Attentive Recurrent Neural Network for Weak-supervised Multi-label Image Classification

Liang Li, Shuhui Wang∗
Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China
{lili.wangshuhui}@ict.ac.cn

Shuqiang Jiang, Qingming Huang∗
Institute of Computing Technology, CAS
University of Chinese Academy of Sciences, China
sqjiang@ict.ac.cn,qmhuang@ucas.ac.cn

ABSTRACT
Multi-label image classification is a fundamental and challenging task in computer vision, and recently achieved significant progress by exploiting semantic relations among labels. However, the spatial positions of labels for multi-labels images are usually not provided in real scenarios, which brings insuperable barrier to conventional models. In this paper, we propose an end-to-end attentive recurrent neural network for multi-label image classification under only image-level supervision, which learns the discriminative feature representations and models the label relations simultaneously. First, inspired by attention mechanism, we propose a recurrent highlight network (RHN) which focuses on the most related regions in the image to learn the discriminative feature representations for different objects in an iterative manner. Second, we develop a gated recurrent relation extractor (GRRE) to model the label relations using multiplicative gates in a recurrent fashion, which learns to decide how multiple labels of the image influence the relation extraction. Extensive experiments on three benchmark datasets show that our model outperforms the state-of-the-arts, and performs better on small-object categories and under the scenario with large number of labels.

CSC CONCEPTS
• Computing methodologies → Object recognition; Supervised learning by classification; Neural networks;

KEYWORDS
Multi-label image classification; attentive recurrent neural network; label relations modeling; reinforcement learning

ACM Reference Format:

∗Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM ’18, October 22–26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5665-7/18/10 ... $15.00
https://doi.org/10.1145/3240508.3240649

Figure 1: The RHN generates a sequence of candidate glimpses which can highlight the specific regions, and the emission indicator adaptively select which glimpse should be emitted to the real prediction set for further label relation modeling and classification.

1 INTRODUCTION
Multi-label image classification is one of the most challenging problems in computer vision. Since every real-world image normally abounds with rich semantic information, the goal of multi-label classification is to annotate multiple visual concepts appearing in the image, such as objects and scenes. This task can greatly benefit a wide variety of applications such as image semantic search, human-computer interaction and automatic driving, so it has attracted much research interest [17, 18, 21, 30].

To solve multi-label image classification, researchers [13, 18, 20, 21, 48] try to overcome two main challenges, (1) how to learn the discriminative feature representations and (2) how to effectively model the complex label relations. These derive from the following reasons. First, the intrinsic property of multi-label image, i.e. diverse visual concepts, makes the learning of effective feature representations difficult and important. Second, the label relation could help better capture the co-occurrence pattern among different labels, which greatly boost the classification results. As a consequence, many existing methods try to exploit these challenges for improving the classification performance.

For the first challenge, a lot of methods have been developed to explore effective feature representations, including discriminative models [18, 37] which transform this task into multiple single-label classification problem and learn the features for each label, nearest neighbor models [15] which predict labels based on the assumption that visually similar images are more likely to share common labels, and clustering models [33] which learn the final representations by employing the max-pool or fisher vector to cluster low-level features. These methods give the whole image the same weights...
when recognizing different objects. However, it might not be optimal for the images with diverse objects, because a ground truth label might be closely related to only specific region of the image. To simulate the capability of human vision system of selectively focusing on specific parts of image for visual concepts [28], many attention-based methods have been proposed and achieved success in different tasks [39, 44, 45]. Despite that existing attention methods can identify salient regions corresponding to labels with high probability, they lack the intrinsic mechanism to avoid the association error between regions and labels.

For the second one, it refers to an observation that the labels associated with natural images do not appear in isolation, they appear correlatively and interact with each other. Many previous methods have been proposed to exploit the label relations by sparse learning [31], graphical model [19] and correlation matrix [41, 48], etc. Recently, RNN based model [14, 21, 36] is utilized to capture such correlations due to its effectiveness of modeling sequential data. However, these methods suffer from the error propagation problem [29], which means that since current label relies on its predecessors, a single false prediction might be propagated and even reinforced along the sequence, and finally largely degrades the performance.

To address these challenges, this paper proposes an end-to-end attentive recurrent neural network for weak-supervised multi-label image classification, which can learn the discriminative feature representations and model the label relations simultaneously. Figure 1 shows the workflow of our attentive neural network and Figure 2 reveals the technical architecture of the proposed method. First, we propose a recurrent highlight network (RHN) to imitate an attention mechanism for generating a sequence of candidate glimpses, which can focus on specific regions of image to learn more effective feature representation. Furthermore, since some candidate glimpses might focus on trivial or noisy parts of image, we introduce a REINFORCE algorithm to learn the emission indicator to decide whether the candidate glimpse should be emitted to the real prediction set as emitted glimpse. Second, we propose a gated recurrent relation extractor (GRRE) to deal with the error propagation problem. It employs all other emitted glimpses as inputs to better model the label correlation and utilizes multiplicative gates as a contribution indicator on this emitted glimpses. The ablation experiments demonstrate the GRRE plays an important role in our model for performance improvement.

Compared with some region proposal based methods [7, 25, 43] that require the ground truth of bounding boxes which need expensive labor cost and might not be available in practice, our RHN can work well with weak image-level supervisions. Moreover, since our model can highlight related regions, it greatly benefits the performance of small visual objects; due to the effective emitted glimpses and the robust label correlation mining, our model is good at multi-label classification under the scenario with large number of labels; experimental results validate the above powers. More in-depth studies and visualizations in the Section 4 demonstrate the effectiveness of the component part of our model. We will release code to facilitate future research.

2 RELATED WORK

Multi-label image classification plays an important role in bridging the gap between low-level features and high-level semantic concept. Many methods have been proposed to learn the mapping between image features and labels for this work. Nearest neighbor based models [9, 12, 15] computed the similarities between training samples and the query sample. A novel topic-model based latent variable regression approach is presented [27] to capture correlations between image features and labels. Recently, some CNN based multi-label classification models have shown competitive performances. Gong et al. [8] combined convolutional architectures with approximate top-k ranking objectives. On top of powerful CNN features, Uricchio et al. [33] used Fisher vector and PCA to obtain more effective image features in order to boost classification performances. ResNet [10] built a well established residual learning framework, where the layers are reformulated as learning residual functions with reference to the layer inputs. However, the above methods left the label correlation untouched.

Various approaches have been proposed to utilize the label correlation for multi-label classification. A common approach [2] was to use the graph structured learning method, which incorporates the topological constraints of relation graph. Recently, with the success of recurrent neural networks, more and more methods tried to apply the RNN based methods (LSTM/GRU) to model the label correlation. Wang et al. [36] proposed a CNN-RNN model to learn a joint embedding to characterize the label dependency and image-label relevance. In [11], they developed a structured inference neural network which models the label relations from
Attention models are inspired by the human vision perception and achieve great successes in different tasks, such as image captioning [45], fine-grained classification [39], etc. Recurrent Attention Model (RAM) [23] was proposed to learn the gaze strategies on cluttered digit classification tasks. It is further extended to multiple digit recognition [1] by processing a multi-resolution crop of the input image.

3 METHOD

3.1 Main Architecture

The architecture of the proposed method is described in Figure 2, which includes three main components: recurrent highlight network (RHN), emission indicator, and gated recurrent relation extractor (GRRE). The RHN first generates a sequence of candidate glimpses which focus on different regions of multi-label image. Since some candidate glimpses might focus on trivial parts or noisy regions, we introduce the emission indicator to decide whether to emit current candidate glimpse. The emission indicators are non-differentiable components of our model, so we design the REINFORCE algorithm to learn the policy for emission indicators. In order to model the label relationship robustly, we propose GRRE, which uses all other emitted glimpses as inputs and applies the multiplicative gates to control the weights from all other emitted glimpses to current emitted glimpses. If the candidate glimpse is chosen to be emitted, we also explore the context features of the emitted glimpse via the context extractor. Finally, the label is predicted based on the three parts, i.e., the emitted glimpse, the context features of the emitted glimpse and the label relations from all other emitted glimpses to current emitted glimpse.

3.2 Recurrent Highlight Network (RHN)

Since the ground truth labels might closely relate to some specific regions in the multi-label image, we propose the attention based recurrent highlight network (RHN), to direct the high-resolution attention to the most related parts. RHN generates a sequence of candidate glimpses in the recurrent manner with the guidance of guide network. We now describe the component parts in details.

Guide Network. To make the candidate glimpses concentrate on different regions of image, guide network is introduced to ensure that the current glimpse explicitly knows the glimpses that has been produced by RHN. It contains a Softmax function and a Deconvolutional layer. Given the feature \( X_t \in \mathbb{R}^{512} \) from fully connected layer with a ReLU nonlinearity at last glimpse \( t - 1 \), we first design a Softmax function to suppress the high activations at candidate glimpse \( t - 1 \), which drives the current glimpse to focus on other regions.

\[
l_{t-1} = \frac{\exp(X_{t-1}^{\text{max}}) - \exp(X_{t-1})}{\sum_j \exp(X_j)} \in \mathbb{R}^{512}
\]

where \( X_{t-1}^{\text{max}} \) is the maximum activation value in \( X_{t-1} \). The \( l_{t-1} \) is used to direct the current glimpse to focus on other regions. Then we apply one deconvolutional layer to obtain the attention maps \( A_t \in \mathbb{R}^{7\times7\times512} \) to guide the generation of current candidate glimpse.

\[
A_t = F(l_{t-1} \theta \mathcal{F}), A_t \in \mathbb{R}^{7\times7\times512}
\]

\[
V_t = V_t \otimes A_t, V_t \in \mathbb{R}^{7\times7\times512}
\]

where \( F(\cdot) \) is the deconvolutional transfer function and \( \theta \mathcal{F} \) are its learnable parameters. The \( V_t \in \mathbb{R}^{7\times7\times512} \) is the features extracted from VGG16 and \( \otimes \) is the element-wise multiplication. The ReLU nonlinearity operation is performed following the deconvolutional layer. Finally, we get the features of current candidate glimpse \( X_t \in \mathbb{R}^{512} \) by applying the fully connected layer on top of the weighted visual features \( V_t \). The visualization in Section 4.6 shows that the \( V_t \) can direct the high resolution attention to different regions of multi-label image at different glimpses.

Context Extractor. After getting the features of current candidate glimpse, we introduce the context extractor to capture the context features based on the attentional regions of current candidate glimpse. The component parts of context extractor are illustrated in the Figure 3. The feature \( V_t \in \mathbb{R}^{7\times7\times512} \) is first sent into two different convolution layers, i.e., a convolution layer with kernel size of \( 1 \times 1 \times 256 \) and a dilated convolution layer [46] with kernel size of \( 3 \times 3 \times 256 \), to capture context features of different scales. Then we concatenate the outputs of the two convolution layers and apply the average pooling layer on top of the concatenation to get the contextual feature \( C_t \in \mathbb{R}^{512} \).

3.3 Emission Indicator

The emission indicator is designed to indicate whether a candidate glimpse should be emitted for final label prediction. Actually in the training datasets, there are only image label information about the whole image while there are no bounding box information about each label in the image. Thus the emission indicator is non-differentiable component of our model due to the lack of ground truth annotations. In this paper, we introduce the Reinforcement Learning that enables learning in non-differentiable setting. We
now describe the REINFORCE algorithm briefly and the reward function that we propose to learn effective decision policies for emission indicator.

**REINFORCE.** In reinforcement learning [38], an agent continually interacts with environment by observing the state and choosing an action according to its policy. The agent also receives reward. Thus, the goal of REINFORCE can be expressed as

$$J(\theta) = \sum_{a \in \mathcal{G}} p_\theta(a)r(a)$$

(3)

where the $\mathcal{G}$ is the space of action sequence. $p_\theta(a)$ denotes a distribution over $a \in \mathcal{G}$ and $\theta$ is its parameters. The $r(a)$ is the reward assigned to possible action sequence. We want to learn the parameters $\theta$ that maximize the reward of a sequence of emission indicator outputs.

$$\nabla J(\theta) \approx \frac{1}{K} \sum_{i=1}^{K} \sum_{t=1}^{T} \nabla \log p_\theta(\alpha_t^i|h_{1:t}, \alpha_{t-1}^i) R_t^i$$

(4)

where the $p_\theta$ is the agent’s policy, which is a learned distribution over candidate glimpses conditioned on the interaction thus far. $\alpha_t$ is the policy’s current action, i.e., the emission indicator in our case. $h_{1:t}$ is the history of past states including current state, and $\alpha_{t-1}$ is the history of past actions. For a sequence of $T$ candidate glimpses, $R_t = \sum f_{ij}$ is the cumulative future reward from the current glimpse onward. The approximate gradients are computed by running $K$ sequences under current policy.

Typically, a baseline reward $b_t$ is subtracted from the returned reward in order to reduce the variance of the expected gradient. A common method to obtain the baseline is to learn a value function $b_t = \mathbb{E}[R_t] [32]$ via a separate network. Thus, the gradient equation becomes:

$$\nabla J(\theta) \approx \frac{1}{K} \sum_{i=1}^{K} \sum_{t=1}^{T} \nabla \log p_\theta(\alpha_t^i|h_{1:t}, \alpha_{t-1}^i)(R_t^i - b_t^i)$$

(5)

Since we obtain the approximate gradients, we can easily incorporate the emission indicator into the standard back propagation training. In our implementation, the binary emission indicator output $\epsilon_t = f_e(X_t; \theta_e)$, where $f_e$ is a fully connected layer followed by a sigmoid non-linearity function and $\theta_e$ denotes its parameters. At training time, $f_e$ is used to parameterize a Bernoulli distribution from which $\epsilon_t$ is sampled.

**Reward.** In our model, we want to learn policies for emission indicator outputs that lead to multi-label classification with high precision and recall. Thus, we propose a reward function $R_T$ that aims to maximize the true positive classifications while minimizing false classifications:

$$R_T = \begin{cases} 
R_e & \text{if } N_e = 0 \\
N_e R_+ - N_e R_- & \text{otherwise}
\end{cases}$$

(6)

The reward is returned at the final (Tth glimpse in our case) candidate glimpse. $N_e$ denotes the number of emitted glimpses. If $N_e = 0$ which means the emission indicators do not emit any glimpse, we utilize a negative reward $-R_e$ to punish this case in order to encourage the emission indicator to emit the appropriate candidate glimpses. On the other hand, if the number of emitted glimpses is not zero, we set another reward function to give the true classification a positive reward $R_e$ and the false classification a punishment $R_-$, respectively. $N_e$ and $N_-$ are the number of true classifications and false classifications, respectively. In the implementation, we set $R_+$ larger than the $R_-$. More specifically, we set $R_+ = 1$, $R_- = 0.1$ and $R_e = 10$.

We therefore train the emission indicator by the proposed reward function with REINFORCE, and learn effective decision policies for emission indicator to lead to high classification performance.

### 3.4 Gated Recurrent Relation Extractor (GRRE)

Recently RNN-based methods [14, 36] have achieved great progress in exploiting the label correlation in a sequential manner. However, these models suffer from the error propagation problem [29]. Specifically, since each label relies on its predecessors, a single false prediction might be propagated and possibly even reinforced along the label sequence. On the other hand, the relations among multiple labels commonly exist in a many-to-one manner beyond a pairwise manner. For instance, the "boat" is meanwhile closely related to "sea" and "island", and both of them can promote positive support on "boat" prediction in certain degrees.

To address these issues, we design a robust gated recurrent relation extractor (GRRE), which is inspired by the flow control in recurrent architectures like GRU or LSTM. First, it encodes the internal representation with a tanh activation. Then it employs the features from all other emitted glimpses and features from the last time-step as the inputs to compute the contributions to current glimpse via a Softmax function. Finally, a multiplicative gate is introduced to assign importance to different emitted glimpses.

In consistency with Eqn.(6), we assume that the emission indicators emit $N_e$ glimpses in total. We now describe the pipeline of relation extraction at the $i$th emitted glimpse as illustrated in Figure 4.
\[ h_k = \text{Tanh}(W_{hxk}X_k + W_{hfk}f_k^{(i-1)}) \]  
(7)

\[ z = \text{Softmax}((W_{xzi}X_i + W_{zf1}f_1^{(i-1)}), \ldots, (W_{xzi}X_i + W_{zfNc}f_{Nc}^{(i-1)})) \]  
(8)

where \( h_k \in \mathbb{R}^{512} \) is an internal representation based on the current input \( X_k \in \mathbb{R}^{312} \) and weighted features \( f_k^{(i-1)} \in \mathbb{R}^{512} \) extracted from the last time step \( (i-1 \text{ in this case}) \) at the \( k \text{th emitted glimpse} \) with a tanh activation function. \( z \in \mathbb{R}^{Nc-1} \) denotes the contribution of other \((Nc-1)\) emitted glimpses to current emitted glimpse, which is learned by using the Softmax function based on the concatenation of \( X_k \in \mathbb{R}^{312} \) and \( f_k^{(i-1)} \in \mathbb{R}^{512} \). \( f_k^{(i-1)} \in \mathbb{R}^{512} \) is the weighted features of other emitted glimpses except the current \( i \text{th emitted glimpse} \), which is computed via the multiplicative gates based on \( z_k \) and \( h_k \), noting that \( k \in (1, 2, \ldots, Nc) \). Finally, we obtain the label relation \( O_i \in \mathbb{R}^{512} \) of the \( i \text{th emitted glimpse} \) from all other emitted glimpses by summing them up. \( W[\cdot] \) are the parameters to be learned and \([\cdot, \ldots, \cdot] \) is the concatenator operation. We therefore learn the explicit label relation by using the above equations in recurrent fashion.

Since all other emitted glimpses are used as inputs and the multiplicative gates can adaptively govern the contribution from other emitted glimpses, our GRRE is robust against the error propagation problem caused by single false prediction.

3.5 Training

The label prediction for the \( i \text{th emitted glimpse} \) is based on the three parts, i.e., the attentional features \( X_i \in \mathbb{R}^{512} \) extracted from the recurrent highlight network, the context features \( C_i \in \mathbb{R}^{512} \) from context extractor and the label relation \( O_i \in \mathbb{R}^{512} \) from GRRE.

\[ P_i = f(W_{xpi}X_i + W_{cip}C_i + W_{op}O_i), P_i \in \mathbb{R}^D \]  
(9)

where \( W[\cdot] \) are the neural network parameters which are to be learned and \( D \) is the size of label vocabulary. Here \( f(\cdot) \) is the Softmax function. Then the negative log probability is applied on the prediction results as the loss function.

\[ L = \frac{1}{N_c} \sum_{i=1}^{N_c} - \log \hat{P}_i \]  
(10)

where \( \hat{P}_i \) is the probability of ground truth label. We therefore train the proposed model with standard back propagation method under weak image-level supervisions in an end-to-end manner. The summary of the proposed method is presented in Algorithm 1.

4 EXPERIMENTS

In this section, we evaluate our model at the task of multi-label image classification on three benchmark datasets, i.e. NUS-WIDE dataset [4], MS-COCO dataset [22] and ESP Game dataset [35].

4.1 Implementation Details

The 16-layer VGG network is pretrained on ImageNet dataset [6]. We adopt Adam [16] to optimize the model in an end-to-end manner.

Learning rate is scheduled as \( 10^{-3} \) and a staircase weight decay is applied after 10 epochs. \( \beta_1 \) and \( \beta_2 \) in Adam are set to 0.8 and 0.999. We fix VGG network at the beginning and fine-tune it after 20 epochs with a relatively small learning rate \( 10^{-5} \). We set the length of candidate glimpse sequence \( T = 10 \) and the number of sampling times in emission indicator \( K = 4 \), which are learned through cross-validation. All parameters in the GRRE are initialized from zero-mean Gaussian with standard deviations 0.1 and the biases are set to 0. The batch size in our model is 64. Our model is implemented using the Torch framework [5] on one Nvidia GTX Titan X.

4.2 Evaluation Measures and Compared Methods

For fair comparison, we follow previous work [9] to evaluate the performance. The precision (\( P_L \)) and recall (\( R_L \)) of all labels are applied as evaluation metrics. Based on these two measures, we compute F1-score \( F_L = 2 * P_L * R_L / (P_L + R_L) \), which achieves trade-off between precision and recall. Similarly, we also report the precision \( P_i \), recall \( R_i \) and F1-score \( F_i \) over all testing images. We compare with the following methods.

- **WARP** [8]: trains the multi-label classification with a weighted approximate ranking loss.
- **WeakCNN** [26]: treats the task as a binary classification, which adds adaptation layers to learn to roughly localize objects or distinctive parts for each label with weakly image-level supervisions.
- **CNN-RNN** [36]: utilizes the RNN to implicitly model label relation, then use the concatenation of visual features of whole image from CNN and label embedding from RNN to classify.
- **CNN-LSTM [42]**: learns the relative importance of the labels by the LSTM and trains the LSTM with same visual features at each timestep.
- **RIA [14]**: extracts the visual features from the whole image and feed it into LSTM at the first timestep to mine the label relation.
- **RAM [23]**: adaptively selects a sequence of regions or locations by reinforcement learning, and learns an RNN model for extracting information from an image.
- **ResNet-50 [10]**: apply the original ResNet-50 fine-tuned on each of the datasets.
- **ResNet50+DCNN [30]**: use ResNet-50 as the input feature, and construct the objective function with max-margin objective, max-correlation objective, and correntropy loss for classification.
- **LSEP [21]**: implement image classification by pairwise ranking based on a log-sum-exp function.
- **SRN [48]**: use ResNet-50 as the input to generate attention maps, and learn the underlying relations via learnable convolutions for classifying.

To analyze the contributions of different model components, we design some variants of our model as follows:

- **Ours w/o CE** removes the context extractor (CE).
- **Ours w/o EI** removes the emission indicator (EI).
- **Ours w/o GRRE** removes the GRRE.
- **Ours w/o GRRE w. RNN** replaces the GRRE with normal RNN.
- **Ours w/o CE w/o GRRE** removes both the context extractor and GRRE.

### 4.3 Performance on NUS-WIDE

NUS-WIDE dataset is widely used as the benchmark for image classification and multimedia retrieval due to its large amount of images and high quality annotations. There are 209,347 labelled images and all these images are manually annotated into 81 concepts by human. We follow the same split of NUS-WIDE dataset as [8]. The number of training images is 150,000 and the rest 59,347 images are testing images. We only use 81-tag annotations for training without using attribute annotations and metadata.

We compare the performance of our method against state-of-the-art approaches and our variants in Table 1. Since the number of labels in each image is different, we take the top \( k \) generated labels to evaluate performance and set \( k = 3 \) in the NUS-WIDE dataset. Compared with these methods using the ranking framework [4, 8, 21], our method achieves significant improvement. Compared with CNN-RNN based models [14, 36, 42], ResNet and its variant [10, 30], our model also gain a favorable performance improvement, which indicates the advantages of RHN and GRRE over normal CNN with CNN-RNN based models [14, 36, 42]. Ours w/o CE and Ours w/o GRRE w. RNN. Third, the context features also can improve the classification performance but not as much as GRRE by comparing Ours w/o CE and Ours. Four, the utility of emission indicator learnt by the REINFORCE algorithm is validated by removing the emission indicator, and experiment shows that the performance of Ours has a valuable improvement by comparing with the method Ours w/o EI.

From the above result, we can find that our model with GRRE performs much better than with normal RNN in all terms of evaluation measures, which demonstrates that GRRE model can help effectively capture the label relationship.

### 4.4 Performance on MS-COCO

MS-COCO dataset is a large benchmark for various tasks such as object detection, segmentation, recognition and captioning. For multi-label classification, we use the object annotations as the ground truth labels. There are 80 object types and about 3 object labels per image. We use 82,783 training images as the training set and 40,504 validation images as the testing set.

Table 2 shows the comparison results on MS-COCO dataset. Compared with methods using CNN features, such as WARP [8], CNN-Softmax, ResNet-50[10], our method achieves consistently

<table>
<thead>
<tr>
<th>Method</th>
<th>( P_L )</th>
<th>( R_L )</th>
<th>( F_L )</th>
<th>( P_I )</th>
<th>( R_I )</th>
<th>( F_I )</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN[4]</td>
<td>32.6</td>
<td>19.3</td>
<td>24.3</td>
<td>42.9</td>
<td>53.4</td>
<td>47.6</td>
<td>49.6</td>
</tr>
<tr>
<td>KNN[4] with VGG16</td>
<td>35.1</td>
<td>30.5</td>
<td>32.6</td>
<td>45.9</td>
<td>57.2</td>
<td>50.8</td>
<td>52.2</td>
</tr>
<tr>
<td>WARPI[8]</td>
<td>31.7</td>
<td>35.6</td>
<td>33.5</td>
<td>48.6</td>
<td>60.5</td>
<td>53.9</td>
<td>54.9</td>
</tr>
<tr>
<td>CNN-LSTM[42]</td>
<td>36.7</td>
<td>38.2</td>
<td>37.4</td>
<td>49.2</td>
<td>61.0</td>
<td>54.5</td>
<td>57.3</td>
</tr>
<tr>
<td>CNN-RNN[36]</td>
<td>40.5</td>
<td>30.4</td>
<td>34.7</td>
<td>49.9</td>
<td>61.7</td>
<td>55.2</td>
<td>56.1</td>
</tr>
<tr>
<td>ResNet-50[10]</td>
<td>39.6</td>
<td>43.8</td>
<td>41.6</td>
<td>51.5</td>
<td>63.2</td>
<td>56.8</td>
<td>59.6</td>
</tr>
<tr>
<td>ResNet50+DCNN[30]</td>
<td>42.1</td>
<td>44.9</td>
<td>43.5</td>
<td>52.6</td>
<td>65.4</td>
<td>58.3</td>
<td>61.9</td>
</tr>
<tr>
<td>RIA[14]</td>
<td>39.2</td>
<td>41.6</td>
<td>40.4</td>
<td>51.3</td>
<td>62.9</td>
<td>56.5</td>
<td>59.5</td>
</tr>
<tr>
<td>LSEP[21]</td>
<td>48.4</td>
<td>42.7</td>
<td>45.4</td>
<td>59.3</td>
<td>60.8</td>
<td>60.0</td>
<td>60.6</td>
</tr>
<tr>
<td>Ours w/o CE w/o GRRE</td>
<td>34.2</td>
<td>31.6</td>
<td>32.8</td>
<td>45.3</td>
<td>55.9</td>
<td>50.1</td>
<td>47.5</td>
</tr>
<tr>
<td>Ours w/o GRRE</td>
<td>35.4</td>
<td>37.7</td>
<td>36.5</td>
<td>48.3</td>
<td>60.1</td>
<td>53.6</td>
<td>52.8</td>
</tr>
<tr>
<td>Ours w/o CE</td>
<td>43.4</td>
<td>45.0</td>
<td>44.2</td>
<td>52.7</td>
<td>65.6</td>
<td>58.4</td>
<td>61.5</td>
</tr>
<tr>
<td>Ours w/o EI</td>
<td>42.8</td>
<td>44.6</td>
<td>43.7</td>
<td>49.9</td>
<td>63.4</td>
<td>55.8</td>
<td>58.3</td>
</tr>
<tr>
<td>Ours w/o GRRE w. RNN</td>
<td>41.5</td>
<td>43.8</td>
<td>42.6</td>
<td>51.4</td>
<td>65.7</td>
<td>57.7</td>
<td>60.9</td>
</tr>
<tr>
<td>Ours</td>
<td>44.2</td>
<td>49.3</td>
<td>46.6</td>
<td>53.9</td>
<td>68.7</td>
<td>60.4</td>
<td>63.5</td>
</tr>
</tbody>
</table>

Table 1: Performance evaluation on NUS-WIDE dataset for \( k = 3 \).

<table>
<thead>
<tr>
<th>Method</th>
<th>( P_L )</th>
<th>( R_L )</th>
<th>( F_L )</th>
<th>( P_I )</th>
<th>( R_I )</th>
<th>( F_I )</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-Softmax[36]</td>
<td>59.0</td>
<td>57.0</td>
<td>58.0</td>
<td>60.2</td>
<td>62.1</td>
<td>61.1</td>
<td>47.4</td>
</tr>
<tr>
<td>WARP[8]</td>
<td>59.3</td>
<td>52.5</td>
<td>55.7</td>
<td>59.8</td>
<td>61.4</td>
<td>60.7</td>
<td>49.2</td>
</tr>
<tr>
<td>CNN w/o RNN[36]</td>
<td>65.3</td>
<td>54.5</td>
<td>59.3</td>
<td>68.5</td>
<td>61.3</td>
<td>65.7</td>
<td>57.2</td>
</tr>
<tr>
<td>CNN-LSTM[42]</td>
<td>66.7</td>
<td>54.2</td>
<td>59.8</td>
<td>68.1</td>
<td>65.0</td>
<td>66.5</td>
<td>62.3</td>
</tr>
<tr>
<td>CNN-RNN[36]</td>
<td>66.0</td>
<td>55.6</td>
<td>60.4</td>
<td>69.2</td>
<td>66.4</td>
<td>67.8</td>
<td>61.2</td>
</tr>
<tr>
<td>WeakCNN[26]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>62.8</td>
</tr>
<tr>
<td>RIA[14]</td>
<td>63.2</td>
<td>56.6</td>
<td>59.7</td>
<td>70.4</td>
<td>65.3</td>
<td>67.9</td>
<td>63.1</td>
</tr>
<tr>
<td>ResNet50[10]</td>
<td>66.5</td>
<td>55.8</td>
<td>60.7</td>
<td>71.9</td>
<td>66.3</td>
<td>70.0</td>
<td>64.3</td>
</tr>
<tr>
<td>LSEP[21]</td>
<td>73.5</td>
<td>56.4</td>
<td>63.8</td>
<td>76.3</td>
<td>61.8</td>
<td>68.3</td>
<td>62.9</td>
</tr>
<tr>
<td>Ours</td>
<td>71.9</td>
<td>59.6</td>
<td>65.2</td>
<td>74.3</td>
<td>69.7</td>
<td>71.8</td>
<td>66.7</td>
</tr>
</tbody>
</table>

Table 2: Performance evaluation on MS-COCO dataset.
much better classification performances because our RHN learns the more specific feature representation than methods using the whole image with uniform weights to extract visual features for different labels. Compared with similar attention-based methods [26] with image-level supervision, our method shows about 4% improvement as we also capture the label relation by GRRE.

Compared with RNN-based methods which aim to exploit the label relation to achieve impressive performance, such as CNN-RNN [36], CNN-LSTM [42], RIA [14] and LSEP [21], our method also gains a great performance improvement (about 4% on MAP) because our GRRE better models the label correlation than RNN (LSTM/GRU) methods, which is more robust against the error propagation problem.

Experiments on small objects. Unlike previous methods which treat the whole image with uniform weights, our proposed RHN can focus on the most related regions of image for corresponding labels. It is beneficial to improve the performance of visually small objects because the features extracted from the whole image with uniform weights tend to characterize the dominant objects and ignore the small objects. Furthermore, the label relation also helps to seek the small objects based on the co-occurrence pattern after the dominant objects have been classified. To verify the efficiency of recognizing small objects of our model, we sample some visually small objects in the MS-COCO dataset, including "bottle", "cell phone", "bowl", "fork", "knife" and "mouse", and then compare their per-class precision and recall scores with CNN-RNN [36] and RIA [14]. The comparison results are shown in Figure 5. We can observe that our method achieves significant improvement among all the small objects than the baseline methods.

4.5 Performance on ESP-Game

The ESP-Game dataset contains images annotated by a game that two players are required to assign labels to the same image without communication, which results in 268 different labels and about 20,000 images. For a fair comparison, we use the same training and testing split as TagProp [9]. Since this dataset contains about 5 labels per image, we set the top ranked label list $k = 5$ to evaluate different methods. Furthermore, we only report precision ($P_k$) and recall ($R_k$) over all labels, because these baseline methods do not provide the evaluation results over all testing images. Additionally, the number of labels with non-zero recall value ($N^+$) is also reported for evaluation.

![Figure 5: Comparison performances on small objects between CNN-RNN [36], RIA [14] and Ours.](image)

Table 3: Performance evaluation on ESP Game dataset for $k = 5$. HC denotes hand-crafted features.

<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>$P_k$</th>
<th>$R_k$</th>
<th>$F_k$</th>
<th>$N^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TagProp[9]</td>
<td>HC</td>
<td>39</td>
<td>27</td>
<td>32</td>
<td>239</td>
</tr>
<tr>
<td>SLED[3]</td>
<td>HC</td>
<td>49</td>
<td>30</td>
<td>37</td>
<td>253</td>
</tr>
<tr>
<td>CNN-R[24]</td>
<td>Caffe-Net</td>
<td>44.5</td>
<td>28.5</td>
<td>34.7</td>
<td>248</td>
</tr>
<tr>
<td>2PKNN[34]</td>
<td>VGG-16</td>
<td>40</td>
<td>23</td>
<td>29</td>
<td>250</td>
</tr>
<tr>
<td>RIA [14]</td>
<td>VGG-16</td>
<td>32</td>
<td>32</td>
<td>31</td>
<td>249</td>
</tr>
<tr>
<td>RAM [23]</td>
<td>FCN</td>
<td>35.5</td>
<td>30.6</td>
<td>32.9</td>
<td>251</td>
</tr>
<tr>
<td>CNN-LSTM[42]</td>
<td>VGG-16</td>
<td>41.6</td>
<td>31.1</td>
<td>35.6</td>
<td>253</td>
</tr>
<tr>
<td>CCA-KNN[24]</td>
<td>VGG-16</td>
<td>46</td>
<td>36</td>
<td>40.4</td>
<td>260</td>
</tr>
<tr>
<td>ResNet-50[10]</td>
<td>ResNet-50</td>
<td>48.7</td>
<td>35.5</td>
<td>41.1</td>
<td>258</td>
</tr>
<tr>
<td>SRN [48]</td>
<td>ResNet-50</td>
<td>51.2</td>
<td>37.1</td>
<td>43.0</td>
<td>261</td>
</tr>
<tr>
<td>Ours</td>
<td>VGG-16</td>
<td><strong>51.3</strong></td>
<td><strong>37.8</strong></td>
<td><strong>43.5</strong></td>
<td><strong>261</strong></td>
</tr>
</tbody>
</table>

Except for these baseline methods described in Section 4.2, we also compare with other methods, including TagProp [9] which predicts labels by taking a weighted combination of nearest neighbor tags, SLED [3] which employs the dictionary representation to mine the label co-occurrence relation, CCA-KNN [24] and 2PKNN [34] which are both nearest neighbor based methods, and RAM [23] which is a visual attention model.

Table 3 shows the comparison results. First, compared with methods using hand-crafted features [3, 9], our method achieves better performance in all evaluation terms. Second, since the ESP-Game dataset contains 268 different labels, it makes the modeling of label relation in ESP-Game dataset more sensitive and difficult than the NUS-WIDE dataset (81 labels) and the MS-COCO dataset (80 labels). Thus, we observe that our method significantly surpasses CNN-RNN-based models [14, 42] with about 9% precision improvement, because the proposed GRRE performs better than the RNN based relation extractor under the complex and practical situation, which weakens the effect of single false label prediction by using multiplicative gates. Third, compared with the visual attention model RAM [23], the precision of our method has the promotion of 15.8%. This benefit from that our RHN use both soft-attention and emission indicator to generate glimpses, while RAM use the crop-region based attention model to obtain glimpses. As shown in Figure 6, our glimpses focus on the most related regions of image for the corresponding labels. Four, compared with other methods using deep features of the whole images [3, 10, 24, 34], our method gains a favorable improvement because RHN learns the more discriminative feature representations. Our model also outperform SRN ([48], ResNet-based Spatial Regularization Net), which benefit from both the effective feature representation from RHN and the robust label correlation mining from GRRE. This indicates the effectiveness of the proposed method under the scenario with large number of labels.

4.6 Visualization

To intuitively illustrate the focused regions at different emitted glimpses, we visualize some images in Figure 6 using Deconvolutional network [47]. In our implementation, we select the feature maps with maximum activation values in the weighted features $V_t$.
As the first example shows in Figure 6, our model first concentrates on the "boat" part, and then shifts its attention to the "sea" at the second emitted glimpse and the "island" at the third emitted glimpse, noting that the visualization of attention does not involve the ground truth labels. This result indicates that the proposed RHN is able to focus on the most related regions of multi-label image to learn the specific visual features for the corresponding labels. Some small objects, such as the "plate" in the second example of Figure 6, is able to focus on the most related regions of multi-label image to learn the specific visual features for the corresponding labels. This result indicates that the proposed RHN is able to focus on the most related regions of multi-label image to learn the specific visual features for the corresponding labels. Some small objects, such as the "plate" in the second example of Figure 6, is able to focus on the most related regions of multi-label image to learn the specific visual features for the corresponding labels.

Besides, in this supplementary materials, we show more experimental results of our proposed end-to-end attentive recurrent neural network for weak-supervised (image-level labels without object bounding box) multi-label image classification, where the discriminative feature representations and the label relations are learned simultaneously. A recurrent highlight network is proposed to learn the more specific features for diverse objects, and then the emission indicator trained by REINFORCE is designed to select the appropriate candidate glimpse. A gated recurrent relation extractor is proposed to capture the label relation without suffering from the error propagation problem. Extensive evaluations on three benchmark datasets demonstrate that our method has the significant improvement over state-of-the-art.

ACKNOWLEDGMENT
This work was supported in part by National Natural Science Foundation of China: 61771457, 61732007, 61532016, 61632016, 61620106009, 61672497, U1636214 and 61650202, in part by National Basic Research Program of China (973 Program): 2015CB351800, and in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013.

REFERENCES


Figure 6: Visualization of regions with high activations at different emitted glimpses. The first column is the original image, and the rest are the high activation maps and corresponding highlighted multi-label images.


