Weighted Mutual Learning with Diversity-Driven Model Compression

Aalborg University, Denmark.
University of New South Wales, Australia.
East China Normal University, China.
Shandong First Medical University, China.
Miao Zhang, Li Wang, David Campos, Wei Huang, Chenjuan Guo, Bin Yang

PosterID:

I. Introduction

1.1. Background:

- Knowledge Distillation distils knowledge from a teacher model to help the training of a student model, by steering the student's logits towards teacher's logits. Online distillation enables peers learn from each other.
- Network Pruning, including unstructured and structured pruning, compresses networks by removing parameters with minimal performance degradation.

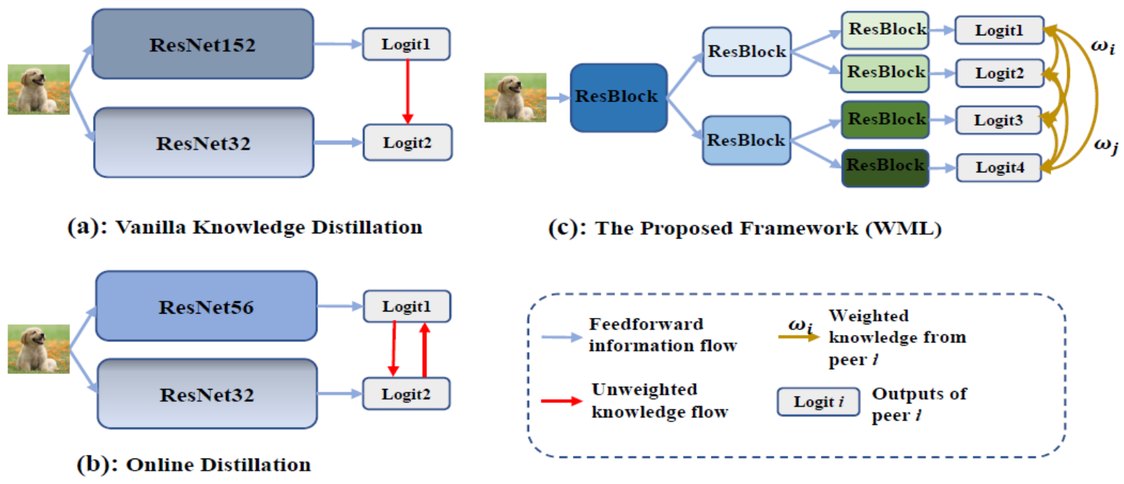

Fig.1 Pictorial depiction of Knowledge distillation, online distillation, and the proposed framework (WML).

1.2. Contributions:

- Proposes a Weighted Mutual Learning with Diversity-Driven Model Compression (WML) for online distillation, where a hierarchical structure and the structured network pruning are leveraged to enhance diversify and reduce the memory consumption.
- Formulates the weighted mutual learning based online distillation as a bilevel optimization problem, and the hypergradient of optimizing the weights is derived with a close-form.

II. Solution Framework

2.1. Overview Structure and Loss Function

Fig. 1 (c) depicts the overall structure of the proposed WML framework, which is similar to the binary tree that the branches at the same levels are copies of each other, and each path is considered as a peer model. WML directly learns from each other, and considers the linear combination of knowledge from all peers with overall loss:

$$loss = (1-\alpha)\sum_{i=1}^{M}\omega_i\mathcal{L}_{CE}(z_i,Y) + \alpha\sum_{i=1}^{M}\sum_{j=1}^{M}\omega_j KL(z_i,z_j), \qquad (1)$$

2.2. Diversifying Peer Models with Channel Pruning

To further reduce the memory requirements, we introduce the channel pruning, the most commonly-used paradigm in structured pruning, to our framework, which also encourage the diversity among homogenized peers. In this paper, we propose a relative importance, $\mathcal{I}_l = \frac{1}{N_l}\sum_{f\in l}\mathcal{I}_f$, to determine importance of each layer at initialization.

2.3. Weighted Distillation with Bi-level Formulation

We consider a dynamic weighting method to determine the importance of each peer, and the weighted mutual learning for online distillation can be formulated as a bilevel optimization problem.

$$\min_{\omega} \ \mathcal{L}_{CE}(\sum_{i=1}^{M}\omega_i z_i^*, Y) \quad \text{s.t. } \theta^* = \text{argmin}_\theta\,(1-\alpha)\sum_{i=1}^{M}\omega_i\mathcal{L}_{CE}(z_i,Y) + \alpha\sum_{i=1}^{M}\sum_{j=1}^{M}\omega_j KL(z_i,z_j),$$

where the gradient for the outer loop is defined as:

$$\nabla_\omega\mathcal{L}_2 = \frac{\partial\mathcal{L}_2}{\partial\omega} + \frac{\partial\mathcal{L}_2}{\partial\theta^*}\frac{\partial\theta^*(\omega)}{\partial\omega} = \frac{\partial\mathcal{L}_2}{\partial\omega} - \gamma\frac{\partial\mathcal{L}_2}{\partial\theta}\frac{\partial^2\mathcal{L}_1}{\partial\omega\partial\theta}.$$

The following **Theorem 1** shows that the second-order term for the outer loop gradient can be analyzed in a closed-form without Taylor expansion approximation when L_1 is with a specific structure.

Theorem 1 *With one-step unroll learning paradigm, the gradeint for ω_i in Eq.(3) is formulated as:*

$$g_{\omega_i} = \nabla_{\omega_i}\mathcal{L}_2 = \frac{\partial\mathcal{L}_2}{\partial\omega_i} - \gamma\frac{\partial\mathcal{L}_2}{\partial\theta}\frac{\partial^2\mathcal{L}_1}{\partial\omega_i\partial\theta} = \frac{\partial\mathcal{L}_2}{\partial\omega_i} - \gamma\frac{\partial\mathcal{L}_2}{\partial\theta}\frac{\partial\mathcal{L}_a}{\partial\theta}^{\mathsf{T}}, \qquad (5)$$

where $\mathcal{L}_a = (1-\alpha)\mathcal{L}_{CE}(z_i,Y) + \alpha\sum_{j=1}^{M}KL(z_j,z_i)$.

Algorithm 1 Weighted Mutual Learning (WML) for Online Distillation

1: **Input**: Dataset $\{(x_n,y_n)\}_{n=1}^{N}$; Given pruning ratios for each peer model $\{p_1,...,p_M\}$.
2: Initialized hierarchical model θ^0 and peer weights ω^0;
3: Calculate the filter importance with SNIP at initialization, and prune peer models i based on the given pruning ratios;
4: **for** $k = 1,...,K$ **do**
5: With the peer importance ω^k, run T steps of SGD to update the model parameters θ with the weighted loss function in Eq.(1);
6: Calculate the gradient for ω^k based on Eq.(5);
7: Run one step of mirror descent and update ω^k to get ω^{k+1} based on Eq.(7);
8: **end for**
9: **output:** M models with outputs $\{z_1,...,z_M\}$ the weights for peers ω.

III. Results

3.1. Comparison with Self-Distillation

For fair comparisons, we consider 4 peer models in our WML, and also prune peer models into similar model size as classifiers in self-distillation.

Table 1: Top-1 accuracy (%) comparison results with self-distillation on CIFAR100.

Networks	Methods	Baseline	Model1	Model2	Model3	Model4	Ensemble
ResNet18	DSN [28]	77.09	67.23	73.80	77.75	78.38	79.67
	SD [59]	77.09	67.85	74.57	78.23	78.64	79.67
	SCAN [60]	77.09	**71.84**	**77.74**	78.62	79.13	80.46
	WML	77.09	71.15	75.65	**78.88**	**79.38**	**80.56**
ResNet50	DSN [28]	77.68	67.87	73.80	74.54	80.27	80.67
	SD [59]	77.68	68.23	74.21	75.23	80.56	81.04
	SCAN [60]	77.68	73.69	78.34	80.39	80.45	81.78
	WML	77.68	**78.58**	**79.69**	**80.81**	**81.24**	**82.57**
ResNet101	DSN [28]	77.98	68.17	75.43	80.98	81.01	81.72
	SD [59]	77.98	69.45	77.29	**81.17**	81.23	82.03
	SCAN [60]	77.98	72.26	79.26	80.95	81.12	82.06
	WML	77.98	**79.60**	**81.16**	81.14	**81.46**	**83.03**

3.2. Comparison with Online Distillation

Table 3 compares the proposed WML with several online distillation methods, including RKD, CTSL-MKT, DML, ONE, and self-distillation (SD) on CIFAR10, CIFAR100, and ImageNet.

Table 3: Top-1 accuracy (%) comparison results with online-distillation.

Datasets	Models	Baseline	KD[25]	RKD[40]	MKT[40]	DML[63]	ONE[64]	SD[59]	**WML**
CIFAR10	ResNet18	94.25	94.67	94.98	95.33	95.19	95.56	95.87	**95.97**
	ResNet50	94.69	94.56	95.23	95.76	95.73	95.85	96.01	**96.17**
CIFAR100	ResNet18	77.09	77.79	76.43	77.46	77.54	77.87	78.64	**79.38**
	ResNet50	77.42	79.33	78.02	78.52	78.31	78.52	80.56	**81.24**
ImageNet	MobileNetV2	71.52	72.23	-	72.46	72.29	72.20	72.37	**72.76**

3.3. Comparison with Channel Pruning Approaches

we compared the performance of the pruned models generated by our WML with existing channel pruning methods.

Table 4: Comparison with channel pruning methods for ResNet-32 and ResNet-56 on CIFAR10

Models	Method	Baseline Acc	Pruned Acc	Acc Drop↓	FLOPs Drop↓
ResNet32	LCCL [8]	92.33%	90.74%	1.59%	31.2%
	SFP [22]	92.63%	92.08%	0.55%	41.5%
	GFP40 [37]	93.51%	92.86%	0.65%	40.1%
	GFP50 [37]	93.51%	92.64%	0.87%	50.1%
	PScratch [50]	93.18%	92.18%	1.00%	50.0%
	FPGM [24]	92.63%	92.31%	0.32%	41.5%
	WML30	92.63%	92.99%	**-0.36%**	24.4%
	WML40	92.63%	92.23%	**0.40%**	45.4%
	WML50	92.63%	92.10%	**0.53%**	50.0%
ResNet56	PFEC [31]	93.04%	93.06%	-0.02%	27.6%
	LCCL [8]	94.35%	92.81%	1.54%	37.9%
	HRank [34]	93.26%	93.52%	-0.26%	29.3%
	NISP [56]	93.26%	93.01%	0.25%	35.5%
	AMC [23]	92.80%	91.90%	0.90%	50.0%
	GFP40 [37]	93.68%	93.54%	0.14%	40.2%
	GAL [35]	93.26%	93.38 %	-0.12%	37.6%
	WML30	93.26%	93.93%	**-0.67%**	24.7%
	WML40	93.26%	93.46%	**-0.20%**	41.7%
	WML50	93.26%	92.68%	**0.58%**	49.0%

3.4. Ablation studies

we first conduct ablation studies to investigate whether the dynamic weighting strategy and peer pruning help improve the performance of online distillation, and then show the effectiveness of the weighted ensemble for WML.

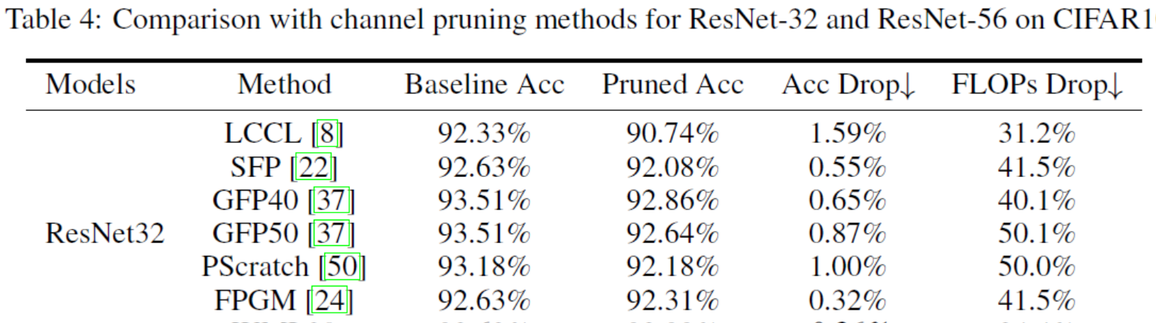

Figure 2: Ablation studies on pruning ratio, step size η and peer importance.

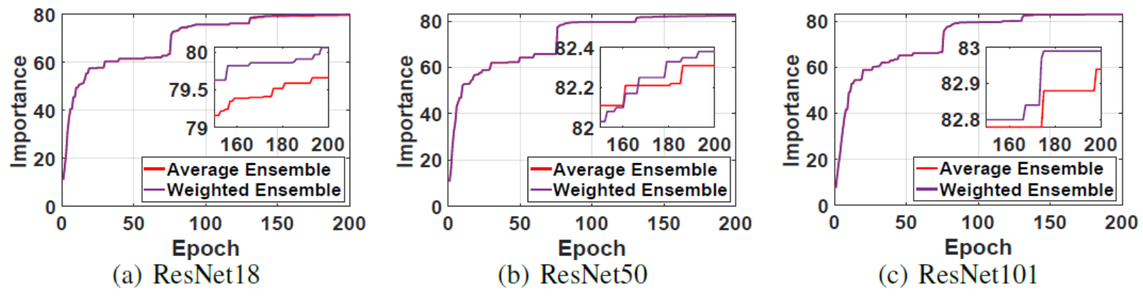

Figure 5: Track the average-ensemble and weighted-ensemble for WML.