Reinforced Multi-Label Image Classification by Exploring Curriculum

Shiyi He,[1,3] **Chang Xu,**[2] **Tianyu Guo,**[1,3] **Chao Xu,**[1,3] **Dacheng Tao**[2]
[1]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, China
[2]UBTECH Sydney AI Centre, SIT, FEIT, University of Sydney, Australia
[3]Cooperative Medianet Innovation Center, Peking University, China
shiyiHe@pku.edu.cn, c.xu@sydney.edu.au, tianyuguo@pku.edu.cn
xuchao@cis.pku.edu.cn, dacheng.tao@sydney.edu.au

Abstract

Humans and animals learn much better when the examples are not randomly presented but organized in a meaningful order which illustrates gradually more concepts, and gradually more complex ones. Inspired by this curriculum learning mechanism, we propose a reinforced multi-label image classification approach imitating human behavior to label image from easy to complex. This approach allows a reinforcement learning agent to sequentially predict labels by fully exploiting image feature and previously predicted labels. The agent discovers the optimal policies through maximizing the long-term reward which reflects prediction accuracies. Experimental results on PASCAL VOC2007 and 2012 demonstrate the necessity of reinforcement multi-label learning and the algorithm's effectiveness in real-world multi-label image classification tasks.

Introduction

Traditional single-label image classification deals with images that are associated with a single label from a finite set of disjoint labels. However, real-world images often correspond to more than one label. For example, an image can belong to "table" as well as "bottle". Recently, this multi-label image classification problem has attracted significant attention in various applications, such as semantic annotation of images (Gong et al. 2013; Wang et al. 2017), and videos (Oh et al. 2015; Wang et al. 2016b) and structured prediction (Lampert 2011).

The most common approach to solve the problem is the binary relevance method (BR) (Luaces et al. 2012). BR transforms a multi-label problem into multiple binary classification problems. However, BR fails to model the dependencies between multiple labels, which widely exist in real-world multi-label data, e.g. one image containing "sky" has a great probability containing "clouds". One simple yet effective approach to handle label dependencies is the classifiers chains (CC) method (Read et al. 2011). It incorporates label correlations implicitly by training a chain of binary classifiers in a given sequential order, with the feature set of each binary classifier augmented by labels that have been previously trained. Besides CC, lots of different techniques have been developed to discover label dependences in multi-label learning. For example, in the max-margin multi-label classifier (M3L) (Hariharan et al. 2010), label dependencies are represented by pairwise label correlations that are computed from the training set. CNN-RNN (Wang et al. 2016a) framework was proposed to learn a joint image-label embedding to characterize the semantic label dependencies.

Most existing algorithms simultaneously estimate all possible labels for an example. Though classifier chain method makes sequential prediction but it requires the pre-defined label sequence. In contrast, humans and animals learn much better when the examples are not randomly presented but organized in a meaningful order which illustrates gradually more concepts, and gradually more complex ones. This idea called curriculum learning (Bengio, Collobert, and Weston 2009) is inspired by the way children are taught: start with easier concepts (for example, recognizing clearly visible objects) and build up to more complex ones (for example, recognizing partially occluded objects). This learning strategy can be formalized into multi-label image classification tasks as well where a series of objects are recognized in a sequential order from easy to complex. Moreover, the size and location of objects influence the recognition difficulty and complexity of images to a great extent. Thus the prediction order should be designed related to each single image following its inter curriculums, instead of a simple pre-defined one computed on the whole training set. Similar to our work, ML-TLLT (Chen et al. 2016b), a multi-label propagation algorithm, also employed the curriculum mechanism to manipulate the propagation sequence from simple examples to more difficult ones. In other words, it tries to organize examples to learn each label. Instead, we organize labels to depict each image.

Besides, real-world image tagging system used to recommend labels for users to annotate their uploaded images. Instead of accepting all the recommended labels, users are more likely to select some labels from the recommended set according to their understandings of the images. However, the user feedbacks have rarely been investigated by traditional multi-label learning algorithms.

In this paper, we propose a novel Reinforced Multi-label Image Classification (RMIC) by exploring curriculum method to solve multi-label image classification problem. Consistent with curriculum learning mechanism, we compose the image feature and previously predicted labels as a new state to predict the next label. The multi-label system recommends labels to users and collects user feedbacks as reward to update the multi-label model. Our RMIC method allows an agent learning to label images from easy to complex by taking user feedbacks as partial observation of ground-truth labels. In particular, this agent aims to discover the optimal policy that maximizes the long-term reward which reflects prediction accuracies. To transform the original multi-label problem into Reinforcement Learning (RL) framework, we propose a scheme that enables the agent take actions (labels) without duplicates in an episode in both training and testing stage. This scheme divides the complete action (label) set into taken set and untaken set. At each step, the agent takes action on untaken set following ϵ-greedy policy and then updates the division. Unlike most existing works, our method does not label image in a pre-defined order, but learns a specific curricular sequence for each image to predict from easy to complex. This means our proposed RMIC method can discover the inter curricular label order related to image context in reinforcement learning process and it can be trained end-to-end based on Deep Q-learning algorithm (Mnih et al. 2015).

Related Work

During the past few years, a large number of multi-label image classification models have been developed. A popular baseline for multi-label classification is binary relevance (BR) (Luaces et al. 2012) which simply treats each label as a separate binary classification problem. However, its performance can be poor when strong label dependencies exist. In some practical applications, label dependencies are known as a priori or can be easily estimated, and thus various methods have been designed to improve performance by exploiting label dependencies. For example, the classifier chain algorithm (CC) (Read et al. 2011) learned a chain of binary classifiers, where each classifier predicts whether the current label exists given each input feature and the previously predicted labels. CC algorithm preserves inter label dependencies but the results can vary for different orders of chains. In order to solve this problem and increase accuracy, CC-DP (Liu and Tsang 2015)searched the globally optimal label order for CC and CC-Greedy (Liu and Tsang 2015) found a globally optimal CC. There are also many multi-label learning methods that can automatically discovery label dependencies from the training data. For example, PrML (You et al. 2017) explored and exploited label relationships by inheriting all the merits of privileged information and low-rank constraints. SLL (You et al. 2016) utilized the label self-representation to model the label relationship.

Curriculum learning aims to improve the learning performance by designing suitable curriculums from easy to complex for the stepwise learner. This learning approach was proposed by (Bengio, Collobert, and Weston 2009), which hypothesized that curriculum learning had both an effect on the speed of convergence of the training process as well as finding a better local minima in the case of non-convex criteria. Self-paced learning is a learning regime proposed by Kumar et al (Kumar, Packer, and Koller 2010), which can be regarded as an implementation of curriculum learning. This regime determines its inner curriculum dynamically to adjust to the learning pace of the learner. Curriculum learning and self-paced learning have been applied to various applications. For example, MMCL (Chen et al. 2016a) employed the curriculum learning methodology by investigating the difficulty of classifying every unlabeled image in the semi-supervised image classification task. The method proposed by (Svetlik et al. 2016) automatically generated a curriculum as a directed acyclic graph to improve the performance of the reinforcement learning agents.

Reinforcement Learning (RL) provides an appealing framework for addressing a wide variety of planning and control problems (Mnih et al. 2015). For example, a policy gradient RL approach was proposed for locomotion of a four-legged robot (Kohl and Stone 2004). Obstacle detection with a monocular camera can be formulated as a RL problem as well (Michels, Saxena, and Ng 2005). Most recently, deep Q-learning algorithm has been successfully applied to make decisions in ATARI games (Mnih et al. 2015). Besides, Google DeepMind proposed a new search algorithm by integrating Monte-Carlo tree search with deep RL, and it beat the world's best human player in the game of Go (Silver et al. 2016). Since then, a number of deep RL have been studied in various problems, e.g. optimal or near-optimal policies to localize objects (Jie et al. 2016; Caicedo and Lazebnik 2015), action-conditional video prediction for Atari Games (Oh et al. 2015) and sequence to sequence learning for text generation (Guo 2015).

Reinforced Multi-label Image Classification

We define $\mathcal{X} \in \mathbb{R}^d$ as an input domain. Let $\mathcal{Y} = \{1, 2, \cdots, m\}$ be a finite set of m possible labels. Consider $\mathbf{x} \in \mathcal{X}$ as an input data instance and $\mathbf{y} \subseteq \mathcal{Y}$ as the target classes associated with this input. If there exists K labels associated with $\mathbf{x}$, then $\mathbf{y} = \{y_1, y_2, \cdots, y_K\}$. y_i corresponds to the i-th label belonging to $\mathbf{x}$. Traditional supervised methods used to be developed with well-annotated data. Given examples accompanied with ground-truth labels, $\{(\mathbf{x}, \mathbf{y})\}_{i=1}^{n}$ pairs are already known. They train a model that maps input $\mathcal{X}$ to output $\mathcal{Y}$ as in the ordinary classification problems.

However, in the real-world image tagging task, an image tagging system not only recommends labels for images, but also collects feedbacks from users to update the labels. We therefore in this paper consider the sequential image annotation problem. In the i-th step of image annotation, this system recommends a label z_i and gets a feedback p_i from users, where $p_i \in \{-1, +1\}$, indicates the goodness of the recommended labels determined by users. If an image $\mathbf{x}$ is associated with K labels, in the training stage, we only consider recommendations $\mathbf{z} = \{z_1, z_2, \cdots, z_K\}$ and feedbacks $\mathbf{p} = \{p_1, p_2, \cdots, p_K\}$ during k interactions. In this sequential learning setting, the ground-truth labels are not directly provided for examples, and the algorithm can only get some feedbacks through the interaction with users to know whether the current prediction is right or not, which is completely different from the classical supervised setting with paired example features and ground-truth labels. The aim of the proposed algorithm is to exploit user feedbacks to learn a curricular model not only for accurately labeling images from easy to complex, but to utilize previous predictions for next label estimation.

Multi-label Image Classification as a Markov Decision Process

Since ground-truth labels are not given in our problem, it is difficult to adopt supervised methods for help. We cast the whole sequential image annotation procedure as a Markov Decision Process (MDP) since this setting provides a formal framework to model a sequential decision making process which can be solved by reinforcement learning algorithms. The image tagging system is treated as an agent in reinforcement learning, the process of an image is seen as an episode, and feedbacks from users are reward to the agent. And then, this procedure can be decomposed as the agent's label prediction process. This agent aims to learn a curricular policy that maximizes the total discounted reward which reflects the overall accuracies of labeling each image.

We adopt the formalism of this deterministic MDP $(\mathcal{X}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the set of possible actions (i.e., $|\mathcal{A}| < \infty$), $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow [-1, 1]$ is the reward of a state-action pair. $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow T(\mathcal{S})$ is the transition achieved by taking an action in a given state, and $\gamma \in (0, 1)$ is a discount factor. A deterministic curricular policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ is a mapping from states to actions. The actions $\mathcal{A}$, state $\mathcal{S}$, transitions $\mathcal{T}$ and reward $\mathcal{R}$ are detailed as follows.

Action: Action of the agent is to select a label for an image at each time step. The action set $\mathcal{A}$ is the same as the label set $\mathcal{Y}$, i.e. $\mathcal{A} = \mathcal{Y} = \{1, 2, \cdots, m\}$. During the training phase, the agent takes an action $a, a \in \mathcal{A}$ and receives positive or negative reward r from the environment. For each image, the number of acting steps depends on the number of labels and the resulting labels of the image consist of the individual label predicted at each step. In the test phase, the agent acts according to the learned curricular policy to sequentially predict labels from easy to complex for an unseen test image.

State: The state is represented as a tuple of two elements: the image feature f and the action history h. This tuple $s = (f, h), s \in \mathcal{S}$ summarizes the observed image and the history of actions since the beginning of the curricular prediction sequence. The image feature f is extracted from layer "fc6" of a VGG-16 model (Simonyan and Zisserman 2014), which is pre-trained on ImageNet and fine-tuned on the multi-label datasets in experiments. According to the suggestions in (Wei et al. 2016), directly using features extracted from networks pre-trained on ImageNet is not appropriate, since object categories in ImageNet and the multi-label datasets in hand are usually different. Compared to single-label classification, there exist various and complicated interactions among objects in term of the semantics and spatial locations. It is therefore necessary to fine-tune the feature extractor for a better investigation of the intrinsic relationship between labels of an image.

The action history h is a real vector that tells which actions have been taken in the past. Moreover, it implies a simple yet powerful formulation of label dependencies, which is beneficial for determining the next possible label. We encode the action vector following (Saberian and Vasconcelos 2011) by transforming an M-array (i.e. $\{1, \cdots, M\}$) into M distinct unit codewords $\mathcal{E} = \{e^1, e^2, \cdots, e^M\} \in \mathbb{R}^d$ that are as dissimilar as possible.

$$\begin{cases} max_{d, e^1, \cdots, e^M}[min_{i \neq j} ||e^i - e^j||^2] \\ s.t \quad ||e^k|| = 1 \quad \forall k = 1, \cdots, M. \\ \qquad e^k \in \mathbb{R}^d \quad \forall k = 1, \cdots, M. \end{cases}$$

Since M points $e^1, \cdots, e^M$ lie in an, at most, $M - 1$ dimensional subspace of $\mathbb{R}^d$, thus $min(d) = M - 1$. Besides, according to (Saberian and Vasconcelos 2011), there is no benefit in increasing d beyond $M - 1$. Thus, if the multi-label dataset contains M categories, each action is represented by a $(M - 1)$ dimensional vector where all values are real and the location of maximum absolute value corresponds to the taken action. n past actions are encoded in the state, which means $h \in \mathbb{R}^{(n * M - 1)}$. The value of n depends on the average number of labels for each image computed on the target database.

Transitions. The transitions $\mathcal{T}$ denote next-state function where every state is thought to be a possible consequence of taking an action in a state. Our proposed MDP transitions $\mathcal{T}$ are deterministic which means the new state is specified for each state and action pair. In an episode, transitions do not change the current image feature f but the action history h:

$$\mathcal{T}(s, a) = \mathcal{T}((f, h), a) = (f, h')$$

h' denotes the action history modified by action a.

Reward: In the real-world life, feedbacks are diverse. However, to simplify the problem and simulate the above sequential image annotation procedure, we generate the user feedbacks as follows. If the label recommended by the image tagging system for an image is acceptable by users, a positive feedback (reward) is returned to the system, otherwise the feedback (reward) is negative. In particular, we clipped all positive reward r at 1 (i.e. $r = 1, r \in R$) and all negative reward r at -1 (i.e. $r = -1, r \in R$). The system is supposed to recommend labels without duplication for an image and the implementation details will be given later. Fig.1 gives an example of the sequential label prediction process. The process starts with an initial default. Each decision corresponds to predicting a unique label. The figure shows the possible prediction path in continuous steps and the reward generated in this process.

Deep Q-learning for Reinforced Multi-label Image Classification

The optimal curricular policy of maximizing the summarization of discounted rewards over episodes (i.e. images) can be discovered with reinforcement learning techniques. Considering the high-dimensional continuous features of images and the model-free formulation of the user feedbacks, we resort to the deep Q-learning algorithm (Mnih et al. 2015) which has been demonstrated to discover optimal policies that generalize well to unseen inputs. This deep Q-learning

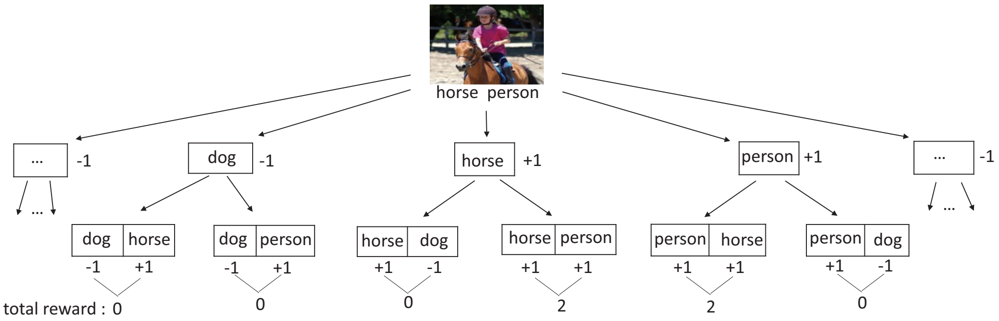

Figure 1: An example of sequential label prediction process. The ground-truth labels of the given image are “horse” and “person”. In the first step, the agent takes the action “horse” or “person” and gets reward “+1”, otherwise the reward is “-1”. This process repeats in the next step. And then, the optimal prediction sequence (policy) with “horse, person” or “person,horse” would result in maximum total reward “+2”. Moreover, the agent takes no duplicated actions in an episode for example, if the “horse” is taken in the first step, the agent would not take “horse” in the next steps.

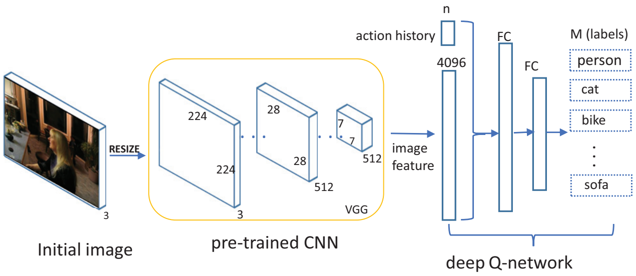

Figure 2: Illustration of our deep Q-network. There exist M categories in the target database and we decode n past actions as action history. The pre-trained CNN is employed as the image representation, and the previously predicted labels are encoded into action history. Then the two parts compose the current state as the input of Q-network and the output of the Q-network returns as the current label to update the model.

algorithm (Mnih et al. 2015) utilizes a deep neural network as a function approximator to estimate the value for each state-action pair. Besides, we use the pre-trained CNN on the ImageNet dataset as the feature extractor instead of learning the full feature hierarchy of the convolutional network similar to (Jie et al. 2016). Note that the pre-trained CNN has been fine-tuned on the multi-label datasets in experiments. During the training stage, we only need to update the parameters of the Q-Network, which makes the Q learning faster and more stable. The detailed architecture of our deep Q-network is illustrated in Fig. 2. In typical reinforcement learning model, starting from an episode, the agent takes one action from the whole action set at each step. There is often no constraint on actions to be taken and the agent can repeat the same actions in an episode. However, in original multi-label image classification problem, the label set for an image contains no duplicate elements. Thus when transforming this problem into the RL framework, the agent is supposed to take actions without duplicates in an episode no matter in the training phase or the test stage.

In the proposed reinforced multi-label image classification task, the agent’s behavior is ϵ-greedy during the training process. Let B be the action set consisting of the actions that have been taken before for the current image, and define $C = A \setminus B$. Different from deep Q-learning algorithm (Mnih et al. 2015), the ϵ-greedy policy of our algorithm is conducted on the action sub-set C, instead of the complete action set A. Specifically, the agent follows the greedy policy $a = argmax_{a' \in C} Q(s, a')$ with probability $1 - \epsilon$ and selects a random action from C with probability ϵ. At the beginning of each episode (i.e. image), B is set as $\emptyset$ and $C = A$. The agent takes action following the ϵ-greedy policy for the current state s, and then gets the next state s' and the reward r. After the reward has been received by the agent, one interaction is ended and the algorithm updates the current state $s = s'$, taken action set $B = B \cup \{a\}$ and untaken action set $C = C \setminus \{a\}$. The agent acts follows this process over episodes. The constraint on action sub-sets B and C is consistent with the multi-label image classification problem, where the predicted labels are requested to be unique. While in policy learning, we also incorporate a replay memory D following (Mnih et al. 2015) to store experiences of the past episodes, which allows one transition to be used in multiple model updates and breaks the short-time strong correlations between training samples. The deep Q-learning algorithm is off-policy and the untaken action set C is different in agent’s learning and acting stage. Thus, the next action a' is recorded and used directly for policy update. We represent each experience as (s, a, r, s', a') as (Mnih et al. 2015), which has additionally include the next action a' into the original four tuple (s, a, r, s') as (Mnih et al. 2015). Each time Q learning update is applied, a mini batch randomly sampled from the replay memory D is used for training. The loss function is the expectations of mean-squared error between approximated target value $r + \gamma \hat{Q}(s', a'; \theta)$ and Q network output value $Q(s, a)$.

$$L(\theta) = \mathbb{E}_{(s,a,r,s',a') \sim D}[(r + \hat{Q}(s', a'; \theta) - Q(s, a; \theta))^2]$$

θ represents the weights of the proposed deep Q-network and the update of these weights at the i-th iteration θ_i is descirbed as follows:

$$\theta_{i+1} = \theta_i + \alpha(r + \gamma\hat{Q}(s', a'; \theta_i) - Q(s, a; \theta_i)) \bigtriangledown_{\theta_i} Q(s, a; \theta_i)$$

where α is the learning rate and γ is the discount factor.

The deep Q-learning for reinforced multi-label image classification algorithm is shown as Algorithm 1.

Algorithm 1 Deep Q-learning for RMIC

```
Initialize replay memory $D$, the whole action set $A$
  for episode = $1, M$ do
    for each image do
      initialize a state $s_1$ with the image and empty the taken
      action set $B$ and untaken action set $C$.
      for $t = 1, T$ do
        Compute untaken action set $C = A \setminus B$
        Select action $a_t$ from $C$ on $\epsilon$-greedy policy
        Execute $a_t$, observe reward $r_t$, next state $s_{t+1}$
        Put $a_t$ into taken action set $B$
        Re-compute $C = A \setminus B$
        Select next action $a_{t+1}$ from $C$ on $\epsilon$-greedy policy
        Store transition $(s_t, a_t, r_t, s_{t+1}, a_{t+1})$ in $D$
        Sample random minibatch of transitions
        $(s_j, a_j, r_j, s_{j+1}, a_{j+1})$ from $D$
        Update $Q$ based on current and next state-action pair.
        $y_j = \begin{cases} r_j & \text{if episode terminates at step } j+1 \\ r_j + \gamma\hat{Q}(s_{j+1}, a_{j+1}; \theta) & \text{otherwise} \end{cases}$
        Update network weights $\theta$ following the gradient
        $(y_j - Q(s_j, a_j; \theta)) \cdot \bigtriangledown Q(s_j, a_j; \theta)$
        Update $s_{t+1} \leftarrow s_t$ and target network $\hat{Q} \leftarrow Q$
      end for
    end for
  end for
```

Implementation Details

The proposed algorithm was evaluated on the PASCAL VOC2007 and PASCAL VOC2012 datasets (Everingham et al. 2010). The output layer of Q-network is a linear layer with a single output for each valid action or label. Since there are 20 categories in VOC database, the three fully-connected layers' neurons of deep Q-network were set as 512, 128 and 20, respectively. Each action was represented by a 19-dimensional vector and the action history h encoded 2 past actions. We trained the network for 3 epochs and each epoch was ended after the agent had interacted with all training images. During the ϵ-greedy training, was annealed linearly set from 1 to 0.2 over the first 2 epochs to progressively allow the agent to use its own learned model. Then ϵ was fixed to 0.2 in the last epoch, so the agent further adjusted its network parameters. We assumed that the number of labels presenting in an image was given in the training phase. While in the test stage, the label numbers were unknown and we limited the number of steps to obtain the desired number of labels for each image. The mini batch size was set to 32. The algorithm was implemented on the publicly available Keras platform on a single NVIDIA GeForce Titan X GPU with 12GB memory.

Experimental Results

In this paper, we conduct comprehensive experiments on PASCAL VOC2007 and VOC2012 classification benchmarks to evaluate the proposed method. These two databases contain 9,963 and 22,531 images respectively, and are divided into *train*,*val* and *test* subsets. We merge the *train* set with *val* set into *trainval* set and conduct our experiments on the *trainval / test* splits (5,011/4,952 for VOC 2007 and 11,540/10,991 for VOC2012). The evaluation metrics include *Average Percision*(AP) and mean of AP(mAP).

Label curriculum exploration

We first evaluate the influence of the label prediction order on the multi-label learning algorithms' performance on the PASCAL VOC2007 dataset. After counting label occurrence frequencies in the training data, it is supposed that more frequent labels tend to appear earlier than less frequent ones (Wang et al. 2016a). Based on this assumption, we design a simple label order with no curriculum characteristic, which is feed into the proposed algorithm RMIC to learn the policy. This RMIC-fixed variant thus has a different reward function with that of standard RMIC. For example, if the pre-defined label order is "person" and "horse" while the algorithm sequentially predicts "horse" and "person", the rewards received will be "-1" and "-1". Positive rewards can only be obtained when each sequentially predicted label is right. However, in our standard RMIC method, the reward can be "+1" and "+1" either for the predictions "horse" and "person" or "person" and "horse". Apart from the different reward setting during training, the state representation, action set, behavior policy of RMIC-fixed and RMIC methods are the same. We thus compared RMIC-fixed with our proposed method in terms of evaluation metrics, including *mAP*, class-level averaged precision and example-level averaged precision (*C-P* and *E-P*), class-level averaged recall and example-level averaged recall (*C-R* and *E-R*), and class-level averaged *F1* and example-level averaged *F1* (*C-F1* and *E-F1*). According to Table 1, it is observed that the proposed RMIC method outperforms the RMIC-fixed variant in terms of nearly all the evaluation metrics. In the proposed RMIC method, the RL agent learns a curricular order to predict from easy to complex based on the current image context and previous prediction history. However, the agent in RMIC-fixed receives the positive reward "+1" only when it has selected the label which is at the corresponding location of the pre-defined order. The performance improvement of the proposed RMIC algorithm demonstrates its effectiveness in discovering the more optimal curricular label order to predict from easy to complex, which is an advantage over other methods which require a number of attempts to determine the best label sequence either in training or test stage. We next proceed to analyze label occurrence frequency in

Method	C-P	E-P	C-F1	C-R	E-R	E-F1	mAP
RMIC-fixed	42.7	87.7	57.4	42.3	89.6	57.5	81.5
RMIC	43.7	88.8	58.6	43.1	91.3	59.4	84.5

Table 1: The influence of label prediction order on the prediction results

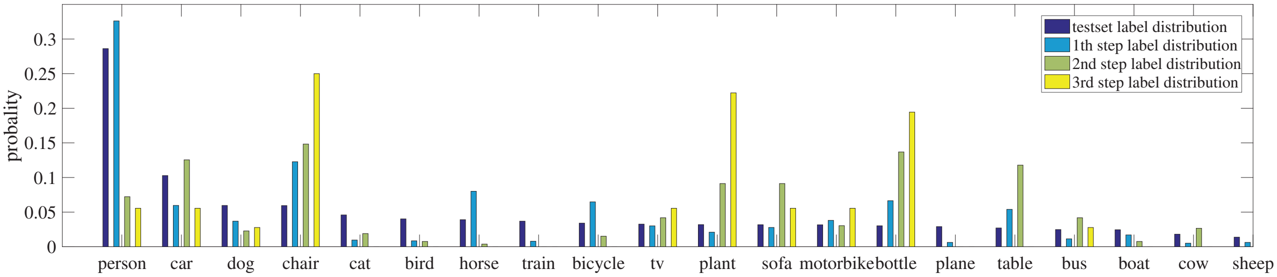

Figure 3: The label appearance numbers of the 20 categories on the VOC2007 dataset

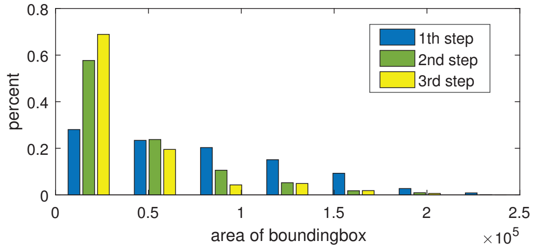

Figure 4: Size distribution of bounding boxes for objects prediction at the first three prediction steps

the VOC2007 dataset to reveal the relationship between label frequencies and their learned curricular appearance order. In Fig.3, purple pillars represent the ground-truth appearance numbers of different labels, while the rest three color ones are the predicted label appearance numbers in the proposed RMIC algorithm at the first three steps. It is observed that the agent generally predicts labels based on the ground-truth label appearance numbers (the purple ones) at the first step. In other words, more frequent labels appear earlier than those less frequent ones . However at the following steps, this trend gets weaker and the agent predicts those infrequent labels with the help of previously predicted ones. For example, “person”, “car”, and “dog” instances are more commonly found in the training set, thus most of images associated with these labels are recognized earlier. While those instances associated with few images like “bottle”, “plant” and “tv” are usually recognized with the help of those already recognized ones at the next two steps. Besides, we analyze the sizes of objects that are predicted at the first three steps. The information of bounding boxes provided in the dataset are used for a more intuitive illustration in Fig.4. It is observed that larger objects are usually recognized first, which then promotes recognition of smaller objects at the next two steps. For example, “chair”, “plant” and “bottle” often appear as appendages of “person”, since the recognition result of “person” provides evidences for the existence of these smaller instances. In Fig.3, infrequent instances like “boat”, “cow” and “sheep” are recognized at the first or second step, and this is mainly because of their larger object sizes in images. Taking Fig.3 and Fig.4 together, we suggest that our proposed RMIC method can predict labels from easy to complex following the curriculum mechanism, and utilizes previous easy predictions to promote predictions of later complex ones.

Image classification results

We compare our the proposed RMIC algorithm with state-of-the-art fully supervised multi-label learning method. The scores computed by the Q-network estimate the confidence on labels. In the test stage, since there is no feedback of users, we directly use the calculated Q-values without iterations as the predicted confidence scores. Classification results on the VOC2007 and VOC2012 datasets are shown in Table 2 and Table 3, respectively. INRIA (Harzallah, Jurie, and Schmid 2010) is built on the transitional feature extraction-coding-pooling pipeline. AGS (Dong et al. 2013) and AMM (Song et al. 2011) employ grounding-truth bounding box information for training. HCP-Alex (Wei et al. 2016) extracts region proposals to fine-tune the features pre-trained on ImageNet. (Sermanet et al. 2013) and (Chatfield et al. 2014) proposed CNN-SVM pipeline for multi-label classification. PRE-1000C* and PRE-1512* (Oquab et al. 2014) described a weakly supervised convolutional neural network directly for object annotation and classification. SPP*(He et al. 2014) equips the network with another pooling strategy which reduces the scale effect on classification task. Whole images can be feed into CNN-RNN (Wang et al. 2016a) framework and this framework can be trained end-to-end. We also conducted the 5*2 CV test in terms of metric mAP on the database of VOC2007. The mean value is 84.57%. This smaller standard deviation indicates the stability of the algorithm’s performance. Besides, all these comparison algorithms have trained in the fully supervised setting, where all the information of ground-truth labels are provided. However, according to the reported results, we find that the performance of our RMIC method can perform comparable or even better than these supervised methods, even though the RMIC method has only used the partially supervised information through the user feedbacks.

Prediction results on different epochs

We show example prediction results on the VOC2007 dataset using the RMIC algorithm trained at different epochs in the training phase in Fig.5. We find that from the first epoch to the third epoch, the predicted labels become more distinguishable, especially for images associated with more ground-truth labels such as the fifth image. We also note that

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
INRIA (Harzallah, Jurie, and Schmid 2010)	77.2	69.3	56.2	66.6	45.5	68.1	83.4	53.6	58.3	51.1	62.2	45.2	78.4	69.7	86.1	52.4	54.4	54.3	75.8	62.1	63.5
AGS (Dong et al. 2013)	82.2	83.0	58.4	76.1	56.4	77.5	88.8	69.1	62.2	61.8	64.2	51.3	85.4	80.2	91.1	55.2	60.0	69.7	83.6	77.0	71.1
AMM (Song et al. 2011)	84.5	81.5	65.0	71.4	52.2	76.2	87.2	68.5	63.8	55.8	65.8	55.6	84.8	77.0	91.1	55.2	60.0	69.7	83.6	77.0	71.3
Razavianetal.* (Sermanet et al. 2013)	88.5	81.0	83.5	82.0	42.0	72.5	85.3	81.6	59.9	58.5	66.5	77.8	81.8	78.8	90.2	54.8	71.1	62.6	87.4	71.8	73.9
PRE-1000C* (Oquab et al. 2014)	88.5	81.5	87.9	82.0	47.5	75.5	90.1	87.2	61.6	75.7	67.3	85.5	83.5	80.0	95.6	60.8	76.8	58.0	90.4	77.9	77.7
Chatfield et al.* (Chatfield et al. 2014)	95.3	90.4	92.5	89.6	54.4	81.9	91.5	91.9	64.1	76.3	74.9	89.7	92.2	86.9	95.2	60.7	82.9	68.0	95.5	74.4	82.4
SPP* (He et al. 2014)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	82.4
HCP-Alex* (Wei et al. 2016)	95.4	90.7	92.9	88.9	53.9	81.9	91.8	92.6	60.3	79.3	73.0	90.8	89.2	86.4	92.5	66.9	86.4	65.6	94.4	80.4	82.7
CNN-RNN (Wang et al. 2016a)	96.7	83.1	94.2	92.8	61.2	81.2	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
RMIC	97.1	91.3	94.2	57.1	86.7	90.7	93.1	63.3	83.3	76.4	92.8	94.4	91.6	95.1	92.3	59.7	86.0	69.5	96.4	79.0	84.5

Table 2: Classification Results (AP in %) comparison on the PASCAL VOC2007 dataset

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
NUS-PSL (Zitnick and Dollr 2014)	97.3	84.2	80.8	85.3	60.8	89.9	86.8	89.3	75.4	77.8	75.1	83.0	87.5	90.1	95.0	57.8	79.2	73.4	94.5	80.7	82.2
Zeiler et al.* (Silver et al. 2016)	96.0	77.1	88.4	85.8	55.8	85.8	78.6	91.2	65.2	74.4	67.7	87.8	86.0	85.1	90.9	52.2	83.6	61.1	91.8	76.1	79.0
PRE-1000C* (Oquab et al. 2014)	93.5	78.5	87.7	80.9	57.3	85.0	81.6	89.4	66.9	73.8	62.0	89.5	83.2	87.6	95.8	61.4	79.0	54.3	88.0	78.3	78.8
PRE-1512* (Oquab et al. 2014)	94.6	82.9	88.2	84.1	60.3	89.0	84.4	90.7	72.1	86.8	69.0	92.1	93.4	88.6	96.1	64.3	86.6	62.3	91.1	79.8	82.8
Chatfieldetal.* (Chatfield et al. 2014)	96.8	82.5	91.5	88.1	62.1	88.3	81.9	94.8	70.3	80.2	76.2	92.9	90.3	89.3	95.2	57.4	83.6	66.4	93.5	81.9	83.2
HCP-Alex* (Wei et al. 2016)	97.7	83.2	92.8	88.5	60.1	88.7	82.7	94.4	65.8	81.9	68.0	92.6	89.1	87.6	92.1	58.0	86.6	55.5	92.5	77.6	81.8
RMIC	98.0	85.5	92.6	88.7	64.0	86.8	82.0	94.9	72.7	83.1	73.4	95.2	91.7	90.8	95.5	58.3	87.6	70.6	93.8	83.0	84.4

Table 3: Classification Results (AP in %) comparison on the PASCAL VOC2012 dataset

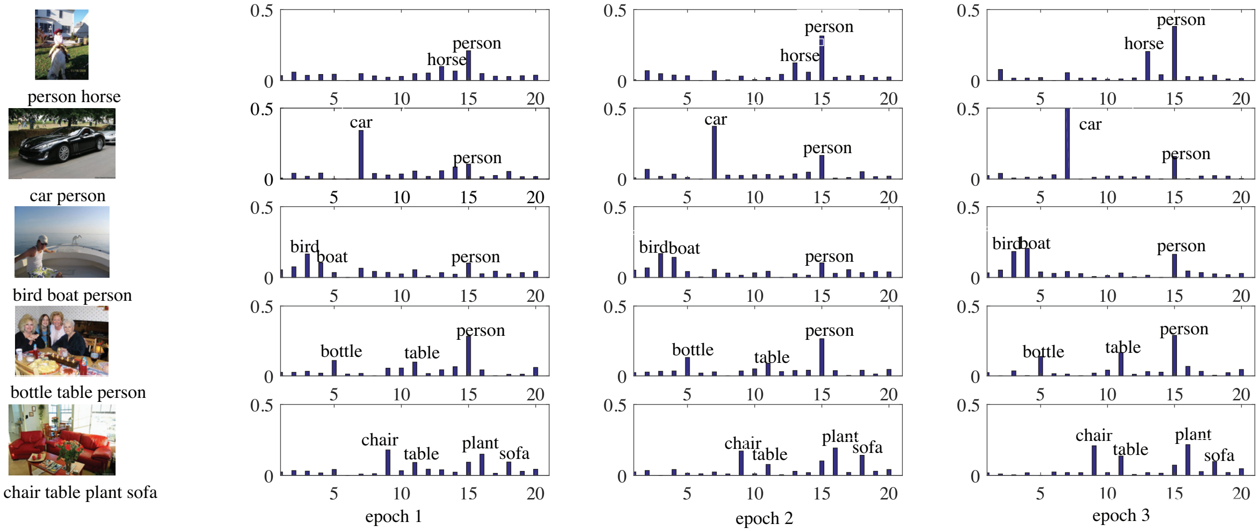

Figure 5: Examples of prediction results on the VOC2007 dataset at different epochs

easily predicted objects will be beneficial for the prediction of complex ones. For example, “car” in the second image promotes the prediction of “person”, and in the fourth image. Though the existences of “table” and “bottle” are not very obvious, the label “person” helps the recognition of “bottle”, and “bottle” promotes the existence probability of “table”. These indicate that our RMIC method learns to annotate images from easy to complex not only based on the object context, but also on the relationship of objects, so that the algorithm can determine the optimal curricular order .

Conclusion

In this paper, we propose a novel Reinforced Multi-label Image Classification (RMIC) approach imitating human curriculum mechanism to label image from easy to complex. This approach allows a reinforcement learning agent to sequentially predict labels by taking the user feedbacks as partial observation of ground-truth labels. This agent fully exploits image feature and previously predicted labels as a new state to predict the next label, and it aims to learn an optimal curricular policy with the goal of determining the most accurate labels of the target images. Experimental results show that our method can explore the optimal relationship of objects and actively utilize their dependencies to determine the curricular prediction order in the multi-label classification task. Moreover, experimental results on PASCAL VOC2007 and VOC2012 demonstrate that this RMIC approach achieves superior performance to the state-of-the-art supervised learning methods.

Acknowledgments

We thank supports of NSFC 61375026 and 2015BAF15B00, and ARC Projects: FL-170100117, DE-180101438, DP-180103424, DP-140102164, LP-150100671

References

Bengio, Y.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *International Conference on Machine Learning*, 41–48.

Caicedo, J. C., and Lazebnik, S. 2015. Active object localization with deep reinforcement learning. In *IEEE International Conference on Computer Vision*, 2488–2496.

Chatfield, K.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Return of the devil in the details: Delving deep into convolutional nets. *Computer Science*.

Chen, G.; Tao, D.; Maybank, S. J.; Wei, L.; Kang, G.; and Jie, Y. 2016a. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing* 25(7):3249–3260.

Chen, G.; Tao, D.; Liu, W.; and Liu, W. 2016b. Teaching-to-learn and learning-to-teach for multi-label propagation. In *Thirtieth AAAI Conference on Artificial Intelligence*, 1610–1616.

Dong, J.; Xia, W.; Chen, Q.; Feng, J.; Huang, Z.; and Yan, S. 2013. Subcategory-aware object classification. In *Computer Vision and Pattern Recognition*, 827–834.

Everingham, M.; Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* 88(2):303–338.

Gong, Y.; Jia, Y.; Leung, T.; Toshev, A.; and Ioffe, S. 2013. Deep convolutional ranking for multilabel image annotation. *CoRR* abs/1312.4894.

Guo, H. 2015. Generating text with deep reinforcement learning. *Computer Science* 40(4):1–5.

Hariharan, B.; Zelnik-Manor, L.; Vishwanathan, S. V. N.; and Varma, M. 2010. Large scale max-margin multi-label classification with priors. In *International Conference on Machine Learning*, 423–430.

Harzallah, H.; Jurie, F.; and Schmid, C. 2010. Combining efficient object localization and image classification. In *IEEE International Conference on Computer Vision*, 237–244.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 37(9):1904–1916.

Jie, Z.; Liang, X.; Feng, J.; Jin, X.; Lu, W.; and Yan, S. 2016. Tree-structured reinforcement learning for sequential object localization. In *Advances in Neural Information Processing Systems 29*, 127–135.

Kohl, N., and Stone, P. 2004. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings*, 2619–2624 Vol.3.

Kumar, M. P.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. In *International Conference on Neural Information Processing Systems*, 1189–1197.

Lampert, C. H. 2011. Maximum margin multi-label structured prediction. *Advances in Neural Information Processing Systems* 289–297.

Liu, W., and Tsang, I. W. 2015. On the optimality of classifier chain for multi-label classification. In *International Conference on Neural Information Processing Systems*, 712–720.

Luaces, O.; Dez, J.; Barranquero, J.; Coz, J. J. D.; and Bahamonde, A. 2012. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence* 1(4):303–313.

Michels, J.; Saxena, A.; and Ng, A. Y. 2005. High speed obstacle avoidance using monocular vision and reinforcement learning. In *International Conference*, 593–600.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; and Ostrovski, G. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529.

Oh, J.; Guo, X.; Lee, H.; Lewis, R. L.; and Singh, S. P. 2015. Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, 2863–2871.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Weakly supervised object recognition with convolutional neural networks. *Vision Par Ordinateur Et Reconnaissance De Formes*.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine Learning* 85(3):333–359.

Saberian, M. J., and Vasconcelos, N. 2011. Multiclass boosting: Theory and algorithms. *Advances in Neural Information Processing Systems* 2124–2132.

Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; and Lecun, Y. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *Eprint Arxiv*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van, d. D. G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; and Lanctot, M. 2016. Mastering the game of go with deep neural networks and tree search. *Nature* 529(7587):484.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *Computer Science*.

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

Svetlik, M.; Leonetti, M.; Sinapov, J.; Shah, R.; Walker, N.; and Stone, P. 2016. Automatic curriculum graph generation for reinforcement learning agents. In *AAAI Conference on Artificial Intelligence*.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016a. Cnn-rnn: A unified framework for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2285–2294.

Wang, Y.; Xu, C.; You, S.; Tao, D.; and Xu, C. 2016b. Cnnpack: Packing convolutional neural networks in the frequency domain. In *Conference on Neural Information Processing Systems*.

Wang, Y.; Chang, X.; Shan, Y.; Chao, X.; and Tao, D. 2017. Dct regularized extreme visual recovery. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* 26(7):3360.

Wei, Y.; Xia, W.; Lin, M.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; and Yan, S. 2016. Hcp: A flexible cnn framework for multi-label image classification. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 38(9):1901–1907.

You, S.; Xu, C.; Wang, Y.; Xu, C.; and Tao, D. 2016. Streaming label learning for modeling labels on the fly. *CoRR* abs/1604.05449.

You, S.; Xu, C.; Wang, Y.; Xu, C.; and Tao, D. 2017. Privileged multi-label learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 3336–3342.

Zitnick, C. L., and Dollr, P. 2014. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, 391–405.