Supplementary Material: Adaptive Consistency Regularization for Semi-Supervised Transfer Learning

Abulikemu Abuduweili[1,2]*, Xingjian Li[1,3]*, Humphrey Shi[2], Cheng-Zhong Xu[3], Dejing Dou[1]

[1]Big Data Lab, Baidu Research, [2]SHI Lab, University of Oregon,
[3]State Key Lab of IOTSC, Department of Computer Science, University of Macau

Appendix A. Additional Information of Proposed Method

In this paper, we propose two regularization methods: Adaptive Knowledge Consistency (AKC) between the source and target model and Adaptive Representation Consistency (ARC) between labeled and unlabeled examples.

A.1. Adaptive Knowledge Consistency

The AKC regularization can be incorporated with supervised or unsupervised transfer learning methods. As shown in Figure 2, we constrain the weighted sample-level consistency (Kullback–Leibler divergence or mean square error) of feature-representation between the pre-trained source feature extractor and the target feature extractor using both the labeled and unlabeled samples. The weight of each sample was determined by the entropy of the pre-trained source model's prediction.

A.2. Adaptive Representation Consistency

The ARC regularization can be used to transfer or learning from scratch semi-supervised methods. As shown in Figure 3, we constrain Maximum Mean Discrepancies between representations' distribution of selected labeled and selected unlabeled samples. Only confident (labeled and unlabeled) samples with high confidence scores will be selected to regularize the distribution of (labeled and unlabeled) data representation. A high confident sample means that the input sample is more likely to fall into the target model's trust region with low entropy of the prediction. To maintain a sufficient number of samples used in ARC regularization, we impose a replay buffer to save recent selected confident samples.

A.3. Intuitive Explanation of ARC

As shown in Figure 1, although there's no systematic bias between labeled and unlabeled samples, the risk of sampling bias can be severe when labeled samples are scarce. Without ARC, features learned by unlabeled and labeled data may deviate from each other, but still simultaneously satisfy their constrains due to DNN's great memorizing capacity. As observed in the plots, this hurts discrimination as misclassification increase even among seen unlabeled samples (left plot), while learned representations induce better decision boundary if labeled samples match the population (right plot).

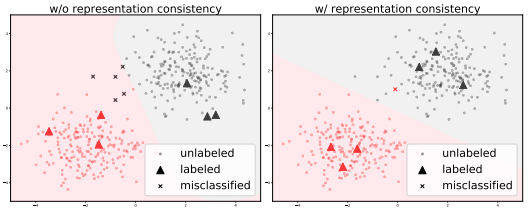

Figure 1. Illustration of why enforcing representation consistency helps the model generalize when labeled samples are scarce. Red and black spots denote unlabeled samples.

Appendix B. Additional Experiments

B.1. Descriptions about Datasets

- CUB-200-2011: The CUB-200-2011 dataset contains 200 fine-grained classes of birds with 11,788 images in total (about 30 images per class for training set and 30 images per class for validation set). In our experiment, we construct the labeled training set with the sample size of $n \in \{2000, 1000, 400, 200\}$, and use the rest images as unlabeled training set.
- MIT Indoor-67: Indoor-67 has 67 scene categories. In each category, there are 80 images for training and 20 images for testing. In our experiment, we construct the labeled training set with the sample size of $n \in \{1340, 670, 134\}$, and use the rest images as unlabeled training set.
- MURA: MURA is a dataset of musculoskeletal radiographs, which contains 40,561 images from 14,863 pa-

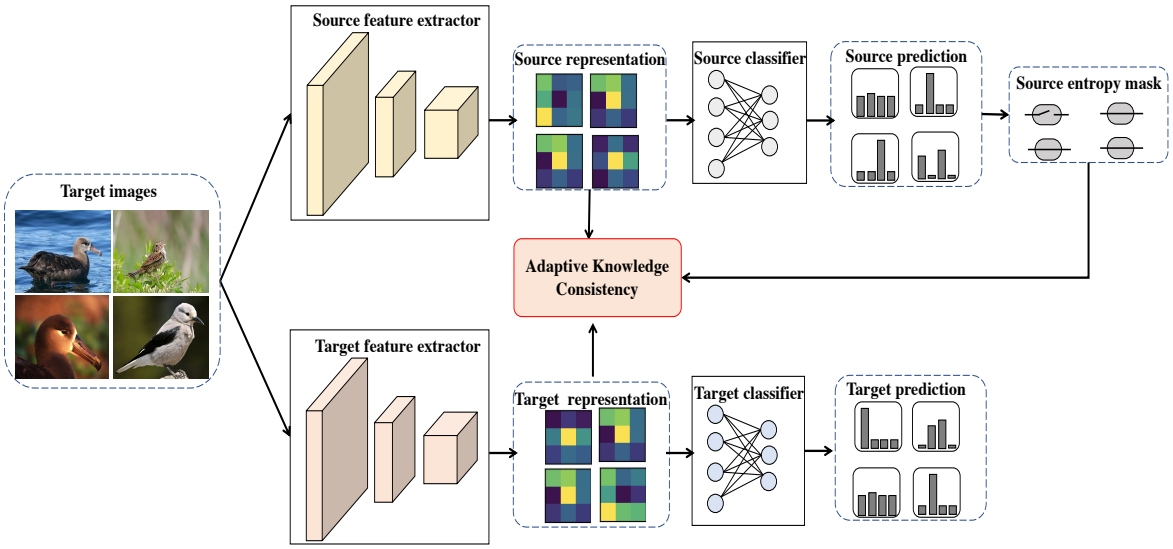

Figure 2. Adaptive knowledge consistency between the source and target model.

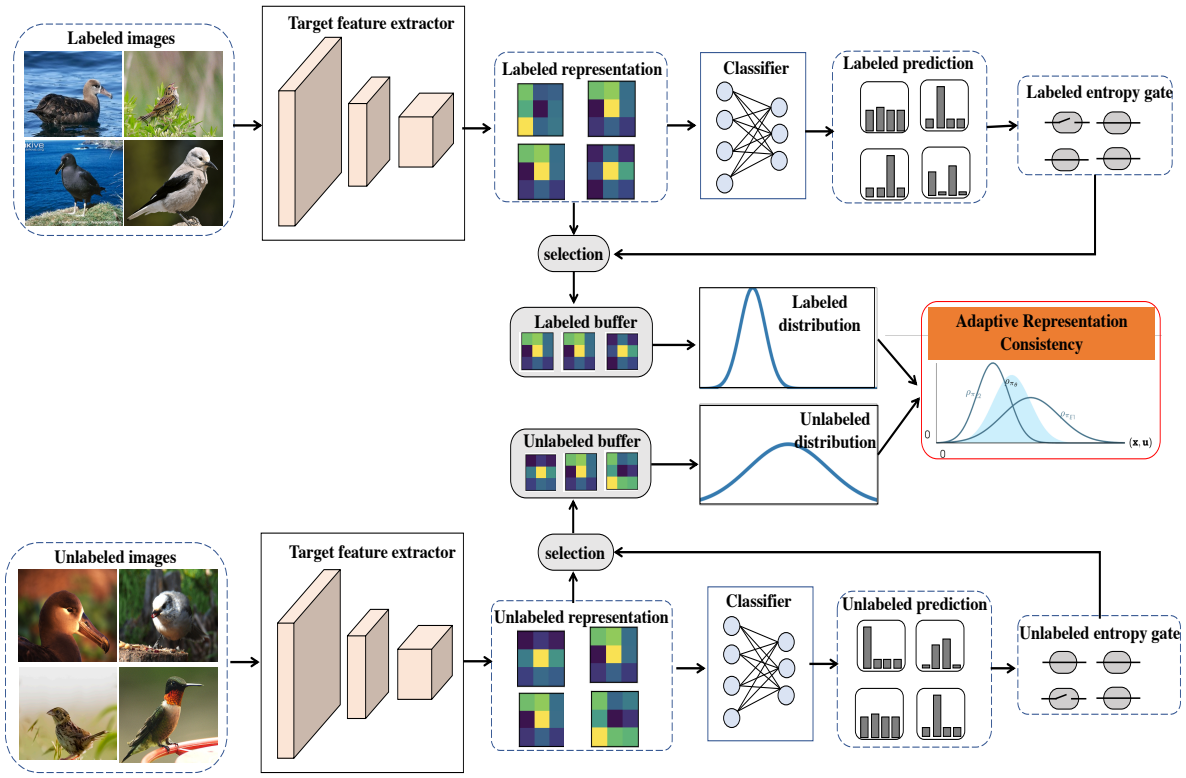

Figure 3. Adaptive representation consistency between labeled data distribution and unlabeled data distribution.

tient studies. X-ray images are collected from seven parts of human body: elbow, finger, forearm, hand, humerus, shoulder, and wrist. The goal of this dataset is to distinguish normal musculoskeletal studies from abnormal ones (a study often contains more than one image). This paper follows the experiment setting of [**?**]: to simply classify normal and abnormal radiographs (one image). For the MURA dataset, We construct the labeled training set with the sample size of $n \in \{1000, 400\}$, and use the rest images as unlabeled training set.

- CIFAR-10: The CIFAR-10 dataset is composed of 60,000 images of 10 classes with the size of 32x32.

#label	4000		250		40	
Method	From Scratch	Transfer	From Scratch	Transfer	From Scratch	Transfer
Pseudo label	16.09	7.04	49.78	12.92	79.51	25.62
Mean teacher	9.19	6.43	32.32	14.03	74.43	24.67
MixMatch	6.42	5.52	11.05	10.01	47.54	21.50
FixMatch	4.26	4.24	5.07	5.04	13.81	9.05

Table 1. Comparison of error rate using SSL methods with and without transfer learning.

50,000 images are used for training and 10,000 are used for testing.

Methods \#label	1340	670	134
Supervised labeled	68.94	63.35	44.28
Pseudo label	71.68	63.77	39.28
Mean teacher	71.34	64.37	43.05
MixMatch	73.14	68.58	44.65
FixMatch	74.27	68.31	44.13
AKC	71.93	66.64	46.79
ARC	72.72	66.94	46.67
AKC+ARC	73.31	67.44	47.11
MixMatch +AKC+ARC	75.54	70.30	48.54
FixMatch +AKC+ARC	76.64	70.61	48.34

Table 2. Classification accuracy of proposed AKC, ARC, and baselines on Indoor-67 dataset.

B.2. Results on Indoor-67

The experimental results on Indoor-67 dataset are listed in Table 2. Similar to the results of CUB-200-2011 dataset, the method of combining AKC with ARC achieves the best or comparable performance among previous-best baseline methods. In the case of 1340 (and 134) labeled sample size, by utilizing AKC and ARC regularization techniques in FixMatch, the performance is increased by 3.2% (and 9.54%) than vanilla FixMatch.

B.3. Empirical study about balancing AKC and ARC

We measure the increased accuracy after introducing AKC or ARC on three different Office-Home datasets*. Generally, as observed in Fig 4, AKC is relatively more useful as the discrepancy between the source and target dataset reduces†, while ARC contributes more with more unlabeled samples provided.

B.4. The time efficiency of our method

The proposed AKC and ARC involves almost only extra computation for knowledge distillation in the standard semi-supervised learning framework, which is much more computational efficient than complex operations used in modern SSL methods like MixMatch and FixMatch. Moreover, adding AKC+ARC on top of these competitive methods requires little additional cost as most operations can be reused. For example, combining AKC+ARC and FixMatch only increase 3% running time compared with the original FixMatch. The actual running time per iteration (in seconds) is measured on CUB-200, as shown in Table 3.

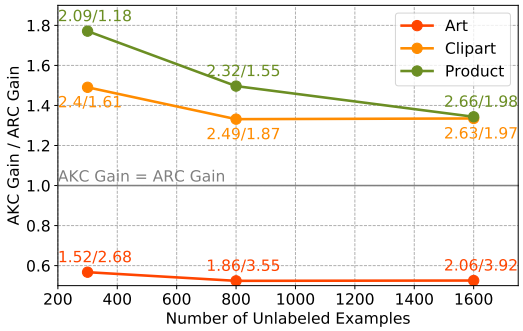

Figure 4. Comparison of AKC and ARC gain on Office-Home.

Method	MM	FM	AKC	ARC	AC	FMAC
Time(s)	0.629	0.563	0.531	0.513	0.562	0.580

Table 3. Running time per iteration for the CUB-200 experiment evaluated with Tesla V100 GPU. MM: MixMatch, FM: FixMatch, AC: AKC+ARC, FMAC: FM+AC.

B.5. Effectiveness of transfer learning in semi-supervised setting

We studied the effectiveness of transfer learning in some SSL methods on CIFAR-10 dataset, as shown in table 1. As can be seen, transfer learning could considerably improve the performance of SSL methods compared to learning from scratch, especially when labeled examples are insufficient. For example, given only 40 labels, transfer learning improves the performance of the leading SSL method FixMatch by 34.5% on CIFAR-10. Thus, the effectiveness of transfer learning in semi-supervised settings was underestimated in the previous works. With the Imprinting technique and proper training strategy, transfer learning could lead to a noticeable improvement.

*https://www.hemanthdv.org/officeHomeDataset.html

†*Art* is the most dissimilar with ImageNet due to its particular textures.