

# Learning Local Instance Constraint for Multi-label Classification

Shang Luo<sup>1,2</sup>, Xiaofeng Wu<sup>1,2</sup>, Bin Wang<sup>1,2(✉)</sup>, and Liming Zhang<sup>1,2</sup>

<sup>1</sup> Key Laboratory for Information Science of Electromagnetic Waves,  
Fudan University, Shanghai, China

{sluo15, xiaofengwu, wangbin, lmzhang}@fudan.edu.cn

<sup>2</sup> Research Center of Smart Networks and Systems, School of Information  
Science and Technology, Fudan University, Shanghai, China

**Abstract.** Compared to single-label image classification, multi-label image classification outputs unknown-number objects of different categories for an input image. For image-label relevance in multi-label classification, how to incorporate local information of objects with global information of label representation is still a challenging problem. In this paper, we propose an end-to-end Convolutional Neural Network (CNN) based method to address this problem. First, we leverage CNN to extract hierarchical features of input images and the dilated convolution operator is adopted to expand receptive fields without additional parameters compared to common convolution operator. Then, one loss function is used to model local information of instance activations in convolutional feature maps and the other to model global information of label representation. Finally, the CNN is trained end-to-end with a multi-task loss. Experimental results show that the proposed proposal-free single-CNN framework with a multi-task loss can achieve the state-of-the-art performance compared with existing methods.

**Keywords:** CNN · Multi-label classification · Multi-task

## 1 Introduction

Single-label image classification, which just outputs a dominant label from a predefined label set for an input image, has been studied during the past years. However, real-world images mostly contain multiple objects of different categories, thus multi-label image classification needs to be considered for real-world images and usually it is a more complex and challenging task.

In recent years, Convolutional Neural Network (CNN) [1] has achieved great success in single-label image classification [2–4]. Inspired by this, recent state-of-the-art works for multi-label image classification are mainly involved with CNN and these methods can be generally categorized into two types: based on proposals and based on

---

This work was supported in part by the National Natural Science Foundation of China under Grant 61572133.

multi-network. The first type of methods [5–7] has a multi-stage pipeline in training phase that first generate object proposals for an input image, and then makes predictions from features extracted by a CNN for each proposal. Although proposal based methods can produce high quality proposals, most of these proposals are redundant and thus proposal selection is required to reduce computation. The second type of methods [7, 8] trains a fusion model of multiple CNNs or CNN combining with Recurrent Neural Network (RNN). These multi-network models usually have more parameters to tune and in practice are harder to converge. Moreover, combination of local information of objects and global information of label representation is not considered in these methods.

To address the problems above, in this paper we propose a proposal-free single-CNN based multi-label classification framework with a multi-task loss. Firstly, a CNN is used to extract hierarchical features for an input image. By directly taking an image as input instead of multiple region proposals, the redundant proposal extraction process is avoided. Secondly, the dilated convolution operation is adopted to expand receptive fields without additional parameters compared to common convolutional operation, which will benefit further global information representation. Thirdly, inspired by [9], with stronger activations in convolutional feature maps of higher layers generally corresponding to positions of object instances in the image, bounding box annotation (ground-truth rectangle tightly enclosed an object) of each instance can be considered as local constraint information with strong label. To leverage this insight into multi-label classification, the CNN model is trained with a multi-task loss composed of two loss functions: one is to model local information of instance activations in convolutional feature maps and the other model global information of label representation.

The main contributions of our work can be briefly summarized as follows:

- An end-to-end proposal-free method with single-CNN framework for multi-label image classification is proposed.
- The dilated convolution operation is adopted to expand receptive fields for aggregating multi-scale contexture information without additional parameters.
- A multi-task loss is utilized to leverage local information of object instances and global information of label representation to enhance the discriminative capability of CNN.

The rest of this paper is organized as follows. The proposed method is given in Sect. 2, in which the basic structure of CNN, the dilated convolution operator and the multi-task loss are described in details. Section 3 shows experimental results on two widely used datasets and the performance comparisons of the proposed method with the state-of-the-art methods. Finally, concluding remarks are drawn in Sect. 4.

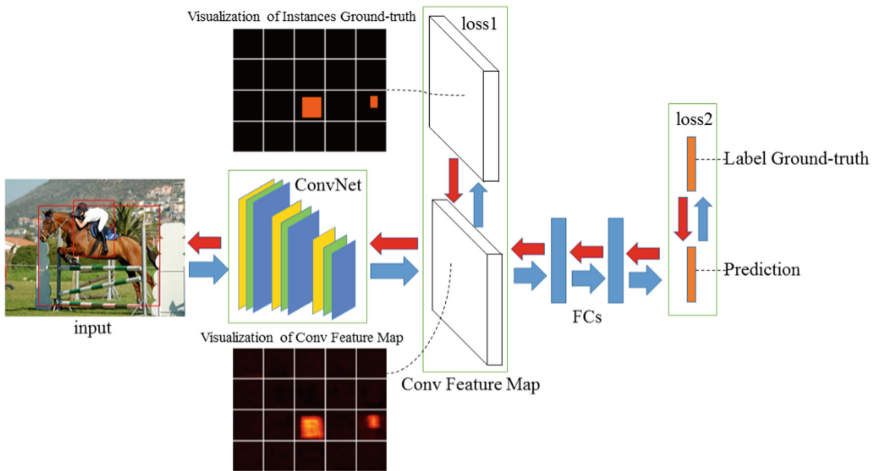
## 2 Our Method

To address the multi-label image classification problem, we propose an end-to-end proposal-free single-CNN based framework with a multi-task loss. Figure 1 shows that our method comprises three main parts: hierarchical feature learning of CNN (ConvNet

in Fig. 1), local instance constraint on convolutional feature maps (loss1 in Fig. 1) and global presentation in label space for classification (loss2 in Fig. 1). Contributing to hierarchical feature learning, the basic structure of CNN and the dilated convolution operation are separately described in Subjects. 2.1 and 2.2, and a multi-task loss composed of two loss functions is elaborated in Subject 2.3.

### 2.1 Basic Structure of Convolutional Neural Network

A CNN is generally composed of several convolutional and pooling layers (denoted as C layers and P layers) to extract hierarchical features from the original inputs or receptive fields, subsequently with several fully connected layers (denoted as FC layers) followed for specific tasks, as shown in Fig. 2.



**Fig. 1.** Framework of the proposed method with a multi-task loss (loss1 represents local instance constraint and loss2 represents global representation in label space). Blue arrows indicate forward computation and red arrows indicate backward computation in CNN. Red rectangles in input represent bounding box annotations. Black dashed lines show description of output in CNN. (Color figure online)

Assumed that a CNN is constructed with  $L$  layers and the output of the  $l$ -th layer is denoted as  $\mathbf{I}^l$ , where  $l \in \{1, 2, \dots, L\}$ , layer, specifically  $\mathbf{I}^0$  denotes the input data. As shown in Fig. 2, the input data is connected locally to a convolutional where a 2-D convolution operation is performed with convolutional kernels  $\mathbf{W}_l^c$  and a bias term  $\mathbf{b}_l^c$  is added to the resultant feature maps. To model nonlinearities in CNN, an activation function  $\partial(\cdot)$  is generally performed following convolutional layers. Then, a pooling operation  $\text{pool}(\cdot)$  is usually followed to achieve shift-invariance by reducing the resolution of the feature maps. The general C-P block of CNN can be formulated as

$$\mathbf{I}^l = \text{pool}(\partial(\mathbf{I}^{l-1} * \mathbf{W}_i^c + \mathbf{b}_i^c)), \quad (1)$$

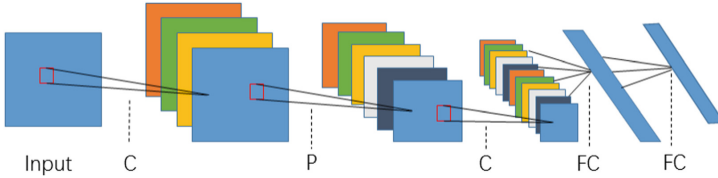
where  $*$  denotes the convolution operation. After some C-P blocks, hierarchical features are further transformed into 1-D feature vector by the FC layers. The FC layers connect all neurons in the previous layer to each single neuron of the current layer to generate global semantic information. Denoting weight as  $\mathbf{W}_i^{fc}$  and bias as  $\mathbf{b}_i^{fc}$ , an FC layer computation can be formulated as follows:

$$\mathbf{I}^l = \partial(\mathbf{I}^{l-1} \bullet \mathbf{W}_i^{fc} + \mathbf{b}_i^{fc}) \quad (2)$$

The output of the last FC layer is usually fed to an output layer using certain operations for specific tasks, for example, softmax operation is used for multi-class classification. Suppose we have  $N$  desired input-output pairs  $\{(\mathbf{x}^n, \mathbf{y}^n); n \in [1, 2, \dots, N]\}$ , where  $\mathbf{x}^n$  is the  $n$ -th input data and  $\mathbf{y}^n$  is its corresponding target label and  $\mathbf{t}^n$  is the corresponding output of CNN. Denoting  $\boldsymbol{\theta}$  as all the parameters of CNN, the loss of CNN can be computed as

$$L = \frac{1}{N} \sum_{n=1}^N \ell(\boldsymbol{\theta}; \mathbf{y}^n, \mathbf{t}^n) \quad (3)$$

Training a CNN can be seen as an optimization of function mapping, i.e., to minimize the loss of CNN, and generally, stochastic gradient descent (SGD) is used to find the best fitting set of parameters.



**Fig. 2.** Common CNN architecture with convolution layers (C), pooling layers (P) and fully connected layers (FC).

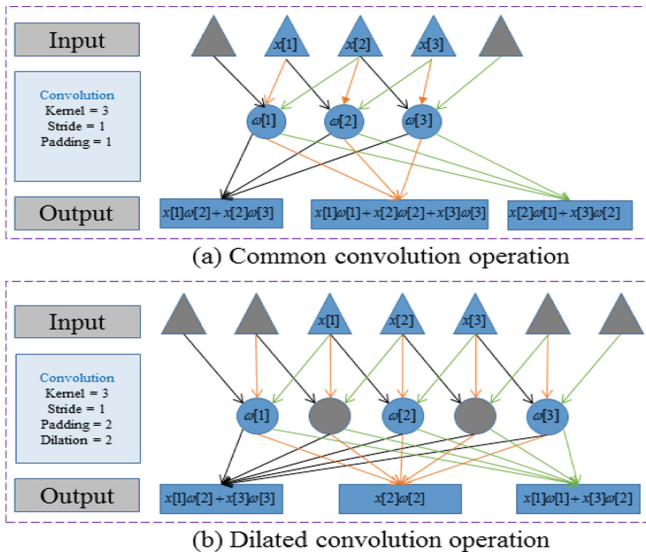
## 2.2 Dilated Convolutional Neural Network

Compared to common convolution operation, the dilated convolution operator is used to gain context information like cross-layer connection in [10]. Unlike the deconvolutional layer [10], with dilation rate [11] in CNN, the dilated convolution operation can apply the same convolutional kernel at different scales without additional memory and loss of information. Combining with proper parameter stride and padding in convolution operation, the dilated convolution operation can be used for multi-scale context information, which is demonstrated to be superior to cross-layer connection [12].

Considering one-dimensional convolution operator with a kernel  $\omega[m]$  of length  $M$  for a 1-D input signal  $x[i]$ , the output  $y[i]$  is defined as

$$y[i] = \sum_{m=1}^M x[i + d \cdot m] \omega[m], \tag{4}$$

where  $d$  is the dilation rate for input sampling. Thus common convolution operation can be seen as a special case of dilated convolution with a dilation rate of 1. In practice, as shown in Fig. 3, the dilated convolution operator with kernel size of  $k \times k$  and dilation rate of  $d$  just inserts  $d - 1$  zeros between consecutive filter values, transforming kernel size of  $k$  to  $k + (k - 1)(d - 1)$  without additional computation and memory.



**Fig. 3.** Illustration of common convolution operation and dilated convolution operation in one dimension. (a) Common convolution (dilation rate of 1). (b) Dilated convolution (dilation rate of 2, insert zero between adjacent filter values).

Due to dilation rate in convolution operation, the effective kernel size increases, but the number of filter parameters remains the same because of insertion of zero values. By aggregating dilated convolution in a chain of layers with proper stride and padding, a CNN can produce feature maps with desired resolution and larger receptive fields, which contains more context information and benefits for semantic representation.

### 2.3 Multi-task Loss

The proposed single-CNN framework is trained with a multi-task loss composed of two loss functions. The first loss  $L_{\text{act}}$  involves a  $H \times W \times C$  convolutional feature maps, in which each  $H \times W$  plane represents an activation map of the category. The second loss  $L_{\text{cls}}$  involves a discrete probability over  $C$  categories.

Each input image is labeled with a multi-label ground-truth and instances ground-truth. A multi-task loss  $L$  is used to jointly train for multi-label classification:

$$L = L_{\text{act}} + \lambda L_{\text{cls}}, \quad (5)$$

where the hyper-parameter  $\lambda$  controls the balance between the two task losses.

**Local Instance Constraint.** As discovered in [9], CNN can learn hierarchical features due to its deep architecture, and higher complex features are sensitive to local structures in the input images. Following these works, we propose a loss function that considers precise instance location and activation values in convolutional feature maps, allowing the network to capture local structures of each individual object instance.

Based on [12], the dilated convolution operator described in Sect. 2.2 is employed to expand receptive fields, and after the last convolution operation, a  $1 \times 1$  convolutional layer with the same number of filters as the number of categories is adopted. In this way, as shown in Fig. 1, each plane of convolutional feature maps stands for one specific category, thus higher activations in specific feature map indicate higher existing probability of the category. For local instance constraint, a Euclidean distance based loss function is adopted for penalizing the position with no object and constraining the activation values where there are objects corresponding to the category. Thus, for  $N$  training samples, the loss function  $L_{\text{act}}$  is Euclidean distance between convolutional feature map  $f^{c,i}$  and sum of instance bounding box masks  $\sum_{t=1}^{T(c,i)} b_t^{c,i}$  over  $C$  categories, which can be expressed as:

$$L_{\text{act}} = \sum_{i=1}^N \sum_{c=1}^C \left\| f^{c,i} - \sum_{t=1}^{T(c,i)} b_t^{c,i} \right\|, \quad (6)$$

where  $b_t^{c,i} \in \{0, 1\}$  (1 indicates the position with instances and 0 indicates the position without instance) is the  $t$ -th instance bounding box mask for category  $c$  and  $T(c, i)$  is the number of instances in the category  $c$  in the  $i$ -th image. There may exist overlapped instances in each individual category and we encoded its overlapped regions of instances by summing all of the individual binary masks to make the loss function  $L_{\text{cls}}$  surely aware of the higher activation values of objects.

**Global Label Representation.** For global representation, previous works mainly choose Euclidean distance [6, 7] or cross-entropy [8] for distance metric, but no work discusses pros and cons of the two metric learning for multi-label image classification. For each input image with a ground-truth class label  $\mathbf{u}$  and predicted class label  $\mathbf{v}$ , by adopting Euclidean distance the loss function  $L_{\text{cls}}$  can be defined as:

$$L_{cls} = \frac{1}{N} \sum_{i=1}^N \sqrt{\sum_{c=1}^C (u^{c,i} - v^{c,i})^2}, \quad (7)$$

and by adopting cross-entropy the loss function  $L_{cls}$  can be defined as:

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C (u^{c,i} \log v^{c,i}), \quad (8)$$

where  $u^{c,i}$  is the ground-truth label indicator for category  $c$  for  $i$ -th image and  $v^{c,i}$  corresponds to its prediction. The two losses will be compared in Subject 3.3.

### 3 Experimental Results

#### 3.1 Datasets and Baseline

Our method is evaluated on the VOC datasets [13], which is widely used as benchmark datasets for multi-label object recognition task. Following [5–8], VOC 2007 and VOC 2012 are chosen as our experimental datasets, which has been split into 3 parts: TRAIN, VAL and TEST. Like [6–8], we take TRAIN and VAL as our training datasets and TEST for model evaluation. Details of these datasets are shown in Table 1, in which the 20 classes are *airplane (aero)*, *bike*, *bird*, *boat*, *bottle*, *bus*, *car*, *cat*, *chair*, *cow*, *table*, *dog*, *horse*, *motorbike (motor)*, *person*, *plant*, *sheep*, *sofa*, *train* and *television (tv)*. The evaluation metric is average precision (AP) and mean average precision (mAP). In particular, for VOC 2007 TEST, the scores are evaluated with standard VOC evaluation package and for VOC 2012 TEST, the scores are evaluated on VOC evaluation server.

**Table 1.** Datasets information.

Dataset	#TRAINVAL	#TEST	#Classes
VOC 2007	5011	4952	20
VOC 2012	11540	10991	20

We compare the proposed method with several state-of-the-art approaches [6–8, 15–17, 19] in terms of metric mAP and the results are shown in Sect. 3.3.

#### 3.2 Parameters Configuration

Our CNN architecture is based on VGG16 [3], which is pre-trained on ImageNet. Following DeepLab [12], layer fc6 and fc7 are converted into convolutional layers and the dilated convolution operator is employed in layers conv5\_1, conv5\_2, conv5\_3, and fc6. More details of CNN architecture can be seen in Table 2. We fine-tune the VGG model from [12] using SGD with initial learning rate  $10^{-5}$ , 0.9 momentum,

0.0005 weight decay through caffe deep learning framework [14]. The hyper-parameter in  $\lambda$  Eq. 5 is set to 1 in all experiments.

### 3.3 Multi-label Classification Results

**Multi-label Image Classification on VOC 2007.** Table 3 reports our experimental results compared with the state-of-the-arts on VOC 2007. In the upper part of Table 3 above the double strike, we compared with those methods without using bounding box annotations for training, while the lower part shows the methods with bounding box information. For the state-of-the-art methods, INRIA [15] and FV [16] are hand-crafted based methods, and CNN-SVM [17] uses OverFeat [18] as a feature extractor, and the rest are CNN-based methods mainly fine-tuning pre-trained models on ImageNet.

**Table 2.** CNN architecture of the proposed method.

No.	Convolution					Pooling
	Kernel	Stride	Padding	Dilation	#Filter	
1	3	1	1	1	64	No
2	3	1	1	1	64	Yes
3	3	1	1	1	128	No
4	3	1	1	1	128	Yes
5	3	1	1	1	256	No
6	3	1	1	1	256	Yes
7	3	1	1	1	512	No
8	3	1	1	1	512	No
9	3	1	1	1	512	Yes
10	3	1	2	2	512	No
11	3	1	2	2	512	No
12	3	1	2	2	512	Yes
13	4	1	4	6	4096	No
14	1	1	1	1	4096	No
15	1	1	1	1	20	No

**Table 3.** Average precision (AP in %) comparison for the state-of-the-art methods on VOC 2007 TEST. The upper part shows the results of methods without using ground-truth annotations and the lower part shows methods with bounding box information.

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
INRIA[15]	77.2	69.3	56.2	66.5	45.5	68.1	83.4	53.6	58.3	51.1	62.2	45.2	78.4	69.7	86.1	52.4	54.4	54.3	75.8	62.1	63.5
FV[16]	75.7	64.8	52.8	70.6	30.0	64.1	77.5	55.5	55.6	41.8	56.3	41.7	76.3	64.4	82.7	28.3	39.7	56.6	79.7	51.5	58.3
CNN-SVM[17]	88.5	81.0	83.5	82.0	42.0	72.5	85.3	81.6	59.9	58.5	66.5	77.8	81.8	78.8	90.2	54.8	71.1	62.6	87.4	71.8	73.9
I-FT[6]	91.4	84.7	87.5	81.8	40.2	73.0	86.4	84.8	51.8	63.9	67.9	82.7	84.0	76.9	90.4	51.5	79.9	54.1	89.5	65.8	74.4
HCP-1000C[6]	95.1	<b>90.1</b>	92.8	89.9	51.5	80.0	<b>91.7</b>	91.6	57.7	77.8	70.9	89.3	89.3	85.2	93.0	64.0	85.7	62.7	94.4	78.3	81.5
CNN-RNN[8]	<b>96.7</b>	83.1	<b>94.2</b>	<b>92.8</b>	61.2	82.1	89.1	<b>94.2</b>	64.2	<b>83.6</b>	70.0	<b>92.4</b>	<b>91.7</b>	84.2	<b>93.7</b>	59.8	<b>93.2</b>	75.3	<b>99.7</b>	<b>78.6</b>	<b>84.0</b>
PRE-1000C[19]	88.5	81.5	87.9	82.0	47.5	75.5	90.1	87.2	61.6	75.7	67.3	85.5	83.5	80.0	95.6	60.8	76.8	58.0	90.4	77.9	77.7
d-CNN	90.2	85.5	88.5	87.5	53.7	80.7	85.0	85.4	60.3	73.0	78.0	82.9	84.8	81.9	88.6	65.4	82.9	73.8	90.1	76.7	79.7
CNN-L-GE	91.1	86.0	89.1	87.0	55.3	83.4	88.6	88.4	67.5	77.8	80.0	86.2	86.8	85.1	92.0	65.4	83.2	78.1	92.5	77.9	82.1
CNN-L-GC	91.7	87.9	89.8	88.6	<b>61.3</b>	<b>84.0</b>	90.1	88.4	<b>71.3</b>	83.3	<b>82.4</b>	89.2	89.6	<b>86.9</b>	93.4	<b>71.0</b>	85.8	<b>79.6</b>	<b>93.8</b>	<b>78.6</b>	<b>84.3</b>



From Table 3 it can be seen that the CNN-based methods outperform the hand-crafted methods with a large margin of more than 10%, which indicates that hierarchical features of CNN greatly benefits for image representation. PRE-1000C [19] fine-tunes pre-trained models on ImageNet with limited VOC data. Compared with PRE-1000C, 2% improvement can be achieved by our d-CNN (CNN with dilated convolution operation) which takes advantage of dilated convolution operator to learn more semantic information. HCP-1000C [6] is a proposal-based method that relies on proposal extraction method to prepare input patches. Compared with HCP-1000C, both our CNN-L-GE (CNN with local instance constraint and global representation of Euclidean distance) and CNN-L-GC (CNN with local instance constraint and global representation of cross-entropy metric) get higher mAP, which shows a positive effect on multi-task learning because the two tasks, separately involving with local and global information, influence each other through shared parameters. In terms of loss function measuring global representation, cross-entropy achieves a further 2.2% performance than that of Euclidean distance, which verifies the discovery that Euclidean distance is not suitable for distance metric of sparse data in high dimension [20]. Compared with the state-of-the-art method CNN-RNN that uses CNN and RNN to model label dependency and image-label representation, our CNN-L-GC with only one network achieves competitive performance, which demonstrates the effectiveness of the multi-task learning both the local and global information. In particular, the proposed method outperforms the state-of-the-art methods with a large margin when the objects are nearly squared (i.e., *bus*, *chair*, *table*, *motor*, *plant*, and *sofa*), mainly due to local instance constraint from bounding box annotations.

**Multi-label Image Classification on VOC 2012.** Table 4 reports our experimental results compared with the state-of-the-art methods on VOC 2012. Similar to Table 3, we compare with methods without using bounding box annotations in the upper part and methods with bounding box information in the lower part.

**Table 4.** Average precision (AP in %) comparison for the state-of-the-art methods on VOC 2012 TEST. The upper part shows the results of methods without using ground-truth annotations and the lower part shows methods with bounding box information.

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
I-FT[6]	94.6	74.3	87.8	80.2	50.1	82.0	73.7	90.1	60.6	69.9	62.7	86.9	78.7	81.4	90.5	45.9	77.5	49.3	88.5	69.2	74.7
HCP-1000C[6]	<b>97.7</b>	83.0	<b>93.2</b>	87.2	59.6	88.2	81.9	94.7	66.9	81.6	68.0	93.0	88.2	87.7	92.7	59.0	85.1	55.4	93.0	77.2	81.7
HCP-2000C[6]	97.5	84.3	93.0	89.4	62.5	90.2	84.6	<b>94.8</b>	69.7	<b>90.2</b>	74.1	<b>93.4</b>	<b>93.7</b>	88.8	93.3	59.7	<b>90.3</b>	61.8	94.4	78.0	84.2
PRE-1000C[19]	93.5	78.4	87.7	80.9	57.3	85.0	81.6	89.4	66.9	73.8	62.0	89.5	83.2	87.6	95.8	61.4	79.0	54.3	88.0	78.3	78.7
PRE-1512C[19]	94.6	82.9	88.2	84.1	60.3	89.0	84.4	90.7	72.1	86.8	69.0	92.1	93.4	88.6	<b>96.1</b>	64.3	86.6	62.3	91.1	79.8	82.8
FeV[7]	96.8	<b>87.8</b>	88.7	87.2	<b>63.8</b>	<b>92.3</b>	86.2	92.3	72.4	82.0	76.0	91.9	90.3	90.3	95.2	61.2	82.6	65.6	92.8	<b>84.4</b>	84.0
CNN-L-GE	91.1	86.0	89.1	87.0	55.3	83.4	<b>88.6</b>	88.4	67.5	77.8	<b>80.0</b>	86.2	86.8	85.1	92.0	<b>65.4</b>	83.2	<b>78.1</b>	92.5	77.9	82.1
CNN-L-GC	<b>97.7</b>	85.2	91.1	<b>90.0</b>	62.7	91.7	86.1	94.4	<b>75.8</b>	84.5	79.7	92.8	92.4	<b>91.9</b>	<b>96.1</b>	64.9	85.7	77.2	<b>95.4</b>	82.6	<b>85.9</b>

The multi-label classification results on VOC 2012 in terms of mAP are consistent with those in Table 3. Compared with HCP-2000C [6] pre-trained on ImageNet with 2000 categories and PRE-1512C [19] pre-trained on ImageNet with 1512 categories, our CNN-L-GC pre-trained on ImageNet with only 1000 categories outperforms the two state-of-the-art methods by 1.7% and 3.1%. Compared with the state-of-the-art

proposal-based FeV [7] with two-stream CNN, our CNN-L-GC has an improvement of 1.9%. Similar to results on VOC 2007, the proposed method takes advantage of squared objects because of local instance constraint with bounding box annotations.

## 4 Conclusions

In this paper, we presented an end-to-end proposal-free single-CNN based method multi-label image classification framework with a multi-task loss. Without region proposals extraction, the training phase of our work is a single-stage pipeline. Compared with the existing works, our method adopted the dilated convolution operation to expand receptive fields without additional parameters. Further, the proposed method utilized instance constraint for local information and cross-entropy metric for global information representation at the same time to leverage a multi-task learning for boosting the discriminative capacity of CNN. The experimental results on VOC 2007 and VOC 2012 showed that the proposed method achieved the state-of-the-art performance.

## References

1. LeCun, Y.: Gradient based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2323 (1998)
2. Krizhevsky, A.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105. Nips Foundation, San Diego (2012)
3. Simonyan, K.: Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, pp. 1–14 (2015)
4. He, K.: Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. IEEE, New Jersey (2016)
5. Girshick, R.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587. IEEE, New Jersey (2014)
6. Wei, Y., et al.: CNN: single-label to multi-label. *arXiv preprint* (2014). [arXiv:1406.5726](https://arxiv.org/abs/1406.5726)
7. Yang, H.: Exploit bounding box annotations for multi-label object recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–288. IEEE, New Jersey (2016)
8. Wang Jiang, F.: CNN-RNN: A unified framework for multi-label image classification. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294. IEEE, New Jersey (2016)
9. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *ECCV 2014*. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-10590-1\\_53](https://doi.org/10.1007/978-3-319-10590-1_53)
10. Long, J.: Fully convolutional networks for semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440. IEEE, New Jersey (2015)
11. Yu, F.: Multi-scale context aggregation by dilated convolutions. *International Conference on Learning Representation*, pp. 1–13 (2016)

12. Chen, L.C., et al.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS (2014). arXiv preprint [arXiv:1412.7062](https://arxiv.org/abs/1412.7062)
13. Everingham, M.: The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **88**(2), 303–338 (2010)
14. Jia, Y.: Caffe: convolutional architecture for fast feature embedding. In: *ACM international conference on Multimedia*, pp. 675–678. ACM, New York (2014)
15. Harzallah, H.: Combining efficient object localization and image classification. In: *International Conference on Computer Vision*, pp. 237–244. IEEE, New Jersey (2009)
16. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010*. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-15561-1\\_11](https://doi.org/10.1007/978-3-642-15561-1_11)
17. Sharif Razavian A.: CNN features off-the-shelf: an astounding baseline for recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 806–813. IEEE, New Jersey (2014)
18. Sermanet, P.: Overfeat: integrated recognition, localization and detection using convolutional networks. In: *International Conference on Learning Representations*, pp. 1–16 (2014)
19. Oquab, M.: Learning and transferring mid-level image representations using convolutional neural networks. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1717–1724. IEEE, New Jersey (2014)
20. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) *ICDT 2001*. LNCS, vol. 1973, pp. 420–434. Springer, Heidelberg (2001). [https://doi.org/10.1007/3-540-44503-X\\_27](https://doi.org/10.1007/3-540-44503-X_27)