Supplementary Materials for Fair Feature Distillation for Visual Recognition

Sangwon Jung[1]*, Donggyu Lee[1]*, Taeeon Park[1]* and Taesup Moon[2]†
[1]Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, Korea
[2]Department of Electrical and Computer Engineering, Seoul National University, Seoul, Korea
{s.jung, ldk308, pte1236}@skku.edu, tsmoon@snu.ac.kr

1. Implementation Details

For all datasets, we trained all methods for 50 epochs with a mini-batch size of 128 using Adam optimizer with an initial learning rate 0.001 and decaying it by a factor of 10 if no improvement in the test loss for 10 consecutive epochs. Also, all results were averaged over 4 different random runs.

1.1. Network Architecture for CIFAR-10S

We employed a simple convolutional neural network having six convolutional layers with the kernel size of 3×3, followed by two fully connected hidden layers with ReLU [1] activations. The number of channels was set to 32, 32, 64, 64, 128, and 128 for each convolutional layer, respectively. Dropout [4] and max-pooling were applied after every two convolutional layers.

1.2. Hyperparameters for Main Results

For fair comparison, we did the extensive search for one hyperparameter of each method including ours and baselines. We set one parameter to search for and fixed others using a suitable strategy for baselines having more than two hyperparameters. For HKD and FitNet, we focused on finding the optimal T, a temperature to soften the output, while we gradually annealed the strength of output distillation for both methods and fixed feature distillation strength for FitNet to 1 like in [3]. For AD, we tune the strength of the adversary loss while fixing the learning rate of it to 0.003, a commonly used value. For the variants of SS, we do the same search strategy as the knowledge distillation methods. For the variants of AD, we fixed the all hyperparameters of the knowledge distillation to the best values found in the experiment for single distillations and searched the strength of the adversary loss to control the balance between two methods. The values for hyperparameters used to report the results in the manuscript are in Table 1. In Table 1, we denote the strength for each method as λ.

*Equal contribution.
†Corresponding author (E-mail: tsmoon@snu.ac.kr)

1.3. Details on AD+FitNet

Three combined methods of the third class of baselines in the manuscript, except for AD+FitNet, are naturally implemented, but implementing AD+FitNet requires modification to FitNet. More specifically, FitNet originally has two stages of training, the hint training for feature distillation and the KD training for output distillation. However, since this stage-wise training of FitNet has difficulty to being incorporated with the mini-max game with an adversary in AD, we modify the two stages training FitNet to one stage FitNet by minimizing the output and feature distillation loss simultaneously, as in [5]. Then, we integrate the loss of an adversary of AD into the loss of one stage FitNet to implement AD+FitNet.

1.4. Hyperparmeters for t-SNE

Hyperparameters of t-SNE feature visualization for (Figure 5, manuscript) are as follows : dimension of the embedded space (3), perplexity (200), early exaggeration(1.0), maximum number of iterations (250), metric (cosine), random state (5). For other factors, we remained default in scikit-learn [2].

Table 1: Hyperparameters for experiments.

Methods\Dataset	CIFAR-10S	UTKFace	CelebA
HKD	T (1)	T (3)	T (5)
FitNet	T (1)	T (5)	T (1)
AT	λ (1)	λ (30)	λ (1)
NST	λ (30)	λ (3)	-
AD	λ (0.001)	λ (0.01)	λ (10)
SS+HKD	T (3)	T (5)	T (3)
SS+FitNet	T (3)	T (10)	T (3)
AD+HKD	T (1) λ (1e-4)	T (3) λ (30)	T (5) λ (10)
AD+FitNet	T (1) λ (1e-3)	T (5) λ (1)	T (1), λ (1)
MFD	λ (3)	λ (3)	λ (7)

2. Result Tables

Table 2, 3 and 4 show the detail results. The number in the parenthesis with $\pm$ sign stands for the standard deviation

Table 2: Average accuracy (%) and DEO (%) with standard deviation on CIFAR-10S.

	Accuracy	DEO_A	DEO_M
Teacher	79.62 (±0.14)	15.63 (±0.44)	31.32 (±1.47)
HKD	80.34 (±0.35)	15.54 (±0.67)	34.12 (±2.21)
FitNet	81.66 (±0.20)	14.83 (±0.26)	32.28 (±1.59)
AT	79.00 (±0.99)	15.57 (±0.71)	31.25 (±1.20)
NST	79.70 (±0.99)	15.11 (±0.75)	30.87 (±2.38)
SS	82.69 (±0.22)	3.29 (±0.30)	7.13 (±1.36)
AD	62.49 (±30.32)	11.59 (±6.75)	23.07 (±13.36)
SS+HKD	82.27 (±0.33)	10.15 (±0.20)	20.37 (±1.14)
SS+FitNet	81.73 (±0.39)	10.35 (±0.47)	20.92 (±0.54)
AD+HKD	79.27 (±0.33)	16.19 (±0.50)	33.25 (±0.72)
AD+FitNet	79.59 (±0.37)	15.90 (±0.51)	32.47 (±1.66)
MFD	**82.77 (±0.14)**	**2.73 (±0.41)**	**6.08 (±0.91)**

Table 3: Average accuracy (%) and DEO (%) with standard deviation on UTKFace.

	Accuracy	DEO_A	DEO_M
Teacher	74.54 (±1.07)	21.92 (±1.36)	39.25 (±2.86)
HKD	**76.17 (±0.58)**	22.5 (±0.76)	41.25 (±3.49)
FitNet	75.23 (±0.52)	21.50 (±1.59)	40.00 (±4.64)
AT	75.17 (±0.82)	22.67 (±3.41)	40.50 (±6.87)
NST	75.10 (±0.39)	22.75 (±0.49)	42.00 (±4.18)
SS	75.23 (±0.87)	24.33 (±1.75)	38.50 (±2.29)
AD	74.67 (±1.01)	20.42 (±1.55)	36.00 (±2.55)
SS+HKD	76.08 (±0.42)	21.92 (±1.07)	37.50 (±2.05)
SS+FitNet	75.5 (±0.99)	21.92 (±1.75)	38.00 (±2.06)
AD+HKD	69.48 (±3.21)	18.75 (±1.93)	32.50 (±4.15)
AD+FitNet	70.23 (±6.64)	21.17 (±6.03)	33.75 (±6.06)
MFD	74.69 (±0.69)	**17.75 (±1.38)**	**28.50 (±1.80)**

Table 4: Average accuracy (%) and DEO (%) with standard deviation on CelebA.

	Accuracy	DEO_A	DEO_M
Teacher	78.33 (±0.08)	21.04 (±0.48)	21.81 (±0.13)
HKD	78.64 (±0.37)	21.56 (±0.92)	22.54 (±0.60)
FitNet	78.62 (±0.20)	20.66 (±0.81)	21.70 (±0.56)
AT	78.63 (±0.22)	21.28 (±0.28)	22.24 (±0.51)
SS	79.67 (±0.36)	4.87 (±0.69)	5.22 (±0.81)
AD	76.10 (±1.12)	**2.51 (±2.12)**	**3.34 (±3.09)**
SS+HKD	79.95 (±0.42)	8.41 (±1.78)	8.27 (±1.83)
SS+FitNet	79.77 (±0.28)	9.31 (±1.77)	8.61 (±2.23)
AD+HKD	80.31 (±0.30)	3.40 (±2.46)	4.05 (±2.86)
AD+FitNet	**80.60 (±0.14)**	5.12 (±1.67)	5.51 (±1.64)
MFD	80.15 (±0.29)	5.46 (±0.95)	5.86 (±0.83)

of each metric obtained from 4 independent runs.

References

[1] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning (ICML)*, 2010. 1

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 1

[3] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[4] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 1

[5] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020. 1