Learning Efficient Object Detection Models with Knowledge Distillation

Method

01

Overall Structure

02

Knowledge Distillation for Classification with Imbalanced Classes

03

Knowledge Distillation for Regression with Teacher Bounds

04

Hint Learning with Feature Adaptation

Overall Structure

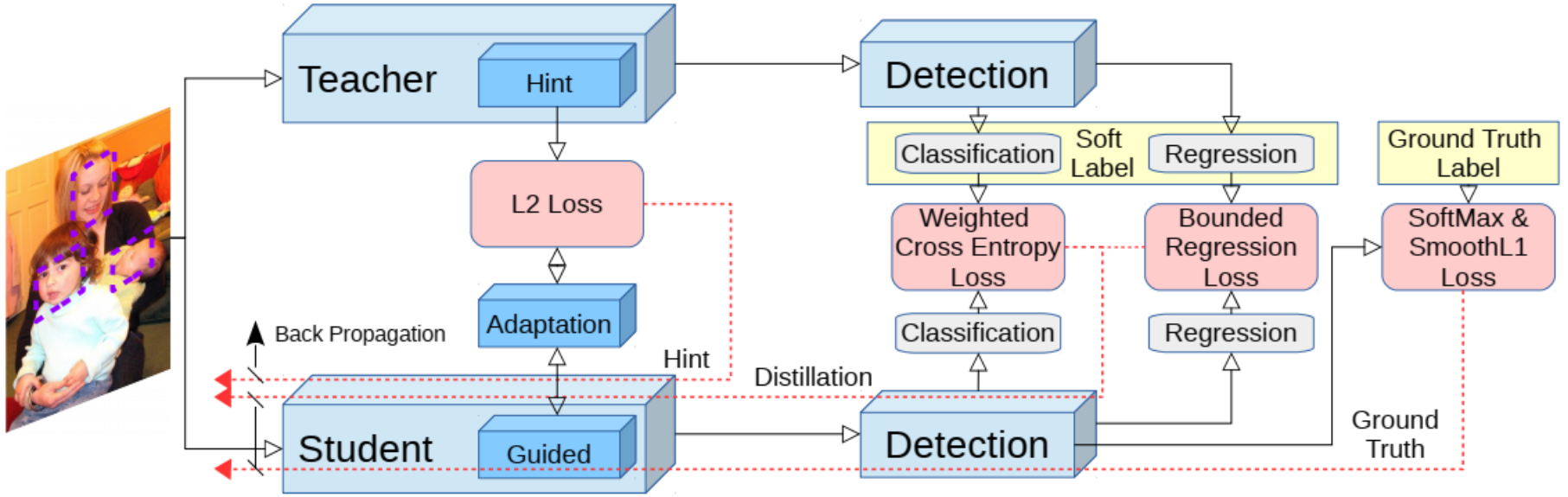

Figure 1: The proposed learning skeme on visual object detection task using Faster-RCNN, which mainly consists of region proposal network (RPN) and region classification network(RCN). The two networks both use multi-task loss to jointly learn the classifier and bounding-box regressor. We employ the final output of the teacher's RPN and RCN as the distillation targets, and apply the intermediate layer outputs as hint. Red arrows indicate the backpropagation pathways.

$$L_{RCN} = \frac{1}{N}\sum_i L_{cls}^{RCN} + \lambda\frac{1}{N}\sum_j L_{reg}^{RCN}$$

$$L_{RPN} = \frac{1}{M}\sum_i L_{cls}^{RPN} + \lambda\frac{1}{M}\sum_j L_{reg}^{RPN}$$

$$L = L_{RPN} + L_{RCN} + \gamma \boldsymbol{L_{Hint}}$$

Knowledge Distillation for Classification with Imbalanced Classes

Loss Function

$$L_{cls} = \mu L_{hard}(P_s, y) + (1-\mu)\boldsymbol{L_{soft}(P_s, P_t)}$$

Distillation Loss

$$L_{soft}(P_s, P_t) = -\sum w_c P_t \log P_s$$

Knowledge Distillation for Regression with Teacher Bounds

$$L_b(R_s, R_t, y) = \begin{cases} \|R_s - y\|_2^2, & \text{if } \|R_s - y\|_2^2 + m > \|R_t - y\|_2^2 \\ 0, & \text{otherwise} \end{cases}$$

$$L_{reg} = L_{sL1}(R_s, y_{reg}) + \nu L_b(R_s, R_t, y_{reg})$$

Hint Learning with Feature Adaptation

Ablation Study

$$L_{Hint}(V, Z) = \|V - Z\|_2^2$$

$$L_{Hint}(V, Z) = \|V - Z\|_1$$

	Baseline	L2	L2-B	CLS	CLS-W	Hints	Hints-A	L2-B+CLS-W	L2-B+CLS-W+Hints-A
PASCAL	54.7	54.6	55.9	57.4	57.7	56.9	58	58.4	59.4
KITTI	49.3	48.5	50.1	50.8	51.3	50.3	52.1	51.7	53.7

Table 4: The proposed method component comparison, i.e., bounded L2 for regression (L2-B, Sec.3.3) and weighted cross entropy for classification (CLS-W, Sec.3.2) with respect to traditional methods, namely, L2 and cross entropy (CLS). Hints learning w/o adaptation layer (Hints-A and Hints) are also compared. All comparisons take VGG16 as the teacher and Tucker as the student, with evaluations on PASCAL and KITTI.

Experiments

01

Overall Performance

02

Speed-Accuracy Trade off in Compressed Models

03

Discussion

Overall Performance

Student	Model Info	Teacher	PASCAL	COCO@.5	COCO@[.5,.95]	KITTI	ILSVRC
Tucker	11M / 47ms	-	54.7	25.4	11.8	49.3	20.6
		AlexNet	57.6 (+2.9)	26.5 (+1.2)	12.3 (+0.5)	51.4 (+2.1)	23.6 (+1.3)
		VGGM	58.2 (+3.5)	26.4 (+1.1)	12.2 (+0.4)	51.4 (+2.1)	23.9 (+1.6)
		VGG16	59.4 (+4.7)	28.3 (+2.9)	12.6 (+0.8)	53.7 (+4.4)	24.4 (+2.1)
AlexNet	62M / 74ms	-	57.2	32.5	15.8	55.1	27.3
		VGGM	59.2 (+2.0)	33.4 (+0.9)	16.0 (+0.2)	56.3 (+1.2)	28.7 (+1.4)
		VGG16	60.1 (+2.9)	35.8 (+3.3)	16.9 (+1.1)	58.3 (+3.2)	30.1 (+2.8)
VGGM	80M / 86ms	-	59.8	33.6	16.1	56.7	31.1
		VGG16	63.7 (+3.9)	37.2 (+3.6)	17.3 (+1.2)	58.6 (+2.3)	34.0 (+2.9)
VGG16	138M / 283ms	-	70.4	45.1	24.2	59.2	35.6

Table 1: Comparison of student models associated with different teacher models across four datasets, in terms of mean Average Precision (mAP). Rows with blank (-) teacher indicate the model is without distillation, serving as baselines. The second column reports the number of parameters and speed (per image, on GPU).

	High-res teacher		Low-res baseline		Low-res distilled student	
	mAP	Speed	mAP	Speed	mAP	Speed
AlexNet	57.2	1,205 / 74 ms	53.2	726 / 47 ms	56.7(+3.5)	726 / 47 ms
Tucker	54.7	663 / 41 ms	48.6	430 / 29 ms	53.5(+4.9)	430 / 29 ms

Table 2: Comparison of high-resolution teacher model (trained on images with 688 pixels) and low-resolution student model (trained on 344 pixels input), on PASCAL. We report mAP and speed (both CPU and GPU) of different models. The speed of low-resolution models are about 2 times faster than the corresponding high-resolution models, while achieving almost the same accuracy when our distillation method is used.

Speed-Accuracy Trade off in Compressed Models

FLOPS(%)	20	25	30	**37.5**	45
Finetune	30.3	49.3	51.4	54.7	55.2
Distillation	35.5(+5.2)	55.4(+6.1)	56.8(+5.4)	59.4(+4.7)	59.5(+4.3)

Table 3: Compressed AlexNet performance evaluated on PASCAL. We compare the model fine-tuned with the ground truth and the model trained with our full method. We vary the compression ratio by FLOPS.

Discussion

		Baseline	Distillation	Hint	Distillation + Hint
PASCAL	Trainval	79.6	78.3	80.9	83.5
	Test	54.7	58.4	58	59.4
COCO	Train	45.3	45.4	47.1	49.6
	Val	25.4	26.1	27.8	28.3

Table 5: Performance of distillation and hint learning on different datasets with Tucker and VGG16 pair.

- ✔ Distillation improves generalization
- ✔ Hint helps both learning and generalization

Thank You