Queen Mary University of London

vision semantics

Self-Referenced Deep Learning

Xu Lan[1] Xiatian Zhu[2] Shaogang Gong[1]

x.lan@qmul.ac.uk eddy@visionsemantics.com s.gong@qmul.ac.uk

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

ACCV 2018 2 – 6 December 2018 Perth Western Australia

1. Introduction

Cross Entropy Hard vs. Soft Class Labels:

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

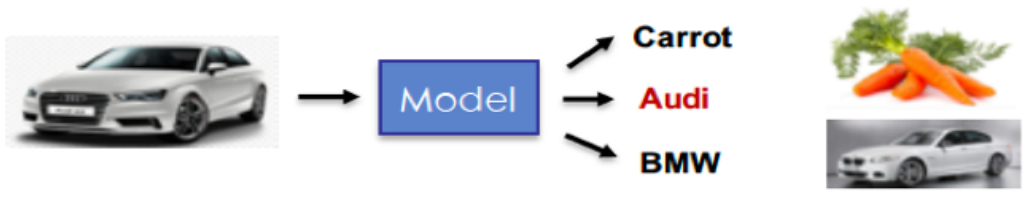

Table 1: The label information and the model predictions

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.

Contributions:

- Investigate for the first time knowledge distillation and fast optimisation in the model training using a unified deep learning approach
- Present a stage-complete learning rate decay schedule for SRDL.
- introduce a random model restart scheme for SRDL.

Solution: Knowledge Distillation

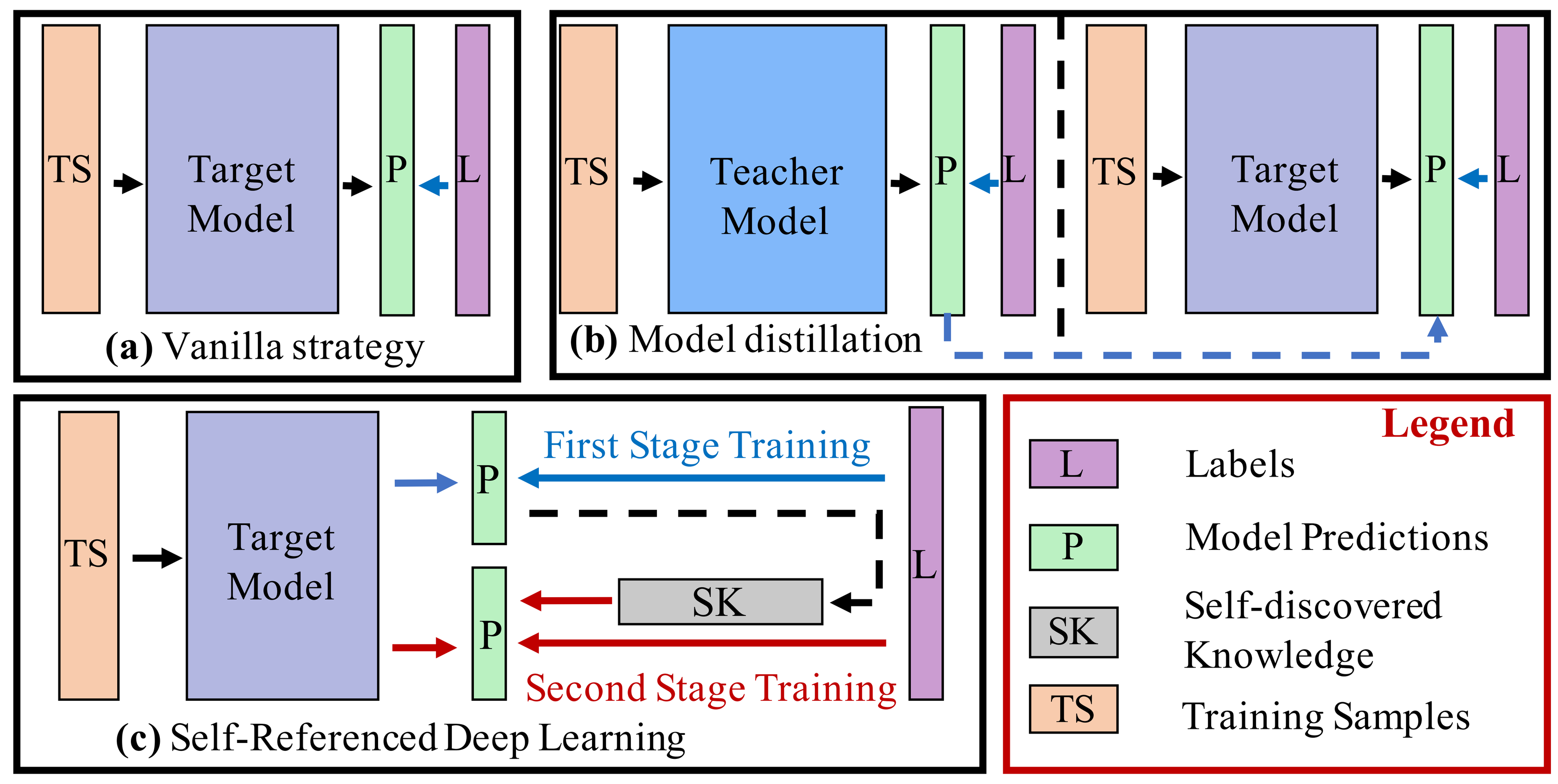

Figure1: Illustration of different deep network learning methods. (a) The vanilla training ; (b) Knowledge Distillation training ; (c) The proposed Self Reference Deep learning (SRDL).

2. Methodology

Self-Referenced Deep Learning

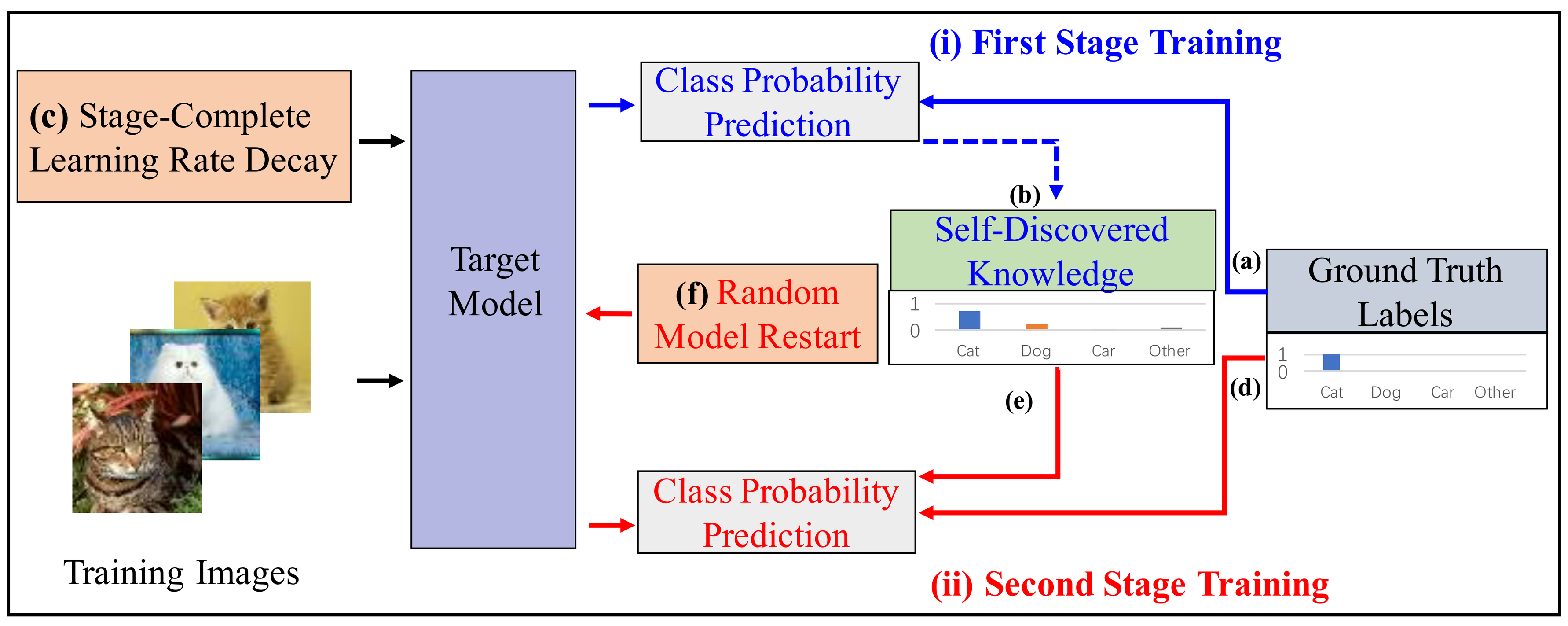

Figure 2: Overview of our proposed Self-Referenced Deep Learning (SRDL)

First Stage Learning:

- In first stage of SRDL, we train the deep model θ by cross-entropy loss.
- To maximise the quality of self-discovered knowledge, we introduce Figure 2 (c) a pass-complete learning rate decay schedule.

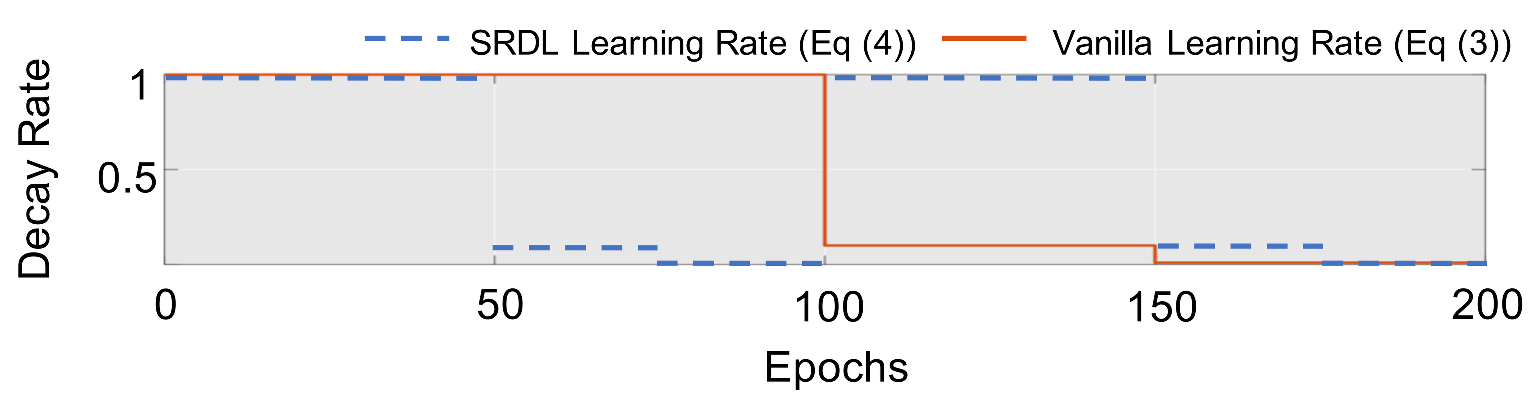

Figure 3: Illustration of a vanilla learning rate step-decay function and the proposed stage-complete learning rate step-decay schedule.

Second Stage Learning:

- We start second stage training with randomly initialised model parameters.
- Continuously optimize the target model for the other half epochs by the joint supervision of both Figure2 (d) the label data and Figure2 (e) self-discovered intermediate knowledge in an end-to-end manner.

$$R_{kl} = \sum_{j=1}^{C} \tilde{p}(j|\boldsymbol{x},\boldsymbol{\theta}^*) \log \frac{\tilde{p}(j|\boldsymbol{x},\boldsymbol{\theta}^*)}{\tilde{p}(j|\boldsymbol{x},\boldsymbol{\theta})}. \qquad \mathcal{L} = \mathcal{L}_{ce} + T^2 * R_{kl}$$

Algorithm 1. Self-Referenced Deep Learning

Input: Labelled training data $\mathcal{D}$; Training epochs M;
Output: Trained CNN model $\boldsymbol{\theta}$;
(I) First stage learning
Initialisation: t=1; Random model $\boldsymbol{\theta}$ initialisation;
while $t \leq 0.5 * M$ **do**
  (i) Update the learning rate ϵ_t (Eq (4));
  (ii) Update $\boldsymbol{\theta}$ by cross-entropy loss (Eq (2));
end
Knowledge Extraction Induce per-sample class probability predictions (Eq (5));
(II) Second stage learning
Initialisation: t=1; Random model $\boldsymbol{\theta}$ restart;
while $t \leq 0.5 * M$ **do**
  (i) Update the learning rate ϵ_t (Eq (4));
  (ii) Update $\boldsymbol{\theta}$ by soft-feedback referenced loss (Eq (7));
end

3. Experiments

Comparison with the Vanilla Learning Strategy

Dataset	# Param	CIFAR10		CIFAR100		Tiny ImageNet	
Metrics		Acc	TrCost	Acc	TrCost	Acc	TrCost
ResNet-32+vanilla	0.5M	92.53	0.08	69.02	0.08	53.33	0.32
ResNet-32+**SRDL**		**93.12**	0.08	**71.63**	0.08	**55.53**	0.32
Gain (SRDL-vanilla)		+0.59	0	+2.61	0	+2.20	0
WRN-28-10+vanilla	36.5M	94.98	12.62	78.32	12.62	58.38	50.48
WRN-28-10+**SRDL**		**95.41**	12.62	**79.38**	12.62	**60.80**	50.48
Gain (SRDL-vanilla)		+0.43	0	+1.06	0	+2.42	0
DenseNet-BC+vanilla	25.6M	**96.68**	10.24	**82.83**	10.24	**62.88**	40.96
DenseNet-BC+**SRDL**		**96.87**	10.24	**83.59**	10.24	**64.19**	40.96
Gain (SRDL-vanilla)		+0.19	0	+0.76	0	+1.31	0

Table 2: Comparison between SRDL and vanilla learning on image classification

Comparison with Knowledge Distillation

Target Net	Method	Teacher Net	CIFAR10		CIFAR100		Tiny ImageNet	
			Acc	TrCost	Acc	TrCost	Acc	TrCost
ResNet-32 (0.5M)	Vanilla	N/A	92.53	**0.08**	69.02	**0.08**	53.33	**0.32**
	KD	WRN-28-10 (36.5M)	**92.83**	12.70	**72.58**	12.70	**56.80**	50.80
		ResNet-110 (1.7M)	92.75	0.30	71.17	0.30	55.06	1.20
	SRDL	N/A	**93.12**	**0.08**	**71.63**	**0.08**	**55.53**	**0.32**

Table 3: Comparison between SRDL and Knowledge Distillation (KD)

Evaluation on Person Instance Recognition

Query Type	Single-Query		Multi-Query	
Metrics (%)	Rank-1	mAP	Rank-1	mAP
ResNet-50+vanilla	87.5	69.9	91.4	78.5
ResNet-50+**SRDL**	**89.3**	**73.5**	**93.1**	**81.5**
Gain (SRDL-vanilla)	+1.8	+3.6	+1.7	+3.0
DenseNet-121+vanilla	**90.1**	74.0	**93.6**	81.7
DenseNet-121+**SRDL**	**91.7**	**76.8**	**94.2**	**83.5**
Gain (SRDL-vanilla)	+1.6	+2.8	+0.6	+1.8

Table 4: Evaluation of person re-id (instance recognition) on Market-1501.

Component Analysis and Discussion

Decay Strategy	Accuracy (%)
Stage-Incomplete	58.11
Stage-Complete	**71.63**

Table 5: Stage-complete schedule

Random Restart	Accuracy (%)
✗	69.73
✓	**71.63**

Table 6: Random model restart.

4. Conclusion

- SRDL train more discriminative small and large networks with little extra computational cost.
- The results validate the performance superiority of SRDL training.

5. Reference

[1] Hinton et al. : Distilling the knowledge in a neural network.