Reducing Domain Gap via Style-Agnostic Networks

Hyeonseob Nam HyunJae Lee Jongchan Park Wonjun Yoon Donggeun Yoo
Lunit Inc.
{hsnam,hjlee,jcpark,wonjun,dgyoo}@lunit.io

Abstract

*Deep learning models often fail to maintain their performance on new test domains. This problem has been regarded as a critical limitation of deep learning for real-world applications. The root cause of the vulnerability to domain changes is that the model tends to be biased to image styles (*i.e. *textures). To tackle this problem, we, the team Lunit, propose Style-Agnostic Networks (SagNets) to force the model to focus more on image contents (*i.e. *shapes) shared across domains but ignore image styles. SagNets consist of three novel components: style adversarial learning, style blending and normalization consistency learning, each of which prevents the model from making decisions based upon style information. In collaboration with a few additional training techniques and an ensemble of several model variants, the proposed method won the 1st place in the semi-supervised domain adaptation task of the Visual Domain Adaptation 2019 (VisDA-2019) Challenge*[1].

1. Introduction

Despite the success of deep neural networks learned with large-scale labeled data, their performance often drops significantly when they confront data from a new test domain, which is known as the problem of *domain shift* [22]. For successful deployment to ever-changing real-world scenarios, it has become crucial to make the model robust against to the domain shift.

A recent line of studies has explored the relationship between a model's robustness and the style (*i.e.* texture) of an image [9, 20, 15, 21, 2]. Geirhos et al. [9] showed that standard CNNs for image classification are biased towards styles. When the model is learned to concentrate on image contents (*i.e.* shapes), the model becomes more robust under variable image distortions. Furthermore, [20, 15] demonstrated that adjusting the proportion of style information on features helps overcome domain differences.

[1] http://ai.bu.edu/visda-2019/

From the previous studies, we assume that the style information easily changes between domains compared with the content information. Inspired by this, we propose *Style-Agnostic Networks (SagNets)* to prevent the model from making decisions based on styles and make it focus more on the contents. Our approach is applicable to all the problems suffered by the heterogeneous domains, such as domain adaptation (DA) [22] and domain generalization (DG) [16]. In this manuscript, however, we focus on the task of semi-supervised domain adaptation (SSDA) [25] in which a few target labels are available in addition to large-scale labeled source data and unlabeled target data.

We shall start by describing the baseline architecture in Section 2 and introduce our SagNets in Section 3. We also provide details of additional training techniques and model variants for ensemble in Section 4 and 5, respectively. Then we present the results of semi-supervised domain adaptation in Section 6 and conclude the manuscript in Section 7.

2. Baseline Architecture

We use ResNet-152 [10] pretrained on ImageNet [24] as our backbone CNN architecture. To tackle the distribution gap between different domains and utilize unlabeled target data, we integrate two additional components into our base domain adaptation framework: Minimax Entropy (MME) [25] and Adaptive Batch Normalization (AdaBN) [17].

2.1. Minimax Entropy (MME)

We adopt MME [25] as our baseline adaptation method, which alternatingly moves class prototypes toward the target data distribution by maximizing the entropy of predictions and updates features to be better clustered around the estimated prototypes by minimizing it. It also includes a similarity-based framework inspired by [5], where the classification is made upon the similarity between a normalized feature vector and class prototypes. This framework is effective for harnessing few-shot labeled examples provided in the semi-supervised domain adaptation setting.

2.2. Adaptive Batch Normalization (AdaBN)

Batch Normalization (BN) [13] is a common module used in current deep networks, which normalizes feature responses to stabilize and accelerate training. When a model is learned with multiple domains, it has been reported that using separate BN modules for individual domains helps to align their feature distributions [17, 3, 4]. We call this method AdaBN, following [17]. As AdaBN make each domain centered at zero, it is effective for reducing the style gap between domains. Hence we opt for AdaBN as our base normalization module.

3. Style-Agnostic Networks

Motivated by [9], we aim to make the model's decision less depend on the style of an image to improve its robustness across different domains. It is well known in style transfer literature [8, 12] that the feature statistics (*e.g.* mean and variance) of a CNN effectively capture the style of an image. Based on this observation, we propose three novel techniques to remove style bias from a CNN: style adversarial learning, style blending, and normalization consistency training.

3.1. Style Adversarial Learning

We employ an adversarial learning framework to obtain style-invariant feature representation. Note that a well-known approach for reducing the domain gap is to impose an adversarial loss on predicting domain labels [7]. However, this approach does not fit our goal as we do not have explicit labels of styles. Instead, we constraint the style features (*i.e.* the mean and variance of convolutional features) to be not capable of discriminating the object class labels by introducing a novel adversarial loss. It can be also viewed as defending against adversarial attacks on styles, in order to make the network more robust to arbitrary style changes.

We apply the style adversarial loss only to lower layers, since the features statistics of higher layers cannot be free from object class information. To this end, given a ResNet comprising four stages, we extract features $\mathbf{X} \in \mathbb{R}^{H\times W\times d}$ from the second stage and take its channel-wise mean and variance $\mu, \sigma \in \mathbb{R}^d$ as style features. Then we construct a style discriminator $\mathbf{D}$ (with a form of $2d-1024-1024-C$) which takes the style features $[\mu, \sigma] \in \mathbb{R}^{2d}$ as input and learns to predict the class probability $\mathbf{p} \in \mathbb{R}^C$. The style feature extractor $\mathbf{G}$, composed of the ResNet layers up to the second stage, are trained to pool the discriminator by inserting a gradient reversal layer [7] between $\mathbf{G}$ and $\mathbf{D}$. Consequently, the low-level features of the network (*i.e.* $\mathbf{G}$) are encouraged to exploit contents rather than styles for the final prediction of class labels.

This approach can be applied to both labeled and unlabeled examples in domain adaptation. For labeled examples either on the source or target domain, $\mathbf{D}$ and $\mathbf{G}$ are trained by minimizing and maximizing the cross-entropy loss, respectively. In case of unlabeled examples on the target domain, $\mathbf{D}$ is not trained but $\mathbf{G}$ can be still trained by maximizing the entropy of the prediction from $\mathbf{D}$, which decreases the confidence of style-based decisions.

3.2. Style Blending

To further make the model agnostic to style, we introduce a novel style blending which randomizes style information. Style Blending is performed on feature space by interpolating style information between different samples. Given a random pair of samples (i, j) within a mini-batch, it changes the feature map $\mathbf{X}_i$ of sample i to

$$\mathbf{X}_i := \hat{\sigma}_i \cdot \left(\frac{\mathbf{X}_i - \mu_i}{\sigma_i}\right) + \hat{\mu}_i, \quad \text{s.t.} \quad \begin{aligned} \hat{\mu}_i &= \alpha \cdot \mu_i + (1-\alpha)\cdot \mu_j, \\ \hat{\sigma}_i &= \alpha \cdot \sigma_i + (1-\alpha)\cdot \sigma_j, \end{aligned} \tag{1}$$

where $\alpha \sim \text{Uniform}(0, 1)$. By randomizing style features, the network is encouraged to neglect the style information when predicting object class labels. The style blending module is inserted right after the first convolution layer and the first stage to train the network to be more agnostic to bottom-level style information which are rarely relevant to object categories.

3.3. Normalization Consistency

One of the effective approaches for leveraging unlabeled data for semi-supervised learning is consistency/smoothness enforcing [19, 27], which let the model prediction invariant to small perturbations on data. We apply this approach to our semi-supervised domain adaptation problem in which we introduce a similar consistency loss to make the model prediction invariant to style variations. Instead of directly perturbing the style on image pixels, we perturb the style on intermediate features by simply applying different normalization schemes. For each training example, we obtain two prediction vectors from the network: one normalized with mini-batch statistics by the AdaBN modules but the other with moving-averaged statistics. The consistency between the two final predictions are estimated by KL divergence and minimized for all unlabeled data. Following [27], training signal annealing with log schedule and confidence-based masking with a threshold 0.5 are applied to prevent overfitting.

4. Learning Details

In a semi-supervised domain adaptation problem where only a few labeled examples are available for training, the use of a large model such as ResNet-152 could lead to

the memorization of certain examples. Thus, we adopt synthetic data augmentation and Mixup [29] to generalize the model to unseen images. To fully leverage the unlabeled data, we introduce a simple semi-supervised learning method which repeats the fine-tuning with pseudo labels of the unlabeled data.

4.1. Synthetic Data Augmentation

CyCADA [11] showed that domain adaptation methods performed on feature-level sometimes fail to capture low-level domain disparity. In order to tackle this issue, we have trained CycleGAN [30] to translate the source domain to the style of the target domain in pixel-level and vice versa. We remove the identity loss from the original loss term in CycleGAN, as one domain is quite far from the other. Unlike CyCADA which solely uses target-stylized source images and target images, we have additionally utilized source images and source-stylized target images. With these additional data which imitate the different style from the original images, we can avoid over-fitting and improve generalization to the domain bias.

4.2. Inter- and Intra-Domain Mixup

Mixup [29] is a simple data augmentation method and has shown improvements in classification tasks, noisy label problems, and GAN tasks. The images and labels of two samples are blended to create a mixed image and a mixed soft label. A naive introduction of Mixup is the intra-domain mixup, where only the images from the same domain are mixed. An extended version is the inter-domain mixup, where we mix images regardless of the domains. There are four different domains for Mixup to choose: *source*, *target*, *source to target* (CyCADA), *target to source* (CyCADA). During Mixup training, we only use the mixed samples for supervision. For the pseudo label training, we also add the pseudo-labeled samples to Mixup training.

4.3. Iterative Pseudo Labeling

[28, 1, 14] have demonstrated that a simple pseudo labeling method is effective for domain adaptation. However, MSTN [28] and clustering-based pseudo labels [1] do not show clear improvements in our experiments, possibly due to the high complexity of the given task. Instead, a simple pseudo-label method [14] with a labeling threshold has yielded significant improvements. Given a learned model, we assign a pseudo label to an unlabeled sample if the prediction score is higher than 0.6. This procedure is repeated several times until the final loss converges.

5. Model Variants

For extra performance improvement, we train multiple models and ensemble their results. While keeping the backbone network as ResNet-152, we construct two variants equipped with Batch-Instance Normalization (BIN) [20] and Style-based Recalibration Module (SRM) [15], respectively, both of which have been shown to be robust against style variations.

Table 1: Results on VisDA-2019 SSDA where the source is *real* and the target is *sketch* (*i.e.* the validation phase).

Method	Accuracy (%)
Baseline (Sec.2)	46.56
SagNet (Sec.3)	55.70
SagNet+synthetic (Sec.4.1 and 4.2)	60.73
SagNet+synthetic+pseudo (Sec.4.3)	62.51
SagNet+synthetic+pseudo+ensemble (Sec.5)	63.08

5.1. Batch-Instance Normalization (BIN)

BIN [20] is a normalization method which combines the benefits from BN [13] and Instance Normalization (IN) [26]. Based on the property that IN removes the style of each image while BN maintains it, BIN learns to selectively remove unnecessary styles but keep important styles, which alleviates style variation under domain shifts.

5.2. Style-based Recalibration Module (SRM)

We also utilize SRM [15] which is an architectural unit that adaptively recalibrates intermediate feature maps by exploiting their style information. It estimates per-channel recalibration weight from style feature then performs a channel-wise recalibration. By explicitly incorporating the styles into CNN representation, SRM can alleviate the inherent style disparity between source and target domain.

6. Experiments

We demonstrate the effectiveness of SagNets and training techniques for semi-supervised domain adaptation task with DomainNet [23] dataset. It consists of 345 categories with 0.6 million images of six distinct domains. For data augmentation, the input images are randomly cropped to 224x224 patches then random horizontal flipping and AutoAugment [6] of policy learned on ImageNet are applied. The networks are trained by SGD with an initial learning rate of 0.002, a momentum of 0.9, and a weight decay of 0.0001. We train the networks for 30,000 iterations with a cosine annealing schedule [18].

Table 1 demonstrates the results on the semi-supervised domain adaptation task where *real* domain is used as a source and *sketch* domain is used as a target. SagNet significantly boosts the performance of the baseline architecture with minimal computation overhead. Furthermore, our training techniques and ensemble of model variants also bring meaningful performance improvement.

7. Conclusion

We have presented Style-Agnostic Networks (SagNets) that is robust against style variation caused by domain shifts. SagNets is learned to concentrate more on contents rather than styles in decision making process. We have also introduced a few additional training techniques and model variants for further performance improvement. Our experiments have demonstrated the effectiveness of SagNets in reducing the disparity between domains with DomainNet dataset. The principle how we let the network concentrate on image contents could be applied to other problems such as improving robustness of neural network to corruptions and perturbations.

References

[1] Visda 2018 challenge openset classification winner presentation. https://ai.bu.edu/visda-2018/assets/attachments/Visda18-classification-JD-AI.pdf. Accessed: 2019-10-01. 3

[2] Francis Brochu. Increasing shape bias in imagenet-trained networks using transfer learning and domain-adversarial methods. *arXiv preprint*, 2019. 1

[3] Fabio Maria Cariucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulò. Autodial: Automatic domain alignment layers. In *ICCV*, 2017. 2

[4] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, 2019. 2

[5] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019. 1

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. In *CVPR*, 2018. 3

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 2

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. 1, 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *ICML*, 2018. 3

[12] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 2, 3

[14] D. H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013. 3

[15] HyunJae Lee, Hyo-Eun Kim, and Hyeonseob Nam. Srm: A style-based recalibration module for convolutional neural networks. In *ICCV*, 2019. 1, 3

[16] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 1

[17] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint*, 2016. 1, 2

[18] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2016. 3

[19] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 2018. 2

[20] Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. In *NeurIPS*, 2018. 1, 3

[21] A Emin Orhan and Brenden M Lake. Improving the robustness of imagenet classifiers using elements of human visual cognition. *arXiv preprint*, 2019. 1

[22] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 2009. 1

[23] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. *arXiv preprint*, 2018. 3

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[25] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, 2019. 1

[26] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint*, 2016. 3

[27] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint*, 2019. 2

[28] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *ICML*, 2018. 3

[29] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 3

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3