Supplementary Material for WAMDA: Weighted Alignment of Sources for Multi-source Domain Adaptation

Surbhi Aggarwal
surbhia@iisc.ac.in
Jogendra Nath Kundu
jogendrak@iisc.ac.in
R. Venkatesh Babu
venky@iisc.ac.in
Anirban Chakraborty
anirban@iisc.ac.in

Department of Computational and Data Sciences,
Indian Institute of Science
Bangalore, India

1 Architecture Details

We first elaborate upon the architecture used in our proposed method.

1.1 Pre-adaptation architecture

In the pre-adaptation stage, for each source S_i, we train the source-specific feature extractor F_{S_i}, source-specific classifier Q_{S_i} and the domain classifier D_{S_i}. We describe the implementation details of each of these components.
The architecture of F_{S_i} is as follows:

- ImageNet pre-trained ResNet-50 till average pool layer
- Linear FC (2048, 1024) + ELU
- Linear FC (1024, 1024) + BatchNorm + ELU
- Linear FC (1024, f_dim) + ELU
- Linear FC (f_dim, f_dim) + BatchNorm + ELU

Here, f_dim refers to the dimensionality of source feature space. Since this hyperparameter can differ with datasets, we have provided the value used for each dataset in Sections **2-5**.
The architecture of Q_{S_i} is as follows, where $n\_classes$ refers to the number of classes:

- Linear FC (f_dim, $n\_classes$)

The architecture of D_{S_i} is as follows:

- Linear FC (f_dim, $f\_dim/2$) + ELU + Linear FC ($f\_dim/2$, 2)

1.2 Adaptation architecture

In the adaptation stage, we train the weighted aligned source encoder E_{S_i} for each source S_i, and the target encoder E_T and the target classifier Q_T. We describe the implementation

details of each of these components.
The architecture of E_{S_i} is as follows:

- Linear FC (f_dim, 1024) + BatchNorm + ELU
- Linear FC (1024, 1024) + BatchNorm + ELU
- Linear FC (1024, c_dim) + BatchNorm + ELU
- Linear FC (c_dim, c_dim) + BatchNorm + ELU

Here, c_dim refers to the dimensionality of target features.
The architecture of E_T is as follows:

- ImageNet pre-trained ResNet-50 till average pool layer
- Linear FC (2048, 1024) + ELU
- Linear FC (1024, 1024) + BatchNorm + ELU
- Linear FC (1024, c_dim) + BatchNorm + ELU
- Linear FC (c_dim, c_dim) + BatchNorm + ELU

The architecture of Q_T is as follows:

- Linear FC (c_dim, c_dim) + ELU + Linear FC (c_dim, $n\_classes$)

2 Office-Home

2.1 Training details

The training details for Office-Home are as follows:

- In Office-Home, we used f_dim=256, c_dim=256.
- In the pre-adaptation stage, Adam optimizer is used for training the components F_{S_i}, Q_{S_i}. The ResNet-50 backbone of F_{S_i} is also fine-tuned. The learning rate for optimizing the backbone is 1e-5, and the learning rate for the other parameters is 1e-4.
- In the adaptation stage, we didn't fine-tune the ResNet-50 backbone of E_T. The parameters of E_{S_i}, Q_T and the remaining parameters of E_T were fine-tuned using Adam with learning rate of 1e-4.

2.2 Relative source relevance

In the results discussion, we refer to the intrinsic source relevance orders for any target. Here we discuss how we can obtain a ground truth of the order of source relevance for a given target. Since the actual order of source relevance is generally unavailable, we can obtain an approximate ground-truth order by observing the *source-only* accuracy of each target with each source. The source-only accuracy is obtained by passing the target samples directly to each source-specific feature extractor (F_{S_i}) and obtaining predictions from the source-specific classifier (Q_{S_i}). Thus, if domain gap is less between the source and target, the corresponding source-only accuracy will be high, and vice-versa. We assume the order obtained by comparing the source-only accuracy as the ground-truth, and we refer to this order while analysing the results. The ground-truth of source relevance as obtained by comparing the source-only accuracy is presented in Table 1. Please note, that this ground-truth order is not used in adaptation, rather it is used only to verify the order as obtained by the estimates of source relevance scores, and to do analysis of results.

	CPR →A	APR →C	ACR →P	ACP →R
Source-1	C = 55.2	A = 51.6	A = 66.1	A = 72.9
Source-2	P = 52.9	P = 51.7	C = 64.7	C = 65.0
Source-3	R = 65.6	R = 53.8	R = 78.6	P = 73.2
GroundTruth Order	R > C ≈ P	R ≈ P ≈ A	R > A ≈ C	P ≈ A > C

Table 1: Source-only adaptation accuracy (%) on *Office-Home* dataset.

	DW →A	AW →D	AD →W
Source-1	D = 64.7	A = 74.5	A = 74.5
Source-2	W = 64.7	W = 98.8	D = 96.8
GroundTruth Order	D ≈ W	W > A	D > A

Table 2: Source-only accuracy (%) on *Office-31* dataset.

3 Office-31

3.1 Training details

The training details for Office-Home are as follows:

- In Office-31, we used *f_dim*=256, *c_dim*=256.
- In the pre-adaptation stage, Adam optimizer is used for training the components F_{S_i}, Q_{S_i}. The ResNet-50 backbone of F_{S_i} is not fine-tuned. The other trainable parameters are optimized with a learning rate of 1e-4.
- In the adaptation stage, we didn't fine-tune the ResNet-50 backbone of E_T. The parameters of E_{S_i}, Q_T and the remaining parameters of E_T were fine-tuned using Adam with learning rate of 1e-4.

3.2 Relative source relevance

The ground-truth of source relevance as obtained by comparing the source-only accuracy is as shown in Table 2.

4 Office-Caltech

The training details for Office-Home are as follows:

- In Office-Caltech, we used *f_dim*=256, *c_dim*=256.
- In the pre-adaptation stage, Adam optimizer is used for training the components F_{S_i}, Q_{S_i}. The ResNet-50 backbone of F_{S_i} is also fine-tuned. The learning rate for optimizing the backbone is 1e-5, and the learning rate for the other parameters is 1e-4.
- In the adaptation stage, we didn't fine-tune the ResNet-50 backbone of E_T. The parameters of E_{S_i}, Q_T and the remaining parameters of E_T were fine-tuned using Adam with learning rate of 1e-4.

4.1 Relative source relevance

The ground-truth of source relevance as obtained by comparing the source-only accuracy is presented in Table 3.

	CDW →A	ADW →C	ACW →D	ACD →W
Source-1	C = 95.5	A = 89.6	A = 98.0	A = 90.5
Source-2	D = 93.5	C = 89.6	C = 92.9	C = 93.9
Source-3	W = 93.6	W = 89.6	W = 100	D= 100
GroundTruth Order	C > W ≈ D	A ≈ C ≈ W	W > A > C	D > C > A

Table 3: Source-only accuracy (%) on *Office-Caltech* dataset.

Standard	**Method**	IPQRS →C	CPQRS →I	CIQRS →P	CIPRS →Q	CIPQS →R	CIPQR →S	**Avg**
No Adapt	ResNet	39.6	8.2	33.9	11.8	41.6	23.1	26.4
Single Best	DAN	39.1	11.4	33.3	**16.2**	42.1	29.7	28.6
	MCD	42.6	19.6	42.6	3.8	50.5	33.8	32.2
Source Combine	DAN	45.4	12.8	36.2	15.3	48.6	34.0	32.1
	MCD	54.3	22.1	45.7	7.6	58.4	43.5	38.5
Multi-Source	$M^3SDA-\beta$	58.6	**26.0**	**52.3**	6.3	62.7	**49.5**	**42.6**
	Ours	**59.3**	21.8	52.1	9.5	**65.0**	47.7	**42.6**

Table 4: Classification accuracy (%) on *DomainNet* dataset.

5 DomainNet

5.1 Training details

- In DomainNet, we used f_dim=512, c_dim=256.
- Instead of ResNet-50, we used ImageNet pretrained ResNet-101.
- In the pre-adaptation stage, Adam optimizer is used for training the components F_{S_i}, Q_{S_i}. The ResNet-101 backbone of F_{S_i} is also fine-tuned. The learning rate for optimizing the backbone is 1e-5, and the learning rate for the other parameters is 1e-4.
- In the adaptation stage, we fine-tune the ResNet-101 backbone of E_T with a learning rate of 1e-5. The parameters of E_{S_i}, Q_T and the remaining parameters of E_T were fine-tuned using Adam with learning rate of 1e-4.
- We slightly modified the distillation-entropy loss for DomainNet. In all other datasets, we have used L1-loss to perform distillation between pseudo-softmax and prediction of Q_T. However, for DomainNet, we converted the pseudo-softmax to one-hot encoding, such that the class for which the score is maximum is labelled as 1 and rest as 0. Using this one-hot encoding of the pseudo-softmax, which we term as *pseudo-label*, we perform distillation using cross-entropy loss. This modification is used in DomainNet only.

5.2 Results

The results of our method on DomainNet is presented in Table 4. We observe that our overall performance is competitive with the performance of M^3SDA, showing that the proposed approach is applicable even on diverse and challenging datasets.