Towards Oracle Knowledge Distillation with Neural Architecture Search

报告人：陈宇航

时间：2020年8月20日

Outline

- Motivation
- Method
- Search space
- Experiments

motivation

number of ensemble	ResNet-32			DenseNet-40-12		
	Teacher	Student	T-S	Teacher	Student	T-S
1	69.11	-	-	74.30	-	-
2	73.77	73.84	-0.07	77.47	77.82	-0.35
3	75.57	74.12	1.45	78.70	78.03	0.67
4	76.36	74.10	2.26	79.32	78.16	1.16
5	76.87	74.67	2.20	79.77	78.43	1.34

Problem:
The accuracy of teacher and student improves gradually in general as the number of models increases while students mostly fail to reach accuracy of teachers and its differences are getting larger

Contribution:
Proposed NAS to addresses capacity issue in KD

Outline

- Motivation
- Method
- Search space
- Experiments

Method

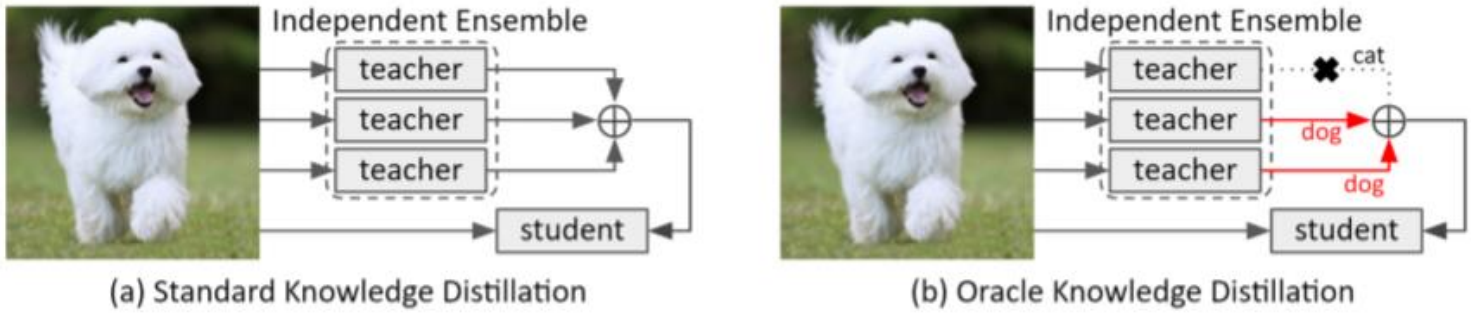


Figure 1: Comparison between standard KD and our proposed OD for the ensemble-based teacher model. In our approach, we train a student network from only the correct models (red arrows) to imitate the oracle predictions of ensemble teacher.

$$\mathcal{L}_{\text{OD}} = \begin{cases} \mathcal{L}_{\text{KD}}(l_s^{(i)}, \bar{l}_t^{(i)}, y^{(i)}) & \text{if } \sum_{j=1}^{N} u_j^{(i)} > 0 \\ \mathcal{L}_{\text{CE}}(l_s^{(i)}, y^{(i)}) & \text{otherwise} \end{cases}, \quad (4)$$

where

$$\bar{l}_t^{(i)} = \frac{\sum_{j=1}^{N} u_j^{(i)} l_{t,j}^{(i)}}{\sum_{j=1}^{N} u_j^{(i)}}.$$

Method

$$S = \{s_1, s_2, \ldots, s_k\},$$

$$\hat{S} = \{s_1, \hat{o}_1, s_2, \hat{o}_2, \ldots, s_k, \hat{o}_k\}.$$

Operation:

skip

convolutions with filter sizes 3×3 and 5×5

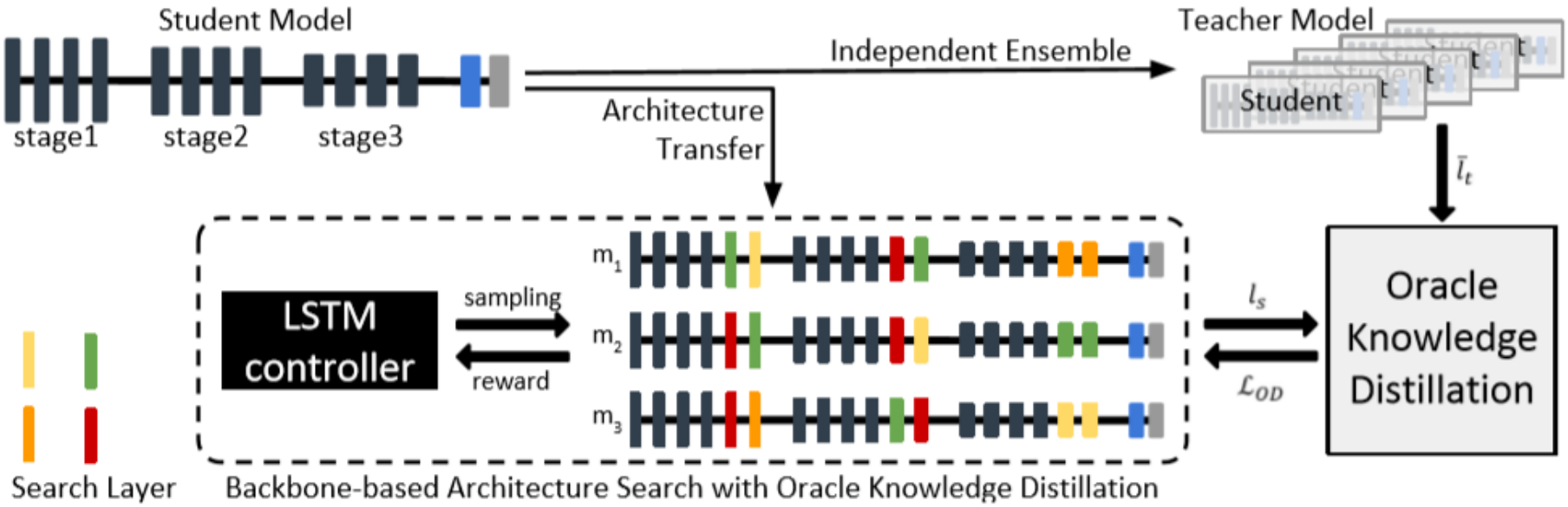
depthwise-separable convolutions with filter sizes 3×3 and 5×5

max pooling 3x3

average pooling 3×3

$$\mathbf{m}^* = \underset{m}{\operatorname{argmax}}\, R(\mathbf{m}), \quad \text{s.t. } |\mathbf{m}| \leq M, \qquad (7)$$

Method



1 use ensemble of independently learned multiple student networks as teacher net work

2 use student network as backbone network

3 LSTM controller provides candidate networks by sampling add-on operations at the end of individual stages in the student

4 Train the candidate networks with OD, and use the validate accurate as reward

5 Chose the best candidate networks as new backbone network

Outline

- Motivation
- Method
- Experiments

Experiment

	Model	$\mathcal{L}_S$	$\mathcal{L}_T$	CIFAR-100 Accuracy	CIFAR-100 Memory	TinyImageNet Accuracy	TinyImageNet Memory	Network identified by
M1	Teacher	-	$\mathcal{L}_{CE}$	76.87	2.35M	62.59	2.38M	-
M2	Student	-	$\mathcal{L}_{CE}$	69.11 ± 0.24	0.47M	54.14 ± 0.65	0.48M	
M3			$\mathcal{L}_{KD}$	74.67 ± 0.10		**58.68 ± 0.09**		
M4			$\mathcal{L}_{OD}$	**74.77 ± 0.02**		58.66 ± 0.25		
M5	ResNet-62	-	$\mathcal{L}_{CE}$	72.06 ± 0.31	0.96M	58.62 ± 0.16	0.97M	Man-Made
M6			$\mathcal{L}_{KD}$	**76.09 ± 0.20**		61.05 ± 0.31		
M7			$\mathcal{L}_{OD}$	75.89 ± 0.19		**61.25 ± 0.14**		
M8	ResNet-110	-	$\mathcal{L}_{CE}$	73.77 ± 0.19	1.73M	60.24 ± 0.45	1.74M	
M9			$\mathcal{L}_{KD}$	**76.77 ± 0.52**		62.03 ± 0.03		
M10			$\mathcal{L}_{OD}$	76.68 ± 0.17		**62.66 ± 0.53**		
M11	NAS	$\mathcal{L}_{CE}$	$\mathcal{L}_{CE}$	74.55 ± 0.51	0.97M	62.01 ± 0.60	0.90M	AutoML
M12			$\mathcal{L}_{KD}$	76.85 ± 0.33		62.10 ± 0.17		
M13			$\mathcal{L}_{OD}$	**77.05 ± 0.23**		**62.57 ± 0.11**		
M14	KDAS (ours)	$\mathcal{L}_{KD}$	$\mathcal{L}_{CE}$	74.56 ± 0.35	0.93M	**62.92 ± 0.10**	0.95M	
M15			$\mathcal{L}_{KD}$	76.97 ± 0.08		62.34 ± 0.10		
M16			$\mathcal{L}_{OD}$	**77.04 ± 0.33**		62.73 ± 0.09		
M17	KDAS (ours)	$\mathcal{L}_{OD}$	$\mathcal{L}_{CE}$	75.14 ± 0.26	0.89M	62.60 ± 0.11	0.87M	
M18			$\mathcal{L}_{KD}$	76.92 ± 0.33		62.17 ± 0.12		
M19			$\mathcal{L}_{OD}$	**77.27 ± 0.11**		**63.04 ± 0.17**		

Experiment

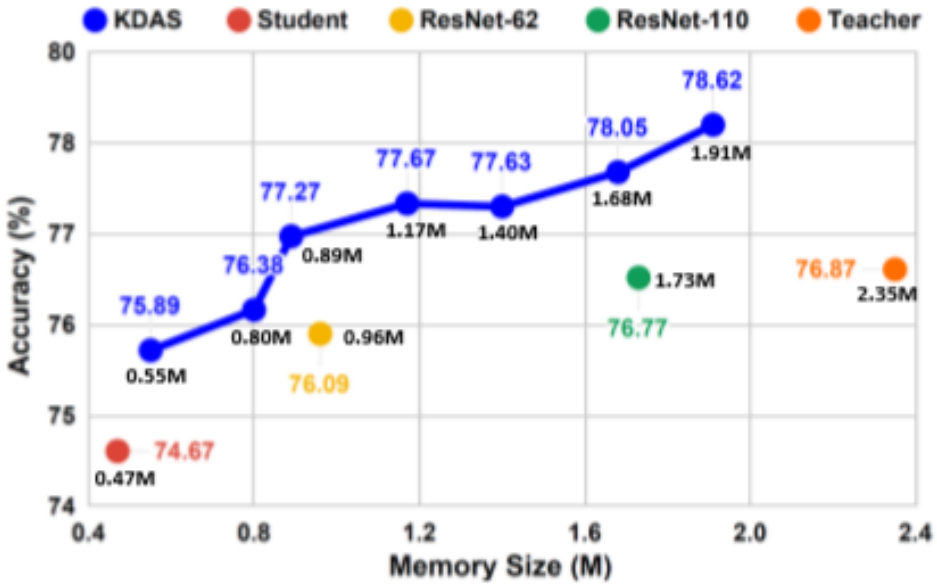


Figure 3: Accuracies varying memory size of networks given by KDAS on the CIFAR-100 dataset with the backbone student network ResNet-32.

Experiment

Table 3: Results with various networks on the CIFAR-100 dataset. We use ResNet-218, WideResNet-76-1, WideResNet-28-2, WideResNet-28-4 networks as MMN of student ResNet-110, WideResNet-40-1, WideResNet-16-2 networks, and WideResNet-16-4 networks, respectively. Numbers in red and blue denote the best and second-best models including the teacher model.

Method	$\mathcal{L}_S$	$\mathcal{L}_T$	ResNet-110		WideResNet-40-1		WideResNet-16-2		WideResNet-16-4	
			Accuracy	Memory	Accuracy	Memory	Accuracy	Memory	Accuracy	Memory
Teacher	-	$\mathcal{L}_{CE}$	**79.24**	8.67M	**77.53**	2.85M	**77.77**	3.52M	**79.49**	13.86M
Student	-	$\mathcal{L}_{CE}$	73.77 ± 0.19	1.73M	69.96 ± 0.15	0.57M	71.16 ± 0.30	0.70M	75.17 ± 0.24	2.77M
Student	-	$\mathcal{L}_{KD}$	76.77 ± 0.52	1.73M	74.72 ± 0.23	0.57M	75.42 ± 0.04	0.70M	78.59 ± 0.34	2.77M
MMN	-	$\mathcal{L}_{KD}$	77.39 ± 0.21	3.48M	76.48 ± 0.15	1.15M	76.97 ± 0.05	1.48M	79.28 ± 0.16	5.87M
KDAS	$\mathcal{L}_{OD}$	$\mathcal{L}_{OD}$	**79.01 ± 0.28**	2.73M	**76.70 ± 0.25**	1.14M	**77.83 ± 0.23**	1.30M	**79.79 ± 0.24**	5.47M

Table 4: Performance comparison with other KD algorithms on the CIFAR-100 dataset. We use a single ResNet-110 network as a teacher model. The red-colored number means the highest accuracy.

Student	CE	KD	DML	BSS	TAKD	KDAS (0.91M)
ResNet-62 (0.96M)	71.73 ± 0.03	74.57 ± 0.18	72.98 ± 1.07	73.06 ± 0.53	75.18 ± 0.13	**75.82 ± 0.32**
ResNet-68 (1.05M)	71.77 ± 0.06	74.82 ± 0.09	73.39 ± 0.70	73.43 ± 0.21	75.45 ± 0.12	

Experiment

Table 5: Training accuracy of single ResNet-32 network on CIFAR-100 and TinyImageNet datasets. We also present the percentage of training examples in terms of the number of models that predict correctly.

Dataset	# of models that predict correctly					Training
	1	2	3	4	5	Acc.
CIFAR-100	0.5	1.0	2.6	8.9	86.9	94.04
TinyImageNet	6.3	6.8	8.7	14.3	49.6	70.28

The End!