Unified Robust Semi-Supervised Variational Autoencoder

Xu Chen [1]

Abstract

In this paper, we propose a novel noise-robust semi-supervised deep generative model by jointly tackling noisy labels and outliers simultaneously in a unified robust semi-supervised variational autoencoder (URSVAE). Typically, the uncertainty of of input data is characterized by placing the uncertainty prior on the parameters of probability density distributions in order to ensure the robustness of the variational encoder towards outliers. Subsequently, a noise transition model is integrated naturally into our model to alleviate the detrimental effects of noisy labels. Moreover, a robust divergence measure is employed to further enhance the robustness, where a novel variational lower bound is derived and optimized to infer the network parameters. By proving that the influence function of the proposed evidence lower bound is bounded, the enormous potential of the proposed model in the classification in the presence of the compound noise is demonstrated. The experimental results highlight the superiority of the proposed framework by the evaluating on image classification tasks and comparing with the state-of-the-art approaches.

1. Introduction

The recent success of training deep neural network has been heavily relying on the collection of large datasets of high quality labels. Semi-supervised deep neural network (Kingma et al., 2014) serves as an effective alternative approach to alleviate the challenges of labeling large datasets (Tanno et al., 2019)(Maaløe et al., 2017). However, these alternative solutions inevitably introduce the label noise due to human fatigue and the subjective nature of annotators, which can deteriorate the performance of deep neural network. Moreover, the problem of semi-supervised deep learning with noisy labels becomes more complicated when the input image contains outliers where their behaviors are far away from normal samples (Bora et al., 2018). For instance, due to radiation or patient motions during the imaging process, the deviations in neuroimaging data are generated. In parallel, Bayesian deep learning (BDL) (Li & Gal, 2017)(Huang et al., 2018) has drawn extensive attention due to its capability of transforming the problem of posterior inference of a BDL model into the optimization of an objective function, which is expressed as an expectation of an analytic function of latent variables. The question that is then asked is: how can we design a unified deep generative model in order to counter noisy labels and outliers in one shot simultaneously in a semi-supervised Bayesian deep learning ? In this paper, we address this challenging problem by *developing a unified robust semi-supervised variational autoencoder*.

Towards learning with noisy labels, semi-supervised learning with noisy labels have achieved remarkable success including the work at (Malach & Shwartz, 2017)(Jiang et al., 2018)(Patrini et al., 2017a). In (Kaneko et al., 2019), label-noise robust generative adversarial network is proposed by incorporating a noise transition model to learn a clean label conditional generative distribution even when training labels are noisy. Targeting at reducing the noise on the input data with generative models, AmbientGAN (Bora et al., 2018) considered the task of learning an implicit generative model given only lossy measurements of samples from the distribution of interest. Despite the separate progress of learning with noisy labels and outliers, their connection has not been well explored in semi-supervised learning settings. Previous work on learning with noisy labels (Patrini et al., 2017a) usually focuses on correcting the loss function based on estimation of the noise transition matrix. However, the limitations of these approaches is that they can not handle the case with high noise ratio for corrupted labels. Moreover, none of them has considered the influences of sample outliers. To overcome these challenges, our work take a different perspective by proposing a robust solution with variational Bayesian approaches. In (Hou & Galata, 2008), a variational Bayesian approach has been applied to the problem of robust estimation of gaussian mixtures from noisy input data. However, the method described in (Hou & Galata, 2008) is mainly designed for clustering in an unsupervised learning setting and therefore it does not take into

[1] Cary, NC, USA. Correspondence to: Xu Chen <steven.xu.chen@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

account of the more complicated and interesting problem of robust semi-supervised learning with noisy labels in deep neural network.

The work (Maaløe et al., 2017) has leveraged auxiliary variables to enhance deep generate models for more accurate variational approximation for semi-supervised learning when the labels and the data are noise-free. Inspired by that, we propose a novel hierarchical variational inference model to characterize the uncertainties of the data with Gaussian mixture models in light of outliers. In the work of Futami et al (Futami et al., 2018), an outlier-robust pseudo-Bayesian variational method by replacing the Kullback-Leibler divergence used for data fitting to a robust divergence such as the β-divergence. Realizing the effectiveness of robust divergence (Futami et al., 2018) to overcome outliers, our work leverages the robust divergence on the hierarchical model in the semi-supervised learning by incorporating a noise transition model for noisy labels and the robust divergence jointly to counter the compound noise.

1.1. Summary of Contributions

The major contribution of this paper can be summarized as:

1. To the best of our knowledge, this is the first work that a novel robust semi-supervised deep generative model is proposed to handle noisy labels and sample outliers in one shot simultaneously.

2. To alleviate the compound noise, the proposed model proposes to leverage three denoising schemes to ensure the robustness of the proposed system:

    - First, URSVAE places the uncertainty priors on the parameters of the mixture components to model the noisy input data and meanwhile adapt noise transition models to characterize the noisy labels in the robust semi-supervised learning. The proposed novel denoising schemes and architecture have effectively reduced the detrimental effect of the compound noise.
    - Subsequently, the robust β-divergence is employed to replace Kullback-Leibler divergence used for data fitting to infer the network parameters and a novel evidence lower bound for sem-supervised learning is derived. Theoretical analysis on the influence function of the lower bound ensures the robustness of the proposed model.
    - Our URSVAE achieves significantly improved performance in the classification of data with noisy labels and outliers by evaluation on multiple benchmark datasets and comparison with the state-of-the-art approaches.

1.2. Related Work

Compared to the work using auxiliary deep generative models (Maaløe et al., 2017) to strengthen the expressive power of the generative model, our work extend the graphical models by placing the uncertainty priors on the first and second order statistics of the Gaussian mixture models and deriving the novel ELBO based on the robust β-divergence, aiming at resolving a more challenging problem for robust semi-supervised learning. Our work also differs from (Futami et al., 2018) in the sense that, the work (Futami et al., 2018) studies the theory of variational inference using robust divergence, while our work cohesively adapts uncertainty prior and noise transitional model into a robust semi-supervised variational autoencoder and demonstrate the enormous potential of our model in image classification with compound noise. Meanwhile, in the area of robust learning with deep neural work, the work (Li et al., 2019) tackles the learning with noise problem by applying a noise-tolerant training algorithm relying on a meta-learning update. P-correction (Yi & Wu, 2019) mitigates the influence from the noisy labels by training an end-to-end framework in order to update network parameters and label estimations dynamically. More work along this direction recently includes Iterative-CV (Chen et al., 2019) which employs cross-validation to randomly split noisy datasets and adopts Coteaching(Yu et al., 2019) techniques to train DNNs robustly against noisy labels. More recently, Dividemix (Li et al., 2020) characterizes the per-sample loss using a mixture model to dynamically split the training data into a labeled set with clean samples and an unlabeled set with noisy samples so as to train two diverged networks simultaneously. However, all of these work donot explore the challenging issue of semi-supervised learning with noisy labels and outliers simultaneously.

2. Our Approach

2.1. Variational AutoEncoder

Variational autoencoder (Diederik & Welling, 2013) has been recently proposed as a powerful solution for semi-supervised learning. Typically, variational inference with deep learning from powerful probabilistic models are constructed by an inference neural network $q(z|x)$ and a generative neural network $p(x|z)$. The generative model and the inference model are parameterized by θ and ϕ respectively. Subsequently, the network parameters are estimated by optimizing the evidence lower bound in the variational inference. For unsupervised learning, the evidence lower bound (ELBO) for vanilla VAE is represented as (Maaløe et al., 2017):

$$\ln p(x) \geq \mathrm{E}_{q(z|x)}[\ln p(x|z)] - \mathrm{D}_{KL}[q(z|x)//p(z)] \quad (1)$$

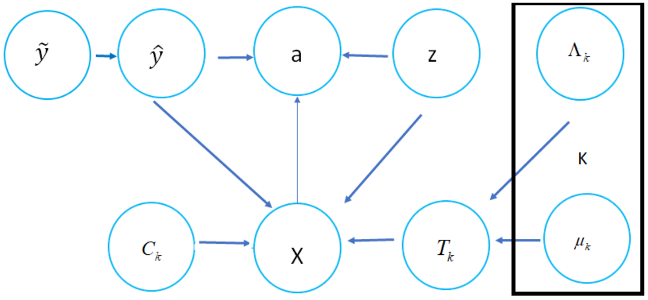


Figure 1. Probabilistic generative model for the proposed unified robust semi-supervised variational autoencoder (URSVAE), where x denotes the observations, $\tilde{y}$ and $\hat{y}$ refer to noisy labels and corrected labels respectively, z and a stand for the latent variable and auxiliary variable for VAE respectively, T_k represents the mean of the kth Gaussian component which satisfies the normal distribution with the mean μ_k and the precision matrix Λ_k and C_k represents the variance of the kth Gaussian component.

2.2. Unified Robust Semi-supervised Variational AutoEncoder (URSVAE)

Now we introduce our novel framework of unified robust semi-supervised deep generative model with the main focus of constructing a semi-supervised deep generative model that is robust to both of the outliers and noisy labels. As the outliers of input data poses a serious problem for the generative tasks in the sampling process for the learnt distribution, motivated by the fact that the student-t distribution is more robust to the outliers than Gaussian distribution by constraining the shapes of the mixture components from collapsing, shown in Fig.1, the proposed deep generative model integrates the uncertainty model by modeling the input data X with a Gaussian mixture model and placing the uncertainty prior on the parameters of the mixture components. Here the number of mixture components K is determined from the number of classes in the labeled data. Specifically, x_n is a noisy measurement of its true position and is a draw from the Gaussian mixture model, where the mean of each Gaussian component T_k is unknown and the variance C_k is known. In order to characterize the uncertainty and the outliers from the input, the Gaussian prior is placed on the top of the mean for each Gaussian component. Namely, T_k satisfies the normal distribution with the mean μ_k and the precision matrix Λ_k. ω_i is the latent variable for the ith data point specifying which Gaussian it came from and π is the mixing weight for the Gaussian mixture model. Specifically, a Normal-Wishart prior (Murphy, 2007) is placed on the mean and precision of the Gaussian components: $p(\mu, \Lambda) = \prod_{k=1}^{K} N(\mu_k|m_0, (\beta_0\Lambda_k)^{-1})W(\Lambda_k|W_0, \nu_0)$, where m_0 is set to be zero and β_0 is set to be a very small value. $W(\Lambda|W, \nu)$ is a Wishart distribution with scale matrix W and ν degrees of freedom. For a semi-supervised learning setting, the labels y are either unknown (for unlabeled data) or noisy (for labeled data). The generative model is then defined as: $p_\theta(x|\omega, T, C)p_\theta(T|Z, \mu, \Lambda)p_\theta(a|z, \tilde{y}, x)p_\theta(x|\tilde{y}, z)p(z)p(\tilde{y})$ $p(\omega)$. Define p_θ as the deep neural network with parameters θ and y as the ground truth of class label. For unlabeled data, y is considered as a latent variable. Further denote $\mathrm{Cat}(.)$ as a multinomial distribution and in this paper we reasonably assume that the labels are of categorical distribution and the proposed model applies to other distributions for the latent variable y. In order to fit more complicated posteriors for the marginal distribution $p(z|x)$, motivated by (Maaløe et al., 2017), we extend the variational distribution with auxiliary variables a, so that the generative model is invariant to marginalization over a $p(x, z, a, \omega, T, C, \mu, \Lambda) = p_\theta(x|\omega, T, C)p_\theta(T|Z, \mu, \Lambda)p_\theta(a|z, x)p_\theta(x, z)$. To mitigate the influence with the noisy labels, we introduce $\tilde{y}$ as the corrupted labels and $\hat{y}$ as the clean label and not observed during the training which is connected with the $K \times K$ noise transition matrix M where estimation of the noise transition matrix has been addressed in previous methods (Sukhbaatar et al., 2015)(Patrini et al., 2017b). In particular, $M = (M_{i,j}) \in [0,1]^{c\times c}(\sum_i M_{i,j} = 1)$. Further denote the corrupted labels as $\tilde{y}$ and the corrected labels as $\hat{y}$. The proposed generative model can then be expressed as:

$$
\begin{aligned}
p(z) &= N(z|0, I) \ , \\
p(\tilde{y}) &= \mathrm{Cat}(\tilde{y}|\eta) \ , \\
p_\theta(a|z, \tilde{y}, x) &= f(a; z, \tilde{y}, x) \ , \\
p_\theta(x|z, \tilde{y}) &= f(x; z, \tilde{y}, \theta) \ , \\
p(x|\omega, T, C) &= \textstyle\prod_{n=1}^{N_u} \prod_{k=1}^{K} \pi_k^{\omega_{nk}} N(x_n|t_k, C_k)^{\omega_{nk}} \\
&\quad \textstyle\prod_{n=1}^{N_l} N(t_n|t_{\tilde{y}_n}, C_{\tilde{y}_n}) \\
p(T_k|\mu_k, \gamma_k) &= N(t_k|\mu_k, \Lambda_k^{-1}) \\
p(\omega|\pi) &= \textstyle\prod_{n=1}^{N_u} \prod_{k=1}^{K} \pi_k^{\omega_{nk}} \\
p(\hat{y} = i|\tilde{y} = j) &= M_{ij} \ ,
\end{aligned} \tag{2}
$$

The inference model can be represented as:

$$
\begin{aligned}
q_\phi(a, z, \mu, \Lambda, T, \tilde{y}, \omega|x) &= q(z|a, \tilde{y}, x)q(a|x) \\
&\quad q(\tilde{y}|a, x)q(T, \mu, \Lambda, \omega|x)
\end{aligned} \tag{3}
$$

which can be further factorized as

$$q_\phi(z|a, \tilde{y}, x) = N(z|\mu_\phi(a, \tilde{y}, x), diag(\sigma^2)), \tag{4}$$

$$q_\phi(\tilde{y}|a, x) = \mathrm{Cat}(\tilde{y}|\eta_\phi(a, x)), \tag{5}$$

$$q_\phi(\mu_k, \Lambda_k) = q(\mu_k|\Lambda_k)q(\Lambda_k) \tag{6}$$

In order to compute $q(T, \mu, \Lambda, \omega|x)$, we utilize the mean-field approximation approach (Bishop, 2006) to factorize all the latent variables and parameters:

$$
\begin{aligned}
&q(T, \mu, \Lambda, \omega|x) \approx q(T|\omega, x)q(\omega|x)q(\mu, \Lambda|x) = \\
&[\textstyle\prod_i Cat(\omega_i|r_i)][\mathrm{Dir}(\pi|\alpha)\prod_k N(\mu_k|m_k, (\beta_k\Lambda_k)^{-1}) \\
&W(\Lambda_k|W_k, \nu_k)
\end{aligned} \tag{7}
$$

To characterize the normal distributions of $p_\theta(a|z, \tilde{y}, x)$, $p_\theta(x|z, y)$ and $q_\phi(z|a, \tilde{y}, x)$, two separate outputs from the top deterministic layer in each deep neural network are defined as $\mu_{\phi\vee\theta}(.)$ and $\log \sigma^2_{\phi\vee\theta}(.)$. Thus, the reparameterization trick (Diederik & Welling, 2013) can be employed to approximate the expectations from the output.

2.3. Robust Divergences for VAE

Given the uncertainty prior on our model, we now briefly introduce the robust divergence to further alleviate the impact from outliers. Introduced theoretically at (Basu et al., 1998), the β-divergence between two functions g and f are defined as

$$D_\beta(g \parallel f) = \tfrac{1}{\beta}\int g(x)^{1+\beta}dx + \int f(x)^{1+\beta}dx \\ -\tfrac{\beta+1}{\beta}\int g(x)f(x)^\beta dx \qquad (8)$$

When $\beta \to 0$, the β-divergence converges to KL-divergence, $\lim_{\beta\to 0} D_\beta(g \parallel f) = D_{KL}(g \parallel f)$. As described in (Futami et al., 2018), minimizing the β-divergence from the empirical distribution $\hat{p}(x)$ to $p(x;\theta)$

$$\arg\min_\theta D_\beta(\hat{p}(x) \parallel p(x;\theta)) \qquad (9)$$

we obtain

$$\frac{1}{N}\sum_{i=1}^{N} p(x_i;\theta)^\beta \frac{\partial \ln p(x_i;\theta)}{\partial\theta} - \mathrm{E}_{p(x;\theta)}[p(x;\theta)^\beta \frac{\partial \ln p(x;\theta)}{\partial\theta}]$$

As the probability densities of outliers are usually much smaller than those of inliers, the first term of the above equation is the likelihood weighted according to the power of the probability density for each sample, which effectively suppress the likelihood of outliers. This estimator is also called as M-estimator (Huber & Ronchetti, 2011), which provides provably superior performance in various machine applications (Li & Gal, 2017).

2.4. Variational Lower Bound on β-Divergence for URSVAE

In order to infer the network parameters for robust optimization, we shall now derive the variational lower bound based on β-divergence (β-ELBO) for our URSVAE. The β-ELBO for VAE using robust β-divergence can be characterized as:

$$ELBO_\beta = \mathrm{D}_{KL}[q(z|x)//p(z)] - \\ \int q(z|x)(-Nd_\beta(p(x|z)//\hat{p}(x))) \qquad (10)$$

and

$$d_\beta(p(x|z)//\hat{p}(x)) = -\frac{\beta+1}{\beta N}\sum_{i=1}^{N} p(x_i|z)^\beta + \int p(x|z)^{1+\beta}dx$$

For labeled data, the variational lower bound for the proposed URSVAE model can be represented as:

$$\log p(x,\tilde{y}) = \textstyle\int_a\int_z\int_T\int_\mu\int_\Lambda\int_\omega \\ \log(x,\tilde{y},a,z,T,\mu,\Lambda,\omega)dadzdTd\mu d\Lambda d\omega \geq \\ \mathrm{E}[\log(p_\theta(a,z,T,\mu,\Lambda,\omega,x,\tilde{y}))] - \\ \mathrm{E}[q_{\phi(a,z,T,\mu,\Lambda,\omega|x,\tilde{y})}] = \mathrm{E}[\log(p_\theta(a,z,T,\mu,\Lambda,\omega,x,\tilde{y}))] \\ -\mathrm{E}[q_{\phi(a|x)}] - \mathrm{E}[q_{\phi(z|a,\tilde{y},x)}] - \mathrm{E}[q_{\phi(T|\mu,\Lambda,x)}] \\ -\mathrm{E}[q_{\phi(\mu,\Lambda)}] - \mathrm{E}[q_{\phi(\omega|\pi)}]$$

The above inequality can be rewritten as

$$\log p(x,\tilde{y}) \geq \mathrm{E}_{q_{\phi(a,z,T,\mu,\Lambda|x,y)}}[\log \tfrac{p_\theta(a,z,T,\mu,\Lambda,x,\tilde{y})}{q_{\phi(a,z,T,\mu,\Lambda|x,\tilde{y})}}] = \\ \mathrm{E}_{q_\phi(a,z,T,\mu,\Lambda|x,\tilde{y})}[\log(p_\theta(x,\tilde{y}|a,z,T,\mu,\Lambda))] + \\ \mathrm{D}_{KL}[q(a,z,T,\mu,\Lambda|x,\tilde{y})//p(a,z,T,\mu,\Lambda)] \qquad (11)$$

In order to attenuate the influence of outliers, defining the set for all latent variables as $H = \{a, z, T, \omega, \mu, \Lambda\}$ and employing the technique from (Futami et al., 2018) by replacing KL-divergence with β-Divergence, the β-ELBO for labeled data L_β can be represented as:

$$L_\beta = \textstyle\int q(H|x,\tilde{y})(-\frac{\beta+1}{\beta}\sum_{i=1}^{N} p(\tilde{y}_i|H;x_i)^\beta \\ +N\int p(\tilde{y}|H;x)^{1+\beta}d\tilde{y}) + \mathrm{D}_{KL}[q(H|x,\tilde{y})//p(H)] \qquad (12)$$

The above equation computes the lower bound and learn the network parameters from noisy labels and outliers based on labeled data, where the first term enhances the the robustness to the sample outliers as the reconstruction error relying on β-divergence for labeled data and the second term regularizes $q(H|x,\tilde{y})$ to be close to the prior $p(H)$ as the prior regularization error. Denote L_β^{Dec} as the reconstruction error from the decoder based on log likelihood for labeled data and L^{prior} as the prior regularization error. Therefore,

$$L_\beta = L_\beta^{Dec} + L_{prior} \qquad (13)$$

In order to remedy the influence of noisy labels $\tilde{y}$, our goal is to construct the evidence lower bound based on the clean label $\hat{y}$ using the noise transition model. Particularly, we reformulate the equation as

$$q(H|x,\tilde{y}) = \frac{\sum_{\hat{y}^r} p(\tilde{y}=\tilde{y}^r|\hat{y}=\hat{y}^r)q(\hat{y}|H,x)q(H|x)}{\sum_{\hat{y}^r} p(\tilde{y}=\tilde{y}^r|\hat{y}=\hat{y}^r)q(\hat{y}|x)}, \\ q(\tilde{y}|H,x) = \frac{\sum_{\hat{y}^r} M_{\tilde{y}^r,\hat{y}^r}q(\hat{y}|H,x)q(H|x)}{\sum_{\hat{y}^r} M_{\tilde{y}^r,\hat{y}^r}q(\hat{y}|x)} \qquad (14)$$

For unlabeled data, by introducing the variational distribution for $\tilde{y}$ as $q_\phi(a, x|\tilde{y})$, the variational lower bound for the proposed URSVAE can be represented as

$$\log p(x) = \textstyle\int_a\int_z\int_T\int_\mu\int_\Lambda\int_{\tilde{y}} \log(x,\tilde{y},a,z,T,\mu,\Lambda)da$$

$$
\begin{aligned}
&dzdTd\mu\Lambda d\tilde{y} \geq \mathrm{E}_{q_\phi(a,\tilde{y},z,T,\mu,\Lambda|x)}[\log \tfrac{p_\theta(a,z,T,\mu,\Lambda,x,\tilde{y})}{q_\phi(a,z,T,\mu,\Lambda,\tilde{y}|x)}] \\
&= \mathrm{E}[\log(p_\theta(a,z,T,\mu,\Lambda,\omega,x,\tilde{y}))] - \mathrm{E}[q_{\phi(a,z,T,\mu,\Lambda,\omega,\tilde{y}|x)}] \\
&= \mathrm{E}[\log(p_\theta(a,z,T,\mu,\Lambda,\omega,x,\tilde{y}))] - \mathrm{E}[q_{\phi(a|x)}] - \mathrm{E}[q_{\phi(\tilde{y}|a,x)}] \\
&-\mathrm{E}[q_{\phi(z|a,\tilde{y},x)}] - \mathrm{E}[q_{\phi(T|\mu,\Lambda,x)}] - \mathrm{E}[q_{\phi(\mu,\Lambda)}] - \mathrm{E}[q_{\phi(\omega|\pi)}]
\end{aligned}
$$

Similarly, replacing the KL-divergence with β-Divergence and augmenting the latent variable set with the unknown labels y as $H_u = \{a, z, y, T, \omega, \mu, \Lambda\}$, the β-ELBO for unlabeled data in our model can be cast as

$$
\begin{aligned}
U_\beta = \textstyle\int q(H_u|x)(-\tfrac{\beta+1}{\beta}\sum_{i=1}^{N} p(x_i|H_u)^\beta + \\
\textstyle\int p(x|H_u)^{1+\beta}dx) + \mathrm{D}_{KL}[q(H_u|x)//p(H_u)]\ ,
\end{aligned}
$$

Similarly, we can further write the above equation as

$$
U_\beta = U_\beta^{Dec} + U_{prior}\ . \qquad (15)
$$

Practically, L_β^{Dec} and U_β^{Dec} are calculated via Monte Carlo sampling. The robustness of our proposed ELBO can be guaranteed leveraging the influence function (IF) (Huber & Ronchetti, 2011). As IF is widely used to analyze how much contamination affects estimated statistics, it is straightforward to show that given the perturbation on the empirical cumulative distribution caused by outliers, the IF of out posterior distribution is bounded, where the detailed theoretical analysis can be found at (Futami et al., 2018). The full objective for the proposed unified robust semi-supervised variational autoencoder (URSVAE) is therefore:

$$
L_{URSVAE} = L_\beta^{Dec} + \lambda_1 U_\beta^{Dec} + L_{prior} + \lambda_1 U_{prior} \quad (16)
$$

We would like to emphasize that our framework advances the model from (Maaløe et al., 2017) by introducing three novel components to effectively counter the outliers and noisy labels including:

- Uncertainty prior on the parameters of Gaussian mixture models to model the noisy input data.
- Noise transition model to correct the noisy labels so that the generative model and the inference model are conditioned on clean labels.
- Robust divergence to minimize the impact from the outliers in the optimization of ELBO.

2.5. Influence Function

Define $G(x)$ as a empirical cumulative distribution given by $\{x_i\}_{i=1}^n$ and denote the perturbed version of G at z as $G_{\varepsilon,z}(x) = (1-\varepsilon)G(x) + \Delta_z(x)$, where ε is the contamination portion and $\Delta_z(x)$ is the point mass at x. Given a statistic T. the influence function (IF) is defined as (Futami et al., 2018)

$$
IF(z, T, G) = \frac{\partial T(G_{z,\varepsilon}(x))}{\partial \varepsilon}\ |_{\varepsilon=0} \qquad (17)
$$

Table 1. Unsupervised test log-likelihood using different learning algorithms on the permutation invariant MNIST dataset (with 20% outliers) the normalizing flows VAE (VAE+NF), importance weighted auto-encoder (IWAE), variational Gaussian pro-cess VAE (VAE+VGP), Ladder VAE (LVAE) with FT denot-ing the finetuning procedure (Sønderby et al., 2016) and auxiliary deep generative models (Maaløe et al., 2017) and our method (β=0.2), where L represents the number of stochastic latent layers $z_1, \ldots, z_L$ and IW characterizes the importance weighted samples during training.

Method	$-\log p(x)$
VAE+NF(Miyato et al., 2015), L=1	-89.35
IWAE, L=1, IW=1 (Burda et al., 2015)	-90.26
IWAE, L=1, IW=50 (Burda et al., 2015)	-88.36
IWAE, L=2, IW=1 (Burda et al., 2015)	-89.71
IWAE, L=2, IW=50 (Burda et al., 2015)	-86.43
VAE+VGP, L=2 (Tran et al., 2015)	-85.79
LVAE, L=5, IW=1 (Sønderby et al., 2016)	-85.08
ADGM, L=1, IW=1 (Maaløe et al., 2017)	-84.67
ADGM, L=1, IW=2 (Maaløe et al., 2017)	-84.34
URSVAE, L=1, IW=1	-83.12
URSVAE, L=1, IW=2	-82.86

Thus, the IF of L_{URSVAE} is given by

$$
\begin{aligned}
&(\tfrac{\partial^2 L_{URSVAE}}{\partial H_u^2})^{-1}\tfrac{\partial}{\partial H_u}\mathrm{E}_{q(H_u)}[\mathrm{D}_{KL}[q(H_u|x)//p(H_u)] \\
&+N(\tfrac{\beta+1}{\beta}p(z|H_u)^\beta - \textstyle\int p(x|H_u)^{1+\beta}dx] \\
&+(\tfrac{\partial^2 L_{URSVAE}}{\partial H^2})^{-1}\tfrac{\partial}{\partial H}\mathrm{E}_{q(H)}[\mathrm{D}_{KL}[q(H|x)//p(H)] \\
&+N(\tfrac{\beta+1}{\beta}p(\tilde{y}|x,H)^\beta - \textstyle\int p(\tilde{y}|x,H)^{1+\beta}d\tilde{y}]\ , \qquad (18)
\end{aligned}
$$

It is straightforward to show that the above result is always bounded, namely our system is robust to the compound noise (the outliers on the data x and the labels y).

3. Experimental Results

Dataset: We evaluate the performance on five benchmark datasets including the MNIST, CIFAR-10, CIFAR-100 as benchmark datasets for image classification. The proposed algorithm is also evaluated with two real world large scale image datasets including Clothing1M and WebVision1.0. For Clothing1M dataset, it includes 1 million training images obtained from online shopping websites and labels are generated from surrounding texts. WebVision includes 2.4 million images collected from the internet using the 1,000 concepts in ImageNet ILSVRC12. Similar to the previous work (Chen et al., 2019), the baseline methods on the first 50 classes of the Google image subset using the inception-resnet v2 (Szegedy et al., 2017) are compared.

Competing approaches: We evaluate the robustness and accuracy performance of the proposed algorithm by comparing with multiple baseline methods and the state-of-the-art approaches. The baseline methods consist of ADGM(Maaløe

Table 2. Comparison of classification accuracy using different learning algorithms on the CIFAR-10(C10) and CIFAR-100(C100) datasets with varying levels of label noise and sample outliers. We re-implement all methods under the same setting based on public code.

Dataset	C10	C10	C10	C10	C100	C100	C100	C100
Label Noise, Outliers(%)	50, 10	80, 10	50, 20	80,20	50, 10	80,10	50, 20	80,20
VAT(Miyato et al., 2015)	68.2	62.3	78.6	73.1	48.7	28.3	47.2	25.3
LadderNet(Rasmus et al., 2015)	70.3	67.7	68.1	51.3	43.5	46.7	38.5	36.2
ADGM(Maaløe et al., 2017)	72.6	69.3	69.6	67.8	53.7	46.5	52.3	41.3
Coteaching(Yu et al., 2019)	85.1	65.8	81.3	61.5	53.1	21.9	51.6	20.7
M-correct(Arazo et al., 2019)	79.7	62.9	75.7	72.6	67.3	48.6	65.6	47.3
P-correct(Yi & Wu, 2019)	86.3	75.1	74.5	72.3	65.3	58.3	61.6	46.9
Meta-Learn(Li et al., 2019)	87.9	84.0	75.2	72.6	59.1	46.6	58.7	45.3
Dividemix (Li et al., 2020)	91.3	90.7	90.5	88.7	72.1	57.9	68.2	56.8
RGAN(Kaneko et al., 2019)	90.3	88.6	89.2	87.5	77.5	62.3	71.6	59.8
AmbientGAN (Bora et al., 2018)	87.5	83.6	84.9	81.6	67.2	64.5	63.8	60.3
URSVAE	**94.3**	**92.7**	**93.7**	**93.2**	**79.6**	**65.6**	**76.5**	**64.3**

et al., 2017) VAT (Miyato et al., 2015) and Ladder Network (Rasmus et al., 2015). The state-of-art approaches include Dividemix (Li et al., 2020), M-correction (Arazo et al., 2019), P-correction (Yi & Wu, 2019), Meta-Learning (Li et al., 2019), Coteaching (Yu et al., 2019), AmbientGAN (Bora et al., 2018) and rGAN (Kaneko et al., 2019). Among these approaches, Meta-Learning (Li et al., 2019) applies a noise-tolerant training algorithm relying on a meta-learning update. P-correction (Yi & Wu, 2019) tackles the noisy labels by training an end-to-end framework which can update network parameters and label estimations as label distributions. Iterative-CV (Chen et al., 2019) applies cross-validation to randomly split noisy datasets and adopts Coteaching(Yu et al., 2019) techniques to train DNNs robustly against noisy labels. Dividemix (Li et al., 2020) models the per-sample loss with a mixture model to dynamically divide the training data into a labeled set with clean samples and an unlabeled set with noisy samples and trains two diverged networks simultaneously. In the training of AmbientGAN (Bora et al., 2018), the output of the generator is passed through a simulated random measurement function to cope with lossy measurement. rGAN (Kaneko et al., 2019) incorporate a noise transitional model to learn a clean label generative distribution, where WGAN-GP (Gulrajani et al., 2017) is uitlized to ensure the training statability in the GAN configuration.

Table 3. Comparison of classification accuracy using different learning algorithms on the CIFAR-10 (with asymmetric 40% label noise and 10% outliers) and Clothing1M datasets (real-world noise and 10% outliers).

Dataset	C10	Clothing1M
VAT(Miyato et al., 2015)	73.6	59.42
M-correction(Arazo et al., 2019)	87.1	70.53
Joint-Optim(Tanaka et al., 2018)	87.6	71.35
P-correction(Yi & Wu, 2019)	86.3	72.81
Meta-Learning (Li et al., 2019)	87.9	73.01
Dividemix (Li et al., 2020)	91.3	73.16
RGAN(Kaneko et al., 2019)	86.9	71.97
AmbientGAN (Bora et al., 2018)	87.3	70.55
URSVAE	**94.7**	**79.65**

Implementation Details: For all the benchmark with variational autoencoders on MNIST, CIFAR-10 and CIFAR-100 datasets, we parameterize the deep neural network with three sets of 5 by 5 fully convolutional, ReLU and pooling layers followed by two fully connected hidden layers where each pair of layers contains the hidden units as $dim(h) = 500$ or $dim(h) = 1000$. Moreover, in order to have the fair comparison the dimensions of the auxiliary variables a, namely $dim(a, z) = 100$. and the latent variable z are set to be the same as ADGM(Maaløe et al., 2017). The network is trained with SGD using a batch size of 128. A momentum of 0.9 is set with a weight decay of 0.0005. The network is trained for 300 epochs. We set the initial learning rate as 0.02, and reduce it by a factor of 10 after 150 epochs. The warm up period is 10 epochs for CIFAR-10 and 30 epochs for CIFAR-100. For the Clothing1M and WebVision datasets, we utilize a similar architecture of Resnet-18 (He et al., 2015) but adding the uncertainty prior for the input, the noise transition model along with the auxilary variables as our encoder, by encoding a 256 by 256 RGB image into 512 feature maps of size 8 by 8. λ_1 is set to be the ratio of the number of unlabeled samples versus the number of labeled samples. All parameters are initialized using the same scheme described as (Glorot & Bengio, 2010). β varies from 0.1 to 0.4 where the best performance is reported.

Typically, Both CIFAR-10 and CIFAR-100 contain 50K training images and 10K test images of size 32 by 32. Each dataset is randomly sampled and divided into three disjointed subsets including the labeled set (5% samples), unlabeled set (75% samples) and test set (20% samples). We vary the outliers percentages from 10% to 20% and apply them to the datasets where the outliers are created by randomly removing 10% to 20% features in the data and replacing with zeros. Two types of label noises are studied in the experiments including symmetric and asymmetric label noise. In particular, the symmetric noise is ranging from 20% to 80% and generated by randomly replacing the labels for a percentage of training data with all possible labels. The asymmetric noise is created to simulate the real-world label noise where the corrupted label consists of the labels from the most similar class with respect to the ground truth (e.g. ”horse” to ”deer”, ”truck” to ”automobile”, ”bird” to ”airplane”).

Performance Evaluations on Classification: Fig.2 demonstrates the comparison of the t-SNE visualization with two dimensions for the auxiliary latent space for the CIFAR-10 dataset using ADGM (Maaløe et al., 2017) and our URSVAE with 10% sample outliers and 20% noisy labels, where the index number indicates classes 0 to 9. Each number locates on the median position of the corresponding vectors and the outliers are marked with squares. The embeddings from 10 distinct clusters using our method corresponds to true class labels instead of noisy labels. As illustrated by Fig.2, URSVAE finds a better separation among different classes in the presence of compound noise due to the utilization of uncertainty prior, noise transition model and robust divergence which justifies the robustness of our method to label noise and sample outliers. Fig.3 provides exemplary images with noisy labels detected by our URSVAE method from Clothing1M dataset, where the false labels are above the images in orange and the true labels are below the images in light blue. These examples demonstrates the efficacy of URSVAE in detection of images with noisy labels.

Figure 2. Comparison of the t-SNE visualization with two dimensions for the auxiliary latent space for the CIFAR-10 dataset using ADGM (Maaløe et al., 2017) and our URSVAE (β=0.15) with 10% sample outliers and 20% noisy labels, where the index number indicates classes 0 to 9. Each number locates on the median position of the corresponding vectors and the outliers are marked with squares. The embeddings from 10 distinct clusters using our method corresponds to true class labels, which justifies the robustness of our method to label noise and sample outliers.

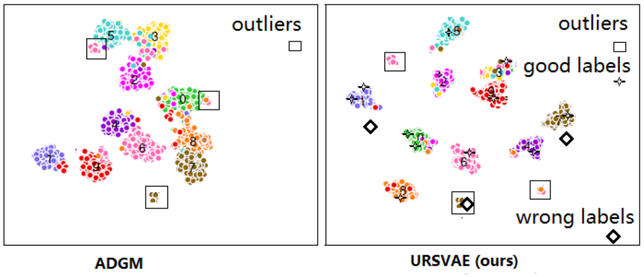
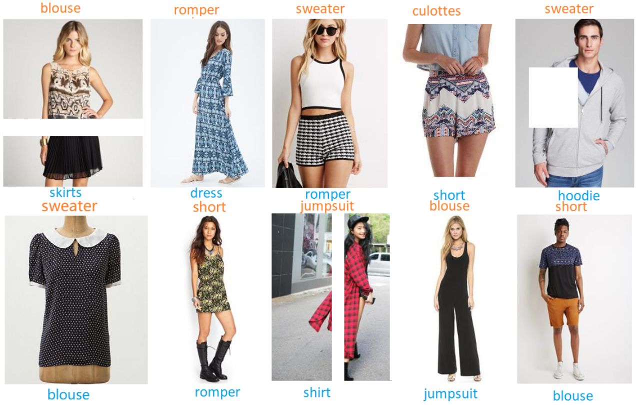


Figure 3. Exemplary images with compound noise detected by URSVAE from Clothing1M dataset, where the false labels are noted in orange and the true labels are noted in light blue.



We report the lower bound for the β-ELBO U_β on the unlabeled data with 5000 importance weighted samples where the similar setting as (Rasmus et al., 2015) with warm up, batch normalization and 1 Monte Carlo and IW sample for training. The percentage of outliers is set to be 20%. Table 1 demonstrates the log-likelihood scores for the permutation invariant MNIST dataset. The results shown in Table 1 indicates the the proposed method has strong expressive power by performing better than other methods in terms of log-likelihood due to the utilization of the uncertainty prior and the robust divergence in the inference. From Table 2, we see that our URSVAE is suffering significantly less from the compound noise than the other competing approaches due to our effective denoising schemes, which helps explain the state-of-the-art performance of our approach. The better performance with respect to Dividemix (Li et al., 2020) and RGAN(Kaneko et al., 2019) is mainly because our method efficiently rejects the outliers by placing a smaller or zero weight on them. While the performance gain compared to AmbientGAN(Bora et al., 2018) can be attributed to the fact that URSVAE successfully suppresses the samples with noisy labels by integrating the noise transition model in the optimization. Table 3 provides the comparison of classification accuracy using different learning algorithms on the CIFAR-10 (with asymmetric 40% label noise and 10% outliers) and Clothing1M datasets (real-world noise) along with standard deviation (in brackets). 40% asymmetric label noise is selected because certain classes become theoretically indistinguishable for asymmetric noise larger than 50%. Joint-Optim (Tanaka et al., 2018) jointly optimize the sample labels and the network parameters. As it can be seen from Table 3 that our method works nicely with asymmetric noise and real-world noise. With the optimization of the proposed β-ELBO, our method significantly outperforms

Table 4. Comparison of top-1 (top-5) accuracy with different state-of-the-art methods on the WebVision validation dataset and the ImageNet ILSVRC12 validation datasets training on (mini)WebVision dataset with 20% outliers.

Dataset	WebVision	WebVision	ILSVRC12	ILSVRC12
Metric	top1	top5	top1	top5
Coteaching(Yu et al., 2019)	62.75	83.61	60.73	83.56
F-correction(Patrini et al., 2017a)	60.73	81.64	56.81	81.72
Decoupling(Malach & Shwartz, 2017)	61.37	82.95	57.93	81.38
MentorNet(Jiang et al., 2018)	62.78	80.92	57.52	79.51
Iterative-CV (Chen et al., 2019)	64.87	84.03	61.31	83.79
Dividemix (Li et al., 2020)	75.68	87.73	72.87	85.61
RGAN(Kaneko et al., 2019)	74.39	85.68	71.35	84.78
URSVAE	**77.35**	**90.68**	**75.41**	**90.68**

Table 5. Ablation study results in terms of testing accuracy(%) on CIFAR-10(C10) and CIFAR-100(C100) for our method.

Dataset	C10	C10	C10	C10	C10	C100	C100	C100	C100
Noise type	Sym.	Sym.	Sym.	Sym.	Asym.	Sym.	Sym.	Sym.	Sym.
Label Noise, Outliers(%)	50, 10	80, 10	50, 20	80,20	40, 10	50, 10	80,10	50, 20	80,20
Ours w/o uncertainty prior	91.7	90.7	91.5	88.7	92.3	77.3	72.0	63.8	60.9
Ours w/o robust divergence	89.6	92.6	93.4	92.1	91.6	77.6	74.9	61.2	61.7
Ours w/o noise transition	86.5	90.7	91.9	88.7	91.3	78.5	75.1	63.2	62.8
Ours	**95.7**	**93.3**	**93.7**	**92.9**	**95.1**	**81.3**	**67.1**	**77.8**	**67.9**

the best competitor Dividemix by 3.8% and 5.2% respectively on CIFAR-10 and Clothing1M datasets. In contrast, most approaches from the competitors cannot address the issues from outliers and label noise simultaneously.

Table 4 illustrates the comparison of top-1 (top-5) accuracy with different state-of-the-art methods on the WebVision validation dataset and the ImageNet ILSVRC12 validation datasets training on (mini)WebVision dataset with 20% outliers. Here top-5 accuracy is an extension to top-1 accuracy where instead of computing the probability that the most probable class label equals to the ground truth label, the probability that the group truth label is in the top 5 most probable labels is calculated. Specifically, in MentorNet(Jiang et al., 2018), an auxiliary teacher network is pre-trained and used to drop samples with noisy labels for its student network which is used for image recognition. Our method again achieves the best performance with respect to other competing methods due to its capability of mitigating the compound noise simultaneously in one shot. **Ablation Study:** Here we provide some details on ablation study in Table 5. Our method w/o uncertainty prior excludes the uncertainty prior from the model. Hence the performance degradation (especially with 20% outliers) suggests the importance of the proposed hierarchical structure for variational inference by placing the uncertainty prior. Secondly, our method w/o robust divergence replaces β-divergence with the regular KL-divergence for ELBO in the optimization, which validates the contribution of using robust divergence for countering the problem of sample outliers because more outliers would be mistakenly classified without robust divergence. Moreover, our method w/o noise transition is utilizing the same network architecture as URSVAE except omitting the noise transition model, which indicates the effectiveness of adapting the noise transition model in the classification to alleviate the detrimental effect of noisy labels by conditioning on clean labels. Finally, by tuning β in the robust divergence, the classification performance of URSVAE is improved by around 2.2% on average. The training time of URSVAE on CIFAR-10 evaluated on a single Nvidia V100 GPU is 4.9 hours.

4. Conclusion

A novel robust semi-supervised variational autoencoder under noisy labels and outliers is proposed. With the aid of uncertainty priors and noise transition models, URSVAE has demonstrated the robust performance in the presence of noisy labels and outliers. Moreover, the proposed robust divergence in variational inference further enhanced the robustness by minimizing the β-ELBO. Evaluations on multiple benchmark and real-world datasets demonstrate the efficiency and robustness of URSVAE compared to the state-of-the-art approaches.

References

Arazo, E., Ortego, D., Albert, P., and Connor, N. Unsupervised Label Noise Modeling and Loss Correction. *International Conference on Machine Learning*, 2019.

Basu, A., Harris, I., Hjort, N., and Jones, M. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 1998.

Bishop, C. Pattern Recognition and Machine Learning. *Springer*, 2006.

Bora, A., Price, E., and Dimakis, A. AmbientGAN: Generative models from lossy measurements. *Interational Conference of Learning Representations*, 2018.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance Weighted Autoencoders. *arXiv preprint-arXiv:1509.00519*, 2015.

Chen, P., Liao, B., Chen, G., and Zhang, S. Understanding and utilizing deep neural networks trained with noisy labels. *International Conference on Machine Learning*, 2019.

Diederik, P. and Welling, M. AutoEncoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Futami, F., Sato, I., and Sugiyama, M. Variational Inference based on Robust Divergences. *Twenty-First International Conference on Artificial Intelligence and Statistics*, 2018.

Glorot, X. and Bengio, Y. Understanding the dif-ficulty of training deep feedforward neural networks. *AISTATS*, 2010.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of Wasserstein GANs. *Neural Information Processing Systems*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *Computer Vision and Pattern Recognition*, 2015.

Hou, S. and Galata, A. Robust Estimation of Gaussian Mixtures from Noisy Input Data. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

Huang, H., Li, Z., He, R., Sun, Z., and Tan, T. IntroVAE: Introspective variational autoencoders for photographic image synthesis. *Neural Information Processing Systems*, 2018.

Huber, P. and Ronchetti, E. Robust Statistics. *Wiley Series in Probability and Statistics*, 2011.

Jiang, L., Zhou, Z., Leung, T., Li, L., and Fei-Fei, L. Mentornet: Learning datadriven curriculum for very deep neural networks on corrupted labels. *International Conference on Machine Learning*, 2018.

Kaneko, T., Ushiku, Y., and Harada, T. Label-Noise Robust Generative Adversarial Network. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. Semi-supervised Learning withDeep Generative Models. *Neural Information Processing Systems*, 2014.

Li, J., Wong, Y., Zhao, Q., and Kankanhalli, M. Learning to Learn from Noisy Labeled Data. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Li, J., Socher, R., and Hoi, S. DIVIDEMIX: Learning with noisy labels as semi-supervised learning. *International Confernce on Learning Respresentation*, 2020.

Li, Y. and Gal, Y. Dropout inference in Bayesian neural networks with alpha-divergences. *International Conference on Machine Learning*, 2017.

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. Auxiliary Deep Generative Models. *International Conference on Machine Learning*, 2017.

Malach, E. and Shwartz, S. Decoupling "when to update" from " how to update". *Neural Information Processing Systems*, 2017.

Miyato, T., Maeda, S.-i. M. K., Nakae, K., and Ishii, S. Distributional Smoothing with Virtual Adversarial Training. *arXiv preprint arXiv:1507.00677*, 2015.

Murphy, K. Conjugate Bayesian analysis of the Gaussian distribution. *Lecture notes on UC Berkeley*, 2007.

Patrini, G., Rozza, A., Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017a.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017b.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. Semi-supervised learning with ladder networks. *Neural Information Processing Systems*, 2015.

Sønderby, C., Raiko, T., Maaløe, L., Sønderby, S., and Winther, O. Ladder variational autoen-coders. *arXiv preprint arXiv:1602.02282.*, 2016.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. *ICLR workshop*, 2015.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. *AAAI*, 2017.

Tanaka, D., Ikami, D., Yamasaki, T., and Aizawa, K. Joint optimization framework for learning with noisy labels. *International Conference on Computer Vision and Patterson Recognition*, 2018.

Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C., and Silberman, N. Learning from noisy labels by regularized estimation of annotator confusion. *Computer Vision and Pattern Recognition*, 2019.

Tran, D., Ranganath, R., and Blei, D. Variational Gaussian process. *arXiv preprintarXiv:1511.06499*, 2015.

Yi, K. and Wu, J. Probabilistic end-to-end noise correction for learning with noisy labels. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., and M.Sugiyama. How does disagreement help generalization against label corruption? *International Conference on Machine Learning*, 2019.