CNN: Single-label to Multi-label

Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, *Senior Member, IEEE*
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) no explicit hypothesis label is required; 4) the shared CNN may be well pre-trained with a large-scale single-label image dataset, e.g. ImageNet; and 5) it may naturally output multi-label prediction results. Experimental results on Pascal VOC2007 and VOC2012 multi-label image datasets well demonstrate the superiority of the proposed HCP infrastructure over other state-of-the-arts. In particular, the mAP reaches 84.2% by HCP only and 90.3% after the fusion with our complementary result in [47] based on hand-crafted features on the VOC2012 dataset, which significantly outperforms the state-of-the-arts with a large margin of more than 7%.

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

arXiv:1406.5726v3 [cs.CV] 9 Jul 2014

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 [14], [18], [10]. For image representation and classification, conventional approaches utilize carefully designed hand-crafted features, e.g., SIFT [32], along with the bag-of-words coding scheme, followed by the feature pooling [25], [44], [37] and classic classifiers, such as Support Vector Machine (SVM) [4] and random forests [2]. 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 [26], [21], [24], [36]. Among these architectures, one of the greatest breakthroughs in image classification is the deep convolutional neural network (CNN) [24], 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.

Multi-label image classification is however a more general and practical problem, since the majority of real-world images are with more than one objects of different categories. Many methods [37], [6], [12] have been proposed to address this more challenging problem. The success of CNN on single-label image classification also sheds some light on the multi-label image classification problem. However, the CNN model cannot be trivially extended to cope with the multi-label image classification problem in an interpretable manner, mainly due to the following reasons. Firstly, the implicit assumption that foreground objects are roughly aligned, which is usually true for single-label images, does not always hold for multi-label images. Such alignment facilitates the design of the convolution and pooling infrastructure of CNN for single-label image classification. However, for a typical multi-label image, different categories of objects are located at various positions with different scales and poses. For example, as shown in Figure 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. Secondly, the interaction between different objects in multi-label images, like partial visibility and occlusion, also poses a great challenge. Therefore, directly applying the original CNN structure for multi-label image classification is not feasible. Thirdly, due to the tremendous parameters to be learned for CNN, a large number of training images are required for the model training. Furthermore, from single-label to multi-label (with n category labels) image classification, the label space has been expanded from n to 2^n, thus more training

Yunchao Wei is with Department of Electrical and Computer Engineering, National University of Singapore, and also with the Institute of Information Science, Beijing Jiaotong University, e-mail: wychao1987@gmail.com.
Yao Zhao is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China.
Bingbing Ni is with the Advanced Digital Sciences Center, Singapore.
Wei Xia, Junshi Huang, Jian Dong and Shuicheng Yan are with Department of Electrical and Computer Engineering, National University of Singapore.

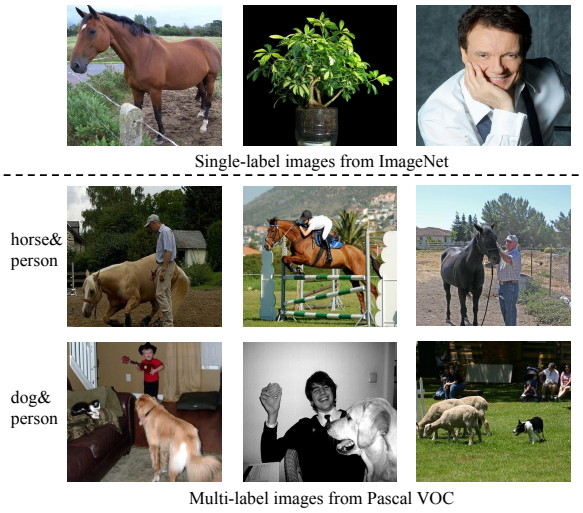

Fig. 1. Some examples from ImageNet [10] and Pascal VOC 2007 [13]. The foreground objects in single-label images are usually roughly aligned. However, the assumption of object alignment is not valid for multi-label images. Also note the partial visibility and occlusion between objects in the multi-label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

data is required to cover the whole label space. For single-label images, it is practically easy to collect and annotate the images. However, the burden of collection and annotation for a large scale multi-label image dataset is generally extremely high.

To address these issues and take full advantage of CNN for multi-label image classification, in this paper, we propose a flexible deep CNN structure, called Hypotheses-CNN-Pooling (HCP). HCP takes an arbitrary number of object segment hypotheses as the inputs, which may be generated by the sate-of-the-art objectiveness detection techniques, e.g., binarized normed gradients (BING) [8], and then a shared CNN is connected with each hypothesis. Finally the CNN output results from different hypotheses are aggregated by max pooling 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 [12], [5], [15], [35], 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 the 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 reserved and the noisy ones are ignored. 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.

- *No explicit hypothesis label is required for training.* The state-of-the-art CNN models [15], [35] utilize the hypothesis label for training. They first compute the Intersection-over-Union (IoU) overlap between hypotheses and ground-truth bounding boxes, and then assign the hypothesis with the label of the ground-truth bounding box if their overlap is above a threshold. In contrast, the proposed HCP takes an arbitrary number of hypotheses as the inputs without any explicit hypothesis labels.

- *The shared CNN 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.

- *The HCP outputs are intrinsically multi-label prediction results.* HCP produces a normalized probability distribution over the labels after the softmax layer, and the the predicted probability values are intrinsically the final classification confidence values for the corresponding categories.

Extensive experiments on two challenging multi-label image datasets, Pascal VOC 2007 and VOC 2012, well demonstrate the superiority of the proposed HCP infrastructure over other state-of-the-arts. The rest of the paper is organized as follows. We briefly review the related work of multi-label classification in Section 2. Section 3 presents the details of the HCP for image classification. Finally the experimental results and conclusions are provided in Section 4 and Section 5, respectively.

2 Related Work

During the past few years, many works on various multi-label image classification models have been conducted. These models are generally based on two types of frameworks: bag-of-words (BoW) [19], [37], [6], [12], [5] and deep learning [35], [16], [38].

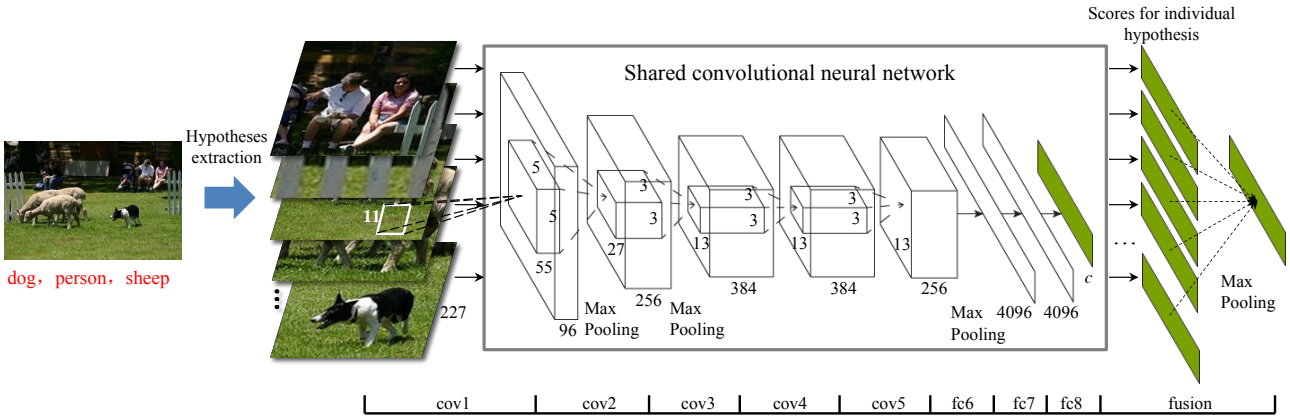

Fig. 2. An illustration of the infrastructure of the proposed HCP. For a given multi-label image, a set of input hypotheses to the shared CNN is selected based on the proposals generated by the state-of-the-art objectness detection techniques, e.g., BING [8]. The shared CNN has a similar network structure to [24] except for the layer fc8, where c is the category number of the target multi-label dataset. 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. 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.

2.1 Bag-of-Words Based Models

A traditional BoW model is composed of multiple modules, e.g., feature representation, classification and context modelling. For feature representation, the main components include hand-crafted feature extraction, feature coding and feature pooling, which generate global representations for images. Specifically, hand-crafted features, such as SIFT [32], Histogram of Oriented Gradients [9] and Local Binary Patterns [34] are firstly extracted on dense grids or sparse interest points and then quantized by different coding schemes, e.g., Vector Quantization [33], Sparse Coding [45] and Gaussian Mixture Models [20]. These encoded features are finally pooled by feature aggregation methods, such as Spatial Pyramid Matching (SPM) [25], to form the image-level representation. For classification, conventional models, such as SVM [4] and random forests [2], are utilized. Beyond conventional modelling methods, many recent works [19], [42], [39], [6], [5] have demonstrated that the usage of context information, e.g., spatial location of object and background scene from the global view, can considerably improve the performance of multi-label classification and object detection.

Although these works have made great progress in visual recognition tasks, the involved hand-crafted features are not always optimal for particular tasks. Recently, in contrast to hand-crafted features, *learnt* features with deep learning structures have shown great potential for various vision recognition tasks, which will be introduced in the following subsection.

2.2 Deep Learning Based Models

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) [26] has demonstrated an extraordinary ability for image classification [21], [27], [29], [24], [30] on single-label datasets such as CIFAR-10/100 [23] and ImageNet [10].

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 similar network structure to [24]. 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.

Fortunately, some recent works [11], [15], [35], [40], [38], [17] 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.* [40] and Razavian *et al.* [38] 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. However, different from the single-label image, objects in a typical multi-label image are generally less-aligned, and also often with partial visibility and occlusion as shown

in Figure 1. Therefore, global CNN features are not optimal to multi-label problems. Recently, Oquab *et al.* [35] 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, these methods 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.

In contrast, the proposed HCP infrastructure in this paper requires no ground-truth bounding box information for training and is robust to the possibly noisy and/or redundant hypotheses. Different from [35], [15], no explicit hypothesis label is required during the training process. Besides, we propose a hypothesis selection method to select a small number of high-quality hypotheses (10 for each image) for training, which is much less than the number used in [15] (128 for each image), thus the training process is significantly sped up.

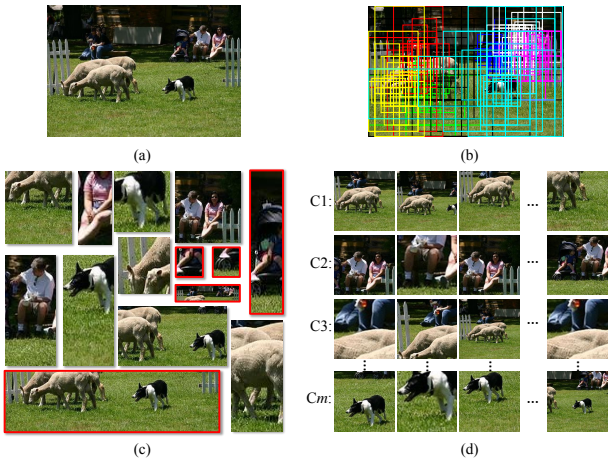

Fig. 3. (a) Source image. (b) Hypothesis bounding boxes generated by BING. Different colors indicate different clusters, which are produced by normalized cut. (c) Hypotheses directly generated by the bounding boxes. (d) Hypotheses generated by the proposed HS method.

3 HYPOTHESES-CNN-POOLING

Figure 2 shows the architecture of the proposed Hypotheses-CNN-Pooling (HCP) deep network. We apply the state-of-the-art objectness detection technique, i.e., BING [8], to produce a set of candidate object windows. A much smaller number of candidate windows are then selected as hypotheses by the proposed hypotheses extraction 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 a 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 cost more computational time and need more powerful hardware (e.g., RAM and GPU). Thus a small hypothesis number is required 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 computational 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 methods [31], [7], [46], [1], [3], [43] have been proposed to tackle the hypotheses detection problem. [31], [7], [46] are based on salient object detection, which try to detect the most attention-grabbing (salient) object in a given image. However, these methods are not applicable to HCP, since saliency based methods are usually applied to a single-label scheme while HCP is a multi-label scheme. [1], [3], [43] are based on objectness proposal (hypothesis), which generate a set of hypotheses to cover all independent objects in a given image. Due to the large number of proposals, such methods are usually quite time-consuming, which will affect the real-time performance of HCP.

Most recently, Cheng *et al.* [8] proposed a surprisingly simple and powerful feature called binarized normed gradients (BING) to find object candidates by using objectness scores. This method is faster (300fps

on a single laptop CPU) than most popular alternatives [1], [3], [43] and has a high object detection recall rate (96.2% with 1,000 hypotheses). Although the number of hypotheses (i.e., 1,000) is very small compared with a common sliding window paradigm, it is still very large for HCP.

To address this problem, we propose a hypotheses selection (HS) method to select hypotheses from the proposals extracted by BING. A set of hypothesis bounding boxes are produced by BING for a given image, denoted by $H = \{h_1, h_2, ..., h_n\}$, 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 b_i and b_j, which can be defined as

$$W_{ij} = \frac{|h_i \cap h_j|}{|h_i \cup h_j|}, \quad (1)$$

where $|\cdot|$ is used to measure the number of pixels. The normalized cut algorithm [41] is then adopted to group the hypothesis bounding boxes into m clusters. As shown in Figure 3(b), different colors indicate different clusters. We empirically filter out those hypotheses with small areas or with high height/width (or width/height) ratios, as those shown in Figure 3(c) with red bounding boxes. For each cluster, we pick out the top k hypotheses with higher predictive scores generated by BING and resize them into square shapes. As a result, mk hypotheses, which are much fewer than those directly generated by BING, will be selected as the inputs of HCP for each image.

3.2 Initialization of HCP

In the proposed HCP, the architecture of the shared CNN is similar to the network described in [24]. The shared CNN contains five convolutional layers and three fully-connected layers with 60 million parameters. Therefore, 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 unaffordable. 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 image of multi-label is firstly cropped into many hypotheses and each hypothesis is assumed to contain at most one object based on the architecture of HCP.

However, directly using the parameters pre-trained by ImageNet to initialize the shared CNN is not appropriate, due to the following reasons. Firstly, both the data amount and the object categories between ImageNet and the target muliti-label dataset are usually different. Secondly, there exist very diverse and complicated interactions among the objects in a multi-label image, which makes multi-label classification more challenging than single-label classification. To better initialize the shared CNN, based on the pre-trained parameters by ImageNet, fine-tuning is enforced to adjust the parameters.

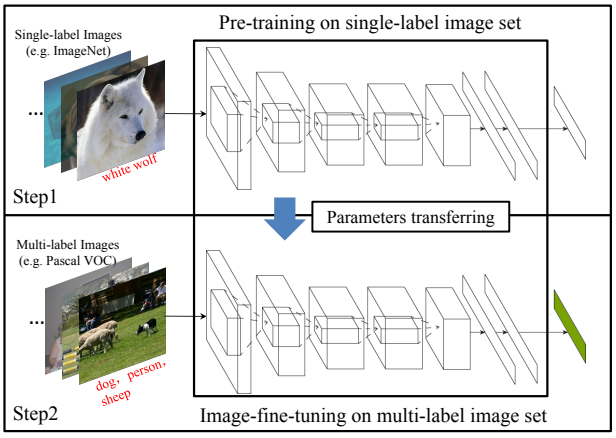

Fig. 4. The initialization of HCP is divided into two steps. The shared CNN is first pre-trained on a single-label image set, e.g., ImageNet and then fine-tuned on the target multi-label image set using the entire image as input. Parameters pre-trained on ImageNet are directly transferred for fine-tuning except for the last fully-connected layer, since the category numbers between these two datasets are different.

As shown in Figure 4, the initialization process of HCP is divided into two steps.

Step1: Pre-training on single-label image set. We use the ImageNet [10] to pre-train the shared CNN. Given an image, we first resize it into 256×256 pixels. Then, we extract random 227×227 patches (and their horizontal reflections) from the given image and train our network based on these extracted patches. Each extracted patch is pre-processed by subtracting the image mean, and fed into the first convolutional layer of the CNN. Indicated by [24], the output of the last fully-connected layer is fed into a 1,000-way softmax layer with the multinomial logistic regression as the loss function, to produce a probability distribution over the 1,000 classes. For all layers, we use the rectified linear units (ReLU) as the nonlinear activation function. We train the network by using stochastic gradient descent with a momentum of 0.9 and weight decay of 0.0005. To overcome overfitting, each of the first two fully-connected layers is followed by a drop-out operation with a drop-out ratio of 0.5. The learning rate is initialized as 0.01 for all layers and reduced to one tenth of the current rate after every 20 epoches (90 epoches in all).

Step2: Image-fine-tuning on multi-label image set. To adapt the pre-trained model on ImageNet to HCP, the entire images from a multi-label image set, e.g., Pascal VOC, are then utilized to further adjust the parameters. The image-fine-tuning (I-FT) process is similar with the pre-training except for several details listed as follows.

Each image is resized into 256×256 pixels without cropping. Since the category number of Pascal VOC

is not equal to that of ImageNet, the output of the last fully-connected layer is fed into a c-way softmax which produces a probability distribution over the c class labels. Different from the pre-training, squared loss is used during I-FT. Suppose there are N images in the multi-label image set, and $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \cdots, y_{ic}]$ is the label vector of the i^{th} image. $y_{ij} = 1$ $(j = 1, \cdots, c)$ 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 $\boldsymbol{p_i} = [p_{i1}, p_{i2}, \cdots, p_{ic}]$. And then the cost function to be minimized is defined as

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

During the I-FT process, as shown in Figure 4, the parameters of the first seven layers are initialized by the parameters pre-trained on ImageNet and the parameters of the last fully-connected layer are randomly initialized with a Gaussian distribution $G(\mu, \sigma)(\mu = 0, \sigma = 0.01)$. The learning rates of the convolutional layers, the first two fully-connected layers and the last fully-connected layer are initialized as 0.001, 0.002 and 0.01 at the beginning, respectively. We executed 60 epoches in total and decreased the learning rate to one tenth of the current rate of each layer after 20 epoches (momentum=0.9, weight decay=0.0005).

By setting the different learning rates for different layers, the updating rates for the parameters from different layers also vary. The first few convolutional layers mainly extract some low-level invariant representations, thus the parameters are quite consistent from the pre-trained dataset to the target dataset, which is achieved by a very low learning rate (i.e., 0.001). Nevertheless, in the final layers of the network, especially the last fully-connected layer, which are specifically adapted to the new target dataset, a much higher learning rate is required to guarantee a fast convergence to the new optimum. Therefore, the parameters can better adapt to the new dataset without clobbering the pre-trained initialization. It should be noted that the I-FT is a critical step of HCP. We tried without this step and found that the performance on VOC 2007 dropped dramatically.

3.3 Hypotheses-fine-tuning

All the $l = mk$ hypotheses are fed into the shared CNN, which has been initialized as elaborated in Section 3.2. For each hypothesis, a c-dimensional vector can be computed as the output of the shared CNN. Indeed, the proposed HCP is based on the assumption that each hypothesis contains at most one object and all the possible objects are covered by some subset of the extracted hypotheses. Therefore, the number of hypotheses should be large enough to cover all possible diversified objects. However, with more hypotheses, noise (hypotheses covering no object) will inevitably increase.

To suppress the possibly noisy hypotheses, a cross-hypothesis max-pooling is carried out to fuse the outputs into one integrative prediction. Suppose $\boldsymbol{v}_i (i = 1, ..., l)$ 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

$$\boldsymbol{v}^{(j)} = \max(\boldsymbol{v}_1^{(j)}, \boldsymbol{v}_2^{(j)}, \ldots, \boldsymbol{v}_l^{(j)}), \tag{3}$$

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

The cross-hypothesis max-pooling is a crucial step for the whole HCP framework to be robust to the noise. If one hypothesis contains an object, 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 reserved, while the values from the noisy hypotheses will be ignored.

During the hypotheses-fine-tuning (H-FT) process, the output of the fusion layer is fed into a c-way softmax layer with the squared loss as the cost function, which is defined as Eq. (2). Similar as I-FT, we also adopt a discriminating learning rate scheme for different layers. Specifically, we execute 60 epoches in total and empirically set the learning rates of the convolutional layers, the first two fully-connected layers and the last fully-connected layer as 0.0001, 0.0002, 0.001 at the beginning, respectively. We decrease the learning rates to one tenth of the current ones after every 20 epoches. The momentum and the weight decay are set as 0.9 and 0.0005, which are the same as in the I-FT step.

3.4 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 firstly generate the input hypotheses of the given image based on the proposed HS 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 to produce the final prediction. As shown in Fig. 5, the second row and the third row indicate the generated hypotheses and the corresponding outputs from the shared CNN. For each object independent 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. 5, the high responses (i.e., *car*, *horse* and *person*), which can be considered as the predicted labels, are reserved.

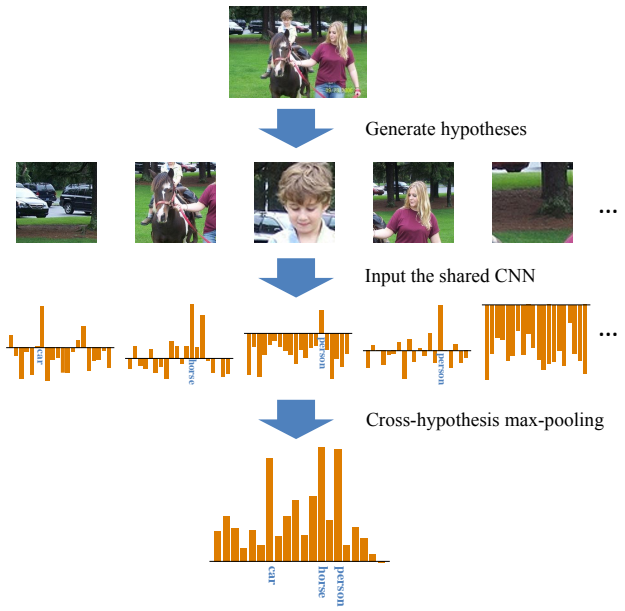

Fig. 5. An illustration of the proposed HCP for a VOC 2007 test image. The second row indicates the generated hypotheses. The third row indicate the predicted results for the input hypotheses. The last row is predicted result for the test image after cross-hypothesis max-pooling operation.

4 Experimental Results

In this section, we present the experiments to validate the effectiveness of our proposed Hypotheses-CNN-Pooling (HCP) framework for multi-label image classification.

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 benchmark 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.

Instead of using two GPUs as in [24], we conduct experiments 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]. The initialization of the shared CNN is based on the parameters pre-trained on the 1,000 classes and 1.2 million images of ILSVRC-2012.

We compare the proposed HCP with the state-of-the-art approaches. Specifically, the competing algorithms are generally divided into two types: those based on hand-crafted features and those based on *learnt* features.

- **INRIA [19]:** Harzallah *et al.* proposed a contextual combination method of localization and classification to improve the performance for both. Specifically, for classification, image representation is built on the traditional feature extraction-coding-pooling pipeline, and object localization is built on sliding-widow approaches. Furthermore, the localization is employed to enhance the classification performance.

- **FV [37]:** The Fisher Vector representation of images can be considered as an extension of the bag-of-words. Some well-motivated strategies, e.g., L2 normalization, power normalization and spatial pyramids, are adopted over the original Fisher Vector to boost the classification accuracy.

- **NUS:** In [5], Chen *et al.* presented an Ambiguity-guided Mixture Model (AMM) to seamlessly integrate external context features and object features for general classification, and then the contextualized SVM was further utilized to iteratively and mutually boost the performance of object classification and detection tasks. Dong *et al.* [12] proposed an Ambiguity Guided Subcategory (AGS) mining approach, which can be seamlessly integrated into an effective subcategory-aware object classification framework, to improve both detection and classification performance. The combined version of the above two, NUS-PSL [47] received the winner prizes of the classification task in PASCAL VOC 2010-2012.

- **CNN-SVM [38]:** OverFeat [40], which obtained very competitive performance in the image classification task of ILSVRC 2013, was released by Sermanet *et al.* as a feature extractor. Razavian *et al.* [38] employed OverFeat, which is pre-trained on ImageNet, to get CNN activations as the off-the-shelf features. The state-of-the-art classification result on PASCAL VOC 2007 was achieved by using linear SVM classifiers over the 4,096 dimensional feature representation extracted from the 22^{nd} layer of OverFeat.

- **I-FT:** The structure of the shared CNN follows that of Krizhevsky *et al.* [24]. The shared CNN was first pre-trained on ImageNet, and then the last fully-connected layer was modified into 4096×20, and the shared CNN was re-trained with squared loss function on PASCAL VOC for multi-label classification.

- **PRE-1000C and PRE-1512 [35]:** Oquab *et al.* pro-

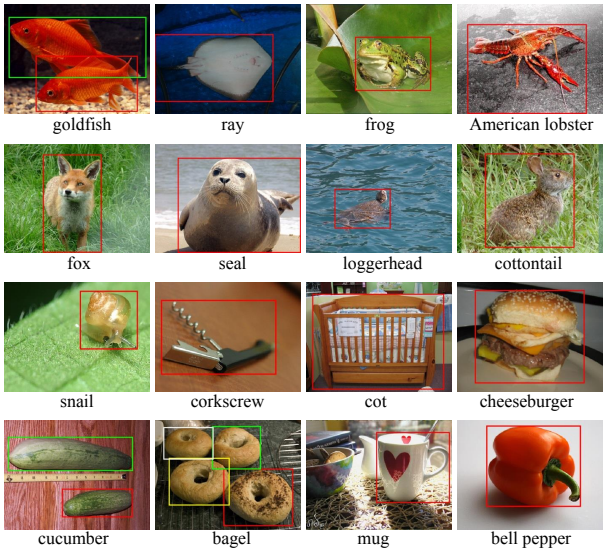

Fig. 6. Exemplar images with ground-truth bounding boxes from the detection dataset of ILSVRC 2013.

posed to transfer image representations learned with CNN on ImageNet to other visual recognition tasks with limited training data. The network has exactly the same architecture as in [24]. Firstly, the network is pre-trained on ImageNet. Then the first seven layers of CNN are fixed with the pre-trained parameters and the last fully-connected layer is replaced by two adaptation layers. Finally, the adaptation layers are trained with images from the target PASCAL VOC dataset. PRE-1000C and PRE-1512 mean the transferred parameters are pre-trained on the original ImageNet dataset with 1000 categories and the augmented one with 1512 categories, respectively. For PRE-1000C, 1.2 million images from ILSVRC-2012 are employed to pre-train the CNN, while for PRE-1512, 512 additional ImageNet classes (e.g., *furniture, motor vehicle, bicycle etc.*) are augmented to increase the semantic overlap with categories in PASCAL VOC. To accommodate the larger number of classes, the dimensions of the first two fully-connected layers are increased from 4,096 to 6,144.

4.2 Hypotheses Extraction

In [8], BING[1] has shown a good generalization ability on the images containing object categories that are not used for training. Specifically, [8] trained a linear SVM using 6 object categories (i.e., the first 6 categories in VOC dataset according to alpha order) and the remaining 14 categories were used for testing. The experimental results in [8] demonstrate that the transferred model almost has the same performance with that using all categories (all the 20 categories in PASCAL VOC) for training.

Since the proposed HCP is independent of the ground-truth bounding box, no object location information can be used for training. Inspired by the generalization ability test in [8], the detection dataset of ILSVRC 2013 is used as augmented data for BING training. It contains 395,909 training images with ground-truth bounding box annotation from 200 categories. To validate the generalization ability of the proposed framework for other multi-label datasets, the categories as well as their subcategories which are semantically overlapping with the PASCAL VOC categoires are removed[2].

For a fair comparison, we follow [8] and only use randomly 13,894 images (instead of all images) from the detection dataset of ILSVRC 2013 for model training. Some selected samples are illustrated in Figure 6, from which we can see that there are big differences between objects in PASCAL VOC and the selected ImageNet samples. After training, the learnt BING-ImageNet model is used to produce hypotheses for VOC 2007 and VOC 2012. We test the object detection rate with 1000 proposals on VOC 2007, which is only 0.3% lower than the reported result (i.e., 96.2%) in [8].

Considering the computational time and the limitation of hardware, we proposed a hypotheses selection (HS) method to filter the input hypotheses produced by BING-ImageNet. As elaborated in Section 3.1, we cluster the extracted proposals into 10 clusters based on their bounding box overlapping information by normalized-cut [41]. The hypotheses are filtered out, which have smaller areas than 900 pixels or larger height/width (or width/height) ratios than 4. Some exemplar hypotheses extracted by the proposed HS method are shown in Figure 7. We sort the hypotheses for each cluster based on the predicted objectness scores and show the first five hypotheses.

During the training step, for each training image, the top k hypotheses from each cluster are selected and fed into the shared CNN. We experimentally vary $k = 1, 2, 3, 4, 5$ to train the proposed HCP on VOC 2007 and observe that the performance changes only slightly on the testing dataset. Therefore, we set $k = 1$ (i.e., 10 hypotheses for each image) for both VOC 2007 and VOC 2012 during the training stage to reduce the training time. To achieve high object recall rate, 500 hypotheses (i.e., $k = 50$) are extracted from each test image during the testing stage. On VOC 2012, the hypotheses-fine-tuning step takes roughly 20 hours. For each testing image, about 1.5 second is cost.

1. http://mmcheng.net/bing/

2. The removed categories include *bicycle, bird, water bottle, bus, car, domestic cat, chair, table, dog, horse, motorcycle, person, flower pot, sheep, sofa, train, tv or monitor, wine bottle, watercraft, unicycle, cattle and cart.*

Fig. 7. Exemplar hypotheses extracted by the proposed HS method. For each image, the ground-truth bounding boxes are shown on the left and the corresponding hypotheses are shown on the right. C1-C10 are the 10 clusters produced by normalized-cut [41] and hypotheses in the same cluster share similar location information.

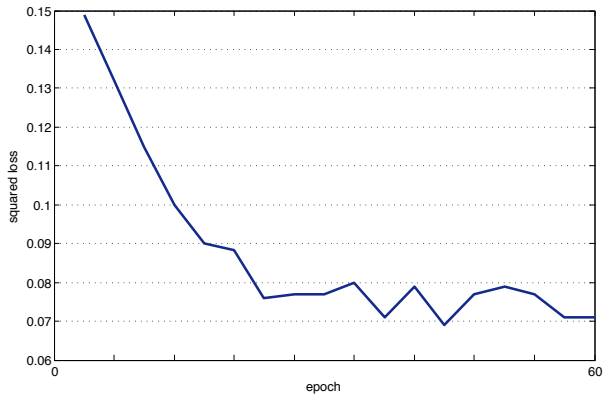

Fig. 8. The changing trend of the squared loss during the I-FT step on VOC 2007.

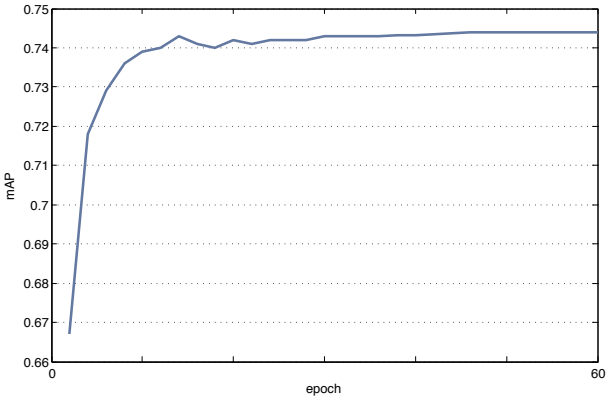

Fig. 9. The changing trend of mAP scores during the I-FT step on VOC 2007. The mAP converges fast to 74.4% after almost 15 epoches on *test* dataset.

4.3 Initialization of HCP

As discussed in Section 3.2, the initialization process of HCP consists of two steps: Pre-training and Image-fine-tuning (I-FT). Since the structure setting of the shared-CNN is almost consistent with the pre-trained model implemented by [22], we apply the pre-trained model on ImageNet by [22] to initialize the convolutional layers and the first two fully-connected layers of the shared CNN. For I-FT, we use images from PASCAL VOC to re-train the shared CNN. As shown in Figure 4, the pre-trained parameters for the first seven layers are transferred to initialize the CNN for fine-tuning. The last fully-connected layer with 4096×20 parameters is randomly initialized with Gaussian distribution.

Actually, similar as Gong *et al.* [16], the class labels of a given image can also be predicted by the fine-tuned model at the I-FT stage. Figure 8 shows the changing trends of the squared loss of different epoches on VOC 2007 during I-FT. The corresponding change of mAP score on the *test* dataset is shown in Figure 9. We can see that the mAP score based on the image-fine-tuned model can achieve 74.4% on VOC 2007, which is more competitive than the scheme of CNN features with SVM classifier [38].

4.4 Image Classification Results

Image Classification on VOC 2007: Table 1 reports our experimental results compared with the state-of-the-arts on VOC 2007. The upper part of Table 1 shows the methods not using ground-truth bounding box information for training, while the lower part of the table shows the methods with that information. Besides, CNN-SVM, I-FT, HCP-1000C, HCP-2000C and PRE-1000C are methods using additional images for training from an extra dataset, i.e., ImageNet, and the other methods only utilize PASCAL VOC data for training. In specific, HCP-1000C indicates that the initialized parameters of the shared CNN are pre-trained on the 1.2 million images from 1000 categories of ILSVRC-2012. Similar as [35], for HCP-2000C, we augment the ILSVRC-2012 training set with additional 1,000 ImageNet classes (about 0.8 million images) to improve the semantic overlap with classes in the Pascal VOC dataset.

From the experimental results, we can see that the CNN based methods which utilize additional images from ImageNet have a 2.6%$\sim$13.9% improvement compared with the state-of-the-art methods based on hand-crafted features, i.e., 71.3% [5]. By utilizing the ground-truth bounding box information, a remarkable improvement can be achieved for both deep learning based methods (PRE-1000C *vs.* CNN-SVM and I-FT) and hand-crafted feature based methods (AGS and AMM *vs.* INRIA and FV). However, bounding box annotation is quite costly. Therefore, approaches requiring ground-truth bounding boxes cannot be transferred to the datasets without such annotation. From Table 1, it can be seen that the proposed HCP has a significant improvement compared with the state-of-the-art performance even without bounding box annotation i.e., 81.5% *vs.* 77.7% (HCP-1000C *vs.* PRE-1000C [35]). Compared with HCP-1000C, 3.7% improvement can be achieved by HCP-2000C. Since the proposed HCP requires no bounding box annotation, the proposed method has a much stronger generalization ability to new multi-label datasets.

Figure 10 shows the predicted scores of images for different categories on the VOC 2007 testing dataset using models from different fine-tuning epoches. For each histogram[3], orange bars indicate predicted scores of the ground-truth categories. We show the predictive scores at the 1^{st} and the 60^{th} epoch during I-FT and H-FT stages. For the first row, it can be seen that the predictive score for the *train* category gradually increases. Besides, for the third row, it can be seen that there are three ground-truth categories

3. For each histogram, categories from left to right are *plane, bike, bird, boat, bottle, bus, car, cat, chair, cow, table, dog, horse, motor, person, plant, sheep, sofa, train and tv.*

TABLE 1
Classification results (AP in %) comparison for state-of-the-art approaches on VOC 2007 (trainval/test). The upper part shows methods not using ground-truth bounding box information for training, while the lower part shows methods with that information. * indicates methods using additional data (i.e., ImageNet) for training.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
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
FV[37]	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*[38]	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*	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*	95.1	90.1	92.8	89.9	51.5	80.0	91.7	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
HCP-2000C*	**96.0**	**92.1**	**93.7**	**93.4**	**58.7**	**84.0**	**93.4**	**92.0**	**62.8**	**89.1**	**76.3**	**91.4**	**95.0**	**87.8**	**93.1**	**69.9**	**90.3**	**68.0**	**96.8**	**80.6**	**85.2**
	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
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[5]	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
PRE-1000C*[35]	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

TABLE 2
Classification results (AP in %) comparison for state-of-the-art approaches on VOC 2012 (trainval/test). The upper part shows methods not using ground-truth bounding box information for training, while the lower part shows methods with that information. * indicates methods using additional data (i.e., ImageNet) for training.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
I-FT*	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
LeCun-ICML*[28]	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
HCP-1000C*	97.7	83.0	93.2	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*	97.5	84.3	93.0	89.4	62.5	90.2	84.6	94.8	69.7	90.2	74.1	93.4	93.7	88.8	93.3	59.7	90.3	61.8	94.4	78.0	84.2
	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
NUS-PSL[47]	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
PRE-1000C*[35]	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*[35]	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
HCP-2000C+NUS-PSL*	**98.9**	**91.8**	**94.8**	**92.4**	**72.6**	**95.0**	**91.8**	**97.4**	**85.2**	**92.9**	**83.1**	**96.0**	**96.6**	**96.1**	94.9	**68.4**	**92.0**	**79.6**	**97.3**	**88.5**	**90.3**

in the given image, i.e., *car, horse, person*. It should be noted that the *car* category is not detected during 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 fine-tuning stage, the entire image is treated as the input, which may lead to ignorance of some background categories.

Image Classification on VOC 2012: Table 2 reports our experimental results compared with the state-of-the-arts on VOC 2012. LeCun *et al.* [28] reported the classification results on VOC 2012, which achieved the state-of-the-art performance without using any bounding box annotation. Compared with [28], the proposed HCP-1000C has an improvement of 2.7%. Both pre-trained on the ImageNet dataset with 1,000 classes, HCP-1000C gives a more competitive result compared with PRE-1000C [35] (81.7% *vs.* 78.7%).

From Table 2, it can be seen that the proposed HCP-1000C is not as competitive as NUS-PSL [47] and PRE-1512 [35]. This can be explained as follows. For NUS-PSL, which got the winner prize of the classification task in PASCAL VOC 2012, model fusion from both detection and classification is employed to generate the integrative result, while the proposed HCP-1000C is based on a single model without any fusion. For PRE-1512, 512 extra ImageNet classes are selected for CNN pre-training. In addition, the selected classes have intensive semantic overlap with PASCAL VOC, including *hoofed mammal, furniture, motor vehicle, public transport, bicycle*. Therefore, the greater improvement of PRE-1512 compared with HCP-1000C is reasonable. By augmenting another 1,000 classes, our proposed HCP-2000C can achieve an improvement of 1.4% compared with PRE-1512.

Finally, the comparison in terms of rigid and articulated categories among NUS-PSL, PRE-1512 and HCP-2000C is shown in Table 3, from which it can be seen that the hand-crafted feature based scheme, i.e., NUS-PSL, outperforms almost all CNN feature based schemes for rigid categories, including *plane, bike, boat, bottle, bus, car, chair, table, motor, sofa, train, tv*, while for articulated categories, CNN feature based schemes seem to be more powerful. Based on theses results, it can be observed that there is strong complementarity between hand-crafted feature based schemes and CNN feature based schemes. To verify this assumption, a late fusion between the predicted scores of NUS-PSL (also from the authors of this paper) and

Fig. 10. Samples of predicted scores on the VOC 2007 testing dataset using models from different fine-tuning epochs (i.e., I-FT-1^{st} epoch, I-FT-60^{th} epoch, H-FT-1^{st} epoch, and H-FT-60^{th} epoch).

HCP is executed to make an enhanced prediction for VOC 2012. Incredibly, the mAP score on VOC 2012 can surge to 90.3% as shown in Table 2, which demonstrates the great complementarity between the traditional framework and the deep networks.

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.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012.
[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
[3] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(7):1312–1328, 2012.
[4] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Trans. Intelligent Systems and Technology*, 2(3):27, 2011.
[5] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classication. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2014.
[6] Q. Chen, Z. Song, Y. Hua, Z. Huang, and S. Yan. Hierarchical matching with side information for image classification. In *Computer Vision and Pattern Recognition*, pages 3426–3433, 2012.
[7] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *International Conference on Computer Vision*, pages 1529–1536, 2013.

TABLE 3
Comparison in terms of rigid categories and articulated categories among NUS-PSL, PRE-1512 and HCP-2000C.

	class	Comparison						
		NUS-PSL	PRE-1512	HCP-2000C	NUS-PSL *vs.* PRE-1512	mean	NUS-PSL *vs.* HCP-2000C	mean
Rigid	plane	97.3	94.6	97.5	2.7	3.0	-0.2	1.5
	bike	84.2	82.9	84.3	1.3		-0.1	
	boat	85.3	84.1	89.4	1.2		-4.1	
	bottle	60.8	60.3	62.5	0.5		-1.7	
	bus	89.9	89.0	90.2	0.9		-0.3	
	car	86.8	84.4	84.6	2.4		2.2	
	chair	75.4	72.1	69.7	3.3		5.7	
	table	75.1	69.0	74.1	6.1		1.0	
	motor	90.1	88.6	88.8	1.5		1.3	
	sofa	73.4	62.3	61.8	11.1		11.6	
	train	94.5	91.1	94.4	3.4		0.1	
	tv	80.7	79.8	78.0	0.9		2.7	
Articulated	bird	80.8	88.2	93.0	-7.4	-5.9	-12.2	-7.1
	cat	89.3	90.7	94.8	-1.4		-5.5	
	cow	77.8	86.8	90.2	-9.0		-12.4	
	dog	83.0	92.1	93.4	-9.1		-10.4	
	horse	87.5	93.4	93.7	-5.9		-6.2	
	person	95.0	96.1	93.3	-1.1		1.7	
	sheep	79.2	86.6	90.3	-7.4		-11.1	
	plant	57.8	64.3	59.7	-7.4		-0.9	

[8] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *Computer Vision and Pattern Recognition*, 2014.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[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 Vision and Pattern Recognition*, pages 827–834, 2013.

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):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. *Computer Vision and Image Understanding*, 106(1):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. In *International Conference on Learning Representations*, 2014.

[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 *Computer Vision and Pattern Recognition*, pages 237–244, 2009.

[20] P. Hedelin and J. Skoglund. Vector quantization based on gaussian mixture models. *IEEE Trans. Speech and Audio Processing*, 8(4):385–401, 2000.

[21] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *International Conference on Computer Vision*, pages 2146–2153, 2009.

[22] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

[23] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Computer Science Department, University of Toronto, Tech. Rep*, 2009.

[24] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, pages 1106–1114, 2012.

[25] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.

[26] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. In *Neural Information Processing Systems*, 1990.

[27] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition*, volume 2, pages II–97, 2004.

[28] Y. LeCun and M. Ranzato. Deep learning tutorial. In *International Conference on Machine Learning*, 2013.

[29] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *International Conference on Machine Learning*, pages 609–616, 2009.

[30] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations*, 2014.

[31] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011.

[32] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[33] N. M. Nasrabadi and R. A. King. Image coding using vector quantization: A review. *IEEE Trans. Communications*, 36(8):957–971, 1988.

[34] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[35] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. *arXiv*, 2013.

[36] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *International Conference on Computer Vision*, pages 2056–2063, 2013.

[37] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, pages 143–156, 2010.

[38] 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.

[39] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *European Conference on Computer Vision*, pages 1–15. 2012.

[40] 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.

[41] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[42] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *Computer Vision and Pattern Recognition*, pages 1585–1592, 2011.

[43] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[44] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition*, pages 3360–3367, 2010.

[45] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[46] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Computer Vision and Pattern Recognition*, pages 1155–1162, 2013.

[47] S. Yan, J. Dong, Q. Chen, Z. Song, Y. Pan, W. Xia, H. Zhongyang, Y. Hua, and S. Shen. Generalized hierarchical matching for subcategory aware object classification. In *Visual Recognition Challange workshop, European Conference on Computer Vision*, 2012.