A Study of Self Distillation for Mango Image Classification

Wei-Chi Chen
National Cheng Kung University
Tainan, Taiwan
iphone31302@gmail.com

Wei-Ta Chu
National Cheng Kung University
Tainan, Taiwan
wtchu@gs.ncku.edu.tw

***Abstract*—We study a knowledge transfer approach called self distillation on a mango image dataset. Taking the deepest part of a convolutional neural network as the teacher, the self distillation approach transfers the relatively richer knowledge of the deepest part to shallow parts of this network, which are viewed as the students. We verify that this approach is effective in the target mango image dataset. Furthermore, we propose two more losses to improve performance considering data characteristics. In the discussion, we not only verify effectiveness of self distillation, but also point out weakness of the current approach, which unveils potential improvement for self distillation in the future.**

***Index Terms*—self distillation, model compression, image classification**

I. Introduction

Convolutional neural networks (CNN) have been widely adopted in image classification, semantic segmentation, object detection, and many other applications. Good performance mostly comes from over-parameterized networks trained on a large amount of training data. Despite of promising performance, training an over-parameterized network like ResNet [1] even on GPU devices is computationally expensive. Moreover, the space required to store millions of network parameters prevents deploying the constructed network into edge devices for real-world applications.

To reduce size of an over-parameterized network, model compression techniques like knowledge distillation [2] have been proposed. The idea is guiding the learning of a relatively lightweight network (called student network) by an over-parameterized network (called teacher network), such that the student network can achieve performance similar to the teacher network. Although knowledge distillation has been demonstrated as an effective model compression technique, the main barrier is how to design and train proper teacher networks. Existing works mainly assume that the teacher network has been well trained in advance, and put more focus on transferring knowledge from the teacher to the student.

In [3], Zhang et al. proposed the idea of self distillation. A CNN like ResNet-18 can be divided into several groups of residual blocks. The main idea is taking the deepest part as the teacher, which guide enhancing the shallow parts. Starting from a relatively simple model like ResNet-18, which is usually taken as a student network in other works, this method makes the student network learn from herself, without the need of a complex and computationally expensive teacher network. According to [3], the self distillation scheme largely reduces the training time (because no teacher network is required for training), and the self-distillated network performs well or even better than conventional knowledge distillation.

In this work, we investigate performance of self-distillated models on a mango image dataset. These mangos are classified into A, B, and C classes according to their quality. Based on the method proposed in [3], we design loss functions specially designed to the mango dataset, and investigate performance variations of the self distillation scheme.

The rest of this paper is organized as follows. Sec. II provides literature survey on knowledge distillation. Sec. III describes details of the self distillation scheme and our proposed improvement. Sec. IV presents experimental results, followed by the concluding remarks shown in Sec. V.

II. Related Works

The problem of highly-demanded resource, thus impeding model deployment in real-world applications, emerges as the rapid development of deep neural networks. How to reduce model size and required resource thus becomes urgent, and many model compression techniques have been proposed. In this work, we focus on one of the most popular approaches: knowledge distillation.

The idea of knowledge distillation is transferring knowledge of a complex neural network (usually called teacher network) into a simpler network (usually called student network), so that the student network can perform as well as the teacher network [2]. The number of parameters in the student network is smaller, and we attempt to achieve similar performance by a simpler model through knowledge distillation. Since the pioneering work [2], many variants have been proposed. Xu et al. [4] proposed to adopt a conditional generative adversarial network to evaluate the approximate error between the teacher network and the student network. Zhang et al. [5] proposed mutual learning between students. They first transferred knowledge from the teacher network to multiple students, and then transfer knowledge between student networks. Park et al. [6] proposed the concept of relational knowledge distillation. In addition to transferring knowledge from the teacher network to the student network, the relationship between results by the teach network should be similar to the the relationship between results by the student network. Mirzadeh et al. [7] proposed

that the performance gap between the teacher network and the student network can be reduced by introducing "teacher assistant" (TA) networks. The complexity of TA networks is in-between the teacher and the student. Knowledge of the teacher can be gradually transferred to TAs and then to the student, such that the student can work better. They conducted theoretical and empirical analysis about this simple idea.

Most works focus on how to transfer knowledge from a complex teacher network to a simple student network, by assuming the teacher network is available in advance. However, training a teacher network for a target domain is not trivial nor cost-effective. Zhang et al. [3] thus proposed the self distillation approach, where only a simple network is needed. The idea is that the deepest part of the simple network can be the teacher for the shallow parts. In this paper, we will investigate the self distillation approach on a mango image dataset. Considering data statistics, we propose two more losses and study performance variations.

III. Self Distillation

A. Network Architecture

Fig. 1 shows the network architecture of the self distillation approach [3] with slight modification by adding two more losses. We take ResNet-18 model [1] as the main instance to explain the self distillation method. This model can be divided into four sections according to residual blocks. The idea of self distillation is taking the deepest section (the 4th section) as the teacher to guide learning of the shallow sections (the 1st to the 3rd sections). The intuition behind this design is that the deepest section has the richest knowledge and is able to achieve the best classification result. The deepest section thus can be the teacher of the shallow sections.

To implement this idea, outputs of each section are connected with a fully-connected layer and a softmax layer, so that each section can be viewed as a classifier. The 1st to the 3rd classifiers can be trained as student models via distillation from the 4th classifier. The distillation can be guided from threefold.

- The classification results yielded by different sections should be similar. Cross entropy loss from labels are calculated to measure this, and thus the knowledge hidden in the dataset is introduced to not only the deepest section but also shallow sections.
- The KL divergence of softmax outputs between students and the teacher is calculated. Slightly different from knowledge from labels, the KL divergence directly measures similarity between softmax outputs.
- Different sections make their classification based on different levels of feature maps. These feature maps conceptually represent the same image, and implicit knowledge of the deepest feature maps can be introduced to improving feature extraction in shallow sections. To compare feature maps between different sections, outputs of residual blocks are in fact connected to a bottleneck layer, as shown in Fig. 1. The L2 losses between feature maps of the deepest section and each shallow section are calculated.

B. Loss Functions

Here we formally define the loss functions mentioned above. In addition, we consider the characteristics of the mango image dataset, and further design the triplet loss and the ordinal loss. Let $\Theta = \{\theta_{i/C}\}_{i=1}^{C}$ denote the classifiers in the target network, which is divided into C sections, and thus conceptually C classifiers are included. The softmax output of the ith classifier is denoted as q^i, and the softmax output of the deepest classifier is denoted as q^C. Given an input image x, the whole network finally outputs the predicted label $\hat{y}$ based on q^C.

The first item mentioned in Sec. III-A is mathematically defined as the summed cross entropy between the predicted label and the softmax outputs of shallow classifiers:

$$\mathcal{L}_c = \sum_{i=1}^{C} crossentropy(q^i, \hat{y}). \tag{1}$$

The second term is defined as the summed KL divergence between the softmax output of the Cth classifier and each shallow classifier:

$$\mathcal{L}_k = \sum_{i=1}^{C} KL(q^i, q^C). \tag{2}$$

Notice that the cross entropy in eqn. (1) is calculated between the softmax output and predicted labels, while the KL divergence in eqn. (2) is calculated between softmax outputs.

The third term is defined as the summed L2 distance between the deepest section and each shallow section:

$$\mathcal{L}_\ell = \sum_{i=1}^{C} \|F_i - F_C\|_2^2, \tag{3}$$

where F_i and F_C denote features (output by the bottleneck layer) fed to the classifier θ_i and θ_C, respectively.

In addition to the three losses proposed in [3], here we propose two more losses considering data characteristics.

Triplet loss. We propose to use the triplet loss in place of the cross entropy. Assume that two inputs x_p and x_q belong to the same class, while the input x_s belongs to the other class. The difference between outputs corresponding to x_p and x_q thus should be smaller than the difference between outputs corresponding to x_p and x_s. The triplet loss is defined as:

$$\mathcal{L}_t = \sum_{i=1}^{C} \max(\|F_i^{(p)} - F_i^{(q)}\|_2^2 - \|F_i^{(p)} - F_i^{(s)}\|_2^2 + \delta, 0), \tag{4}$$

where $F_i^{(p)}$ denotes the feature maps obtained by the ith classifier for the sample x_p. The term δ is the predefined margin parameter representing the difference between samples in different classes. The loss in eqn. (4) is smaller when the samples (x_p and x_q) in the same class yield similar feature maps ($F_i^{(p)}$ and $F_i^{(q)}$), and the samples (x_p and x_s) in different classes yield distinct feature maps ($F_i^{(p)}$ and $F_i^{(s)}$).

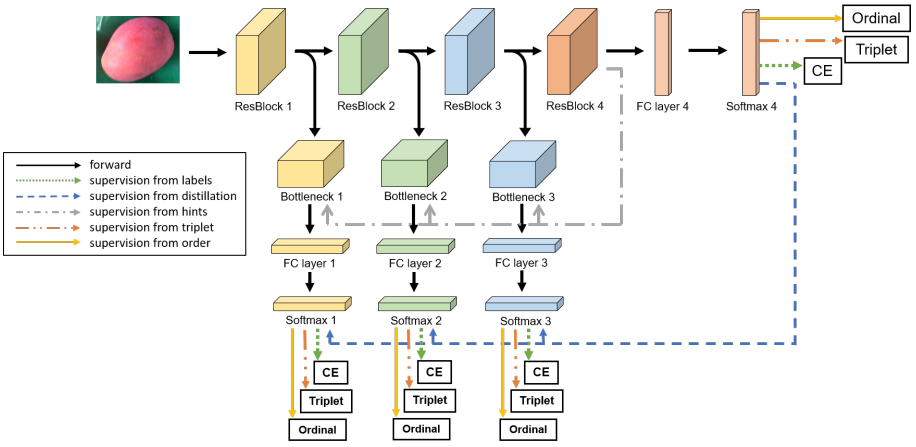

Fig. 1. Architecture of the self distillation approach.

Ordinal loss. We further propose the ordinal loss considering the characteristics of the mango image dataset. Fig. 2 shows sample images. Images from the top row to the bottom row are categorized into classes A, B, and C, respectively. Class A represents the best quality, in terms of color, shape, damage degree, etc. Class B represents the second best, while class C is the worst. Label of each image is manually defined by professional farmers.

Because of this ordinal characteristics, mis-classifying class-A images into class C (or vice versa) should be given higher penalty than mis-classifying class A into class B. Assume that three inputs x_a, x_b, and x_c belong to classes A, B, and C, respectively. The ordinal loss is defined as:

$$\mathcal{L}_o = \sum_{i=1}^{C} \max(\|F_i^{(a)} - F_i^{(b)}\|_2^2 - \|F_i^{(a)} - F_i^{(c)}\|_2^2 + \Delta, 0), \quad (5)$$

where $F_i^{(a)}$ denotes the feature maps obtained by the ith classifier for the sample x_a. The term Δ is the predefined margin parameter representing the difference between the A-B pairs and the A-C pairs.

The overall loss is the weighted sum of these losses:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_c + \alpha\mathcal{L}_k + \lambda\mathcal{L}_\ell + \beta\mathcal{L}_t + \gamma\mathcal{L}_o, \quad (6)$$

where the parameters α, λ, β, and γ are empirically set as 0.005, 0.01, 0.5, and 0.5, respectively. The margin parameters δ and Δ are both set as 2.0.

C. Training Details

Based on the losses mentioned above, the self-distillated ResNet-18 is trained based on the training data, with mini-batch size 256. Random horizontal flip, vertical flip, and random rotation are used for data augmentation. The network parameters are determined by the SGD optimizer. The initial learning rate is 0.01, with the weight decay parameter set as 1e-4. The momentum value is set as 0.9.

After training, each sub-classifier can be independently used to do classification. Following [3], we can also obtain an ensemble result by simply adding the weighted outputs of the softmax layer in each classifier.

Fig. 2. Sample mango images. Images from the top row to the bottom row are categorized into classes A, B, and C, respectively.

IV. Evaluation

A. Mango Image Dataset

Mango is one of the most famous and high-priced fruits exported from Taiwan. Different grades of mangos can be sold at highly-varied prices, and thus grading mangos before exporting and selling is important. The mango image dataset was collected at three fruit collection facilities in Fangshan, Pingtung, i.e., the southernmost county of Taiwan. Each mango was put on data collectors' hand or the conveyor. According to the quality of mango, in terms of color, shape, and damage degree, mangos are graded into class A (best), class B (medium), and class C (worst). This dataset totally consists of 52,000 mango images. Among them, 45,000 images are for training, and 7,000 images are for validation.

This dataset was created for a competition called AI CUP 2020 in Taiwan. To faithfully reflect the realistic situation farmers would face, these images were collected by low-cost consumer cameras. To quickly make a large-scale collection, the collector may take a video (rather than a still image

TABLE I
PERFORMANCE VARIATIONS ON MANGO IMAGE CLASSIFICATION.

Methods	Classifier 1/4	Classifier 2/4	Classifier 3/4	Classifier 4/4	Ensemble
Baseline ResNet-18			79.195		
Self-dis. ResNet-18 [3]	79.867	79.974	80.252	80.176	81.733
Self-dis. ResNet-18 + TO	80.233	80.362	<u>80.972</u>	<u>81.281</u>	**82.324**

for each mango) sequentially capturing a series of mangos. Selected screenshots of the video are included in the dataset. Fig. 2 shows some sample images. As can be seen, the mango images would suffer from motion blur, sensor noise, and luminance variant. These samples demonstrate the technical challenge of this task.

B. Experimental Results

Table I shows performance variations on mango image classification, in terms of classification accuracy. The first row (Baseline ResNet-18) shows performance of the ResNet-18 model trained from scratch. The second row (Self-dis. ResNet-18) shows performance of self-distillated ResNet-18, according to the design mentioned in [3]. Although the four sub-classifiers achieve slightly worse performance than a normal ResNet-18, the ensemble version outperforms the baseline (81.733 vs. 80.949). Two observations can be made:

- We verify that the self distillation approach outperforms the same network trained from scratch.
- In [3], they evaluated the self distillation approach based on ImageNet and CIFAR100 datasets. In that case, Classifier 3/4 and Classifier 4/4 already outperformed the baseline ResNet-18 on the CIFAR100 dataset, and Classifier 4/4 outperformed the baseline on the ImageNet dataset. However, it is not the case for the mango images. We think this may be due to the difference of data characteristics. Comparing with CIFAR100 and ImageNet, the task of classifying mango images is a fine-grained image classification problem. The result in the second row thus shows that the self distillation approach provides performance gain over baseline, but it also shows shortage of the current design.

The third row (Self-dis. ResNet-18 + TO) shows the self-distillated ResNet-18 when we further consider the triplet loss (T) and the ordinal loss (O). With the help of these losses, we see the ensemble version further improves the original self distillation method. Interestingly, we see that Classifier 3/4 and Classifier 4/4 have already surpassed the baseline ResNet-18 in this case. This shows effectiveness of the proposed two losses, and also unveils the potential of further improvement for fine-grained classification in the future.

Fig. 3 shows confusion matrices of the basic self distillation approach and the proposed improvement. Classifying class B and class C is relatively more difficult.

V. CONCLUSION

We have verified the effectiveness of self distillation on a fine-grained image classification dataset, i.e., mango images.

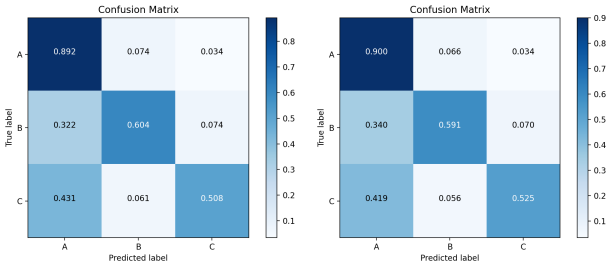

Fig. 3. Confusion matrices of the basic self distillation approach (left) and the improved one (right).

The main idea of self distillation is taking the deepest part of a network as the teacher, which transfers relatively richer knowledge to shallow parts of the network, i.e., the students. Considering data characteristics, we propose two more losses to further improve performance. We not only verify generality of the self distillation approach, but also point out its weakness. In the future, more knowledge transfer techniques in the self distillation paradigm can be studied, and generality of related methods can be investigated.

ACKNOWLEDGMENT

This work was partially supported by Qualcomm Technologies, Inc. under the grant number B109-K027D, and by the Ministry of Science and Technology, Taiwan, under the grant 108-2221-E-006-227-MY3, 107-2923-E-194-003-MY3, and 109-2218-E-002-015.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proceedings of Advances in Neural Information Processing Systems*, 2014.

[3] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of International Conference on Computer Vision*, 2019.

[4] Z. Xu, Y.-C. Hsu, and J. Huang, "Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks," in *Proceedings of International Conference on Learning Representations Workshop*, 2018.

[5] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2018.

[6] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2019.

[7] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of The AAAI Conference on Artificial Intelligence*, 2020.