An Attention Model Based on Spatial Transformers for Scene Recognition

Shuxuan Guo*, Li Liu*, Wei Wang†, Songyang Lao*, and Liang Wang†, ‡,
*Science and Technology on Information Systems Engineering Laboratory
National University of Defense Technology
Email: $gsxuan6688@163.com, liuli\_nudt@nudt.edu.cn, laosongyang@vip.sina.com$
†Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition
‡Center for Excellence in Brain Science and Intelligence Technology
Institute of Automation, Chinese Academy of Sciences
Email: $\{wangwei, wangliang\}@nlpr.ia.ac.cn$

***Abstract*—Scene recognition is an important and challenging task in computer vision. We propose an end-to-end pipeline by combing CNNs with explicit attention model to determine several meaningful regions of original images for scene recognition. In the proposed pipeline, the spatial transformer network is leveraged as the attention module, which can automatically learn the scales and movements of centers of attention windows. As for feature extraction, the basic CNN architecture is utilized. Furthermore, the stronger descriptors of scenes are constructed by feature fusion. The highlight of our proposed network is that it is capable to localize discriminative regions from an image in a data-driven manner without any additional supervision. We conduct experiments on a subset of the Places205 database to evaluate the performance of the proposed basic network and the involved parameters. Our model achieves state-of-the-art top-1 accuracy 82.10% on the evaluation dataset comparing with fine-tuned PlacesCNN (80.98%). We find that our model is able to learn informative attention regions for discriminating scene categories.**

I. Introduction

Recognizing scenes is not an easy task, which is significantly attributed to the difficulty in representing scenes due to their variability, ambiguity and the wide range of illumination, where view points and scale changes may apply. As deep learning develops, a CNN trained on ImageNet [1] significantly outperforms the hand-crafted features in scene recognition. In order to further improve the performance of this task, Zhou et al. [2] introduced a large-scale database called Places to support CNNs for scene recognition, and established state-of-the-art results on several scene datasets. In [3], Zhou et al. made an attempt to understand the implicit objects detectors in CNNs. However, despite great progresses, the performance of large-scale scene recognition is still far from satisfying, and what contributes to discriminating scenes categories in images is inexplicable.

The attention mechanism subtending human visual system (HVS) is the capacity to learn and focus on distinctive samples from what humans have seen according to given visual tasks. When it comes to distinguishing scenes in images, humans do not pay attention to the entire image at once. Actually, they tend to take advantages of informative objects, regions and the relationship between them.

We propose an end-to-end CNN framework by modeling explicit visual attention for scene recognition in this paper. Our method is derived from the spatial transformer network [4] that was originally proposed to actively spatially transform feature maps in CNN. In our model, we incorporate it as attention model via detecting meaningful regions in images, then the network generates discriminative descriptors for scenes through feature fusion. To evaluate the performance of our method, we do experiments on a subset of the Places205 database.

The contributions of this paper are twofold :

1. We are the first to attempt to explore explicit attention model for scene recognition, by which the network can automatically fix its gaze on meaningful regions in images just under supervision of image-level labels.

2. We propose to combine CNN features extracted from local regions with those from original image to strengthen scene representation. Experimental results demonstrate the effectiveness of our model.

The remainder of the paper is organized as follows: related works on scene recognition and attention modeling are briefly reviewed in Section II. Details of our proposed method are described in Section III. Section IV presents and evaluates the performance of our proposed method and its variants, and visualizes some results in experiments. Finally, Section V concludes our work.

II. Related works

In this section, we review prior works covering scene recognition and visual attention modeling.

Scene Recognition: There are numerous approaches devoted to the scene recognition task. Ariadna Quattoni et al. [5] proposed a model based on hand-crafted features for indoor scene recognition. Fei-Fei Li et al. [6] built a bayesian hierarchical model for learning natural scene categories. With the development of CNNs, several different CNN architectures have been applied to recognize scenes. For example, Zhou et

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.

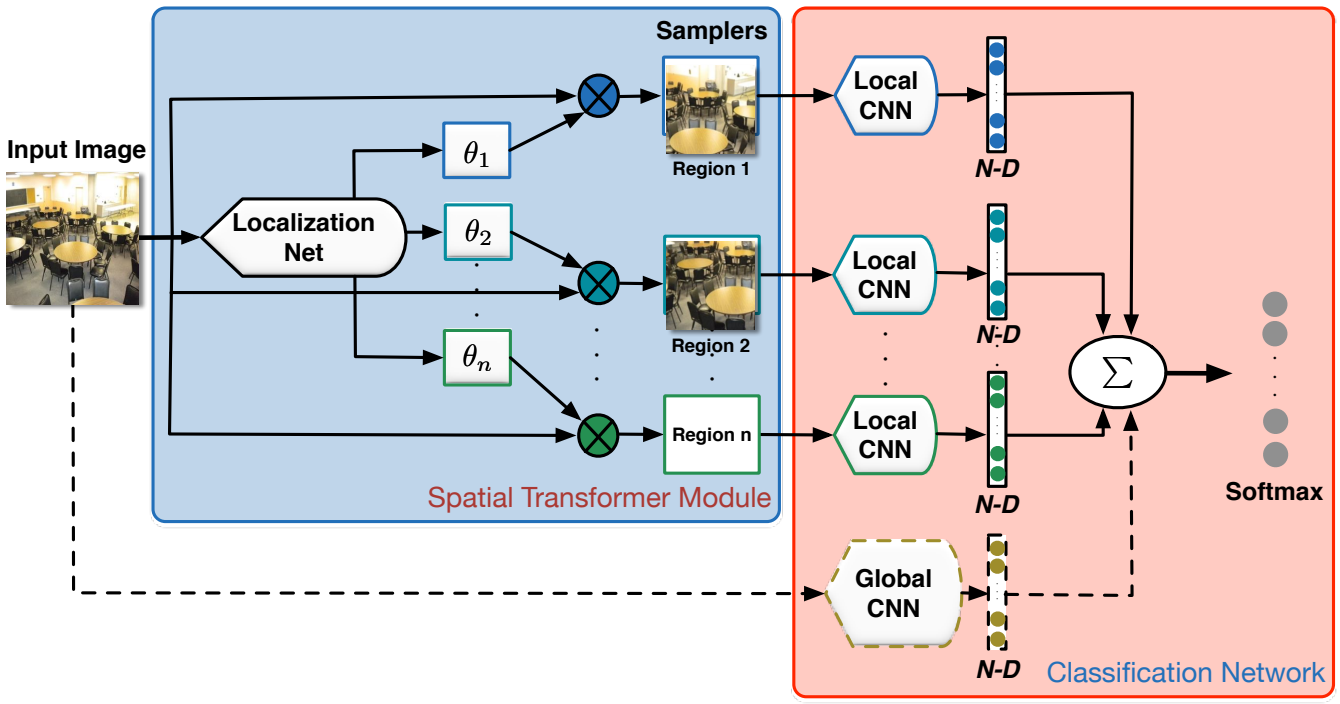

Fig. 1. The overall end-to-end framework of ST-PlacesCNN. **Spatial Transformer Module:** the localization network is shared by all n spatial transformers with the same input, which produce spatial transformation parameters θ_i $(i = 1, 2, ..., n)$. And its output is the set of detected regions. **Classification Network:** it consists of two kinds of feature extracting networks. One is for regions to extract local features, while the other is for the original input to extract global features. Afterwards, these features are fused together. Finally, there is a softmax classifier for final recognition.

al. [2] trained the classical ImageNet-CNN on a scene-centric dataset (Places), and Wang et al. [7] trained three VGGNet models, namely VGGNet-11, VGGNet-13, and VGGNet-16 on the large-scale Places205 dataset. Recent works on scene recognition are mainly object-centered methods, exploring how to utilize candidate objects or region proposals as precursors combing with CNNs at the classification stage. These regions are often gained manually or extracted by supervised-trained part detectors. In [8], they are generated from MCG (Multi-scale Combination Grouping [9]). On one hand, the main drawback of these methods is splitting the end-to-end learning pipeline into separate steps, which may lose information among the candidate proposals. As a consequence, these defined parts may not be optimal for scene recognition. On the other hand, annotating parts is significantly more challenging than collecting image labels. Although a series of state-of-the-art results on popular benchmark datasets (MIT Indoor 67 [5], SUN397 [10], Places205 [2]) have been achieved, CNN features are used rudimentarily. It is difficult to figure out what is important for scene understanding and how to produce better scene representation.

Visual Attention Modeling: Previous works have made some progress on visual attention modeling. Some researchers have concentrated on specific tasks in toy or constrained environments, such as detecting simple shapes [11] [12]. While some others have been interested in less constrained environments, specifically the fine-grained categorization task [13]. Most recently, Ba et al. [14] have utilized visual attention to recognize multiple objects in images. Xu et al. [15] applied two different attention models based on LSTM to the image caption task. Besides, Zhou et al. [16] show that the CNNs are capable to learn a form of implicit attention somewhere they respond more strongly to some parts of an image than others. Jaderberg et al. [4] proposed the spatial transformer networks, which can be regarded as spatial attention model.

In order to deal with weaknesses of recent scene recognition approaches, our models are trained end-to-end from input-output pairs only with image-level labels, and the attention regions are located top-to-bottom driven by the final task explicitly. Additionally, feature fusion is introduced to enhance scene representations so that our model can be expected to yield better performance.

III. Our Model

The convolutional neural network combining with spatial transformer (ST-PlacesCNN) is detailed in this section. The architecture of the network is shown in Fig. 1. It can be split into two parts, namely spatial transformer module regarded as visual attention model to focus on the discriminative regions in images and sequential classification network containing several subCNNs, which extract features for feature fusion and final scene classification. The ST-PlacesCNN can be trained in an end-to-end way without any manual labels of attention regions in images.

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.

A. Spatial Transformer Network

Spatial transformer network is initially introduced to actively spatially transform feature maps within CNN in [4]. It consists of three parts, namely a localization network, a grid generator and a sampler. Our method extends the work of Jaderberg et. al. [4] for less constrained visual recognition, specifically scene classification in images.

In our visual attention-based model, the spatial transformer network is utilized as attention model, which is able to learn the scales and locations of some discriminative attention windows for informative regions and crop them out from original images automatically. Therefore, a constrained 2D affine transformation is used.

$$A_\theta = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \tag{1}$$

where s_x, s_y denote the scales of attention windows, and t_x, t_y represent the movements of the centric position of attention windows on x-axis and y-axis, respectively.

As for visual attention, the point-wise transformation is :

$$\begin{pmatrix} {x_i}^s \\ {y_i}^s \end{pmatrix} = A_\theta \begin{pmatrix} {x_i}^t \\ {y_i}^t \\ 1 \end{pmatrix} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \begin{pmatrix} {x_i}^t \\ {y_i}^t \\ 1 \end{pmatrix} \tag{2}$$

where $\begin{pmatrix} {x_i}^s \\ {y_i}^s \end{pmatrix}$ are the sampling points on input feature maps in the source coordinate system, and $\begin{pmatrix} {x_i}^t \\ {y_i}^t \end{pmatrix}$ are the corresponding transformed points in the target coordinate system. We use height and width normalized coordinates, such that $-1 \leq {x_i}^s, {y_i}^s, {x_i}^t, {y_i}^t \leq 1$.

Scenes can be distinguished according to whether they contain particular objects or regions. However, the sizes and positions of them may vary significantly in different instances. Besides, some other objects or regions in images are not correlated with the scene label, which could be neglected. Consequently, multiple spatial transformers can be used in a parallel manner to detect the objects or regions containing information about scenes, which can reduce noises and enhance the subsequent scene classification task.

B. Convolutional Neural Networks

The deep convolutional neural networks are used to produce transformation parameters of visual attention windows, and extract features of both original images and local detected regions in our model.

1) Localisation Network: As shown in the left part of Fig. 1, there are several parallel affine transformers, which are utilized as attention modules in the architecture. Each transformer learns the scales s_x, s_y and the movements of centric positions (t_x, t_y) of attention windows on x-axis and y-axis, respectively. In order to make the whole learning process more meaningful and more effective, we impose some constrains on these parameters as follows:

$$\begin{cases} s_x = S_x \times sigmoid(s_{lx}) \\ s_y = S_y \times sigmoid(s_{ly}) \\ t_x = T_x \times tanh(t_{lx}) \\ t_y = T_y \times tanh(t_{ly}) \end{cases} \tag{3}$$

in which,

$$\begin{cases} S_x + T_x = 1 \\ S_y + T_y = 1 \end{cases} \tag{4}$$

where $t_{lx}, t_{ly}, s_{lx}, s_{ly}$ are learned from the localization network. As a result, the scales of attention windows are restrained within S_x, S_y of the image size, and the centric positions of attention windows are limited: $-T_x \leq t_x \leq T_x$, $-T_y \leq t_y \leq T_y$.

2) Feature Extractor: Given the original input images and corresponding sets of attention regions, we employ the fine-tuned PlacesCNNs with the output layer removed as feature extractors, in which we use batch normalization [17] in convolutional layers and fully-connected layers. The outputs of feature extractors are with same dimension. Sequentially, fusion layers to be introduced in IV-B and a new output layer are appended. Note that one or more extra layers integrating local features, and/or integrating local features with global features during the fusion step could be added.

C. Feature Fusion

Features of global original images and local attention regions are obtained through subCNN streams. In this section, the fusion of these features is introduced to generate more robust and enhanced representations of scenes.

For each input image, its final representation R can be obtained:

$$\begin{cases} R = W_g F_g + \sum_{i=1}^{n} W_i F_{li} \\ W_g + \sum_{i=1}^{n} W_i = 1 \end{cases} \tag{5}$$

where F_g represents global features, while F_{li} is local features of $Region_i$. The equation (5) above is formulated to calculate the weighting sum of global features and local features of all attention regions. The output of the fusion layer is the input for the last softmax classification layer.

IV. Experiments

In this section, experimental details and results analysis are presented. We use multiple spatial transformers in parallel to perform scene recognition and evaluate the model on a subset of the Places205 dataset (Places20) for illustration. In our experiments, we only use image-level labels to train models.

A. Evaluation Dataset

The Places205 dataset is released in [2] to support deep learning methods applied to the scene recognition task, covering 205 scene categories. The train set contains 2,448,873 images, with the minimum 5,000 and the maximum 15,000

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.

0: classroom, 1: bedroom, 2: shoe_shop, 3: bookstore, 4: dining_room, 5: kitchen, 6: living_room, 7: music_studio, 8: art_studio, 9: office, 10: desert/sand, 11: desert/vegetation, 12:baseball_field, 13:stadium/football,14: ocean, 15:rainforest, 16: underwater/coral_reef, 17: lighthouse, 18:gas station, 19: sky

Fig. 2. Instances from each scene category of Places20 dataset.

images per category. The validation set contains 100 images per category and the test set contains 200 images per category.

Places20: To facilitate our research on learning visual attention regions, we select 20 classes from the Places205 dataset, dubbed Places20 here. It has 276,640 training images, 2,000 validating images and 4,000 testing images,and consists of 10 indoor scenes and 10 outdoor scenes shown in Fig. 2. Note that we do not struggle to deal with the huge datasets. Our aim is to declare the capability of our model discovering discriminative objects or regions automatically when trained just on image labels, so that they can boost the performance of scene recognition through reducing noises irrelevant to scenes and making the utmost of informative regions related to scenes for classification. We use a small dataset Places20 as a start to verify our findings.

B. Methods

We evaluate the following methods for comparison.

1. PlacesCNN fine-tuned on Places20: PlacesCNN is pre-trained on the Places205 dataset in [2]. The architecture of PlacesCNN is the same as the one used in the Caffe reference network (AlexNet). Based on this pre-trained network, we apply batch normalization [17] on 'conv1', 'conv3', 'con5', 'fc6' and 'fc7' layers. Afterward, we replace the 205-classes output layer with a 20-classes output layer. We fine-tune the alternative architecture called Places20CNN on Places20 dataset. The result of fine-tuned PlacesCNN is consider as the baseline for comparison with our proposed methods.

2. ST-placesCNN with different scales of spatial transformer modules: We compare different max scales (0.3, 0.5) of attention windows to verify the conjecture that our network can learn some salient regions of original images containing discriminative information for scene recognition.

3. ST-placesCNN with different numbers of spatial transformer modules: In our model, the number of transformers can be varied. We do experiments on networks with 1 or 2 parallel spatial transformers, which are parameterized for visual attention and act on the input image respectively.

4. ST-placesCNN combining global features with local features: The set of attention regions may lose some information because of the restriction to the attention process. We propose to combine global features with local features to generate stronger descriptors of scenes and to see if better performance can be achieved.

C. Experimental Setup

Initially, we fine-tune placesCNN on the Places20 dataset with batch normalization to obtain Places20CNN as the baseline. For data preprocessing, all images are resized to 454×454 resolution, and then are downsampled to 227×227 that are subtracted by the pixel mean as the inputs of the localization network.

All spatial transformers share the same localization network that is derived from Places20CNN in the following way. In order to preserve spatial information, the last classification layer, pooling layer and 2 fully-connected layers are removed. The output of the truncated CNN has 13×13 spatial resolution with 256 feature channels. Sequentially, an $128D$ fully-connected layer is added, and N fully-connected layers with $4D$ output are used to produce transformer parameters, where N is the number of transformers (in our experiments, $N = 1$ or 2). Through the transformer layers, there are several attention regions cropped from original images, then features of them and input images are extracted via individual subCNN streams. Afterwards, all the features are fused together. Finally they are classified with a single softmax layer.

All models were trained with the adadelta algorithm. We initialize the localization network and the feature extracting networks with the parameters of the corresponding layers in Places20CNN. In order to reduce overfitting, we set dropout with 0.5 after each fully connected layer and set the weight decay as 10^{-5}. We evaluate two ST-PlacesCNNs with different numbers of spatial transformers (1 or 2), two 1ST-PlacesCNNs with different max scales of spatial transformers (0.3 or 0.5), and a 1ST-g-PlacesCNN combining 1 spatial transformer with a global feature extractor on Places20 dataset.

D. Results and Analysis

The top-1 accuracy on the Places20 dataset of all methods is shown in TABLE I. The released model PlacesCNN only achieves 53.95% on Places20. Our proposed 1ST-g-Places20CNN(1ST, scale: 0.5, global) achieves

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.

82.10% slightly higher than 80.98% given by the baseline Places20CNN(bn).

TABLE I
TOP-1 ACCURACY

(Notes: ST–spatial transformer, bn–batch normalization, scale–the max scale of attention window, global–global features of original image.)

Method	Top-1 Accuracy
PlacesCNN	53.95%
Places20CNN(bn)	80.98%
2ST-Places20CNN(2ST, scale: 0.5)	74.03%
1ST-Places20CNN(1ST, scale: 0.5)	69.20%
1ST-Places20CNN(1ST, scale: 0.3)	58.15%
1ST-g-Places20CNN(1ST, scale: 0.5, global)	**82.10%**

We first discuss the influence of the attention window scale. The max scales (in Equation (3) S_x, S_y) are both set as 0.3 or 0.5 for 1 spatial transformer for comparison. The performance improves significantly with the increase of scales, from 58.15% of a small scale (0.3) to 69.20% of a large scale (0.5). In addition, it is worth noting that no matter how large the attention window is, our model learns to focus on the most informative regions in images. As shown in Fig. 3, two bounding boxes almost cover the same region in positive instances A-(a), A-(b), B-(a) and B-(b). In instance A-(c), the model with an attention window scaled 0.5 achieves the correct label, which detects the region containing the most important object – bed in the bedroom scene category, while the attention window scaled 0.3 outperform that scaled 0.5 in B-(c) because the former focuses on majority of lighthouse in the image. During our experiments, we find that it is difficult to automatically learn the scales of attention windows and the model tends to cover more area in images restricted to constrains.

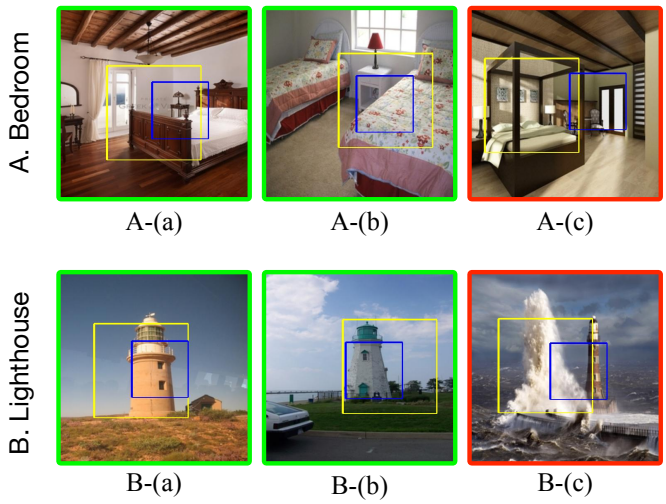

Fig. 3. Classification results of three instances from the bedroom and the lighthouse scene category, respectively. The images with a green boundary are positive instances, while those with a red boundary are negative. Besides, the yellow, blue bounding boxes in images are denoted as attention windows with different scales, namely 0.5, 0.3, respectively (Best viewed in color).

Secondly, the performances of different numbers of spatial transformers in ST-PlacesCNNs is investigated. Top-1 accuracy for one, two spatial transformers with attention window scaled 0.5 are 58.15%, 74.03%, respectively, demonstrating that two spatial transformers produce remarkably better result than one spatial transformer only. Scene recognition benefits more from several attention windows. Some meaningful salient regions in images located by 2ST-Places20CNN(2ST, scale: 0.5) are shown in Fig. 4. The spatial transformers wok well when they focus on discriminative regions of scenes. For instance, the model fixes its gaze on bodies of lighthouse; while for classroom images, it pays more attention to desks and chairs. It is interesting to observe that there are some overlaps between two regions, which may embody more essential information of scene categories.

Finally, we find that fusing local and global CNN features can further improve performance. The performance of 1ST-g-PlacesCNN with global information is slightly better than that of pure 1ST-PlacesCNN. The max scale of spatial transformer is set to 0.5. Visualization of scene recognition performance of 1ST-g-PlacesCNN on each category is illustrated in Fig. 5. We can make two observations: 1) It is easy to distinguish indoor scenes from outdoor scenes. 2) Scene recognition would make mistakes when two scenes are similar, such as, desert/sand and desert/vegetation, which are sub-classes in the desert scene. Our method is capable to fix its gaze on the discriminative regions to assist fine-grained scene recognition.

Discovering discriminative attention regions containing enough information from the original images is the key of our proposed method to recognize scenes. It is reasonable to believe that the classification performance increases as the scale or the number of attention windows grows because more information in images is utilized. Moreover, there are some overlaps that can be observed covering important information about scenes. With the help of visual attention, our networks are potential to actively harvest the most meaningful and informative regions in images and generate stronger descriptor for scene categories in a data-driven manner without any additional supervision.

V. CONCLUSION

In this work, we have proposed to explore explicit attention model – spatial transformer for scene recognition task. We have combined an CNN architecture with spatial transformers to detect meaningful and informative regions from original images to generate better descriptor of scenes through feature fusion. The experimental results show that our method outperforms the basic PlacesCNN model. Future work will focus on improving the explicit attention modeling and applying our models to other scene datasets.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, (NIPS 2012)*, pp. 1106–1114.

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.

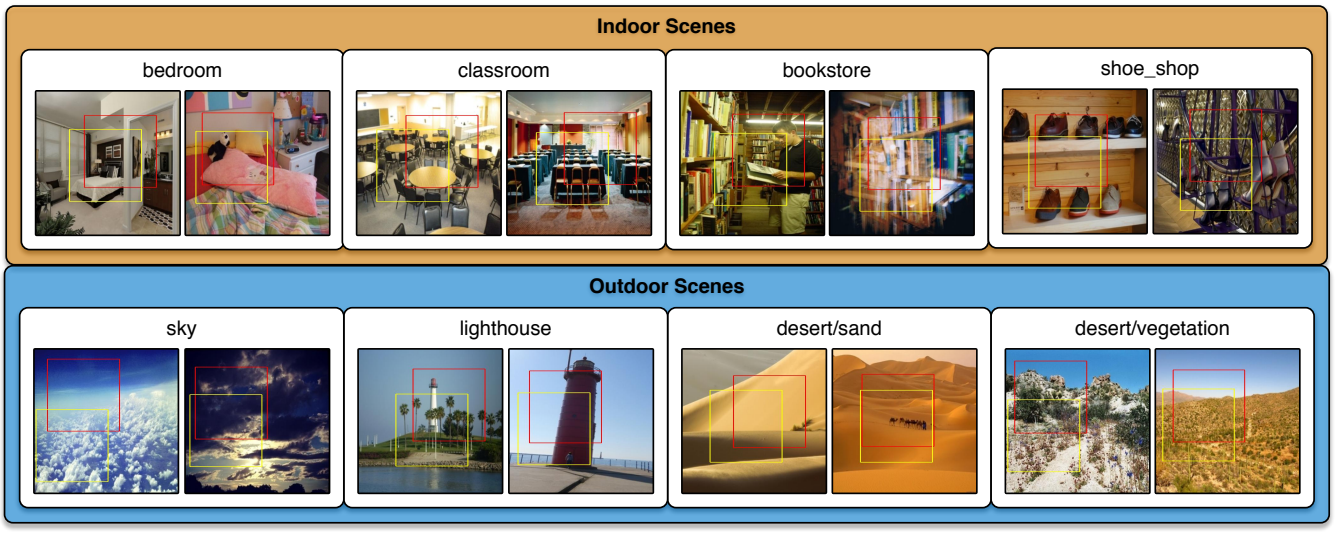

Fig. 4. Discriminative regions of some instances detected by 2ST-Places20CNN on the Places20 dataset.

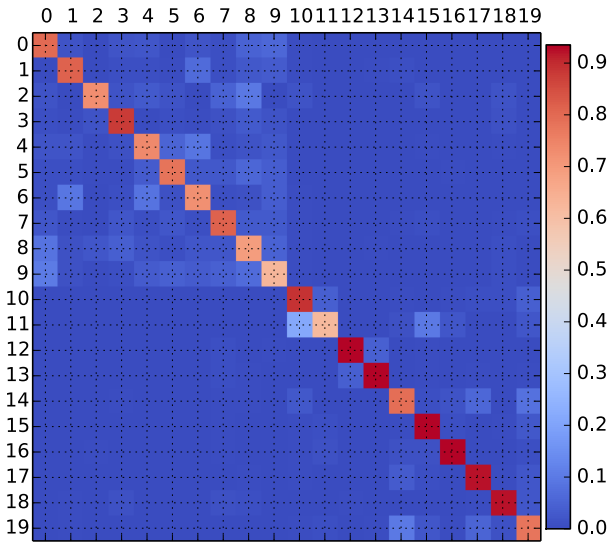

Fig. 5. Classification results of 1ST-g-PlacesCNN on the Places20 dataset.

[2] B. Zhou, À. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, (NIPS 2014)*, pp. 487–495.

[3] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *CoRR*, vol. abs/1412.6856, 2014.

[4] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, (NIPS 2015)*, pp. 2017–2025.

[5] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 413–420.

[6] F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, pp. 524–531.

[7] L. Wang, S. Guo, W. Huang, and Y. Qiao, "Places205-vggnet models for scene recognition," *CoRR*, vol. abs/1508.01667, 2015.

[8] R. Wu, B. Wang, W. Wang, and Y. Yu, "Harvesting discriminative meta objects with deep CNN features for scene classification," in *2015 IEEE International Conference on Computer Vision, (ICCV 2015)*, pp. 1287–1295.

[9] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Computer Vision and Pattern Recognition*, 2014.

[10] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proceedings of the 23th IEEE Conference on Computer Vision and Pattern Recognition, (CVPR 2010)*, pp. 3485–3492.

[11] J. Schmidhuber and R. Huber, "Learning to generate artificial fovea trajectories for target detection," *Int. J. Neural Syst.*, vol. 2, no. 1-2, pp. 125–134, 1991.

[12] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas, "Learning where to attend with deep architectures for image tracking," *Neural Comput.*, vol. 24, no. 8, pp. 2151–2184, Aug. 2012.

[13] P. Sermanet, A. Frome, and E. Real, "Attention for fine-grained categorization," *CoRR*, vol. abs/1412.7054, 2014.

[14] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *CoRR*, vol. abs/1412.7755, 2014.

[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning, (ICML 2015)*, pp. 2048–2057.

[16] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *CoRR*, vol. abs/1512.04150, 2015.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning, (ICML 2015)*, pp. 448–456.

Preprint submitted to 23rd International Conference on Pattern Recognition.
Received April 20, 2016.