Sharpen Focus: Learning with Attention Separability and Consistency Supplementary Material

Lezi Wang[1], Ziyan Wu[2], Srikrishna Karanam[2], Kuan-Chuan Peng[2],
Rajat Vikram Singh[2], Bo Liu[1], and Dimitris N. Metaxas[1]
[1]Rutgers University, New Brunswick NJ
[2]Siemens Corporate Technology, Princeton NJ
{lw462, lb507, dnm}@cs.rutgers.edu, {ziyan.wu, srikrishna.karanam, kuanchuan.peng, singh.rajat}@siemens.com

1. Implementation details

As noted in the main paper, our experiments contain three parts: (a) generic image classification evaluation on three datasets: CIFAR-100 [6], Caltech-256 [3] and ILSVRC2012 [8], (b) fine-grained image classification evaluation on the CUB-200-2011 [12] dataset, and (c) multi-label image classification evaluation on the PASCAL VOC 2012 [2] dataset. We perform experiments using PyTorch [7] and NVIDIA Titan X GPUs. We do not search in the hyperparameter space for the best hyperparameters and instead use the same training parameters as those in the corresponding baselines. Complete experimental details about training and the five datasets we used are provided in Table 1.

1.1. Generic image classification

CIFAR-100: The image is padded by 4 pixels on each side, filled with 0 value resulting in a 40×40 image. A 32×32 crop is randomly sampled from an image or its horizontal flip, with the per-pixel RGB mean value subtracted. We adopt the same weight initialization method following [4] and train the ResNet using Stochastic Gradient Descent (SGD) [1] with a mini-batch size of 128. We use a weight decay of 0.0005 with a momentum of 0.9 and set the initial learning rate to 0.1. The learning rate is divided by 10 at 81 and 122 epochs. The training is terminated after 160 epochs.

Caltech-256: There is no official training/testing data split. We follow the work in [3] to randomly select 25 images per category as the testing set and 30, 60 images per category as training. We remove the last (257-th) category "clutter," keeping the 256 categories which describe specific objects. We use VGG-19 [10] and ResNet-18 [4] as the baseline models. For the training of both the baseline and our proposed method, we use a weight decay of 0.001 with a momentum of 0.9 and set the initial learning rate to 0.01. To speed up the model training, we adopt cyclic cosine annealing [5] with a cycle of one to train the network for 20 epochs.

ILSVRC2012: We conduct large-scale image classification experiments using the ImageNet ILSVRC2012 dataset [8]. The evaluation is conducted on the images of the ILSVRC2012 validation set. We use ResNet-18 [4] as the baseline model. We use SGD [1] with a mini-batch size of 256 to train the network. The initial learning rate is set as 0.1 and weight decay of 0.0001 with a momentum of 0.9. The learning rate is divided by 10 at 30 and 60 epochs. The training is terminated after 90 epochs.

1.2. Fine-grained image classification

We follow the training pipeline from [11] to choose ResNet-50 and ResNet-101 as the baseline models. The input images are resized to 448×448 for both training and testing and we apply standard augmentation for training data, *i.e*. mirror, and random cropping. The SGD [1] is used to optimize the networks. The learning rate is decayed by 0.1 after 30 and 60 epochs.

1.3. Multi-class image classification

We use ResNet-18 with the Multi-Label-Soft-Margin loss as our baseline model. Cyclic cosine annealing [5] with the cycle of 1 is used to speed up the training. The total number of training epochs is 20.

2. Multi-label image classification results

For PASCAL VOC 2012, besides the mean Average Precision (mAP) shown in the main paper, we also provide the results for each category in Table 2. We notice that ResNet-18 guided by our ICASC$_{\mathcal{A}_{ch}}$ supervision gives the best performance in most of the categories, resulting in the best overall mAP score. When using Grad-CAM [9] as the attention guidance, the ICASC$_{Grad-CAM}$ also outperforms the baseline method ResNet-18, which further validates

dataset	CIFAR-100 [6]	Caltech-256 [3]	ILSVRC2012 [8]	CUB-200-2011 [12]	PASCAL VOC 2012 [2]
# classes	100	256	1000	200	20
image size	32×32	299×299	224×224	448×448	299×299
# images	60000	30607	∼1.3M	11788	15000
# training images	50000	7680/15360	1.2M	5994	5717
# testing images	10000	6400	50000	5794	5823
training batch size	128	16	256	10	16
weight decay	0.0005	10^{-3}	10^{-4}	10^{-4}	10^{-3}
momentum	0.9	0.9	0.9	0.9	0.9
initial learning rate	0.1	0.01	0.1	10^{-3}	0.01
# training epochs	160	20	90	90	20
evaluation metric	Top-1 Accuracy	Top-1 Accuracy	Top-1 Accuracy	Top-1 Accuracy	mean Average Precision

Table 1: The details of the dateset and training parameters.

class\method	ResNet18	+ ICASC$_{Grad-CAM}$	+ ICASC$_{\mathcal{A}_{ch}}$
aeroplane	95.16	96.33	**96.85**
bicycle	76.18	80.82	**82.41**
bird	92.92	94.69	**95.17**
boat	84.82	87.91	**89.13**
bottle	53.32	60.66	**61.07**
bus	89.81	91.73	**92.25**
car	77.41	80.11	**81.74**
cat	93.91	95.63	**96.28**
chair	68.53	73.04	**73.69**
cow	57.41	67.85	**71.12**
diningtable	67.35	73.07	**73.64**
dog	88.18	91.70	**92.62**
horse	73.58	80.89	**84.21**
motorbike	82.36	86.09	**87.51**
person	95.69	96.22	**96.44**
pottedplant	46.38	56.27	**57.75**
sheep	78.79	84.93	**86.15**
sofa	54.83	64.00	**64.63**
train	92.05	95.05	**95.56**
tvmonitor	80.07	**85.32**	85.24
mAP	77.44	82.12	**83.17**

Table 2: Categorical and mean Average Precision (mAP) (%) for our PASCAL VOC 2012 image classification experiment. The highest scoring entry in each row is shown in bold.

the effectiveness of our proposed attention-driven learning framework ICASC.

3. Additional qualitative results

We show additional qualitative results for our proposed method in Figure 1 and Figure 2. Each figure shows four examples, where for each example, we show the input image and the ground-truth class in the first column, the top-5 categorical attention maps for the baseline in the top row of the adjacent columns, and those with our approach in the bottom row. In Figure 1, where the images are in high resolution, the baseline method is ResNet-18 and our method is ResNet-18 + ICASC$_{\mathcal{A}_{ch}}$. In Figure 2, the baseline method is ResNet-110 and our method is ResNet-110 + ICASC$_{\mathcal{A}_{ch}}$. In all the figures, the ground-truth class attention map is marked using a red bounding box. There will be no marked attention map if the ground-truth class is not in the top-5 predictions. These figures show that our discriminative attention achieves better attention separability, with our model attending to regions that tell different categories apart. On the other hand, we observe visual confusion with the baseline, with high responses in the attention maps located at similar spatial locations among different categories.

As can be seen from these figures, since discriminative attention is our principled learning objective, attention responses given by our method across the top-5 categories are more separable than those from the baseline method, and our trained model is able to attend to semantically discriminative parts of the ground-truth objects, resulting in the better classification results. For example, in the top left "cake" example in Figure 1, for both "cake" and "fried egg," the baseline method attends to the central areas, containing the fruits and the cream around, which leads to visual confusion and misclassification of the image as "fried egg," whereas our method attends to the central part (fruits and cream) for "cake" and the right part (cream) for "fried egg," classifying the image as "cake" correctly. Additionally, in Figure 2, our method brings the ground-truth class to the top-1 which is out of top-5 predictions in the baseline method.

References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 1

[2] Mark Everingham, Luc Van Gool, Chris Williams,

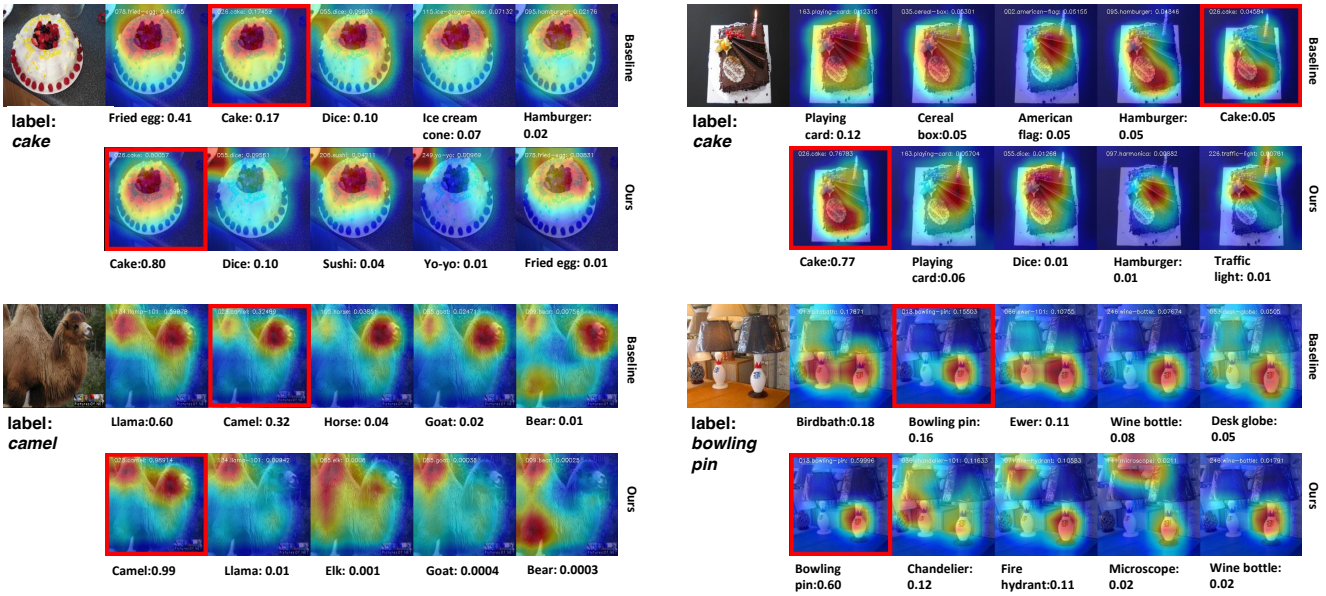

Figure 1: Improvements in top-1 predictions with our method (ResNet-18 + ICASC$_{\mathcal{A}_{ch}}$) when compared to the baseline (ResNet-18). Top row: ResNet-18; bottom row: ResNet-18 + ICASC$_{\mathcal{A}_{ch}}$.

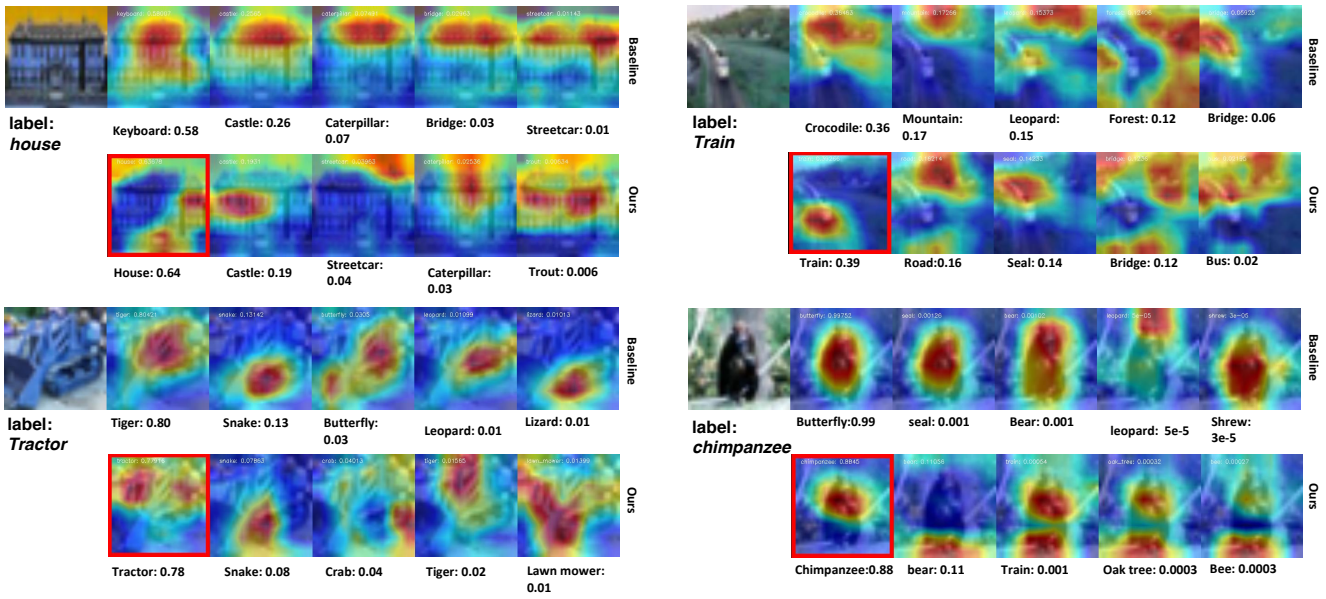

Figure 2: Improvements in top-1 predictions with our method (ResNet-110 + ICASC$_{\mathcal{A}_{ch}}$) when compared to the baseline (ResNet-110). Top row: ResNet-110; bottom row: ResNet-110 + ICASC$_{\mathcal{A}_{ch}}$.

John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 1, 2

[3] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 1, 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *ICLR*, 2017. 1

[6] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Cite-

seer, 2009. 1, 2

[7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 1

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 2

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 1

[10] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations Workshop*, 2014. 1

[11] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, 2018. 1

[12] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 2