Short Papers

HCP: A Flexible CNN Framework for Multi-Label Image Classification

Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, *Senior Member, IEEE*, and Shuicheng Yan, *Senior Member, IEEE*

Abstract—Convolutional Neural Network (CNN) has demonstrated promising performance in single-label image classification tasks. However, how CNN best copes with multi-label images still remains an open problem, mainly due to the complex underlying object layouts and insufficient multi-label training images. In this work, we propose a flexible deep CNN infrastructure, called Hypotheses-CNN-Pooling (HCP), where an arbitrary number of object segment hypotheses are taken as the inputs, then a shared CNN is connected with each hypothesis, and finally the CNN output results from different hypotheses are aggregated with max pooling to produce the ultimate multi-label predictions. Some unique characteristics of this flexible deep CNN infrastructure include: 1) no ground-truth bounding box information is required for training; 2) the whole HCP infrastructure is robust to possibly noisy and/or redundant hypotheses; 3) the shared CNN is flexible and can be well pre-trained with a large-scale single-label image dataset, e.g., ImageNet; and 4) it may naturally output multi-label prediction results. Experimental results on Pascal VOC 2007 and VOC 2012 multi-label image datasets well demonstrate the superiority of the proposed HCP infrastructure over other state-of-the-arts. In particular, the mAP reaches 90.5% by HCP only and 93.2% after the fusion with our complementary result in [12] based on hand-crafted features on the VOC 2012 dataset.

Index Terms—Deep Learning, CNN, Multi-label Classification

1 Introduction

SINGLE-LABEL image classification, which aims to assign a label from a predefined set to an image, has been extensively studied during the past few years [10], [14], [18]. For image representation and classification, conventional approaches utilize carefully designed hand-crafted features, e.g., SIFT [29], along with the bag-of-words coding scheme, followed by the feature pooling [24], [32], [39] and classic classifiers, such as Support Vector Machine (SVM) [5] and random forests [3]. Recently, in contrast to the hand-crafted features, *learnt* image features with deep network structures have shown their great potential in various vision recognition tasks [21], [23], [25]. Among these architectures, one of the greatest breakthroughs in image classification is the deep convolutional neural network (CNN) [23], which has achieved the state-of-the-art performance (with 10% gain over the previous methods based on hand-crafted features) in the large-scale single-label object recognition task, i.e., ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [10] with more than one million images from 1,000 object categories.

- *Y. Wei is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China, and the Department of Electrical and Computer Engineering, National University of Singapore. E-mail: wychao1987@gmail.com.*
- *Y. Zhao is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China. E-mail: yzhao@bjtu.edu.cn.*
- *B. Ni is with the Department of Electronic Engineering, Shanghai Jiaotong University, China. E-mail: bingbing.ni@adsc.com.sg.*
- *W. Xia, M. Lin, J. Huang, J. Dong and S. Yan are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. E-mail: {weixiaee, mavenlin, junshi.huang, djtcut}@gmail.com, eleyans@nus.edu.sg.*

Manuscript received 11 Feb. 2015; revised 9 Aug. 2015; accepted 5 Oct. 2015. Date of publication 25 Oct. 2015; date of current version 11 Aug. 2016.
Recommended for acceptance by A. Vedaldi.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2491929

Multi-label image classification is, however, a more general and practical problem, since the majority of real-world images contain objects from multiple different categories. For a typical multi-label image, different categories of objects are located at various positions with different scales and poses. Furthermore, the different composition and interaction between objects in multi-label images, like partial visibility and occlusion, also increase the complexity of the problem, which requires more annotated data to cover the different situations. For example, as shown in Fig. 1, for single-label images, the foreground objects are roughly aligned, while for multi-label images, even with the same label, i.e., *horse and person*, the spatial arrangements of the *horse* and *person* instances vary largely among different images. Compared to single-label images which are practically to collect and annotate, the burden of annotation for a large-scale multi-label image dataset is much heavier. Many methods [8], [12], [32] have been proposed to address this more challenging problem. The success of CNN on single-label image classification also sheds some light on the solution of the multi-label image classification problem. Generally, the CNN can well handle images with well-aligned objects, while it is relatively inaccurate in predicting images with objects severely mis-aligned or occluded. Therefore, by relaxing the multi-label problem into several single-label tasks and alleviating the issues of mis-alignment and occlusion, the great discriminating ability of the CNN model can be better exploited.

Recently, many hypothesis-based methods have been proposed for detection [9] and segmentation [40], [41]. By generating a pool of hypotheses of either bounding boxes or segments, the multi-label problem can be transformed into several sub-tasks of single-label prediction. Since object hypotheses generally have higher confidence of objectness, which means they are more likely to contain certain semantic objects, after cropping and normalization, both mis-alignment and occlusion can be somewhat alleviated. Motivated by the idea of hypothesis and the great single-label classification performance of the traditional CNN models, in this paper, we propose a flexible deep CNN structure, called Hypotheses-CNN-Pooling (HCP). HCP takes an arbitrary number of object segment hypotheses (H) as the inputs, which may be generated by the sate-of-the-art objectiveness detection techniques, e.g., binarized normed gradients (BING) [9] or EdgeBoxes [44], and then a shared CNN (C) is connected with each hypothesis. Finally, to aggregate the single-label CNN predictions from different hypotheses into multi-label results, a novel pooling layer (P) is integrated into the proposed CNN model to give the ultimate multi-label predictions. Particularly, the proposed HCP infrastructure possesses the following characteristics:

- *No ground-truth bounding box information is required for training on the multi-label image dataset.* Different from previous works [7], [12], [30], which employ ground-truth bounding box information for training, the proposed HCP requires no bounding box annotation. Since bounding box annotation is much more costly than labelling, the annotation burden is significantly reduced. Therefore, the proposed HCP has a better generalization ability when transferred to new multi-label image datasets.
- *The proposed HCP infrastructure is robust to noisy and/or redundant hypotheses.* To suppress the possibly noisy hypotheses, a cross-hypothesis max-pooling operation is carried out to fuse the outputs from the shared CNN into an integrative prediction. With max pooling, the high predictive scores from those hypotheses containing objects are preserved and the noisy ones are discarded. Therefore, as long as one hypothesis contains the object of interest, the noise can be suppressed after the cross-hypothesis pooling. Redundant hypotheses can also be well addressed by max pooling.

Fig. 1. Some examples from ImageNet [10] and Pascal VOC 2007 [13]. The foreground objects in single-label images are usually roughly aligned. However, the assumption of object alighment is not valid for multi-label images. Also note the partial visibility and occlusion between objects in the multi-label images.

- *The shared CNN is flexible and can be well pre-trained with a large-scale single-label image dataset*. To address the problem of insufficient multi-label training images, based on the Hypotheses-CNN-Pooling architecture, the shared CNN can be first well pre-trained on some large-scale single-label dataset, e.g., ImageNet, and then fine-tuned on the target multi-label dataset. Besides, the architecture of the shared CNN is flexible, and various advanced CNNs, e.g., Network-in-Network [28], Spatial Pyramid Pooling Net [20], Very Deep Net [36] and GoogLeNet [37], can be employed as the shared CNN.
- *The HCP outputs are intrinsically multi-label prediction results*. HCP produces a normalized probability distribution over the labels after the softmax layer, and the predicted probability values are intrinsically the final classification confidence for the corresponding categories.

2 Related Work

Deep learning tries to model the high-level abstractions of visual data by using architectures composed of multiple non-linear transformations. Specifically, deep convolutional neural network (CNN) [25] has demonstrated an extraordinary ability for image classification [20], [21], [23], [26], [27], [28], [37] on single-label datasets (e.g., ImageNet [10]) and event detection [42].

More recently, CNN architectures have been adopted to address multi-label problems. Gong et al. [16] studied and compared several multi-label loss functions for the multi-label annotation problem based on a network structure similar to [23]. However, due to the large number of parameters to be learned for CNN, an effective model requires lots of training samples. Therefore, training a task-specific convolutional neural network is not applicable on datasets with limited numbers of training samples.

Some recent works [6], [11], [15], [17], [30], [33], [34], [36] have demonstrated that CNN models pre-trained on large datasets with data diversity, e.g., ImageNet, can be transferred to extract CNN features for other image datasets without enough training data. Pierre et al. [34] and Razavian et al. [33] proposed a CNN feature-SVM pipeline for multi-label classification. Specifically, global images from a multi-label dataset are directly fed into the CNN which is pre-trained on ImageNet, to get CNN activations as the off-the-shelf features for classification. Chattield et al. [6] explored the effect of CNN representations based on different CNN architectures for multi-label classification task. Simonyan et al. [36] extracted and aggregated image descriptors over a wide range of scales based on two Very Deep Convolutional Networks, which achieved the state-of-the-art performance on the Pascal VOC datasets with SVM classifier.

Besides, Oquab et al. [30] and Girshick et al. [15] presented two proposal-based methods for multi-label classification and detection. Although considerable improvements have been made by these two approaches, they highly depend on the ground-truth bounding boxes, which may limit their generalization ability when transferred to a new multi-label dataset without any bounding box information. Specifically, all hypotheses with $\geq$0.5 IoU overlap with a ground-truth box are treated as positives for the category of that box and the rest are treated as negatives in [15]. These labeled hypotheses are then utilized to fine-tune the pre-trained CNN. In contrast, the proposed HCP infrastructure requires no ground-truth bounding box information for training and is robust to the possibly noisy and/or redundant hypotheses. Different from [15], [30], no explicit hypothesis label is required during the training process. Instead, a novel hypothesis selection method is proposed to select a small number of high-quality hypotheses for training.

3 Hypotheses-CNN-Pooling

Fig. 2 shows the architecture of the proposed Hypotheses-CNN-Pooling (HCP) deep network. We apply the objectness detection

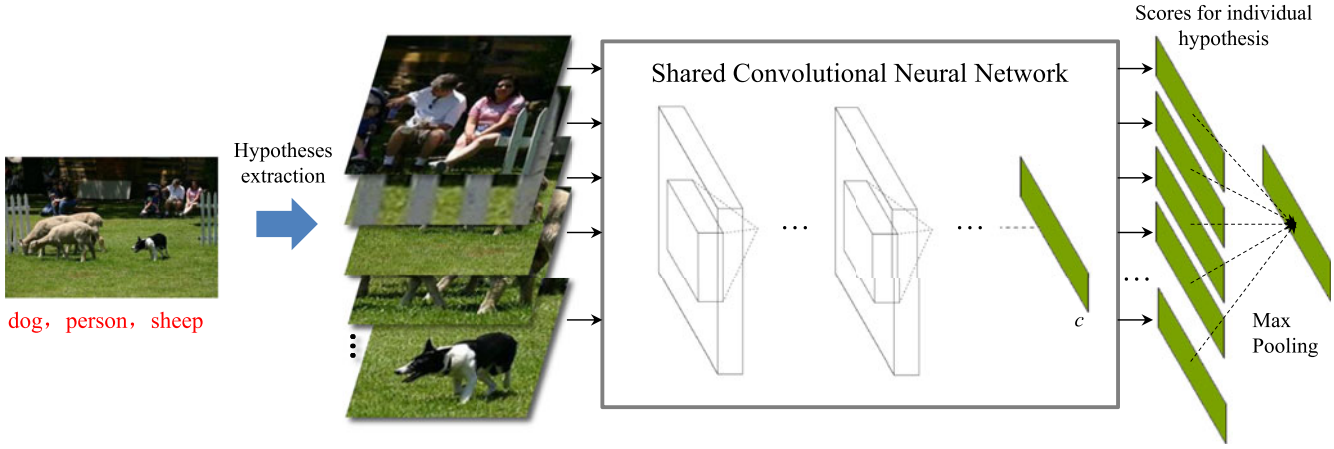

Fig. 2. An illustration of the infrastructure of the proposed HCP. For a given multi-label image, a set of input hypotheses to the shared CNN is selected based on the proposals generated by the state-of-the-art objectness detection techniques, e.g., BING [9] or EdgeBoxes [44]. We feed the selected hypotheses into the shared CNN and fuse the outputs into a c-dimensional prediction vector with cross-hypothesis max-pooling operation, where c is the category number of the target multi-label dataset. The shared CNN is firstly pre-trained on the single-label image dataset, e.g., ImageNet and then fine-tuned with the multi-label images based on the squared loss function. Finally, we retrain the whole HCP to further fine-tune the parameters for multi-label image classification.

technique, e.g., BING [9] or EdgeBoxes [44], to produce a set of candidate object windows. A much smaller number of candidate windows are then selected as hypotheses by the proposed hypotheses selection method. The selected hypotheses are fed into a shared convolutional neural network (CNN). The confidence vectors from the input hypotheses are combined through a fusion layer with max pooling operation, to generate the ultimate multi-label predictions. In specific, the shared CNN is first pre-trained on a large-scale single-label image dataset, i.e., ImageNet and then fine-tuned on the target multi-label dataset, e.g., Pascal VOC, by using the entire image as the input. After that, we retrain the proposed HCP with squared loss function for the final prediction.

3.1 Hypotheses Extraction

HCP takes an arbitrary number of object segment hypotheses as the inputs to the shared CNN and fuses the prediction of each hypothesis with the max pooling operation to get the ultimate multi-label predictions. Therefore, the performance of the proposed HCP largely depends on the quality of the extracted hypotheses. Nevertheless, designing an effective hypotheses extraction approach is challenging, which should satisfy the following criteria:

High object detection recall rate. The proposed HCP is based on the assumption that the input hypotheses can cover all single objects of the given multi-label image, which requires a high detection recall rate.

Small number of hypotheses. Since all hypotheses of a given multi-label image need to be fed into the shared CNN simultaneously, more hypotheses requires more computational resources (e.g., RAM and GPU). Thus a small hypothesis number is preferred for an effective hypotheses extraction approach.

High computational efficiency. As the first step of the proposed HCP, the efficiency of hypotheses extraction will significantly influence the performance of the whole framework. With high efficiency, HCP can be easily integrated into real-time applications.

In summary, a good hypothesis generating algorithm should generate as few hypotheses as possible in an efficient way and meanwhile achieve as high recall rate as possible.

During the past few years, many objectness proposal (hypothesis) methods [1], [2], [4], [9], [38], [44] have been proposed to generate a set of hypotheses to cover all independent objects in a given image. We experimentally adopt two proposal methods, i.e., BING [9] and EdgeBoxes [44], for hypotheses generation due to their high computational efficiency and high object detection recall rate. Although the number of hypotheses generated by BING or EdgeBoxes is very small compared with a common sliding window paradigm, it is still very large for HCP training. To address this problem, we propose a hypotheses selection (HS) method to select hypotheses from the generated proposals. Denote the generated hypothesis bounding boxes for a given image as $H = \{h_1, h_2, \ldots, h_n\}$, where n is the hypothesis number. An $n \times n$ affinity matrix W is constructed, where W_{ij} $(i, j <= n)$ is the IoU scores between h_i and h_j, which can be defined as

$$W_{ij} = \frac{|h_i \cap h_j|}{|h_i \cup h_j|}, \tag{1}$$

where $|\cdot|$ is used to measure the number of pixels. The normalized cut algorithm [35] is then adopted to group the hypothesis bounding boxes into m clusters. As shown in Fig. 3b, different colors indicate different clusters. We empirically filter out those hypotheses with small areas (<900 pixels) or with high height/width (or width/height) ratios (>4), as those shown in Fig. 3c with red bounding boxes. For each cluster, we pick out the top 1 hypothesis with the highest predictive score generated by BING or EdgeBoxes, and resize it into a square shape. As a result, m hypotheses, which are much fewer than those directly generated by proposal methods, will be selected as the inputs of HCP for each image.

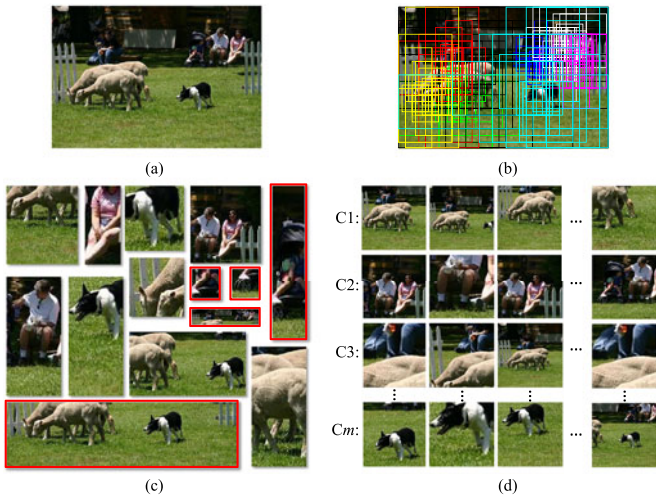

Fig. 3. (a) Source image. (b) Hypothesis bounding boxes generated by BING. Different colors indicate different clusters, which are produced by normalized cut. (c) Hypotheses directly generated by the bounding boxes. (d) Hypotheses generated by the proposed HS method.

3.2 Training HCP

In the proposed HCP, any state-of-the-art CNN model [23], [28], [36], [37] can be employed as the shared CNN. Take Alex Net [23] as an example, which contains five convolutional layers and three fully-connected layers with 60 million parameters. Without enough training images, it is very difficult to obtain an effective HCP model for multi-label classification. However, to collect and annotate a large-scale multi-label dataset is generally difficult. Fortunately, a large-scale single-label image dataset, i.e., ImageNet, can be used to pre-train the shared CNN for parameter initialization, since each multi-label image is firstly cropped into many hypotheses and each hypothesis is assumed to contain at most one object based on the architecture of HCP.

The initialization process of HCP mainly includes two steps. First, the shared CNN is initialized with the parameters pre-trained on ImageNet. Second, the final fully-connected layer of the network (which is trained for 1000-way ImageNet classification) is replaced with a c-way fully-connected layer, where c is the category number of the target multi-label dataset, and an image-fine-tuning (I-FT) process is adopted to initialize the final fully-connected layer by utilizing the target multi-label image set as inputs.

After the initialization, hypotheses-fine-tuning (H-FT) is carried out based on the proposed HCP framework. Specifically, all the m hypotheses as elaborated in Section 3.1 for the training image are fed into the shared CNN. To suppress the possibly noisy hypotheses, a cross-hypothesis max-pooling is carried out to fuse the outputs into one integrative prediction. Suppose $\boldsymbol{v}_i (i = 1, \ldots, m)$ is the output vector of the i^{th} hypothesis from the shared CNN and $\boldsymbol{v}_i^{(j)} (j = 1, \ldots, c)$ is the j^{th} component of $\boldsymbol{v}_i$. The cross-hypothesis max-pooling in the fusion layer can be formulated as

$$\boldsymbol{v}^{(j)} = \max\left(\boldsymbol{v}_1^{(j)}, \boldsymbol{v}_2^{(j)}, \ldots, \boldsymbol{v}_m^{(j)}\right), \tag{2}$$

where $\boldsymbol{v}^{(j)}$ can be considered as the predicted value for the j^{th} category of the given image.

It should be noted that I-FT is an important step for HCP training. The reason is that, for each ground truth label, one instance shall be selected to represent this class after cross-hypothesis max-pooling operation. Without reasonable parameters for the last fully-connected layer, the initial link may be incorrect, which may cause the CNN model stuck at a local optimum. In addition, the cross-hypothesis max-pooling is a crucial step for the robustness of the whole HCP framework to the noise. If one hypothesis contains an object,

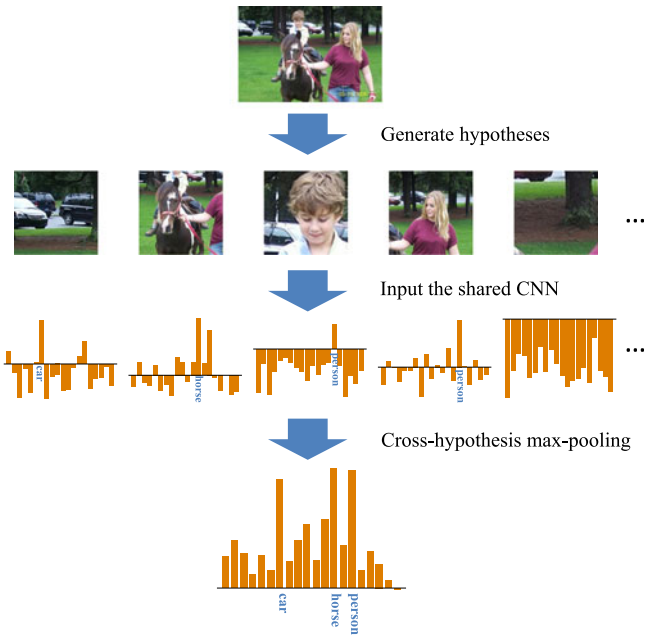

Fig. 4. An illustration of the proposed HCP for a VOC 2007 test image. The second row indicates the generated hypotheses. The third row indicates the predicted results for the input hypotheses. The last row is the predicted result for the test image after cross-hypothesis max-pooling.

the output vector will have a high response (i.e., large value) on the j^{th} component, meaning a high confidence for the corresponding j^{th} category. With cross-hypothesis max-pooling, large predicted values corresponding to objects of interest will be preserved, while the values from the noisy hypotheses will be suppressed.

For both I-FT and H-FT, we experimentally utilize the squared loss as the loss function. Suppose there are N images in the multi-label image set, and $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{ic}]$ is the label vector of the i^{th} image. $y_{ij} = 1$ $(j = 1, \ldots, c)$ if the image is annotated with class j, and otherwise $y_{ij} = 0$. The ground-truth probability vector of the i^{th} image is defined as $\hat{\boldsymbol{p}}_i = \boldsymbol{y}_i / ||\boldsymbol{y}_i||_1$ and the predictive probability vector is $\boldsymbol{p_i} = [p_{i1}, p_{i2}, \ldots, p_{ic}]$. And then the cost function to be minimized is defined as

$$J = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{c} (p_{ik} - \hat{p}_{ik})^2. \tag{3}$$

For I-FT, the last fully-connected layer is randomly initialized with a Gaussian distribution $G(\mu, \sigma)(\mu = 0, \sigma = 0.01)$. The learning rates of the last fully-connected layer and other layers are experimentally initialized as 0.01 and 0.001, respectively. For H-FT, the learning rates of the last fully-connected layer and other layers are set as 0.001 and 0.0001. We carry out 30 training epoches for both I-FT and H-FT, and decrease the learning rates to one tenth of the current ones after every 10 epoches. The momentum and the weight decay are set as 0.9 and 0.0005.

3.3 Multi-label Classification for Test Image

Based on the trained HCP model, the multi-label classification of a given image can be summarized as follows. We first generate the input hypotheses of an given image based on the hypothesis extraction method. Then, for each hypothesis, a c-dimensional predictive result can be obtained by the shared CNN. Finally, we utilize the cross-hypothesis max-pooling operation accompanied with softmax to produce the final prediction. As shown in Fig. 4, the second row and the third row indicate the generated hypotheses and the corresponding outputs from the shared CNN. For each object hypothesis, there is a high response on the corresponding category (e.g., for the first hypothesis, the response on *car* is very high). After cross-hypothesis max-pooling operation, as indicated by the last row in Fig. 4, the high responses (i.e., *car*, *horse* and *person*), which can be considered as the predicted labels, are preserved.

TABLE 1
The Improvements of Hypothesis Fine-Tuning Based on Two Kinds of Shared CNNs (mAP in %).

Shared CNN		Alex Net	VGG Net
VOC 2007	I-FT	74.4	84.5
	HCP	82.7	90.9
	Improvement	8.3	6.5
VOC 2012	I-FT	74.7	84.8
	HCP	81.8	90.5
	Improvement	7.1	5.7

4 Experimental Results

4.1 Datasets and Settings

We evaluate the proposed HCP on the PASCAL Visual Object Classes Challenge (VOC) datasets [13], which are widely used as the benchmarks for multi-label classification. In this paper, PASCAL VOC 2007 and VOC 2012 are employed for experiments. These two datasets, which contain 9,963 and 22,531 images respectively, are divided into *train*, *val* and *test* subsets. We conduct our experiments on the *trainval*/*test* splits (5,011/4,952 for VOC 2007 and 11,540/10,991 for VOC 2012). The evaluation metric is *Average Precision* (AP) and mean of AP (mAP), complying with the PASCAL challenge protocols. We experimentally validate the proposed method based on two CNN models, i.e., Alex Net [23] and VGG Net (16 layers) [36]. We directly apply the parameters pre-trained by Jia et al. [22] and Simonyan et al. [36] with 1,000 ImageNet classes to initialize the CNN models. For hypothesis-fine-tuning, the number of bounding box clusters m is set as 15. Detailed justifications of model components are provided in the supplementary material. All experiments are conducted on one NVIDIA GTX Titan GPU with 6GB memory and all our training algorithms are based on the code provided by Jia et al. [22].

4.2 Image Classification Results

Comparison with I-FT. Table 1 shows the details of improvement from I-FT to HCP. It can be observed that, based on the proposed HCP framework, the classification performance can be further improved by at least 5.7%. The results of I-FT and HCP are based on using single center crop and 500 Edgeboxes hypotheses for testing, respectively. Fig. 5 shows an example of the testing results based on different models. It can be seen that there are three ground-truth categories in the given image, i.e., *car*, *horse*, *person*. It should be noted that the *car* category is not detected during image-fine-tuning while it is successfully recovered in HCP. This may be because the proposed HCP is a hypotheses based method and both foreground (i.e., *horse*, *person*) and background (i.e., *car*) objects can be equivalently treated. However, during the I-FT stage, the entire image is treated as the input, which may lead to ignorance of some background objects. We also test the I-FT model by using 500 hypotheses, but the improvement is very limited. Please refer to the supplementary material for more details .

Comparisons of using different number of hypotheses for testing. Table 2 shows the testing results by varying the number (from 50 to 500) of hypotheses during the testing stage on VOC 2007. We compare BING [9][1] with Edgeboxes [44] based on Alex Net and

1. Our method is independent of the ground-truth bounding box. Therefore, to train the object detector of BING, the detection dataset of ILSVRC 2013 is used as augmented data. We removed those images which are semantically overlapping with PASCAL VOC categories and randomly selected 13,894 images for training.

TABLE 2
Classification Results (mAP in %) by Varying the Number of Hypotheses During the Testing Stage on VOC 2007

Number	Alex Net		VGG Net	
	BING	EdgeBoxes	BING	EdgeBoxes
T-50	81.0	81.1	89.9	89.9
T-100	81.6	81.7	90.3	90.3
T-150	81.9	82.0	90.4	90.6
T-200	82.1	82.3	90.5	90.7
T-250	82.1	82.4	90.5	90.8
T-300	82.1	82.4	90.6	90.8
T-400	82.1	82.6	90.6	90.8
T-450	82.1	82.6	90.6	90.9
T-500	82.2	82.7	90.6	90.9

VGG Net. It can be observed that EdgeBoxes performs slightly better than BING. In addition, along with the decreasing of the hypothesis number, the performance of both proposal generators is very stable (from 500 to 50, only 1%∼1.6% drop). Therefore, even with a little number of hypotheses, our method can still achieve satisfactory performance. Specifically, with top-50 hypotheses, the performance is 89.9% based on VGG Net. This result still outperforms [36] (i.e., 89.3%) and the testing for one image can be done within 2s.

Comparison with the state-of-the-art methods. Table 3 and Table 4 report our experimental results compared with the state-of-the-arts on VOC 2007 and VOC 2012, respectively. The upper and the bottom parts of Table 3 and Table 4 show the results produced by single model and combined models, respectively. Besides, the methods marked with * are those using additional images, i.e., ImageNet, for training. All our results are obtained by utilizing top-500 hypotheses of each testing image generated by Edgeboxes

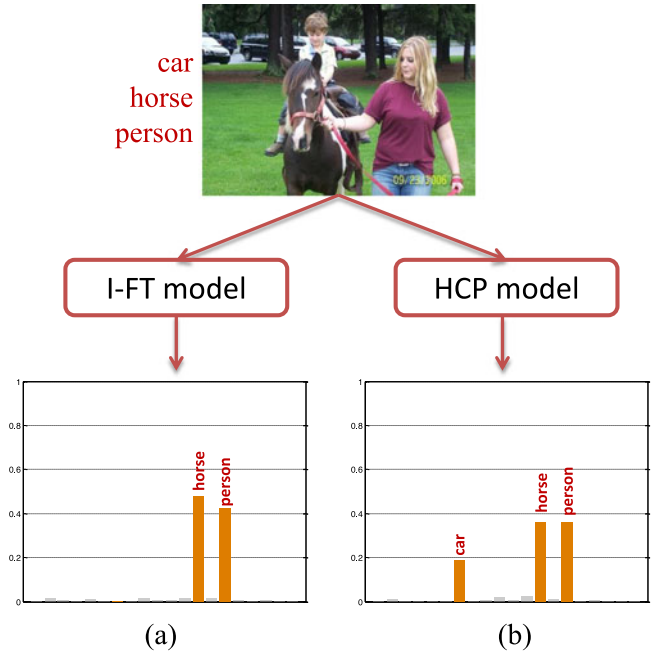

Fig. 5. (a) The predicted result based on I-FT model. (b) The predicted result based on HCP model.

TABLE 3
Classification Results (AP in %) Comparison for State-of-the-Art Approaches on VOC 2007 Test

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Single Model:																					
INRIA[19]	77.2	69.3	56.2	66.6	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
AGS[12]	82.2	83.0	58.4	76.1	56.4	77.5	88.8	69.1	62.2	61.8	64.2	51.3	85.4	80.2	91.1	48.1	61.7	67.7	86.3	70.9	71.1
AMM[7]	84.5	81.5	65.0	71.4	52.2	76.2	87.2	68.5	63.8	55.8	65.8	55.6	84.8	77.0	91.1	55.2	60.0	69.7	83.6	77.0	71.3
Razavian et al.* [34]	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
PRE-1000C* [31]	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
Chatfield et al.* [6]	95.3	90.4	92.5	89.6	54.4	81.9	91.5	91.9	64.1	76.3	74.9	89.7	92.2	86.9	95.2	60.7	82.9	68.0	95.5	74.4	82.4
SPP* [20]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	82.4
VGG-16-SVM* [37]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.3
VGG-19-SVM* [37]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.3
HCP-Alex*	95.4	90.7	92.9	88.9	53.9	81.9	91.8	92.6	60.3	79.3	73.0	90.8	89.2	86.4	92.5	66.9	86.4	65.6	94.4	80.4	82.7
HCP-VGG*	**98.6**	**97.1**	**98.0**	**95.6**	**75.3**	**94.7**	**95.8**	**97.3**	**73.1**	**90.2**	**80.0**	**97.3**	**96.1**	**94.9**	**96.3**	**78.3**	**94.7**	**76.2**	**97.9**	**91.5**	**90.9**
Combined Models:																					
VGG-16-19-SVM* [37]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.7

TABLE 4
Classification Results (AP in %) Comparison for State-of-the-Art Approaches on VOC 2012 Test

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Single Model:																					
NUS-PSL[44]	97.3	84.2	80.8	85.3	60.8	89.9	86.8	89.3	75.4	77.8	75.1	83.0	87.5	90.1	95.0	57.8	79.2	73.4	94.5	80.7	82.2
Zeiler et al.* [44]	96.0	77.1	88.4	85.5	55.8	85.8	78.6	91.2	65.0	74.4	67.7	87.8	86.0	85.1	90.9	52.2	83.6	61.1	91.8	76.1	79.0
PRE-1000C* [31]	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-1512* [31]	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	96.1	64.3	86.6	62.3	91.1	79.8	82.8
Chatfield et al.* [6]	96.8	82.5	91.5	88.1	62.1	88.3	81.9	94.8	70.3	80.2	76.2	92.9	90.3	89.3	95.2	57.4	83.6	66.4	93.5	81.9	83.2
Oquab et al.* [32]	96.7	88.8	92.0	87.4	64.7	91.1	87.4	94.4	74.9	89.2	76.3	93.7	95.2	91.1	**97.6**	66.2	91.2	70.0	94.5	83.7	86.3
VGG-16-SVM* [37]	99.0	88.8	95.9	93.8	73.1	92.1	85.1	97.8	79.5	91.1	83.3	97.2	96.3	94.5	96.9	63.1	93.4	75.0	97.1	87.1	89.0
VGG-19-SVM* [37]	99.1	88.7	95.7	93.9	73.1	92.1	84.8	97.7	79.1	90.7	83.2	97.3	96.2	94.3	96.9	63.4	93.2	74.6	97.3	87.9	89.0
HCP-Alex*	97.7	83.2	92.8	88.5	60.1	88.7	82.7	94.4	65.8	81.9	68.0	92.6	89.1	87.6	92.1	58.0	86.6	55.5	92.5	77.6	81.8
HCP-VGG*	99.1	92.8	97.4	94.4	79.9	93.6	89.8	98.2	78.2	94.9	79.8	97.8	97.0	93.8	96.4	74.3	**94.7**	71.9	96.7	88.6	90.5
Combined Models:																					
VGG-16-19-SVM* [37]	99.1	89.1	96.0	94.1	74.1	92.2	85.3	97.9	79.9	92.0	83.7	97.5	96.5	94.7	97.1	63.7	93.6	75.2	97.4	87.8	89.3
HCP-VGG+[44]*	**99.8**	**94.8**	**97.7**	**95.4**	**81.3**	**96.0**	**94.5**	**98.9**	**88.5**	**94.1**	**86.0**	**98.1**	**98.3**	**97.3**	97.3	**76.1**	93.9	**84.2**	**98.2**	**92.7**	**93.2**

as inputs. The testing time of Alex Net and VGG Net is about 3s/image and 10s/image on a GPU including proposal generation (EdgeBoxes: 0.25s/image).

From the experimental results, we can see that the performance from single HCP-VGG model is better than all previous methods. Specifically, in [36], the pre-trained VGG models are firstly applied to extract visual features over a wide range of image scales ($Q \in 256, 384, 512, 640, 768$), and then are aggregated (5 scales and 50 cropped patches for each scale) by averaging to generate the final image representations, which achieves the state-of-the-art performance with the SVM classifier. As can be seen from Table 3 and Table 4, our single model results outperform [36], on both model architectures as well as their combined models. Indicated by Table 2, our single model can reach 90.8% by using the same number (i.e., 250) of hypotheses in testing, which is 1.5% increase over the single model of [36]. More detailed comparative analyses are provided in the supplementary material.

On VOC 2012, some new state-of-the-art results are achieved by `MVMI-DSP` and `Tencent-BestImage` on the public leaderboard[2], whose results are 90.7% and 90.4%, respectively. However, as illustrated by their descriptions, both results are obtained through some combination. To make further improvement, a late fusion between the predicted scores of HCP-VGG and our previous model NUS-PSL [12] (which obtained the winner prize of the classification task in PASCAL VOC 2012) is executed. Incredibly, the mAP score produced by the combination of these two models can surge to 93.2%, which outperforms all other methods.

5 Conclusions

In this paper, we presented a novel Hypotheses-CNN-Pooling (HCP) framework to address the multi-label image classification problem. Based on the proposed HCP, CNN pre-trained on large-scale single-label image datasets, e.g., ImageNet, can be successfully transferred to tackle the multi-label problem. In addition, the proposed HCP requires no bounding box annotation for training, and thus can easily adapt to new multi-label datasets. We evaluated our method on VOC 2007 and VOC 2012, and verified that significant improvement can be made by HCP compared with the state-of-the-arts. Furthermore, it is proved that late fusion between outputs of CNN and hand-crafted feature schemes can incredibly enhance the classification performance.

Acknowledgments

This work is supported in part by National Basic Research Program of China (No.2012CB316400), Fundamental Scientific Research Project (No.K15JB00360), National NSF of China (61210006, 61532005).

References

[1] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.
[2] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 328–335.
[3] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[4] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1312–1328, Jul. 2012.
[5] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.
[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.
[7] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 13–27, Jan. 1, 2015.
[8] Q. Chen, Z. Song, Y. Hua, Z. Huang, and S. Yan, "Hierarchical matching with side information for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 3426–3433.
[9] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 3286–3293.
[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.
[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.
[12] J. Dong, W. Xia, Q. Chen, J. Feng, Z. Huang, and S. Yan, "Subcategory-aware object classification," in *Computer Vis. Pattern Recog.*, 2013, pp. 827–834.
[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[14] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Comput. Vis. Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.
[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *arXiv preprint arXiv:1311.2524*, 2013.
[16] Y. Gong, Y. Jia, T. K. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multi label image annotation," *arXiv preprint arXiv:1312.4894*, 2013.
[17] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," *arXiv preprint arXiv:1403.1840*, 2014.
[18] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.
[19] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 237–244.
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 346–361.
[21] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 2146–2153.
[22] Y. Jia. (2013). Caffe: An open source convolutional architecture for fast feature Embedding [Online]. Available: http://caffe.berkeleyvision.org/
[23] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
[24] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2006, vol. 2, pp. 2169–2178.
[25] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems*. San Mateo, CA, USA: Morgan Kaufmann, 1990.
[26] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2004, vol. 2, pp. II-97–II-104.
[27] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 609–616.
[28] M. Lin, Q. Chen, and S. Yan, "NetÂ-work in netÂ-work," *arXiv preprint arXiv:1312.4400*, 2013.
[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[30] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1717–1724.
[31] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Weakly supervised object recognition with convolutional neural networks," INRIA, Le Chesnay, France, Tech. Rep. HAL-01015140, 2014.
[32] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.
[33] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," *arXiv preprint arXiv:1403.6382*, 2014.
[34] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.
[35] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

2. http://host.robots.ox.ac.uk:8080/leaderboard/

[38] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.
[39] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3360–3367.
[40] W. Xia, C. Domokos, L. F. Cheong, and S. Yan, "Background context augmented hypothesis graph for object segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 4, pp. 582–594, Sept. 2015.
[41] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, and S. Yan, "Semantic segmentation without annotating segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, Australia, Dec. 2013, pp. 2176–2183.
[42] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1798–1807.
[43] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[44] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.