Discriminative Sample Generation for Deep Imbalanced Learning

Ting Guo[1], **Xingquan Zhu**[2*], **Yang Wang**[1] and **Fang Chen**[1]
[1]Faculty of Eng. & Info. Tech., University of Technology Sydney, Australia
[2]Dept. of Computer & Electrical Eng. and Computer Science, Florida Atlantic University, USA
Ting.Guo@uts.edu.au, xzhu3@fau.edu, {Yang.Wang, Fang.Chen}@uts.edu.au

Abstract

In this paper, we propose a discriminative variational autoencoder (DVAE) to assist deep learning from data with imbalanced class distributions. DVAE is designed to alleviate the class imbalance by explicitly learning class boundaries between training samples, and uses learned class boundaries to guide the feature learning and sample generation. To learn class boundaries, DVAE learns a latent two-component mixture distributor, conditioned by the class labels, so the latent features can help differentiate minority class *vs.* majority class samples. In order to balance the training data for deep learning to emphasize on the minority class, we combine DVAE and generative adversarial networks (GAN) to form a unified model, DVAAN, which generates synthetic instances close to the class boundaries as training data to learn latent features and update the model. Experiments and comparisons confirm that DVAAN significantly alleviates the class imbalance and delivers accurate models for deep learning from imbalanced data.

1 Introduction

Learning and classification of data with imbalanced class distributions is a significant challenge for the machine learning community. On one hand, minority class instances often represent the objects of interests in applications [He and Garcia, 2008; Sun *et al.*, 2007; Chen and Shyu, 2011], and require significant attention by the underlying learning algorithms. On the other hand, class distributions have a significant impact to most machine learning methods, and with a disadvantage in populations, minority class instances are essentially more vulnerable being incorrectly classified by learning algorithms [Krawczyk, 2016]. Accordingly, many methods exist to (1) modify sample distributions using random under- or over-sampling [Wang *et al.*, 2015; Chawla *et al.*, 2002; Barua *et al.*, 2012], or (2) modify costs associated to different classes [Yan *et al.*, 2017], such that the learning and classification can emphasize on minority class instances. For most researches in the field, they are carried out in the conventional machine learning context and work on the original feature space for learning and classification.

Recently, deep learning has emerged with superb performance gain, thanks to the powerful feature learning. However, most existing methods are mainly designed under assumptions that training data are relatively balanced and do not take the class distributions into consideration for feature learning. A couple of works have recently addressed deep learning for imbalanced data [Wu *et al.*, 2017; Yan *et al.*, 2015], whereas these methods are all based on sampling approaches working on the original instance space to learn new features. These methods do not have a solution to differentiate class boundaries, so the feature learning cannot be aligned to the class distributions. This essentially deteriorates deep learning performance for imbalanced data.

While learning from imbalanced data has been severely challenged by the imbalance and scarcity of minority class samples, latent space generative models in deep learning, for example, variational autoencoders (VAEs) and generative adversarial networks (GANs) have already provided effective ways to learn mapping from a latent encoding space to a data space [Kingma and Welling, 2014; Goodfellow *et al.*, 2014]. In other words, VAEs or GANs can generate new instances resemble to the original data, and this raises an open question on *whether synthetic minority class samples can be generated to alleviate imbalanced learning challenges*.

Indeed, VAEs and GANs have shown to generate samples close to original data, but such samples are normally not directly usable to help re-train the model. On one hand, VAEs suffer from the effect of $\mathfrak{L}_2$ loss and thus generate blurry images, because $\mathfrak{L}_2$ loss is based on the assumption that the data follow a single Gaussian distribution. When samples have multi-modal distributions, VAE cannot generate samples with sharp edges and fine details. Besides, the squared error is an element-wise measure which lacks higher-level and sufficiently invariant discriminative information to guide the training process. GANs, on the other hand, learn a loss that tries to classify if the output sample is real or fake, while simultaneously training a generative model to minimize this loss [Goodfellow *et al.*, 2014]. GANs suffer from training instability and the quality of generated samples are largely not as good as original data and may cause over-fitting problem. Accordingly, some researches suggested hybrid VAE-

[*] Contacting Author

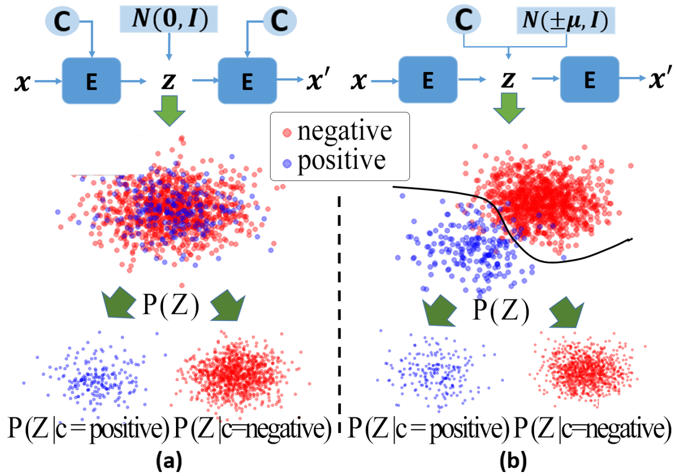

Figure 1: The difference between conditional VAEs (CVAE) and the proposed discriminative VAEs (DVAE) on latent representations z (t-SNE visualization). (a) is CVAE and (b) is DVAE. z of CVAE follows a simple Gaussian distribution while z of DVAE follows a two-component mixture distribution. Both CVAE and DVAE can distinguish positive and negative samples by providing conditional variable c, but DVAE can provide additional capability of margin adjustment and identification because two-component mixture distribution allows DVAE to model the class boundary. As a result, DAVE can generate synthetic minority class samples close to the class boundary to help alleviate class imbalance for deep imbalanced learning.

GAN model by combining a variational autoencoder with a generative adversarial network in which learned feature representations in the GAN discriminator can be used as the basis for the VAE reconstruction objective [Larsen *et al.*, 2016; Dosovitskiy and Brox, 2016; Yu *et al.*, 2017]. In addition, to allow users to have control on the sample distributions, conditional VAE/GAN (CVAE/CGAN) was developed [Mirza and Osindero, 2014; Isola *et al.*, 2017; Sohn *et al.*, 2015; van den Oord *et al.*, 2016; Bao *et al.*, 2017], so users can generate samples close to a user-specified category.

Although CVAE/CGAN can be conditioned to one additional variable c expressing the labels. They are not suitable for imbalanced learning as they are unable to differentiate class boundary in latent space. Take Fig.1 (a) as an example, Z is the latent representation and $P(Z)$ is the latent distribution where positive and negative samples are mixed together and positive samples can only be distinguished from negative ones when specific $c = positive$ is given and thus $P(Z|c = positive) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The above observations motivate us to model latent representation as being drawn from two simple distributions with opposite means, which can be adjusted through the training of adversarial networks. As a result, we can explicitly depict the class boundary in latent space to help improve the discriminator's performance. By clearly modeling class boundary to generate new minority class samples, we can directly learn a classifier by learning discriminative latent features and generating high-quality synthetic minority class samples to alleviate class imbalance for deep imbalanced learning.

The main contribution of the paper is threefold:

Discriminative generative model. We impose a prior over the latent two-component mixture distribution to explicitly differentiate different classes in the latent space.

Adversarial networks for class boundary alignment. We unify jointly learn and align class boundary in the latent space to generate high-quality synthetic samples to help improve the discriminator.

Data augmentation for deep imbalanced learning. Different from the existing sampling-based approaches for imbalanced learning, we propose to generate and include synthetic samples close to the class boundary (support samples) to alleviate the class imbalance. This allows discriminative and generative models to be seamlessly integrated for effective deep imbalanced learning.

2 Problem Definition and Preliminary

2.1 Problem Definition

For deep imbalanced learning, we are given a training set with imbalanced binary class labels. The class presented with very few samples but associated with higher identification importance is referred to as the minority (or positive) class, while the other one is taken as the majority (or negative) class. Mathematically, given a training dataset $D = \{< x_i, y_i > | x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}\}$, $X_{pos} = \{x_i | y_i = 1\}$ and $X_{neg} = \{x_i | y_i = -1\}$, we aim to learn a deep discriminative model (*i.e.* a classifier) from D with the best performance in classifying both minority and majority class test samples.

2.2 Variational Autoencoder

A traditional VAE consists of two networks that encode a data sample x to a latent representation z and decode the latent representation back to data space, respectively:

$$z \sim Enc(x) = Q(z|x), \quad \widetilde{x} \sim Dec(z) = P(x|z) \qquad (1)$$

The VAE regularizes the encoder by imposing a prior over the latent distribution $P(z)$. Typically $z \sim \mathcal{N}(0, \mathbf{I})$ is chosen, where $\mathbf{I}$ is the identity matrix. The VAE loss is negative of the sum of the expected log likelihood (the reconstruction error) and a prior regularization term:

$$\begin{aligned}\mathfrak{L}_{VAE} &= -\mathbb{E}_{Q(z|x)}[\log \frac{P(x|z)P(z)}{Q(z|x)}] \\ &= -\mathbb{E}_{Q(z|x)}[\log P(x|z)] + D_{KL}(Q(z|x)||P(z))\end{aligned} \qquad (2)$$

where D_{KL} is the Kullback-Leibler divergence.

2.3 Generative Adversarial Network

A GAN consists of two networks: a generator $Gen(z)$ maps latent z to data space while a discriminator network assigns probability $y = Dis(x) \in [0, 1]$ that x is an actual training sample and probability $1 - y$ that x is generated by proposed model through $x = Gen(z)$ with $z \sim P(z)$. The GAN objective is to find the binary classifier that gives the best possible discrimination between true and generated data and simultaneously encouraging Gen to fit the true data distribution. We thus aim to maximize the binary cross entropy:

$$\mathfrak{L}_{GAN} = log(Dis(x)) + log(1 - Dis(Gen(z))) \qquad (3)$$

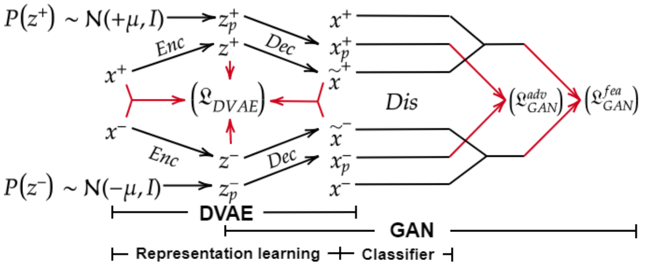

Figure 2: Flow through the DVAAN model during training (Red lines represent terms in the training objectives).

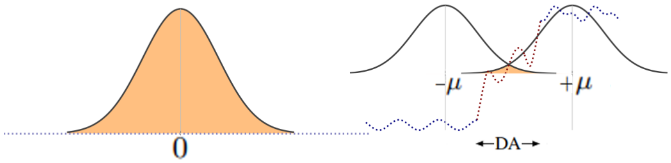

Figure 3: Simple *vs.* two-component distributions. VAE/CVAE uses one simple normal distribution (left panel), so their latent representations from different classes are from the same distribution (orange area). As a result, there is no boundary between different classes, so one cannot distinguish minority samples from majority samples in the latent space (as shown in Fig. 1(a)). In comparison, DVAE (right panel) uses two normal distributions and GAN's discriminator can help learn decision boundary/curve to distinguish samples from different classes (dotted curve in the right panel). This decision boundary can be considered as the class boundary (the black curve in Fig. 1(b)). Our Data Augmentation (DA) strategy will generate samples from the class boundary regions (orange area) to alleviate data imbalance for deep imbalanced learning).

with respect to Dis/Gen with x being a training sample and $z \sim P(z)$.

3 DVAAN for Deep Imbalanced Learning

Fig. 2 shows the proposed Discriminative Variational Autoencoding Adversarial Network (DVAAN), which includes a discriminative variational autoencoding (DVAE) module and a generative adversarial network (GAN) module.

3.1 Discriminative Variational Autoencoding

For existing VAE-based methods, they assume that the latent representation follows a simple normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ without considering the class variety. This is, unfortunately, not the case for real-world data. For example, statistics show that men have an average taller figure than women. If we take height as a feature, men and women will each fall into different intervals and the overall distribution of human is known not following any unimodal distribution [Schilling *et al.*, 2002]. Besides, as the imbalance degree may cause the unreliability of uncovering regularities inherent in a small class, the assumption of one simple distribution over all classes may be biased to the majority.

Alternatively, we regularize the encoder by imposing a prior over the latent two-component mixture distribution $P(z|y)$. It is a conditional probability distribution (CPD). It means for each possible value of y, we would have a $P(z)$. Specifically, $P(z|y = 1) = P(z_{pos})$ and $P(z|y = -1) = P(z_{neg})$, where we force,

$$z_{pos} \sim \mathcal{N}(\mu, \mathbf{I}), \quad z_{neg} \sim \mathcal{N}(-\mu, \mathbf{I}) \tag{4}$$

μ is a chosen parameter. In DVAE, we assume that the latent representations follow a two-component mixture distribution. By forcing discriminative distributions to different classes, the potential discriminative regularities inherent in the minority class can be retained to a greater extent without being lost in the learning preference on the majority class. Besides, it is obvious that for different classes, the distribution of the data should be different. Therefore it makes more sense to choose different distributions for different classes.

As the proposed Discriminative VAE is a supervised generative model, the class information (label) is known for training process. Therefore, we have,

$$\begin{aligned} z \sim Enc(x) = Enc(x|y) &= \begin{cases} Q(z_{pos}|x_{pos}), y = 1, \\ Q(z_{neg}|x_{neg}), y = -1. \end{cases} \\ \widetilde{x} \sim Dec(z) = Dec(z|y) &= \begin{cases} P(x_{pos}|z_{pos}), y = 1, \\ P(x_{neg}|z_{neg}), y = -1. \end{cases} \end{aligned} \tag{5}$$

The loss function of the Discriminative VAE is the expected log likelihood with a regularizer:

$$\mathfrak{L}_{DVAE} = -\mathbb{E}_{Q(z|x,y)}[\log \frac{P(x|z,y)P(z)}{Q(z|x,y)}] = \mathfrak{L}_{ele} + \mathfrak{L}_{KL} \tag{6}$$

with

$$\begin{aligned} \mathfrak{L}_{ele} &= -\mathbb{E}_{Q(z|x,y)}[\log P(x|z,y)] \\ \mathfrak{L}_{KL} &= D_{KL}(Q(z|x,y)||P(z)) \end{aligned} \tag{7}$$

The loss function is similar to Conditional VAE (CVAE) [Sohn *et al.*, 2015], which uses class information y as part of the inputs for both encoder and decoder to model latent variables and data conditioned to some random variables. However, rather than considering y as extra input, DVAE uses y to decide $z_{pos} \sim \mathcal{N}(\mu, \mathbf{I})$ or $z_{neg} \sim \mathcal{N}(-\mu, \mathbf{I})$, which are used as the prior to impose for the latent distribution $P(z)$. It means that z_{pos} and z_{neg} are in the same latent space while the latent representations of CVAE for different classes are in different spaces. By adjusting μ, the boundary of positive and negative classes are changing accordingly.

3.2 DVAAN Framework

The appealing properties of GAN are that its discriminator network implicitly learns a rich top-down similarity metric for classification tasks and this discriminator can be considered as a classifier. We thus propose to transfer the properties of data learned by the discriminator into (1) a more abstract reconstruction error for the DVAE, and (2) a classifier for distinguishing positive samples from negative ones. The end result will be a method that combines the advantage of GAN as a high quality generative & classification model and VAE as a method that produces an encoder of data into the latent space z with minimal information loss.

Specifically, we use the decoder network of the DVAE as the generator network of the GAN, but instead of training a separate discriminator network, we combine the decoder and discriminator into a single network. So our inputs for GAN

Algorithm 1 Training the DVAAN model

```
input: Imbalanced training dataset $D$
output: $\theta_{Dis}$
$\mu$, $\theta_{Enc}$, $\theta_{Dec}$, $\theta_{Dis} \leftarrow$ initialize network parameters;
$\boldsymbol{D}^{+} \leftarrow \{\}$
repeat
    $(\boldsymbol{X}, \boldsymbol{Y}) \leftarrow$ random mini-batch from $\{\boldsymbol{D}, \boldsymbol{D}^{+}\}$
    $\boldsymbol{Z} \leftarrow Enc(\boldsymbol{X})$
    $\tilde{\boldsymbol{X}} \leftarrow Dec(\boldsymbol{Z})$
    $\boldsymbol{Z}_p \leftarrow$ samples from $\mathcal{N}(\pm\mu, \mathbf{I})$
    $\boldsymbol{X}_p \leftarrow Dec(\boldsymbol{Z}_p)$
    $\mathfrak{L}_{ele} \leftarrow -\mathbb{E}_{Q(z|x,y)}[\log P(x|z,y)]$
    $\mathfrak{L}_{KL} \leftarrow D_{KL}(Q(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{Y})||P(\boldsymbol{Z}))$
    $\mathfrak{L}_{GAN}^{adv} \leftarrow \log(1 - \boldsymbol{Y}Dis(\tilde{\boldsymbol{X}})) + \log(1 - \boldsymbol{Y}Dis(\boldsymbol{X}_p))$
    $\mathfrak{L}_{GAN}^{fea} \leftarrow -\mathbb{E}_{Q(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{Y})}[\log P(Dis(\boldsymbol{X})|\boldsymbol{Z},\boldsymbol{Y})]$

    // Update parameters according to gradients
    $\theta_{Enc} \xleftarrow{+} -\bigtriangledown_{\theta_{Enc}}(\mathfrak{L}_{ele} + \mathfrak{L}_{KL} + \lambda_1\mathfrak{L}_{GAN}^{fea})$
    $\theta_{Dec} \xleftarrow{+} -\bigtriangledown_{\theta_{Dec}}(\mathfrak{L}_{ele} + \lambda_1\mathfrak{L}_{GAN}^{fea} + \lambda_2\mathfrak{L}_{GAN}^{adv})$
    $\theta_{Dis} \xleftarrow{+} -\bigtriangledown_{\theta_{Dis}}(\mathfrak{L}_{GAN}^{adv})$
    $\mu \xleftarrow{+} -\Delta\mu$
    $\boldsymbol{D}^{+} \leftarrow Data\ Augmentation(\theta_{Dec}, \theta_{Dis}, \mu)$
until Convergence
```

is $\tilde{x} = Gen(z) = Dec(z)$ where $z \in \{z_{pos}, z_{neg}\}$. The task for the GAN discriminator is twofold: (1) to learn a top-down similarity metric to discriminate positive from negative generated/real samples. We thus aim to maximize/minimize the binary cross entropy: $\mathfrak{L}_{GAN}^{adv}$, which is the adversarial loss that forces the discriminator to label a sample as positive or negative; (2) to minimize the feature-wise reconstruction loss, $\mathfrak{L}_{GAN}^{fea}$, by evaluating the difference between original and reconstruction in the space of the hidden layers of the discriminator.

$$\begin{aligned}\mathfrak{L}_{GAN}^{adv} &= \log\left(1 - Dis(Dec(z_{pos}))\right) + \log\left(1 + Dis(Dec(z_{neg}))\right)\\ \mathfrak{L}_{GAN}^{fea} &= -\mathbb{E}_{Q(z|x,y)}[\log P(Dis(x)|z,y)]\end{aligned} \tag{8}$$

We train our combined model with the triple criterion:

$$\mathfrak{L} = \mathfrak{L}_{DVAE} + \mathfrak{L}_{GAN}^{adv} + \mathfrak{L}_{GAN}^{fea} \tag{9}$$

Class Boundary Alignment

In DVAE, we introduce a latent two-component mixture distribution $P(z|y)$ to depict the latent representation in hidden space. As μ is a predefined parameter, if μ is too large, the generator network is hard to learn identical latent representations z_{pos} and z_{neg} because of inseparability of original data representation. While if μ is too small, it is hard for discriminator network to identify the positive samples from negative ones as the inseparability of latent representation z_{pos} and z_{neg}. In this section, we will propose an update strategy for μ based on the adversarial loss. It is a modification of the adversarial loss that forces the class boundary of latent representation to better minimize both top-down reconstruction loss ($\mathfrak{L}_{GAN}^{fea}$) and final classification loss ($\mathfrak{L}_{GAN}^{adv}$).

$$\Delta\mu = tanh(\mathfrak{L}_{GAN}^{fea} - \gamma\mathfrak{L}_{GAN}^{adv}) \tag{10}$$

$tanh()$ is the hyperbolic tangent function and we use a parameter γ to weight the ability to update μ. So based on Eq. (10), we can update μ by using $\mu_{k+1} = \mu_k - \Delta\mu$, where k is the iterations. It is obvious that if the discriminator network is hard to identify positive reconstructed samples from negative ones ($\mathfrak{L}_{GAN}^{adv}$ accounts for greater), we increase μ to let latent distributions be far away from each other ($\Delta\mu > 0$) to make latent representations more discriminative; while if the decoder/generator is hard to reconstruct original samples with small reconstruction loss ($\mathfrak{L}_{GAN}^{fea}$ accounts for greater), we decrease μ to make latent representations from different classes close to each other and therefore ensuring high-quality reconstruction ability.

Data Augmentation

Benefit from the strong generative ability of GAN and the control ability of the forced two-component mixture distribution of DVAE, we can generate synthetic samples close to the class boundary, and these samples can be considered similar to support vectors defined in SVM [Cortes and Vapnik, 1995].

Specifically, as we force our latent representation to follow a two-component mixture distribution by given μ, the samples generated from $\boldsymbol{Z}^0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are highly likely to be the support samples. Accordingly, we first generate a latent representation z from $\boldsymbol{Z}^0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and calculate the distance between z to μ and $-\mu$ respectively. If z is close to μ we set a pseudo label to z as $y = 1$ and $y = -1$ otherwise. Then we generate the support samples $\tilde{x} = Dec(z)$, and we will get another pseudo label $y' = Dis(\tilde{x})$. If $y' = y$, we have the confidence to consider $\tilde{x}$ as an additional training samples.

Algorithm 1 lists the training procedure of DVAAN.

4 Experiments

4.1 Benchmark Data

MNIST. is a handwritten digit database commonly used for deep learning. It has a training set of 60,000 examples, and a test set of 10,000 samples[1].

Fashion-MNIST. is a dataset of Zalando's article images—consisting of a training set of 60,000 examples and a test set of 10,000 samples. Each example is a 28x28 grayscale image, associated with a label from 10 classes[2].

CIFAR-10. is one of the most widely used datasets for deep learning. It contains 60,000 32x32 color images in 10 different classes, representing airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks[3].

TIS. is used to predict the Translation Initiation Sites (TIS) at which the translation from a messenger RNA to a protein sequence was initiated. The dataset has two classes and 927 features[4]. The detailed process, including a biological explanation, is detailed in [Liu and Wong, 2003].

Because MNIST, Fashin-MNIST and CIFAR-10 are multi-class datasets, we choose two similar classes from each

[1] http://yann.lecun.com/exdb/mnist/

[2] https://github.com/zalandoresearch/fashion-mnist

[3] https://www.cs.toronto.edu/~kriz/cifar.html

[4] https://www.comp.nus.edu.sg/~wongls/courses/cs2220/2008/TIS-data.html

dataset. Digit 5 (positive) and Digit 6 (negative) are chosen from MNIST, Sandal and Sneaker are chosen from Fashion-MNIST, and cat (positive) and dog (negative) are chosen from CIFAR-10. For all four datasets, we randomly sample a percentage of positive training samples to create different imbalance levels. For example, an imbalance ratio 1:10 is created by selection 10% of positive samples and keep 100% negative samples (except for Table 2 which we use one-against-all meaning we keep all positive samples and all samples in other classes are used as negative).

4.2 Baseline Methods & Settings

For all baselines, the encoder network consists of four convolution layers followed by two fully-connected layers. The convolution layers have 64, 92, 128 and 256 kernels with the filter size of 5×5, 5×5, 3×3 and 3×3. The decoder network (Generator network) consists of 2 fully-connected layers, followed by 4 deconv layers with 2-by-2 unsampling and corresponding architectures to the encoder network (both kernels and filter size). For the Discriminator network, we use the same discriminator network as the VAE+GAN [Larsen *et al.*, 2016]. The batch normalization layer [Ioffe and Szegedy, 2015] is also applied after each convolution layer. The model is implemented using Tensorflow toolbox.

VAE-GAN. is an updated version of VAE-GAN [Larsen *et al.*, 2016] for binary classification task. The difference is the input of GAN's discriminator is the reconstructed positive and negative samples and the aim of GAN's discriminator is to distinguish positive samples from negative ones.

VAE-GAN$_b$. uses over-sampling techniques for mini-batch selection to balance the class distribution. This serves as a baseline of traditional sampling methods for deep learning from imbalanced data.

AC-GAN. adds a classifier layer to the discriminator and a conditioning vector to the generator [Odena *et al.*, 2016]. The discriminator is then trained to minimize classification error in addition to the traditional GAN objective of real *vs.* fake. This allows labels to be leveraged and provides additional information to the GAN. In addition, it allows the generator to be conditioned to produce certain classes of samples.

CVAE-GAN. combines a variational auto-encoder with a generative adversarial network for synthesizing images in fine-grained categories [Bao *et al.*, 2017]. This approach models a sample as a composition of the label and latent attributes in a probabilistic model.

DVAAN, DVAAN$_\mu$ and DVAAN$_a$. DVAAN is the full version of our proposed algorithm. To validate the effectiveness of the class differentiation in the latent space, we fix $\mu = 3$ and denote this method by DVAAN$_\mu$. Meanwhile, to validate the data augmentation, we drop the data augmentation component from DVAAN (so no synthetic data are added to the training set), and denote this variant by DVAAN$_a$.

4.3 Classification Performance Comparison

Table 1 reports the AUC scores of different methods. The results show that VAE-GAN has the worst performances because no treatment was used to tackle imbalanced data in learning. Meanwhile, the results from VAE-GAN$_b$ show that over-sampling is effective to help improve the learning performance on imbalanced data, mainly because the balanced sample distributions of mini-batch selection in VAE-GAN$_b$ helps avoid the preference of optimizing reconstruction and adversarial loss of the samples from the majority class.

Methods	*MNIST*	*Fashion-MNIST*	*CIFAR-10*	*TIS*
VAE-GAN	91.52%	87.12%	89.87%	71.38%
VAE-GAN$_b$	96.17%	88.35%	92.19%	74.14%
AC-GAN	98.49%	92.07%	94.26%	80.62%
CVAE-GAN	96.83%	91.44%	95.05%	78.45%
DVAAN$_\mu$	96.28%	90.34%	93.37%	83.56%
DVAAN$_a$	97.76%	91.75%	94.83%	84.69%
DVAAN	**98.71%**	**92.33%**	**96.17%**	**85.51%**

Table 1: AUC scores of different methods on four imbalanced datasets (Imbalance Ratio = 1:5)

Methods	*MNIST*	*Fashion-MNIST*	*CIFAR-10*
VAE-GAN	93.94%	88.14%	91.08%
VAE-GAN$_b$	97.03%	88.75%	94.14%
AC-GAN	98.61%	89.30%	**98.07%**
CVAE-GAN	98.76%	90.26%	96.96%
DVAAN$_\mu$	96.90%	89.71%	95.77%
DVAAN$_a$	98.33%	91.81%	96.28%
DVAAN	**99.13%**	**92.59%**	97.91%

Table 2: AUC scores of different methods on three imbalanced datasets (Imbalance Ration=1:9).

Overall, AC-GAN and CVAE-GAN achieve a better performance comparing to VAE-GAN as they take label information into consideration to train a better generative model and therefore high-quality generated samples could help improve the performance of the discriminator.

In Fig. 4, we report the change of the latent two-component distribution during the training. For MNIST, the converged μ value is very small which means that the latent distribution over two classes tend to follow a unimodal distribution. This explains why VAE-GAN$_\mu$, which forces a simple unimodal distribution on latent representations, has a relatively good performance on MNIST as well. On the contrary, TIS has a relatively large converged μ value meaning that the latent distributions of positive and negative classes are different, so patterns between classes are overlapping at certain levels in feature space. As a result, discriminative rules are hard to induce so the model tend to increase μ to make patterns more separated for features. The flexibility of μ allows our model to better handle separability problem of imbalanced learning.

Comparing DVAAN$_a$ and DVAAN in Table 1, it is clear that data augmentation indeed helps improve the classification accuracy as synthetic samples generated close to the class boundary help discriminators learn better discriminative model. Similar to the core idea of SVM, the training aim of our discriminator is to find a mapping so that the support samples (vectors) of the separate categories are divided by a clear gap that is as wide as possible.

In Table. 2, we change the setting to one-against-all. The

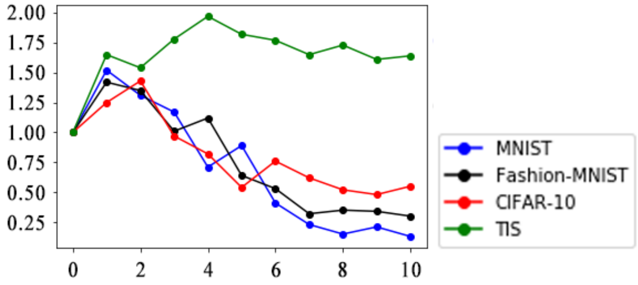

Figure 4: The change of μ values (y-axis) *w.r.t.* the iterations.

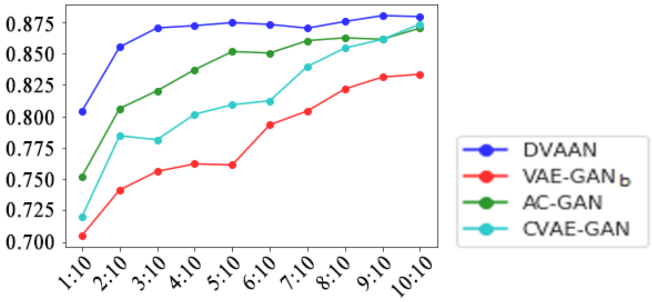

Figure 5: The AUC scores comparison *w.r.t.* different imbalance ratios on TIS. 1:10 means # of positive samples *vs.* # of negative samples are 1:10.

results show that our algorithm still remain good performance even if negative classes have complex distributions.

Figure. 5 reports the performance on TIS dataset by using different imbalance ratios. The results show that our model have better performance, especially for high imbalance ratios. This confirms that data augmentation, using synthetic samples to help characterize class boundary, is particularly useful when positive samples are sparse. Different from traditional sampling methods, DVAAN generates samples resemble to the training data of the same class and close to the class boundary and uses them to alleviate class imbalance.

4.4 Class Boundary Learning Results

In Fig. 6, we use DVAAN to generate samples from different latent distributions on MNIST dataset, so we can visually examine the quality of generated samples. Specifically, the first row and last row show the samples generated from two-component latent distribution. These samples can be clearly recognized with their belonging classes meaning that the learned latent representations indeed follow our proposed two-component latent distribution. The middle two rows are generated from a simple distribution with $\mu = 0$. These samples are more likely to be the ones close to the class boundary in latent space. For example, the second row of Fig. 6 is hardly recognized to belong to positive (digit 5) or negative (digit 6) class. It means these samples tend to share similar features/feature values. Adding these samples to the training data can help discriminator better learn the separating capability based on the shared similar features/feature values.

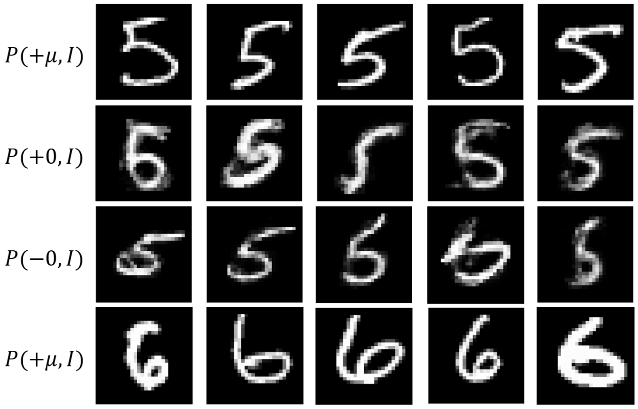

Figure 6: Samples generated from different latent distributions.

Methods	*MNIST*	*Fashion-MNIST*	*CIFAR-10*	*TIS*
VAE-GAN-1	0.0713	0.1734	0.1524	0.7321
VAE-GAN-2	0.0137	0.0824	0.0932	0.4213
CVAE-GAN	**0.0031**	**0.0427**	**0.0371**	0.4136
DVAAN-1	0.0115	0.0767	0.0726	0.6211
DVAAN-2	0.0072	0.0719	0.0623	0.4213
DVAAN-3	0.0059	0.0503	0.0569	**0.3836**

Table 3: Average Mean Squared Construction Errors of different models (Imbalance Ratio=1:5)

4.5 Reconstruction Error Comparison

Table 3 reports the reconstruction errors of VAE-based methods. The results show that CVAE-GAN has the least average MSE on MNIST, Fashion-MNIST, and CIFAR-10, because CVAE-GAN take label information as the latent condition for both encoder and decoder of the model. Even though the latent representations for positive and negative classes are in same latent space, they are somehow independent as the forced simple distribution is conditional to different classes and therefore CVAE enjoys a more flexible learning process for each single class. It is worth noting that DVAAN focuses on finding features best separating positive and negative samples rather than finding features with least information loss.

5 Conclusion

In this paper, we proposed a Discriminative Variational Autoencoding Adversarial Network (DVAAN) for deep imbalanced learning. We argued existing deep learning methods for imbalanced data cannot differentiate class boundaries, so their feature learning cannot be aligned to the class distributions to differentiate minority class from the majority class. In comparison, DVAAN combines a discriminative generative model and an adversarial network to explicitly learn class boundary in the latent representation space. As a result, it can control the generation of synthetic samples to alleviate class imbalance for deep imbalanced learning.

Acknowledgements

This research is sponsored in part by the US National Science Foundation through Grants No. IIS-1763452 and CNS-1828181.

References

[Bao *et al.*, 2017] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: Fine-grained image generation through asymmetric training. In *Prof. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, Oct 2017.

[Barua *et al.*, 2012] Sukarna Barua, Md. Monirul Islam, Xin Yao, and Kazuyuki Murase. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering*, 26(2):405–425, 2012.

[Chawla *et al.*, 2002] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[Chen and Shyu, 2011] Chao Chen and Mei-Ling Shyu. Clustering-based binary-class classification for imbalanced data sets. In *Proc. of the 12th IEEE Intl. Conf. on Information Reuse and Integration*, pages 384–389, 2011.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[Dosovitskiy and Brox, 2016] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Proc. of the 2016 Advances in Neural Information Proc. Systems*, pages 658–666, 2016.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. of the 2014 Advances in Neural Information Proc. Systems*, pages 2672–2680, 2014.

[He and Garcia, 2008] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, 9:1263–1284, 2008.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc.of the 2017 Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. of the International Conf. on Learning Representations*, 2014.

[Krawczyk, 2016] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[Larsen *et al.*, 2016] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proc. of the 33nd International Conf. on Machine Learning*, pages 1558–1566, 2016.

[Liu and Wong, 2003] Huiqing Liu and Limsoon Wong. Data mining tools for biological sequences. *Journal of Bio. and Computational Biology*, 1(1):139–167, 2003.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Odena *et al.*, 2016] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.

[Schilling *et al.*, 2002] Mark F Schilling, Ann E Watkins, and William Watkins. Is human height bimodal. *The American Statistician*, 56(3), 2002.

[Sohn *et al.*, 2015] Kihyuk Sohn, Xinchen Yan, and Honglak Lee. Learning structured output representation using deep conditional generative models. In *Proc. of the 2015 Advances in Neural Information Proc. Systems*, pages 3483–3491, 2015.

[Sun *et al.*, 2007] Yanmin Sun, Mohamed S.Kamel, Andrew K.C. Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007.

[van den Oord *et al.*, 2016] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *Proc. of the 2016 Advances in Neural Information Proc. Systems*, pages 4790–4798, 2016.

[Wang *et al.*, 2015] S. Wang, L. L. Minku, and X. Yao. Resampling-based ensemble methods for online class imbalance learning. *IEEE Transactions on Knowledge & Data Engineering*, 27(5):1356–1368, 2015.

[Wu *et al.*, 2017] Fei Wu, Xiao-Yuan Jing, Shiguang Shan, Wangmeng Zuo, and Jing-Yu Yang. Multiset feature learning for highly imbalanced data classification. In *Proc. of the 31st AAAI Conf. on Artificial Intelligence*, pages 1583–1589, 2017.

[Yan *et al.*, 2015] Yilin Yan, Min Chen, Mei-Ling Shyu, and Shu-Ching Chen. Deep learning for imbalanced multimedia data classification. In *Proc. of the 2015 IEEE International Symposium on Multimedia*, pages 483–488, 2015.

[Yan *et al.*, 2017] Yan Yan, Tianbao Yang, Yi Yang, and Jianhui Chen. A framework of online learning with imbalanced streaming data. In *Proc. of the 31st AAAI Conf. on Artificial Intelligence*, pages 2817–2823, 2017.

[Yu *et al.*, 2017] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proc. of the 31st AAAI Conf. on Artificial Intelligence*, pages 2852–2858, 2017.