Journal of Artificial Intelligence Practice (2016) 1: 25-29
Clausius Scientific Press, Canada

BLSTM Recurrent Neural Network for Object Recognition

Yalan Qin[1, a]

[1]College of Computer and Information Science, Southwest University, Chongqing, 400715, China
[a]email: CystalQin1477@gmail.com

Keywords: Multi-object Relationship; Object Recognition; BLSTM

Abstract: Multi-object relationship information can help eliminate some incorrect combinations or locations of objects. Moreover, it is favorable to extract scene information for object recognition. In this paper, we introduce a new way to generate image representation and propose a deep learning framework to fuse the contextual dependencies among objects and scene information in an image. It adopts a bidirectional long short-term memory recurrent neural network (BLSTM-RNN) to deal with the problem of variable-length sequence produced by local detectors in different images. Then it is applied to the existing tree context model for further recognition. Experimental results on SUN09 dataset show that our model outperforms the state-of the-art object localization methods.

1. Introduction

Recently, standard single detectors [1] have been focusing on identifying particular object categories locally. Then only the information inside each local candidate window is used in this way. Relationship information among candidate windows and information out of the candidate window are discarded. In this paper, image representation generation is introduced and a framework fusing scene information and contextual dependency exploited by BLSTM-RNN [3] among objects within an image (BLSTM-Context) for object presence prediction is proposed in Figure1.

Firstly we use information of outputs by local detectors such as confidence and location as input to the given framework. Then it is as the input to the BLSTM-RNN model. Moreover, the properties of the specific scenes, which can give us associatively contextual information about the existing object scenes is also used. Then it is as the input to the feed-forward neural network model.

Due to different images often obtain different number of outputs by local detectors. Furthermore, the contextual dependencies among objects within an image are needed to be exploited. Then the proposed BLSTM-Context model is used to obtain fixed-size semantic image representation. The obtained image representation is utilized as features to achieve higher recognition accuracies. Then we apply it to tree context model [6] (BLSTM-Tree) for further prediction.

2. Input of BLSTM-Context model

Firstly, we run M standard single detectors [1] in the given image to output a set of N candidate windows $\varphi = \{x_1, x_2, \dots, x_N\}$. $x_i = \{c_i, s_i, l_i^r, l_i^e, l_i^a, l_i^h\}$ donates i-th candidate window. c_i indicates the object label of the i-th candidate window, where $c_i \in \{1,2,\dots,M\}$. The symbol s_i denotes the confidence produced by the c_i-th object detector. l_i^r (l_i^e) is the horizontal (vertical) coordinate of the center of the i-th window. l_i^a is the width of the i-th window and l_i^h is the height of the i-th

window. In order to describe the scale changes of these candidate windows, the method in [2] is used by us to transform them into the 3-D-world coordinate

$$L_x = \frac{l_i^r}{l_i^h} H_0 \, , L_y = \frac{l_i^g}{l_i^h} H_0 \, , L_z = \frac{1}{l_i^h} H_0 \tag{1}$$

where H_0 denotes the physical height of object 0. And $W_i = \{c_i, s_i, L_i\}$ is defined to represent the i-th window as in [2]. After the definition of the window representation, we encode object label c_i of a window in one-hot mode. And data of 114 dimensions including s_i, L_y and $\log(L_z)$ is used to represent a detector window. Meanwhile, scene information is denoted with 512 dimensions.

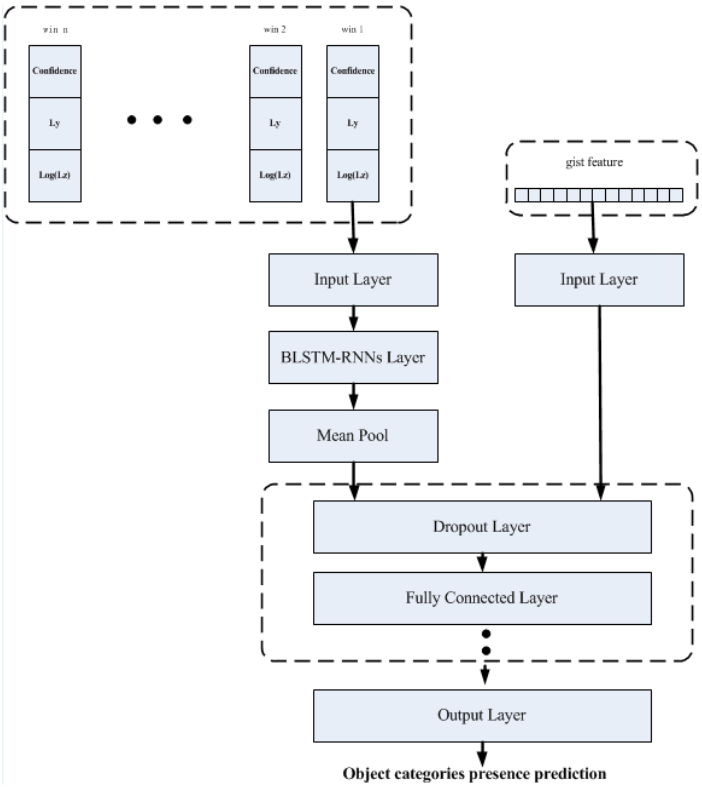


Fig.1. The general deep learning framework for object presence prediction

3. BLSTM-Context model for Image Representation

BLSTM-RNNs are powerful at incorporating long periods of contextual information from both directions without suffering from vanishing gradients. In our work, BLSTM-RNNs are used to learn semantic representations for images as in Figure2. Given a variable-length sequence of detector window representations and the sequence in opposite direction as input, BLSTM-RNNs produce a fixed-length image representation. We not only make use of the sequence summarization property of BLSTM-RNNs but also consider the global semantics of each image. Here, we average each forward hidden state h_t^f and backward hidden state h_t^b to produce the image representation.

To be more precise, the process of generating an image representation works as follows. For image i, we first generate all its window representations $W = \{x_1, x_2, ..., x_t, ..., x_n\}$. Then it is as input to the BLSTM-RNN model. In each time step t, BLSTM-RNN processes the input x_t and previous cell state C_{t-1} to generate the current forward hidden state h_t^f. Meanwhile, the backward hidden state h_t^b is also generated. After obtaining all of hidden states $\{h_1^f, h_1^b, h_2^f, h_2^b, ..., h_n^f, h_n^b\}$, we calculate their average and output it as $i's$ representation. Furthermore, the scene information of an image is also considered to represent the image.

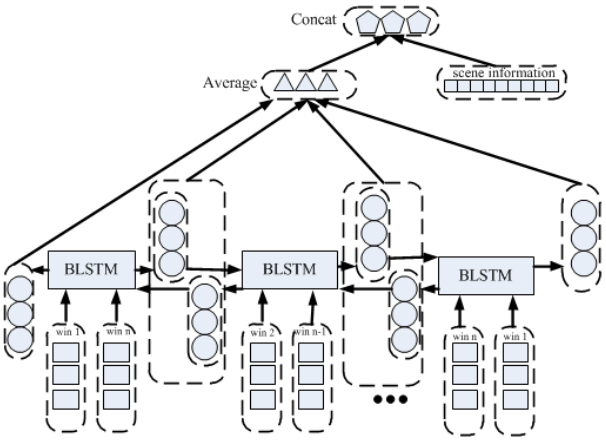


Fig.2. The process of generating image representations via BLSTM-RNNs

4. BLSTM-Context model for Object Category Prediction

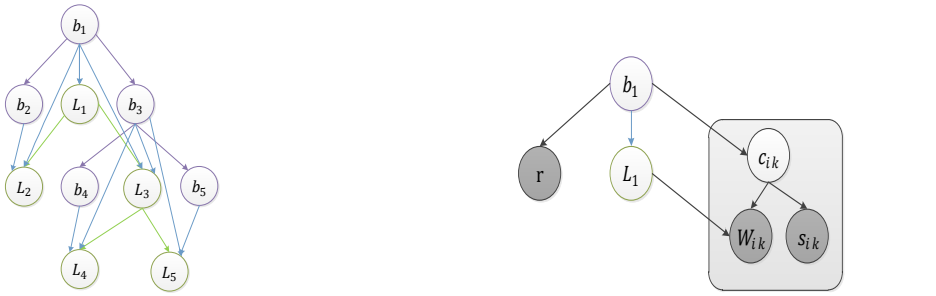
In this paper, we consider the semantic image representations as features. Then the obtained image representation is fed to the deep learning structure as in Figure1. The Dropout Layer is used for reducing over-fitting in neural networks. The output layer has a size equals to the number of object categories. Sigmoid activation function used in this layer is to give network responses between 0 and 1 at every time step. And we train the neural network as an object categories prediction problem. The mean squared error between target categories T_i and predicted categories P_i is used as the cost function. Given a training set D, the training objective is to estimate the network parameters by minimizing the loss function:

$$f_{loss} = \sum_{i \epsilon D} \sum_{j=1}^{C} (P_j(i) - T_j(i))^2 \quad (2)$$

where D indicates the training data, i represents an image, and C refers to the number of object categories. Note that $P(i)$ has C dimensions.

5. BLSTM-Tree model Generation and Prediction

In [6], the prior model consists of a tree-structure graphical model which is learned from co-occurrence statistics of object categories. Moreover, it provides efficient integration of different sources of contextual information such as location and scale relationship as in Figure3 (a). Meanwhile, the measurement model integrates produced semantic image representation and local detector output as measurements as in Figure3 (b). The symbol $\mathbf{r}$ indicates image representation, and local detector provides location W_{ik} and score s_{ik}. And the binary variable c_{ik} denotes whether the candidate window is correct.



(a) Prior model (b) Measurement model for object $\mathbf{i}$

Fig.3. Graphical model representations for parts of tree context model

In Figure3 (a), we denote the joint probability of all binary variables factored according to the tree structure $p(b) = p(b_{root}) \prod_i p(b_i | b_{pa(i)})$, where $pa(i)$ denotes the parent of object i. And symbol b without a subscript indicates a collection of all corresponding variables: $b = \{b_i\}$. Given image representation i, candidate window location $W = \{W_{ik}\}$ and their scores $s = \{s_{ik}\}$, we infer the presence of objects $b = \{b_i\}$, correct window detections $c = \{c_{ik}\}$ and expected locations of all objects $L = \{L_i\}$ by solving the optimization problem as follows:

$$\hat{b}, \hat{c}, \hat{L} = arg_{b,c,L} \max p(b, c, L | r, W, s) \quad (3)$$

And the inference process in [6] is used to obtain the final object presence prediction and localization prediction results

6. Test results

In this section, we will evaluate our method for object recognition on the SUN09 dataset which contains 8684 images with 111 object categories. Then we apply our proposed BLSTM-Tree model to object localization and presence prediction.

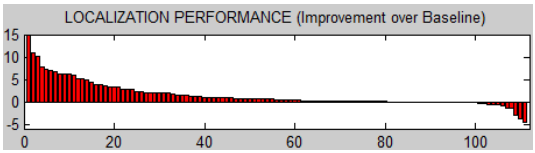
Standard single detectors are used as the baseline local detectors. For comparative evaluation, we compare with the method in [6]. Table 1 lists the average precision for object localization and presence prediction on SUN09 dataset.

Table 1 Mean AP (Averaged across All Object Categories) for Localization and Presence Prediction

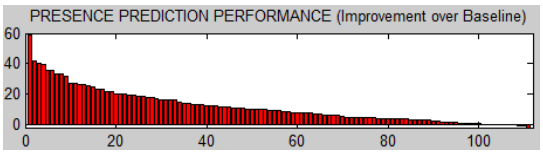
MEAN AP	Localization	Presence Prediction
Baseline	13.31	6.82
Tree Context	20.74	7.91
BLSTM-Tree Context	24.98	8.30

From Table 1, we get the observations that the average precision of the proposed method is higher than that of the approach in [6], which use the tree model for prediction on SUN09 dataset.

To further describe the performance of the proposed method, we also list the improvement of our method over the baseline and tree context model in terms of each object category respectively.

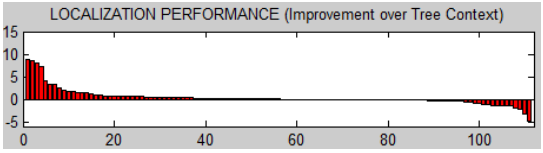


Index of object category

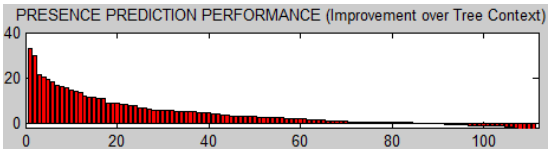


Index of object category

Fig.4. Improvement of localization and presence in terms of our method over baseline.



Index of object category



Index of object category

Fig.5. Improvement of localization and presence in terms of our method over tree context.

7. Summary

This paper has proposed a BLSTM-Tree model for object localization and presence prediction.

To employ the relationships among different object categories in BLSTM-Context model, we introduce the BLSTM-RNN which is good at dealing with unfixed sequence to mine the contextual dependencies among different objects in an image.

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[2] D. Hoiem, A. A. Efros, and M. Hebert, “Putting objects in perspective,”Int. J. Comput. Vis., vol. 80, no. 1, pp. 3–15, 2008.

[3] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5): 602-610.

[4] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, “Exploiting hierarchical context on a large database of object categories,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), San Francisco, CA, USA,Jun. 2010, pp. 129–136.

[5] Doetsch P, Kozielski M, Ney H. Fast and robust training of recurrent neural networks for offline handwriting recognition [C]//Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. IEEE, 2014: 279-284.

[6] M. J. Choi, A. Torralba, and A. S. Willsky, “A tree-based context model for object recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34,no. 2, pp. 240–252, Feb. 2012.