Wasserstein Contrastive Representation Distillation: Supplementary Material

Liqun Chen[1]*, Dong Wang[1]*, Zhe Gan[2], Jingjing Liu[2], Ricardo Henao[1], Lawrence Carin[1]
[1]Duke University, [2]Microsoft Corporation
{liqun.chen, dong.wang, ricardo.henao, lcarin}@duke.edu, {zhe.gan, jingjl}@microsoft.com

A. WCoRD Algorithm

The detailed implementation of the proposed Wasserstein Contrastive Representation Distillation (WCoRD) method is summarized in Algorithm 1.

Algorithm 1 The proposed WCoRD Algorithm.

Input: A mini-batch of data samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$.
Extract features $\boldsymbol{h}^T$ and $\boldsymbol{h}^S$ from the teacher and student networks, respectively.
Construct a memory buffer $\mathcal{B}$ to store previous computed features.
Global contrastive knowledge transfer:
    Max. the GCKT loss in Eqn. (11) over $\boldsymbol{\theta}_S$ and $\boldsymbol{\phi}$.
Local contrastive knowledge transfer:
    Min. the LCKT loss in Eqn. (13) over $\boldsymbol{\theta}_S$.
Min. the task-specific loss over $\boldsymbol{\theta}_S$.

B. Baseline Methods and Model Architectures

B.1. Baseline Methods

We compare WCoRD with a number of baseline distillation methods, detailed below.

- Fitnets: Hints for thin deep nets [11];
- Knowledge Distillation (KD) [5];
- Attention Transfer (AT) [19];
- Like what you like: Knowledge distilllation via neuron selectivity transfer (NST) [6];
- A gift from knowledge distillation: fast optimization, network minimization and transfer learning (FSP) [18];
- Learning deep representations with probabilistic knowledge transfer (PKT) [9];
- Paraphrasing complex network: network compression via factor transfer (FT) [7];
- Similarity-preserving knowledge distillation (SP) [17];
- Correlation congruence (CC) [10];
- Variational information distillation for knowledge transfer (VID) [1];
- Relational knowledge distillation (RKD) [8];
- Knowledge transfer via distillation of activation boundaries formed by hidden neurons (AB) [4];
- Contrastive representation distillation (CRD) [16] via NCE [2].

Note that the hyper-parameter setup for these baseline methods follows the setup in CRD [16].

B.2. Model Architectures

In experiments, we utilize the following model architectures.

- Wide Residual Network (WRN) [20]: WRN-d-w represents wide ResNet with depth d and width factor w.
- resnet [3]: We use ResNet-d to represent CIFAR-style resnet with 3 groups of basic blocks, each with 16, 32, and 64 channels, respectively. In our experiments, resnet8x4 and resnet32x4 indicate a 4 times wider network (namely, with 64, 128, and 256 channels for each of the blocks).
- ResNet [3]: ResNet-d represents ImageNet-style ResNet with bottleneck blocks and more channels.
- MobileNetV2 [12]: In our experiments, we use a width multiplier of 0.5.
- vgg [13]: The vgg networks used in our experiments are adapted from their original ImageNet counterpart.
- ShuffleNetV1 [21], ShuffleNetV2 [15]: ShuffleNets are proposed for efficient training and we adapt them to input of size 32x32.
- InceptionNet-v3 [14] is used for the teacher network in the priviledge distillation experiment.

*Equal contribution

Teacher **Student**	**WRN-40-2** **WRN-16-2**	**WRN-40-2** **WRN-40-1**	**resnet56** **resnet20**	**resnet110** **resnet20**	**resnet110** **resnet32**	**resnet32x4** **resnet8x4**	**vgg13** **vgg8**
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36
CRD	75.48 $\pm$ 0.09	74.14 $\pm$ 0.22	71.16 $\pm$ 0.17	71.46 $\pm$ 0.09	73.48 $\pm$ 0.13	75.51 $\pm$ 0.18	73.94 $\pm$ 0.22
CRD+KD	75.64 $\pm$ 0.21	74.38 $\pm$ 0.11	71.63 $\pm$ 0.15	71.56 $\pm$ 0.16	73.75 $\pm$ 0.24	75.46 $\pm$ 0.25	74.29 $\pm$ 0.12
WCoRD	75.88$\pm$ 0.07	74.73 $\pm$ 0.17	71.56 $\pm$ 0.13	71.57 $\pm$ 0.09	73.81$\pm$ 0.11	75.95 $\pm$ 0.11	74.55$\pm$ 0.18
WCoRD+KD	76.11 $\pm$ 0.15	74.72$\pm$ 0.14	71.92$\pm$ 0.17	71.88$\pm$ 0.15	74.20$\pm$ 0.20	76.15$\pm$ 0.14	74.72$\pm$ 0.13

Table 1: Results with standard deviation of both the CRD and WCoRD methods.

λ_1	0	0.05	0.07	0.1	0.15	0.2	0.5	0.8	1.0
Result	79.12	80.11	82.15	83.50	83.33	83.78	84.2	**84.5**	84.3

Table 2: AUC (%) of student network ResNet-8x4 with different λ_1 values on the GCKT term.

C. Additional Results

In Table 1, we report additional results of the baseline distillation methods when combined with KD, and the standard deviation of the results of both CRD and WCoRD, with or without KD. Our method achieves better performance.

We also tested the importance of the GCKT module in WCoRD. We fixed the LCKT module by choosing $\lambda_2 = 0.1$, and then we adjust λ_1 from 0 to 1.0. Results are summarized in Table 2. Our model is fairly robust towards different choices of λ_1. Also, without the help of the GCKT module, models only with LCKT cannot obtain a very good performance.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *CVPR*, 2019. 1

[2] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *AISTATS*, 2010. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[4] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *AAAI*, 2019. 1

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPS Deep Learning and Representation Learning Workshop*, 2015. 1

[6] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017. 1

[7] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *NeurIPS*, 2018. 1

[8] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 1

[9] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, 2018. 1

[10] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *ICCV*, 2019. 1

[11] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *ICLR*, 2014. 1

[12] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 1

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 1

[15] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *CVPR*, 2019. 1

[16] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2020. 1

[17] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, 2019. 1

[18] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017. 1

[19] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 1

[20] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 1

[21] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 2018. 1