Iterative Relation Reasoning for Multiple Object Recognition

Tianyu Zhou* and Zhenjiang Miao

School of Computer and Information Technology, Beijing Jiaotong University, Haidian District, 100044, Beijing, China

*E-mail: 12120360@bjtu.edu.cn

Abstract. We present a novel model for multiple object recognition. Our model combines current deep learning recognition systems with object category relation information. This model is mainly inspired by spatial memory network [1], which treats multiple object recognition task as an iterative process, reusing some region feature as context information. We extend this work by presenting a statistical-based category relation model to measure object category semantic relevance. With the spatial memory and category relation model as relation reasoning modules, our model takes context information. Our model achieves 4.4% per-instance and 4.9% per-class absolute improvement of average classification accuracy over plain convolutional neural networks on ADE dataset.

1. Introduction

Deep convolutional neural networks have achieved remarkable results in image object detection task in recent years. General CNN-based detection methods first generate candidate visual regions, and then classify these regions according to their CNN feature. In multiple object recognition task, most methods deal with each region individually, ignoring the correlations between objects. In fact, objects in the same scene share context and semantic relation information, which can be important clue for recognition [2, 3].

In this paper, we consider modeling and using this relation information from two aspects: co-occurrence object category distribution and iterative reasoning. In the first aspect, there are usually certain correlations between specific object categories such as windows and doors are always appearing at the same time, or some object categories are more likely to be in a specific scene such as a plane in the sky. Modeling this kind of category relation information can help better recognize objects. The second aspect refers to the recognition process of the human visual system. Humans usually recognize objects from primary to secondary. Recognized objects can be a clue for others. Based on these two points we design an iterative relation reasoning model to improve the results of multiple object detection. In our approach, we use Faster R-CNN [4] as our base object detection network. We propose a category relation model to measure category correlation, and use a modified spatial memory network [1] to achieve iterative reasoning. We test our iterative relation reasoning method on ADE dataset [5]. Experiment results show that our method can improve the average recognition accuracy, especially for some hard examples such as small size, occlusion and deformed objects.
2. Related Work
Since AlexNet [6] made a success of image classification, deep learning has become the mainstream of image processing tasks. The evolution of convolutional neural networks (CNN) in these years has made the deep models’ feature expression ability stronger. For object detection task, some CNN-based methods have specific modules to solve detection problems. R-CNN method [7] makes detection problem as region prediction and classification problem. R-CNN first generates many candidate regions, which may contain a target object, and then classifies these regions with CNN. This framework has two main steps so is called two-stage method. The improved R-CNN methods [4, 8] reduce redundancy in feature calculations, and use region proposal network to propose regions which highly improved model efficiency. Different from two-stage R-CNN series methods, one-stage approaches like YOLO [9] and SSD [10] directly run detection over dense grid sampling locations. One-stage methods are faster than two-stage methods without region proposal step but has a little reduce of accuracy.

In the methods mentioned above, each region of interest is classified individually without considering the scene context. As far as we know, an object always has close correlations with other objects in the same scene. This contextual information plays an important role in human visual recognition system, especially when object does not have sufficient appearance features. Analyzing scene information can help us recognize target object more accurately [2, 3]. Much work has been done on the context modeling [11-13] in the early years. In recent years, some research is trying to model contextual information using deep convolutional networks. Inside outside network (ION) [14] uses spatial recurrent neural networks to integrate contextual information outside the region of interest. DeepIDNet [15] proposes to refine the detection scores with global contextual information. MultiRegion CNN [16] extracts features from regions surrounding the object proposals as contextual information. In these related methods, contextual information is used in the form of convolutional features, but the correlation between each two objects is not explicitly expressed. Spatial Memory Network (SMN) [1] is a relation-based model. This method uses spatial memory to assemble object features into a pseudo image representation, which can be utilized and updated in a sequential reasoning architecture.

Although SMN consider the relation between two objects in a specific image, the category semantic is overlooked. Category distribution from a large number of images can reflect the semantic relevance between different object categories [9, 17]. Our work mainly extends the work of SMN. We propose a relation model to measure category correlation, and use it in the iterative relation reasoning.

3. Relation Reasoning
Our detection model is based on Faster R-CNN model with two additional modules. A category relation model defines the semantic relevance between different object categories. Spatial memory network is used to store previous context information and refine recognition results in an iterative process. We show the relation reasoning framework in Figure 1. The details of our main network and modules will be explained in the following sections.

**Figure 1.** Framework of our whole model. Spatial memory and category relation model are two key modules of relation reasoning. Multiple object regions are classified iteratively.
3.1. Base Network
The base network does not depend on specific detection model, and we just build it on top of Faster R-CNN as the SMN original paper [1] does. In our task, we just need to classify target regions, so we modified the original model by removing region proposal network and bounding box regression module. For feature extraction we use VGG16 [18] as the base network, which has 13 convolutional layers and 5 max-pooling layers. Given the input image height $h$ and width $w$, the last convolutional layer feature size $(h', w')$ is approximately $1/16$ of the original image size. The region classification network uses two fully connected layers.

3.2. Spatial Memory
One key module is spatial memory. As mentioned in [1], the idea of SMN is to use a spatial memory $S$ to store the information of previously detected objects. $S$ is a 3-dimensional tensor, the size of which is same as the last convolutional layer, and the depth $d$ is 512. Each cell of $S$ stores the CNN features of object instances, and the locations of cells reflect spatial information of object instances. Once the object feature information is stored in $S$, it can be used to predict other objects. For example, it would be easy to recognize a “window” once a “door” is recognized.

$S$ is updated each time region feature input. Both high-level and middle-level features are used. For the location of region $r$, middle-level feature is cropped and resized to a $7 \times 7$ square, while high-level feature is appended to all the locations. Then we use two $1 \times 1$ convolutions to fuse the information. In the same region location, memory $S$ is cropped and resized to a $7 \times 7$ square $s_r$. We use a gated recurrent unit (GRU) [19] to iteratively update $S$:

$$s'_r = u \cdot s_r + (1 - u) \cdot \sigma(W_{fr} \cdot r + W_z \cdot z + b)$$  \hspace{1cm} (1)

In the equation above, $s'_r$ is updated memory $s_r$, $u$ is update gate and $z$ is reset gate in the definition of GRU. $W_f, W_r$ and $b$ are convolutional weights and bias, $r$ is the feature of region $r$. $\sigma(\cdot)$ is activation function, $\cdot$ is entry-wise product. This update process continues until all regions have been calculated.

During recognition $S$ is the input of a reasoning network, which has 3 convolutional layers and 2 fully connected layers. Such memory $S$ preserves multiple region feature with spatial information, which effectively improved reasoning result.

3.3. Category Relation Model
Another important module in our approach is category relation model. From an intuitive understanding, objects with high semantic relevance may appear in the same scene. For example, “car”, “motorbike” and “traffic light” are common objects in a street view image, while “desk”, “computer” and “keyboard” are more likely to be present in an indoor scene image. If an object instance cannot be classified easily from its appearance features, what other objects around it may be important clue to help reasoning. Our category relation model is built from statistical-based category distribution, which aims to measure this semantic relevance between objects and refine recognition result during context reasoning.

To build this model, first we count the category frequencies from image dataset annotations. We define matrix $M$ as the original category correlation matrix. Given $N$ as the category number of the whole dataset, $M$ is a $N \times N$ matrix. Each $W_y \in M$ represents the correlation degree of the ith and the jth category. We initialize $M$ from every training set image labels. Given an image $I$, its labels $L^I = \{I_1, ..., I_l\}$ represent the category ids in the label set. For image $I$, let $c' (N \times N)$ represent the correlation of categories in $L^I$. For all category ids in label set, we set $c'_{ij}$ to 1 if both ith and jth id are present in $L^I$; we set we set $c'_{ij}$ to a small negative value $-\varepsilon$ as negative correlation if only one of $i$ and $j$ is in $L^I$; the rest of $c'_{ij}$ are zeroes. So the assignment of $c'_{ij}$ is given as:

$$c'_{ij} = \begin{cases} 
1 & (i, j \in L^I) \\
-\varepsilon & (i \in L^I, j \notin L^I \text{ or } i \notin L^I, j \in L^I) \\
0 & (i, j \notin L^I)
\end{cases}$$  \hspace{1cm} (2)
The value of $\epsilon$ depends on the dataset and is used to suppress low category relevance reasoning. For all images $I_1, \ldots, I_k$, the sum of $c_{ij}$ is the correlation degree:

$$W_{ij} = \sum_{t=1}^{k} c_{ij}$$ (3)

After summation and normalization, we get the final $M$ as category relation model. This model is used in the testing process. For each region $r$ in an input image, the final region category score vector $s$ ($N+1$ dim) can be represented as:

$$s = f(\omega_f s_f + \omega_s s_r)$$ (4)

in which $s_f$ is score from final convolutional features including spatial memory, and $s_r$ is score from category relation model. $\omega_f$ and $\omega_s$ are score weights depending on classification confidence. $f(\cdot)$ is the mapping function to make sure each value of the final score $s$ is in $[0, 1]$. $s_f$ directly comes from the output of last fully connected layer, and $s_r$ can be calculated by this equation:

$$s_r = \sum_{i \in C} A_{i}M_{sf}$$ (5)

$A$ is a binary adjacency matrix in which the previously detected objects number location is set to 1. $M$ is our category relation matrix, and $C$ is top-k possible category set by feature score $s_f$.

With the help of this category relation model, objects’ category semantic relevance is utilized to make up for the lack of feature classification.

4. Experiment

In order to evaluate our approach, we test our model on a real scene image dataset. Our main task is region classification, which aims to assign correct category labels to ground-truth region bounding boxes. This section tells the details of experiment steps and results.

4.1. Dataset

Since we need sufficient object instance annotations to build category relation model, we choose ADE dataset [5]. The training set has 20, 210 images, and the validation set has 2, 000 images is split into val-set and test-set with 1, 000 images each. The val-set is for adjusting parameters and the test-set is for testing only. Some data processing refers to [20]. We use the original raw names of object categories as categorization target. Because too many infrequent object instances are bad samples, we filter out these categories, leaving categories only more than 5 instances.

For building category relation model, the number of each object category should not be too small because a sparse relation model is difficult to represent correlations correctly. Then we further merge categories by their higher-level category names of raw names in wordnet annotations. After deduplication we finally get the number of categories 218 as the dimension of category relation model.

Since ADE is not a detection but segmentation dataset, we need to convert segmentation annotations to detection bounding boxes. For all target object categories, each object instance is created a bounding box.

4.2. Network Implementation Details

We implement a TensorFlow version of Faster R-CNN, using VGG16 [7] pre-trained on ImageNet [21] as image classifier. During training and testing process, all images are enlarged to shorter size 600 pixels. The ground-truth boxes are used as RoIs to compute region features. We feed the spatial memory $S$ with conv4 layer as mid-level features, and use conv5 layer as final features for prediction.

For training process, we adopt stochastic gradient descent and batch normalization. We use dynamic learning rate: $5e^{-4}$ as initial for 280K iterations and $5e^{-5}$ for the rest 40K iterations. We choose $1e^{-4}$ as weight decay and 0.9 as momentum.
4.3. Results
We have 3 sets of experimental results on the ADE test-set. First we test the original CNN recognition model without spatial memory and category relation model as baseline. Then we test the base model with spatial memory. At last, we add category relation model. Table 1 shows the results of all sets of experiments. From the results we can see our iterative relation reasoning modules perform much better than the baseline on ADE test-set. The spatial memory module achieves 70.8% average classification accuracy per-instance and 37.5% per-class, 3.8% per-instance and 4.4% per-class absolute improvement compared to the baseline. The final accuracy goes to 71.4% per-instance and 38.0% per-class when category relation model added. The overall absolute improvement is 4.4% per-instance and 4.9% per-class.

Table 1. Results on ADE test-set.

| Method                                | Average Classification Accuracy |
|---------------------------------------|---------------------------------|
|                                       | per-instance | per-class   |
| Baseline VGG16                        | 0.670         | 0.331       |
| Baseline + spatial memory             | 0.708         | 0.375       |
| Baseline + spatial memory + category relation model | 0.714         | 0.380       |

5. Conclusion and Discussion
We present a novel multiple object recognition method, which combines CNN with category relation information to improve average recognition accuracy. A statistical-based category relation model is designed to measure object category semantic relevance. In a relation reasoning model, relevant objects should not be recognized individually, so we add a spatial memory network to achieve iterative reasoning. We compared the original CNN recognition model with our model. Experiment results show that both spatial memory and category relation model can improve the overall average recognition accuracy, especially for some hard examples such as small size, occlusion and deformed objects. Between the two modules, the effect of spatial memory is more obvious because of the similar feature form and fine fusion of convolutions.

Acknowledgments
This work is supported by the NSFC 61672089, 61273274, 61572064, and National Key Technology R&D Program of China 2012BAH01F03.

Reference
[1] Chen X, Gupta A. Spatial memory for context reasoning in object detection[J]. arXiv preprint arXiv:1704.04224, 2017.
[2] Biederman I, Mezzanotte R J, Rabinowitz J C. Scene perception: Detecting and judging objects undergoing relational violations[J]. Cognitive psychology, 1982, 14(2): 143-177.
[3] Hock H S, Gordon G P, Whitehurst R. Contextual relations: the influence of familiarity, physical plausibility, and belongingness[J]. Perception & Psychophysics, 1974, 16(1): 4-8.
[4] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.
[5] Zhou B, Zhao H, Puig X, et al. Semantic understanding of scenes through the ADE20K dataset[J]. International Journal of Computer Vision, 2016: 1-20.
[6] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
[7] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[8] Girshick R. Fast r-cnn[C] // Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[10] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C] // European conference on computer vision. Springer, Cham, 2016: 21-37.

[11] Torralba A. Contextual priming for object detection[J]. International journal of computer vision, 2003, 53(2): 169-191.

[12] Oliva A, Torralba A. The role of context in object recognition[J]. Trends in cognitive sciences, 2007, 11(12): 520-527.

[13] Divvala S K, Hoiem D, Hays J H, et al. An empirical study of context in object detection[C] // Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 1271-1278.

[14] Bell S, Lawrence Zitnick C, Bala K, et al. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2874-2883.

[15] Ouyang W, Wang X, Zeng X, et al. Deepid-net: Deformable deep convolutional neural networks for object detection[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2403-2412.

[16] Gidaris S, Komodakis N. Object detection via a multi-region and semantic segmentation-aware cnn model[C] // Proceedings of the IEEE International Conference on Computer Vision. 2015: 1134-1142.

[17] Rabinovich A, Vedaldi A, Galleguillos C, et al. Objects in context[C] // Computer vision, 2007. ICCV 2007. IEEE 11th international conference on. IEEE, 2007: 1-8.

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[19] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.

[20] Chen X, Li L J, Fei-Fei L, et al. Iterative Visual Reasoning Beyond Convolutions[J]. arXiv preprint arXiv:1803.11189, 2018.

[21] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.