DEEP MULTI-CONTEXT NETWORK FOR FINE-GRAINED VISUAL RECOGNITION

Xinyu Ou[1,2,3], *Zhen Wei*[2,4], **Hefei Ling*[1], *Si Liu*[2], *Xiaochun Cao*[2]

[1]Huazhong University of Science and Technology, School of Computer Science and Technology
[2]Chinese Academy of Sciences, Institute of Information Engineering, SKLOIS
[3]Yunnan Open University, YNGBZX
[4]University of Electronic Science and Technology of China, Yingcai Experimental School
{ouxinyu,lhefei}@hust.edu.cn, zhen.wei@hotmail.com, {liusi, caoxiaochun}@iie.ac.cn

ABSTRACT

In this paper, we tackle the FINE-GRAINED VISUAL RECOGNITION problem by proposing a deep multi-context framework. We employ deep Convolutional Neural Networks to model features of objects in images. Global context and local context are both taken into consideration, and are jointly modeled in a unified multi-context deep learning framework. To cleanse the relatively dirty data for training, a regional proposal method is designed to make the multi-context modeling suited for fine-grained visual recognition in the real world. Furthermore, recently proposed contemporary deep models are used, and their combination is investigated. Our approaches are evaluated on MSR-IRC 2016 and further assessed on the more complex validation set. The results show significant and consistent improvements over the baseline.

Index Terms— Multi-Context, Object Proposal with Multi-Crop, Multi-Model, Fine-Grained

1. INTRODUCTION

Thanks to the advance of deep learning algorithms, significant progresses have been made in visual recognition [1–3] in the past several years. But, there is still a considerable gap from these academic innovations and practical intelligent services. Fine-grained recognition task such as identifying the breed of a dog, is quite challenging because the visual differences between the categories are small and can be easily overwhelmed by those caused by factors such as pose, viewpoint, or location of the object in the image. Specifically, the difficultly of fine-grained classification comes from the fact that discriminative features are localized not just on a foreground object, but more importantly on object parts. For examples, the differences between "Siberian husky and "Alaskan husky are the shape of ears, but this characteristic not very obvious.

For this work, a bottom-up process is to propose image regions that contain parts of certain objects. The parts are often defined manually and the detectors are trained in a supervised manner. Selective search or EdgeBox are used for proposing such regions. Recently variants of such models [4] shown significantly improve over earlier work. A drawback of these approaches is that annotating parts is significantly more challenging than collecting image labels, and manually defined parts may not be suit for the recognition task.

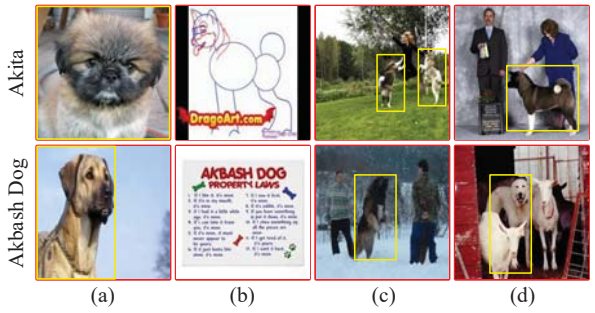

Fig. 1. Examples to show importance of context in real-world datasets.

In real-world, data is always complicated and changeable, a problem is how to find where the object is. An appropriate scope of context is very important to help an attentive object stand out from image meanwhile keep those non-salient objects suppressed in background. In Figure 1(c-d), high-level knowledge tells us information about dogs, persons, sheep, blackboard and trees, but can not answer which are attentive objects. If a local context(the yellow boxes) is adopted in order to determine the attention then all these objects are highlighted as attentive objects. This becomes general object detection problems, and finding attentive objects can be regarded as a regional proposal process. Due to the extremely large variation in positions and scales of objects of different categories, a global context (including the full image) is more suitable to determine object semantics. Because a global context takes all objects in an image into account, and only with a global context we can model the contrast between all objects. As shown in Figure 1(d-top), if the "Akita" and the person,

Zhen Wei contributed equally to this work and should be considered co-first authors.

Hefei Ling is the corresponding author.

978-1-4799-7082-7/15/$31.00

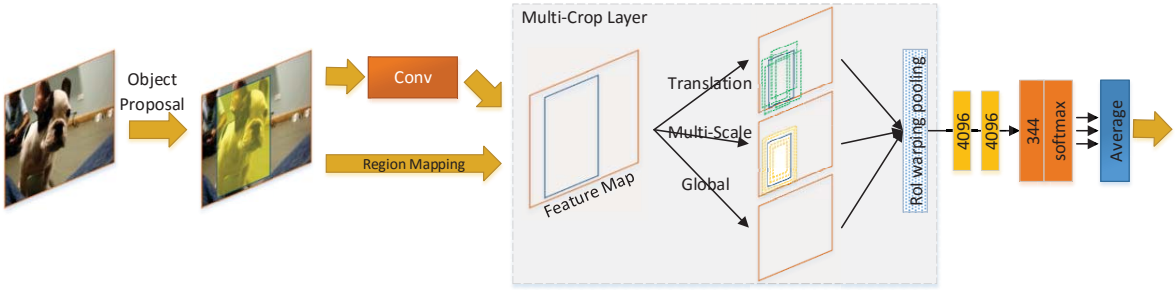

Fig. 2. Illustration of the proposed architecture for Fine-Grained Visual Recognition.

blackboard are considered together, then only the "Akita" is classified as the attentive object. In addition, it is known that deep models are also powerful in learning global contextual.

Based on the above motivations, a deep multi-context framework is proposed, we have three major contributions:

1. An integrated deep model with *multi-context* is designed to produce object features. The global context is utilized to model features in full-image, while the local context is used for prediction in objective areas.

2. To filter out noisy data and prevent over-fitting, a *multi-crop* based object proposal strategy is proposed to generate variety of samples in training and testing.

3. In order to make full use of the complementarity of different models, *multi-model* combination is used to revise the prediction bias of single model.

2. METHODS

An elegant and effective solution is generated feature expression of the original images, where computation costs are nearly equal to one forward for all of the context regions. By sharing convolutions, the marginal cost for computing feature is small(e.g., 0.3s per image). For the convolution network, the vanilla model we used in MSR-IRC 2016, i.e. the VGG-16 [2], is adopted. It is flexible to incorporate any of other contemporary deep models (such as AlexNet [1], GoogLeNet, ResNet [3]) into to our framework, and in this paper we investigate combine these contemporary architectures. Figure 2 illustrates an overview of our proposed network architecture.

2.1. Multi-Context Model

While the global-context model aims to robustly model object feature with few large errors, a large amount of noise data and confused background will hurt the accuracy; local-context is designed to look at details of objects, it focuses on a smaller scope as related to RoI as input to refine the object prediction without interference information. For all the context region, we perform RoI pooling by a differentiable RoI warping layer [5] followed by standard max pooling. The operation of RoI warping layer can be written as a linear transformation on the full-image feature map $F(\theta)$:

$$F_i^{RoI}(\theta) = G(B_i(\theta))F(\theta)) \tag{1}$$

Here $F(\theta)$ is reshaped as a n-dimensional vector, with $n = W \times H$ for a full-image feature map of a spatial size which is the last convolutional layer. Given a predicted box $B_i(\theta)$ centered at $(x_i(\theta), y_i(\theta)$ with width $w_i(\theta)$ and height $h_i(\theta)$, an RoI warping layer interpolates the features inside the box and outputs a feature $F_i^{RoI}(\theta)$ of a fixed spatial resolution $W' \times H'$ by $G(\star)$ (28×28 in this paper). After the ROI warping layer, a max pooling layer is then applied to produce a lower-resolution output, e.g., 7×7 for VGG-16.

In this paper, the local-context model shares the same deep structure with the global-context model, other deep structures can also be flexibly incorporated in the context model. Overall, prediction of an arbitrary image input is performed by estimating the fusion probability:

$$score(I_{lc}, I_{gc}) = P(C|I_{lc}, I_{gc}; \theta)) \tag{2}$$

where I_{lc} and I_{gc} are output of the last layer of the local context model and the global context model respectively. C is the prediction of softmax over the total categories. In our approach, the parameters in our framework can be decomposed to several parts, i.e.$\theta = \{W_{lc}, W_{gc}, \alpha, \beta\}$, where W_{lc} are last-layer parameters for local-context modeling, W_{gc} are last-layer parameters in the neural network for global-context modeling, and α, β are parameters of an ambiguity modeling function controlling the weights of global-context modeling and local-context modeling(typical $\alpha = \beta = 1$ in this paper).

2.2. Object Proposal with Multi-Crop

In order to solve the effects of noise data, especially to those who do not belong to the dog breed, see in Figure 1 (b). A simple way is to make use of detection to reduce the interference of outlier samples. Due to the dataset does not include any bounding box, we cannot train the detection network with

multi-context network together. As an alternative, we utilize Faster-RCNN [6] as a preprocessing to generate valid data. This approach is used in both training and testing, it improves accuracy over 10 points.

Furthermore, we use the proposal region to achieve data augmentation, we called it object proposal with multi-crop. Data augmentation is commonly with CNNs to reduce overfitting and enhanced diversity. We performed it by replicating the region with a number of transformations. Specifically, for each RoI, we rescale ($k \in [0.8, 1.2]$) and systematic combinations of horizontal and vertical shifts ($(\delta x, \delta y) \in [-20, 20]$). In the training, we randomly selected $9 \times k$ and $16 \times (\delta x, \delta y)$, while $5 \times k$ and $4 \times (\delta x, \delta y)$ are defined in testing.

2.3. Multi-Model Combination

In this section, we will further elaborate on constructing a multi-model combination framework. The motivation for combining different models lies on the fact that a suitable classifier relies heavily on a robust image representation, such as Bilinear model [7]. Contrastively, in our framework individual deep models are fine-tuned separately first in order to use an optimum performance of each model. Then a pooling operation is adapted over every prediction to realize a combination procedure and obtains the final score. The whole process is formulated as:

$$f_{combination}(I) = pooling(f_1(I), f_2(I), ..., f_k(I)) \quad (3)$$

where I is the input image, $f_k(I), (k = 1, 2, 3, ...)$ are inference results of each model, average-pooling is used in this paper, and $f_{combination}(I)$ is the final score.

Comparing to other methods, the multi-model consultation framework provides more flexibilities with a boosted performance: it has an arbitrary number of models and the pooling kernel can be either fixed or learned (such as linear combination) depending on specific tasks and consultant models.

3. DATASETS AND EXPERIMENTS

3.1. Benchmark Datasets

To further motivate and challenge the academic and industrial research community, the large-scale real-world **Clickture-Dog** dataset is used in MSR-IRC. Different from regular benchmark dataset, it has a lot of noise. There are $\sim$10% images are invalid assess by object proposal. Most of these images are not dog breed or object is relatively small. Besides, the long tail of data distribution is very serious, most categories are very few even only 1 picture, while some are more than 5,000. Another serious problem is many images have the wrong label. Figure 1 show some examples. In order to alleviate these problems, we add some reliable images prudently with three extra benchmark datasets(Columbia Dogs [1], Stanford Dog [2], The Oxford-IIIT Pet [3]).

In all, the final dataset involves 344 dog breeds, 85,279 training images, and 36,548 validation images. In the official competition, the evaluation set include $\sim$100 categories, $\sim$10,000 images. Very different from our validation set, the evaluation set does not include noise data.

Table 1. The competition results of MSR-IRC 2016.

rank	team	Precision@5	Description
1	NLPR_CASIA	89.65%	
2	ybt_bj	86.90%	
3	NFS2016	85.00%	
4	WestMountain	84.75%	Multi-Model*
5	rucmm	84.55%	
6	CASIIE-Asgard	83.4%	Local-Context
	...		
10	FrenchBulldog	71.25%	

Note *: The Multi-Model used in MSR-IRC is only achieved multi-model consultation with different local-context models.

3.2. Implementation Details

Our implementation follows the practice in [1, 2]. The randomly crop, horizontal flip, and per-channel mean subtracted are used. All models are trained by fine-tuning for 60 epoches. We use SGD with a mini-batch size of 64, a weight decay of 0.0001 and a momentum of 0.9. The learning rate starts from 0.01 and was divided by 10. In testing, standard 10-view crop is used in each region.

3.3. Results on MSR-IRC Dog Evaluation

Our results on MSR-IRC are summarized in Table 1. We sent two teams to take part in the competition, respectively "WestMountain" and "CASIIE-Asgard".

The "CASIIE-Asgard" employed local-context to modeling network. Firstly, Faster RCNN is utilized to filter the raw images, then follow a VGG-16 to implement classification. More specially, most of the time we only input the detection region to the CNN. However, it is obvious that some of the test image without any detection results. For these images, we directly use the original images as input to generate a prediction. Under these settings, we achieve 83.4% top-5 accuracy.

For "WestMountain", a simplified multi-model is used. After the Faster RCNN regions proposal, a "three-stream" architectures has been used to analyze the image. Specifically, we used VGG-16, VGG-19 and ResNet-152 as fusion models with the average pooling for score fusion. Note that, this

[1] http://www.robots.ox.ac.uk/$\sim$vgg/data/pets/
[2] http://vision.stanford.edu/aditya86/ImageNetDogs/
[3] http://www.robots.ox.ac.uk/$\sim$vgg/data/pets/

Table 2. The results of our methods on validation set.

Methods	Precision@1	Precision@5
Local-Context w/o MC	64.76%	82.00%
Global-Context w/o MC	62.19%	83.68%
Multi-Context w/o MC	64.77%	86.09%
Local-Context w/ MC	64.82%	82.14%
Global-Context w/ MC	62.19%	83.68%
Multi-Context w/ MC	65.37%	86.29%
Multi-Model w/ MC	68.02%	87.83%

multi-model used in MSR-IRC is only achieveed by different local-context. Without multi-crop, this model achieved an 84.75% top-5 accuracy on evaluation set.

3.4. Results on MSR IRC Dog Validation

Beyond the competition, in order to further improve the performance, more approaches have been investigated, as illustrated in Table 2. We separate the global-context without multi-crop as a baseline model denoted as "Global-Context w/o MC, it achieves top-1 and top-5 accuracy of 62.19% and 83.68%. Unfortunately, our submitted in MSR-IRC with Local-Context model is the worst models on validation set. We think the main reason is that our local-context model is mainly designed for the dataset which contains lots of noise data. However, the evaluation set is only contains dog breeds. As shown in Table 2, multi-context model consistently outperforms the single-context models on the validation set. Especially on the Top-1 accuracy, the multi-context model increases the accuracy by 3% denoted as "Multi-Model w/ MC. It is clearly shown that multi-context model refines the erroneous predictions of the single-context model.

We measure the performance of the object proposal with multi-crop method with different context strategies on validation set. In the training stage, this strategy is used to augment dataset by default. Evaluation results on most models have similar characteristics, by implementing average of all the regions, the accuracy slightly outperforms initial version.

Finally, our framework is flexible to incorporate other contemporary deep models. Similar to MSR-IRC, VGG-16, VGG-19 and ResNet-152 are used, and only convolutional layers had been replaced. But, with greatly different from previous methods, multi-context models are used for consultative, while the competition is combined with local-context models. The final model is integrated with multi-context, multi-crop and multi-model and obtained the best performance 87.83%.

4. CONCLUSIONS

In this paper, we propose a deep multi-context network to recognize fine-grained object. Firstly, we introduce a deep multi-context network for fine-grained visual recognition. Global context and local context are integrated into a unified multi-context framework for feature extraction. Global and local-context share the same deep structure and parameters. Secondly, due to the noise data source, an effective data preprocessing and augmentation method is suggested, that using a multi-crop strategy on proposal regions to eliminate noise data and increase data diversity. Finally, some recently proposed contemporary deep Convolutional networks in image classification task are employed to multi-model combination. Experiments validate the effectiveness of the component in our framework, and show our approach significantly and consistently outperform the baseline.

5. ACKNOWLEDGEMENT

This work was supported by Natural Science Foundation of China (U1536203, 61272409, 61572493), the Major Scientific and Technological Innovation Project of Hubei Province (2015AAA013), and "Strategic Priority Research Program" of the Chinese Academy of Sciences (XDA06010701).

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 2012, 2012.

[2] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *Eprint Arxiv*, 2014.

[3] Shaoqing Ren Kaiming He, Xiangyu Zhang and Jian Sun, "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016.

[4] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell, "Part-based r-cnns for fine-grained category detection," *Lecture Notes in Computer Science*, vol. 8689, pp. 834–849, 2014.

[5] Jifeng Dai, Kaiming He, and Jian Sun, "Instance-aware semantic segmentation via multi-task network cascades," *CoRR*, vol. abs/1512.04412, 2015.

[6] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[7] Tsung Yu Lin, Aruni Roychowdhury, Subhransu Maji, and Lin, "Bilinear cnn models for fine-grained visual recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1449–1457.