1. Proof of Lemma 1

Lemma 1.1. Assume that $\mathcal{L}_{D_k}$ is differentiable and $(W_k^*, \phi_k^*)$ is the unique minimizer of $\mathcal{L}_{D_k}(\theta, W_k, \phi_k) + \frac{1}{2} ||g(W_k) - g(W_C)||_F^2$. Then the gradient components of the meta-loss $F_k(\theta, W_C)$ with respect to $\theta$ and $W_C$ are given by $\frac{\partial F_k}{\partial \theta} = \frac{\partial \mathcal{L}_{D_k}(\theta, W_k, \phi_k)}{\partial \theta}$ and $\frac{\partial F_k}{\partial W_C} = \lambda \left( \frac{\partial g(W_C)}{\partial W_C} \right)^T [g(W_C) - g(W_k)]$, which are no Hessian information.

Proof. First, since $\mathcal{L}_{D_k}$ is differentiable and $(W_k^*, \phi_k^*) = \arg \min_{W_k, \phi_k} \mathcal{L}_{D_k}(\theta, W_k, \phi_k) + \frac{1}{2} ||g(W_k) - g(W_C)||_F^2$, from the first-order optimality condition we know that, when $W_k = W_k^*$ and $\phi_k = \phi_k^*$.

$$
\frac{\partial \mathcal{L}_{D_k}(\theta, \theta_k^*, \phi_k^*)}{\partial \theta} + \lambda \left( \frac{\partial g(W_k^*)}{\partial \theta} \right)^T [g(W_k^*) - g(W_C)] = 0,
$$
$$
\frac{\partial \mathcal{L}_{D_k}(\theta, \theta_k^*, \phi_k^*)}{\partial \phi_k} = 0.
$$

(1)

Next, we know that:

$$
F_k(\theta, W_C) = \min_{W_k, \phi_k} \mathcal{L}_{D_k}(\theta, W_k, \phi_k) + \frac{1}{2} ||g(W_k) - g(W_C)||_F^2
$$
$$
= \mathcal{L}_{D_k}(\theta, W_k, \phi_k) + \frac{\lambda}{2} ||g(W_k^*) - g(W_C)||_F^2,
$$

(2)

where $(W_k^*, \phi_k^*) = \arg \min_{W_k, \phi_k} \mathcal{L}_{D_k}(\theta, W_k, \phi_k) + \frac{1}{2} ||g(W_k) - g(W_C)||_F^2$.

Then, we compute the gradient components of $F_k(\theta, W_C)$ with respect to $\theta$ and $W_C$. From the chain rule, we have (3) and (4), where \(1\) and \(2\) hold according to (1).

2. Dataset and Implementation details

2.1. More details of Meta-Dataset

Meta-Dataset [3] contains different types of datasets with different categories. Some datasets contain natural images, like ImageNet, Flowers and Birds, but other datasets consist of special types of images. Quick Draw and Omniglot, their images are some black handwritten characters on a white background. Images of textures present perceptual features with different sense of quality, not having a single object in the image. Traffic signs involves a variety of traffic signs. MSCOCO is similar to ImageNet, but lower resolution.

2.2. Implementation details

In this section, we first introduce the implementation details of our baseline model multiple SDLs (single-domain models) and MDL (a multi-domain model) by optimizing (5) and (6), respectively, in our experimental comparison part. We use ResNet-18 as backbone, in line with the other models.

$$
\min_{\theta, \phi_k} \mathcal{L}_{D_k}(\theta, \phi_k), \quad k = 1, 2, ..., N.
$$

(5)

$$
\min_{\phi_k} \sum_{k=1}^{N} \mathcal{L}_{D_k}(\theta, \phi_k).
$$

(6)

In our experiment, $N$ equals 8, and the eight datasets are ImageNet, Omniglot, Aircraft, Birds, Textures, Quick Draw, Fungi, VGG Flower from Meta-dataset respectively, also called seen datasets.

We follow the training protocol of precious works [1, 2]. For SDLs, we use SGD with momentum and adjust the learning rate using cosine annealing. See the Line 2 to 9 of the Table 1, in which we set the learning rate, the weight decay, the annealing frequency, the batch size and the maximum number of training iterations (Max iteration) for each subdataset from Meta-Dataset. These results are finalized
and we can find that the optimal model of the specific
compute and compare the average performance. See the Ta-
each SDL on each subdataset from Meta-Dataset. And We
zation and inter-domain generalization, that is to evaluate
layer at each iteration, and the number of update steps is 2.
train hyperparameters as mentioned above is shown in the
train a separate model on the eight seen datasets, and the
mentations and random crops. And for MDL, we need to
augmentation in the training stage, like random color aug-
by evaluating the performance on the validation set of dif-
ferent values of hyperparameters. Meanwhile, we use data
mentation in the training stage, like random color aug-
3.3. Impact of gradient update step number in the
fitting. We use
as tuned. The results of different
m of the basis vector matrix W is a hyperparameter to be
tuned. The results of different m on Meta-Datasets are
shown in the Table 3. We find that settings with larger pa-
parameter values give better results than smaller values, prob-
ably because subspaces with higher dimension contain more
feature information, while too high value leads to over-
fitting. We use 384 in our model.
3.2. Impact of the subspace dimension on general-
ization
In our method, the subspace dimension, i.e., the rank
dataset is the model trained by the dataset or the ImageNet
model. We can also see that the SDL trained by ImageNet
outperforms other SDLs substantially, which may be due to
the large number and variety of images in ImageNet.
3.2. Impact of the subspace dimension on general-
ization
In our method, the subspace dimension, i.e., the rank
m of the basis vector matrix W is a hyperparameter to be
tuned. The results of different m on Meta-Datasets are
shown in the Table 3. We find that settings with larger pa-
parameter values give better results than smaller values, prob-
ably because subspaces with higher dimension contain more
feature information, while too high value leads to over-
fitting. We use 384 in our model.
3.2. Impact of the subspace dimension on general-
ization
In our method, the subspace dimension, i.e., the rank
m of the basis vector matrix W is a hyperparameter to be
tuned. The results of different m on Meta-Datasets are
shown in the Table 3. We find that settings with larger pa-
parameter values give better results than smaller values, prob-
ably because subspaces with higher dimension contain more
feature information, while too high value leads to over-
fitting. We use 384 in our model.
3.3. Impact of gradient update step number in the
lower-layer optimization.
Number of steps for gradient update in the lower layer of
our meta-problem is set as a hyperparameter n. The results

\begin{align}
\frac{\partial F_k}{\partial \theta} &= \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial \theta} + \left( \frac{\partial W_k^*}{\partial \theta} \right)^T \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial W_k^*} + \left( \frac{\partial \phi_k^*}{\partial \theta} \right)^T \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial \phi_k^*} \\
&+ \lambda \left( \frac{\partial g(W_k)}{\partial W_k} \right)^T \left[ g(W_k^*) - g(W_k) \right] \\
&= \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial \theta} + \left( \frac{\partial W_k^*}{\partial \theta} \right)^T \left\{ \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial W_k^*} + \lambda \left( \frac{\partial g(W_k)}{\partial W_k} \right)^T \left[ g(W_k^*) - g(W_k) \right] \right\} \\
&+ \left( \frac{\partial \phi_k^*}{\partial \theta} \right)^T \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial \phi_k^*} + 0 + 0 \\
&= \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial \theta} + \left( \frac{\partial W_k^*}{\partial \theta} \right)^T \left\{ \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial W_k^*} + \lambda \left( \frac{\partial g(W_k)}{\partial W_k} \right)^T \left[ g(W_k^*) - g(W_k) \right] \right\} \\
&+ \left( \frac{\partial \phi_k^*}{\partial \theta} \right)^T \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial \phi_k^*} \tag{3}
\end{align}

\begin{align}
\frac{\partial F_k}{\partial W_C} &= \left( \frac{\partial W_k^*}{\partial W_C} \right)^T \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial W_k^*} + \left( \frac{\partial \phi_k^*}{\partial \theta} \right)^T \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial \phi_k^*} \\
&+ \lambda \left( \frac{\partial g(W_C)}{\partial W_C} \right)^T \left[ g(W_k^*) - g(W_k) \right] \\
&= \left( \frac{\partial W_k^*}{\partial W_C} \right)^T \left\{ \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial W_k^*} + \lambda \left( \frac{\partial g(W_k)}{\partial W_k} \right)^T \left[ g(W_k^*) - g(W_k) \right] \right\} \\
&+ \left( \frac{\partial \phi_k^*}{\partial \theta} \right)^T \frac{\partial L_{D_k}(\theta, W_k^*, \phi_k^*)}{\partial \phi_k^*} + 0 + \lambda \left( \frac{\partial g(W_C)}{\partial W_C} \right)^T \left[ g(W_k^*) - g(W_k) \right] \\
&= \lambda \left( \frac{\partial g(W_C)}{\partial W_C} \right)^T \left[ g(W_C) - g(W_k^*) \right] \tag{4}
\end{align}
Table 1. Training hyperparameters of multiple SDLs and MDL on Meta-Dataset. The first column represent different SDLs on different subdataset of Meta-Dataset.

| Model       | Learning rate | Weight decay | Annealing frequency | Batch size | Max iteration |
|-------------|---------------|--------------|---------------------|------------|---------------
| ImageNet    | $3 \times 10^{-2}$ | $7 \times 10^{-4}$ | 48000               | 64         | 480000        |
| Omniglot    | $3 \times 10^{-2}$ | $7 \times 10^{-4}$ | 3000                | 16         | 5000          |
| Aircraft    | $3 \times 10^{-2}$ | $7 \times 10^{-4}$ | 3000                | 8          | 5000          |
| Birds       | $3 \times 10^{-2}$ | $7 \times 10^{-4}$ | 3000                | 16         | 5000          |
| Textures    | $3 \times 10^{-2}$ | $7 \times 10^{-4}$ | 1500                | 32         | 5000          |
| Quick Draw  | $1 \times 10^{-2}$ | $7 \times 10^{-4}$ | 48000               | 64         | 480000        |
| Fungi       | $3 \times 10^{-2}$ | $7 \times 10^{-4}$ | 1500                | 32         | 480000        |
| VGG Flower  | $3 \times 10^{-2}$ | $7 \times 10^{-4}$ | 1500                | 8          | 5000          |
| MDL         | $3 \times 10^{-2}$ | $7 \times 10^{-4}$ | 48000               | 64         | 240000        |

Table 2. Results of all SDLs. Mean accuracy and 95% reported. All results are obtained during meta-testing phase. The test tasks of the domain corresponding to the trained model for each column belong to in-domain generalization, while the remaining domains are out-of-domain generalization. We also report overall accuracy for all domains.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>ImageNet</th>
<th>Omniglot</th>
<th>Aircraft</th>
<th>Birds</th>
<th>Textures</th>
<th>Quick Draw</th>
<th>Fungi</th>
<th>VGG Flower</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>$55.78 \pm 1.0$</td>
<td>16.01 ± 0.5</td>
<td>21.18 ± 0.7</td>
<td>26.26 ± 0.8</td>
<td>26.00 ± 0.8</td>
<td>22.53 ± 0.7</td>
<td>32.30 ± 0.9</td>
<td>24.5 ± 0.8</td>
<td></td>
</tr>
<tr>
<td>Omniglot</td>
<td>$65.73 \pm 1.3$</td>
<td>93.20 ± 0.5</td>
<td>56.87 ± 1.3</td>
<td>57.75 ± 1.2</td>
<td>54.37 ± 1.3</td>
<td>77.24 ± 1.0</td>
<td>56.20 ± 1.2</td>
<td>54.0 ± 1.2</td>
<td></td>
</tr>
<tr>
<td>Aircraft</td>
<td>$49.77 \pm 0.9$</td>
<td>17.37 ± 0.5</td>
<td>85.74 ± 0.5</td>
<td>29.48 ± 0.7</td>
<td>23.72 ± 0.6</td>
<td>25.48 ± 0.7</td>
<td>31.08 ± 0.7</td>
<td>24.5 ± 0.6</td>
<td></td>
</tr>
<tr>
<td>Birds</td>
<td>$70.43 \pm 0.8$</td>
<td>13.51 ± 0.5</td>
<td>19.14 ± 0.7</td>
<td>71.24 ± 0.9</td>
<td>22.80 ± 0.7</td>
<td>17.59 ± 0.7</td>
<td>43.16 ± 0.9</td>
<td>27.4 ± 0.8</td>
<td></td>
</tr>
<tr>
<td>Textures</td>
<td>$72.95 \pm 0.6$</td>
<td>29.19 ± 0.5</td>
<td>41.32 ± 0.6</td>
<td>42.73 ± 0.6</td>
<td>60.40 ± 0.7</td>
<td>39.40 ± 0.7</td>
<td>57.27 ± 0.7</td>
<td>42.1 ± 0.7</td>
<td></td>
</tr>
<tr>
<td>Quick Draw</td>
<td>$55.20 \pm 0.9$</td>
<td>52.53 ± 0.9</td>
<td>40.22 ± 1.0</td>
<td>38.66 ± 0.9</td>
<td>41.59 ± 1.0</td>
<td>82.81 ± 0.6</td>
<td>35.57 ± 0.9</td>
<td>39.5 ± 1.1</td>
<td></td>
</tr>
<tr>
<td>Fungi</td>
<td>$42.72 \pm 1.1$</td>
<td>9.80 ± 0.5</td>
<td>13.10 ± 0.6</td>
<td>25.70 ± 0.9</td>
<td>17.59 ± 0.8</td>
<td>11.92 ± 0.6</td>
<td>65.78 ± 0.9</td>
<td>23.4 ± 0.7</td>
<td></td>
</tr>
<tr>
<td>VGG Flower</td>
<td>$86.96 \pm 0.6$</td>
<td>23.66 ± 0.6</td>
<td>47.03 ± 0.8</td>
<td>63.74 ± 0.8</td>
<td>49.23 ± 0.9</td>
<td>35.32 ± 0.8</td>
<td>79.70 ± 0.7</td>
<td>78.2 ± 0.6</td>
<td></td>
</tr>
<tr>
<td>Traffic Sign</td>
<td>$48.28 \pm 1.0$</td>
<td>16.27 ± 0.6</td>
<td>33.92 ± 0.9</td>
<td>35.65 ± 0.9</td>
<td>37.54 ± 1.0</td>
<td>30.79 ± 1.0</td>
<td>26.77 ± 0.7</td>
<td>30.3 ± 0.8</td>
<td></td>
</tr>
<tr>
<td>MSCOCO</td>
<td>$51.99 \pm 1.0$</td>
<td>14.41 ± 0.6</td>
<td>20.10 ± 0.7</td>
<td>25.17 ± 0.8</td>
<td>26.57 ± 0.9</td>
<td>19.39 ± 0.8</td>
<td>30.21 ± 0.9</td>
<td>25.6 ± 1.8</td>
<td></td>
</tr>
<tr>
<td>MNIST</td>
<td>$78.00 \pm 0.6$</td>
<td>92.82 ± 0.4</td>
<td>67.94 ± 0.7</td>
<td>78.18 ± 0.6</td>
<td>72.39 ± 0.7</td>
<td>87.61 ± 0.5</td>
<td>66.03 ± 0.7</td>
<td>72.9 ± 0.7</td>
<td></td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>$69.64 \pm 0.7$</td>
<td>29.89 ± 0.6</td>
<td>38.53 ± 0.7</td>
<td>39.60 ± 0.7</td>
<td>40.78 ± 0.7</td>
<td>38.37 ± 0.7</td>
<td>37.03 ± 0.7</td>
<td>40.3 ± 0.8</td>
<td></td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>$58.56 \pm 1.0$</td>
<td>13.56 ± 0.7</td>
<td>23.24 ± 0.8</td>
<td>29.81 ± 0.9</td>
<td>29.12 ± 0.9</td>
<td>23.62 ± 0.9</td>
<td>27.02 ± 0.9</td>
<td>29.0 ± 1.0</td>
<td></td>
</tr>
</tbody>
</table>

Average 61.7 32.5 39.2 43.40 38.62 39.4 45.6 39.4

Table 3. Quantitative analysis of subspace dimension on Meta-Dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$m$</th>
<th>256</th>
<th>320</th>
<th>384</th>
<th>446</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-domain</td>
<td>72.54</td>
<td>74.31</td>
<td>79.72</td>
<td>77.39</td>
<td></td>
</tr>
<tr>
<td>Out-of-domain</td>
<td>53.32</td>
<td>63.54</td>
<td>69.35</td>
<td>68.97</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>65.15</td>
<td>70.17</td>
<td><strong>75.73</strong></td>
<td>74.15</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Quantitative analysis of gradient update step number on Meta-Dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Step $n$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-domain</td>
<td>79.18</td>
<td>79.72</td>
<td>78.21</td>
<td>77.89</td>
<td></td>
</tr>
<tr>
<td>Out-of-domain</td>
<td>68.79</td>
<td>69.35</td>
<td>67.67</td>
<td>66.35</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>75.18</td>
<td><strong>75.73</strong></td>
<td>74.16</td>
<td>73.45</td>
<td></td>
</tr>
</tbody>
</table>

of different $n$ on Meta-Datasets are shown in the Table 4. A good choice for our model is $n = 2$ according to attempts already made.

References


[2] Wei-Hong Li, Xiaole Liu, and Hakan Bilen. Universal repre-