Supplementary Material: Large-Scale Few-Shot Learning via Multi-Modal Knowledge Discovery

Shuo Wang[1,3], Jun Yue[3], Jianzhuang Liu[3], Qi Tian[4], and Meng Wang[1,2]

[1] School of Computer Science and Information Engineering,
Hefei University of Technology
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3] Noah's Ark Lab, Huawei Technologies
[4] Huawei Cloud BU

1 Visual Knowledge Discovery

In this section, we provide more visual responsive results from the three independent CNNs, Ω_o, Ω_f, and Ω_b. As shown in Fig. A(a), for many base samples, such as "Mousetrap", "Hamster", "Balloon", "French Horn", "Drake", "Acoustic Guitar", "Tricycle", and "Radio Telescope", it is easy to see that Ω_o and Ω_f focus on the regions of the objects, and Ω_b concentrates on the bodies or edges of these objects. By the way, in the "Acoustic Guitar" image, Ω_b can find more guitars with their shapes in the background. On the other hand, these CNNs perform differently on many novel samples as shown in Fig. A(b) and Fig. A(c). For the novel samples in Fig. A(b), the responses of Ω_o are deviated from the objects like "Scuba Diver", "Beigel", "Palace", "Mailbox", "Goblet", "Cicada", "Basketball", and "Pencil Sharpener". When Ω_f is used, we can see that the responses on these instances are shifted to the bodies of the objects. Then we show the importance of Ω_b in Fig. A(c). For many other novel samples, such as "Cassette", "Valley", "Parachute", "Space Heater", "Marimba", "Radiator", "Microwave Oven", and "Home Theater" images, the objects may be segmented as the backgrounds by the unsupervised saliency detection [4]. Thus, Ω_b is necessary to extract useful features from the backgrounds in these cases. It is worth mentioning that, as shown in Fig. A(b), for the objects "Palace", "Mailbox", "Basketball", and "Pencil Sharpener", the responses of both Ω_f and Ω_b are useful to describe the objects.

2 Textual Knowledge Discovery

In this section, we list more examples of the results by the network with the textual knowledge discovery (denoted as "$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Semantic}}$") and the network without it (denoted as "$\mathcal{L}_{\text{CE}}$ Only") in Fig. B. Compared with the predictions of "$\mathcal{L}_{\text{CE}}$ Only", the predicted results of "$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Semantic}}$" are more relevant to the input objects. For example, when the input novel image is a kind of car ("Tow

Fig. A. The responsive regions of three CNNs (ResNets-50 [2]) visualized by Grad-CAM [3] from several novel samples in ImageNet-FS [1].

Novel Samples	Method	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7
Tow Truck	$\mathcal{L}_{CE} + \mathcal{L}_{Semantic}$	Tow Truck	Police Van	Passenger Car	Trolleybus	Police Van	Fountain	Catamaran
	$\mathcal{L}_{CE}$ Only	Police Van	Trolleybus	Catamaran	Fountain	Bobsled	Drum	Tow Truck
Electric Guitar	$\mathcal{L}_{CE} + \mathcal{L}_{Semantic}$	Electric Guitar	Banjo	Accordion	Fountain Pen	Projector	Cornet	Jersey
	$\mathcal{L}_{CE}$ Only	Bikini	Projector	Fountain Pen	Swing	Jersey	Accordion	Doormat
Labrador Retriever	$\mathcal{L}_{CE} + \mathcal{L}_{Semantic}$	Labrador Retriever	Chesapeake Dog	Kuvasz	Weimaraner	Pug	Samoyed	Paper Towel
	$\mathcal{L}_{CE}$ Only	Paper Towel	Chesapeake Dog	Pembroke	Lion	Red Wolf	Oxcart	Labrador Retriever
Cornet	$\mathcal{L}_{CE} + \mathcal{L}_{Semantic}$	Cornet	Banjo	Ladle	Swing	Maypole	Accordion	Barbell
	$\mathcal{L}_{CE}$ Only	Swing	Banjo	Maypole	Totem Pole	Barbell	Bow	Ladle
Fire Salamander	$\mathcal{L}_{CE} + \mathcal{L}_{Semantic}$	Fire Salamander	Tailed Frog	Eft	Snail	Spiny Lobster	Tree Frog	Ringneck Snake
	$\mathcal{L}_{CE}$ Only	Spiny Lobster	Gila Monster	Tailed Frog	Snail	Eft	Fire Salamander	Pineapple
Yorkshire Terrier	$\mathcal{L}_{CE} + \mathcal{L}_{Semantic}$	Yorkshire Terrier	Norwich Terrier	Standard Schnauzer	Giant Schnauzer	Irish Terrier	Wheaten Terrier	African Hunting Dog
	$\mathcal{L}_{CE}$ Only	Paper Towel	Tarantula	Standard Schnauzer	Otter	Norwich Terrier	Yorkshire Terrier	Irish Terrier
Ptarmigan	$\mathcal{L}_{CE} + \mathcal{L}_{Semantic}$	Ptarmigan	Limpkin	Oystercatcher	Ant	Horizontal Bar	Cock	Bustard
	$\mathcal{L}_{CE}$ Only	Ant	Bonnet	Gong	Teddy	Horizontal Bar	Monarch	Swing
Cornet	$\mathcal{L}_{CE} + \mathcal{L}_{Semantic}$	Cornet	Banjo	Accordion	Ladle	Swing	Electric Guitar	Violin
	$\mathcal{L}_{CE}$ Only	Accordion	Swing	Banjo	Shower Cap	Horizontal Bar	Cornet	Jersey
Ant	$\mathcal{L}_{CE} + \mathcal{L}_{Semantic}$	Ant	Monarch	Fly	Cabbage Butterfly	Lycaenid	Hip	Cricket
	$\mathcal{L}_{CE}$ Only	Bell Pepper	Hip	Monarch	Spiny Lobster	Fly	Lycaenid	Broccoli

Fig. B. The recognition results of several novel samples by the networks with and without the textual knowledge discovery, denoted as "$\mathcal{L}_{CE} + \mathcal{L}_{Semantic}$" and "$\mathcal{L}_{CE}$ Only", respectively. In this experiment, $K = 1$. We randomly select one image from each label category for easy understanding of the objects corresponding to the labels.

Truck" here), all the top 5 results by "$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Semantic}}$" are car labels. Although the 6th and the 7th results ("Fountain" and "Catamaran") by "$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Semantic}}$" are not relevant, its overall ranking of the top 7 results is better than that by "$\mathcal{L}_{\text{CE}}$ Only".

References

1. Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. In: ICCV (2017)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
3. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
4. Zhang, J., Zhang, T., Dai, Y., Harandi, M., Hartley, R.: Deep unsupervised saliency detection: A multiple noisy labeling perspective. In: CVPR (2018)