SoFA: Source-data-free Feature Alignment for Unsupervised Domain Adaptation Supplementary Material

Hao-Wei Yeh[1], Baoyao Yang[3], Pong C. Yuen[3], Tatsuya Harada[1,2]
[1]The University of Tokyo [2]RIKEN [3]Hong Kong Baptist University
yeh@mi.t.u-tokyo.ac.jp, {byyang, pcyuen}@comp.hkbu.edu.hk, harada@mi.t.u-tokyo.ac.jp

Table 1. Accuracies (%) of VisDA-C Object Recognition. The sample-wise accuracy is the accuracy averaged over all target samples. The per-class accuracy is computed by taking the average over the 12 accuracies computed within each class.

Method	Sample-wise	Per-class
Source Only	51.96	45.03
sMDA [1]	52.69	44.95
RWA [5]	55.70	49.46
SHOT [3]	29.12	27.28
SHOT-IM [3]	60.59	**60.73**
SoFA (Ours)	64.54	60.44
SoFA student (Ours)	**64.59**	60.48

1. VisDA-C Object Recognition

To show the proposed method is applicable to large-scale dataset, we conduct experiments on the VisDA-C dataset [4], a 12-class object recognition dataset that consists of 152k synthetic source images and 55k real-world target images.

We also consider linear classification in this experiment. As the features taken directly from the ImageNet-pre-trained model cannot adapt well, we first fine-tune the ImageNet-pre-trained ResNet101 [2] on source data, then we use the fine-tuned features before the final linear classifier as input and the fine-tuned model as the source model.

For LA-VAE, a fully-connected layer with dropout is added on top of the latent features as the classifier. We set the dimension of latent features z as 256. The decoder consists of 1 layer of the "fully-connected + batch normalization + Leaky ReLU (alpha=0.2)" module and a final fully-connected layer to reconstruct the input features. The number of channels for the fully-connected layers are set to 2048. In order to make the classes more discriminative to each other, we bound the logarithm of variances of the Gaussian mixtures between ± 1. The overall pipeline is trained for 1500 epochs until convergence, with batch size of 256 and ADAM optimizer with learning rate of 1e-4. The "kl annealing"-like scheduling is also applied in this experiment, in which the weight for the alignment term is set as zero in the first 500 epochs, and gradually ramps up from 0 to 1 over the subsequent 500 epochs.

The results in accuracy are summarized in Table 1. The proposed method achieves higher or comparable accuracy to the existing methods in terms of the per-class accuracy, and outperforms the existing methods in terms of the sample-wise accuracy. Note that SHOT and SHOT-IM need to access the source model parameters during adaptation, while the proposed method only needs source-model predictions. This makes the proposed method more suitable for privacy-protected applications, where not the parameters but only the predictions from the source model are accessible.

The results indicate that the proposed method is also applicable to large-scale datasets.

References

[1] Boris Chidlovskii, Stéphane Clinchant, and Gabriela Csurka. Domain adaptation in the absence of source domain data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 451–460. ACM, 2016.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. *arXiv preprint arXiv:2002.08546*, 2020.

[4] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[5] Twan van Laarhoven and Elena Marchiori. Unsupervised domain adaptation with random walks on target labelings. *arXiv preprint arXiv:1706.05335*, 2017.