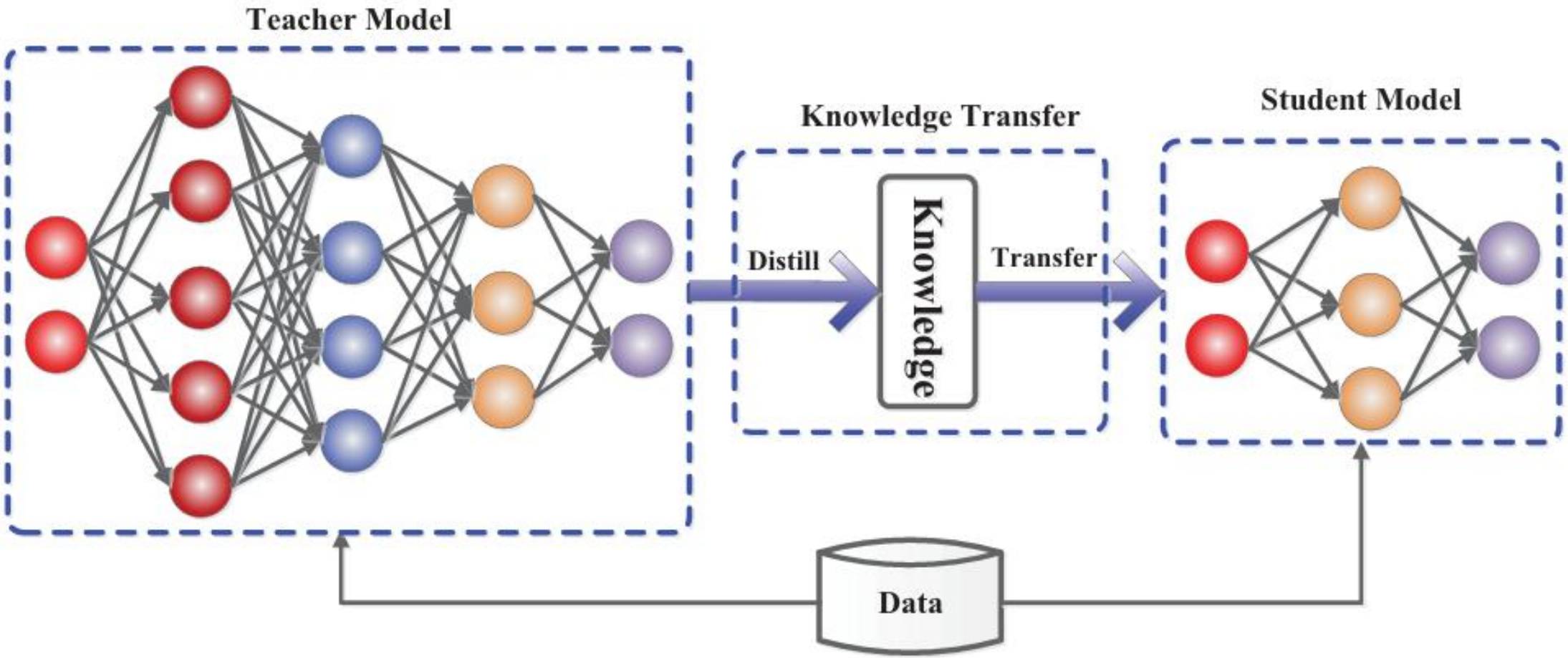
Online Knowledge Distillation with Diverse Peers

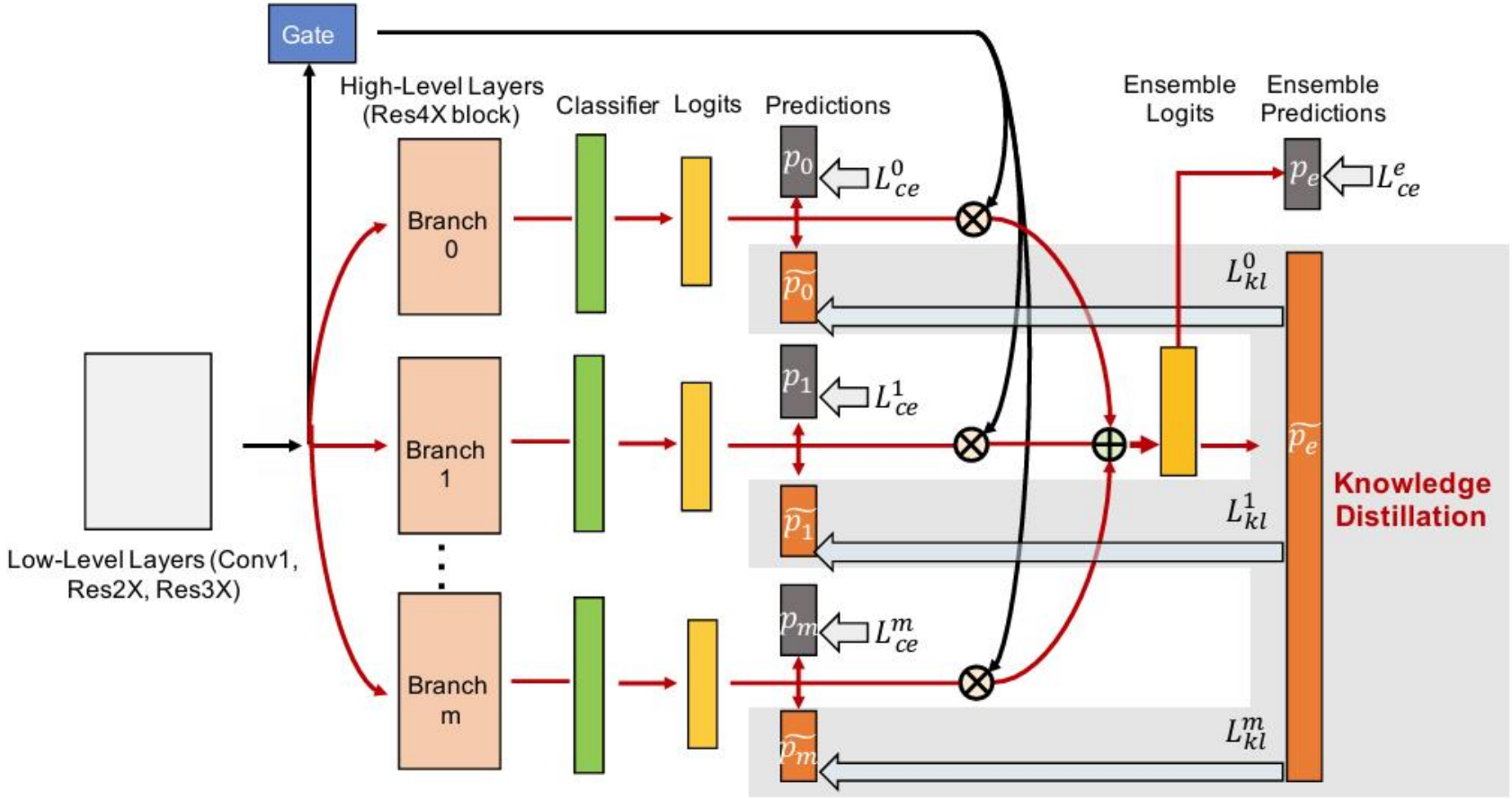
Defang Chen,[1,2] Jian-Ping Mei[3*]
Can Wang,[1,2] Yan Feng,[1,2] Chun Chen[1,2]
[1]College of Computer Science, Zhejiang University, Hang Zhou, China.
[2]ZJU-LianlianPay Joint Research Center.
[3]College of Computer Science, Zhejiang University of Technology, Hang Zhou, China.
defchern@zju.edu.cn, jpmei@zjut.edu.cn, {wcan, fengyan, chenc}@zju.edu.cn

AAAI 2020

Knowledge Distillation

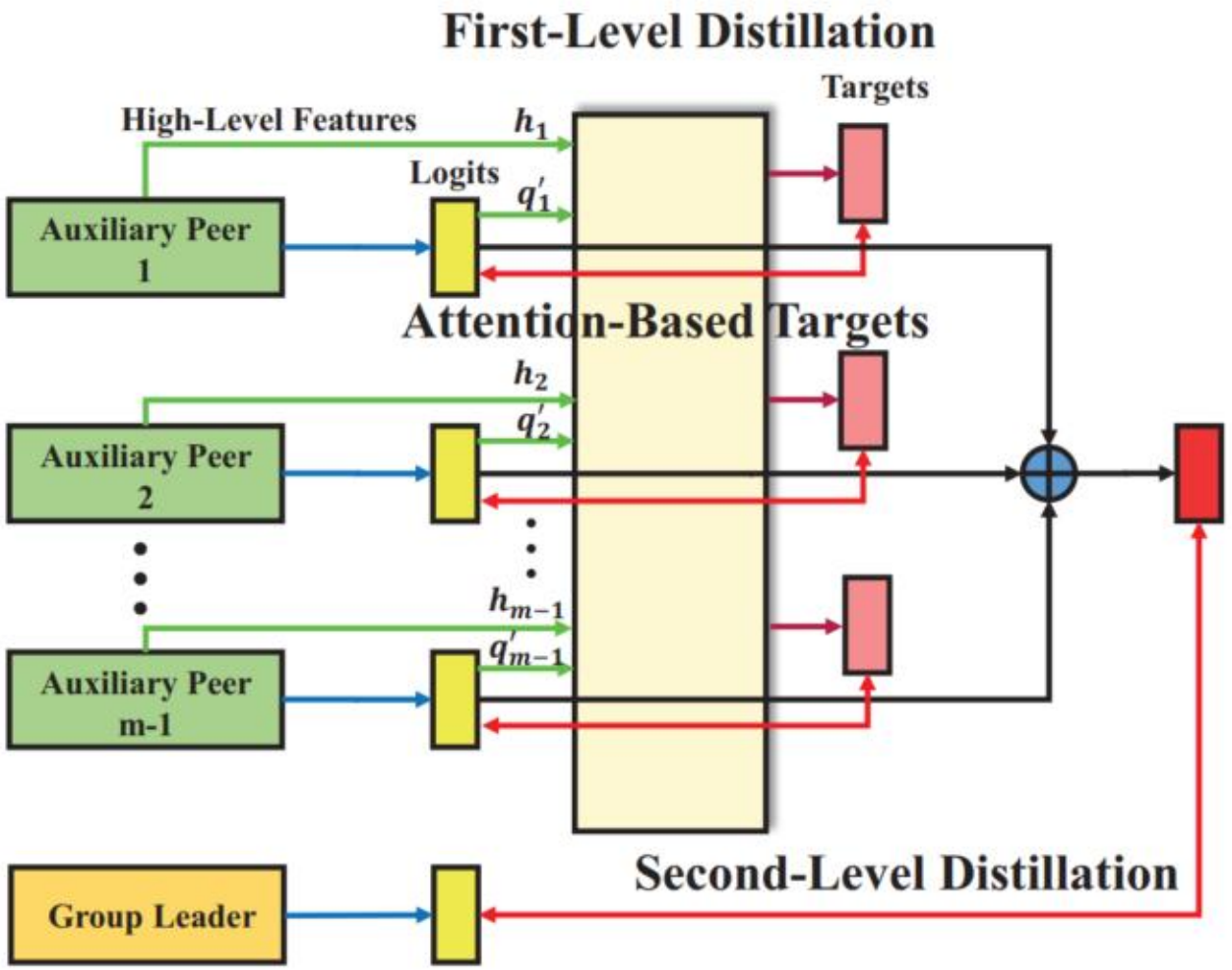


Online Knowledge Distillation[1]



由于Peer预测质量不同，平等地对待所有Peer是不合理的，简单的聚合函数会导致Peer同质化，进而影响组蒸馏效果

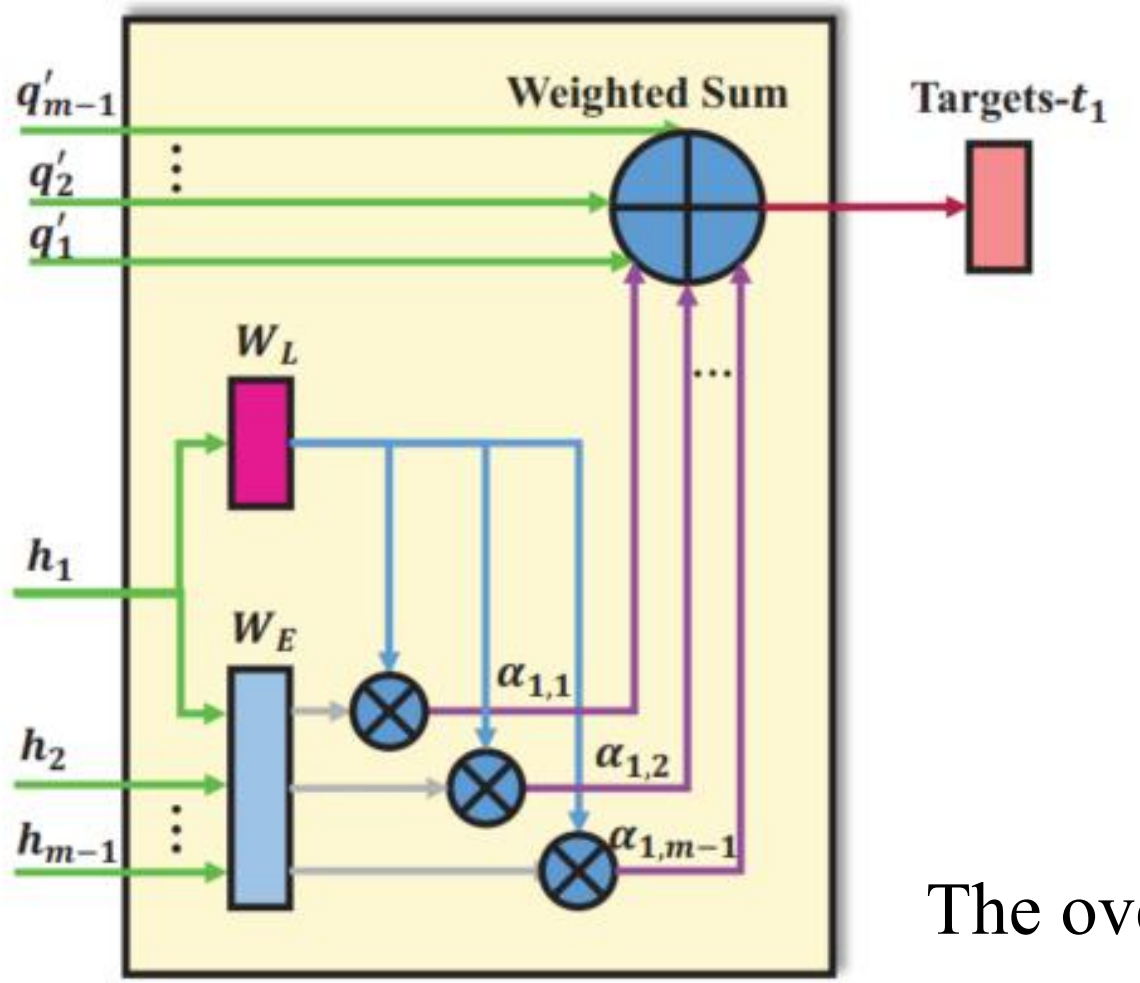
Methods



$$t_a = \sum_{b=1}^{m-1} \alpha_{ab} \cdot q'_b, \quad \text{其中} \sum_b \alpha_{ab} = 1$$

$$\mathcal{L}_{dis1} = \sum_{a=1}^{m-1} KL(t_a, q'_a),$$

Methods



Attention-Based Targets

$$L(\boldsymbol{h}_a) = \mathbf{W}_L^T \boldsymbol{h}_a \quad and \quad E(\boldsymbol{h}_a) = \mathbf{W}_E^T \boldsymbol{h}_a,$$

$$\alpha_{ab} = \frac{e^{L(\boldsymbol{h}_a)^T E(\boldsymbol{h}_b)}}{\sum_{f=1}^{m-1} e^{L(\boldsymbol{h}_a)^T E(\boldsymbol{h}_f)}}.$$

$$\mathcal{L}_{dis2} = KL(\boldsymbol{t}_m, \boldsymbol{q}'_m)$$

The overall loss：

$$\mathcal{L}_{OKDDip} = \sum_{a=1}^{m} \mathcal{L}_{gt}(a) + T^2 \mathcal{L}_{dis1} + T^2 \mathcal{L}_{dis2}$$

Experiment

Table 1: Error rates (Top-1, %) on CIFAR-10. OKDDip: network-based (1st column) and branch-based (2nd column).

Network	Baseline	Ind	DML	CL-ILR	ONE	OKDDip	
DenseNet-40-12	6.87 ± 0.02	6.97 ± 0.03	6.50 ± 0.02	7.02 ± 0.08	6.85 ± 0.15	**5.94 ± 0.05**	6.48 ± 0.12
ResNet-32	6.34 ± 0.03	5.99 ± 0.15	6.18 ± 0.05	6.06 ± 0.07	5.94 ± 0.06	5.62 ± 0.07	**5.58 ± 0.08**
VGG-16	6.12 ± 0.15	6.03 ± 0.01	5.94 ± 0.04	6.22 ± 0.10	6.16 ± 0.08	**5.88 ± 0.04**	**5.87 ± 0.03**
ResNet-110	5.46 ± 0.02	4.95 ± 0.02	5.68 ± 0.03	4.88 ± 0.12	5.02 ± 0.04	**4.54 ± 0.07**	**4.56 ± 0.11**
WRN-20-8	5.27 ± 0.06	5.35 ± 0.02	5.04 ± 0.08	5.12 ± 0.16	5.29 ± 0.02	**4.84 ± 0.07**	5.06 ± 0.04

Table 2: Error rates (Top-1, %) on CIFAR-100. OKDDip: network-based (1st column) and branch-based (2nd column).

Network	Baseline	Ind	DML	CL-ILR	ONE	OKDDip	
DenseNet-40-12	28.97 ± 0.23	29.20 ± 0.09	26.64 ± 0.17	28.61 ± 0.12	28.76 ± 0.18	**26.10 ± 0.03**	28.34 ± 0.02
ResNet-32	28.76 ± 0.08	27.84 ± 0.05	26.47 ± 0.26	27.44 ± 0.05	26.50 ± 0.13	**25.40 ± 0.08**	25.63 ± 0.14
VGG-16	26.19 ± 0.12	25.81 ± 0.18	25.33 ± 0.03	25.62 ± 0.11	25.63 ± 0.03	**24.88 ± 0.06**	25.15 ± 0.19
ResNet-110	24.12 ± 0.20	23.54 ± 0.15	22.50 ± 0.11	21.56 ± 0.09	21.67 ± 0.12	**21.09 ± 0.17**	21.14 ± 0.14
WRN-20-8	22.50 ± 0.44	21.85 ± 0.12	20.21 ± 0.11	20.44 ± 0.13	21.19 ± 0.12	**19.63 ± 0.07**	20.06 ± 0.05

Experiment

Table 3: Error rates (Top-1, %) for ResNet-34 on ImageNet-2012. OKDDip: network-based (1st column) and branch-based (2nd column).

Baseline	DML	CL-ILR	ONE	OKDDip	
26.76	26.03	26.06	25.92	**25.42**	25.60

Table 4: Error rates (Top-1, %) of ensemble predictions with branch-based student models on CIFAR-100.

Network	CL-ILR	ONE	OKDDip	Ind
VGG-16	25.56	25.54	**24.95**	25.62
ResNet-32	27.01	24.90	**23.45**	23.74
ResNet-110	20.19	20.14	**19.54**	20.18

Experiment

(1) w/o SA (random). A random attention matrix with normalization is used to put randomly generated belief among peers.

(2) w/o SA (entropy). we completely remove Ldis1 from objective function and let the peers only learn from Lgt with an entropy term.

(3) w/o SA (mean). Simple average is applied to aggregate the predictions of peers in the first-level distillation.

(4) w/o SA (asymmetry). Another special case that ablates asymmetry of SA by forcing WE and WL as identity matrices.

(5) w/o two-level. The second-level distillation is ablated by removing the group leader from training and inference with a randomly chosen student model.

Table 5: Ablation study: Error rates (Top-1, %) for ResNet-32 on CIFAR-100

w/o SA (random)	w/o SA (entropy)	w/o SA (mean)	w/o SA (asymmetry)	w/o two-level	OKDDip
28.24 ± 0.16	26.71 ± 0.19	26.35 ± 0.14	26.05 ± 0.17	27.79 ± 0.14	$\mathbf{25.63 \pm 0.14}$

Experiment

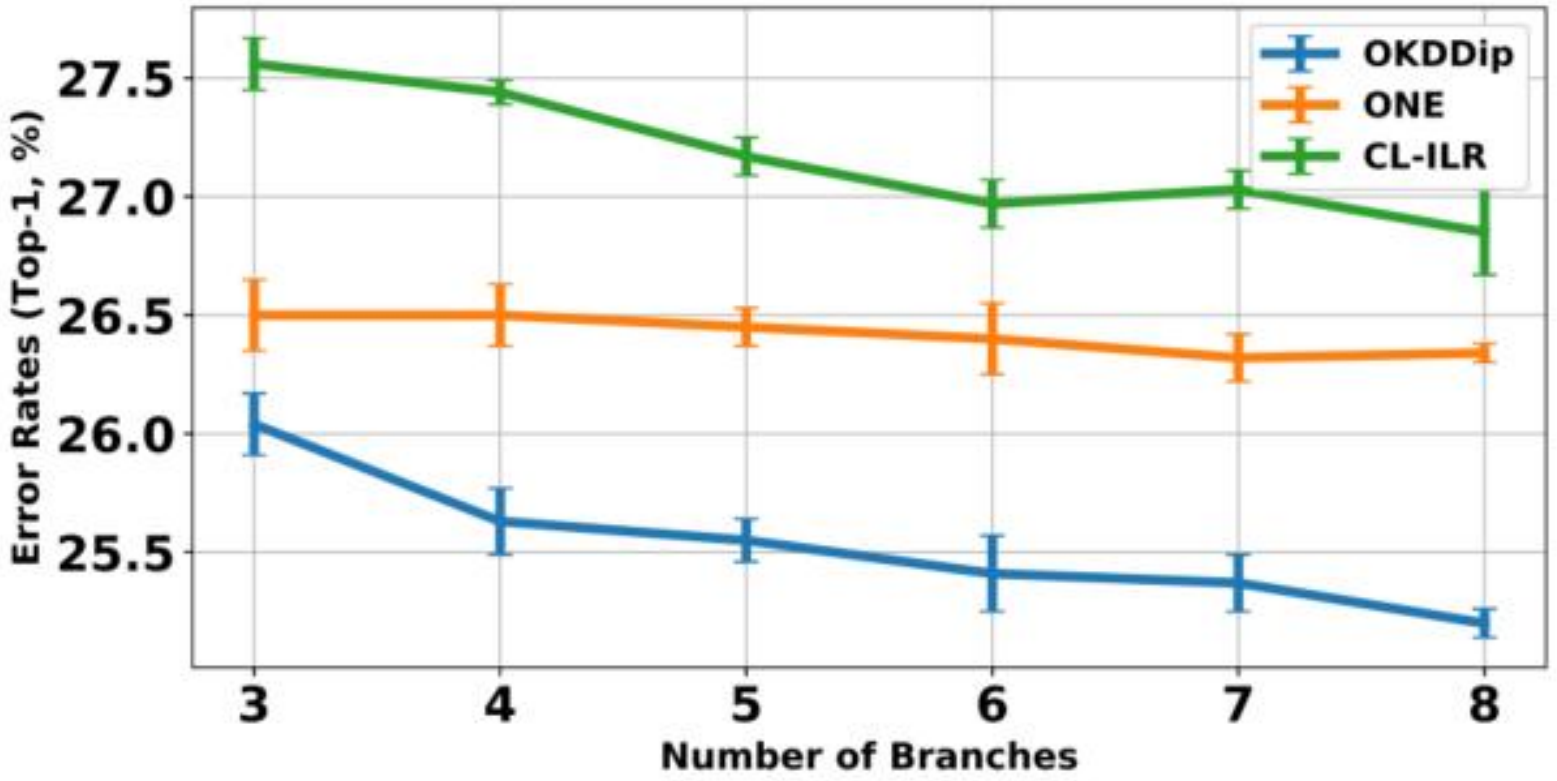


Figure 3: Impact of group size with branch-based ResNet-32 on CIFAR-100.

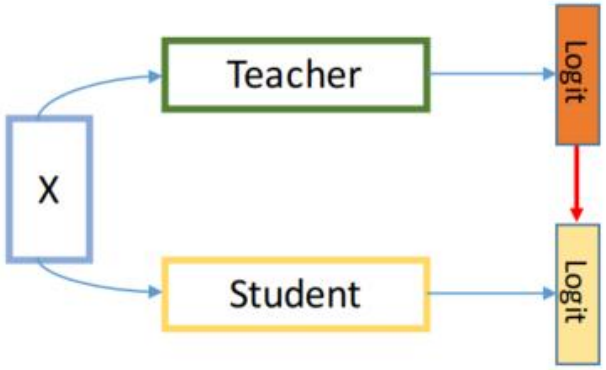
Online Knowledge Distillation via Collaborative Learning

Qiushan Guo[1], Xinjiang Wang[2], Yichao Wu[2], Zhipeng Yu[2], Ding Liang[2], Xiaolin Hu[3], Ping Luo[4]
[1]Beijing University of Posts and Telecommunications [2]SenseTime Group Limited
[3]Tsinghua University [4]The University of Hong Kong
qsguo@bupt.edu.cn xlhu@mail.tsinghua.edu.cn pluo@cs.hku.hk
{wangxinjiang,wuyichao,yuzhipeng,liangding}@sensetime.com

CVPR 2020

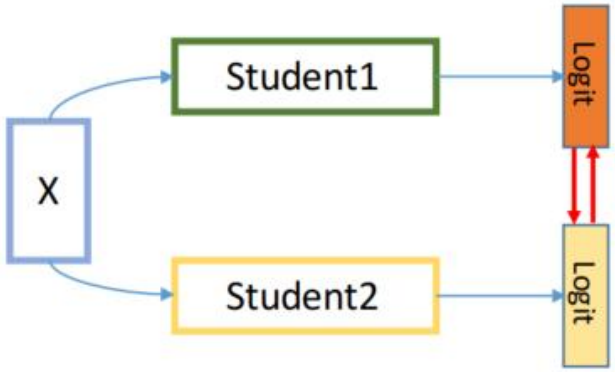
Motivation

The training cost is high.

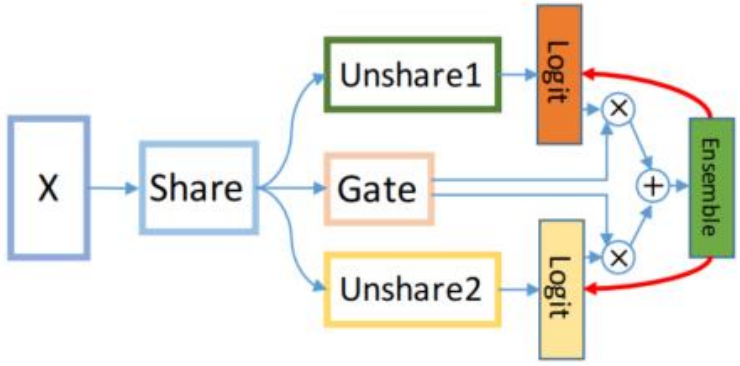


(a) Baseline

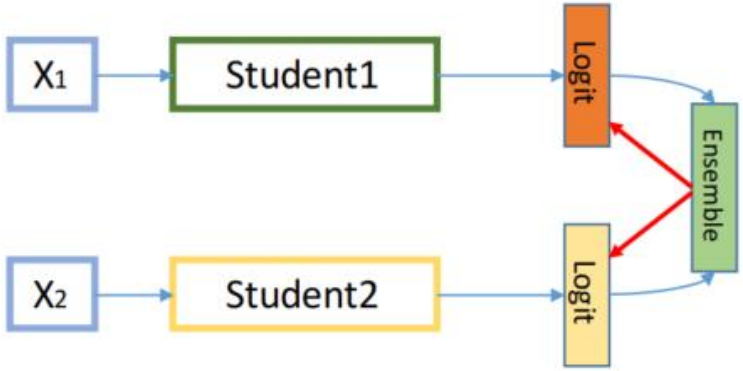
The outputs of the student models may conflict with each other.



(b) DML



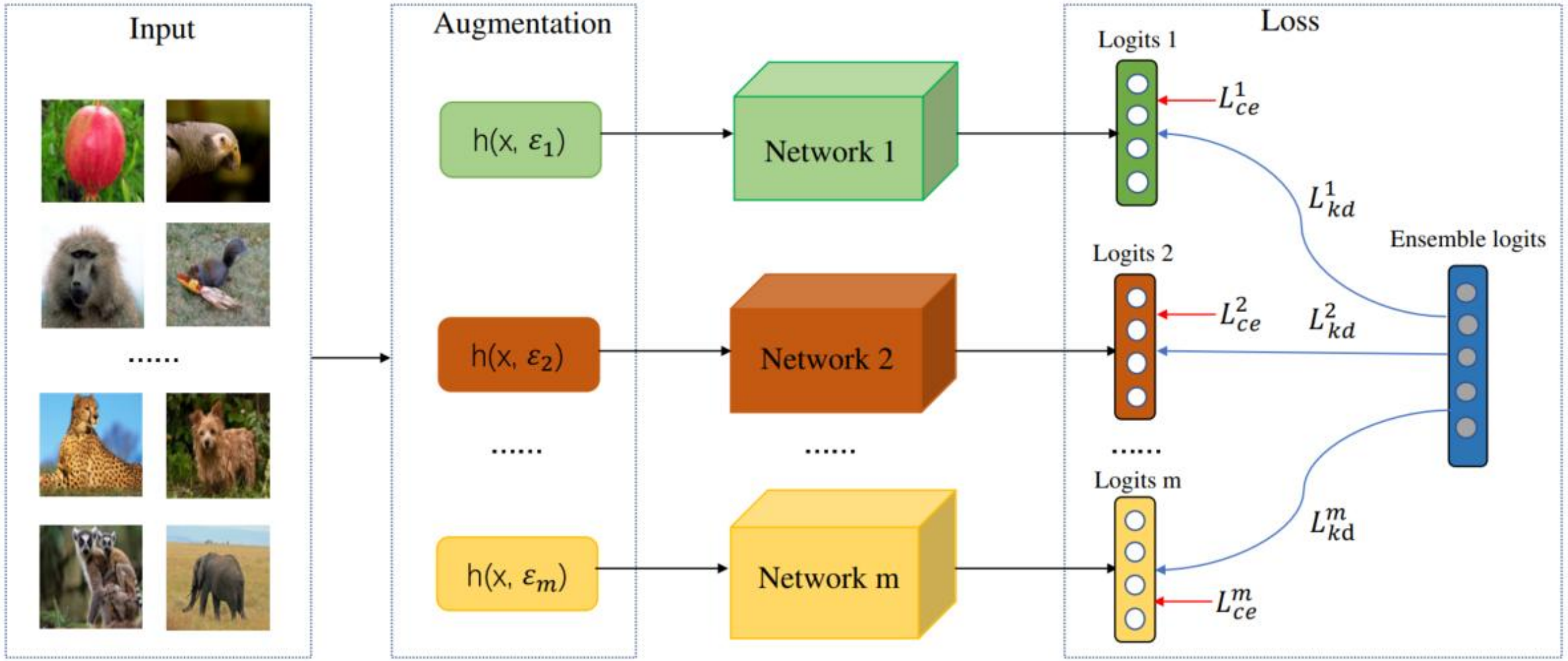
(c) ONE



(d) KDCL

Lack of flexibility due to shared underlying network.

Methods



$$L = \sum_{i=1}^{m} L_{CE}^i + \lambda L_{KD}^i$$

Methods

KDCL-Naive

Find the student output with the minimum loss compared to the one-hot label:

$$\mathbf{z}_t = \mathbf{z}_k, k = \arg\min_i L_{CE}(\mathbf{z}_i, \mathbf{y}),$$

KDCL-Linear

The soft labels obtained using the above method may be of poor quality. Therefore, consider performing a weighted sum of all student outputs to minimize the loss with respect to the hard labels:

$$\min_{\alpha \in \mathbb{R}^m} L_{CE}(\alpha^T \mathbf{Z}, \mathbf{y}), \text{subject to} \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0$$

Methods

KDCL-MinLogit

KDCL-Linear introduces an optimization problem during training, and we hope the network ensemble is efficient.

The differences between logit values determine the probability distribution from the softmax function. Thus, the output probabilities are represented as:

$$\mathbf{p} = softmax(\mathbf{z}) = softmax(\mathbf{z} - z^c)$$

For all student networks, the c-th element of output is set to 0. When the other elements in the logit decrease, the CE loss with the one-hot label will be reduced.

A concise way to generate teacher logits is to select the minimum element in each row of the matrix.

KDCL-General

To assess the generalization quality, it's necessary to have a validation set as a metric. Therefore, a portion of the training set is first selected to serve as the validation set.

The objective is to weight the logits of the student network on validation set, with the requirement that $w_i \in [0,1]\ (i = 1,2,...,m)$ and $\sum_{i=1}^{m} w_i = 1$, such that:

$$E(\mathbf{x}) = (f(\mathbf{x}) - t)^2$$

$$E = \int \left(\sum_{i=1}^{m} w_i f_i(\mathbf{x}) - t \right)^2 p(\mathbf{x}) d\mathbf{x}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j C_{ij},$$

$$C_{ij} = \int (f_i(\mathbf{x}) - t)\,(f_j(\mathbf{x}) - t)\, p(\mathbf{x}) d\mathbf{x}$$

$$\approx \frac{1}{N} \sum_{k=1}^{N} (f_i(\mathbf{x}_k) - t)(f_j(\mathbf{x}_k) - t)$$

$$w_k = \frac{\sum_{j=1}^{m} C_{kj}^{-1}}{\sum_{i=1}^{m} \sum_{j=1}^{m} C_{ij}^{-1}}$$

Experiment

Method	ResNet-50	ResNet-18	Gain
Vanilla	76.8	71.2	0
KD[10]	76.8	72.1	0.9
DML[32]	75.8	71.7	-0.5
ONE[15]	-	72.2	-
CLNN[22]	-	72.4	-
KDCL-Naive	77.5	72.9	2.4
KDCL-Linear	**77.8**	**73.1**	**2.9**
KDCL-MinLogit	**77.8**	**73.1**	**2.9**
KDCL-General	77.1	72.0	1.1

Table 2: Top-1 accuracy rate (%) on ImageNet. All the models are reimplemented with our training procedure for a fair comparison. Gain indicates the sum of the component student network improvement. ONE and CLNN are incompatible with different network structures. Therefore, only the accuracy of ResNet-18 is compared.

Experiment

Method	Top-1	Top-5	Params
Vanilla	71.2	90.0	11.7M
ONE[15]	72.2	90.6	29.5M
KDCL MobileNetV2x1.2	**72.9**	**90.8**	**16.4M**
CLNN[22]	72.4	90.7	40.5M
KDCL ResNet-50	**73.1**	**91.2**	**37.2M**

Table 3: Top-1 and Top-5 accuracy rate (%) on ImageNet. The backbone is ResNet-18. ONE is trained with 3 branches (Res4 block) and CLNN has a hierarchical design with 4 heads. For KDCL, ResNet-18 is trained with a peer network.

Experiment

Model 1	Top-1	Model 2	Top-1	Method
MBV2	72.0	MBV2x0.5	64.8	Vanilla
MBV2	**73.1**	MBV2x0.5	66.2	Linear
MBV2	**73.1**	MBV2x0.5	**66.3**	MinLogit
ResNet-18	71.2	MBV2x0.5	64.8	Vanilla
ResNet-18	71.8	MBV2x0.5*	**65.6**	Linear
ResNet-18	**71.9**	MBV2x0.5*	**65.6**	MinLogit
ResNet-18	71.2	MBV2	72.0	Vanilla
ResNet-18	72.1	MBV2*	**72.8**	Linear
ResNet-18	**72.2**	MBV2*	**72.8**	MinLogit
ResNet-50	76.8	MBV2x0.5	64.8	Vanilla
ResNet-50	77.5	MBV2x0.5*	**67.1**	Linear
ResNet-50	**77.7**	MBV2x0.5*	66.8	MinLogit
ResNet-50*	76.5	ResNet-18*	71.2	Vanilla
ResNet-50*	76.8	ResNet-18*	72.0	Linear
ResNet-50*	**77.0**	ResNet-18*	**72.1**	MinLogit

Table 4: The comparative result of different sub-network on ImageNet validation set. MBV2 is the abbreviation of MobileNetV2. MBV2x0.5 represents the width multiplier is 0.5. ResNet-50* and ResNet-18* are trained for 100 epochs. MBV2* and MBV2x0.5* are trained for 200 epochs.

Experiment

Network	1	2	3	4	5
Top-1 (%)	70.1	71.3	71.61	71.75	71.87

Table 5: KDCL benefits from ensembling more sub-networks. All the networks are ResNet-18 to prevent the impact of network performance differences.

Model	Res-50	Res-18	MBV2	MBV2x0.5	Gain
Vanilla	76.8	71.2	72.0	64.8	0
KDCL	**78.2**	**73.5**	**74.0**	**66.9**	**7.8**

Table 6: Top-1 accuracy rate (%) on ImageNet. ResNet-50 is significantly improved with the knowledge from three compact models.

Experiment

Method	ICL	ResNet-32 Acc %	WRN-16-2 Acc %	Gain
Vanilla		69.9	72.2	0
2 distill 1 [10]		73.3	72.2	3.4
1 distill 2 [10]		69.9	74.5	2.3
DML [32]		73.3	74.8	6.0
ONE [15]		73.6	-	-
CLNN [22]		73.4	-	-
KDCL-Naive		73.7	74.8	6.4
KDCL-Naive	√	73.8	74.9	6.6
KDCL-Linear		73.4	74.6	5.9
KDCL-Linear	√	73.6	74.9	6.4
KDCL-MinLogit		73.0	74.1	5.0
KDCL-MinLogit	√	73.5	74.6	6.0
KDCL-General		74.0	75.2	7.1
KDCL-General	√	**74.3**	**75.5**	**7.7**

Table 7: The comparative and ablative result of our generate distillation method on CIFAR-100 dataset. ICL is invariant collaborative learning. We only report the accuracy of ResNet-32 as ONE and CLNN are incompatible with WRN-16-2.

Thanks