Attention Generating Target Images with Labels for Unsupervised Domain Adaptation

Shuai Fu, YiFan Ye, Jing Chen *
School of Physics & Optoelectronic Engineering Guangdong University of Technology, Guangzhou, China

*Corresponding author e-mail: jchen125@gdut.edu.cn

Abstract. Domain adaptation has recently become a popular and successful field of solving the dependence of the deep learning model on the training set. Previous works proposed unsupervised learning methods based on Generative Adversarial Networks (GANs), which solved the mixed target domain labels problem caused by the neglect of class-level matching, by generating target pairs. And these well-designed models have successfully achieved great performance on different datasets. Moreover, due to the recent widespread use of attention mechanism, we intuitively introduce attention mechanism to the previous model. Our new structure is based on GANs containing attention mechanism and well-designed upampling module, which make our model more robust. Demonstrated by extensive experiments, our new model outperforms the previous works on several standard domain adaptation datasets.

Keywords: transfer learning, domain adaptation, generative adversarial networks, image generation.

1. Introduction
Recently, with the help of large-scale labeled datasets, deep learning has achieved impressive and satisfied success in diverse applications, i.e., computer vision, natural language processing. However, the models, training on various datasets, are occasionally hard to generalize to new unlabeled datasets owing to a phenomenon known as domain shift. The domain shift between training set with label and unlabeled target data would seriously decrease the model’s performance. Spontaneously, how to minimize domain shift raises the problem of domain adaptation.

Intuitively, we can use simple and effective fine-tune strategy on new target dataset to address the problem. However, considering the consumption of time, the expensive of manual labels collection and tune considerable parameters of deep network for various target data set, the fine-tune strategy is probably not satisfied and relatively impractical. Due to this dilemma, the development of domain adaptation has promoted significantly in recent years, which aims to propose effective method by applying the rich supervised information of the source data to the target data.

In most cases, source and target domains share the same label space, but there are some differences in data distribution owing between them due to various reasons. Our work continually follows this setting. Meanwhile, based on the previous works, we focus on unsupervised training strategy, which is more useful and challenging.
Fig. 1. Overall chart. In stage 1, we use annotated source domain samples to train $F$. In state 2, we fix classifier $F$ and train GANs in order to improve the quality of generated images and minimize the proposed semantic loss. In stage 3, we train $F$ and GANs simultaneously. On the one hand, source labeled data and generated target data are used to continue training $F$. On the other hand, GANs are still being trained to generate more realistic target images. All the dash-line network blocks indicate that they have trained at the former stage. At the test stage, we just need to test on the classifier $F$.

Some previous methods aimed to solve this task by focusing on a shared space where domain gap is narrowed by source domain with labeled data and unlabeled target data. This idea to reduce domain gap is achieved by seeking a domain-invariant representation. Many algorithms proposed by researchers successfully solved the problem, but at the same time, it still has certain demerit point which is the solution hard to achieve class-level distribution matching [1].

Recent methods proposed to solve the problem mainly around two major technologies: moment matching method and adversarial learning approach. These two methods have made a significant progress recently. Our work based on GANs and introduce attention mechanism into the original GANs model. Considering the fine-tune strategy, we intuitively divided our model in 3 steps. By training former 2 steps and fixed the part of the parameters, directly using them as pre-trained model in the final training step make our model have better performance.

In summary, the major contributions of this paper are as follows:

- We introduce attention mechanism to generative adversarial networks (GANs). We present an upsampling module in GANs with residual block embedded the attention mechanism.
- Our training process is divided into three main phases in order to achieve better performance in the target domain. It’s worth noting that although we adopt a phased training strategy, our model can still be considered as an end-to-end model.
- Our model demonstrates competitive performance in several standard domain adaptation datasets.

2. Related Work

With the development of labeled large-scale dataset and in-depth study of convolutional neural networks, computer vision field have achieved great success in recent years. However, considerable researches have shown that many current deep networks greatly depend on the annotated data. Unsupervised domain adaptation focuses on two similar distributions of data, one of which is annotated source domain data and the other is unlabeled target domain data. Spontaneously, domain adaptation problems have attracted great attention. Considering the time-consuming manual labeling process, unsupervised domain adaptation is an effective solution. This section is going to discuss recent related papers and works, ranging from unsupervised domain adaptation, adversarial learning, to recent works on attention mechanism.

2.1. Unsupervised Domain Adaptation

Domain adaptation is a machine learning approach in order to make the model more robust on different data sets rather than only the training set. Unsupervised domain adaptation (UDA) setting on a challenging situation that the target domain samples are totally unlabeled. This makes training more difficult than fully-supervised training method. Previous works showed that UDA has been widely used in various fields, e.g., image classification and semantic segmentation. Some of the methods also focusing on measurement of shift between different domains. Those statistic-based distance metrics methods [2, 3] minimize domain distribution discrepancy to exploit transferable domain features.
Our works also utilize diverse loss function inspired by previous works [loss] which they focusing on different levels of convolutional neural networks and various dimensions across target domain and source domain. These methods significantly boost improvement of domain adaptation.

Very recently, self-supervised learning method has been proposed and gained great attention. Accordingly, it shows benefit the adversarial learning model significantly when jointly trained model for semantic segmentation.

With the inspiration of the fully-supervised and self-supervised methods, we finally train our model with some accurate labels which are our model generated. Lastly, in order to fully utilize the label information that our model produce, our training strategy dramatically become semi-supervised.

2.2. Adversarial Learning

Generative Adversarial Networks (GANs) are a type of widely used generative model, whose adversarial idea has exerted great influence on other fields of machine learning. In general, a GAN consists of a generative model, named generator G and a discriminative model, named discriminator D. G is used to generate verisimilar images, while D is used to distinguish whether real images or synthetic images. To be precise, the training process of GANs is a game between G and D and the final game result is that the images generated by G can deceive D. Formally, generator and discriminator models can be trained via the following minimax game:

$$\min_G \max_D V(D,G) = E_{z \sim P(z)}[\log D(z)] + E_{x \sim P(x)}[\log(1 - D(G(x)))].$$

Inspired by the adversarial ideas of GANs, considerable adversarial domain adaptation approaches have been studied and developed, where the feature extractor can learn the domain-invariant and category-discriminative feature representation by the game between the feature extractor and the domain discriminator. Further on, this method can align the features via reverse gradient backpropagation [4] during adversarial learning.

Moreover, adversarial learning can accomplish pixel-level transformation between source domain and target domain, which can also successfully minimize the discrepancy between them. Bousmalis et al. [5] propose a resolution based on adversarial learning while it subjected to assumption that source data and target data have similar annotation. Li et al. proposed a pixel-level GAN-based adaptation method to generate target samples with given labels [1]. The CyCADA [6] is a remarkable bi-directional circular GAN-based unsupervised domain adaptation model.

Our model is also a GAN-based domain adaptation method, similar to PixelDA [5], CyCADA [6] et al. A phased training strategy for GAN-based domain adaptation is proposed by us due to the instability in the adversarial learning.

2.3. Attention Mechanism

Attention mechanism recently has widely been applied to the deep learning technology. VIT [7] show that Transformer attention structure perfectly capable to accomplish complex computer vision task and even outperform than very Deep Convolutional Neural Networks (DCNN).

Previous works also demonstrated that attention layers are fully-capable in any tasks using convolutional neural networks. This mean that attention mechanism can simply replace any DCNN by well-designed methods with even faster training speed and less parameters.

In this paper, we introduced attention mechanism to different frameworks making our model more robust and getting a more satisfied performance.
3. Proposed Method

In this section, we propose a approach for unsupervised domain adaptation. We demonstrate our overall model and show specific stages with some details of loss functions in our training process.

3.1. Framework

Considering the great performance attention mechanism achieved, we intuitively introduce it to previous generative adversarial networks for unsupervised domain adaptation task. As the Fig. 1 shows that, at first stage, we take sourced domains labeled picture and labels which we denote Cs as input. While in our training process we used MNIST data set as our sourced data. First, we train a classifier F on source data which we use Cross Entropy Loss that is denoted as follows:

$$E(x) = -\frac{1}{n} \sum_x [y \ln \alpha + (1 - y) \ln(1 - \alpha)],$$  \hspace{1cm} (1)

Where E indicate Cross Entropy Loss function, n represent whole amounts of samples and $\alpha = \text{sigmoid}(\omega x + \text{bias})$. We successfully trained a high-performance feature extractor on the specific dataset.

Then, in stage 2, we propose a GANs which largely based on previous generative adversarial networks, but we add attention mechanism to the generator as the Fig. 2 shows. We feed generator with (C, Z) where C denote as labels and Z represent noise. Generator G will generate a series of images which we marked as Xg according to the labels. Then discriminator D take Xt (target domains data) and Xg as input and update by loss function lGAN.
\[ l_{\text{GAN}}(D, G; D_T, D_s) = E_{x \sim D_s} [\log D(x)] + E_{c \sim c} [\log(1 - D(G(c, z)))], \] (2)

Where \( c \) is a random one-hot categorical code, and \( z \) is a noise vector.

Simultaneously, \( X_g \) will feed into our trained classifier \( F \) updated by a loss function \( l_{\text{sem}} \) which calculate the discrepancy between ground-truth \( C \) and output label generated by classifier \( F \). To be clarify, classifier \( F \) parameters at this stage are fixed and do not update by \( l_{\text{sem}} \).

### Table 1. Classification Accuracy (%) on Five Settings. The Best Performances Are Indicated With Bold

| Method       | MNIST2USPS | USPS2MNIST | MNIST2MNIST-M | SVHN2MNIST | MNIST2SVHN |
|--------------|------------|------------|----------------|------------|------------|
| MMD [2]      | 81.1       | -          | 76.9           | 71.1       | -          |
| DANN [5]     | 85.1       | 73.0       | 77.4           | 73.9       | 35.7       |
| DRCN [9]     | 91.8       | 73.7       | -              | 82.0       | 40.1       |
| CoGAN [10]   | 91.2       | 89.1       | -              | -          | -          |
| UNIT [11]    | 95.9       | 93.5       | -              | 90.5       | -          |
| DuplexGAN [12]| 96.1       | 98.7       | -              | 92.4       | 62.6       |
| CyCADA [6]   | 95.6       | 96.5       | -              | 90.4       | -          |
| SourceOnly   | 92.4       | 86.1       | 54.2           | 76.4       | 57.3       |
| Label-to-Image-DA (Ours) | **96.3** | **97.1** | **96.5** | **95.4** | **63.7** |

\[ l_{\text{sem}}(G) = E_{c \sim c} E[F(G(c, z)), c]. \] (3)

Stage 3 structure is based on the combination of former two stages. We stack former structure together and add source domain data as classifier \( F \) input to generate new semantic loss and source domain label to generate original source loss. In stage 3 we simultaneously train our generator classifier and discriminator. Even though our main training steps are divided into 3 sections, our whole strategy can still consider as an end-to-end training strategy.

### 4. Experiments

In this section, we take considerable experiments to verify our method on standard domain adaptation tasks, including five experiments: MNIST2USPS, USPS2MNIST, MNIST2MNIST-M, MNIST2SVHN and SVHN2MNIST. In the following experiments, standard unsupervised protocol, where annotated source samples and unlabeled target samples can be used, is applied to all domain adaptation experiments.

#### 4.1. Experiments setting

We carry out our experiments on the PyTorch platform. All digit dataset contains digit images belonging to 10 classes (0–9). MNIST and USPS datasets are two types hand-written digital grayscale images set. Both SVHN and MNIST-M datasets are color digit images with complex backgrounds and different styles. These handwritten digital image datasets exist certain differences, so they are often used to evaluate the performance of the domain adaptation models.

#### 4.2. The Details of Experiment Implementation

In order to facilitate experiments, all the digit images are resized to \( 32 \times 32 \times 3 \). All the network structures we proposed are convolutional neural networks (CNNs). Adam solver is adopted to \( F \), \( D \) and \( G \) with the learning rates 0.001, 0.0002 and 0.0004, respectively.
The final output layer of the generator and the discriminator use the Tanh activation function and the Leaky ReLU activation function, respectively. Specifically, we present an UpResBlock in the generator using the ReLU activation function. We design a similar convolutional neural networks structure in [8] to transfer source knowledges. We present a upsampling generative module, named UpResBlock, in the generator G. In addition, we use a SN layer after a convolutional layer in the discriminator D, where the discriminator is optimized by adding regularization. Fig. 4 shows our proposed structure of discriminator and generator.

4.3. Results and Analysis
As shown in Fig. 3, our model is able to generate verisimilar digital images 0–9 in a specified order, thanks to our carefully designed semantic loss. It has been proved that our model can generate target samples with given labels by transferring the knowledges of the annotated source samples, finally resulting in the improvement of the classifier for the target domain.

The result of image classification of our model in the unsupervised domain adaption is reported in TABLE I. Each row represents the performance of a domain adaption method and each column represents a kind of domain adaptation task. The dashed line indicates that there is no result in this task for that method. Particularly, the ‘Source only’ row represents testing the target domain samples directly with the classifier after training the classifier only with the labeled source domain data. Our model performs very well in all five domain adaptation tasks, achieving the optimal performance in four of them and suboptimal performance in the USPS2MNIST task. Especially, our model still achieves 63.7% accuracy in the MNIST2SVHN task, which is considered as one of the most challenging domain adaptation tasks.

In summary, our model achieves the remarkable performance of images classification in the target domain by generating verisimilar target images with designated labels directly.

5. Conclusion
Considering the accuracy descend caused by the distribution discrepancy of two domains, in this paper, we propose an unsupervised attention based adversarial model which aim to generate verisimilar target images with given labels directly. We first concentrate on training an accurate classifier F on source domain which achieve satisfied performance. Secondly, we train a generator G and discriminator D simultaneously updating by our well-designed semantic loss function and GAN function. We introduce attention mechanism to generator G improving the performance significantly. Lastly, we jointly train and update all models while generator G can generate accurate label-image pairs after a great amount of iteration so we can consider it generated labels as ground-truth labels that can be used in training. This is one of the merits of our training process. We fully utilize generated information as another loss function in final training process. To be clear, even though our training process manually divided into three steps, our model can still be considered as an end-to-end model.

Verified on several datasets, our model demonstrates a competitive performance and experimentally proved robust. However, it still has several deficiencies when two domains have a large domain discrepancy. Generator is unable to generate high-resolution images even if we trained our model on
several large-scale datasets. Our future work will focus on large domain discrepancy problem. We aim to build a more robust model and generate more realistic images with higher performance on various datasets.

References
[1] R. Li, W. Cao, S. Wu and H. -S. Wong, "Generating Target Image-Label Pairs for Unsupervised Domain Adaptation," in IEEE Transactions on Image Processing, vol. 29, 2020, pp. 7997-8011.
[2] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," presented at the Proceedings of the 34th International Conference on Machine Learning(ICML) - Volume 70, Sydney, NSW, Australia, 2017.
[3] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in 2012 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2012, pp. 2066-2073.
[4] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," presented at the Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML) - Volume 37, Lille, France, 2015.
[5] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 95-104.
[6] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, CyCADA: Cycle-consistent adversarial domain adaptation, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, vol. 80, PMLR, Stockholmsmässan, Stockholm Sweden, 2018, pp. 1989–1998.
[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. in press.
[8] R. Shu, H. H. Bui, H. Narui, and S. Ermon, “A DIRT-T approach to unsupervised domain adaptation,” in Proceedings of International Conference on Learning Representations (ICLR), 2018.
[9] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation," Cham, 2016: Springer International Publishing, pp. 597-613.
[10] M. Liu, O. Tuzel, Coupled generative adversarial networks, in: Annual Conference on Neural Information Processing Systems (NIPS), Barcelona, 2016, pp. 469–477.
[11] M. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, 2017, pp. 700–708.
[12] L. L. Hu, M. Kan, S. Shan, and X. Chen, "Duplex Generative Adversarial Network for Unsupervised Domain Adaptation," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1498-1507