Supplementary Material:

Learning What and Where to Transfer

A. Network Architectures and Tasks

For small image experiments (32×32), we use TinyImageNet[1] as a source task, and use CIFAR-10, CIFAR-100 (Krizhevsky & Hinton, 2009) and STL-10 (Coates et al., 2011) datasets as target tasks. CIFAR-10 and CIFAR-100 have 10 and 100 classes containing 5000 and 500 training images for each class, respectively, and each image has 32×32 pixels. STL-10 consists of 10 classes, with 500 labeled images per each class in the training set. Since the original images in TinyImageNet and STL-10 are not 32×32, we resize them into 32×32 when training and testing. We use a pre-trained 32-layer ResNet (He et al., 2016) on TinyImageNet as a source model, and we train 9-layer VGG (Simonyan & Zisserman, 2015), which is the modified architecture used in Srinivas & Fleuret (2018), on CIFAR-10/100 and STL-10 datasets.

For large image experiments (224×224), we use a pre-trained 34-layer ResNet on ImageNet (Deng et al., 2009) as a source model, and consdier Caltech-UCSD Bird 200 (Wah et al., 2011), MIT Indoor Scene Recognition (Quattoni & Torralba, 2009), Stanford 40 Actions (Yao et al., 2011) and Stanford Dogs (Khosla et al., 2011) datasets as target tasks. Caltech-UCSD Bird 200 (CUB200) contains 5k training images of 200 bird species. MIT Indoor Scene Recognition (MIT67) has 67 labels for indoor scenes and 80 training images per each label. Stanford 40 Actions (Stanford40) contains 4k training images of 40 human actions. Stanford Dogs has 12k training images of 120 dog categories. For these target fine-grained datasets, we train 18-layer ResNets.

B. Optimization

All target networks are trained by stochastic gradient descent (SGD) with a momentum of 0.9. We use a weight decay of 10^{-4} and an initial learning rate 0.1 and decay the learning rate with a cosine annealing (Loshchilov & Hutter, 2017): $\alpha_t = \frac{1}{2}(1 + \cos \frac{t}{T}\pi)$ where α_t is the learning rate at epoch t, and T is the maximum epoch. For all experiments, we train target networks for $T = 200$ epochs. The size of mini-batch is 128 for small image experiments, e.g., CIFAR, or 64 for large image experiments, e.g., CUB200. When using feature matching, we use $\beta = 0.5$. For data pre-processing and augmentation schemes, we follow He et al. (2016). We use the ADAM (Kingma & Ba, 2015) optimizer for training the meta-networks f_ϕ, g_ϕ with a learning rate of 10^{-3} or 10^{-4}, and a weight decay of 0 or 10^{-4}. In our meta-training scheme, we observe that $T = 2$ is enough to learn what and where to transfer. We repeat experiments 3 times and report the average performance as well as the standard deviation.

C. Ablation Studies

C.1. Comparison between the meta-networks and meta-weights

Table 1. Classification accuracy (%) of transfer learning using meta-networks or meta-weights.

Target task	CUB200	MIT67	Stanford40
meta-weights	61.75	64.10	58.88
meta-networks	65.05	64.85	63.08

The weights, channel importance $w^{m,n}$ and connection importance $\lambda^{m,n}$, decide amounts of transfer given a sample to meta-networks. One can also learn directly $w^{m,n}$ and $\lambda^{m,n}$ as constant meta-weights using suggested bilevel scheme without meta-networks. Here, we compare the effectiveness of using *meta-networks*, which gives different amount of transfer for each sample, to learning *meta-weights* directly, giving the same importance over all the samples. For fair comparison, we use same hyperparameters as described in Section A and B, except the meta-parameters. As reported in Table 1, the performance of target models using meta-networks outperforms the one using meta-weights up-to 4.2%, which supports the effectiveness of using selective transfer depending on samples.

[1] https://tiny-imagenet.herokuapp.com/

C.2. Comparison between the proposed bilevel scheme and original one

To validate the effectiveness of the suggested bilevel scheme, we perform experiments comparing the performance of target models trained with meta-networks, using the proposed and original bilevel scheme. For a fair comparison, we use $T = 2$ for both methods, and the other hyperparameters, model architectures and the source task are same with the ones in Section A and B. The original scheme obtains significantly lower accuracies than the proposed bilevel scheme (Table 2). With much larger T, e.g., 5$\sim$100, a target model with the original bilevel scheme does not succeed to obtain comparable performance with our bilevel scheme. Moreover the meta-training time for meta-networks is increasing linearly as T increases, thus the original scheme is not applicable to practical scenarios. These results show that the proposed bilevel scheme is more effective for learning meta-networks for selective transfer.

Table 2. Classification accuracy (%) of transfer learning using the original or proposed bilevel schemes.

Target task	CUB200	MIT67	Stanford40
Original	35.38	54.18	53.47
Ours	65.05	64.85	63.08

References

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, 2011.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 2009.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, 2016.

Khosla, A., Jayadevaprakash, N., Yao, B., and Fei-Fei, L. Novel dataset for fine-grained image categorization. In *The 1st Workshop on Fine-Grained Visual Categorization, the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 2011.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *The 3rd International Conference on Learning Representations (ICLR 2015)*, 2015.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *The 5th International Conference on Learning Representations (ICLR 2017)*, 2017.

Quattoni, A. and Torralba, A. Recognizing indoor scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 2009.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *The 3rd International Conference on Learning Representations (ICLR 2015)*, 2015.

Srinivas, S. and Fleuret, F. Knowledge transfer with Jacobian matching. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., and Fei-Fei, L. Human action recognition by learning bases of action attributes and parts. In *The IEEE International Conference on Computer Vision (ICCV 2011)*, 2011.