Multi-Label Image Classification via Feature / Label Co-Projection

Shiping Wen, Weiwei Liu, Yin Yang, Pan Zhou, Zhenyuan Guo, Zheng Yan,
Yiran Chen *Fellow, IEEE*, and Tingwen Huang, *Fellow, IEEE*

Abstract—This paper presents a simple and intuitive solution for multi-label image classification, which achieves the competitive performance on the popular COCO and PASCAL VOC benchmarks. The main idea is to capture how humans perform this task: we recognize both labels (i.e., objects and attributes) and the correlation of labels at the same time. Here, label recognition is performed by a standard ConvNet pipeline, whereas label correlation modeling is done by projecting both labels and image features extracted by the ConvNet to a common latent vector space. Specifically, we carefully design the loss function to ensure that (i) labels and features that co-appear frequently are close to each other in the latent space and (ii) conversely, labels / features that do not appear together are far apart. This information is then combined with the original ConvNet outputs to form the final prediction. The whole model is trained end-to-end, with no additional supervised information other than the image-level supervised information. Experiments show that the proposed method consistently outperforms previous approaches on COCO and PASCAL VOC in terms of mAP, macro/micro precision, recall, and F-measure. Further, our model is highly efficient at test time, with only a small number of additional weights compared to the base model for direct label recognition.

Index Terms—Multi-label Classification, Label Embedding, Neural Network, Deep Learning.

1 Introduction

Multi-label image classification is a fundamental task in computer vision with numerous applications [1], [2], [3], [4], [5], [6]. In this task, each input image is associated with a set of labels, where the universe of all possible labels are given, but the number of labels matching an image is often not known beforehand, and can vary from image to image. For example, in Figure 1a, the image clearly matches labels such as “person”, “tennis racket” and “tennis ball”. The output of multi-label classification is usually represented as a binary vector, in which each bit indicates the presence or absence of a label in the given image.

There has been a plethora of methods for multi-label image classification. Yet, few of them reflect how humans approach this problem. To illustrate, consider Figure 1b, which covers up the left half of the image in Figure 1a. To a human, the image here presents a *context* (e.g., from the pose of the man and the position of his racket) that strongly suggests the existence of a tennis ball. The photo would be rather unsatisfying if it does not show a ball, and downright bizarre if instead of a ball, there is a sheep or the face of a celebrity at the left side of the image. Meanwhile, the context alone may be insufficient to identify all matching labels. Figure 1a, for instance, also matches the label “chair”, which is not obvious from the context, and needs to be recognized from its own visual features. However, The style of the machine algorithm is quite different from that of human beings to understand data. It is much harder for

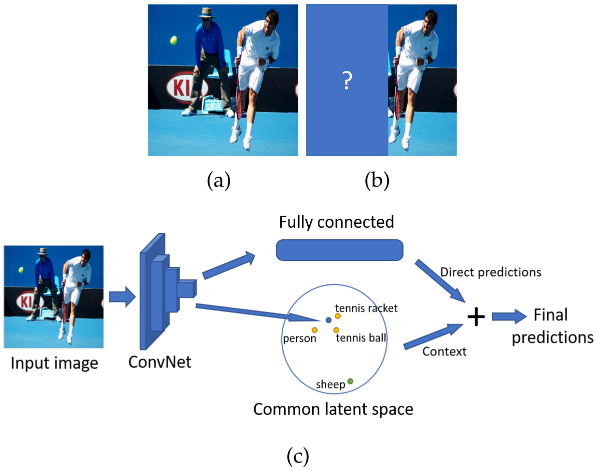

Fig. 1. (a) an input image associated with labels person, tennis racket and tennis ball; (b) right half of the same image, from which the presence of tennis ball can be inferred; (c) proposed neural network pipeline that combines both direct predictions from a ConvNet and contextual information extracted by projecting image features and labels to a common vector space.

- *Shiping Wen is with School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, E-mail: wenshiping@uestc.edu.cn. Weiwei Liu is with School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China, E-mail: liuvv@hust.edu.cn. Yin Yang is with College of Science and Engineering, Hamad Bin Khalifa University, E-mail: yyang@hbku.edu.qa. Pan Zhou is with the School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China, E-mail: panzhou@hust.edu.cn. Zhenyuan Guo is with College of Mathematics and Econometrics, Hunan University, Changsha, Hunan, China, E-mail: zyguo@hnu.edu.cn. Yiran Chen is with Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA, E-mail: yiran.chen@duke.edu. Tingwen Huang is with Science Program, Texas A & M University at Qatar, E-mail: tingwen.huang@qatar.tamu.edu.*
- *This work was supported by the Natural Science Foundation of China under Grant 61673187. This publication was made possible by NPRP grants: NPRP 9-466-1-103 from Qatar National Research Fund. The statements made herein are solely the responsibility of the author. (Corresponding author: Yin Yang.)*

algorithm to recognize negligible objects like ball than to identify large objects like person. In order to relate image data to corresponding label, we propose to map image and its label to the same latent space. In latent space, we can explicitly model the label correlation information.

In this paper we propose a novel solution that captures the above intuitions, and combines both direct label recognition with image feature extraction, as illustrated in Figure 1c. Specifically, a ConvNet pipeline extracts features from the input image, which are fed to a fully connected layer for direct label recognition. Meanwhile, these image features, as well as the labels associated with the image, are projected to a common vector space through embedding. There is a certain correlation among labels that often appear in the same image. Therefore, in this latent space, we require that (i) the projection of image features should be close to those of the associated labels, as well as features from images associated with the correlated labels and (ii) conversely, the projected image features should be far apart from labels that are not associated with image. These requirements are enforced through our well designed loss function, which also includes classification loss of the final predictions. In our implementation, the final prediction is simply the sum of the direct predictions and the feature learned from latent space.

A naive approach for multi-label classification is to construct a binary classifier for each label [7], which disregards the correlation among labels completely. Similarly, methods based on region proposals, e.g., HCP [8] improves the accuracy of direct label recognition by focusing on relevant image patches; yet, this method fails to capture label semantics information. A refined solution by Wang et al [2] applies visual attention to model spatial and semantic correlations between labels. None of these methods, however, ignore that exploit label correlation. In our implementation, we use the plain-old ConvNet [9] for direct label recognition; the above techniques could potentially further enhance the accuracy of our model.

Among methods that aim to model the label dependencies, earlier attempts mainly focus on utilizing label correlations (e.g., [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]) as auxiliary information. One problem with this idea is that it fails to capture *visual correlation*: for instance, the label "person" by itself is not strongly correlated with "tennis ball", but the specific visual features of the person (e.g., his pose and attire) in Figure 1b do suggest the presence of a tennis ball. Recent work by Yeh et al. [30] performs label embedding through an autoencoder, and additionally projects ConvNet features to the embedding space through Canonical Correlation Analysis. This approach, however, does not contain a direct label recognition module. Lastly, another line of work applies recurrent networks, e.g., [4], [31], which recognizes labels *sequentially*, e.g., first a person, then a tennis racket, and thirdly a tennis ball. Earlier labels then provide context for later ones. Intuitively, humans normally do not identify objects or attributes sequentially, except for solving puzzles. Instead, we construct a holistic mental picture of the image context, as in the proposed solution.

We have experimentally evaluated the proposed solution on the popular COCO [32] and PASCAL VOC [33] benchmark datasets. The result demonstrate that our solution consistently and significantly outperforms existing methods on various metrics, including mean average precision (mAP), micro / macro precision, recall, and F-measure. Finally, our solution is highly efficient at test time, since it only introduces $2048 \times C$ additional weights to the base ConvNet model, where C is the number of possible labels.

2 Related Work

Multi-label classification is a fundamental problem in machine learning, with a wide range of applications in computer vision, text topic categorization, music retrieval, and gene analysis. One strategy to approach multi-label classification is to transform the problem to multiple single-label classification tasks (e.g., [34], [35], [36]), which can be either binary of multi-class. Those methods can be categorized as first-order strategy and ignore correlation among labels. There are second-order strategy [36], [37], [38], [39] and high-order strategy methods [40], [41], [42], [43]. Other methods adapt single-label classifiers such as decision trees [44], boosting [45], K-nearest neighbors [46] and neural networks [47]. These methods, however, are not designed for large-scale image classification problems and fail to exploit label correlation.

In addition, other researchers proposed to relate image features and label domain data in a latent space and learn label correlation in latent space. To achieve this, C2AE [30] introduces DNN architecture to canonical correlation analysis and autoencoder model. C2AE [30] builds the embedding space through an autoencoder on the labels, and then projects the image features extracted by a ConvNet to the same latent space, via Canonical Correlation Analysis (CCA). C2AE also lacks a direct label recognition module with the assumption that the number of labels associated with an image is known in advance. Further, methods based on embedding also used in image retrieval [48], visual-semantic embedding [49] and neural language task [50]. However, what different with this embedding method is that relation learning is considered in our solution by our designed ranking loss. The idea of modeling context by constructing a latent vector space for labels and image features has also been explored in previous methods, e.g., using SVD [51], compressed sensing [52] and SLEEC [1]. A common problem with these earlier approaches is that they lack a modern, ConvNet-based direct label recognition module. As explained in Section 1, not all labels can be inferred from the context (such as "chair" in Figure 1), and direct recognition is necessary for such labels.

Deep learning provides a new feasibility solution for large-scale image multi-label classification. Most deep learning methods designed CNN-RNN architecture to solve multi-label classification by learning semantic information or capturing global dependencies among learned features [3], [4], [31], [53]. HCP [8] follows an object detection pipeline that generates region proposals, and applies a classifier to each region proposal for multi-label classification. WSD [54] proposed to improve multi-label classification performance by distilling knowledge from weakly-supervised detection task without bounding box. SRN [2] used spatial regularization learning attention maps for multi-label recognition. To further exploit label correlation information, DDPP [55] proposed DPP module to capture label-correlations while incorporate external knowledge about label co-occurrence. CorrLog [56] explicitly modeled the pairwise correlation between labels and improved the performance of multi-label recognition. Further, CGL [57] modeled formulate multi-label problem as conditional graphical lasso inference problem and focused on image feature when exploiting label correlation. Therefore, label correlation becomes hot topic for multi-label problem.

To summarize, previous methods, to our knowledge, miss either explicit context construction, or a ConvNet-based direct label recognition module; meanwhile, many of them require the knowledge of number of labels associated with the image. The proposed solution, presented next, combines both context and direct recognition, and can identify an arbitrary number of labels from an image.

3 Proposed Solution

The proposed solution contains three main components: a feature extractor, a feature / label co-projector that map both image features and labels to the same latent vector space, and a classifier that combines direct label recognition results using the feature extractor with contextual information extracted from latent vector space. Figure 2 shows the overall architecture of the proposed framework.

The feature extractor extracts visual features from the input image, which can be performed with a standard ConvNet pipeline commonly used for single-label image recognition tasks. These features can be viewed as abstract representations of visual contents in the image. From these features, we can build a direct label recognizer for each label, e.g., with a fully-connected layer on top of the visual features. In addition, features from deeper neural network layer have richer semantic information and are more abstract.

The feature / label co-projector is responsible for embedding image convolutional features and corresponding label, respectively, as explained in Section 1. Both projector can be viewed as encoder. The feature / label co-projector takes feature extractor's features and labels as inputs, respectively. Specifically, the projector component maps both visual features and one-hot-encoded labels to the same latent space. In the latent space, we can explicitly model label correlation. Then, metric learning method is used to force the distance between correlated embedding vectors from image feature and label are small than non-correlated ones. Meanwhile, our well designed constrained ranking loss ensures that the mapping correctly reflects the semantic relationships between images and labels. Finally, we extract feature of image feature embedding network as part of feature for label prediction.

Lastly, the classifier combines direct label recognition results (one confidence value per class) with the image context from this latent mapping. In our implementation, the combination is an element-wise sum for simplicity. The whole model can be trained end-to-end with no additional data other than the images and ground truth labels in the training set.

3.1 Feature Extractor

As explained earlier, the feature extractor can be done with any standard ConvNet pipeline for single-label classification. Our implementation employs ResNet-101 [9], which achieves competitive performance (7.1 top-5 error) on the ImageNet dataset. We remove the last pooling layer and the last classification layer and use the features map from last convolution layer, as the inputs for our classification and embedding branches.

Formally, let $D = \{x_1, x_2, ..., x_i, ..., x_n\}^{d*n}$ denotes the set of images with corresponding label $Y = \{y_1, y_2, ..., y_i, ..., y_n\}$, $y \in \{0,1\}^{C*n}$, where y_i is a C dimension label vector for image x_i. Meanwhile, let d, C and n denote the image data dimension, total label number and image dataset size respectively; y_{il} is $+1$ when x_i has the $l-$th label, and 0 otherwise. We feed the image x to the feature extractor f_{cnn} to get the image features F_x:

$$F_x = f_{cnn}\left(x; \theta_{cnn}\right), F_x \in R^{14*14*2048} \tag{1}$$

3.2 Feature / Label Embedding

The embedding components of our solution captures the correlations between the image and its labels, as well as between different labels and features from different images. For this purpose, we design two mapping networks that embed visual features and labels to the common latent space, respectively. The projections of features and labels in this space are then adjusted through back-propagation, using the proposed constrained ranking loss function, detailed later in Section 4.

Following common embedding network designs, we design a convolution network for projecting visual features from the feature extractor, and another pipeline consisting of fully-connected layers for projecting one-hot-encoded labels. In general, our framework can work with any such projection pipelines, and our specific implementation is detailed later in Section 5.1. In particular, for image feature projection, we firstly use a convolution layer f_{conv} to map the image feature F_x. The role of the convolutional layer is to turn image features $F_x \in R^{14\times14\times2048}$ into a form $F_e \in R^{14\times14\times C}$ that is easier to optimize and understand. Each channel of F_e represents the corresponding object class feature. If the label is included in the image, the corresponding channel has a larger activation.

Formally, let f_{im} and f_l denote the convolution network (for image feature projection)) and the fully connect networks (for label projection), respectively. We can get the embedding representation F_e and L_e as follows:

$$F_e = f_{im}\left(f_{conv}\left(F_x; \theta_{conv}\right); \theta_{im}\right), F_e \in R^C \tag{2}$$

$$F_{conv} = f_{conv}\left(F_x; \theta_{conv}\right), F_{conv} \in R^{14*14*C} \tag{3}$$

$$L_e = f_l\left(y, \theta_l\right), L_e \in R^C \tag{4}$$

During training, the projected vectors are adjusted through the proposed constrained ranking loss function, elaborated in Section 4. Intuitively, in the latent space, we aim to move the projections of image closer to the projections of its associated labels (which we call positive labels), and away from the projections of labels not associated with the image (negative labels). Meanwhile, labels that are semantically correlated are moved close together through the training process, so are semantically correlated image features.

Lastly, we fuse the features from the image and the mapping network to calculate the final prediction. Specifically, in the main classification module, we use the fusion of global max pooling and global average pooling operation to reduce the dimension of image feature, and a fully connect layer is followed to compute the initial prediction. We add them together to get the final predicted confidence $P \in R^{N*C}$. Max pooling can find the activation of small objects in image, but average pooling can find the activation of bigger objects. The fusion of both pooling is helpful to find all labeled object. For our channel-wise pooling, global max pooling is employed.

$$P = f_{pool}\left(f_{conv}\left(F_x;\theta_{conv}\right)\right) + f_c\left(f_{pool}\left(F_x\right);\theta_c\right) \tag{5}$$

4 Loss Function

4.1 Multi-Label Soft Margin

In order to optimize our proposed framework, we use the Multi Label Soft Margin classification loss and constrained ranking loss as our loss function. Firstly, multi-label can be viewed as a one-to-many classification problem between image and its labels. Note that we assume the general setting where the number of labels corresponding to each image is unknown. Previous works such as [58] incorporate a label decision module into the model, which estimate the optimal confidence thresholds for each visual concept. The Multi Label Soft Margin chooses 0 as the label thresholds instead of estimating the label thresholds. This makes it easier to optimize and more stable.

Specifically, our Multi Label Soft Margin creates a criterion that optimizes a multi-label one-vs-all loss based on cross entropy between inputs X and the ground truth Y:

$$\begin{aligned} Loss(\mathrm{x},\mathrm{y}) = & -\sum_i y_i * \log((1+e^{-F(x_i)})^{-1}) \\ & + (1-y_i) * \log(\tfrac{e^{-F(x_i)}}{1+e^{-F(x_i)}}) \end{aligned} \tag{6}$$

where F denotes the mapping for image x to label y. Ideally, F should have $F(x_i) = y_i$ for i in range N. Since the Multi Label Soft Margin loss is based on cross entropy, it cannot capture the label dependency in multi-label mission.

4.2 The Constrained Ranking Loss

To exploit feature / label correlations, we design a constrained ranking loss to capture the label dependency. The ranking loss has been studied in the the pre-deep-learning multi-label classification setting, such as SVM [37]. The ranking loss mining multi-label data is computed in [7], where the ranking loss averages over the samples, and the number of label pairs are incorrectly ordered, such as true labels have a lower score than false labels. And the lowest achievable ranking loss is 0. The ranking loss used in this method indicates the number of irrelevant labels that are higher than the relevant labels. However, not all the labels are considered simultaneously; instead, only the incorrectly ranked labels are considered. In fact, the label correlation is naturally local where the subsets of images share the correlation rather than all image instances. Huang, et al. measure the similarity between image instances in the label space rather than the feature space because the image instances with the same label share the same correlation [59].

Meanwhile, researchers map the label into a low dimension or high dimension latent space [60], [61] to solve multi-label classification. All these methods can be viewed as label embedding. In the latent space, the correlation between labels can be implicitly exploited. In our proposed solution, we use a deep convolution network $U : \mathbb{R}^{\mathbb{HW}} \to \mathbb{R}^{\mathbb{C}}$ to map the image feature maps to a latent space and a fully connected network $V : \mathbb{R}^{\mathbb{C}} \to \mathbb{R}^{\mathbb{C}}$ to map the corresponding labels to the same latent space. H and W point at the height and width of corresponding image. Let $U(f)$ denote the embedded image features and $V(y)$ for embedded labels. Furthermore, we design a constrained ranking loss to measure the similarity between embedded images and labels. We consider all positive label and the negative label simultaneously. In the embedding space, let $d(f_i^+, y_j^+)$ denote the distance between embedded positive features and embedded positive labels. And let $d(f_i^+, y_k^-)$ denote the distance between embedded positive features and the embedded negative labels. The y_i^+ and y_k^- denotes the embedded positive labels and negative labels. We expect the distance $d(f_i^+, y_j^+)$ to be smaller than the distance $d(f_i^+, y_k^-)$, with a large margin of δ which is set as 0.5 here. This leads to the following formulation:

$$\begin{aligned} & d\left(f_i^+\ , y_j^+\right) + \delta \leq d\left(f_i^+\ , y_k^-\right) \\ & \forall y_j^+ \in Y^+, \forall y_k^- \in Y^- \end{aligned} \tag{7}$$

In our solution ,F_e and L_e which introduced in section 3.2 3.2 is point at f_i and y_i respectively. Here, $d(f, y)$ denotes the Euclidean distance between image features and label features. Intuitively, in the same latent space, the positive features and corresponding labels have the similar embedding and have large margin with negative labels.

We also define the constraints for the label side.

$$\begin{aligned} & d\left(y_i^+\ , y_j^+\right) + \delta \leq d\left(y_i^+\ , y_k^-\right) \\ & \forall y_j^+ \in Y^+, \forall y_k^- \in Y^- \end{aligned} \tag{8}$$

These constraints ensure that the embedded positive labels are as close as possible with each other, and as far away as possible from the embedded negative labels. We then add the constraints terms corresponding to our baseline ranking loss function:

$$\begin{aligned} Loss_r = & \lambda_1 * \sum_{i,j,k} \left[\delta + d\left(f_i^+\ , y_j^+\right) - d\left(f_i^+\ , y_k^-\right)\right]_+ \\ & + \lambda_2 * \sum_{i,j,k} \left[\delta + d\left(y_i^+\ , y_j^+\right) - d\left(y_i^+\ , y_k^-\right)\right]_+ \end{aligned} \tag{9}$$

where λ_1 and λ_2 are the hyperparameters to balance the ranking loss. We set both to 0.5.

Our constrained ranking loss can measure the similarity between the embedded labels and features. We simultaneously consider all the ranked labels, because minimizing the above loss function is equivalent to maximizing the predicted value of all positive label attribute pairs while

TABLE 1
Comparison results of average precision and mAP of other methods and our method on the MSCOCO dataset. The red front is used to mark the best results.

	ALL							TOP-3					
Methods	MAP	F1-C	P-C	R-C	F1-O	P-O	R-O	F1-C	P-C	R-C	F1-O	P-O	R-O
WARP	-	-	-	-	-	-	-	55.7	59.3	52.5	60.7	59.8	61.4
CNN-RNN	-	-	-	-	-	-	-	60.4	66.0	55.6	67.8	69.2	66.4
RDAR	-	-	-	-	-	-	-	67.4	79.1	58.7	72.0	84.0	63.0
RARL	-	-	-	-	-	-	-	66.2	78.8	57.2	71.1	84.0	61.6
VGG	67.8	63.3	72.0	56.4	68.9	76.8	62.4	60.4	75.1	50.5	66.4	81.5	66.0
Ours(VGG)	72.9	68.8	75.5	63.1	73.3	79.5	67.9	65.8	80.0	55.9	70.6	85.8	60.0
ResNet101	75.2	69.5	80.8	63.4	74.4	82.2	68.0	65.9	84.3	57.4	71.7	86.5	61.3
ResNet-SRN	77.1	71.2	81.6	65.4	75.8	82.7	69.9	67.4	85.2	58.8	72.9	87.4	62.5
Ours(Resnet)	81.1	75.8	81.2	70.8	78.1	83.6	73.3	72.7	86.4	62.9	75.1	88.7	65.1

TABLE 2
Comparison of average precision and mAP of other methods and our method on VOC dataset. The best evaluation value is highlighted in red front.

Methods	Aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	MAP
CNN-SVM	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	87.2	71.8	73.9
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
VeryDeep	98.9	95.0	96.8	95.4	69.7	90.4	93.5	96.0	74.2	86.6	87.8	96.0	96.3	93.1	97.2	70.0	92.1	80.3	98.1	87.0	89.7
RLSD	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
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
FeV+LV	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
RDAR	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
RARL	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.3	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92.0
Ours	99.9	98.4	97.8	98.8	81.2	93.7	97.1	98.4	82.7	94.6	87.1	98.1	97.6	96.2	98.8	83.2	96.2	84.7	99.1	93.5	93.8

TABLE 3
Architecture of the image feature projection network in our implementation

Output Size	Layer
$14 \times 14 \times C$	conv, $(1 \times 1, C, 1, 1)$
$7 \times 7 \times C$	conv, $(3 \times 3, C, 2, 1)$
$7 \times 7 \times C/4$	conv, $(3 \times 3, C/4, 2, 1)$
$7 \times 7 \times C$	conv, $(3 \times 3, C, 1, 1)$
$1 \times C$	maxpooling

minimizing the predicted value of all negative label attribute pair, which implicitly forces the label co-occurrence information to be retained. Moreover, the positive and negative labels will be gathered together respectively in the latent space. Therefore, the local label dependency can be implicitly exploited. If other losses such as common ranking loss, cross entropy loss or the mean square error loss are considered, the local label correlation cannot be modeled and exploited.

The loss function is the sum of classification loss and the constrained ranking loss. It is shown as follows:

$$Loss = \alpha * Loss_{cls} + \beta * Loss_{r} \tag{10}$$

where α and β are the hyperparameters, we simply set both of them as 1.

5 Experiments

We have implemented the proposed solution and evaluated it on two popular benchmark datasets: PASCAL VOC 2007 [33], which contains 20 different object labels, and MS COCO 2014 [32], which contains 80 different object labels. We also compare our results with the those reported in previous research papers. In the following, we present the implementation of the proposed solution and the model training process, evaluation metrics, evaluation results, and result visualizations.

5.1 Model Implementation and Training

The proposed solution is implemented using PyTorch (available at pytorch.org). As shown in Figure 2, the feature extractor of our model is implemented using ResNet-101 [9], pre-trained using the ImageNet dataset [62]. Specifically, we removed the last two layers (i.e., global average pooling and 1000-way classification full-connected, respectively), and added instead (i) a new global max pooling layer and (ii) C-way fully-connected layers, where C denotes the number of object categories, which is 20 and 80 in PASCAL VOC and MS COCO datasets, respectively.

We set the size of each input image to 448×448. Then, after the ResNet-101 pipeline, the extracted feature maps (i.e., before the pooling layer) has size $14 \times 14 \times 2048$. These features are fed to the feature/label co-projector branch, which uses a small ConvNet to embed these features to a latent vector space. Table 3 lists the detailed layers of this neural net for image feature projection.

Regarding label projection, we use two fully-connected layers to to embed one-hot-encoded label vectors to the same latent vector space as the image features, as shown in Table 4. Then, the proposed ranking loss is used to model the correlation between embedded labels and image features. Finally, we obtain the final prediction results by aggregating the outputs the direct label recognition (i.e., ResNet-101) and feature/label co-projector branches as shown in Figure 2. The specific aggregation in our implementation is a simple element-wise sum.

At test time, the feature/label co-projection module no longer applies, since the label for a test image is unknown. Hence, we simply remove the network layers that project image features and labels to a common latent space. Note that at test time, compared with our base model, i.e.,

TABLE 4
Architecture of the label projection network in our implementation

Output Size	Layer
$C/2$	fc, $(C, C/2)$
$C/4$	fc, $(C/2, C/4)$
$C/2$	fc, $(C/4, C/2)$
C	fc, $(C/2, C)$

ResNet-101, the proposed solution only introduces $2048 \times C$ additional weights. Hence, the proposed model is highly efficient; yet it achieves state of the art performance as shown in later subsections.

Model training. The proposed deep neural network is trained end-to-end with the training set of the data and no additional information. To demonstrate the robustness of the proposed solution, we used simple training techniques without much hyperparameter tuning. Specifically, during training, we simply re-size each raw input images from the dataset to 448×448, with no other data augmentation. The training steps are performed by a SGD optimizer, with momentum 0.9 and weight decay 1e-4, respectively. We used different learning rates for different network layers. In particular, we set the learning rate of features extraction layers (i.e., ResNet-101) to 0.001, and the learning rate of the other layers as 0.01. The reason is that the ResNet-101 layers have already been pre-trained on ImageNet data, and using a small learning rate is necessary for transfer learning.

5.2 Evaluation Metrics

Following a recent paper [2], we evaluate the proposed solution using 7 metrics for multi-label classification performance: mean average precision (mAP), macro/micro precision (P-C/P-O), macro/micro recall (R-C/R-O) and macro/micro F measure ($F1-C/F1-O$). Specifically, mAP is the mean value of average precision [63] for each class, where average precision is calculated by the average fraction of relevant labels ranked higher than one other relevant label. Macro precision (denoted as P-C) is evaluated by averaging per-class precision measurements. Micro precision (P-O) is an overall measure that counts true predictions for all images over all classes. Formally, they are defined as follows:

$$PO = \frac{\sum_{i}^{C} TP_i}{\sum_{i}^{C} (TP_i + FP_i)}, PC = \frac{1}{C} \sum_{i}^{C} \frac{TP_i^C}{TP_i^C + FP_i^C} \quad (11)$$

where TP is the number of true positives and FP the number of false positives for each class, respectively. The recall and F1-score metrics are defined as:

$$RO = \frac{\sum_{i}^{C} TP_i}{\sum_{i}^{C} (TP_i + FN_i)}, RC = \frac{1}{C} \sum_{i}^{C} \frac{TP_i^C}{TP_i^C + FN_i^C} \quad (12)$$

$$F1O = \frac{2 * (PO * RO)}{PO + RO}, F1C = \frac{2 * (PC * RC)}{PC + RC} \quad (13)$$

where FN denotes the number of false negatives for each class. The F score can be viewed as a weighted average of the precision and recall. For $F1$, the precision and recall have the same weight. All 7 evaluation metrics used in the experiments have range between 0 and 1, with higher values indicating better performance.

5.3 Evaluation Results

We compare the proposed solution against previous multi-label image classification methods on MS COCO 2014 [32] and PASCAL VOC 2007 benchmark datasets [33]. The results are shown in Tables 1 (for COCO) and 2 (for VOC), respectively.

Specifically, on the MS COCO dataset, we compare our solution against reported results (directly from their respective papers) for WARP [64], CNN-RNN [4], RDAR [31], RARL [3], and SRN [2]. Our results in Table 1 also includes the performance of the base model of our solution, i.e., ResNet-101. Note that some methods require the knowledge of the number of labels associated with an image; consequently, they cannot predict the set of *all* labels for a given image. Therefore, we also include the results for top-3 labels.

Clearly, the proposed solution outperforms its base model ResNet-101 on all evaluation metrics. We observe that the base model is in fact a strong baseline, which, by itself, outperforms several earlier approaches. More importantly, with two exceptions (P-C for all labels and R-O for top-3 labels), the proposed solution achieves the best performance on all evaluation metrics, usually with significant performance gaps. Notably, our mAP is 81.1%, compared to the previous best 77.1% obtained by a recent work SRN [2]; similarly, our F1 scores are also several percentage points higher than the best previous results. Hence, these evaluation results firmly establish the proposed model as the new state of the art for multi-label classification on MS COCO.

Another evaluation dataset used in the experiments, i.e., PASCAL VOC 2007, contains 9963 images of 20 different object categories, split into a training set of 5011 images and a validation set of 4952 images. On this dataset, we compare the result of our solution against the reported results (again directly from their respective papers) of the following methods: CNN-SVM [65], CNN-RNN [4], VeryDeep [66], RLSD [67], HCP [68], RDAR [31] and RARL [3]. The ressults, as shown in Table 2, list the average precision for each of the 20 classes, as well as the mAP score. In terms of overall mAP, our method significantly outperforms the previous best result obtained by RARL [31]. Note that RARL involves a complicated network architecture involving ConvNet, RNN, and attention, whereas the proposed method has a much simpler architecture, and much fewer weights at test time. Finally, for specific classes, our method achieves the highest average precision for the majority of the classes; for the remaining classes, the performance of our model is also highly competitive. Therefore, we achieve a new state of the art on PASCAL VOC 2007.

5.4 Ablation Experiments

To evaluate our model, we decompose our deep neural network and valid the effect of image/label co-projector in COCO dataset. Ablation for backbone: In our experiments, we use Resnet101 as backbone of our model following SRN. We can easily know from Tables 1 and 2 that we achieve

TABLE 5
Experiment results about the effect of extra branch for our model. TBA and TBC denote results of additional branch are directly added to the classification results of the main classification branch, and features of two branches are cascaded to each other, respectively. Resnet baseline comes from SRN.

	MAP	F1-C	P-C	F1-O	F1-O	P-O	R-O
ResNet101	75.2	69.5	80.8	63.4	74.4	82.2	68.0
TBA	73.0	68.2	76.9	61.3	73.3	81.4	66.6
TBC	75.5	70.7	77.6	65.0	75.1	81.3	69.8
Ours	81.1	75.8	81.2	70.8	78.1	83.6	73.3

TABLE 6
Experiment results about the effect of constrained ranking loss. MSE denotes mean square error loss. CRL is our proposed constrained ranking loss.

	MAP	F1-C	P-C	F1-O	F1-O	P-O	R-O
Ours(MSE)	79.2	74.0	81.4	67.8	77.7	85.3	71.2
Ours(CRL)	81.1	75.8	81.2	70.8	78.1	83.6	73.3

competitive performance compared with other great methods. Further, in order to rule out the impact of additional extra branch on our model, we discard the feature map part of the model and leave the rest to complete the experiment. We consider the case where the results of additional branch are directly added to the classification results of the main classification branch, or the features of two branches are cascaded to each other. Corresponding results are shown in Table 5. Ablation experiments were done on the COCO dataset.

In addition, we also compared proposed constrained ranking loss with mean square error loss. Experiments results are shown in Table 6. We can easily know that constrained ranking loss is working well in our method.

5.5 Visualizations

CAM Visualizations. To provide further insights into the proposed solution, we visualize the features extracted by the feature extractor of our method using CAM [69], which shows the attention map of of these features. Meanwhile, to demonstrate that our feature/label co-projection module correctly learns the correlations between image features and different labels, we also visualize the features from the $f_{conv}(F_x)$ layer in the feature/label co-projector. Figure 3 shows the visualization results on the MS COCO dataset. From these results, we observe that the label correlation is well represented in the features $f_{conv}(F_x)$ with the help of label embedding and our constrained ranking loss. The results on PASCAL VOC lead to similar conclusions, and are omitted for brevity.

6 Conclusion

We present a simple and intuitive solution to the fundamental problem of multi-label image recognition, which combines direct label recognition using a base model (ResNet-101 in our implementation) and a feature/label co-projection module that explicitly models the context of the image. Our implementation of the proposed method achieves state of the art performance on two popular benchmark dataset: MS COCO and PASCAL VOC, while being highly efficient at test time, with only $2048 \times C$ additional weights compared to the base model, where C is the number of possible classes.

As future work, we plan to further improve the proposed solution using effective techniques such as visual attention. Meanwhile, we intend to investigate multi-label classification in other contexts, e.g., with abstract attribute labels, and for other types of challenging data such as video.

References

[1] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, “Sparse local embeddings for extreme multi-label classification,” in *Advances in Neural Information Processing Systems*, 2015, pp. 730–738.

[2] 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.

[3] T. Chen, Z. Wang, G. Li, and L. Lin, “Recurrent attentional reinforcement learning for multi-label image recognition,” *arXiv preprint arXiv:1712.07465*, 2017.

[4] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, “Cnn-rnn: A unified framework for multi-label image classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294.

[5] Y. Li, C. Huang, C. C. Loy, and X. Tang, “Human attribute recognition by deep hierarchical contexts,” in *European Conference on Computer Vision*. Springer, 2016, pp. 684–700.

[6] Z. Yan, W. Liu, S. Wen, and Y. Yang, “Multi-label image classification by feature attention network,” *IEEE Access*, vol. 7, pp. 98 005–98 013, 2019.

[7] G. Tsoumakas, I. Katakis, and I. Vlahavas, “Mining multi-label data,” in *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 667–685.

[8] 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 transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1901–1907, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] B. Hariharan, L. Zelnik-Manor, M. Varma, and S. Vishwanathan, “Large scale max-margin multi-label classification with priors,” in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. Citeseer, 2010, pp. 423–430.

[11] G. Tsoumakas, A. Dimou, E. Spyromitros, V. Mezaris, I. Kompatsiaris, and I. Vlahavas, “Correlation-based pruning of stacked binary relevance models for multi-label learning,” in *Proceedings of the 1st International Workshop on Learning from Multi-label Data*, 2009, pp. 101–116.

[12] X. Shu, J. Tang, G.-J. Qi, Z. Li, Y.-G. Jiang, and S. Yan, “Image classification with tailored fine-grained dictionaries,” *IEEE transactions on circuits and systems for video technology*, vol. 28, no. 2, pp. 454–467, 2018.

[13] J. Tang, X. Shu, G.-J. Qi, Z. Li, M. Wang, S. Yan, and R. Jain, “Tri-clustered tensor completion for social-aware image tag refinement,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1662–1674, 2017.

[14] X. Zeng, S. Wen, Z. Zeng, and T. Huang, “Design of memristor-based image convolution calculation in convolutional neural network,” *Neural Computing and Applications*, vol. 30, pp. 502–508, 2018.

[15] S. Wen, R. Hu, Y. Yang, T. Huang, Z. Zeng, and Y. Song, “Memristor-based echo state network with online least mean square,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 9, pp. 1787–1796, 2019.

[16] S. Wen, X. Xie, Z. Yan, T. Huang, and Z. Zeng, “General memristor with applications in multilayer neural networks,” *Neural Networks*, vol. 103, pp. 142–148, 2018.

[17] S. Wen, H. Wei, Y. Yang, Z. Guo, Z. Zeng, T. Huang, and Y. Chen, “Memristive LSTM networks for sentiment analysis,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–10, 2019.

[18] Y. Cao, S. Wang, Z. Guo, T. Huang, and S. Wen, “Synchronization of memristive neural networks with leakage delay and parameters mismatch via event-triggered control,” *Neural Networks*, vol. 119, pp. 178–189, 2019.

[19] S. Wang, Y. Cao, T. Huang, Y. Chen, P. Li, and S. Wen, "Sliding mode control of neural networks via continuous or periodic sampling event-triggering algorithm," *Neural Networks*, vol. 0, pp. 1–11, 2019.

[20] S. Wen, H. Wei, Z. Yan, Z. Guo, Y. Yang, T. Huang, and Y. Chen, "Memristor-based design of sparse compact convolutional neural networks," *IEEE Transactions on Network Science and Engineering*, vol. 99, pp. 1–11, 2019.

[21] S. Wen, M. Z. Chen, X. Yu, Z. Zeng, and T. Huang, "Fuzzy control for uncertain vehicle active suspension systems via dynamic sliding-mode approach," *IEEE Transactions on Systems, Man and Cybernetics: Systems*, vol. 47, pp. 24–32, 2017.

[22] S. Wen, T. Huang, X. Yu, M. Z. Chen, and Z. Zeng, "Aperiodic sampled-data sliding-mode control of fuzzy systems with communication delays via the event-triggered method," *IEEE Transactions on Fuzzy Systems*, vol. 24, pp. 1048–1057, 2016.

[23] M. Dong, S. Wen, Z. Zeng, Z. Yan, and T. Huang, "Sparse fully convolutional network for face labeling," *Neurocomputing*, vol. 331, pp. 465–472, 2019.

[24] S. Wang, Y. Cao, T. Huang, and S. Wen, "Passivity and passification of memristive neural networks with leakage term and time-varying delays," *Applied Mathematics and Computation*, vol. 361, pp. 294–310, 2019.

[25] G. Ren, Y. Cao, S. Wen, Z. Zeng, and T. Huang, "A modified elman neural network with a new learning rate," *Neurocomputing*, vol. 286, pp. 11–18, 2018.

[26] S. Wen, W. Liu, Y. Yang, Z. Zeng, and T. Huang, "Generating realistic videos from keyframes with concatenated gans," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 2337–2348, 2019.

[27] Y. Cao, Y. Cao, S. Wen, Z. Zeng, and T. Huang, "Passivity analysis of reaction-diffusion memristor-based neural networks with and without time-varying delays," *Neural Networks*, vol. 109, pp. 159–167, 2019.

[28] X. Xie, S. Wen, Z. Zeng, and T. Huang, "Memristor-based circuit implementation of pulse-coupled neural network with dynamical threshold generator," *Neurocomputing*, vol. 284, pp. 10–16, 2018.

[29] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, and Z. Zeng, "Clu-cnns: Object detection for medical images," *Neurocomputing*, vol. 350, pp. 53–59, 2019.

[30] 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, pp. 2838–2844.

[31] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 464–472.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[33] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2007 (voc 2007) results (2007)," 2008.

[34] J. Read, L. Martino, and D. Luengo, "Efficient monte carlo methods for multi-dimensional learning with classifier chains," *Pattern Recognition*, vol. 47, no. 3, pp. 1535–1546, 2014.

[35] J. Read, L. Martino, P. M. Olmos, and D. Luengo, "Scalable multi-output label prediction: From classifier chains to classifier trellises," *Pattern Recognition*, vol. 48, no. 6, pp. 2096–2109, 2015.

[36] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine learning*, vol. 73, no. 2, pp. 133–153, 2008.

[37] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in neural information processing systems*, 2002, pp. 681–687.

[38] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005, pp. 195–200.

[39] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 17–26.

[40] R. Yan, J. Tesic, and J. R. Smith, "Model-shared subspace boosting for multi-label classification," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 834–843.

[41] S. Ji, L. Tang, S. Yu, and J. Ye, "Extracting shared subspace for multi-label classification," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 381–389.

[42] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2004, pp. 22–30.

[43] W. Cheng and E. Hüllermeier, "Combining instance-based learning and logistic regression for multilabel classification," *Machine Learning*, vol. 76, no. 2-3, pp. 211–225, 2009.

[44] Y.-L. Chen, C.-L. Hsu, and S.-C. Chou, "Constructing a multi-valued and multi-labeled decision tree," *Expert Systems with Applications*, vol. 25, no. 2, pp. 199–209, 2003.

[45] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine learning*, vol. 39, no. 2-3, pp. 135–168, 2000.

[46] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[47] ——, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[48] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord, "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings," *arXiv preprint arXiv:1804.11146*, 2018.

[49] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: improved visual-semantic embeddings," *arXiv preprint arXiv:1707.05612*, 2017.

[50] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," *Training*, vol. 720, no. 619-508, p. 2, 2017.

[51] F. Tai and H.-T. Lin, "Multilabel classification with principal label space transformation," *Neural Computation*, vol. 24, no. 9, pp. 2508–2542, 2012.

[52] D. J. Hsu, S. M. Kakade, J. Langford, and T. Zhang, "Multi-label prediction via compressed sensing," in *Advances in neural information processing systems*, 2009, pp. 772–780.

[53] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "Semantic regularisation for recurrent image annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2872–2880.

[54] Y. Liu, L. Sheng, J. Shao, J. Yan, S. Xiang, and C. Pan, "Multi-label image classification via knowledge distillation from weakly-supervised detection," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 700–708.

[55] P. Xie, R. Salakhutdinov, L. Mou, and E. P. Xing, "Deep determinantal point process for large-scale multi-label classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 473–482.

[56] Q. Li, B. Xie, J. You, W. Bian, and D. Tao, "Correlated logistic model with elastic net regularization for multilabel image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3801–3813, 2016.

[57] Q. Li, M. Qiao, W. Bian, and D. Tao, "Conditional graphical lasso for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2977–2986.

[58] Y. Li, Y. Song, and J. Luo, "Improving pairwise ranking for multi-label image classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[59] S.-J. Huang, Z.-H. Zhou, and Z. Zhou, "Multi-label learning by exploiting label correlations locally." in *AAAI*, 2012, pp. 949–955.

[60] X. Li and Y. Guo, "Multi-label classification with feature-aware non-linear label space transformation." in *IJCAI*, vol. 2015, 2015, pp. 3635–3642.

[61] C.-S. Ferng and H.-T. Lin, "Multilabel classification using error-correcting codes of hard or soft bits," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 11, pp. 1888–1900, 2013.

[62] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 248–255.

[63] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," *arXiv preprint arXiv:1609.00288*, 2016.

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

[65] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: an astounding baseline for recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[66] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[67] J. Zhang, Q. Wu, C. Shen, J. Zhang, and J. Lu, “Multi-label image classification with regional latent semantic dependencies,” *IEEE Transactions on Multimedia*, 2018.

[68] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 280–288.

[69] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization.” *CVPR*, 2016.

Shiping Wen received the M. Eng. degree in Control Science and Engineering, from School of Automation , Wuhan University of Technology, Wuhan, China, in 2010, and received the Ph.D degree in Control Science and Engineering, from School of Automation, Huazhong University of Science and Technology, Wuhan, China, in 2013. He is currently a Professor at School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include memristor-based circuits and systems, neural networks, and deep learning.

Weiwei Liu received his B. Eng. degree from Information Engineering school, Henan University of Science and Technology, Wuhan, China in 2017. He is currently working towards the M. Eng. degree from Huazhong University of Science and Technology, Wuhan, China. His research interests include computer version, multilabel classification, generative adversarial network and deep learning.

Yin Yang is currently an Assistant Professor in the College of Science and Engineering, Hamad Bin Khalifa University. His main research interests include cloud computing, database security and privacy, and query optimization. He has published extensively in top venues on differentially private data publication and analysis, and on query authentication in outsourced databases. He is now working actively on cloud-based bigdata analytics, with a focus on fast streaming data.

Zheng Yan received the B.Eng. degree in automation and computer-aided engineering and the Ph.D. degree in mechanical and automation engineering from The Chinese University of Hong Kong, Hong Kong, in 2010 and 2014, respectively. Dr. Yan was a recipient of the Graduate Research Grant from the IEEE Computational Intelligence Society in 2014.

Zhenyuan Guo received the B.S. degree in mathematics and applied mathematics and the Ph.D. degree in applied mathematics from the College of Mathematics and Econometrics, Hunan University, Changsha, China, in 2004 and 2009, respectively.

He was a Joint Ph.D. Student to visit the Department of Applied Mathematics, University of Western Ontario, London, ON, Canada, from 2008 to 2009. From 2013 to 2015, he was a Post-Doctoral Research Fellow with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong. He is currently a Professor with the College of Mathematics and Econometrics, Hunan University. His current research interests include theory of functional differential equations and differential equations with discontinuous right-hands, and their applications to dynamics of neural networks, memristive systems, and control systems.

Pan Zhou (Member, IEEE) received the B.S. degree (Advanced Class) and the M.S. degree from the School of EIC, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2006, and the Ph. D degree from the School of Electrical and Computer Engineering, Georia Institute of Technology (Georgia Tech), Atlanta, GA, USA, in 2011.

He is currently an Associate Professor with School of Cyber Science and Engineering, HUST. He was a senior technical memberat Oracle Inc, America during 2011 to 2013, Boston,MA, USA. He was an Associate Professor with the School of Electronic Information and Communications, HUST, from 2013 to 2019. His current research interest include:security and privacy, machine learning and big data analytics, and informationnetworks.

Yiran Chen received B.S and M.S. from Tsinghua University and Ph.D. from Purdue University in 2005. After five years in industry, he joined University of Pittsburgh in 2010 as Assistant Professor and then promoted to Associate Professor with tenure in 2014, held Bicentennial Alumni Faculty Fellow. He now is a tenured Professor of the Department of Electrical and Computer Engineering at Duke University and serving as the co-director of Duke Center for Evolutionary Intelligence (CEI), focusing on the research of new memory and storage systems, machine learning and neuromorphic computing, and mobile computing systems. Dr. Chen has published one book and more than 300 technical publications and has been granted 93 US patents. He is the associate editor of IEEE TNNLS, IEEE TCAD, IEEE D&T, IEEE ESL, ACM JETC, ACM TCPS, and served on the technical and organization committees of more than 40 international conferences. He received 6 best paper awards and 14 best paper nominations from international conferences. He is the recipient of NSF CAREER award and ACM SIGDA outstanding new faculty award. He is the Fellow of IEEE.

Tingwen Huang is a professor at Texas A & M University-Qatar. He received his B.S. degree from Southwest Normal University (now Southwest University), China, 1990, his M.S. degree from Sichuan University, China, 1993, and his Ph.D. degree from Texas A & M University, College Station, Texas, 2002. After graduated from Texas A & M University, he worked as a Visiting Assistant Professor there. Then he joined Texas A & M University at Qatar (TAMUQ) as an Assistant Professor in August 2003, then he was promoted to Professor in 2013. His research interests include neural networks based computational intelligence, distributed control and optimization, nonlinear dynamics and applications in smart grids. He has published more than three hundred peer-review reputable journal papers, including more than one hundred papers in IEEE Transactions. Currently, he serves as an associate editor for four journals including IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, and Cognitive Computation.

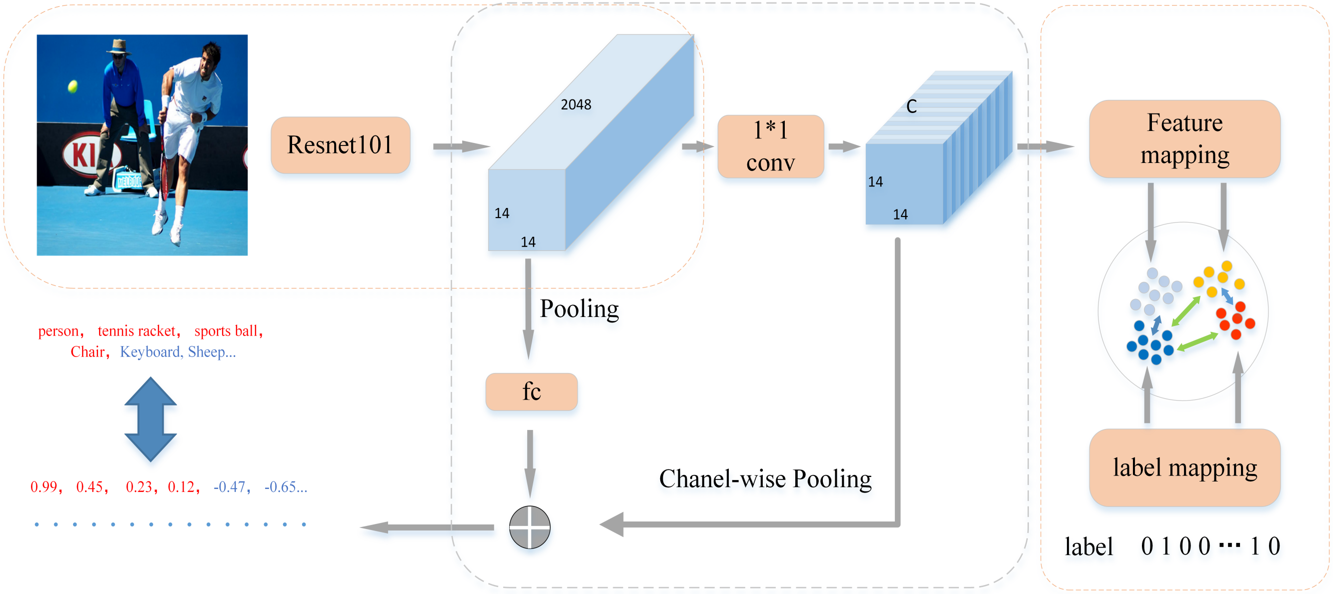

Fig. 2. Overview of the proposed solution for multi-label image classification. Orange squares represent neural network layers, and blue cubes denote the feature maps output by a network pipeline. The proposed network consists of three modules: feature extractor, feature/label co-projector, and classifier. The feature extractor outputs a feature map as a $14 \times 14 \times 2048$ tensor. A subsequent convolution layer then generates the new image features as inputs to the feature/label co-projector, which embed these features and the labels associated with the image to the same latent space, shown as a gray circle. Dots with different colors represent different embedded data: red and blue ones are embedded positive and negative labels, respectively, whereas orange/gray dots are embedded positive/negative image features, respectively. Green (resp., blue) arrows indicate that data should be away from (resp., close to) each other, which are embodied in the proposed loss function.

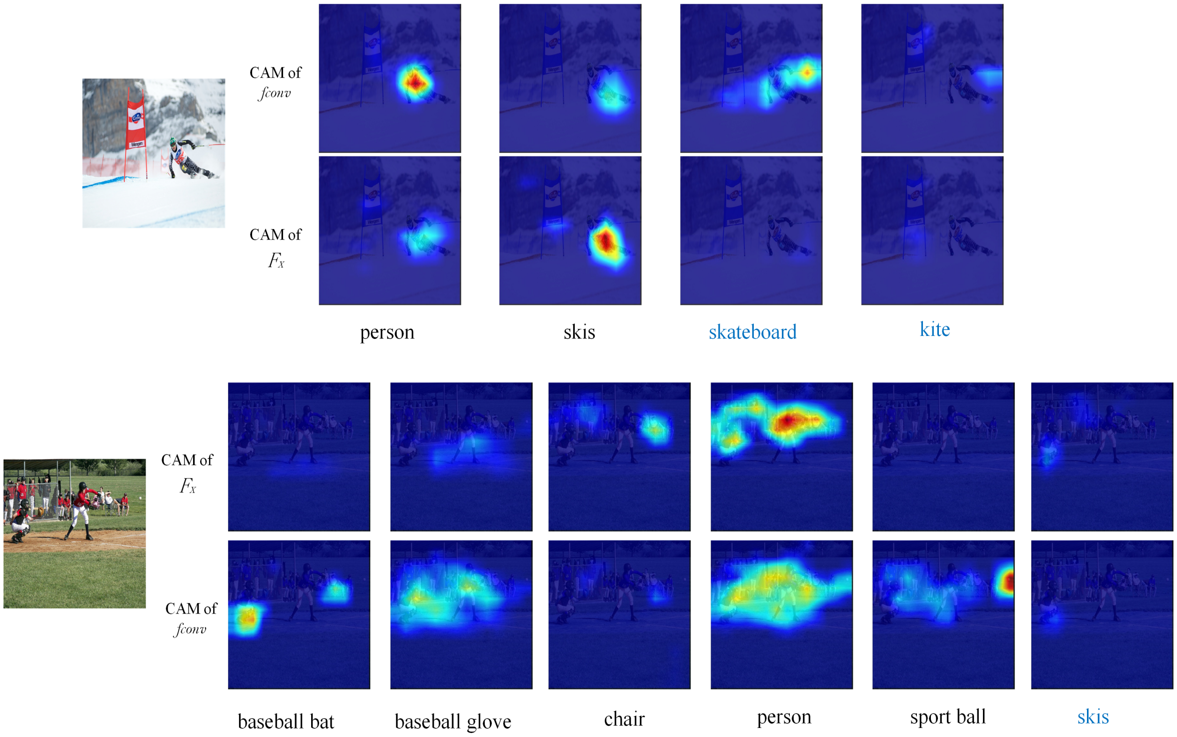

Fig. 3. CAM visualization results. Images and their activation attention maps from F_x with classification layer and f_{conv} are shown in the first and second rows, respectively. A label in blue font indicates that it is not associated with the the corresponding image. The features of f_{conv} are learned using the proposed constrained ranking loss, which captures correlations between image features and labels. From the results, clearly the attention map from f_{conv} has a greater response to people and skis than skateboard in the skiing scene. Meanwhile, in the baseball ball scene, features that are not affected by our constrained ranking loss have a greater response in the human area. On the other hand, the features affected by the constrained ranking loss have clear responses to the bat, glove and the ball in this scene, even though the response to the human is smaller.