Enhancing TripleGAN for Semi-Supervised Conditional Instance Synthesis and Classification

Si Wu, Guangchang Deng, Jichang Li, Rui Li, Zhiwen Yu, Hau-San Wong

School of Computer Science and Engineering, South China University of Technology
Department of Computer Science, City University of Hong Kong

cswusi@scut.edu.cn, csgc@mail.scut.edu.cn, cslijichang@mail.scut.edu.cn
ruili52-c@my.cityu.edu.hk, zhwyu@scut.edu.cn, cshswong@cityu.edu.hk

Abstract

Learning class-conditional data distributions is crucial for Generative Adversarial Networks (GAN) in semi-supervised learning. To improve both instance synthesis and classification in this setting, we propose an enhanced TripleGAN (EnhancedTGAN) model in this work. We follow the adversarial training scheme of the original TripleGAN, but completely re-design the training targets of the generator and classifier. Specifically, we adopt feature-semantics matching to enhance the generator in learning class-conditional distributions from both the aspects of statistics in the latent space and semantics consistency with respect to the generator and classifier. Since a limited amount of labeled data is not sufficient to determine satisfactory decision boundaries, we include two classifiers, and incorporate collaborative learning into our model to provide better guidance for generator training. The synthesized high-fidelity data can in turn be used for improving classifier training. In the experiments, the superior performance of our approach on multiple benchmark datasets demonstrates the effectiveness of the mutual reinforcement between the generator and classifiers in facilitating semi-supervised instance synthesis and classification.

1. Introduction

Significant advances in deep learning techniques have resulted in its wide adoption in a broad set of applications, most notably in computer vision [14] [31] [34] and natural language processing [42]. However, in view of the need to collect a massive amount of labeled data for most fully-supervised deep learning models, semi-supervised learning represents an effective approach to avoid the need for extensive manual annotations. This is due to the capability of semi-supervised learning to capture the characteristics of a dataset through a small number of labeled instances, togeth-
Figure 2. Illustration of the structure of the proposed EnhancedTGAN model for semi-supervised conditional instance synthesis and classification.

In this work, we propose an enhanced TripleGAN (EnhancedTGAN) model for improving semi-supervised conditional instance synthesis and classification. We follow the adversarial training scheme of the TripleGAN model in general, but re-design the overall loss functions of the generator and classifiers. Our feature-semantics matching approach is able to reduce the risk of mode collapse and improve the synthesis of the instances of each class. Specifically, we adopt class-wise mean feature matching to regularize the generator, such that the class-conditional distribution of the synthesized data can match with that of the real data for each class in the latent space learnt by a classifier, instead of the discriminator. In addition, we further include a semantic matching term to ensure the semantics consistency of the synthesized data between the generator and the classifier, which is a prerequisite for improving classifier training. On the other hand, a better classification model can provide more accurate categorical information on a large number of unlabeled instances, which leads to better guidance for the generator. For this purpose, we include two classifiers in our model which operate in a collaborative learning fashion. The classifiers can learn from each other by penalizing the divergence between the predicted class probability distributions, and the consensus predictions on the unlabeled data are more accurate than individual predictions in most cases. As a result, the generator and the classifiers can mutually reinforce each other. The structure of our model is illustrated in Figure 2. Our experiments verify the effectiveness and superiority of the proposed model. The main contributions of this work are summarized as follows:

- We propose a feature-semantics matching approach through which the generator can more effectively learn the class-conditional data distributions.
- In addition to the synthesized high-fidelity instances for training data augmentation, collaborative learning between the classifiers can also lead to more accurate classification on the unlabeled data, which in turn provides better guidance for the generator.
- The proposed enhanced TripleGAN model improves the state-of-the-art results in both semi-supervised instance synthesis and classification on multiple widely used benchmarks.

2. Related Work

Recently, various strategies have been applied to improve semi-supervised deep learning. Unsupervised learning can be used as an auxiliary task for exploring the structure of a dataset, and thus the generalization capability of the classification model can be improved. To formulate the unsupervised learning loss function for unlabeled samples, Rasmus et al. [30] proposed the $\Gamma$-model, in which a consistency regularization term was adopted to penalize the inconsistent predictions of the Ladder network [37] in inputs with...
and without noise. Similar to the G-model, Laine and Aila [15] proposed the Π-model, which regularized the model outputs of the training samples under different dropout and augmentation conditions. To provide more stable training targets for the unsupervised loss, Laine and Aila further proposed the Temporal-Ensembling model. In this model, self-ensembling of the network was applied to complement supervision. The ensemble of predictions at different epochs are expected to be more accurate, and can thus be used as the training targets for unlabeled samples. In contrast to the exponential moving average of the network predictions on each sample, Tarvainen and Valpola [36] proposed the Mean-Teacher model to average the network weights, and the resulting model can be considered as a teacher to provide the training targets for the unlabeled samples. Based on the assumption that similar samples should be located in the same cluster in the latent space, Luo et al. [19] proposed the Smooth Neighbors on Teacher Graphs (SNTG) method to build a graph based on the predictions of a teacher network for measuring the similarity between unlabeled data points. To ensure smoothness on the data manifold, the contrastive loss was used to ensure that neighbors had consistent predictions, while non-neighbors were pushed away from each other. In addition, it is common to incorporate the locally-Lipschitz condition through penalizing inconsistent predictions of unlabeled samples with different perturbations. Miyato et al. [24] [23] proposed the Virtual Adversarial Training (VAT) based regularization method to improve the local smoothness of the predicted class probability distribution by applying perturbation in the adversarial direction with respect to the classification model. VAT can be incorporated into existing semi-supervised learning networks, and yields impressive results. On the other hand, Park et al. [28] developed the Virtual Adversarial Dropout (VaD) approach to reconfigure the neural network and minimize the divergence between the obtained network and the original network for increasing the sparsity of the overall network.

Deep generative models have been recently applied to semi-supervised learning. In [10], Kingma et al. adopted the Variational Auto-Encoder (VAE) model [11] to treat class label as an additional latent variable in the process of learning the generative model. In [27], Odena modified the discriminator network to simultaneously distinguish real samples from synthesized samples and predict the corresponding class labels. Springenberg [35] proposed the Categorical Generative Adversarial Network (CatGAN) to make the discriminator assign high-confidence class labels for the real samples, while forcing the predicted class probability distributions on the synthesized samples to be uniform. Salimans et al. [32] proposed a variety of training techniques to improve the GAN training procedure, which lead to improvement in semi-supervised learning and sample synthesis. Furthermore, Wei et al. [38] improved the Wasserstein GAN [2] by including a consistency term with respect to the discriminator responses for enforcing Lipschitz continuity. To prevent the discriminator from playing two roles of identifying synthesized samples and predicting class labels for real samples in a minimax game, Li et al. [17] incorporated a classifier as an additional player into the game, and proposed the Triple Generative Adversarial Net (TriangleGAN). Dumoulin et al. [5] proposed the Adversarially Learned Inference (ALI) model, in which a generation network learnt the mapping from the latent space to the data space, while an inference network learnt the inverse mapping. These two networks were jointly optimized with a discriminative network in an adversarial process. To learn the joint distribution between samples and labels, Gan et al. [7] proposed the Triangle Generative Adversarial Network (TriangleGAN), in which two generators were adopted to learn the conditional distributions between samples and labels, and two discriminators were used to identify the types of fake pairs between real (fake) samples and fake (real) labels. Instead of matching the real and fake data distributions, Dai et al. [4] proposed a complementary generator which was trained by minimizing the KL divergence between the distributions, such that the generated samples were located in the low-density region in the latent space, and the diversity of the training data was increased.

TriangleGAN is the most related work to our proposed approach. However there are significant differences between them. Although we follow the adversarial training scheme of TriangleGAN in general, we completely re-design the overall loss function of the generator by including feature-semantics matching for effective and efficient learning of the class-conditional data distributions. In addition, we include two classifiers which collaboratively learn from each other to provide more accurate categorical information for the generator. As a result, the generator and classifier mutually reinforce each other to facilitate semi-supervised instance synthesis and classification.

3. Method

Inspired by the method in [6], the maximum mean discrepancy measure was used for training GANs. Feature matching has shown effectiveness in addressing the instability problem in GAN. The objective function defined below can be applied to force the generator \( G \) to synthesize data that matches the statistics of the real data \([32][3]\):

\[
\| E_{x \sim p_{\text{data}}} f_D(x) - E_{z \sim p_z} f_D(G(z)) \|, \tag{1}
\]

where \( p_{\text{data}} \) denotes the distribution of real data \( x \), \( p_z \) denotes the distribution of random vector \( z \), e.g., \( U[0,1] \), \( G(z) \) denotes a synthesized sample from \( z \), and \( f_D(\cdot) \) denotes the features associated with the hidden layer of the discriminator \( D \). The center of synthesized data points is forced to
match that of real data points in the latent space learnt by the discriminator. However, there are two main issues when applying the above formulation to our task. On one hand, the categories of the instances are not taken into account during the process of matching the marginal distributions. On the other hand, in addition to class-conditional instance synthesis, another objective is to perform accurate classification on the unlabeled data, while feature matching in the space learnt by the discriminator cannot directly improve classification. In this section, we introduce the EnhancedTGAN model to improve both semi-supervised conditional instance synthesis and classification.

### 3.1. Feature-Semantics Matching

In our setting, only a small portion of the training samples are labeled. Let \( x \sim p_u \) denote unlabeled samples, and \( (x, y) \sim p_l \) denote the labeled data pair, where \( y \) denotes the label of sample \( x \). Our EnhancedTGAN consists of the following four modules: the generator \( G \), discriminator \( D \), and classifiers \( C_1 \) and \( C_2 \). We slightly modify the adversarial training scheme of the TripleGAN model. Specifically, the generator \( G \) synthesizes new instances by sampling pairs of random vector and class label \((\tilde{z}, \tilde{y})\) from a pre-specified distribution \( p_g \). The two classifiers \( C_1 \) and \( C_2 \) collaboratively learn from each other, and produce the consensus prediction \( \tilde{y} \) of the input data.

To improve class-conditional instance synthesis, we optimize the generator by including the class-wise mean feature matching term defined as follows:

\[
\ell_{\text{feaMat}}(\theta_G) = \sum_k \left\| \mathbb{E}_{(x,y) \sim p_l} [1(y,k)f_{C_1}(x)] - \mathbb{E}_{(\tilde{z},\tilde{y}) \sim p_g} [1(\tilde{y},k)f_{C_1}(G(\tilde{z},\tilde{y}))] \right\|, \tag{2}
\]

where \( k \) denotes the class index, \( f_{C_1}(\cdot) \) denotes the features on the hidden layer of the classifier \( C_1 \), and the function \( 1(\cdot,\cdot) \) returns 1 if the inputs are equal and 0 otherwise. Since the number of labeled instances is small, we can use moving historical averages to obtain more stable means for them. The main advantage of \( \ell_{\text{feaMat}} \) is to avoid the mode collapse problem where the generator always outputs the same point. Another advantage is to increase the separability of different classes of synthesized data. In order to utilize the synthesized samples for training the classifiers, their semantics from the perspectives of the classifier and generator should be consistent. To enforce this consistency, we adopt a semantics matching term to regularize the generator as follows:

\[
\ell_{\text{semMat}}(\theta_G) = \mathbb{E}_{(\tilde{z},\tilde{y}) \sim p_g} [-\tilde{y} \log \tilde{p}_{C}(G(\tilde{z},\tilde{y}))], \tag{3}
\]

where \( \tilde{p}_{C}(x) = \text{avg-pool}(p_{C_1}(x), p_{C_2}(x)) \), \( p_{C_1}(\cdot) \) (\( p_{C_2}(\cdot) \)) denotes the predicted class probability distribution by classifier \( C_1 \) (\( C_2 \)). The average pooling of the classifier predictions can be expected to be more accurate in most cases. After including the adversarial training term with the discriminator, the optimization of the generator can be formulated as follows:

\[
\min_{G} \frac{1}{2} \mathbb{E}_{(z,\tilde{y}) \sim p_g} \left\{ \log(1 - D(G(z,\tilde{y}), \tilde{y})) \right\} + \eta \ell_{\text{feaMat}} + \nu \ell_{\text{semMat}}, \tag{5}
\]

where the weighting factors \( \eta \) and \( \nu \) are used for controlling the relative importance of the corresponding terms.

### 3.2. Collaborative Learning of Classifiers

Different from the TripleGAN model, we include two classifiers in our model, due to the reason that they can provide the training targets of the unlabeled instances for each other via collaborative learning. Existing works have demonstrated that collaborative learning is capable of facilitating semi-supervised classification.

Similar to the given labeled instances, the synthesized instances can also be utilized because of the known labels. The classifier can be enhanced by including the instances from the generator. The loss measure for supervised learning is the cross entropy between the given labels and the predicted distribution. Furthermore, the model can also learn from the unlabeled samples by minimizing the conditional entropy with respect to the posterior probability distribution. Therefore, we define the term for classification evaluation as follows:

\[
\ell_{\text{classify}}(\theta_{C_1}) = \mathbb{E}_{(x,y) \sim p_l} \left[ -y \log p_{C_1}(x) \right] + \mathbb{E}_{(\tilde{z},\tilde{y}) \sim p_g} \left[ -\tilde{y} \log p_{C_1}(G(z,\tilde{y})) \right] + \mathbb{E}_{u \sim p_u} \left[ -p_{C_1}(x) \log p_{C_1}(x) \right]. \tag{6}
\]

The classifiers tend to be confident on the unlabeled samples. To stabilize the estimation of the conditional entropy, a smoothness regularization term \( \ell_{\text{smoReg}} \) is defined as follows:

\[
\ell_{\text{smoReg}}(\theta_{C_1}) = \mathbb{E}_{x \sim p_u} \max_{\|\gamma\| \leq \xi} \text{KL}(p_{C_1}(x) \| p_{C_1}(x + \gamma)), \tag{7}
\]

where the constant \( \xi \) is used to control the intensity of the adversarial perturbation \( \gamma \), and \( \text{KL}(\cdot \| \cdot) \) denotes the Kullback-Leibler (KL) divergence. Similar to [24], the perturbation is generated in the direction most sensitive to the classifier prediction, and the KL divergence is used to measure the prediction difference with respect to the classifier for cases with and without perturbation. As a result, the output of the classifier will become smooth in the neighborhood of unlabeled samples.
As a symmetrized and smoothed version of the KL divergence, we further define a consistency regularization term $\ell_{conReg}$ by adopting the Jensen-Shannon (JS) divergence to measure the similarity between the posterior probability distributions of the two classifiers as follows:

$$
\ell_{conReg}(\theta_{C1}, \theta_{C2}) = \mathbb{E}_{x \sim p_c} [\mathbb{D}_{JS}(p_{C1}(x), p_{C2}(x))] \\
+ \mathbb{E}_{(z, \hat{y}) \sim p_u} [\mathbb{D}_{JS}(p_{C1}(G(z, \hat{y})), p_{C2}(G(z, \hat{y})))] .
$$

As a symmetrized and smoothed version of the KL divergence, $\mathbb{D}_{JS}$ is defined by

$$
\mathbb{D}_{JS}(p_{C1}(x), p_{C2}(x)) = \frac{1}{2} \text{KL}(p_{C1}(x) \| \bar{p}_C(x)) \\
+ \frac{1}{2} \text{KL}(p_{C2}(x) \| \bar{p}_C(x)) ,
$$

In addition, $\mathbb{D}_{JS}(p_{C1}(G(z, \hat{y})), p_{C2}(G(z, \hat{y})))$ has a similar definition. Based on the definitions in Eq.(4) and Eqs.(8-9), minimizing $\ell_{conReg}$ leads to the classifiers producing predictions consistent with the consensus result $\bar{p}_C$.

The classifiers attempt to produce the predicted data pair $(x, \hat{y})$ for fooling the discriminator, where $\hat{y}$ denotes the one-hot label determined by $\bar{p}_C(x)$. We need an adversarial training term for optimizing the classifiers, and the final formulation can be expressed as follows:

$$
\min_{C1, C2} \frac{1}{2} \mathbb{E}_{x \sim p_c} [\bar{p}_C(x) \log(1 - D(x, \hat{y})] ] \\
+ \ell_{classify}(\theta_{C1}) + \lambda \ell_{smoReg}(\theta_{C1}) \\
+ \ell_{classify}(\theta_{C2}) + \lambda \ell_{smoReg}(\theta_{C2}) \\
+ \mu \ell_{conReg}(\theta_{C1}, \theta_{C2}),
$$

where $\lambda$ and $\mu$ are the weighting factors for achieving a balance among the terms.

### 3.3. Adversarial Training

Since we follow the adversarial training scheme of the TripleGAN model in general, the discriminator $D$ learns to distinguish the labeled data pair $(x, y)$ from the synthesized data pair $(G(z, \hat{y}))$ and predicted data pair $(x, \hat{y})$. The corresponding optimization formulation is presented as follows:

$$
\max_{D} \mathbb{E}_{(x, y) \sim p_l} [\log D(x, y)] \\
+ \frac{1}{2} \mathbb{E}_{(z, \hat{y}) \sim p_u} [\log(1 - D(G(z, \hat{y}), \hat{y})] ] \\
+ \frac{1}{2} \mathbb{E}_{x \sim p_c} [\log(1 - D(x, \hat{y})] ] .
$$

The discriminator competes with the generator and classifiers in the minimax game. The generator attempts to syn-
the proposed model aligns better than that of the baseline.

4.2. Benchmark datasets

We further compare EnhancedTGAN with state-of-the-art semi-supervised deep learning models on the MNIST, SVHN and CIFAR-10 benchmarks, which are widely used for evaluation of classification and synthesis. According to the common setting, we perform experiments for the cases in which there are 100, 1000 and 4000 randomly selected labeled instances for MNIST, SVHN and CIFAR-10, respectively. The network architecture of the classifiers in EnhancedTGAN is the same as that in the main competing methods, such as TripleGAN and CT-GAN. The classification results are presented in Table 1. The error rates of the competing methods are taken from the existing literature, except TripleGAN*. TripleGAN* is a strong baseline, and outperforms the original TripleGAN. In all the cases, the proposed EnhancedTGAN achieves more accurate classification results than TripleGAN*. For CIFAR-10 with 4000 labels, EnhancedTGAN surpasses TripleGAN* by a large margin, and significantly reduces the test error rate from 14.65% to 9.42%. Compared with other competing methods, the proposed EnhancedTGAN produces more accurate or comparable classification results in all cases. Figure 4 shows the synthesized samples by our EnhancedTGAN model for the three datasets. We also visualize the t-SNE embedding [20] of the features associated with the last hidden layer of the classifier network in TripleGAN* and our EnhancedTGAN model on CIFAR-10 with 4000 labels. As shown in Figure 5, EnhancedTGAN performs better than TripleGAN* in learning the class-conditional data distributions, since we can observe that the samples are strongly clustered, and the distribution of the synthesized data can match the distribution of the unlabeled data very well.
Table 1. Comparison between our model and the competing methods on semi-supervised classification on the benchmark datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>MNIST</th>
<th>SVHN</th>
<th>CIFAR-10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 labels</td>
<td>All labels</td>
<td>1000 labels</td>
</tr>
<tr>
<td>LadderNetwork[30]</td>
<td>1.06±0.37</td>
<td>0.57±0.02</td>
<td>-</td>
</tr>
<tr>
<td>SPCTN[40]</td>
<td>1.00±0.11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>II-model[15]</td>
<td>0.89±0.15</td>
<td>-</td>
<td>4.82±0.17</td>
</tr>
<tr>
<td>Temporal-Ensembling[15]</td>
<td>-</td>
<td>-</td>
<td>4.42±0.16</td>
</tr>
<tr>
<td>Mean-Teacher[36]</td>
<td>-</td>
<td>-</td>
<td>3.95±0.19</td>
</tr>
<tr>
<td>VAT[24]</td>
<td>-</td>
<td>-</td>
<td>3.74±0.09</td>
</tr>
<tr>
<td>VAdv[28]</td>
<td>-</td>
<td>-</td>
<td>4.16±0.08</td>
</tr>
<tr>
<td>VAdv+VAT[28]</td>
<td>-</td>
<td>-</td>
<td>3.55±0.05</td>
</tr>
<tr>
<td>SNTG+II-model[19]</td>
<td>0.66±0.07</td>
<td>-</td>
<td>3.82±0.25</td>
</tr>
<tr>
<td>SNTG+VAT[19]</td>
<td>-</td>
<td>-</td>
<td>3.83±0.22</td>
</tr>
</tbody>
</table>

Table 2. Ablation study of the proposed model on CIFAR-10 with 4000 labels for investigating the influence of synthesized data, and consistency and smoothness regularization in semi-supervised classification.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test error rate (%) with # labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4000 labels</td>
</tr>
<tr>
<td>w/o GAN</td>
<td>11.92±0.19</td>
</tr>
<tr>
<td>w/o $\ell_{conReg}$</td>
<td>11.47±0.13</td>
</tr>
<tr>
<td>w/o $\ell_{smoReg}$</td>
<td>12.03±0.29</td>
</tr>
<tr>
<td>EnhancedTGAN</td>
<td>9.42±0.22</td>
</tr>
</tbody>
</table>

4.3. Ablation study

We remove the feature-semantics matching terms from the overall loss function of the generator on CIFAR-10 with 4000 labels, and show the classification accuracy on the synthesized data during training in Figure 6. We can observe that the class-wise mean feature matching term is able to boost the classification accuracy, which indicates that the synthesized data can better match the statistics of the real data. The semantics matching term can further improve the training of the generator.

To investigate the effectiveness of our proposed improvement strategies in semi-supervised classification, we perform an ablation study to compare the resulting models when removing the corresponding modules on CIFAR-10 with 4000 labels as shown in Table 2. We first remove the generator and the discriminator to evaluate the classifiers with collaborative learning, and the test error rate rises to 11.92%, which indicates that the synthesized instances are useful for improving classifier training. In addition, we remove the consistency and smoothness regularization terms from the overall loss function of the classifiers to investigate the influence of collaborative learning and the local Lipschitz condition, and a significant performance drop can be observed in both cases. We consider that both synthesized data and regularization are important for improving semi-
supervised classification.

4.4. Face synthesis

To further investigate the capability of the proposed EnhancedTGAN model in performing difficult semi-supervised instance synthesis, we conduct an experiment on the FaceScrub dataset. Since the classes in this dataset contain different numbers of human face images, we select the 100 largest classes in our experiment, and only 20 images sampled randomly in each class are labeled. All the images are resized to $64 \times 64$, and thus we slightly modify the network architectures used previously for this experiment without significantly increasing the number of model parameters. The synthesized human face images are shown in Figure 7. For TripleGAN*, we observe that the variation within a class is relatively small, and the structures of the human faces are lost in some images. On the other hand, the synthesized images of our EnhancedTGAN look realistic and preserve human identities. The corresponding classification results of TripleGAN* and EnhancedTGAN are shown in Table 3.

Table 3. Comparison between the baseline model and the proposed model on semi-supervised classification on FaceScrub-100.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test error rate (%) with # labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2000 labels</td>
</tr>
<tr>
<td>TripleGAN*</td>
<td>18.23 ± 0.56</td>
</tr>
<tr>
<td>EnhancedTGAN</td>
<td>16.08 ± 0.24</td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper, we propose an enhanced TripleGAN model for improving both semi-supervised conditional instance synthesis and classification. Toward this end, we adopt feature-semantics matching to force the generator to effectively learn the class-conditional data distributions, such that the synthesized instances with high-fidelity can be used for training better classifiers. On the other hand, we collaboratively train two classifiers, which can provide more accurate guidance for the generator. The experiment results demonstrate that the proposed model outperforms the original TripleGAN and achieves new state-of-the-art results on multiple benchmark datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Project No. 61502173, U1611461, 61722205, 61751205, 61572199), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11300715), in part by City University of Hong Kong (Project No. 7005055), in part by the Natural Science Foundation of Guangdong Province (Project No. 2016A030310422), in part by Key R&D Program of Guangdong Province (Project No. 2018B010107002), and in part by the Fundamental Research Funds for the Central Universities (Project No. 2018ZD33).
References


