IDA-GAN: A Novel Imbalanced Data Augmentation GAN

Hao Yang， Yun Zhou

Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology

Changsha，410073，China

E-mail: yanghao@nudt.edu.cn

Abstract

Class imbalance is a widely existed and challenging problem in real-world applications such as disease diagnosis, fraud detection, network intrusion detection and so on. Due to the scarce of data, it could significantly deteriorate the accuracy of classification. To address this challenge, we propose a novel Imbalanced Data Augmentation Generative Adversarial Networks (GAN) named IDA-GAN as an augmentation tool to deal with the imbalanced dataset. This is a great challenge because it is hard to train a GAN model under this situation. We address this issue by coupling variational autoencoder along with GAN training. In this paper, specifically, we introduce the variational autoencoder to learn the majority and minority class distributions in the latent space, and use the generative model to utilize each class distribution for the subsequent GAN training. The generative model learns useful features to generate target minority-class samples. Compared with the state-of-the-art GAN model, the experimental results demonstrate that our proposed IDA-GAN could generate more diverse minority samples with better qualities, and it could benefits the imbalanced classification task in terms of several widely-used evaluation metrics on five benchmark datasets: MNIST, Fashion-MNIST, SVHN, CIFAR-10 and GTSRB.

Introduction

Most classification algorithms are only suitable for balanced dataset that is equivalently distributed across different classes. However, in real-world, datasets are often imbalanced, and sometimes the minority-class is extremely important and need to be accurately classified. conventional approaches are oversampling, undersampling and cost-sensitive learning. The main idea of undersampling is to discard most majority-class samples to achieve balance across classes, but this approach might lose lots of information. Costsensitive learning method provides different weights through different types of samples, and it pay more attention to samples in minority-class. Nonetheless, using costsensitive learning, it is hard to obtain an accurate estimate of misclassification cost. Oversampling is also called data augmentation, which achieves balance mainly by adding minorityclass samples. However, traditional data augmentation usually apply some geometric changes on the image classification dataset, e.g. rotation, scaling, translation or mirroring and it might disrupt original relevant features. To address this problem, we introduce an improved GAN model named IDAGAN in this paper.

Methodology

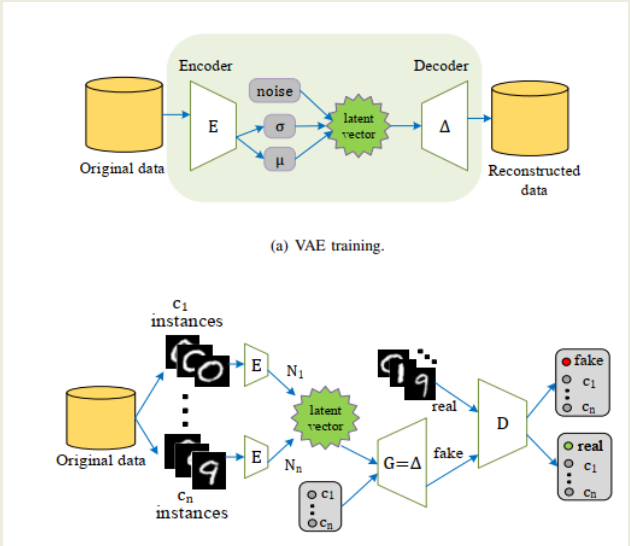

Fig. 1. The training of our proposed IDA-GAN

Experimental Results

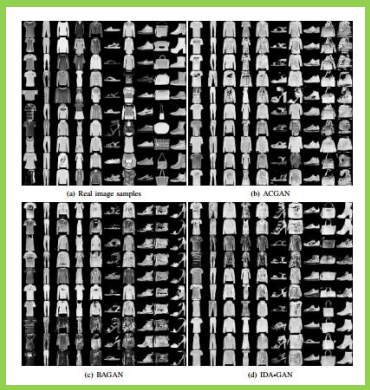

Fig. 2: Synthesized images generated by IDA-GAN.

Method	MNIST			Fashion-MNIST		
	Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
ACGAN	86.34	81.10	78.90	77.15	67.45	66.31
BAGAN	87.09	82.73	80.86	80.28	70.71	69.69
IDA-GAN	**88.45**	**83.25**	**82.56**	**83.33**	**79.72**	**78.85**

Method	SVHN			CIFAR-10			GTSRB		
	Precision (%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
ACGAN	71.93	54.80	55.39	64.31	51.58	48.99	83.04	83.08	81.72
BAGAN	77.55	73.94	72.39	69.42	67.82	61.65	85.55	85.60	84.46
IDA-GAN	**79.32**	**75.59**	**74.44**	**73.77**	**66.01**	**64.36**	**87.20**	**87.53**	**86.41**

Fig. 3: The Comparison of Classification performance in terms of Precision, Recall and F1 score.