{D}_L} = -\text{smooth-ELBO}_{\mathbb{D}_L}(\mathbf{X}_L, \mathbf{y}_L)$
2: $\mathbf{X}_U^0, \mathbf{X}_U^1 = \mathbf{X}_U, \text{OptimalMatch}(\mathbf{X}_U)$
3: $L_{\mathbb{D}_U} = -\text{ELBO}_{\mathbb{D}_U}(\mathbf{X}_U) + w_t\cdot\text{OT}_{\mathbb{D}_U}(\mathbf{X}_U^0, \mathbf{X}_U^1, \lambda)$
4: $L = L_{\mathbb{D}_L} + L_{\mathbb{D}_U}$
5: $\boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}^{(t)} = \text{SGD}(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\phi}^{(t-1)}, \frac{\partial L}{\partial\boldsymbol{\theta}}, \frac{\partial L}{\partial\boldsymbol{\phi}})$

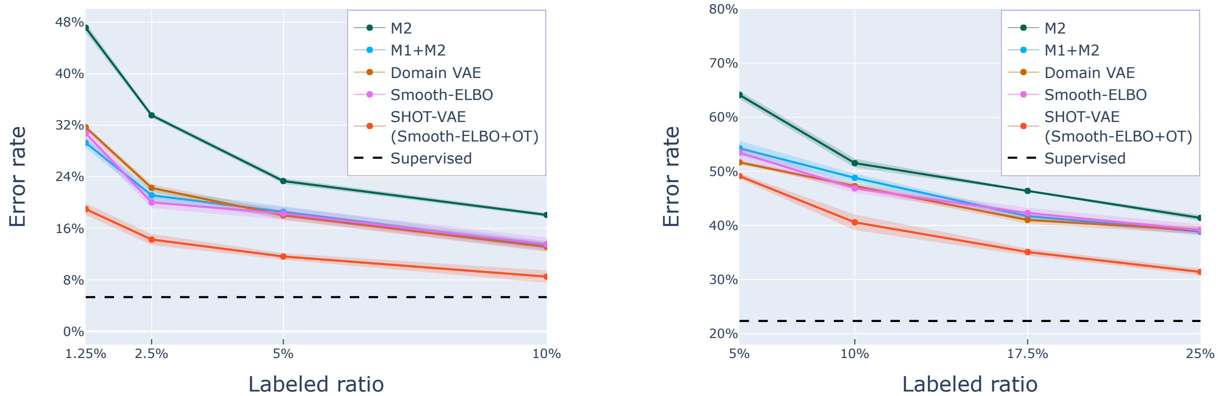


Figure 3: Error rate comparison of SHOT-VAE to baseline methods on CIFAR-10 (left) and CIFAR-100 (right) for a varying number of labels. "Supervised" refers to training with all 50000 training samples and no unlabeled data. Results show that (1) SHOT-VAE surpasses other models with a large margin in all cases. (2) both *smooth-ELBO* and *OT-approximation* contribute to the inference accuracy, reducing the error rate on 10% labels from 18.08% to 13.54% and from 13.54% to 8.51%.

The Implementation Details of SHOT-VAE

The complete algorithm of SHOT-VAE can be obtained by combining the *smooth-ELBO* and the *OT-approximation*, as shown in Algorithm 1. In this section, we discuss some details in training process.

First, the working condition for *OT-approximation* is that the ELBO has reached the bottleneck value. However, quantifying the ELBO bottleneck value is difficult. Therefore, we extend the *warm-up strategy* in $\beta -$ **VAE** (Higgins et al. 2017) to achieve the working condition. The main idea of *warm-up* is to make the weight w_t for OT-approximation increase slowly at the beginning and most rapidly in the middle of the training, i.e. exponential schedule. The function of the exponential schedule is $w_t = \exp(-\gamma \cdot (1 - \frac{t}{t_{\max}})^2)$, where γ is the hyper-parameter controlling the increasing speed, and we use $\gamma = 5$ in all experiments.

Second, the optimal match operation in equation (13) requires to find the most similar $\mathbf{X}_1$ for each $\mathbf{X}_0$ in $\mathbb{D}_U$, which consumes a lot of computation resources. To overcome this, we set a large batch-size (e.g., 512) and use the most similar $\mathbf{X}_1$ in one mini-batch to perform optimal interpolation.

Moreover, calculating the gradient of the expected log-likelihood $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{y}|\mathbf{X})} \log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{z},\mathbf{y})$ is difficult. Therefore, we apply the reparameterization tricks (Rezende et al. 2014; Jang et al. 2017) to obtain the gradients as follows:

$$\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X})} \log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{z}) \approx (\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I}))$$
$$\frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \log p_{\boldsymbol{\theta}}(\mathbf{X}|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{X}) + \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{X}) \odot \boldsymbol{\epsilon}_i).$$

and

$$\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X})} \log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{y}) \approx (\boldsymbol{\delta}_i \sim \mathbf{Gumbel}(\boldsymbol{\epsilon};\mathbf{0},\mathbf{1}))$$
$$\frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathrm{Softmax}(\frac{\log \boldsymbol{\pi}_{\boldsymbol{\phi}}(\mathbf{X}) + \boldsymbol{\delta}_i}{\tau})).$$

Following previous works (Dupont 2018), we used $N = 1$ and $\tau = 0.67$ in all experiments. Moreover, to make the VAE model learn the disentangled representations, we also take the widely-used β-VAE strategy (Burgess et al. 2018) in training process and chose $\beta = 0.01$ in all experiments.

Experiments

In this section, we evaluate the SHOT-VAE model with sufficient experiments on four benchmark datasets, i.e. MNIST, SVHN, CIFAR-10, and CIFAR-100. In all experiments, we apply stochastic gradient descent (SGD) as optimizer with momentum 0.9 and multiply the learning rate by 0.1 at regularly scheduled epochs. For each experiment, we create five $\mathbb{D}_L$-$\mathbb{D}_U$ splits with different random seeds and the error rates are reported by the mean and variance across splits. Due to space limitations, we mainly show results on CIFAR-10 and CIFAR-100; more results on MNIST and SVHN as well as the robustness analysis of hyper-parameters are provided in Appendix F. The code, with which the most important results can be reproduced, is available at Github[2].

Smooth-ELBO Improves the Inference Accuracy

In the above sections, We propose *smooth-ELBO* as the alternative of $-\mathrm{ELBO}_{\mathbb{D}_L}$+ CE loss in equation (6), and analyze the convergence theoretically. Here we evaluate the inference accuracy and convergence speed of *smooth-ELBO*.

[2]https://github.com/FengHZ/SHOT-VAE

Parameter Amount	Method	CIFAR10 (4k)	CIFAR100 (4k)	CIFAR100 (10k)
1.5 M	VAT	13.13	/	37.78
	Π-Model	16.37	/	39.19
	Mean Teacher	15.87	44.71	38.92
	CT-GAN	10.62	45.11	37.16
	LP	11.82	43.73	35.92
	Mixup	10.71(±0.44)	46.61(±0.88)	38.62(±0.67)
	SHOT-VAE	**8.51**(±0.32)	**40.58**(±0.48)	**31.41**(±0.21)
36.5 M	Π-Model	12.16	/	31.12
	Mean Teacher	6.28	36.63	27.71
	MixMatch	**5.53**	35.62	25.88
	SHOT-VAE	6.11(±0.34)	**33.76**(±0.53)	**25.30**(±0.34)

Table 1: Error rate comparison of SHOT-VAE to baseline models on CIFAR-10 and CIFAR-100 with 4k and 10k labels in different parameter amounts. The results show that SHOT-VAE outperforms other advanced methods on CIFAR-100. Moreover, our model is not sensitive to the parameter amount. For example, the accuracy on CIFAR-10 only loses 2% when the model size decreases 24 times.

t/t_{max}	CIFAR-10	CIFAR-100
0.1	0.71(±0.24)%	2.79(±0.6)%
0.5	0.39(±0.13)%	1.61(±0.57)%
1	0.11(±0.04)%	0.46(±0.11)%

Table 2: Relative error of smooth-ELBO.

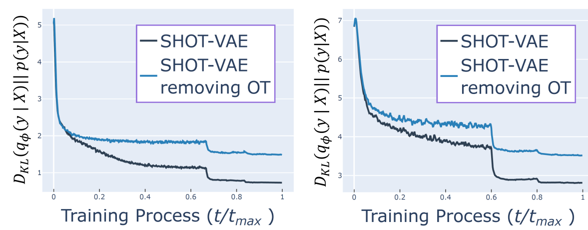
First, we compare the *smooth-ELBO* with other semi-supervised VAE models under a varying label ratios from 1.25% to 25%. As baselines, we consider three advanced VAE models mentioned above: standard semi-supervised VAE (M2), stacked-VAE (M1+M2), and domain-VAE (Ilse et al. 2019).

As shown in Figure 3, *smooth-ELBO* makes VAE model learn better representations from labels, reducing the error rates among all label ratios on CIFAR-10 and CIFAR-100 respectively. Moreover, *smooth-ELBO* also achieves competitive results to other VAE models without introducing additional domain knowledge or multi-stage structures.

Second, we analyze the convergence speed in training process. As mentioned above, the *smooth-ELBO* will converge to the real ELBO when $q_\phi(\mathbf{y}|\mathbf{X}) \to \hat{p}(\mathbf{y}|\mathbf{X})$. Moreover, we also discover that $q_\phi(\mathbf{y}|\mathbf{X})$ converges to $\hat{p}(\mathbf{y}|\mathbf{X})$ in training process. Here we evaluate the convergence speed in training process with the relative error between the *smooth-ELBO* and the real ELBO. As shown in Table 2, the relative error can be very low even at the early stage of training, that is, 0.71% on CIFAR-10 and 2.79% on CIFAR-100, which indicates that the *smooth-ELBO* converges rapidly.

SHOT-VAE Breaks the ELBO Bottleneck

We make two assertions in SHOT-VAE: (1) optimizing ELBO after the bottleneck will make $q_\phi(\mathbf{y}|\mathbf{X})$ get stuck in the local minimum. (2) The OT-approximation can break the ELBO bottleneck by making good estimation of $D_{\mathrm{KL}}(q_\phi(\mathbf{y}|\mathbf{X})\|p(\mathbf{y}|\mathbf{X}))$. We evaluate the assertions



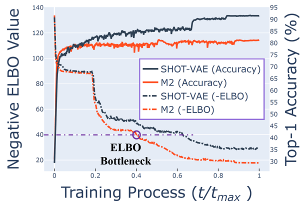
(a) CIFAR-10 (b) CIFAR-100

Figure 4: The $D_{\mathrm{KL}}(q_\phi(\mathbf{y}|\mathbf{X})\|\hat{p}(\mathbf{y}|\mathbf{X}))$ in $\mathbb{D}_U$ with and w/o *OT-approximation*. Results indicate that the OT approximation bridges the gap between $q_\phi(\mathbf{y}|\mathbf{X})$ and $\hat{p}(\mathbf{y}|\mathbf{X})$ in $\mathbb{D}_U$, making $q_\phi(\mathbf{y}|\mathbf{X})$ jump out the local minimum.

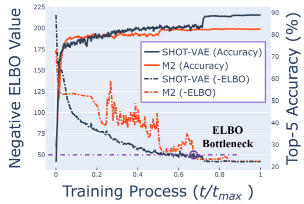
through two stage experiments.

First, to evaluate the "local minimum" assertion, we utilize the label of $\mathbb{D}_U$ to estimate the empirical distribution $\hat{p}(\mathbf{y}|\mathbf{X})$ and calculate $D_{\mathrm{KL}}(q_\phi(\mathbf{y}|\mathbf{X})\|\hat{p}(\mathbf{y}|\mathbf{X}))$ in training process as the metric. Notice these labels are only used to calculate the metric and do not contribute to the model. As shown in Figure 4, we compare the SHOT-VAE with the same model but removing the *OT-approximation*. The results indicate that optimizing ELBO itself without *OT-approximation* will make the gap $D_{\mathrm{KL}}(q_\phi(\mathbf{y}|\mathbf{X})\|\hat{p}(\mathbf{y}|\mathbf{X}))$ stuck into the local minimum, while the *OT approximation* helps the model jump out the local minimum, leading to a better inference of $q_\phi(\mathbf{y}|\mathbf{X})$.

Then, we investigate the relation between the negative ELBO and inference accuracy for SHOT-VAE and M2 model. As shown in Figure 5, for M2 model, the inference accuracy stalls after the ELBO bottleneck. While for SHOT-VAE, optimizing ELBO contributes to the improvement of inference accuracy during the whole training process. Moreover, the SHOT-VAE achieves a much better accuracy than M2, which also indicates that SHOT-VAE breaks the "*ELBO bottleneck*".

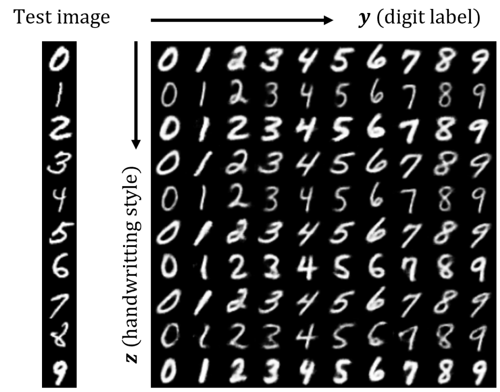


(a) CIFAR-10

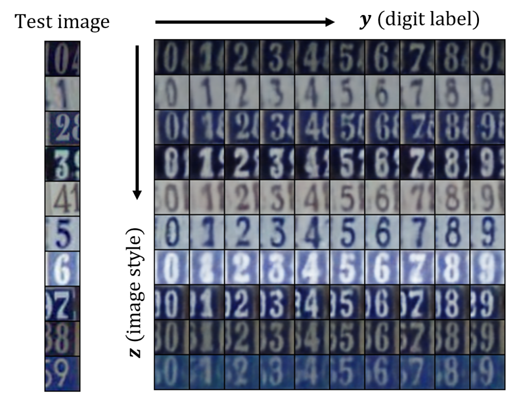


(b) CIFAR-100

Figure 5: Comparison between the negative ELBO value and test accuracy for SHOT-VAE and M2 model. Results indicate that SHOT-VAE breaks the ELBO bottleneck.



(a) MNIST



(b) SVHN

Figure 6: The conditional generation results of SHOT-VAE. The leftmost columns show images from the test set and the other columns show the conditional generation samples with the learned representation. It indicates that $\mathbf{z}$ and $\mathbf{y}$ have learned disentangled representations in latent space, as $\mathbf{z}$ represents the image style and $\mathbf{y}$ represents the digit label.

Semi-supervised Learning Performance

We evaluate the effectiveness of the SHOT-VAE on two parts: evaluations under a varying number of labeled samples and evaluations under different parameter amounts of neural networks.

First, we compare the SHOT-VAE with other advanced VAE models under a varying label ratios from 1.25% to 25%. As shown in Figure 3, both *smooth-ELBO* and *OT-approximation* contribute to the improvement of inference accuracy, reducing the error rate on 10% labels from 18.08% to 13.54% and from 13.54% to 8.51%, respectively. Furthermore, SHOT-VAE outperforms all other methods by a large margin, e.g., reaching an error rate of 14.27% on CIFAR-10 with the label ratio 2.5%. For reference, with the same backbone, fully supervised training on all 50000 samples achieves an error rate of 5.33%.

Then, we evaluate SHOT-VAE under different parameter amounts of neural networks, i.e. "WideResNet-28-2" with 1.5M parameters and "WideResNet-28-10" with 36.5M parameters. As baselines for comparison, we select six current best models from 4 categories: Virtual Adversarial Training (VAT Miyato et al. (2019)) and MixMatch (Berthelot et al. 2019) which are based on data augmentation, Π-model (Laine and Aila 2017) and Mean Teacher (Tarvainen and Valpola 2017), based on model consistency regularization, Label Propagation (LP) (Iscen et al. 2019) based on pseudo-label and CT-GAN (Wei et al. 2018, 2019) based on generative models. The results are presented in Table 1. Besides, we also take the mixup-based method into consideration. In general, the SHOT-VAE model outperforms other methods among all experiments on CIFAR-100. Furthermore, our model is not sensitive to the parameter amount and reaches competitive results even with small networks (e.g., 1.5M parameters).

Disentangled Representations

Among semi-supervised models, VAE based approaches have great advantages in interpretability by capturing semantics-disentangled latent variables. To demonstrate this property, we perform conditional generation experiments on MNIST and SVHN datasets. As shown in Figure 6, we pass the test image through the inference network to obtain the distribution of the latent variables $\mathbf{z}$ and $\mathbf{y}$ corresponding to this image. We then fix the inference $q_\phi(\mathbf{z}|\mathbf{X})$ of continuous variable $\mathbf{z}$, vary $\mathbf{y}$ with different labels, and generate new samples. The generation results show that $\mathbf{z}$ and $\mathbf{y}$ have learned semantic-disentangled representations, as $\mathbf{z}$ represents the image style and $\mathbf{y}$ represents the classification contents. Moreover, by comparing the results through columns, we find that each dimension of the discrete variable $\mathbf{y}$ corresponds to one class label separately.

Conclusions

We investigate one challenge in semi-supervised VAEs that "good ELBO values do not imply accurate inference results". We propose two causes of this problem through reasonable experiments. Based on the experiment results, we propose SHOT-VAE to address the *"good ELBO, bad inference"* problem. With extensive experiments, We demonstrate that SHOT-VAE can break the ELBO value bottleneck without introducing additional prior knowledge. Results also show that our SHOT-VAE outperforms other advanced semi-supervised models. Moreover, the SHOT-VAE model is robust to the parameter amount of models.

Acknowledgments

This research has been supported by National Key Research and Development Program (2019YFB1404802) and National Natural Science Foundation of China (U1866602, 61772456).

References

Berthelot, D.; Carlini, N.; Goodfellow, I. J.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. *CoRR* abs/1905.02249.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in β-VAE. *CoRR* abs/1804.03599.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. AutoAugment: Learning Augmentation Strategies From Data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Davidson, T. R.; Falorsi, L.; Cao, N. D.; Kipf, T.; and Tomczak, J. M. 2018. Hyperspherical Variational Auto-Encoders. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, 856–865.

Doersch, C. 2016. Tutorial on Variational Autoencoders. *CoRR* abs/1606.05908.

Dupont, E. 2018. Learning Disentangled Joint Continuous and Discrete Representations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, 708–718.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Hoffman, M. D.; and Johnson, M. J. 2016. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2.

Ilse, M.; Tomczak, J. M.; Louizos, C.; and Welling, M. 2019. DIVA: Domain Invariant Variational Autoencoder. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*.

Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2019. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5070–5079.

Jang, E.; Gu, S.; Poole, B.; et al. 2017. Categorical Reparameterization with Gumbel-Softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Kingma, D. P.; Mohamed, S.; Rezende, D. J.; Welling, M.; et al. 2014. Semi-supervised Learning with Deep Generative Models. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 3581–3589.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Lee, D.-H. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2.

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. S. 2016. The Variational Fair Autoencoder. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Maaløe, L.; Sønderby, C. K.; Sønderby, S. K.; and Winther, O. 2016. Auxiliary Deep Generative Models. In Balcan, M.; and Weinberger, K. Q., eds., *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, 1445–1453. JMLR.org.

Miyato, T.; Maeda, S.; Koyama, M.; and Ishii, S. 2019. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(8): 1979–1993. doi:10.1109/TPAMI.2018.2858821.

Müller, R.; Kornblith, S.; Hinton, G. E.; et al. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 4696–4705.

Narayanaswamy, S.; Paige, B.; van de Meent, J.; Desmaison, A.; Goodman, N. D.; Kohli, P.; Wood, F. D.; and Torr, P. H. S. 2017. Learning Disentangled Representations with Semi-Supervised Deep Generative Models. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 5925–5935.

Rezende, D. J.; Mohamed, S.; Wierstra, D.; et al. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 1278–1286.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 1195–1204.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold Mixup:

Better Representations by Interpolating Hidden States. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 6438–6447.

Wei, K.; Deng, C.; Yang, X.; et al. 2020. Lifelong Zero-Shot Learning. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 551–557. ijcai.org. doi:10.24963/ijcai.2020/77. URL https://doi.org/10.24963/ijcai.2020/77.

Wei, K.; Yang, M.; Wang, H.; Deng, C.; and Liu, X. 2019. Adversarial Fine-Grained Composition Learning for Unseen Attribute-Object Recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 3740–3748. IEEE. doi:10.1109/ICCV.2019.00384. URL https://doi.org/10.1109/ICCV.2019.00384.

Wei, X.; Gong, B.; Liu, Z.; Lu, W.; and Wang, L. 2018. Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Zhao, S.; Song, J.; Ermon, S.; et al. 2017. InfoVAE: Information Maximizing Variational Autoencoders. *CoRR* abs/1706.02262.