Multi-scale and Discriminative Part Detectors Based Features for Multi-label Image Classification

Gong Cheng[1], **Decheng Gao**[1], **Yang Liu**[2], **Junwei Han**[1*]
[1]School of Automation, Northwestern Polytechnical University, Xi'an, China
[2]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
{chenggong1119, xidianliuyang, junweihan2010}@gmail.com, decheng@mail.nwpu.edu.cn

Abstract

Convolutional neural networks (CNNs) have shown their promise for image classification task. However, global CNN features still lack geometric invariance for addressing the problem of intra-class variations and so are not optimal for multi-label image classification. This paper proposes a new and effective framework built upon CNNs to learn Multi-scale and Discriminative Part Detectors (MsDPD)-based feature representations for multi-label image classification. Specifically, at each scale level, we (i) first present an entropy-rank based scheme to generate and select a set of discriminative part detectors (DPD), and then (ii) obtain a number of DPD-based convolutional feature maps with each feature map representing the occurrence probability of a particular part detector and learn DPD-based features by using a task-driven pooling scheme. The two steps are formulated into a unified framework by developing a new objective function, which jointly trains part detectors incrementally and integrates the learning of feature representations into the classification task. Finally, the multi-scale features are fused to produce the predictions. Experimental results on PASCAL VOC 2007 and VOC 2012 datasets demonstrate that the proposed method achieves better accuracy when compared with the existing state-of-the-art multi-label classification methods.

1 Introduction

Multi-label image classification has attracted particular attention recently driven by its broad applications [Geng and Luo, 2014; George and Floerkemeier, 2014; Gong *et al.*, 2013; Jing *et al.*, 2015; Li *et al.*, 2016a; Li *et al.*, 2016b; Li *et al.*, 2017; Murthy *et al.*, 2016; Tan *et al.*, 2015; Wang *et al.*, 2016; Wei *et al.*, 2014; Wei *et al.*, 2016; Xie *et al.*, 2017b; Yeh *et al.*, 2017; Zhu *et al.*, 2017]. The task of multi-label image classification is to predict the presence or absence of multiple specific object categories in an image. Compared with single-label image classification which has been actively studied in recent years [Herranz *et al.*, 2016; Krizhevsky *et al.*, 2012; Simon *et al.*, 2014; Simonyan and Zisserman, 2015; Szegedy *et al.*, 2015], multi-label image classification is a more practical problem because most of the real-world images usually contain multiple objects from different categories. Besides, as shown in Figure 1, each object class in real-world multi-label images often has large intra-class variations caused by occlusion, scale, viewpoint, illumination, etc., and the composition and interaction between object categories also increase the complexity of the problem, which make the task of multi-label image classification more challenging.

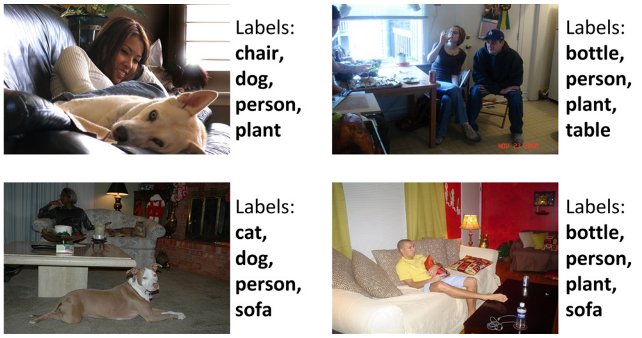

Figure 1: Multi-label images from the PASCAL VOC 2007 dataset. The intra-class variations and the composition and interaction between different object categories make the task of multi-label image classification more challenging.

During the past few years, various deep learning methods especially convolutional neural networks (CNNs) have shown their promise as a universal representation and have dominated most of the recent works on image classification task. However, most research efforts made on image classification mainly focus on addressing the task of single-label image classification. Although several recent works [Oquab *et al.*, 2014; Sharif Razavian *et al.*, 2014; Simonyan and Zisserman, 2015] have demonstrated that a pre-trained CNN model can also be straightforwardly transferred to multi-label image classification, they do not perform well for recognizing complex object layouts and scenes in multi-label images. This mainly because global CNN features still lack geometric invariance for addressing the problem of intra-class variations and so are not optimal for multi-label image classification.

[*] Corresponding author.

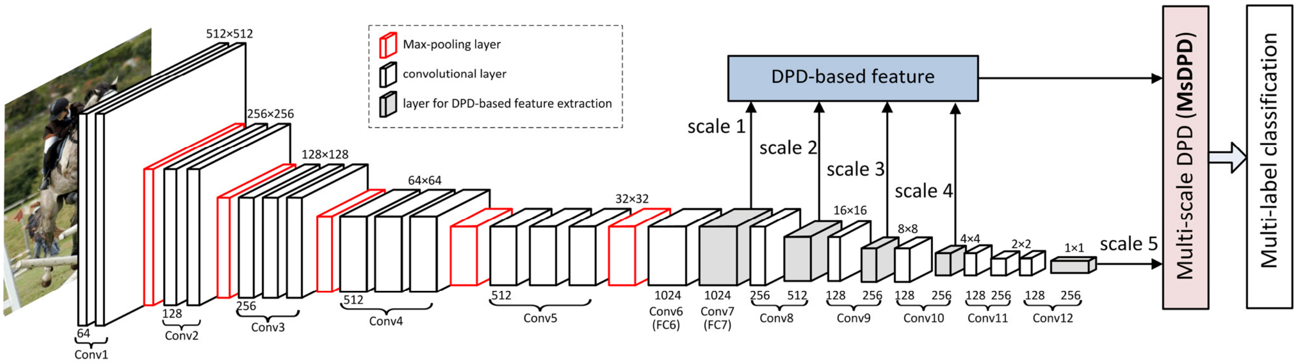

Figure 2: The architecture of the proposed MsDPD-based multi-label image classification framework. At each scale level (denoted by gray blocks), the DPD-based feature representations are learned from the CNN convolutional features by using our proposed optimization method, as shown in Figure 3. The ultimate multi-label predictions are obtained by aggregating the features from different scale levels.

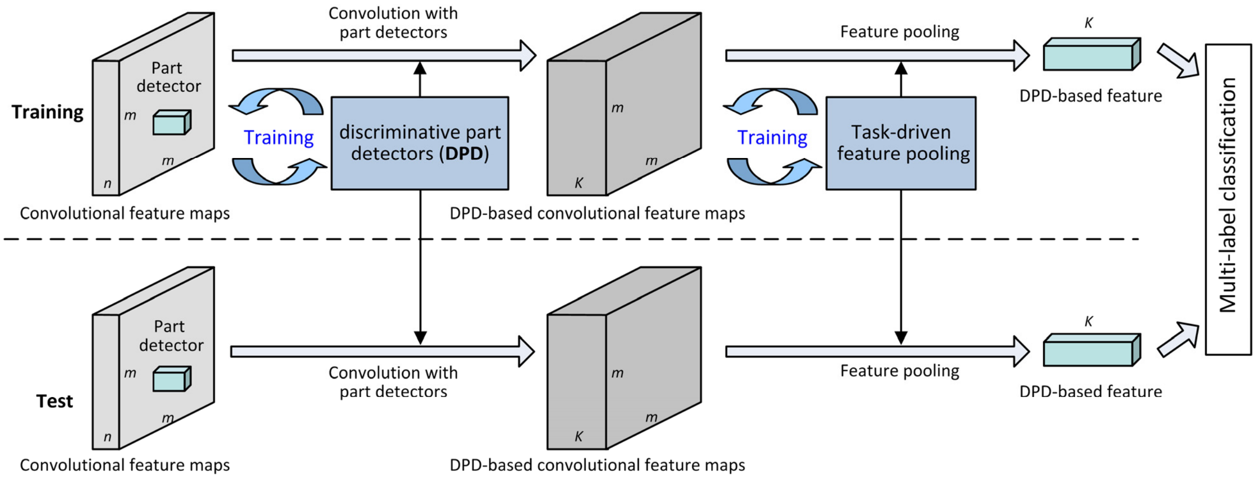

Figure 3: The proposed unified optimization framework for the joint training of discriminative part detectors (DPD) and DPD-based feature representations. Specifically, for a convolutional layer of size $m \times m$ with n channels, we convolve it with K $1 \times 1 \times n$ part detectors to produce a number of part detectors-based feature maps of size $m \times m$ with K channels, followed by a task-driven pooling step to produce the final K-dimensional DPD-based feature representation.

In this paper, we propose a novel and effective framework built upon CNNs to learn multi-scale and discriminative part detectors (MsDPD)-based feature representations for the task of multi-label image classification, as shown in Figure 2. Specifically, at each scale, we first present an object-proposal-free and entropy-rank based scheme to generate and select a number of discriminative part detectors (DPD). Then, we obtain a set of DPD-based feature maps with each feature map representing the occurrence probability of a particular part detector, and learn the pooled DPD-based features by using a task-driven pooling scheme. We formulate the two steps into a unified optimization framework, which trains part detectors incrementally and integrates the learning of feature representations into the classification task, as shown in Figure 3. Finally, the features from different scale levels are aggregated to produce the ultimate multi-label predictions. In the experiments, we evaluate the proposed framework on the PASCAL VOC 2007 and VOC 2012 datasets [Everingham *et al.*, 2015] and achieve state-of-the-art results when compared with the existing multi-label image classification methods.

To sum up, our main contributions are as follows. First, we propose a unified framework by leveraging the highly expressive CNNs to learn a kind of discriminative part detectors-based feature representation, termed MsDPD, to address the problems of intra-class variations faced for multi-label image classification. The proposed approach formulates the training of part detectors and the learning of feature representations into a unified optimization framework by developing a new objective function. Second, we present an entropy-rank based scheme to evaluate the distinctiveness of part detectors and then train part detectors incrementally by mining reliable instances iteratively. Third, we propose a task-driven pooling technique to integrate the learning of feature representation into classification task to improve its generality.

Fourth, different from previous region proposal-based image classification methods [Wei *et al.*, 2016; Wu *et al.*, 2015; Yang *et al.*, 2016], our method does not need ground-truth bounding boxes or object proposals, making the proposed method more efficient and practical. We have confirmed through experiments that the feature representation obtained by using the proposed method is capable of delivering state-of-the-art results on two popular multi-label classification benchmarks including PASCAL VOC 2007 and VOC 2012 datasets.

2 Methodology

While many CNN-based methods have achieved successful results on image classification, most of them are developed for single-label image classification by extracting global CNN features. Inspired by the fact that each object class in multi-label images generally exhibits dramatically different appearances, shapes, occlusions and interactions, we propose to extract discriminative part detectors-based features, a kind of local CNN-based features, to handle the problem of intra-class variations. Here, part detectors are used to capture generalized objects and their parts that are discriminative (being different enough from each other) and representative (occuring frequently enough). As shown in Figure 2, the core task of this proposed method is to learn Multi-scale and Discriminative Part Detectors (MsDPD)-based features for multi-label image classification. Figure 3 illustrates how to learn DPD-based features from each scale of convolutional features denoted by gray blocks. Specifically, we first present an entropy-rank based scheme to generate a number of discriminative part detectors. Then, we obtain part detectors-based convolutional feature maps and generate the pooled feature representations by using a task-driven feature pooling scheme. For ease of optimization we integrate the two steps into a unified framework by developing a new objective function to jointly train part detectors and learn the feature representations. The final multi-label prediction results are obtained by fusing the features from different scale levels.

2.1 Model Architecture

Figure 2 illustrates the overall architecture of our MsDPD framework. The basic configuration of our model is similar to that of Single Shot MultiBox Detector (SSD) [Liu *et al.*, 2016]. The early layers (Conv1 to Conv7) are transferred from the pre-trained VGGNet-16 [Simonyan and Zisserman, 2015], where the convolutional layers Conv6 and Conv7 are converted from the fully-connected layers FC6 and FC7 by using a scheme that is similar to SSD as follows: subsample parameters from FC6 and FC7, change pool5 from 2×2 - s2 to 3×3 - s1, and use the `a trous algorithm to fill the "holes". The fully-connected layers are converted to convolutional ones to cope with the uncertainty for the localization of object parts. These layers are followed by some extra convolutional layers (Conv8 to Conv12) to extract much deeper features and even bigger object parts. The last convolutional layer (Conv12_2) is used to fine-tune the network by using multi-label images. The detailed model parameters can be found in Figure 2.

In this work, DPD-based feature representations are extracted based on four scales of convolutional layers including Conv7, Conv8_2, Conv9_2, and Conv10_2 (denoted by gray blocks in Figure 2), which decrease in size progressively to allow the detections of object parts at multiple scales. Specifically, for a convolutional layer of size $m \times m$ with n channels, we convolve it with K $1 \times 1 \times n$ part detectors to produce a number of part detectors-based convolutional feature maps of size $m \times m$ with K channels, where each feature map represents the occurrence probability of a particular part detector, followed by a task-driven pooling step to produce the final K-dimensional DPD-based feature representation. For ease of reference, we index the four DPD-based feature layers and the last convolutional layer by using scale 1 through scale 5. Next we describe how to train part detectors and learn part detectors-based feature representations.

2.2 Initializing Discriminative Part Detectors (DPD)

To initialize the candidate part detectors that are shared across all image categories, we randomly sample a large number of (about one hundred thousands) pixels from each scale of the convolutional feature maps of all training images. Each pixel from the feature maps can be considered as a "local" CNN feature which has a very large receptive field in the original image, and the length of the pixel equals the channel number of the convolutional features. Then we perform *k*-means clustering over these sampled pixels and only retain sufficiently large clusters to ensure the representativeness, where each cluster corresponds to a to-be-learned candidate part detector.

We consider an object part is discriminative if it only appears frequently in some specific image categories rather than almost all classes. For example, "wheel" will occur in the object classes of "bus" and "car", so the entropy would be low. In contrast, a non-discriminative "sky" could occur uniformly in almost any of the classes with higher entropy. To select discriminative part detectors, we introduce an entropy-rank based scheme to measure the discrimination of each part detector by computing their entropies across all image categories. Specifically, the entropy $E(D)$ for a part detector D is computed by

$$E(D) = -\sum_{c=1}^{C} p_c(D) \log p_c(D) \tag{1}$$

where C is the number of image classes and $p_c(D)$ is the fraction of the members of part detector D that are from the images of the c-th class. Then, we take the entropy as a measure of discrimination of a part detector to select K detectors with low entropy values.

2.3 Training DPD and Learning DPD-based Feature

Let $\mathcal{X} = \{\mathbf{X}_i \mid i = 1, 2, \cdots, N\}$ be the set of training images and $\mathcal{Y} = \{\boldsymbol{Y}_i \mid i = 1, 2, \cdots, N\}$ be the set of image labels of $\mathcal{X}$, where $\mathbf{X}_i = \{\boldsymbol{x}_{ij} \in \mathbb{R}^n \mid j = 1, 2, \cdots, M\}$ is represented by the entries from its CNN convolutional layer of size $m \times m$ with n channels and $M = m^2$, N is the total number of training images, $\boldsymbol{Y}_i \in \mathbb{R}^C$ denotes the ground truth label vector of sample X_i with at least one element being 1, and C is the number of image classes. For a given training image $\mathbf{X}_i$, let

$\mathbf{O}(\mathbf{X}_i) = \{\boldsymbol{O}(\boldsymbol{x}_{ij}) \in \mathbb{R}^K \mid j = 1, 2, \cdots, M\}$ be its DPD-based feature maps, $\boldsymbol{\varphi}(\mathbf{X}_i) \in \mathbb{R}^K$ be its pooled DPD-based feature, $\boldsymbol{P}(\mathbf{X}_i)$ be its multi-label prediction result, $(\mathbf{W}_1, \boldsymbol{b}_1)$ be the parameters of the to-be-learned DPD, and $(\mathbf{W}_2, \boldsymbol{b}_2)$ be the parameters of C multi-label classifiers, where $\mathbf{W}_1 \in \mathbb{R}^{n \times K}$, $\boldsymbol{b}_1 \in \mathbb{R}^K$, $\mathbf{W}_2 \in \mathbb{R}^{K \times C}$, and $\boldsymbol{b}_2 \in \mathbb{R}^C$. Thus, $\boldsymbol{O}(\boldsymbol{x}_{ij})$ and $\boldsymbol{P}(\mathbf{X}_i)$ can be computed by

$$\boldsymbol{O}(\boldsymbol{x}_{ij}) = S\left(\mathbf{W}_1^T \boldsymbol{x}_{ij} + \boldsymbol{b}_1\right) \tag{2}$$

$$\boldsymbol{P}(\mathbf{X}_i) = \sigma\left(\mathbf{W}_2^T \boldsymbol{\varphi}_i + \boldsymbol{b}_2\right) \tag{3}$$

where $S(\boldsymbol{x}) = \exp(\boldsymbol{x}) / \|\exp(\boldsymbol{x})\|_1$ and $\sigma(\boldsymbol{x}) = (1 + \exp(-\boldsymbol{x}))^{-1}$ are the softmax and sigmoid non-linear activation functions, which are used for predicting the occurrence probabilities of DPD and the multi-label classification results, respectively.

As shown in Figure 3, to leverage CNNs to learn effective feature representation, we formulate the training of DPD and the learning of DPD-based feature representation into a unified framework, which jointly trains part detectors incrementally and integrates DPD-based feature learning into classification. To this end, we develop a new objective function as follows, which contains three terms including an image-level classification loss term, a generalized max pooling regularization term, and an object part-level classification loss term:

$$J = \min \begin{pmatrix} J_1(\mathcal{X}, \mathcal{Y}, \mathbf{W}_2, \boldsymbol{b}_2, \boldsymbol{\varphi}) + \dfrac{\lambda_1}{2} J_2(\mathbf{W}_1, \boldsymbol{b}_1, \boldsymbol{\varphi}) \\ + \lambda_2 J_3(\mathbf{W}_1, \boldsymbol{b}_1) \end{pmatrix} \tag{4}$$

where λ_1 and λ_2 are two trade-off parameters that control the relative importance of these three terms.

1) Image-level classification loss term. This term is defined as sigmoid cross-entropy loss function for multi-label image classification. It aims to minimize the classification error for the given training images and is computed by

$$J_1 = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \begin{pmatrix} [\boldsymbol{Y}_i]_c \log\left[\boldsymbol{P}(\mathbf{X}_i)\right]_c + \\ \left(1 - [\boldsymbol{Y}_i]_c\right) \log\left(1 - \left[\boldsymbol{P}(\mathbf{X}_i)\right]_c\right) \end{pmatrix} \tag{5}$$

where $[\boldsymbol{Y}_i]_c$ and $\left[\boldsymbol{P}(\mathbf{X}_i)\right]_c$ denote the c-th entries of $\boldsymbol{Y}_i$ and $\boldsymbol{P}(\mathbf{X}_i)$, respectively.

2) Generalized max pooling regularization term. This term is used to learn the pooled DPD-based feature $\boldsymbol{\varphi}(\mathbf{X}_i)$ from the input $\mathbf{O}(\mathbf{X}_i)$ by enforcing the pooled representation to be close to each column of the input $\mathbf{O}(\mathbf{X}_i)$, which is computed by using the following formula

$$J_2 = \frac{1}{MN} \sum_{i=1}^{N} \left(\sum_{j=1}^{M} \left(\boldsymbol{O}(\boldsymbol{x}_{ij})^T \boldsymbol{\varphi}(\mathbf{X}_i) - 1 \right)^2 + \left\| \boldsymbol{\varphi}(\mathbf{X}_i) \right\|_2^2 \right) \tag{6}$$

Similar to [Murray and Perronnin, 2014; Xie *et al.*, 2015], by using this pooling regularization term, the learned DPD-based feature could enforce the dot product similarity between $\boldsymbol{O}(\boldsymbol{x}_{ij})$ and the pooled feature $\boldsymbol{\varphi}(\mathbf{X}_i)$ to be a constant one. By integrating feature learning into classification, we can use more information from $\mathbf{O}(\mathbf{X}_i)$ to get a task-driven feature representation which is more suitable for classification than traditional pooling strategies such as max/average-pooling.

3) Object part-level classification loss term. This term is defined as softmax cross-entropy loss function for object part classification. It aims to minimize the classification error for the mined object part instances and is computed by

$$J_3 = -\frac{1}{Kt} \sum_{l=1}^{Kt} \sum_{k=1}^{K} \left([\boldsymbol{y}_l]_k \log\left[\boldsymbol{O}(\boldsymbol{x}_l)\right]_k \right) \tag{7}$$

where $\boldsymbol{y}_l \in \mathbb{R}^K$ stands for object part label vector of object part instance $\boldsymbol{x}_l$ with only one element being 1, K is the number of DPD selected in subsection 2.2, and t is the number of high-confident object parts we select in each iteration used for updating each part detector. In such way, we can mine reliable instances iteratively and train the DPD incrementally.

2.4 Optimization

To solve the optimization problem of Eq. (4), we present a simple EM-like iterative minimization method to update $(\mathbf{W}_1, \boldsymbol{b}_1)$, $(\mathbf{W}_2, \boldsymbol{b}_2)$ and $\boldsymbol{\varphi}$ alternatively via stochastic gradient descent method (SGD) [Williams and Hinton, 1986].

1) Initialization. Given K selected part detectors, we initialize the parameters $(\mathbf{W}_1, \boldsymbol{b}_1)$ by using Eq. (7). The DPD-based feature representation $\boldsymbol{\varphi}$ for all images are initialized by using (6) and generalized max pooling method [Murray and Perronnin, 2014]. The parameters $(\mathbf{W}_2, \boldsymbol{b}_2)$ are initialized by using Eq. (5).

2) Updating $(\mathbf{W}_1, \boldsymbol{b}_1)$ and $(\mathbf{W}_2, \boldsymbol{b}_2)$ by fixing $\boldsymbol{\varphi}$. The gradients of the objective function J with respect to the parameters $(\mathbf{W}_1, \boldsymbol{b}_1)$ and $(\mathbf{W}_2, \boldsymbol{b}_2)$ can be computed by

$$\frac{\partial J}{\partial \mathbf{W}_1} = \frac{\lambda_1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} (g_{ij} - 1)\left(\boldsymbol{x}_{ij} \boldsymbol{z}_{ij}^T\right) + \frac{\lambda_2}{Kt} \sum_{l=1}^{Kt} \boldsymbol{x}_l \left[\boldsymbol{O}(\boldsymbol{x}_l) - \boldsymbol{y}_l\right]^T \tag{8}$$

$$\frac{\partial J}{\partial \boldsymbol{b}_1} = \frac{\lambda_1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} (g_{ij} - 1) \boldsymbol{z}_{ij} + \frac{\lambda_2}{Kt} \sum_{l=1}^{Kt} \left[\boldsymbol{O}(\boldsymbol{x}_l) - \boldsymbol{y}_l\right] \tag{9}$$

$$\frac{\partial J}{\partial \mathbf{W}_2} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\varphi}(\mathbf{X}_i) \left(\boldsymbol{P}(\mathbf{X}_i) - \boldsymbol{Y}_i\right)^T \tag{10}$$

$$\frac{\partial J}{\partial \boldsymbol{b}_2} = \frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{P}(\mathbf{X}_i) - \boldsymbol{Y}_i\right) \tag{11}$$

where g_{ij} and $\boldsymbol{z}_{ij}$ are defined as follows

$$g_{ij} \triangleq \boldsymbol{O}(\boldsymbol{x}_{ij})^T \boldsymbol{\varphi}(\mathbf{X}_i) \tag{12}$$

$$\boldsymbol{z}_{ij} \triangleq \boldsymbol{O}(\boldsymbol{x}_{ij}) \odot \boldsymbol{\varphi}(\mathbf{X}_i) - g_{ij} \boldsymbol{O}(\boldsymbol{x}_{ij}) \tag{13}$$

with the operation $\odot$ denoting element-wise multiplication.

Thus, the parameters $(\mathbf{W}_1, \boldsymbol{b}_1)$ and $(\mathbf{W}_2, \boldsymbol{b}_2)$ can be updated by using gradient descent method as follows

$$\mathbf{W}_1 = \mathbf{W}_1 - \mu \frac{\partial J}{\partial \mathbf{W}_1}, \ \boldsymbol{b}_1 = \boldsymbol{b}_1 - \mu \frac{\partial J}{\partial \boldsymbol{b}_1} \tag{14}$$

$$\mathbf{W}_2 = \mathbf{W}_2 - \mu \frac{\partial J}{\partial \mathbf{W}_2}, \ \boldsymbol{b}_2 = \boldsymbol{b}_2 - \mu \frac{\partial J}{\partial \boldsymbol{b}_2} \tag{15}$$

where μ is the learning rate.

3) Updating $\boldsymbol{\varphi}$ by fixing $(\mathbf{W}_1, \boldsymbol{b}_1)$ and $(\mathbf{W}_2, \boldsymbol{b}_2)$. The gradients of the objective function J with respect to $\boldsymbol{\varphi}$ can be computed by

$$\begin{aligned}\frac{\partial J}{\partial \boldsymbol{\varphi}(\mathbf{X}_i)} &= \frac{1}{N}\mathbf{W}_2\left(\boldsymbol{P}(\mathrm{X}_i) - \boldsymbol{Y}_i\right) + \\ &\frac{\lambda_1}{M}\sum_{j=1}^{M}\left[\boldsymbol{O}(\boldsymbol{x}_{ij})^T \boldsymbol{\varphi}(\mathbf{X}_i) - 1\right]\boldsymbol{O}(\boldsymbol{x}_{ij}) + \lambda_1 \boldsymbol{\varphi}(\mathbf{X}_i)\end{aligned} \tag{16}$$

Thus, the parameter $\boldsymbol{\varphi}$ can be updated as follows

$$\boldsymbol{\varphi}(\mathbf{X}_i) = \boldsymbol{\varphi}(\mathbf{X}_i) - \mu \frac{\partial J}{\partial \boldsymbol{\varphi}(\mathbf{X}_i)} \tag{17}$$

2.5 Multi-label Image Classification

After optimizing Eq. (4), the pooled DPD-based features for all training samples are learned at the same time. However, the feature representations of the test images still need to be learned. Since the part detectors make the distributions of training and test data consistent, we can obtain DPD-based features of test images by optimizing Eq. (6). For the ultimate predictions, we concatenate the features from different scale levels to train a set of sigmoid classifiers for prediction.

3 Experiments

In the experiments, we evaluate our method on PASCAL VOC 2007 and VOC 2012 datasets [Everingham *et al.*, 2015], which have been widely used for multi-label image classification by predicting whether the object is present/absent in the image. The performance is measured by using the average precision (AP) and the mean AP over all object classes.

3.1 Parameter Settings

We train the proposed model as shown in Figure 2 by using SGD with initial learning rate of 10^{-4} for the early layers (Conv1 to Conv7), initial learning rate of 10^{-3} for the latter layers (Conv8 to Conv12), momentum of 0.9, weight decay of 0.0005, and batch size of 32. The learning rate decays by 0.1 after 60k iterations and is fixed for the rest 20k iterations. For the training of DPD and DPD-based features, the parameters in Eq. (4) are set to λ_1=1 and λ_2=0.01, and the learning rate μ in Eqs. (14), (15), (17) is set to 0.01.

For the initialization of DPD, we run *k*-means clustering on the convolutional feature maps of scales 1 to 4 with the cluster numbers being set to 2000, 1000, 400, and 200, and then take the entropy-rank based scheme as a measure to select 700, 400, 300, and 200 detectors, respectively.

3.2 Experimental Results

Comparison of features from different scales. We first give the results obtained by using different scales on PASCAL VOC 2007 dataset. Table 1 reports the detailed results. As shown in Table 1, some object classes, such as "bird" and "bottle", fire on small scales and some object categories, such as "person" and "train" fire on big scales. This is because that our MsDPD feature layers are decreased in size progressively to allow the predictions of objects and their parts at multiple scales, thereby for better capture of object variations caused by viewpoint, scale, occlusion, etc. The best results are obtained by fusing the features of different scales.

State-of-the-art CNN-based methods. The following CNN-based methods are used for comparison: VGG-16-SVM and VGG-19-SVM [Simonyan and Zisserman, 2015], ResNet-101-Sigmoid [He *et al.*, 2016], SDE [Xie *et al.*, 2017a], HCP [Wei *et al.*, 2016], CNN-RNN [Wang *et al.*, 2016], and FeV+LV-20-VD [Yang *et al.*, 2016]. The work of [Simonyan and Zisserman, 2015] densely extracts 4096-D CNN features across five image scales {256,384,512,640,748} of the given image with VGG-16 and VGG-19, performs global average pooling on the resulting CNN features, and finally classifies the image with linear SVM classifiers. ResNet-101-Sigmoid trains a multi-label classification system using a pre-trained ResNet-101 model [He *et al.*, 2016] with a sigmoid cross entropy loss function, densely computes sigmoid outputs across five image scales {256,384,512,640,748} of the given image, and finally performs classification by max-pooling the resulting sigmoid outputs as HCP [Wei *et al.*, 2016]. SDE [Xie *et al.*, 2017a] presented a feature learning framework by optimizing the features with the aim of learning selective, discriminative and equalizing representations. HCP [Wei *et al.*, 2016] proposed to address the multi-label classification by extracting object proposals from the given images and the final image-level scores are obtained by max-pooling the scores of the proposals. CNN-RNN [Wang *et al.*, 2016] combined RNNs with CNNs in a unified framework to learn a joint image-label embedding. FeV+LV-20-VD [Yang *et al.*, 2016] proposed a multi-view multi-instance framework to utilize both weak and strong labels (bounding box).

Comparison with state-of-the-art methods on PASCAL VOC 2007 dataset. Table 2 summarizes the results of our MsDPD method and the aforementioned seven state-of-the-art methods on PASCAL VOC 2007 dataset. As shown in Table 2: (i) Compared with global CNN-based approaches such as VGG-16-SVM and VGG-19-SVM [Simonyan and Zisserman, 2015], ResNet-101-Sigmoid [He *et al.*, 2016], and CNN-RNN [Wang *et al.*, 2016], our proposed method obtains significant performance gains of 4.2%, 3.3% and 9.5% in terms of mAP. This shows the superiority of our local CNN-based method. (ii) Compared with other methods, such as HCP [Wei *et al.*, 2016], SDE [Xie *et al.*, 2017a], and FeV+LV-20-VD [Yang *et al.*, 2016], which can be regarded as a kind of local feature based methods, our method still outperforms them with a big margin measured in terms of mAP (at least 2.3%).

Comparison with state-of-the-art methods on PASCAL VOC 2012 dataset. We report our experimental results in Table 3 and compare it with six state-of-the-art CNN-based methods on VOC 2012 dataset. The results are consistent with those on the VOC 2007 dataset. To be specific, we achieve state-of-the-art results for 16 out of 20 object categories. Especially for the difficult categories such as "chair", "cow",

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Scale 1	98.4	96.7	**96.3**	96.5	75.7	97.5	93.7	95.5	77.9	**95.5**	90.9	95.0	97.5	93.8	90.9	77.9	96.3	84.3	99.1	91.8	92.1
Scale 2	98.1	97.2	**96.3**	96.3	64.6	97.8	94.5	**96.5**	79.2	94.3	89.6	95.2	97.3	96.8	90.1	79.7	96.0	87.6	99.2	90.9	91.9
Scale 3	97.3	97.7	94.9	95.6	57.8	95.9	92.2	95.0	74.2	92.2	82.9	94.2	97.0	96.3	91.7	76.0	93.6	82.1	98.7	86.2	89.6
Scale 4	97.1	95.5	93.4	93.0	50.5	90.9	92.6	93.9	72.3	93.0	85.5	93.1	97.0	95.0	93.6	65.3	92.5	84.7	98.8	76.9	87.7
Scale 5	96.4	92.4	92.9	93.4	49.6	89.8	90.2	92.0	69.1	80.4	85.5	89.5	96.6	90.0	91.8	59.6	78.9	85.6	98.7	70.5	84.6
MsDPD	**98.7**	**97.9**	96.2	**96.9**	**76.0**	**97.9**	**95.7**	**96.5**	**81.9**	95.3	**91.5**	**96.3**	**97.6**	**96.9**	**95.6**	**81.6**	**96.7**	**88.4**	99.3	**92.4**	**93.5**

Table 1: Classification results (%) on the PASCAL VOC 2007 test set obtained by using different scales of MsDPD and their fusion (scales 1 to 5). The entries with the best APs for each object category are bold-faced.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
VGG-16-SVM MS†	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.3
VGG-19-SVM MS†	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.3
ResNet-101-Sigmoid MS†	97.6	95.4	94.7	94.5	74.9	91.4	93.9	96.3	77.5	90.0	84.8	94.4	95.2	93.9	98.1	70.4	92.4	82.3	98.3	88.7	90.2
HCP	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
CNN-RNN	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	**99.7**	78.6	84.0
FeV+LV-20-VD	97.9	97.0	96.6	94.6	73.6	93.9	**96.5**	95.5	73.7	90.3	82.8	95.4	**97.7**	95.9	**98.6**	77.6	88.7	78.0	98.3	89.0	90.6
SDE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	91.2
Our MsDPD method	**98.7**	**97.9**	96.2	**96.9**	**76.0**	**97.9**	95.7	96.5	**81.9**	**95.3**	**91.5**	96.3	97.6	**96.9**	95.6	**81.6**	**96.7**	**88.4**	99.3	**92.4**	**93.5**

Table 2: Classification results (%) on the PASCAL VOC 2007 test set obtained by using state-of-the-art CNN-based methods and our proposed MsDPD method. †: MS denotes the results obtained by using a multi-scale scheme with five image scales {256,384,512,640,748}.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
VGG-16-SVM MS†	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 MS†	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
ResNet-101-Sigmoid MS†	98.7	88.9	93.1	92.9	76.3	92.3	85.6	96.6	80.5	85.3	81.9	93.2	94.0	93.2	97.6	64.0	88.1	76.5	97.1	90.3	88.3
HCP	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
FeV+LV-20-VD	98.4	92.8	93.4	90.7	74.9	93.2	90.2	96.1	78.2	89.8	80.6	95.7	96.1	95.3	**97.5**	73.1	91.2	75.4	97.0	88.2	89.4
SDE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	91.1
Our MsDPD method	**99.5**	**94.0**	94.7	**95.3**	**82.8**	**95.3**	**96.9**	96.5	**85.1**	**95.1**	**86.1**	94.7	**98.0**	**95.5**	95.0	**78.8**	**94.8**	**86.5**	**98.8**	**93.1**	**92.8**

Table 3: Classification results (%) on the PASCAL VOC 2012 test set obtained by using state-of-the-art CNN-based methods and our proposed MsDPD method. †: MS denotes the results obtained by using a multi-scale scheme with five image scales {256,384,512,640,748}.

"table", "plant", and "sofa", our method shows good performance. This significant performance gain shows the effectiveness of our DPD-based feature representation.

Ablation experiments. To analyze the importance of each component of our method (part detectors and task-driven pooling), we conducted ablation experiments on the PASCAL VOC 2007 dataset. Table 4 shows the results obtained with part detectors and without part detectors (pool features from Conv7/Conv8_2/Conv9_2/Conv10_2 layers) by using different pooling strategies measured in terms of mAP. As shown in Table 4, task-driven pooling obtains the highest mAP than max-pooling and average-pooling. More importantly, by using our proposed part detectors could obtain big accuracy gains compared with that obtained by directly pooling features from the original convolutional layers.

Without part detectors	Max-pooling	89.2
	Average-pooling	88.2
	Task-driven pooling	90.7
With part detectors	Max-pooling	93.2
	Average-pooling	87.9
	Task-driven pooling	**93.5**

Table 4: Ablation experimental results (mAP, %) on the PASCAL VOC 2007 test set obtained with part detectors and without part detectors by using different pooling strategies.

4 Conclusion

In this paper, we proposed to build upon CNNs to learn part detectors-based features for multi-label image classification. To this end, we first present an entropy-rank based scheme to obtain a set of discriminative part detectors. Then, we generate part detectors-based convolutional feature maps and learn part detectors-based features with a task-driven pooling scheme. For optimization, the aforementioned two steps are formulated into a unified framework by developing a new objective function, which incrementally trains part detectors and integrates the learning of feature representations into the classification task. However, by using the proposed objective function it is difficult to train the whole network end-to-end. Our future work will address this issue.

Acknowledgments

This work was supported in part by the NSFC under Grants 61772425 and 61473231, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2017JM6044, and in part by the Fundamental Research Funds for the Central Universities under Grant 3102018zy023.

References

[Everingham *et al.*, 2015] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV,* 111(1): 98-136, 2015.

[Geng and Luo, 2014] X. Geng and L. Luo. Multilabel ranking with inconsistent rankers. In *CVPR*, 2014.

[George and Floerkemeier, 2014] M. George and C. Floerkemeier. Recognizing products: A per-exemplar multi-label image classification approach. In *ECCV*, 2014.

[Gong *et al.*, 2013] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894,* 2013.

[He *et al.*, 2016] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Herranz *et al.*, 2016] L. Herranz, S. Jiang, and X. Li. Scene recognition with CNNs: objects, scales and dataset bias. In *CVPR*, 2016.

[Jing *et al.*, 2015] L. Jing, L. Yang, J. Yu, and M. K. Ng. Semi-supervised low-rank mapping learning for multi-label classification. In *CVPR*, 2015.

[Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Li *et al.*, 2016a] C. Li, B. Wang, V. Pavlu, and J. Aslam. Conditional bernoulli mixtures for multi-label classification. In *ICML*, 2016a.

[Li *et al.*, 2016b] Q. Li, M. Qiao, W. Bian, and D. Tao. Conditional graphical lasso for multi-label image classification. In *CVPR*, 2016b.

[Li *et al.*, 2017] Y. Li, Y. Song, and J. Luo. Improving Pairwise Ranking for Multi-label Image Classification. In *CVPR*, 2017.

[Liu *et al.*, 2016] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[Murray and Perronnin, 2014] N. Murray and F. Perronnin. Generalized max pooling. In *CVPR*, 2014.

[Murthy *et al.*, 2016] V. N. Murthy, V. Singh, T. Chen, R. Manmatha, and D. Comaniciu. Deep decision network for multi-class image classification. In *CVPR*, 2016.

[Oquab *et al.*, 2014] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[Sharif Razavian *et al.*, 2014] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *CVPRW*, 2014.

[Simon *et al.*, 2014] M. Simon, E. Rodner, and J. Denzler. Part detector discovery in deep convolutional neural networks. In *ACCV*, 2014.

[Simonyan and Zisserman, 2015] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Szegedy *et al.*, 2015] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[Tan *et al.*, 2015] M. Tan, Q. Shi, A. van den Hengel, C. Shen, J. Gao, F. Hu, and Z. Zhang. Learning graph structure for multi-label image classification via clique generation. In *CVPR*, 2015.

[Wang *et al.*, 2016] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. CNN-RNN: A unified framework for multi-label image classification. In *CVPR*, 2016.

[Wei *et al.*, 2014] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. CNN: Single-label to multi-label. *arXiv preprint arXiv:1406.5726,* 2014.

[Wei *et al.*, 2016] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. HCP: A flexible CNN framework for multi-label image classification. *IEEE TPAMI,* 38(9): 1901-1907, 2016.

[Williams and Hinton, 1986] D. Williams and G. Hinton. Learning representations by back-propagating errors. *Nature,* 323(6088): 533-538, 1986.

[Wu *et al.*, 2015] R. Wu, B. Wang, W. Wang, and Y. Yu. Harvesting discriminative meta objects with deep CNN features for scene classification. In *ICCV*, 2015.

[Xie *et al.*, 2015] G.-S. Xie, X.-Y. Zhang, X. Shu, S. Yan, and C.-L. Liu. Task-driven feature pooling for image classification. In *ICCV*, 2015.

[Xie *et al.*, 2017a] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu. SDE: A Novel Selective, Discriminative and Equalizing Feature Representation for Visual Recognition. *IJCV,* 1-24, 2017a.

[Xie *et al.*, 2017b] P. Xie, R. Salakhutdinov, L. Mou, and E. P. Xing. Deep Determinantal Point Process for Large-Scale Multi-Label Classification. In *ICCV*, 2017b.

[Yang *et al.*, 2016] H. Yang, J. Tianyi Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai. Exploit bounding box annotations for multi-label object recognition. In *CVPR*, 2016.

[Yeh *et al.*, 2017] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang. Learning Deep Latent Space for Multi-Label Classification. In *AAAI*, 2017.

[Zhu *et al.*, 2017] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang. Learning Spatial Regularization with Image-level Supervisions for Multi-label Image Classification. In *CVPR*, 2017.