Spatial Attention Network for Few-Shot Learning

Xianhao He*, Peng Qiao, Yong Dou and Xin Niu

National Laboratory for Parallel and Distributed Processing,
National University of Defense Technology

Introduction

Recently, deep learning models such as AlexNet, VGG and ResNet have achieved great success in image classification tasks. However, these models are trained in a supervised manner using large amounts of labeled data. Moreover, these models can only recognize images from specific classes appearing in training data. Furthermore, in a case that we have few training samples from some classes, models trained using these data would perform poorly due to overfitting. Many attempts such as fine-tuning, data augmentation and dropout are proposed to alleviate overfitting, however, this problem still exists. Focusing on above issues, One- or Few-Shot Learning aims to learn new knowledge from one or few instances under the inspiration of human's quick learning ability.

Methodology

We assume that the whole dataset can be divided into D_{base} and D_{novel}. Each category of D_{base} is rich in image examples, while each category of D_{novel} has fewer images. We ensure that D_{base} and D_{novel} have disjoint label space. For D_{novel} set, we randomly sample C disparate classes with M examples for each class, and these data consist of *support set.* In testing episodes, *query* images are made up of n samples with prior knowledge that they have identical label space $\{label_1, label_2, \dots, label_C\}$ with *support set,* and images in *query set* are unseen in *support set.* The task of C-way M-shot classification is to assign labels $\{y_i, y_i \in \{label_1, label_2, \dots, label_C\}\}_{i=1}^{n}$ to images $\{x_i\}_{i=1}^{n}$ in *query set*.

An overview of our network architecture is presented in Fig 1. Our network has two stages: *feature extractor* and *few-shot learning* stage. In *feature extractor* stage , we train a ConvNet based on D_{base} to learn extracting features from input images. In *few-shot learning* stage, attention module is trained to identify informative regions related to target objects.

The purpose of SAN is to find informative regions of input image. To achieve this goal, we fully exploit spatial context of an image via clustering and attention module.

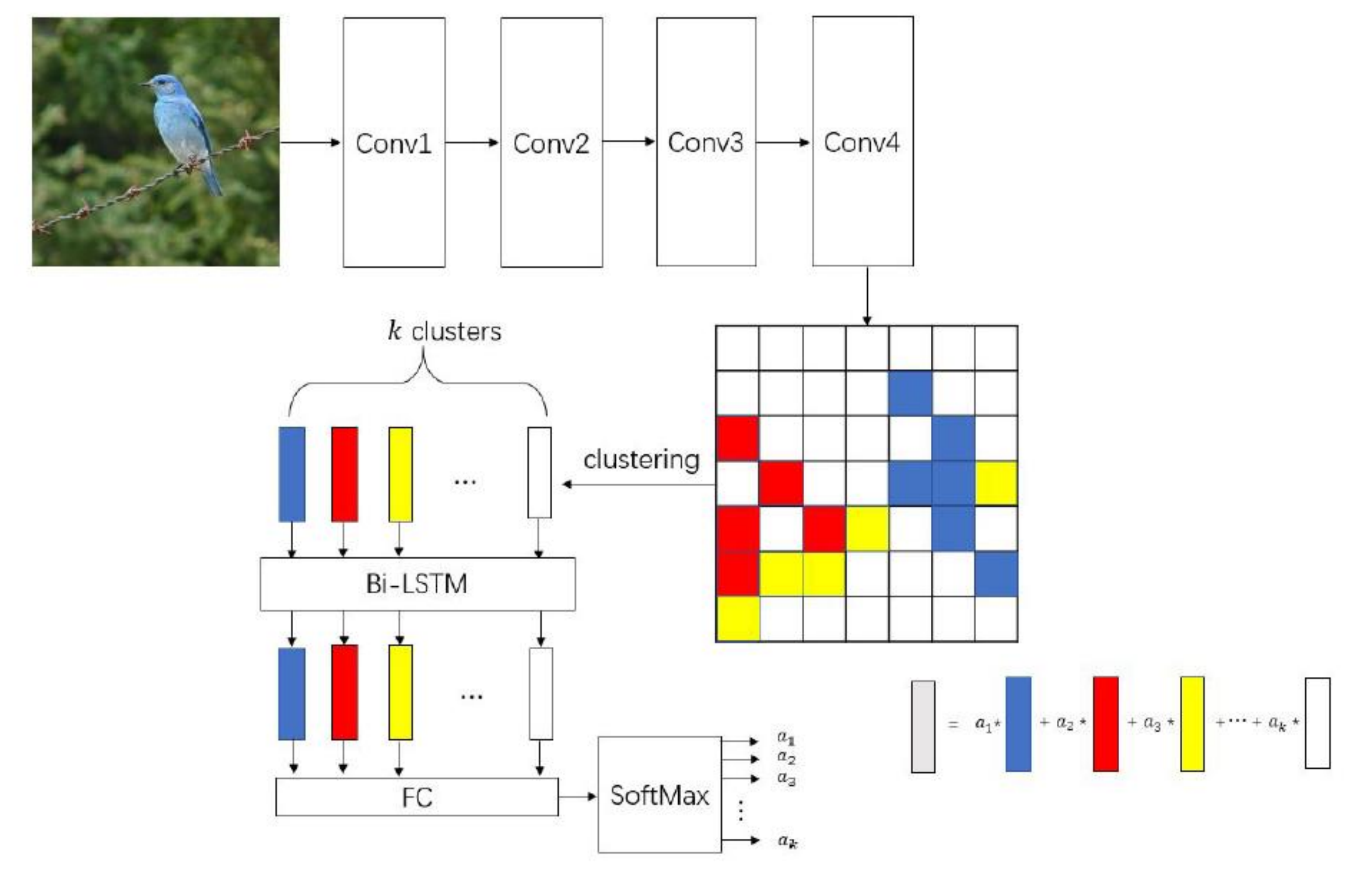


Fig. 1. An illustration of the overview architecture of proposed spatial attention network. It consists of feature extractor and attention module. The figure illustrates an example of attention module applied on one layer in Conv4 Block. Feature map cells drawn with identical color belong to identical cluster. The output feature(gray) is weighted sum of regional features.

Experiment Result

We conduct 5-way 1 and 5-shot classification in all our experiments. Three public benchmark datasets in our experiments are *mini*ImageNet, Caltech-UCSD Birds and *mini*DogsNet.

Prior to experiments, we need to choose cluster number for algorithm. We post the result of selecting cluster number on Table 1, and we choose the best one in later experiments.

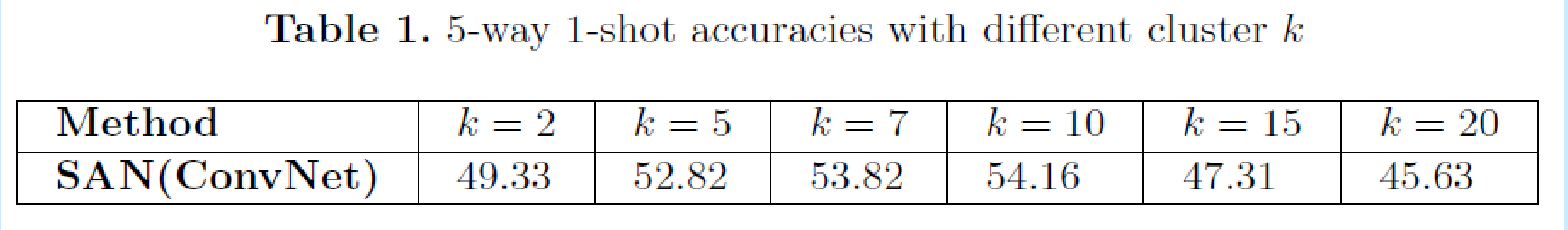


Table 1. 5-way 1-shot accuracies with different cluster k

Method	$k = 2$	$k = 5$	$k = 7$	$k = 10$	$k = 15$	$k = 20$
SAN(ConvNet)	49.33	52.82	53.82	54.16	47.31	45.63

The result of 5-way 1 and 5-shot classification results can be seen on Table 2.

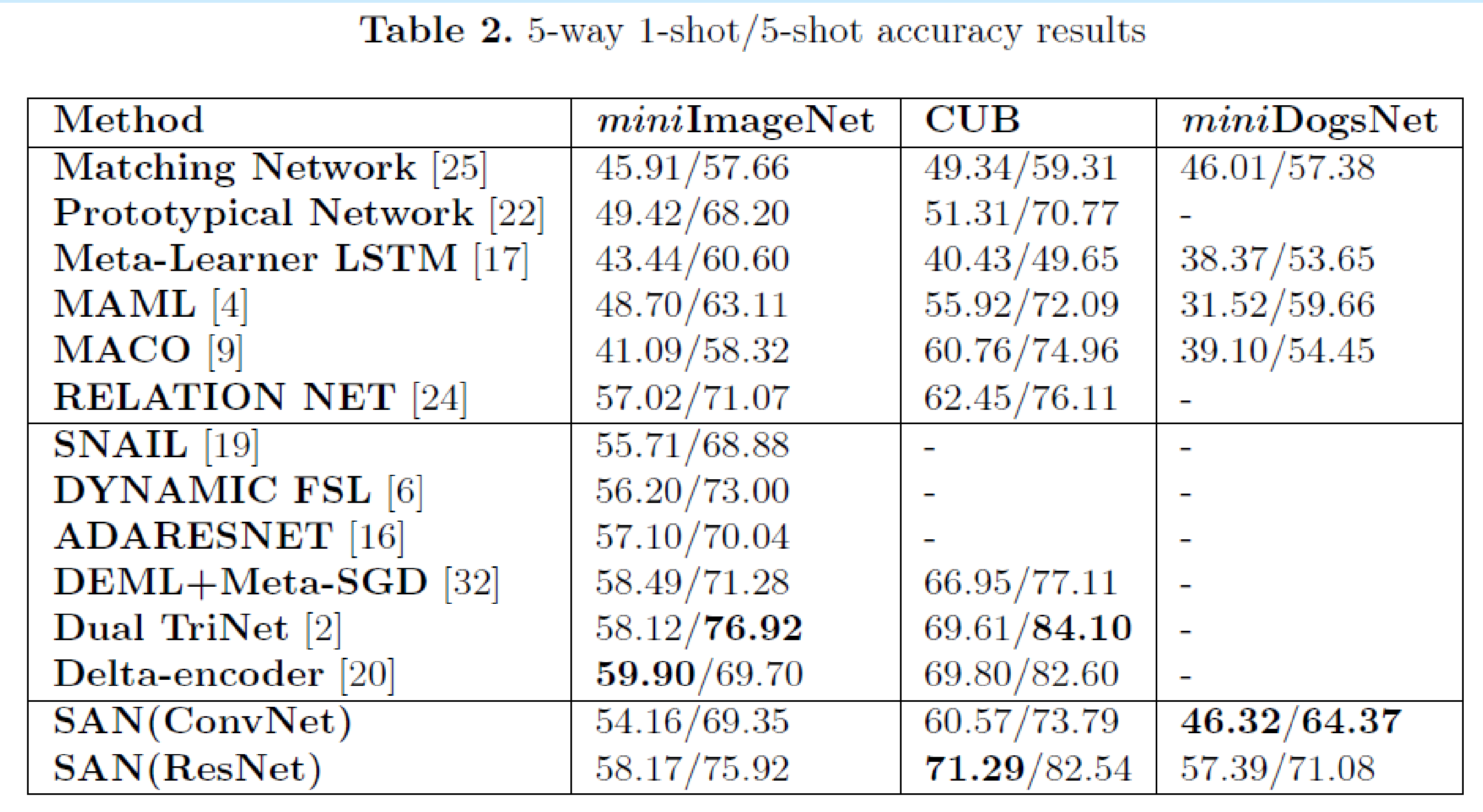
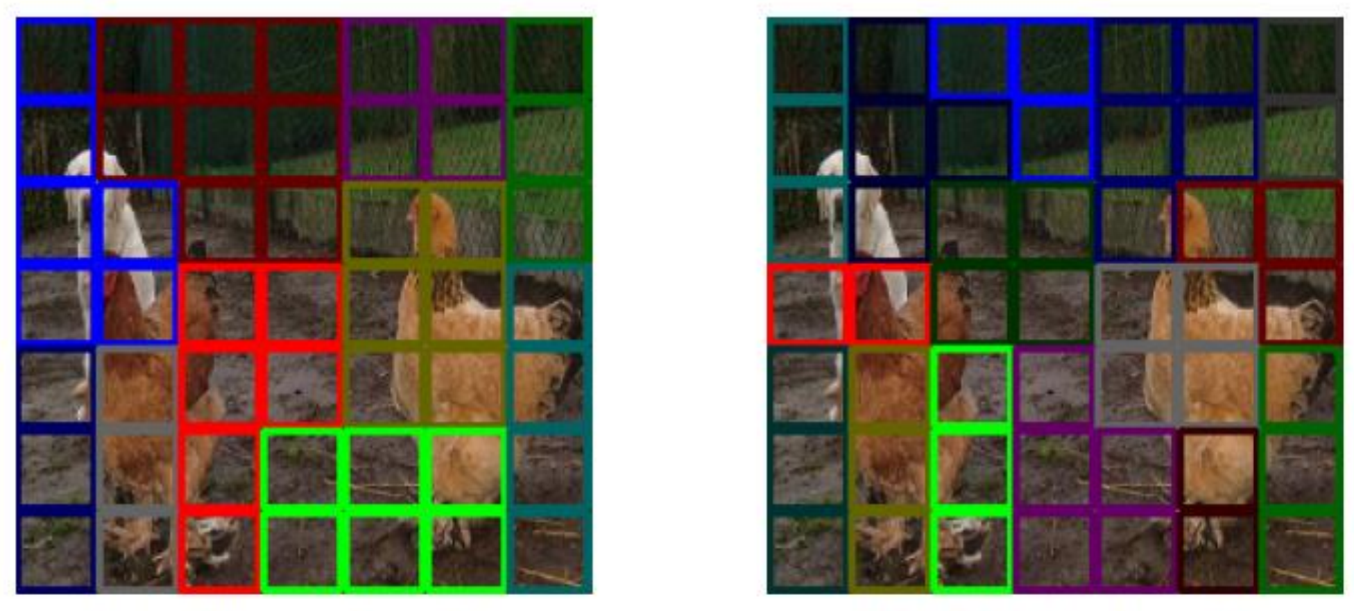


Table 2. 5-way 1-shot/5-shot accuracy results

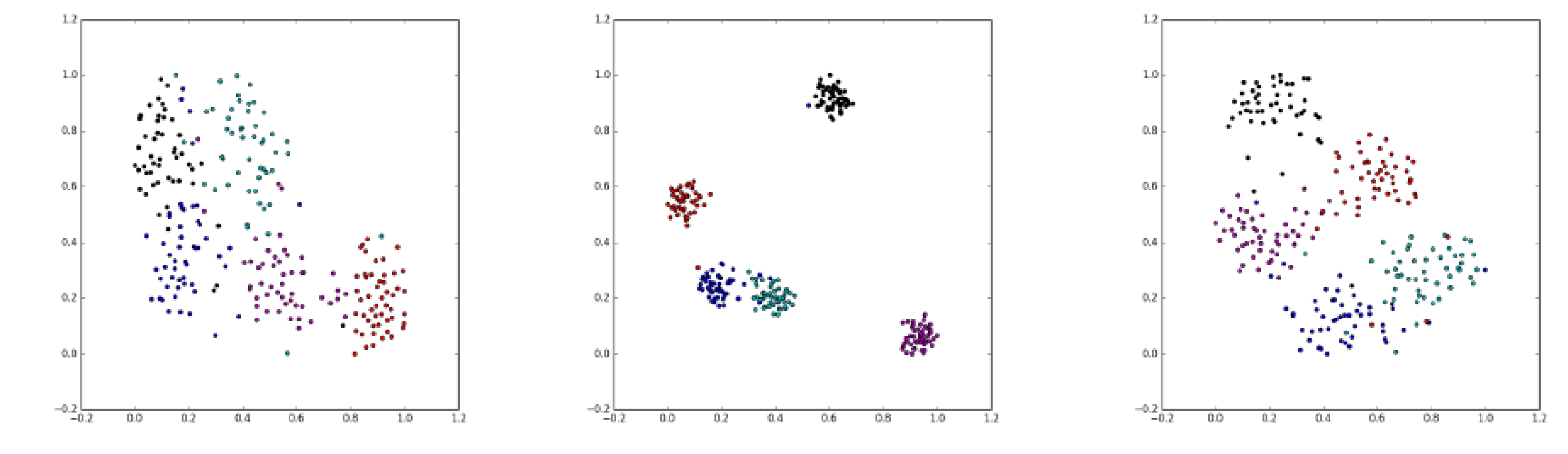
Method	*mini*ImageNet	CUB	*mini*DogsNet
Matching Network [25]	45.91/57.66	49.34/59.31	46.01/57.38
Prototypical Network [22]	49.42/68.20	51.31/70.77	-
Meta-Learner LSTM [17]	43.44/60.60	40.43/49.65	38.37/53.65
MAML [4]	48.70/63.11	55.92/72.09	31.52/59.66
MACO [9]	41.09/58.32	60.76/74.96	39.10/54.45
RELATION NET [24]	57.02/71.07	62.45/76.11	-
SNAIL [19]	55.71/68.88	-	-
DYNAMIC FSL [6]	56.20/73.00	-	-
ADARESNET [16]	57.10/70.04	-	-
DEML+Meta-SGD [32]	58.49/71.28	66.95/77.11	-
Dual TriNet [2]	58.12/**76.92**	69.61/**84.10**	-
Delta-encoder [20]	**59.90**/69.70	69.80/82.60	-
SAN(ConvNet)	54.16/69.35	60.57/73.79	**46.32**/**64.37**
SAN(ResNet)	58.17/75.92	**71.29**/82.54	57.39/71.08

To figure out what attention module learns, we conduct following three visualizations: cluster visualization, feature embedding visualization and attention module visualization, respectively.



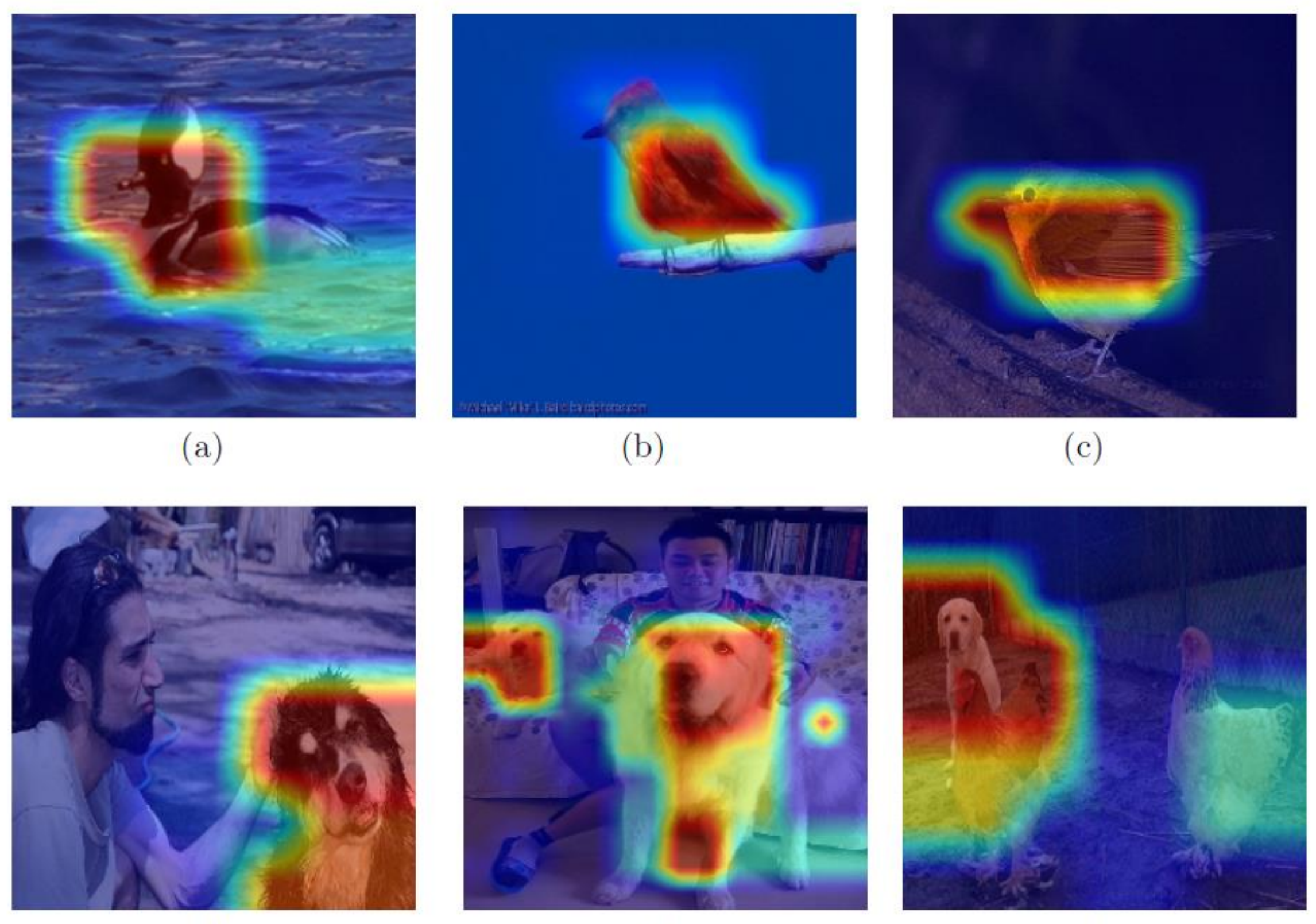
(a) Cluster $k = 10$ (b) Cluster $k = 15$

Fig. 2. Cluster Visialization. Fig 2(a) and 2(b) represent clusters of image contents in dog category. Bounding boxes with different colour indicate different clusters.



(a) *mini*ImageNet (b) CUB (c) *mini*DogsNet

Fig. 3. 2D embedding figure. Figure 3(a), 3(b), and 3(c) represent embeddings in *mini*ImageNet, CUB and *mini*DogsNet respectively. Points with the same color come from identical category.



(a) (b) (c)

(d) (e) (f)

Fig. 4. Visualization of attention module. Fig 4(a), 4(b), 4(c) are from bird category, and Fig 4(d), 4(e), 4(f) are from dog category

Conclusion

- Spatial Attention Network applies attention module onfeature map to generate discriminative features.
- In the embedding space, features in identical category are grouped together.
- Attention module tries to focus on salient image area with target objects.

Contact us

✉

- Xianhao He: hexianhao18@nudt.edu.cn
- Peng Qiao: pengqiao@nudt.edu.cn
- Prof. Yong Dou: yongdou@nudt.edu.cn
- Xin Niu: xinniu@nudt.edu.cn