SALAD : Source-free Active Label-Agnostic Domain Adaptation for Classification, Segmentation and Detection

Divya Kothandaraman[1], Sumit Shekhar[3], Abhilasha Sancheti[1,3],
Manoj Ghuhan[2], Tripti Shukla[3], Dinesh Manocha[1]
University of Maryland College Park[1], Carnegie Mellon University[2],
Adobe Research[3]

A.1. Datasets

In this section, we describe the datasets used in our experiments.

0.1. Classification

- MNIST [4]: MNIST is a handwritten digits dataset, with $60,000$ samples for training and $10,000$ samples for testing. It can be downloaded at http://yann.lecun.com/exdb/mnist/.

- SVHN [10]: SVHN is a house street numbers dataset, and has cropped digits with character wise ground-truth in MNIST format. It has over $600,000$ images and is a much more realistic dataset than MNIST. It can be downloaded at http://ufldl.stanford.edu/housenumbers/.

- VISDA-17 [8]: VISDA is dataset designed for synthetic to real adaptation. The synthetic images are 2D renderings of 3D models generated from various angles and lighting conditions. The real images correspond to natural scene objects. It can be downloaded at http://ai.bu.edu/visda-2017/.

0.2. Segmentation

- GTA5 [9]: GTA5 is a synthetic driving dataset extracted from the computer game Grand Theft Auto. It has 25000 high resolution images. The dataset is available at https://download.visinf.tu-darmstadt.de/data/from_games/.It has 19 classes compatible with CityScapes.

- CityScapes [2]: CityScapes is a real driving dataset collected in Europe. It has 2975 high resolution images for training, and 500 images for testing. The dataset is available for download at https://www.cityscapes-dataset.com/. It has 19 classes.

Table 1: **Synthetic to Real Classification on VISDA:** Our source-free method is on par with state-of-the-art methods that use abundant annotated source data (more than 100k samples).

Method	Source Data	B=10%	B=20%
Random	✓	82.1	87.2
UCN [6]	✓	85.4	90.3
QBC [7]	✓	84.1	89.6
Cluster [3]	✓	83.5	89.6
AADA [11]	✓	84.6	89.7
ADMA [5]	✓	84.8	90.0
SALAD	✗	84.8	89.3

0.3. Document Layout Detection

- PubLayNet [13]: PubLayNet is a large-scale medical documents dataset consisting of images of pages extracted from scientific medical papers. Medical documents are written in a two-column format, with uniform text, figures and tables. It has $360,000$ images and 5 classes. The dataset can be downloaded at https://github.com/ibm-aur-nlp/PubLayNet.

- DSSE [12]: DSSE contains images of pages extracted from magazines, receipts and posters. DSSE is a small dataset with just 150 documents, and has 6 classes, paving way for open-set adaptation from DSSE. The dataset can be downloaded at http://personal.psu.edu/xuy111/projects/cvpr2017_doc.html.

A.2. Synthetic to Real VISDA17 Classification

We conduct experiments on the popular VISDA17 dataset for synthetic to real adaptation. For effective transfer, and to address uncertainty of the target network while sampling, we set $\lambda_G = \lambda_E = 1$ from the second round of sampling. VISDA is a huge dataset with a large variety of samples. Hence, we factor in diversity. The output feature F_T for target samples for clustered using k-means [1]. The mean distance of each target sample from the previously annotated target points gives the diversity score A_D. We set the hyperparameter for diversity score $\lambda_K = 1$ from the second round of sampling. On budgets of 10%, and 20%

of the total target samples, we achieve accuracies of 84.8% and 89.3% respectively. Though SALAD does not use any annotated source data, it achieves accuracies on par with prior work using abundant annotated source data (more than $100k$ samples).

A.3. Implementation details

Hyperparameters: The transformation network τ is a four layer convolutional neural network with kernel size 3, and dilation and padding set to 1. The weight hyperparameter for L_{Tr} is set to 0.1. Our classification models are trained using 1 GPU with 16GB memory. Our document layout detection and segmentation models are trained using 8 GPUs with 16 GB memory each. We use the Stochastic Gradient Descent optimizer for training all our models. All our codes are written using the PyTorch framework. For CityScapes, all images are downsampled by a factor of 2 using bilinear downsampling. Ground truth maps are downsampled by nearest neighbour downsampling. We retain the input image size for our classification experiments. For document layout detection, we resize the images (and appropriately scale the bounding box coordinates) such that the length of the largest size does not exceed 500. We set number of iterations $iter$ to be equal to 3 for MNIST, 1000 for SVHN, 50 for CityScapes.

Codes: In the interest of reproducibility, we release the codes for GATN, including the code for the transformation network, the guided attention modules, and their incorporation within DeepLabv2 for segmentation. We release the train and eval scripts for segmentation as well. We also provide the scripts for H_{AL} with the supplementary zip file. We will make these scripts publicly available upon acceptance of the paper.

We also provide the links to public repositories that we used in our experiments for running SALAD experiments.

- Classification
  - Backbone network: https://github.com/tim-learn/SHOT/blob/master/object/network.py Please follow the procedure in the deeplab_multi.py script in the supplementary zip file to incorporate GATN within the classification backbone.
  - MNIST and SVHN dataloader: https://github.com/tim-learn/SHOT/tree/master/digit/data_load
  - VISDA dataloader: https://github.com/VisionLearningGroup/taskcv-2017-public
  - Training and eval scripts: Please modify the train and eval script in the supplementary zip file to modify code for classification.
- Document Layout Detection
  - Backbone network, train and test scripts: https://github.com/yhenon/pytorch-retinanet
  - PubLayNet dataloader: https://github.com/phamquiluan/PubLayNet/blob/master/training_code/datasets/publaynet.py
  - DSSE dataloader: Use the PubLayNet dataloader to modify.
- Semantic Segmentation
  - Backbone network, train and test scripts: Please check the supplementary zip file.
  - GTA5 and CityScapes dataloaders: https://github.com/wasidennis/AdaptSegNet/tree/master/dataset

References

[1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[3] Ido Dagan and Sean P Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier, 1995.

[4] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[5] Sheng-Jun Huang, Jia-Wei Zhao, and Zhao-Yang Liu. Cost-effective training of deep cnns with active model adaptation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1580–1588, 2018.

[6] Ajay J Joshi, Fatih Porikli, and Nikolaos P Papanikolopoulos. Scalable active learning for multiclass image classification. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2259–2273, 2012.

[7] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004.

[8] 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.

[9] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[10] Pierre Sermanet, Soumith Chintala, and Yann LeCun. Convolutional neural networks applied to house numbers digit classification. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3288–3291. IEEE, 2012.

[11] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 739–748, 2020.

[12] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324, 2017.

[13] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019.