PGASL: PREDICTIVE AND GENERATIVE ADVERSARIAL SEMI-SUPERVISED LEARNING FOR IMBALANCED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern machine learning techniques often suffer from class imbalance where only a small amount of data is available for minority classes. Classifiers trained on an imbalanced dataset, although have high accuracy on majority classes, can perform poorly on minority classes. This is problematic when minority classes are also important. Generative Adversarial Networks (GANs) have been proposed for generating artificial minority examples to balance the training. We propose a class-imbalanced semi-supervised learning algorithm PGASL which can be efficiently trained on unlabeled and class-imbalanced data. In this work, we use a predictive network which is trained adversarially for the discriminator to correct predictions on the unlabeled dataset. Experiments on text datasets show that PGASL outperforms state-of-the-art class-imbalanced learning algorithms by including both predictive network and generator.

1 INTRODUCTION

In many real world applications such as medical data analysis (Cameron et al., 2010), data is often imbalanced as those patients suffering from a certain disease will typically have a very small portion of the population. It is often challenging to train a machine learning model on imbalanced data since a classifier may be trained biased towards the majority classes and lead to poor performance for the minority classes. Although the classification may have overall good performance, this is not preferable especially for some tasks when the minority classes are also important. Addressing class imbalance has become more and more important in many fields. Another feature of these kinds of data is that usually only a small amount of data is labeled, for example, patients who visit hospitals who are under-diagnosed for a certain disease, can't be labeled as negative, thus remain unlabeled. Recently, many deep neural network based semi-supervised learning (SSL) algorithms have shown their ability of utilizing unlabeled data to improve performance. However, most works only focus on one of the above challenges and class-imbalanced semi-supervised learning is still under explored.

In this paper we consider the rare disease detection task in healthcare (Yu et al., 2019) which is a typical imbalanced semi-supervised machine learning task with binary outputs. As GANs have huge success in learning almost real data distributions through adversarial training, we propose a three-player GAN model which can generate artificial minority data as well as extracting minority data from the unlabeled dataset, which helps handling the class imbalance issue. We let the discriminator output probabilities of predictions for each class and we then use a predictive network, which has the same structure as the discriminator but is trained adversarially, to correct predictions of samples on the unlabeled dataset. Finally we use a generator to generate minority samples so that the discriminator will be trained using more balanced data and will be less likely to bias towards the majority class. Our main contributions include:

- We introduce a novel semi-supervised three-players GAN model for imbalanced data.
- We conduct experiments on binary text classification datasets to benchmark PGASL with previous class imbalance learning methods. Results demonstrate that our proposed method can efficiently utilize unlabeled data and handle imbalanced datasets that outperform prior works.

- To the best of our knowledge, this is the first GAN-based algorithm which can handle the class imbalance issue while utilizing the unlabeled dataset at the same time.

2 RELATED WORK

Semi-supervised (SSL) techniques have been proposed to improve performance by utilizing unlabeled data. Pseudo-Labels are used as if they were true labels on unlabeled datasets (Lee et al., 2013), and the models are trained in a manner of entropy minimization (Grandvalet & Bengio, 2004). Consistency regularizations aim at learning similar distributions for local perturbations on unlabeled dataset (Park et al., 2018; Miyato et al., 2018). Several data augmentation techniques in combination with Pseudo-Labels are used to create local perturbations (Berthelot et al., 2019a; Sohn et al., 2020; Berthelot et al., 2019b).

In this paper we want to highlight two GAN-based methods which our paper mainly follows. SSL-GANs use discriminator to classify samples from different classes instead of distinguishing fake samples from real samples. To perform adversarial training, one way is to directly minimize the entropy of the outputs for the discriminator which can help distinguish real samples from fake samples (Springenberg, 2015). On the other hand, one can use an additional class which indicates fake samples so that the discriminator can make predictions and recognize fake samples at the same time(Salimans et al., 2016; Odena, 2016). Other works use an additional network to play the role of distinguishing the fake samples (Mullick et al., 2019). PAN (Hu et al., 2021) replaces the generator with a classifier which mirrors the discriminator in classification on unlabeled dataset to help classify those samples which can be hardly classified by the discriminator. Although PAN was originally proposed for PU learning, it can be extended to semi-supervised learning naturally. We show in this paper, keeping the generator in PAN can have better performance in the class-imbalanced learning setting.

The SSL methods discussed above often assume balanced dataset and may have poor performance for an imbalanced dataset especially when the unlabeled dataset is also imbalanced (He & Garcia, 2009; Sun et al., 2009). Researchers have proposed several techniques for class-imbalanced learning. Re-sampling techniques have been proposed to create a balanced dataset (Chawla et al., 2002; Barandela et al., 2003; Engelmann & Lessmann, 2021b). However, oversampling may cause overfitting while undersampling may cause information loss (Cao et al., 2019). To avoid such issues, researchers have also proposed to balance the training by introducing regularization terms instead of creating balanced dataset (Huang et al., 2016; Jamal et al., 2020; Cui et al., 2019).

Recently, researchers have proposed several class-imbalanced semi-supervised learning (CISSL) techniques to improve performance on imbalanced datasets. Refining the pseudo-labels through an iterative algorithm can reduce bias (Kim et al., 2020; Wei et al., 2021). Introducing an auxiliary balanced classifier(ABC) of a single layer while using a 0/1 mask can balance training (Lee et al., 2021).

3 METHODS

3.1 PROBLEM SETTING

Suppose we have a binary labeled dataset $\mathcal{L} = \{(x_i, y_i), i = 1, \ldots, N_l\}$ where x_i is a $d-$dimensional feature vector and $y_i \in (0, 1)$ is the corresponding binary label. Let $N_l = N_p + N_n$ where N_p is the number of positive data and N_n is the number of negative data. We also have an unlabeled dataset $\mathcal{U} = \{u_i, i = 1, \ldots, N_u\}$ where u_i is also a $d-$dimensional feature vector but the corresponding label is unknown. We denote the label ratio as $\rho = \frac{N_l}{N_l+N_u}$. In this paper we consider the case when $\rho \approx 0.1$ where most of the labels are expensive to obtain. We also consider the extremely class-imbalanced situation where the imbalanced ratio $\gamma = \frac{N_p}{N_p+N_n}$ is very small. We assume the labeled dataset and the unlabeled dataset share the same distribution (e.g. have the same imbalanced ratio). Given the labeled dataset and unlabeled dataset in training we aim at learning a classifier $f : \mathbb{R}^d \to \{0, 1\}$ which can perform well on an imbalanced test set.

3.2 GENERATIVE ADVERSARIAL NETWORK

First proposed in (Goodfellow et al., 2014), generative adversarial networks (GANs) have shown its excellent ability of generating almost true samples which follow the real data distribution. The model simultaneously trains a generative network (often called generator) which generates samples that can be hardly distinguish from the real data to fool the discriminator, and a discriminator which tells if a sample is 'real' or 'fake' (generated by generator). GAN can be often formulated as a minmax game:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[\log(D(x))] + \mathbb{E}_{z\sim P_z(z)}[\log(1-D(G(z)))].$$

GANs can be trained to learn the distribution of minority classes and can then be used to generate samples to oversample minority classes to create balanced dataset (Engelmann & Lessmann, 2021b). Although more focus on generative tasks in general, researchers also demonstrate the good performance of GANs on classification tasks by letting discriminators to make predictions for different classes rather than only distinguish fake samples from real samples. Unlike resampling techniques which create a balanced dataset first and then apply classification algorithms afterward, letting discriminators make predictions allows generator to learn more proper distributions as they are trained together. Also it is easier to adjust the regions of samples which we would like the generator to generate and this is important for imbalance learning. It is often not desirable for generator to only generate samples from minority classes as this may cause overfitting and generating samples from positive samples also helps classification. Denote the ratio of positive and negative samples we expect the generator to generate by ρ_G. The choice of ρ_G is still under investigation. In this paper, we use ρ_G as a hyper-parameter and use grid search to choose the best ρ_G.

3.3 PROPOSED MODEL ARCHITECTURE

The architecture of our framework is shown in Figure 1. We aim at balancing the training through minority oversampling, as such, we propose a GAN model structure which can generate artificial minority samples as well as extracting minority samples from the unlabeled dataset so that more minority samples will be fed into the classifier. The model contains three components, a generator $G(\cdot)$ which takes random noise as input and learns to generate artificial samples from the real data distribution, a discriminator $D(\cdot)$ which makes predictions and a predictive network $C(\cdot)$ which mirrors the discriminator but is trained adversarially. We use the same network structure for both $D(\cdot)$ and $C(\cdot)$ except the dropout rate. We found in practice, letting $C(\cdot)$ has larger dropout rate gives better performance. In this paper we consider binary classification problems and therefore both the predictions from $D(\cdot)$ and $C(\cdot)$ have three classes with 0 as negative, 1 as positive and 2 as 'fake'. In particular, we consider the rare disease scenario where the positive class is the minority.

The discriminator $D(\cdot)$ first learns to classify samples from the labeled dataset correctly using standard entropy minimization for supervised learning. The supervised loss is

$$\ell_{sup}(x,y) = -\mathop{\mathbb{E}}_{x,y\in\mathcal{L}} \log p_D(y|x, y<2),$$

which minimizes the negative log probability of the label, given the labeled data. Then $D(\cdot)$ learns to recognize the artificial samples generated by the generator $G(\cdot)$ by classifying the generated samples as 'fake' (2). This correspond to minimize the loss

$$\ell_{fake}(z) = -\mathop{\mathbb{E}}_{x\in G(z)} \log p_D(2|x).$$

Finally, $D(\cdot)$ has two unsupervised loss. Minimizing the first unsupervised loss

$$\ell^D_{un}(x) = -\mathop{\mathbb{E}}_{x\in\mathcal{U}} \log\left(1 - p_D(2|x)\right)$$

allows $D(\cdot)$ to recognize the unlabeled data by maximizing the probability of a sample from unlabeled dataset is not 'fake'. Minimizing the second unsupervised loss

$$\ell^*_{un}(x) - \mathop{\mathbb{E}}_{x\in\mathcal{U}}[\log p_D(x) - \log p_C(x)]$$

allows $D(\cdot)$ to maximize the distance between the predictions from $D(\cdot)$ and $C(\cdot)$ on the unlabeled dataset. In this paper, we also use KL divergence to measure the distance as in PAN (Hu et al., 2021). Then the total loss for $D(\cdot)$ is:

$$\ell^D = \ell_{sup} + \ell_{fake} + \lambda\left(\ell^D_{un} + \ell^*_{un}\right),$$

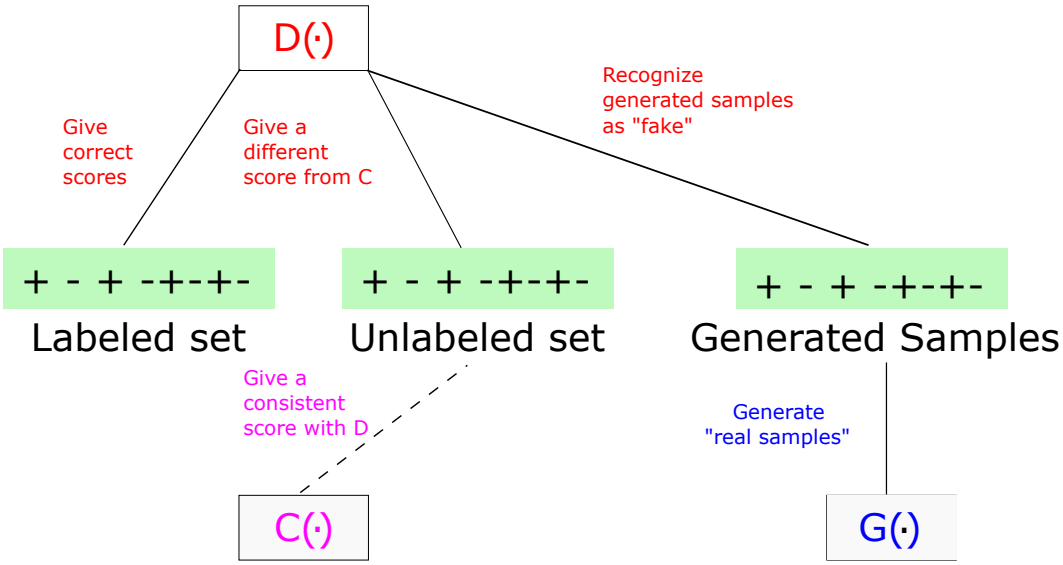

Figure 1: Framework of proposed architecture on semi-supervised learning. The utilities of $D(\cdot)$, $C(\cdot)$ and $G(\cdot)$ are colored in red, purple and blue respectively.

where λ is a hyper-parameter for balancing the supervised loss and the unsupervised loss. To perform adversarial training, the predictive network $C(\cdot)$ learns to shrink the distance between the predictions from $D(\cdot)$ and $C(\cdot)$ by maximizing ℓ_{un}^* so that when reach an equilibrium, $C(\cdot)$ can make predictions with large margin on the unlabeled dataset through adversarial training. Unlike PAN, we keep the existence of the generator for handling the class imbalance issue. Denote the set of positive samples as $\mathcal{P}$ and the set of negative samples as $\mathcal{N}^*$. To train the generator, we sample n data points from $\mathcal{N}^*$ and $\rho_G n$ data points from $\mathcal{P}$ where ρ_G is a hyperparameter. In practice, ρ_G should be greater than 1 if positive is the minority class. Denote the set of the $(\rho_G + 1)n$ data points as $\mathcal{L}_G$. We use the last layer of $D(\cdot)$ as features and map the features of labeled samples and the features of generated samples. This is achieved by minimizing the following feature mapping loss:

$$\ell_{fm} = || \mathop{\mathbb{E}}_{x \in \mathcal{L}_G} D^h(x) - \mathop{\mathbb{E}}_{t \in G(z_1)} D^h(t)||_2^2.$$

In addition, we use the pull away term (Zhao et al., 2016)

$$\ell_{pt} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \left(\frac{f(x_i)^T f(x_j)}{||f(x_i)||||f(x_j)||} \right)^2$$

to increase the diversity of generated samples. Note that the functionality of $C(\cdot)$ is similar to $G(\cdot)$. The difference is that $G(\cdot)$ generate artificial data that is not from training dataset while $C(\cdot)$ extract data from the unlabeled dataset. Both $C(\cdot)$ and $G(\cdot)$ contribute to creating more positive samples for $D(\cdot)$ to balance training.

3.4 Training of PGASL

Algorithm 1 gives the training procedure of PGASL using stochastic optimization algorithms. The loss functions are defined in Section 3.3. The algorithm trains $D(\cdot)$, $C(\cdot)$ and $G(\cdot)$ alternatively. In practice, one can choose to run more iterations for one component in one epoch to boost performance.

4 Results

4.1 Data Source and Pre-processing

In this section, we tested the proposed method PGASL and compare it with state-of-the-art algorithms. In this paper we consider the rare disease scenario which has binary outputs with posi-

Algorithm 1 PGASL for semi-supervised learning

Input: labeled dataset $\mathcal{L}$, unlabeled dataset $\mathcal{U}$, batch size m, hyperparameters λ, ρ_G
Output: D

1: **for** k steps **do**
2: Train D
- Sample a mini-batch $\{(x_i, y_i) \in \mathcal{L}, i = 1, \ldots, m\}$ from labeled dataset and Compute supervised loss $\ell_{sup}(D(x_i), y_i)$.
- Sample a mini-batch $\{(x_i') \in \mathcal{U}, i = 1, \ldots, m\}$ from unlabeled dataset and Compute first unsupervised loss $\ell_{un}^{D}(D(x_i'))$.
- Sample a mini-batch $\{(x_i'') \in \mathcal{U}, i = 1, \ldots, m\}$ from unlabeled dataset and Compute second unsupervised loss $\ell_{un}^{*}(D(x_i''))$.
- Sample noise $z \in \mathcal{N}(0, 1)$ and compute fake loss $\ell_{fake}(D(G(z)))$.
- Update $D(\cdot)$ by minimizing $\ell_D = \ell_{sup} + \lambda(\ell_{un}^{\mathcal{D}} + \ell_{un}^{*}) + \ell_{fake}$.

3: Train C
- Sample a mini-batch $\{(x_i) \in \mathcal{U}, i = 1, \ldots, m\}$ from unlabeled dataset and update $C(\cdot)$ by maximizing the loss $\ell_{un}^{*}(D(x_i))$.

4: Train G
- Sample a mini-batch $\{(x_i^*, 0) \in \mathcal{N}^*, i = 1, \ldots, \frac{m}{\rho_G}\}$, sample a mini-batch $\{(x_i^{*'}, 1) \in \mathcal{P}, i = 1, \ldots, m\}$ and sample noise $z \in \mathcal{N}(0, 1)$ and compute feature mapping loss ℓ_{fm} and pull away term ℓ_{pt}.
- Update $G(\cdot)$ by minimizing $\ell_G = \ell_{fm} + \ell_{pt}$.

5: **end for**

tive class being the minority class. We first created imbalanced versions of two text classification datasets: Amazon polarity reviews (McAuley & Leskovec, 2013) and Yelp polarity reviews (Zhang et al., 2015) using various imbalance ratio γ. We kept all negative data with N_n data points, then we randomly selected $N_p = \frac{\gamma}{1-\gamma} N_n$ positive data points. For the purpose of semi-supervised learning, we randomly selected 10% of the data as labeled data while the rest 90% data as unlabeled data for all datasets. Since the main focus of this paper is semi-supervised learning rather than representation learning, we first used pretrained sentence transformers (Reimers & Gurevych, 2019) to preprocess all the text into sentence embeddings. Then the classification algorithms are applied to the same pretrained sentence embeddings.

We summarize the information of each dataset in Table 1.

Table 1: Number of samples of each class in each dataset

	labeled neg	labeled pos	unlabel neg	unlabel pos	test neg	test pos
Amazon $\gamma = 0.1$	180153	18016	1619847	161985	200000	20000
Amazon $\gamma = 0.05$	179406	8971	1620594	81030	200000	10000
Amazon $\gamma = 0.01$	179848	1799	1620152	16202	200000	2000
Yelp $\gamma = 0.1$	27996	2800	252004	25201	19000	1900
Yelp $\gamma = 0.05$	28004	1401	251996	12600	19000	950
Yelp $\gamma = 0.01$	27939	280	252061	2521	19000	190

4.2 METRICS AND BASELINES

To evaluate our method, we mainly use three metrics: the area under the receiving operator curve(AUC-ROC), the area under the precision recall curve (PR-AUC), and the brier score. The AUC-ROC score, represents the area under the curve of the true positive rate against the false positive rate for each decision threshold, which evaluates a model on ranking predictions in general while a more proper metric for imbalanced data is the PR-AUC score which can be viewed as the average of precision scores calculated for each recall threshold. We also use the brier score which measures the accuracy of probabilistic predictions. The reason we pick these three metrics is because

the evaluations of these scores don't require thresholds and can be directly applied to probabilities. This is important since a threshold of 0.5 usually doesn't work well for imbalanced datasets. We compared the above metrics of the proposed method with the following baselines:

- CWGAN+xgboost: This method first use CWGAN (Engelmann & Lessmann, 2021a) to first oversample the imbalanced datasets and then train xgboost (Chen & Guestrin, 2016) on the oversampled balanced training datasets in a fully supervised learning setting.
- SGAN(Odena, 2016): This method learns discriminators using GAN that output class labels and utilize unlabeled data in a semi-supervised setting. We use the same training of generator for PGASL and SGAN.
- PAN(Hu et al., 2021): This method trains classifier adversarially to the discriminator on unlabeled dataset to improve performance. It was originally proposed for PU learning, we extend it to semi-supervised learning in this paper.
- ABC(Lee et al., 2021): This method attaches a classifier to the last layer of backbone algorithm and uses a 0/1 mask to balance training. We use the same discriminator in PGASL as the backbone algorithm for ABC.

4.3 Training Details

We use fully-connected neural networks with leaky relu activation (Maas et al., 2013) and dropout (Srivastava et al., 2014) for $D(\cdot)$ and $C(\cdot)$ in PGASL. For fair comparison we also use the same network structure for the networks in PAN, discriminator in SGAN and the backbone in ABC. We use grid search for each method to search for the hyper-parameters of the best PR-AUC scores which we consider as the most important metric. Then to reduce the effect of randomness in training, for each method we run 10 repeated experiments with the best hyper-parameters from grid search and we compare the average scores. We use pytorch(Paszke et al., 2019) for implementation and use adam optimizer (Kingma & Ba, 2014) for training.

4.4 Results and analysis

The ROC-AUC scores of each algorithm on test datasets are summarized in Table 2. Except for the Yelp dataset when $\gamma = 0.05$, PGASL has the best scores. Note that all the ROC-AUC scores are large even for the extremely imbalanced dataset. This is because the ROC-AUC score can often give an overly-optimistic performance of a classifier on imbalanced datasets. We believe the PR-AUC score is more proper in the class-imbalance setting and can better measure the performance of the algorithms. The PR-AUC scores are summarized in Table 3. We also summarize the brier scores in Table 4. Note that unlike ROC-AUCs and PR-AUCs, the brier score works like a loss function with 0 representing perfect accuracy. PGASL outperforms SGAN since PGASL has an additional predictive network which increase the prediction accuracy on unlabeled dataset during training. PGASL outperforms PAN since PGASL keeps the generator to generate more balanced data which benefit class-imbalanced learning. Moreover PGASL outperforms the oversampling technique CWGAN+XGB and the state-of-the-art class-imbalanced semi-supervised learning algorithm ABC.

Table 2: Test ROC-AUCs of PGASL model and benchmark models. The scores are averaged over 10 experiments.

	Amazon			Yelp		
Algorithm	$\gamma = 0.1$	$\gamma = 0.05$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 0.05$	$\gamma = 0.01$
CWGAN+XGB	0.896	0.890	0.869	0.944	0.932	0.914
SGAN	0.904	0.895	0.869	0.954	**0.942**	0.938
PAN	0.906	0.895	0.866	0.954	0.941	0.938
ABC	0.903	0.902	0.892	0.954	0.941	0.923
PGASL	**0.916**	**0.909**	**0.909**	**0.956**	0.941	**0.941**

Next we take the Yelp polarity dataset with $\gamma = 0.1$ as an example and show the discriminator supervised training losses, the predictive network training losses and the generator loss in Figure 2.

Table 3: Test PR-AUCs of PGASL model and benchmark models. The scores are averaged over 10 experiments.

	Amazon			Yelp		
Algorithm	$\gamma = 0.1$	$\gamma = 0.05$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 0.05$	$\gamma = 0.01$
CWGAN+XGB	0.589	0.447	0.1530	0.744	0.619	0.302
SGAN	0.605	0.456	0.153	0.784	0.665	0.361
PAN	0.612	0.458	0.151	0.785	0.663	0.354
ABC	0.616	0.487	0.194	0.782	0.661	0.336
PGASL	**0.646**	**0.511**	**0.210**	**0.787**	**0.668**	**0.371**

Table 4: Test brier score of PGASL model and benchmark models. The scores are averaged over 10 experiments. Smaller scores indicate better performance.

	Amazon			Yelp		
Algorithm	$\gamma = 0.1$	$\gamma = 0.05$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 0.05$	$\gamma = 0.01$
CWGAN+XGB	0.0543	0.0342	0.0094	0.0445	0.0299	0.0089
SGAN	0.0567	0.0356	0.0100	0.0378	0.0253	0.0079
PAN	0.0552	0.0362	0.0099	0.0376	0.0255	0.0078
ABC	0.0520	0.0324	0.0088	0.0377	0.0260	0.0079
PGASL	**0.0501**	**0.0316**	**0.0084**	**0.0371**	**0.0251**	**0.0076**

We can see that the discriminator $D(\cdot)$ can successfully learn the correct classification on labeled dataset and the generator $G(\cdot)$ learns to generate diverse samples with similar features to the real data. The predictive network $C(\cdot)$ trained adversarially, although was beat by $D(\cdot)$ at the beginning of training, reach an equilibrium eventually.

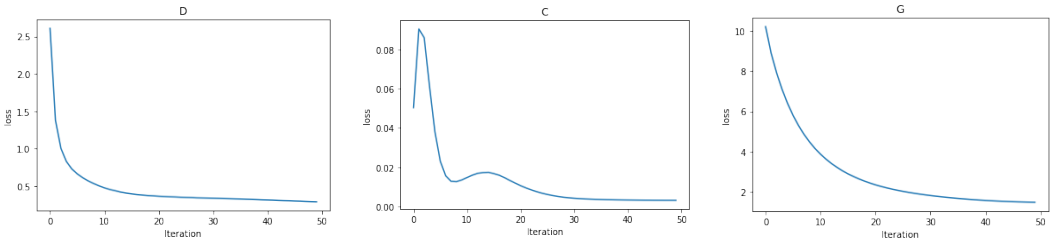

Figure 2: Training losses for $D(\cdot)$,$C(\cdot)$ and $G(\cdot)$

4.5 Visualization for Generator

Finally we use t-SNE (Van der Maaten & Hinton, 2008) to visualize the samples generated by the generator learned above. For the purpose of better visualization, we choose the same number of positive and negative data points from the imbalanced labeled dataset. The visualization result is shown in Figure 3. We can see that the the generator can generate both positive and negative samples which are in the positive and negative clusters of labeled data points respectively.

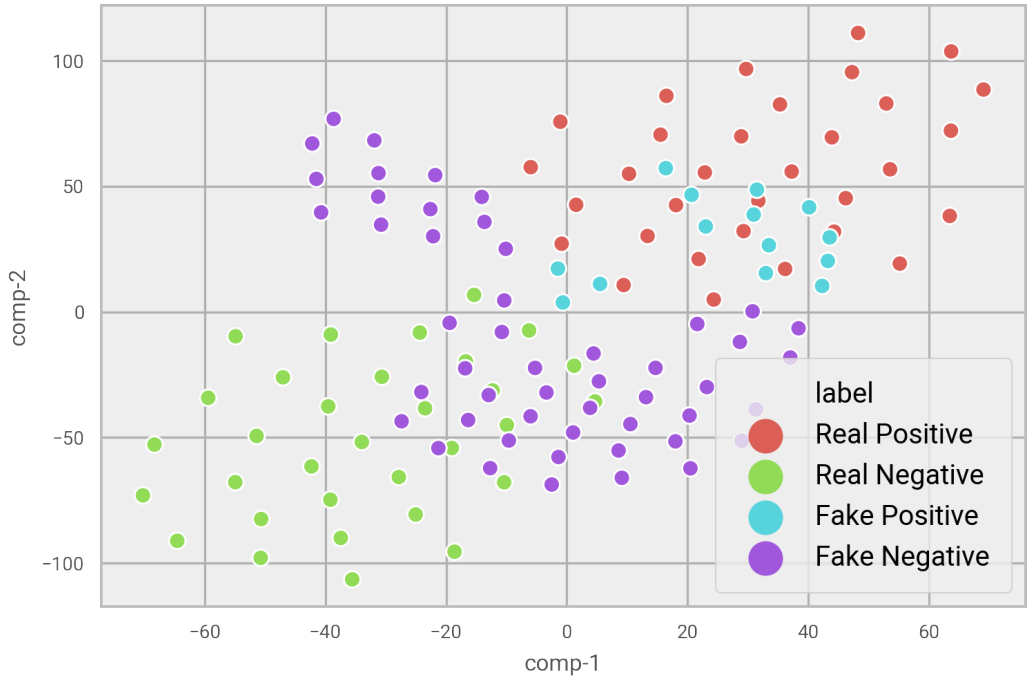

Figure 3: T-sne visualizations where red points and green points represent samples from labeled dataset while blue points and purple points represent artificial samples from generator.

5 Conclusions

In this paper, we proposed a novel GAN-based algorithm called PGASL. PGASL is a three-player model which has an additional predictive network. Unlike other GAN-based algorithms where the adversarial trainings are only performed by discriminators against a generator, the predictive network in PGASL is trained adversarial to the discriminator to boost the performance on the unlabeled dataset. We also showed in the paper that PGASL outperforms other GAN-based models as well as state-of-the-art class-imbalanced semi-supervised learning algorithms on imbalanced datasets. Our future work will be further improve the performance of PGASL by developing more robust training for the generator.

References

Ricardo Barandela, E Rangel, José Salvador Sánchez, and Francesc J Ferri. Restricted decontamination for the imbalanced training sample problem. In *Iberoamerican congress on pattern recognition*, pp. 424–431. Springer, 2003.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019b.

Marcia J Cameron, Micki Horst, Larry W Lawhorne, and Peter A Lichtenberg. Evaluation of academic detailing for primary care physician dementia education. *American Journal of Alzheimer's Disease & Other Dementias®*, 25(4):333–339, 2010.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

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.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Justin Engelmann and Stefan Lessmann. Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, 174:114582, 2021a.

Justin Engelmann and Stefan Lessmann. Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, 174:114582, 2021b.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Wenpeng Hu, Ran Le, Bing Liu, Feng Ji, Jinwen Ma, Dongyan Zhao, and Rui Yan. Predictive adversarial learning from positive and unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7806–7814, 2021.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5375–5384, 2016.

Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7610–7619, 2020.

Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. *Advances in neural information processing systems*, 33:14567–14579, 2020.

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

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

Hyuck Lee, Seungjae Shin, and Heeyoung Kim. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. *Advances in Neural Information Processing Systems*, 34: 7082–7094, 2021.

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, pp. 3. Citeseer, 2013.

Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pp. 165–172, 2013.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1695–1704, 2019.

Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

Sungrae Park, JunKeon Park, Su-Jin Shin, and Il-Chul Moon. Adversarial dropout for supervised and semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29: 2234–2242, 2016.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10857–10866, 2021.

Kezi Yu, Yunlong Wang, and Yong Cai. Modelling patient sequences for rare disease detection with semi-supervised generative adversarial nets. In *International Workshop on Advanced Analysis and Learning on Temporal Data*, pp. 141–150. Springer, 2019.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. *arXiv:1509.01626 [cs]*, September 2015.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.