Knowledge Distillation by On-the-Fly Native Ensemble

Xu Lan[1] x.lan@qmul.ac.uk

Xiatian Zhu[2] eddy@visionsemantics.com

Shaogang Gong[1] s.gong@qmul.ac.uk

[1]Queen Mary University of London, London, UK [2]Vision Semantics Ltd

Advances in Neural Information Processing Systems

1. Introduction

Cross Entropy Hard vs. Soft Class Labels:

$$\mathcal{L}_{ce} = -\sum_{c=1}^{C} \delta_{c,y} \log \left(p(c|\boldsymbol{x}, \boldsymbol{\theta})\right)$$

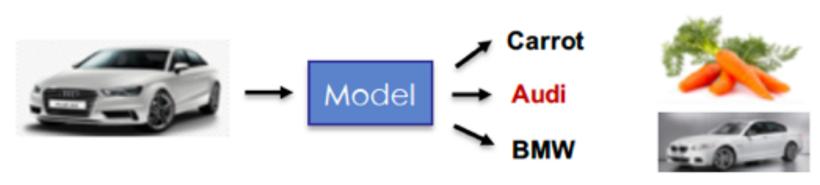


Table: The label information and the Model predictions

	Category	Audi	BMW	Carrot			Audi	BMW	Carrot
Label	Hard Label	1	0	0	Model	Model-A	0.6	0.39	0.01
	Soft Label	0.95	0.049	0.001		Model-B	0.6	0.01	0.39

CE+Hard: $Loss_A = Loss_B$ CE+Soft: $Loss_A < Loss_B$

Drawbacks of Hard Label based Cross Entropy:

- Considering no *correlation* between classes.
- Prone to model *overfitting*

Solution: *Knowledge Distillation*

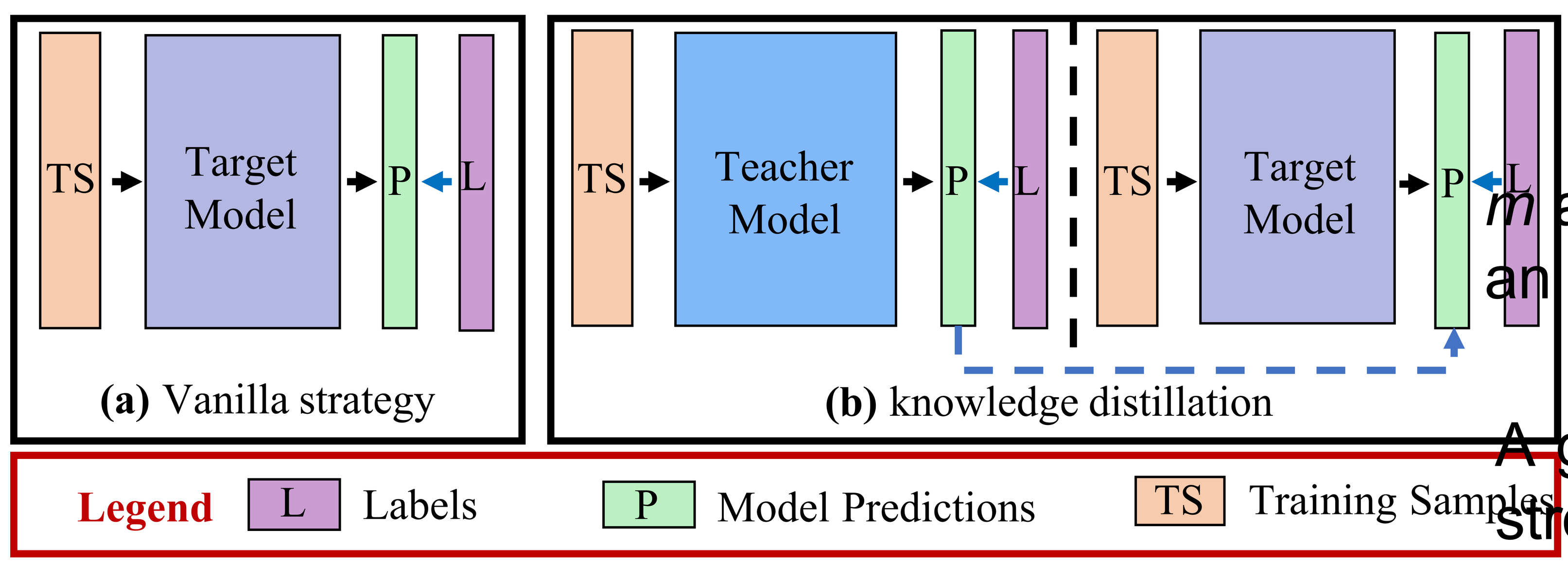


Figure 1: Knowledge distillation and vanilla training

Limitations of Offline Knowledge Distillation:

- Lengthy training time
- Possible teacher model overfitting
- Complex multi-stage training process

Limitations of Online Knowledge Distillation:

- Still need to train multiply models
- Provide limited extra supervision information
- Complex Asynchronous model updating

2. Methodology

Knowledge Distillation by On-the-Fly Native Ensemble

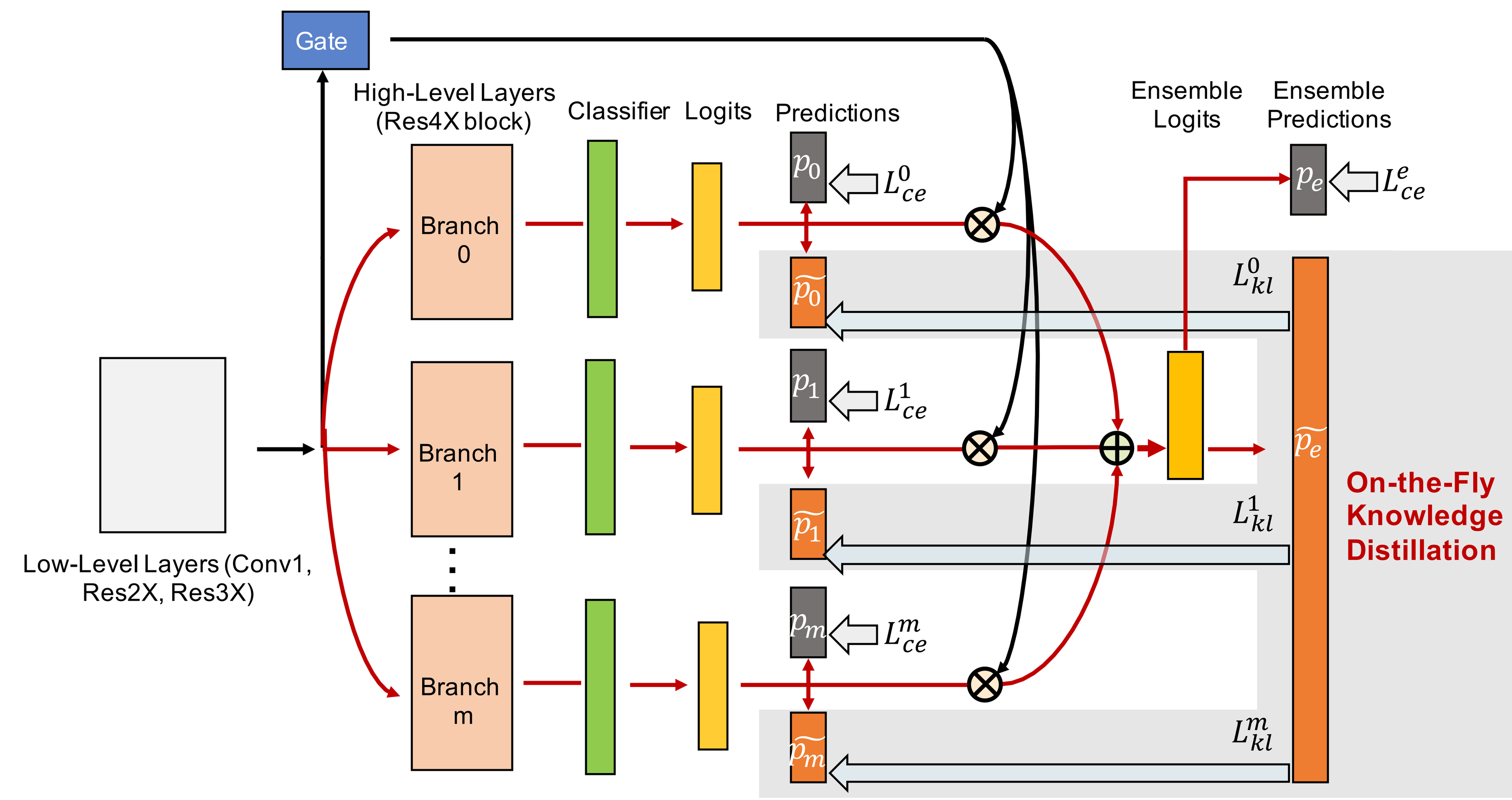


Figure 2: Overview of online distillation training of ResNet-110 by the proposed On-the-Fly Native Ensemble (ONE). With ONE, we reconfigure the network by adding *m* auxiliary branches. Each branch with shared layers makes an individual model, and their ensemble is used to build the teacher model.

Multi-Branch Design:

m auxiliary branches with the same configuration, each serving as an independent efficient classification model.

Gate Network:

A gate which learns to ensemble all (m + 1) branches to build a stronger teacher:

$$\boldsymbol{z}_e = \sum_{i=0}^{m} g_i \cdot \boldsymbol{z}_i$$

On-the-Fly Knowledge Distillation:

Compute soft probability distributions at a temperature of T for branches and the ONE teacher as:

$$\tilde{p}_i(c|\boldsymbol{x}, \boldsymbol{\theta}^i) = \frac{\exp(\boldsymbol{z}_i^c / T)}{\sum_{j=1}^{C} \exp(\boldsymbol{z}_i^j / T)}, \quad \tilde{p}_e(c|\boldsymbol{x}, \boldsymbol{\theta}^e) = \frac{\exp(\boldsymbol{z}_e^c / T)}{\sum_{j=1}^{C} \exp(\boldsymbol{z}_e^j / T)}, \quad c \in \mathcal{Y}$$

Distill knowledge from the teacher to each branch:

$$\mathcal{L}_{kl} = \sum_{i=0}^{m} \sum_{j=1}^{C} \tilde{p}_e(j|\boldsymbol{x}, \boldsymbol{\theta}^e) \log \frac{\tilde{p}_e(j|\boldsymbol{x}, \boldsymbol{\theta}^e)}{\tilde{p}_i(j|\boldsymbol{x}, \boldsymbol{\theta}^i)}$$

Overall Loss Function:

$$\mathcal{L} = \sum_{i=0}^{m} \mathcal{L}_{ce}^i + \mathcal{L}_{ce}^e + T^2 * \mathcal{L}_{kl}$$

3. Experiments

CIFAR and SVHN tests

Method	CIFAR10	CIFAR100	SVHN	Params
ResNet-32	6.93	31.18	2.11	0.5M
ResNet-32 + **ONE**	**5.99±0.05**	**26.61±0.06**	**1.83±0.05**	0.5M
ResNet-110	5.56	25.33	2.00	1.7M
ResNet-110 + **ONE**	**5.17±0.07**	**21.62±0.26**	**1.76±0.07**	1.7M
ResNeXt-29($8\times64d$)	3.69	17.77	1.83	34.4M
ResNeXt-29($8\times64d$) + **ONE**	**3.45±0.04**	**16.07±0.08**	**1.70±0.03**	34.4M
DenseNet-BC(L=190, k=40)	3.32	17.53	1.73	25.6M
DenseNet-BC(L=190, k=40) + **ONE**	**3.13±0.07**	**16.35±0.05**	**1.63±0.05**	25.6M

ImageNet test

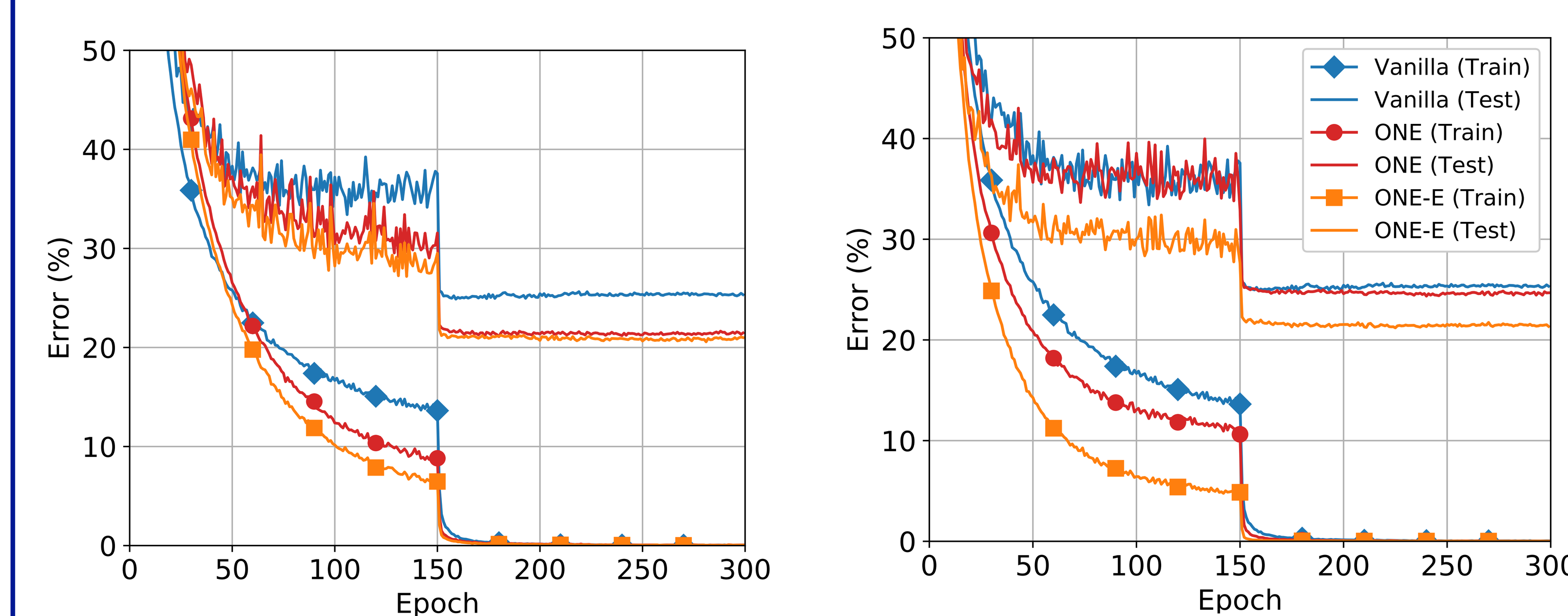
Method	Top-1	Top-5
ResNet-18 [He et al., 2016]	30.48	10.98
ResNet-18 + **ONE**	**29.45±0.23**	**10.41±0.12**
ResNeXt-50 [Xie et al., 2017]	22.62	6.29
ResNeXt-50 + **ONE**	**21.85±0.07**	**5.90±0.05**
SeNet-ResNet-18 [Hu et al., 2017]	29.85	10.72
SeNet-ResNet-18 + **ONE**	**29.02±0.17**	**10.13±0.12**

Knowledge Distillation and Ensemble Comparisons

Target Network	ResNet-32			ResNet-110		
Metric	Error (%)	TrCost	TeCost	Error (%)	TrCost	TeCost
KD [Hinton et al., 2015]	**28.83**	6.43	1.38	N/A	N/A	N/A
DML [Zhang et al., 2017]	29.03±0.22*	**2.76**	1.38	**24.10±0.72**	**10.10**	5.05
ONE	**26.61±0.06**	**2.28**	1.38	**21.62±0.26**	**8.29**	5.05

Network	ResNet-32			ResNet-110		
Metric	Error (%)	TrCost	TeCost	Error (%)	TrCost	TeCost
Snopshot Ensemble [Huang et al., 2017]	27.12	**1.38**	6.90	23.09*	**5.05**	25.25
2-Net Ensemble	26.75	2.76	**2.76**	22.47	10.10	**10.10**
3-Net Ensemble	**25.14**	4.14	4.14	**21.25**	15.15	15.15
ONE-E	**24.63**	**2.28**	**2.28**	**21.03**	**8.29**	**8.29**
ONE	26.61	2.28	1.38	21.62	8.29	5.05

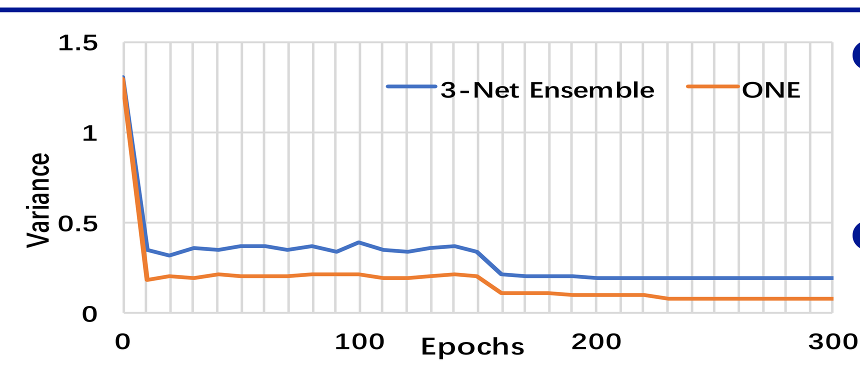
Effect of On-the-Fly Knowledge Distillation



4. Further Analysis

ONE vs. Model Ensemble (ME)

(1) Model variance: average prediction differences between every two models/branches. **(2) Mean model generalisation capability**.



- ONE leads to higher correlations due to the learning constraint from the distillation loss;
- ONE yields superior mean model generalisation capability with lower error rate 26.61 vs 31.07 by ME.

5. Reference

[1] G. Hinton, et al. “Distilling the knowledge in a neural network.” arXiv, 2015.

[2] Y. Zhang, et al. “Deep mutual learning.” CVPR, 2018.