Scene Recognition with Sequential Object Context

Yuelian Wang and Wei Pan(✉)

College of Information Engineering, Capital Normal University, Beijing 100048, China
wangyuelian2355@gmail.com, bjpanwei@163.com

Abstract. Convolutional Neural Networks (CNNs) have been widely used for many computer vision tasks and produce discriminative and rich representations for images or regions of an image. Recognizing scenes requires both local object features and global semantic information as a scene image is usually composed of multiple objects which are organized with specific spatial distribution. To address these problems, in this paper, we propose a deep network architecture which models the sequential object context of scenes to capture object level information. We first detect a set of obejcts in a scene image, and then apply a pre-trained CNN to extract discriminative features for these objects. Then we use a Long Short-Term Memory (LSTM) network to get the context features by progressively receiving all contextual objects. The learned sequential object context incorporates object-object relationship and object-scene relationship in an end-to-end trainable manner. We evaluate our model on two benchmark datasets and achieve promising results compared to state-of-the-art methods.

Keywords: Convolutional neural network · Long short-term memory
Scene recognition · Sequential object context

1 Introduction

Scene recognition is a critical task in the computer vision community. It has a wide range of applications such as assistive human companions, robotic agent path planning, monitoring systems and so on. State-of-the-art approaches in scene recognition are based on the successful combination of deep representations and large-scale datasets. Specifically, deep convolutional neural networks (CNNs) trained on ImageNet [24] and Places [37] have shown significant improvement in performance over methods using hand-engineered features and have been used to set baseline performance for visual recognition.

As a scene images is usually composed of multiple objects which are organized with specific spatial distribution, classifying it requires not only the holistic features of the whole image, but also the local features of objects in the image. However, CNNs learn image features in a layer-wise manner where low layers capture general features that resemble either Gabor filters or color blobs and

J. Yang et al. (Eds.): CCCV 2017, Part III, CCIS 773, pp. 108–119, 2017.
https://doi.org/10.1007/978-981-10-7305-2_10

high layers learn specific features which are semantic and representative even though they greatly depend on the chosen dataset and task [35]. The low layer general features are gradually transformed into high layer powerful features with multiple convolutional layers and pooling layers. This feature learning mechanism of CNNs suggests that they might not be the best suited architectures for classifying scene images where local object features follow a complex distribution in the spatial space. The reason is that the spatial aggregation implementation of pooling layers in a CNN is simple in some extent, and does not retain much information about local feature distributions. When crucial inference happens in the fully connected layers near the top of the CNN, aggregated features fed into these layers are in fact global features that neglect local feature distributions. The global CNN features are not efficient enough to capture contextual knowledge like the complex interaction of objects in a scene.

In addition to the entire image, it has been demonstrated that an image representation based on objects can be very useful in visual recognition tasks for scenes. Li *et al.* [14] propose a high-level image representation where an image is represented as a scale-invariant response map of a large number of pre-trained generic object detectors. The object-based representation carries rich semantic level image information and achieves superior performance on many high level visual recognition tasks. Li *et al.* [15] propose a hierarchical probabilistic graphical model to perform scene classification with the contextual information in form of object co-occurrence is explicitly represented by a probabilistic chain structure. Liao *et al.* [19] propose a architecture which encourages deep neural networks to incorporate object-level information with a regularization of semantic segmentation for scene recognition. Wu *et al.* [31] use a region proposal technique to generate a set of high-quality patches potentially containing objects and then a scene image representation is obtained by pooling the feature response maps of all the learned meta objects at multiple spatial scales to retain more information about their local spatial distribution.

In this paper, we propose a architecture to learn sequential object context which encodes rich object-level context using a LSTM network on top of a set of discriminative objects, as shown in Fig. 1. This architecture attempts to learn powerful semantic representations in scenes by modeling object-object and scene-object relationships within a single system. The intuition is that we human beings first scan objects in an image and then reason the relationships between these objects to decide what scene category the image belongs to. The joint existence of a set of objects in a scene highly influences the final scene category. Additionally, the LSTM units are capable of modeling the relationship between the objects by progressively taking in object features at each time step.

In our framework, we first use a region proposal network to detect a set of objects for each scene image. These objects are sorted by their locations in the image to form the object-based context sequence. And then, we use scene-centric Places CNN to extract features for the whole image to capture global scene information. At the same time, we use object-centric ImageNet CNN to extract features for the detected objects. After this, the representations of the

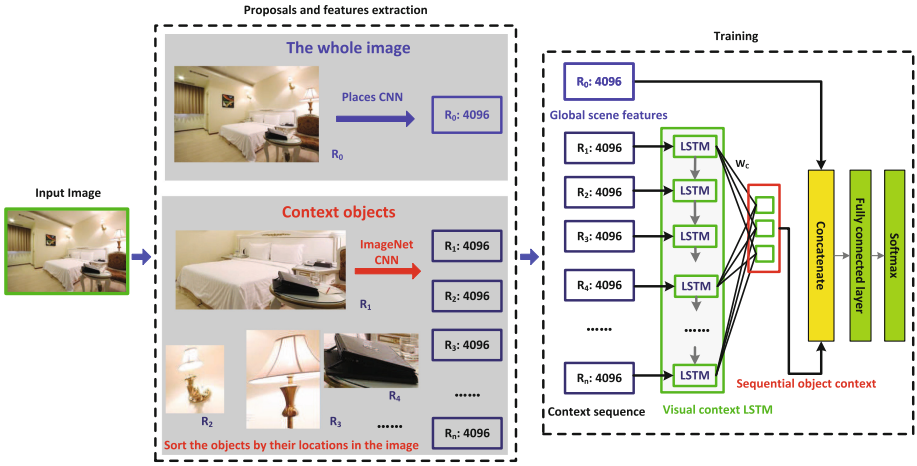

Fig. 1. The framework of our method. Firstly, a set of objects are detected in a scene image. The objects are arranged by their locations in the image. And then we use Places CNN and ImageNet CNN to extract features for the whole image and these objects respectively. The features of these objects are put into the visual context LSTM to form the sequential object context. At last, the global scene features of the whole image and the learned context are concatenated and the combination of them are put into a sub network to classify the image.

object context sequence are put into a LSTM network to form the discriminative and representative sequential object context. At last, the global scene features of the whole image and the learned context are concatenated and the combination of them are put into a sub network to classify the image. In this way, the network can learn about the scene class probability distribution given it has seen a specific set of objects through time. In summary, the main contributions of our paper are as follows:

- We firstly use an LSTM network to explicitly learn sequential object context for scene recognition. The learned discriminative and representative context contains information from all of the objects in the image.
- We empirically show that the sequential object context is complementary to the global scene information extracted form the whole image. Leveraging both global scene features and local sequential object context, our method achieves promising results compared to state-of-the-art methods on many challenging benchmarks.

The rest of the paper is organized as follows. We give a brief overview of related work in Sect. 2. Section 3 describes the proposed method. Section 4 describes the experiments and Sect. 5 concludes the paper.

2 Related Work

Scene recognition. Earlier work in scene recognition focuses on carefully hand crafted representations such as image contours, high contrast points, histogram of oriented gradients and so on [28]. Recently, with the great success of deep convolutional networks, features extracted from CNNs have been the primary candidate in most visual recognition tasks [4,24,37]. As CNNs trained on ImageNet [24] achieve impressive performance in object recognition, CNNs trained on Places [37] get significant performance in scene recognition. In order to get better performance, there are two primary ways to take full advantage of CNN features. The first way is to extract abundant features from local patches and aggregate them into effective scene representations [6,8,31,34]. Usually, these approaches combine multiple local patches and multiple scales features, and these features are pooled using VLAD [8] or Fisher vector [6] encoding. The second way is to leverage features which are extracted from complementary CNNs [10,16]. Thus, these different features can have complementary characteristics. Li *et al.* [16] have demonstrated that the combination of features from deep neural networks with various architectures can significant improve the performance as features obtained from heterogeneous CNNs have different characteristics since each network has a different architecture with different depth and the design of receptive fields. Herranz *et al.* [10] have improved that the concatenation of features extracted from object-oriented and scene-oriented networks results in significant recognition gains. In this paper, we assume that knowledge about objects in a scene image is helpful in scene recognition since objects are main components of scenes. We propose a framework to explore object context for scene recognition.

Context modeling. The utilization of context information for computer vision has attracted a lot of attention. Choi *et al.* [5] propose a graphical model to exploit co-occurrence, position, scale and global context which together is used to identify out-of-context objects in a scene. Torrala *et al.* [27] have shown how to exploit visual context to perform robust place recognition, categorization of novel places, and object priming. Izadinia *et al.* [12] have proposed a method to learn scene structures that can encode three main interlacing components of a scene: the scene category, the context-specific appearance of objects, and their layouts. Recently, RNN-based architectures have been widely used to model context for a lot of visual tasks, such as object detection [2], segmentation [11,18], scene labeling [3,25], human re-identification [29] and so on. The primary mechanism behind RNN is that the connections with previous states enables the network to memorize information from past inputs and thereby capture the contextual dependency of the sequential data. In the same spirit, we use a LSTM network to model context to boost the classification performance for scene images.

CNN-LSTM Models. CNN have been widely used to learn discriminative features for a wide range of visual tasks [18,23,24,30] and Recurrent Neural Networks (RNNs) have been widely used for sequence learning. Recently, a lot of deep architectures use the joint learning of CNN and RNN to get feature

representations as well as their dependencies. Tasks combining visual and language like image captioning [17,30,33] and visual question answering [1,9] use CNN to get image features while use LSTM to generate natural language expressions as image descriptions or answers. As the architectures which are only composed of CNN can get features from large receptive fields but can not allow for finer pixel-level label assignment. Architectures with LSTM components can learn dependencies between pixels and improve agreement among their labels. A lot of work have used CNN-LSTM architectures for scene labeling [3] and semantic segmentation [18]. In this work, we use a CNN-LSTM based network to learn informative and discriminative for scene recognition.

3 Our Method

Overview: As a scene image is usually composed of multiple objects, the context of a scene image encapsulates rich information about how scenes and objects are related to each other. Such contextual information has the potential to enable a coherent understanding of scene images. In order to leverage such informative information, we propose sequential object context incorporating global scene information for scene recognition, as shown in Fig. 1. We first detect a set of objects for each scene image. To model the distribution of these objects, they are sorted by their locations in the image. And then, we use deep neural networks to extract features for the whole image and the detected objects. The representations of the object context sequence are put into a LSTM network to form the discriminative and representative sequential object context. At last, the global scene features of the whole image and the learned context are concatenated and the combination of them are put into a sub network to classify the image. The goal of our method is to complement the deep CNN features extracted from the whole image with local context from objects within a scene. The following sections provide the details of our method and its training procedure.

3.1 Object Proposal Extraction

As we aim to incorporate better local visual context for scene recognition, we first need to detect objects in scene images. We train a Faster R-CNN [23] detector and build our system on top of the detections, as shown in Fig. 2. Because a Region Proposal Network (RPN) in the Faster R-CNN takes an image as input and outputs a set of rectangular object proposals, each with an objectness score, we can select discriminative visual objects depending on the output of the RPN. Specifically, to get local proposals, we select top-n detected objects to represent important local objects according to their class confidence scores obtained from Faster R-CNN. We train our Faster R-CNN model using the VGG-16 convolutional architecture [26]. The model is first pre-trained on ImageNet [24] dataset and then fine-tuned on the training set of MS COCO [20] dataset, as MS COCO contains a lot of images which are composed of multiple objects.

After this, these objects are sorted by their locations in the image with the order of from left to right and top to down. Specially, the entire image is also

considered as a special object and is denoted as R_0. We use $seq(I)$ to denote the initial sequential representations of the local objects, which contains a sequence of representations $seq(I) = \{R_1, R_2, \ldots R_N\}$, where R_1 to R_N are the local objects. R_0 is the corresponding global representation of the entire image.

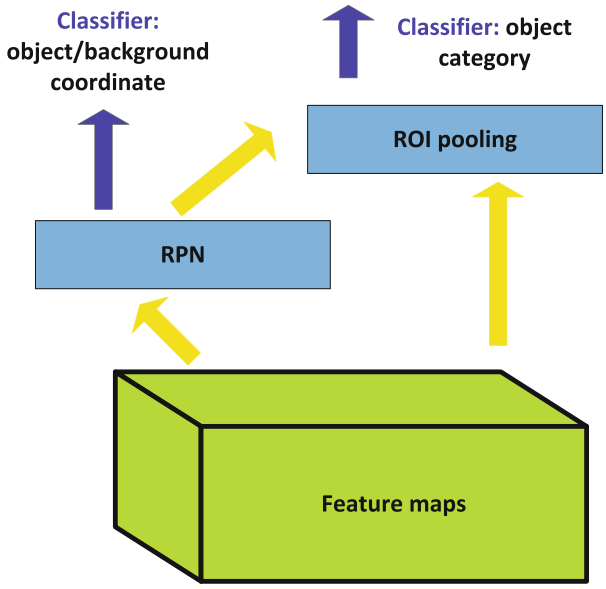

Fig. 2. The object detector in our model. We select top-n detected objects according to their class confidence scores.

3.2 Features Extraction

Convolutional Neural Networks (CNNs) have been widely used for many visual tasks due to its powerful representation ability. In our work, we also use CNNs to extract features for the whole image R_0 and the sequential representations of the local objects $seq(I) = \{R_1, R_2, \ldots R_N\}$. It has been demonstrated that scene-centric knowledge (Places) and object-centric knowledge (ImageNet) are complementary, and the combination of these two can significantly improve the performance [10]. So we use a Palces CNN to extract the 'fc7' layer features for R_0 and use an ImageNet CNN to extract the 'fc7' layer features for sequential local objects $seq(I)$. The features of the entire image is denoted as V_0, where $V_0 = CNN_P(R_0)$. The features of the $i-th$ object is denoted as V_i, where $V_i = CNN_I(R_i)$. The context sequence can be denoted as $seq_V(I) = \{V_1, V_2, \ldots V_N\}$.

3.3 Sequential Object Context Modeling

The core idea of our work is motivated by the previous works [7,21,29] which have demonstrated that the LSTM architectures can model abundant context

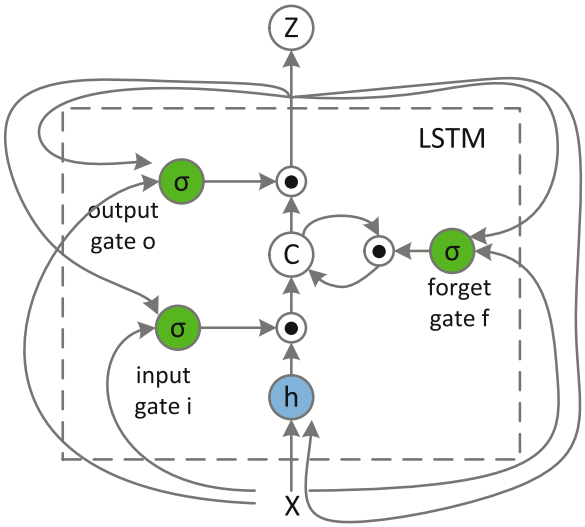

Fig. 3. Diagram of the LSTM network of our model. Our LSTM network continuously receives the detected objects thus progressively capture and aggregate the relevant contextual information.

features for both visual and language tasks. The internal gating mechanisms in the LSTM cells can regulate the propagation of certain relevant contexts, which enhance the discriminative capability of local features. We first introduce the LSTM network which is used in our method and then present the input and output of the LSTM module.

The architectural of the visual context LSTM is illustrated in Fig. 3. It receives the output of the previous time step, as well as the input at the current time step, as the inputs of the current unit. Mathematically, the update equations at time l can be formulated as:

$$i_l = \sigma(\mathbf{W}_{ix}x_l + \mathbf{W}_{im}m_{l-1}) \tag{1}$$

$$f_l = \sigma(\mathbf{W}_{fx}x_l + \mathbf{W}_{fm}m_{l-1}) \tag{2}$$

$$o_l = \sigma(\mathbf{W}_{ox}x_l + \mathbf{W}_{om}m_{l-1}) \tag{3}$$

$$c_l = f_l \odot c_{l-1} + i_l \odot \phi(\mathbf{W}_{cx}x_l + \mathbf{W}_{cm}m_{l-1}) \tag{4}$$

$$m_l = o_l \odot \phi(c_l) \tag{5}$$

where i_l, f_l and o_l represent the input gate, forget gate, and output gate at time step l respectively; c_l is the state of the memory cell and m_l is the hidden state; $\odot$ represents the element-wise multiplication, $\sigma(\cdot)$ represents the sigmoid function and $\phi(\cdot)$ represents the hyperbolic tangent function; $W_{[\cdot][\cdot]}$ denote the parameters of the model.

The LSTM takes in $seq_V(I) = \{V_1, V_2, \ldots V_N\}$ and encodes each object into a fixed length vector. Thus, we have encoding hidden states computed from:

$$h_t = LSTM(seq_V(I)_t, h_{t-1}),\ t = 1, 2, \ldots N \tag{6}$$

Once the hidden representations from all the context objects are obtained, they are combined to obtain the sequential object context $conV$ as shown below:

$$conV = \mathbf{W}_\mathbf{C}^T[(h_1)^T, (h_2)^T, \ldots (h_r)^T, \ldots, (h_N)^T],\ r = 1, 2, \ldots, N \tag{7}$$

where $\mathbf{W_C}$ is the transformation matrix we need to learn and $[\cdot]^T$ indicates the transpose operation.

The final features $V(I)$ used to recognize a scene image are obtained from the combination of the global scene features V_0 and the sequential object context features $conV$. $V(I)$ are put into a sub network which is mainly composed of a fully-connected layer and a softmax layer to classify the image.

$$V(I) = [V_0,\ conV] \tag{8}$$

3.4 Training Details

We train our model on the framework of Caffe [13]. The visual context LSTM and the sub classification network are optimized in a end-to-end manner. For each scene image, we detect $n = 10$ objects to form the consequential object context. We use the mini-batch stochastic gradient descent method with the batch size of 20. The hidden state size of the visual context LSTM is set to 512, and the size of the fully-connected layer of the classification network is 4096.

4 Experiments

4.1 Dataset

To verify the effectiveness of our method, we evaluate the performance of our method on two benchmark datasets: MIT 67 [22] and SUN 397 [32].

MIT 67: MIT Indoor 67 [22] contains 67 categories of indoor images, with 80 images per category available for training as well as $20 * 67$ images for test. Indoor scenes tend to be rich in objects compared to object-centric images, which in general makes the task more challenging.

SUN 397: SUN 397 [32] is a scene benchmark containing 397 categories, including indoor, man-made and natural categories. This dataset is very challenging, not only because of the large number of categories, but also because the more limited amount of training data with 50 images per category for training and 50 images per category for test.

4.2 Quantitative Results

We conduct experiments on two benchmark datasets to qualitatively verify the effectiveness of our method. Our experiments mainly aim to demonstrate the usefulness of the consequential context not to get the best performance, so we just use the Alexnet networks to extract features. Table 1 shows the performance comparison of the proposed algorithm with the baseline algorithm. Alexnet 205 denotes that we use the Alexnet which is trained on Places 205 to extract fc7 features for the entire image and then feed the features to the sub classification network. Alexnet 205 & SOC denotes the combination of the features of the entire image and the learned sequential object context (SOC). It can be seen that the combination of global and local features of the proposed architecture outperforms the global scene information of the baseline algorithm for all the datasets. We get the same results when we use the network trained on Places 365 to extract the features. The comparison of our method with state-of-the-art methods is also show in Table 1. It shows that our method achieves promising results compared to state-of-the-art methods.

Table 1. The recognition performance on MIT 67 and SUN 397 datasets. * indicates that the performance are got with our own implementation. SOC is short for sequential object context.

Method	MIT 67	SUN 397
Alexnet 205*	68.25	54.36
Alexnet 205 & SOC*	69.36	55.78
Alexnet 365*	70.22	56.02
Alexnet 365 & SOC*	**71.86**	**57.72**
ImageNet Alexnet [37]	56.79	42.61
Places 205 Alexnet [37]	68.24	54.32
Places 365 Alexnet [36]	70.72	56.12
MS Orderless Pooling [8]	68.88	51.98

5 Conclusion

In this paper, we propose a deep model to model sequential object context for scene recognition. As scene images are rich in objects and the global scene information extracted from the entire image with CNNs neglect local object distributions, the learned sequential object context features are strong complementary representations which contain full information obtained from local objects. Experimental results show that the combination of the global scene information and the learned local sequential object context significantly improves the recognition performance. By using the LSTM module, our network can selectively

propagate relevant contextual information and thus enhance the discriminative capacity of the local features.

In future work, we will use deeper networks as our feature extractors to get more powerful features. We will also incorporate multi-scale CNN features to our network to get better performance.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant No. 61202027. This work was also funded by the Project of Construction of Innovative Teams and Teacher Career Development for Universities and Colleges Under Beijing Municipality under Grant No. IDHT20150507.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: VQA: visual question answering. Int. J. Comput. Vis. **123**(1), 4–31 (2017)
2. Bell, S., Zitnick, C.L., Bala, K., Girshick, R.: Inside-outside net: detecting objects in context with skip pooling. In: CVPR (2016)
3. Byeon, W., Breuel, T.M., Raue, F., Liwicki, M.R.: Scene labeling with LSTM recurrent neural networks. In: CVPR (2015)
4. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of the British Machine Vision Conference, BMVC 2014 (2014)
5. Choi, M.J., Torralba, A., Willsky, A.S.: Context models and out-of-context objects. Pattern Recogn. Lett. **33**(7), 853–862 (2012)
6. Dixit, M., Chen, S., Gao, D., Rasiwasia, N., Vasconcelos, N.: Scene classification with semantic fisher vectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 2974–2983 (2015)
7. Fernández, S., Graves, A., Schmidhuber, J.: An application of recurrent neural networks to discriminative keyword spotting. In: de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D. (eds.) ICANN 2007. LNCS, vol. 4669, pp. 220–229. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74695-9_23
8. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 392–407. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_26
9. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: elevating the role of image understanding in Visual Question Answering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
10. Herranz, L., Jiang, S., Li, X.: Scene recognition with CNNs: objects, scales and dataset bias. In: CVPR (2016)
11. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 108–124. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_7
12. Izadinia, H., Sadeghi, F., Farhadi, A.: Incorporating scene context and object layout into appearance modeling. In: CVPR (2014)
13. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, MM 2014, pp. 675–678. ACM, New York (2014)

14. Li, L., Su, H., Xing, E., Fei-Fei, L.: Object bank: a high-level image representation for scene classification and semantic feature sparsification. In: Advances in Neural Information Processing Systems (2010)
15. Li, X., Guo, Y.: An object co-occurrence assisted hierarchical model for scene understanding. In: Proceedings of the British Machine Vision Conference (2012)
16. Li, X., Herranz, L., Jiang, S.: Heterogeneous convolutional neural networks for visual recognition. In: Chen, E., Gong, Y., Tie, Y. (eds.) PCM 2016. LNCS, vol. 9917, pp. 262–274. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48896-7_26
17. Li, X., Song, X., Herranz, L., Zhu, Y., Jiang, S.: Image captioning with both object and scene information. In: Proceedings of the 2016 ACM on Multimedia Conference, MM 2016, pp. 1107–1110. ACM, New York (2016)
18. Liang, X., Shen, X., Feng, J., Lin, L., Yan, S.: Semantic object parsing with graph LSTM. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 125–143. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_8
19. Liao, Y., Kodagoda, S., Wang, Y., Shi, L., Liu, Y.: Understand scene categories by objects: a semantic regularized scene classifier using convolutional neural networks. In: IEEE International Conference on Robotics and Automation (ICRA) (2016)
20. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
21. Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., Ward, R.: Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE/ACM Trans. Audio Speech Lang. Process. **24**, 694–707 (2016)
22. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, pp. 413–420 (2009)
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS (2015)
24. Russakvovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Kholsa, A., Bernstein, M., Berg, A., Fei-Fei, L.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
25. Shuai, B., Zuo, Z., Wang, G., Wang, B.: DAG-Recurrent neural networks for scene labeling. In: CVPR (2016)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
27. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: ICCV (2003)
28. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: a survey. Found. Trends. Comput. Graph. Vis. **3**(3), 177–280 (2008)
29. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 135–153. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_9
30. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR (2015)
31. Wu, R., Wang, B., Wang, W., Yus, Y.: Harvesting discriminative meta objects with deep CNN features for scene classification. In: ICCV (2015)

32. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitions, CVPR 2010, pp. 3485–3492 (2010)
33. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: ICML (2015)
34. Yoo, D., Park, S., Lee, J.Y., Kweon, I.S.: Multi-scale pyramid pooling for deep convolutional representation. In: Computer Vision and Pattern Recognition Workshops (CVPRW) (2015)
35. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NIPS (2014)
36. Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., Oliva, A.: Places: an image database for deep scene understanding. arXiv preprint arXiv:1610.02055 (2016)
37. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014, vol. 1, pp. 487–495 (2014)