Knowledge Distillation with a Precise Teacher and Prediction with Abstention

Yi Xu, Jian Pu, Hui Zhao

School of Software Engineering, East China Normal University,
Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University
Email: 51184501068@stu.ecnu.edu.cn, jianpu@fudan.edu.cn, hzhao@sei.ecnu.edu.cn

Introduction

Knowledge distillation[1] has achieved remarkable results in supervised learning.

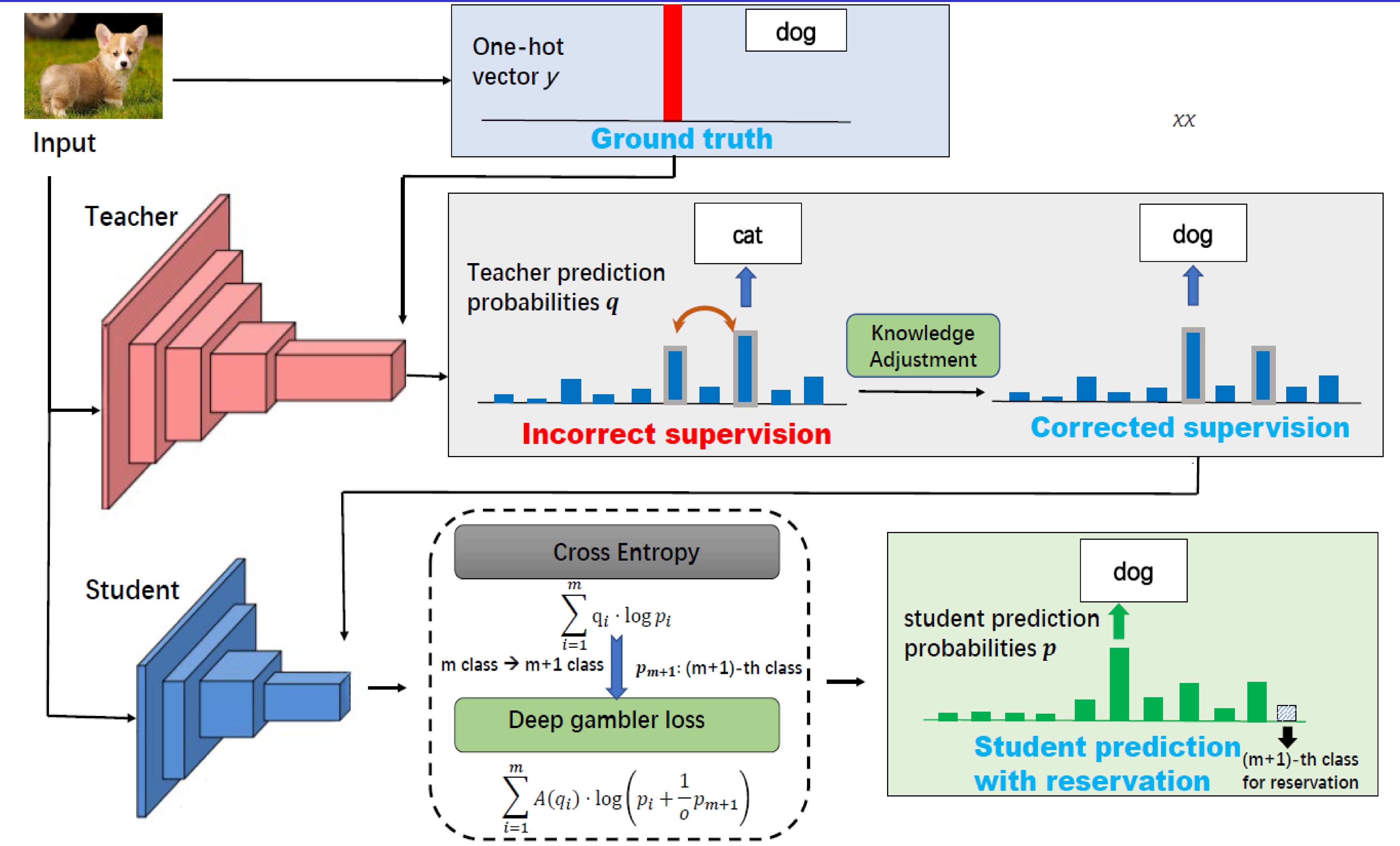
However, there are two major problems with existing knowledge distillation methods.

- Teacher's supervision is sometimes **misleading.**
- Student's prediction is **not accurate** enough.

To address the first issues, we apply **knowledge adjustment** to correct teachers' supervision using ground truth.

For the second problem, we use the **selective classification framework**[3] to train the student model. In particular, the **deep gambler loss**[2] is adopted to predict with reservation by introducing the extra class.

Model



Result

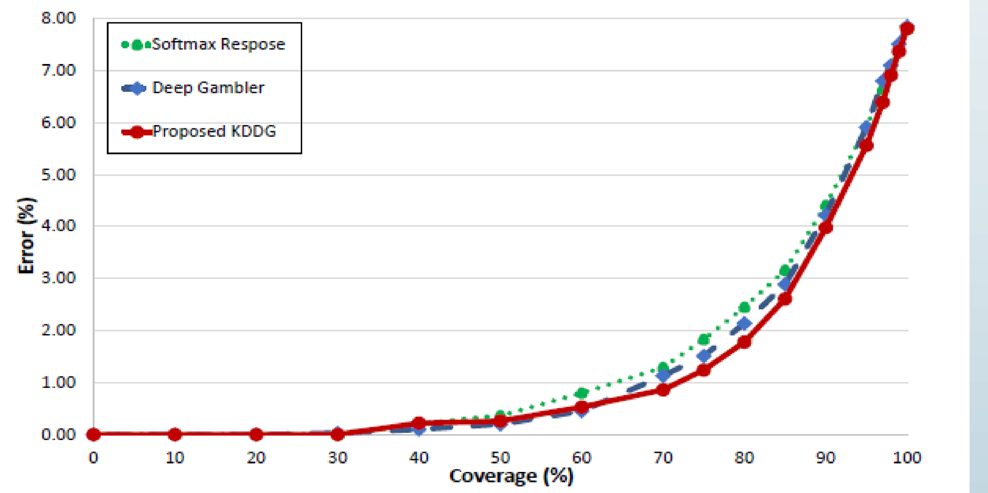
Distillation across Different Network Structures

TABLE I: The comparison of accuracy on four datasets by knowledge distillation across different network structures.

Method	Fashion-MNIST Accuracy(%)	SVHN Accuracy(%)	CIFAR10 Accuracy(%)	CIFAR100 Accuracy(%)
Student(AlexNet)	92.60	94.86	85.57	60.71
Teacher(ResNet50)	93.95	97.47	95.42	76.86
Deep Gambler	92.64	95.02	**87.17**	61.17
Proposed Method	**92.94**	**95.05**	87.11	**61.24**

TABLE II: The comparison of Sum Coverage Error in 0%-100% and 70%-100% by knowledge distillation across different network structures.

Method	Fashion-MNIST		SVHN	
	Sum Coverage Error (0,100)	Sum Coverage Error (70,100)	Sum Coverage Error (0,100)	Sum Coverage Error (70,100)
Softmax Response	130.21	110.70	218.48	149.43
Deep Gambler	119.03	105.63	195.96	135.60
Proposed Method	**114.58**	**86.88**	**181.96**	**124.73**
Method	CIFAR10		CIFAR100	
	Sum Coverage Error (0,100)	Sum Coverage Error (70,100)	Sum Coverage Error (0,100)	Sum Coverage Error (70,100)
Softmax Response	287.52	222.33	1826.63	973.31
Deep Gambler	**265.85**	**217.62**	1612.89	958.59
Proposed Method	276.37	220.54	**1598.24**	**949.50**



(a) AlexNet & ResNet50

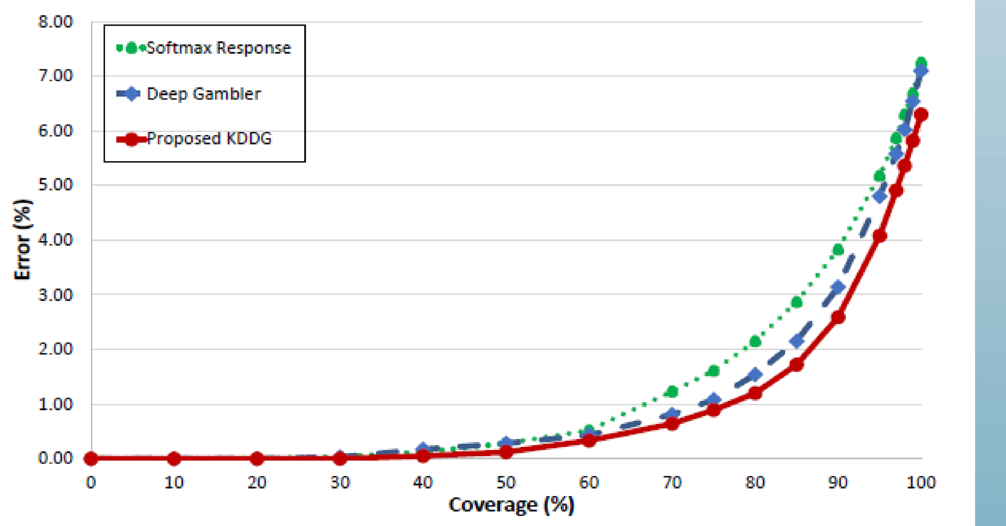
Distillation across Network with Different Depth

TABLE III: The comparison of accuracy on four datasets by knowledge distillation across different network depths.

Method	Fashion-MNIST Accuracy(%)	SVHN Accuracy(%)	CIFAR10 Accuracy(%)	CIFAR100 Accuracy (%)
Student(ResNet18)	93.64	97.25	95.14	76.42
Teacher(ResNet50)	93.95	97.47	95.42	76.86
Deep Gambler	93.79	97.20	95.12	76.54
Proposed Method	**93.92**	**97.41**	**95.42**	**76.86**

TABLE IV: The comparison of Sum Coverage Error in 0%-100% and 70%-100% by knowledge distillation across different network scales.

Method	Fashion-MNIST		SVHN	
	Sum Coverage Error (0,100)	Sum Coverage Error (70,100)	Sum Coverage Error (0,100)	Sum Coverage Error (70,100)
Softmax Response	113.30	98.05	115.18	61.19
Deep Gambler	95.30	82.55	122.61	64.80
Proposed Method	**77.18**	**68.82**	**106.93**	**55.89**
Method	CIFAR10		CIFAR100	
	Sum Coverage Error (0,100)	Sum Coverage Error (70,100)	Sum Coverage Error (0,100)	Sum Coverage Error (70,100)
Softmax Response	61.41	50.11	867.02	521.01
Deep Gambler	63.53	55.94	867.32	519.13
Proposed Method	**59.81**	**47.90**	**853.48**	**516.46**



(b) ResNet18 & ResNet50

Motivation

1. We propose to use knowledge adjustment to revise teacher's incorrect supervision using the ground truth label.
2. We propose to use the deep gambler loss to train the student network in an end-to-end way.
3. We evaluate the proposed method under two knowledge distillation settings. i.e., knowledge distillation across different network structures and distillation across networks with different depths.

Method

Knowledge Adjustment

- The Knowledge Distillation(KD) loss as：

$$\mathcal{L}_{KD} = \alpha\tau^2 \cdot CE(q_\tau, p_\tau) + (1-\alpha)\cdot CE(y, p_\tau) \quad (1)$$

- Swap the incorrect value with the true targets. We only need to operate on the incorrect ones and denote it as an operator A(·). The KD loss becomes:

$$\mathcal{L}_{KD^*} = \tau^2 CE(\mathcal{A}(q_\tau), p_\tau) \qquad (2)$$

Knowledge Distillation with Deep Gambler

- We can add a class to stand for abandoning predictions and reservations according to **Deep Gambler Loss**[2] in selective classification:

$$W(b(f),p) = \sum_i \log[o \cdot f_w(x_i)_j + f_w(x_i)_{m+1}] \quad (3)$$

- We proposed the loss function that utilizes Deep Gambler (DG) loss to the KA method.

$$\mathcal{L} = \sum_i \mathcal{A}(q_\tau^i) \log\left(p_\tau^i + \frac{1}{o} p_\tau^{m+1}\right) \quad (4)$$

Reference

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[2] Z. Liu, Z. Wang, P. P. Liang, R. R. Salakhutdinov, L.-P. Morency, an dM. Ueda, "Deep gamblers: Learning to abstain with portfolio theory," in Advances in Neural Information Processing Systems, pp. 10623–10633,2019.

[3] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems, pp. 4878–4887, 2017.