Training Lightweight yet Competent Network via Transferring Complementary Features

Xiaobing Zhang[1], Shijian Lu[2(✉)], Haigang Gong[1], Minghui Liu[1], and Ming Liu[1]

[1] University of Electronic Science and Technology of China, Sichuan, China
zhangxiaobing@std.uestc.edu.cn, {hggong,csmliu}@uestc.edu.cn, minghuiliuuestc@163.com

[2] Nanyang Technological University, Singapore, Singapore
Shijian.Lu@ntu.edu.sg

Abstract. Though deep neural networks have achieved quite impressive performance in various image detection and classification tasks, they are often constrained by requiring intensive computation and large storage space for deployment in different scenarios and devices. This paper presents an innovative network that aims to train a lightweight yet competent student network via transferring multifarious knowledge and features from a large yet powerful teacher network. Based on the observations that different vision tasks are often correlated and complementary, we first train a resourceful teacher network that captures both discriminative and generative features for the objective of image classification (the main task) and image reconstruction (an auxiliary task). A lightweight yet competent student network is then trained by mimicking both pixel-level and spatial-level feature distribution of the resourceful teacher network under the guidance of feature loss and adversarial loss, respectively. The proposed technique has been evaluated over a number of public datasets extensively and experiments show that our student network obtains superior image classification performance as compared with the state-of-the-art.

Keywords: Knowledge distillation · Transfer learning · Model compression

1 Introduction

Deep neural networks (DNNs) have demonstrated superior performances in various research fields [2,15–17]. However, deeper and larger networks often come with high computational costs and large memory requirements which have impeded effective and efficient development and deployment of DNNs in various resource-constrained scenarios. In recent years, knowledge transfer has attracted increasing interest and several promising networks have been developed through

H. Yang et al. (Eds.): ICONIP 2020, CCIS 1332, pp. 571–579, 2020.
https://doi.org/10.1007/978-3-030-63820-7_65

knowledge distillation (KD) [5], attention transfer (AT) [11], factor transfer (FT) [6], etc. On the other hand, the aforementioned works share a common constrain of feature uniformity where the teacher network is trained with the task-specific objective alone and so learn (and transfer) unitary features and knowledge only. In addition, the *teacher-learned* features are usually optimal for the teacher's performance which may not be the case for the student network due to the large discrepancies in network architecture, network capacity and initial conditions between the teacher and student.

In this paper, we design an innovative network where a teacher network learns and transfers multifarious and complementary features to train a lightweight yet competent student network. The design is based on the observation and intuition that different vision tasks are often correlated and complementary and more resourceful and knowledgeable teachers tend to train more competent students. Our proposed network learns in two phases: 1) knowledge capture; and 2) knowledge transfer as illustrated in Fig. 1. In the first phase, the teacher network is trained under two very different tasks to capture diverse and complementary features. Specifically, an auxiliary image reconstruction task is introduced with which the teacher network can capture structural knowledge and generative latent representations beyond the task-specific features. In the second phase, the student network is trained under the image classification task in a supervised manner. Concurrently, its learned features are modulated and enhanced by feature loss and adversarial loss that facilitate to thoroughly assimilate both pixel-level and spatial-level distributions of the complementary knowledge distilled from the teacher network. With the transferred multifarious features, our teacher can empower a more competent student network in a more efficient manner, more details to be described in Experiments.

The contributions of this work can be summarized from three aspects. First, it designs an innovative knowledge transfer network where a teacher learns and transfers multifarious features to train a lightweight yet competent student. Second, it proposes a novel knowledge transfer strategy where the student is capable of absorbing multifarious features effectively and efficiently under the guidance of feature loss and adversarial loss. Third, our developed network outperforms the state-of-the-art consistently across a number of datasets.

2 Related Work

Knowledge transfer aims to train a compact student network by transferring knowledge from a powerful teacher. Cristian *et al.* [1] first uses soft-labels for knowledge transfer, and this idea is further improved by knowledge distilling by adjusting the temperature of softmax activation function [5]. On the other hand, knowledge distilling relies on label categories and it works only for softmax function. This constraint is later addressed in different ways, e.g. by transferring intermediate features [10,14] or by optimizing the initial weight of student [4].

While the aforementioned methods obtain quite promising results, they train the teacher with a single task and objective and therefore can only transfer

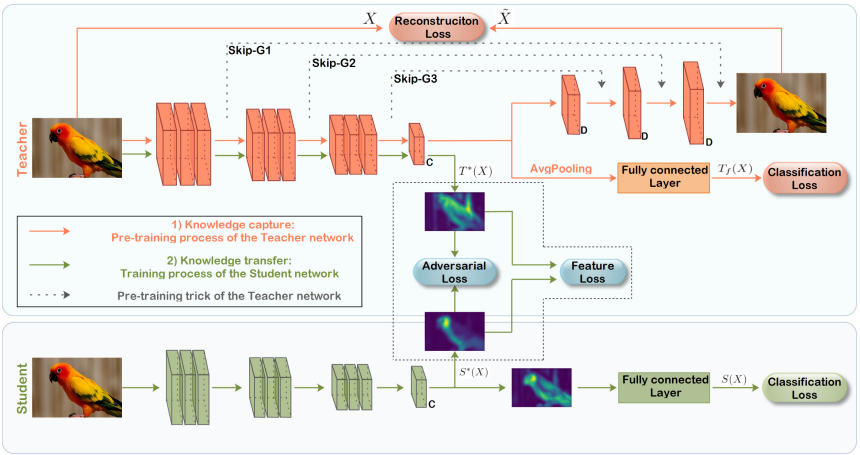

Fig. 1. Architecture of the proposed knowledge transfer network: 1) knowledge capture: a teacher network is first pre-trained with complementary objectives to learn multifarious features; 2) knowledge transfer: a student network is then trained to mimic both pixel-level and spatial-level distribution of the transferred features under the guidance of feature loss and adversarial loss, respectively. C and D denote the convolution operation and deconvolution module for feature alignment and image reconstruction.

task-specific unitary features. Our proposed method addresses this constraint by introducing a reconstruction task to the teacher network for learning and transferring the complementary and generative structural features beyond the task-specific features alone.

3 Proposed Methods

3.1 Learning Multifarious Features

Given a labeled dataset (X, Y), we first pre-train a teacher network T over the dataset for learning multifarious yet complementary features under a classification loss (CL) and a reconstruction loss (RL). The CL will drive T to learn discriminative classification features, whereas RL will drive T to learn generative reconstruction features, more details to be described in the following subsections.

Learning Discriminative Features: In the teacher network, we first include a convolution layer with batch normalization (denoted as 'C' in Fig. 1) for feature alignment. The convolution layer is followed by an averaged pooling and a fully connected layer that produces classification probabilities. Similar to the conventional metric in the classification task, we adopt the cross-entropy function E against labels Y for evaluating the classification result:

$$L_C^t = E(T_f(X), Y) \tag{1}$$

where $T_f(X)$ denotes the output of the fully connected layer and Y denotes the one-hot image-level label of X.

Learning Generative Features: Let $\tilde{X}$ be the reconstructed image by the teacher network that has the same size as the input image X. The RL can be formulated as follows:

$$L_R^t = f(\eta(\tilde{X}), \eta(X)) \tag{2}$$

where η denotes a normalizing operation (i.e. $\eta(\cdot) = \frac{\cdot}{\|\cdot\|_2}$) and f denotes a similarity distance metric function.

In our implemented system, we evaluate the image similarity by using the Kullback-Leibler (KL) divergence that measures how one probability distribution is different from another. Before computing the KL divergence, the cosine similarity of each normalized vector (denoted as $cos(\eta(.)))$ is first computed and the RL can then be evaluated as follows:

$$L_R^t = KL(cos(\eta(\tilde{X})), cos(\eta(X))) = -\frac{1}{n}\sum_{i=1}^{n} cos(\eta(\tilde{X}_i) \log(\frac{cos(\eta(X_i))}{cos(\eta(\tilde{X}_i))}) \tag{3}$$

Learning under the classification and reconstruction tasks alternately thus produces a resourceful and powerful teacher network, which is equipped with multifarious and complementary features for training a lightweight yet competent student network as to be described in the ensuing subsection.

3.2 Transferring Multifarious Features

Once the teacher network converges, we freeze its parameters and train the student network S to absorb the distilled knowledge that actually corresponds to the learned features before the fully connected layer of the teacher network. As illustrated in Fig. 1, the student network is trained with feature loss, adversarial loss and classification loss simultaneously.

For the feature loss, the transferred knowledge $T^*(X)$ from the teacher and the corresponding features $S^*(X)$ from the student are aligned and normalized (i.e. $\eta(\cdot) = \frac{\cdot}{\|\cdot\|_2}$) to calculate the feature metric as:

$$L_{Fea}^s = d(\eta(T^*(X)), \eta(S^*(X))) \tag{4}$$

Here, d can be evaluated by either L_1 or L_2 method to calculate the pixel-level absolute distance between features.

For the adversarial loss, a discriminator D is introduced to distinguish whether the input comes from teacher or student by maximizing the following objective:

$$L_D^s = \min_{S^*(X)} \max_{D} E_{S^*(X)\sim p_S}[log(1 - D(S^*(X)))] + E_{T^*(X)\sim p_T}[log(D(T^*(X)))] \tag{5}$$

where p_T and p_S correspond to the feature distribution of $T^*(X)$ and $S^*(X)$, respectively. Since the discriminator D is composed of fully connected layers with

Table 1. Comparison results of Top-1 mean classification error rate (%) with the unitary feature transferring methods on CIFAR10.

Student	Teacher	Student*	CL+RL					Teacher*
			W/o Skip	Skip-G1	Skip-G2	Skip-G3	Skip-G123	
ResNet20, 0.27M	ResNet56, 0.95M	7.18	6.24	5.92	6.22	6.13	**5.89**	5.78
ResNet20, 0.27M	WRN40-1, 0.66M	7.18	6.54	6.24	**6.10**	6.30	6.21	5.94
VGG13, 9.4M	WRN46-4, 11M	5.82	4.51	4.29	4.38	4.31	**4.21**	4.19
WRN16-1, 0.21M	WRN16-2, 0.97M	7.77	7.42	7.21	7.17	7.25	**7.15**	5.72
Student	Teacher	Student*	AT [11]	KD [5]	FT [6]	AB [4]	OFD [3]	Ours
ResNet20, 0.27M	ResNet56, 0.95M	7.18	7.13	7.19	6.85	6.49	6.32	**5.89**
ResNet20, 0.27M	WRN40-1, 0.66M	7.18	7.34	7.09	6.85	6.62	6.55	**6.10**
VGG13, 9.4M	WRN46-4, 11M	5.82	5.54	5.71	4.84	5.10	4.75	**4.21**
WRN16-1, 0.21M	WRN16-2, 0.97M	7.77	8.10	7.70	7.64	7.58	7.50	**7.15**

convolutional operations, adversarial loss can direct the student to assimilate and mimic the spatial-level relations in the transferred features.

The student network can thus be trained with the three losses as follows:

$$L_C^s = E(S(X), Y) \tag{6}$$

$$L^s = \alpha L_{Fea}^s + \beta L_D^s + L_C^s \tag{7}$$

Where α and β are balance weight parameters. During the student learning process, gradients are computed and propagated back within the student network, guiding it to learn the teacher's knowledge as defined in Eq. 7.

4 Experiments and Analysis

Our proposed network is evaluated over three datasets as follows: CIFAR10 [7] and CIFAR100 [8] are two publicly accessible datasets. They consist of 32×32 pixel RGB images that belong to 10 and 100 different classes, respectively. Both datasets have 50,000 training images and 10,000 test images. ImageNet refers to the large-scale LSVRC 2015 classification dataset, which consists of 1.2M training images and 50 K validation images of 1,000 object classes.

4.1 Implementation Details

During training process, SGD is employed as optimization and weight decay is set to 10^{-4}. On CIFAR dataset, the teacher network is pre-trained with 300 epoch. The learning rate of student drops from 0.1 to 0.01 at 50% training and to 0.001 at 75%. On ImageNet dataset, the student is trained for 100 epoch, with the initial learning rate 0.1 divided by 10 at the 30, 60 and 90 epoch, respectively.

Table 2. Comparison results with the adversarial learning based methods over CIFAR100 dataset.

Model	Top-1 error(%)
ResNet164,2.6M	27.76
ResNet20,0.26M	33.36
ANC [13]	32.45
TSCAN [18]	32.57
KSANC [12]	31.42
KTAN [9]	30.56
Ours	**29.28**

Table 3. Comparison results of Top-1 and Top-5 mean classification error (%) on ImageNet.

Student	Teacher	Test error(%)	
ResNet18	ResNet34	Top-1	Top-5
Student*		29.56	10.60
Teacher*		26.49	8.51
AT [11]		29.3	10.0
FT [6]		28.57	9.71
AB [6]		28.38	9.68
Ours		**28.08**	**9.49**

4.2 Comparisons with the State-of-the-Art

CIFAR10: Comparison results are shown in Table 1, where Student* and Teacher* provide Top-1 mean error rate of the student and teacher while trained from scratch. Two conclusions can be drawn: 1) In the top sub-table, the teacher pre-trained with skip connections 'Skip-G#' can empower the student to achieve the lowest classification error. It is attributed to the skip connection that can supplement the low-level information for the deconvolution modules, with which the teacher can extract and transfer more discriminative features to the student. 2) In the bottom sub-table, our proposed student network consistently outperforms both the original student network 'Student*' and the state-of-the-art results no matter whether the teacher and student networks are of different types. These outstanding performances are largely attributed to the fact that trained with different yet complementary tasks, our teacher network can effectively learn and transfer multifarious and complementary features to the student.

CIFAR100: To prove the generality of our technique, we compare it with the adversarial learning strategy applied methods on CIFAR100. This experiment involves ResNet164/ResNet20 with large depth gap to be teacher/student network pair. All the adversarial learning strategy applied methods obtain relatively good performance. Compared to the KTAN, our model makes noticeable performance with 1.28% improvement. It is largely due to our teacher which can learn multifarious knowledge by training with complementary tasks. As described above, our student acquires the lowest error with the same number of parameters, demonstrating that our model benefits from the multifarious knowledge learning method, as well as different level feature transferring strategy.

ImageNet: We also conduct a large-scale experiment over ImageNet LSVRC 2015 classification task to study its scalability. As results shown in Table 3, the proposed network outperforms the state-of-the-art methods consistently. In addition, our method helps improve the student's Top-1 accuracy by up to 1.48% as compared with the student trained from scratch in the Student* row. This

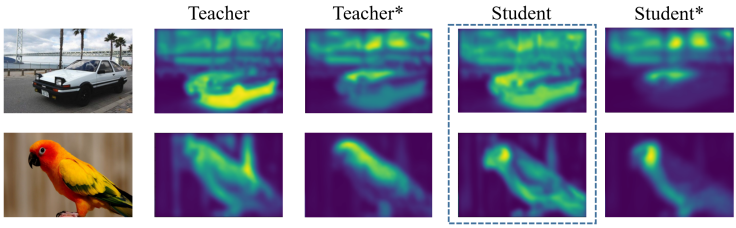

Fig. 2. Teacher and Teacher* columns represent the results from the teacher trained with both classification loss and reconstruction loss, or trained from scratch, respectively. Similarly, results in Student and Student* columns represent the outputs from the student network trained with our proposed teacher or trained from scratch.

Table 4. Ablation results of different transfer loss.

Transfer loss	Test Error(%)	
	CIFAR10	CIFAR100
L_C^s	7.18	31.04
$L_C^s + L_D^s$	6.42	29.62
$L_C^s + L_D^s + L_{L_2}^s$	6.17	28.97
$L_C^s + L_D^s + L_{L_1}^s$	**5.89**	**28.08**

clearly demonstrates the potential adaptability of our proposed method, making promising performance even on the more complex dataset.

4.3 Ablation Studies

Transfer Losses and Transfer Strategies: By comparing the first rows in Table 4, it indicates that adding adversarial loss L_D^s to absorb the shared features clearly improves the student's performance. This is largely attributed to the convolutional structure of the discriminator that can interpret the spatial information in features. In addition, by incorporating the feature loss to measure pixel-level distribution distance, either $L_{L_1}^s$ or $L_{L_2}^s$ shown in the last two rows, it can work as a complement to adversarial loss with distinct performance improvement. By using both adversarial loss and feature loss to capture different level distance between features, our student can assimilate the transferred multifarious features thoroughly with promising performance.

4.4 Discussion

Feature Visualization: As Fig. 2 shows, the teacher network 'Teacher' pretrained with 'CL+RL' focuses on more multifarious features, whereas the same network trained from scratch 'Teacher*' focuses on targeted features only (e.g. bird's beak), leading to the loss of rich contour details. Additionally, the fully

trained 'Student*' fails to learn the sufficient features for correct prediction, resulting in the sub-optimal performance. In contrast, the student network 'Student', under the guidance of the proposed 'Teacher', effectively pays attention to discrimitive and complementary regions (e.g. both bird's head and body parts), indicating and demonstrating the powerful performance of our proposed method.

5 Conclusion

This paper presents a novel knowledge transfer network for model compression in which the teacher can learn multifarious features for training a lightweight yet competent student. The learning consists of two stages, where the teacher is first trained with multiple objectives to learn complementary feature and the student is then trained to mimic both pixel-level and spatial-level feature distribution of the teacher. As evaluated over a number of public datasets, the proposed student network can learn richer and more useful features with better performance.

Acknowledgements. This work is supported in part by National Science Foundation of China under Grant No. 61572113, and the Fundamental Research Funds for the Central Universities under Grants No. XGBDFZ09.

References

1. Bucilua, C., Caruana, R., Niculescumizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 535–541 (2006)
2. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 597–613. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_36
3. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1921–1930 (2019)
4. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3779–3787 (2019)
5. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2014)
6. Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: network compression via factor transfer. In: Advances in Neural Information Processing Systems, pp. 2760–2769 (2018)
7. Krizhevsky, A., Nair, V., Hinton, G.: Cifar-10 dataset
8. Krizhevsky, A., Nair, V., Hinton, G.: Cifar-100 dataset
9. Liu, P., Liu, W., Ma, H., Mei, T., Seok, M.: Ktan: knowledge transfer adversarial network. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2018)
10. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Bengio, Y.: Fitnets: hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2015)

11. Sergey, Z., Nikos, K.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2017)
12. Shu, C., Li, P., Xie, Y., Qu, Y., Dai, L., Ma, L.: Knowledge squeezed adversarial network compression. arXiv preprint arXiv:1904.05100 (2019)
13. Vasileios, B., Azade, F., Fabio, G.: Adversarial network compression. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
14. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4133–4141 (2017)
15. Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M.: Classification-reconstruction learning for open-set recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4016–4025 (2019)
16. Zhang, X., Gong, H., Dai, X., Yang, F., Liu, N., Liu, M.: Understanding pictograph with facial features: end-to-end sentence-level lip reading of Chinese. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9211–9218 (2019)
17. Zhang, X., Lu, S., Gong, H., Luo, Z., Liu, M.: AMLN: adversarial-based mutual learning network for online knowledge distillation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 158–173. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_10
18. Zheng, X., Hsu, Y., Huang, J.: Training student networks for acceleration with conditional adversarial networks. In: BMVC (2018)