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
{liang.li,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.

CCS 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:
Liang Li, Shuhui Wang and Shuqiang Jiang, Qingming Huang*. 2018. Attentive Recurrent Neural Network for Weak-supervised Multi-label Image Classification. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240649

*Corresponding author.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240649

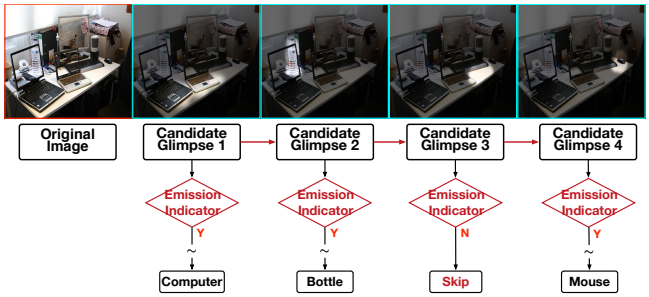

Figure 1: The RHN generates a sequence of candidate glimpses which can highlight the specific regions, and the emission indicator adaptively select which glimpse should be emitted to the real prediction set for further label relation modeling and classification.

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

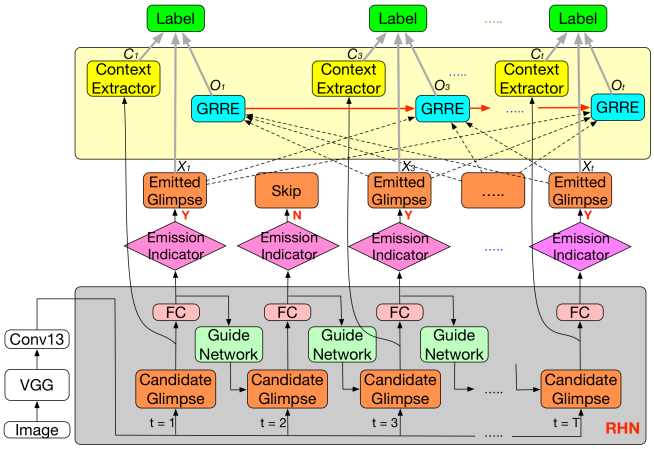

Figure 2: Main architecture of the proposed attentive recurrent neural network. The input image is fed into VGG16 to extract features from the last convolutional layer. Then RHN generates a sequence of candidate glimpses based on the guide network. After that, the emission indicator is learned to decide whether the candidate glimpse should be emitted. The emitted glimpses are sent into the proposed GRRE to model the label relations. Finally, the label is predicted based on three parts, *i.e.*, the features from emitted glimpses, the context features extracted from emitted glimpses via the context extractor, and the label relations from GRRE.

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 leaved 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

hierarchical to within-layer dependencies. Yan *et al.* [42] treated the image annotation as a task of generating ordered tag list, and they used the LSTM to model the long sequential data. Xie *et al.* [40] captured the label-correlations from the training data and incorporating label co-occurrence relations obtained from external knowledge. However, there still exists room for improvement because the above methods fall short on exploiting the specific regions of multi-label image.

Attention models are inspired by the human vision perception and achieve great successes in different tracks, 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 $X_{t-1} \in \mathbb{R}^{512}$ 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.

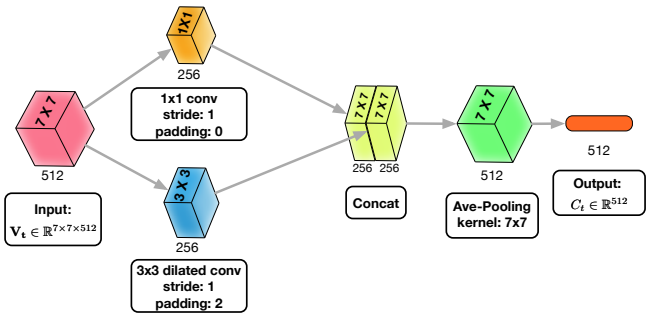

Figure 3: Illustration of the context extractor

$$l_{t-1} = \frac{\exp(X_{t-1}^{\max}) - \exp(X_{t-1})}{\sum_j \exp(X_{t-1}^{j})}, l_{t-1} \in \mathbb{R}^{512} \tag{1}$$

where $X_{t-1}^{\max}$ 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 $\mathbf{A}_t \in \mathbb{R}^{7\times7\times512}$ to guide the generation of current candidate glimpse.

$$\begin{aligned} \mathbf{A}_t &= \mathcal{F}(l_{t-1}; \theta_{\mathcal{F}}), \mathbf{A}_t \in \mathbb{R}^{7\times7\times512} \\ \mathbf{V}_t &= \mathbf{V_I} \otimes \mathbf{A}_t, \mathbf{V_I} \in \mathbb{R}^{7\times7\times512} \end{aligned} \tag{2}$$

where $\mathcal{F}(\cdot)$ is the deconvolutional transfer function and $\theta_{\mathcal{F}}$ are its learnable parameters. The $\mathbf{V_I} \in \mathbb{R}^{7\times7\times512}$ 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 $\mathbf{V}_t$. The visualization in Section 4.6 shows that the $\mathbf{V}_t$ 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 $\mathbf{V}_t \in \mathbb{R}^{7\times7\times512}$ is first sent into two different convolution layers, *i.e.*, a convolution layer with kernel size of $1\times1\times256$ and a dilated convolution layer [46] with kernel size of $3\times3\times256$, 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

$$J(\theta) = \sum_{\alpha \in \mathcal{G}} p_\theta(\alpha) r(\alpha) \tag{3}$$

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

$$\nabla J(\theta) \approx \frac{1}{K} \sum_{i=1}^{K} \sum_{t=1}^{T} \nabla \log \pi_\theta(\alpha_t^i | h_{1:t}^i, \alpha_{1:t-1}^i) R_t^i \tag{4}$$

where the π_θ is the agent's policy, which is a learned distribution over candidate glimpses conditioned on the interaction thus far. α_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 $\alpha_{1:t-1}$ is the history of past actions. For a sequence of T candidate glimpses, $R_t = \sum_t^T r_t$ 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 $b_t = \mathbb{E}_\pi[R_t]$ [32] via a separate network. Thus, the gradient equation becomes:

$$\nabla J(\theta) \approx \frac{1}{K} \sum_{i=1}^{K} \sum_{t=1}^{T} \nabla \log \pi_\theta(\alpha_t^i | h_{1:t}^i, \alpha_{1:t-1}^i)(R_t^i - b_t^i) \tag{5}$$

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 $e_t = f_e(X_t; \theta_e)$, where f_e is a fully connected layer followed by a sigmoid non-linearity function and θ_e denotes its parameters. At training time, f_e is used to parameterize a Bernoulli distribution from which e_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:

$$R_T = \begin{cases} -R_e & \text{if } N_e = 0 \\ N_+ R_+ - N_- R_- & \text{otherwise} \end{cases} \tag{6}$$

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_+ and the false classification a punishment R_-, respectively. N_+ 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.

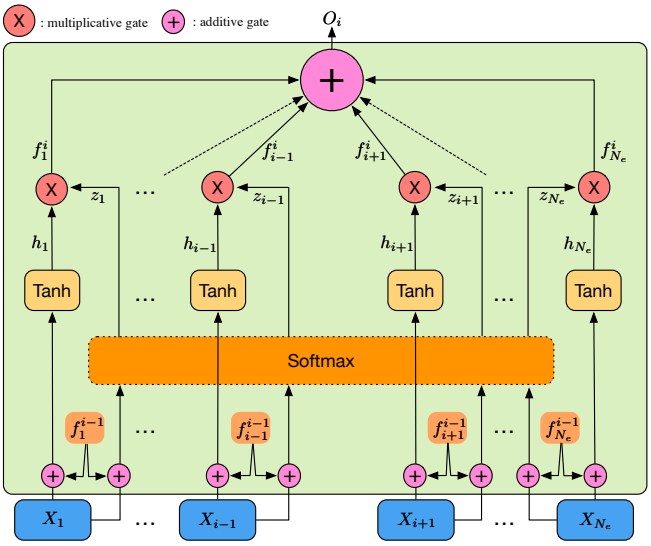

Figure 4: The unit of GRRE at the ith emitted glimpse. It models the label relation from all other emitted glimpses to the ith emitted glimpse.

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 ith emitted glimpse as illustrated in Figure 4.

$$h_k = \text{Tanh}(\mathbf{W}_{hxk}X_k + \mathbf{W}_{hfk}f_k^{i-1}) \tag{7}$$

$$\begin{aligned} z =& \text{Softmax}([(\mathbf{W}_{zx1}X_1 + \mathbf{W}_{zf1}f_1^{i-1}), \ldots, \\ & (\mathbf{W}_{zxN_e}X_{N_e} + \mathbf{W}_{zfN_e}f_{N_e}^{i-1})]) \\ f_k^i =& z_k * h_k,\ O_i = \sum_{k=1}^{N_e \setminus i} f_k^i \end{aligned} \tag{8}$$

where $h_k \in \mathbb{R}^{512}$ is an internal representation based on the current input $X_k \in \mathbb{R}^{512}$ and weighted features $f_k^{i-1} \in \mathbb{R}^{512}$ extracted from the last time step ($i-1$ in this case) at the kth emitted glimpse with a tanh activation function. $z \in \mathbb{R}^{N_e-1}$ denotes the contribution of other (N_e-1) emitted glimpses to current emitted glimpse, which is learned by using the Softmax function based on the concatenation of $X_k \in \mathbb{R}^{512}$ and $f_k^{i-1} \in \mathbb{R}^{512}$. $f_k^i \in \mathbb{R}^{512}$ is the weighted features of other emitted glimpses except the current ith emitted glimpse, which is computed via the multiplicative gates based on z_k and h_k, noting that $k \in (1, 2, \ldots, N_e)\setminus i$. Finally, we obtain the label relation $O_i \in \mathbb{R}^{512}$ of the ith emitted glimpse from all other emitted glimpses by summing them up. $\mathbf{W}_{[\cdot]}$ are the parameters to be learned and $[\cdot, \ldots, \cdot]$ is the concatenation operator. 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 ith 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.

$$P_i = f(\mathbf{W}_{xp}X_i + \mathbf{W}_{cp}C_i + \mathbf{W}_{op}O_i), P_i \in \mathbb{R}^D \tag{9}$$

where $\mathbf{W}_{[\cdot]}$ 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.

$$L = \frac{1}{N_e}\sum_{i=1}^{N_e} -\log \hat{P}_i \tag{10}$$

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.

Algorithm 1 The proposed method

Require: the input image I, emitted glimpse set Φ, emitted signal e
Preprocess: Initialize the parameters in our method
Initialize Φ = null
Extract the features $\mathbf{V_I}$ from VGG16
for each timestep t **do**
  Generate a candidate glimpse g_t based on $\mathbf{V_I}$ by using the RHN in Section 3.2
  $e_t \leftarrow$ Decide whether to emit the candidate glimpse by using emission indicator in Section 3.3
  if e_t = True **then**
    Push the g_t into Φ
    Compute the context features by using the context extractor in Section 3.2
  end if
end for
for each emitted glimpse g in Φ **do**
  Compute the label relation by using the GRRE in Section 3.4
  Compute the label prediction by Eqn.(9)
end for
Output: multiple labels predicted from the image I

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. β_1 and β_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 by 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 the 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 a 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 and RNN (LSTM/GRU). The ablation experiments demonstrate the effectiveness of different model components. From the results, we can find the following observations. First, GRRE can greatly boost the classification performance by comparing the variant **Ours w/o GRRE** and **Ours**. Second, GRRE achieves much better performance than the normal RNN by comparing CNN-RNN based models [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.

Method	P_L	R_L	F_L	P_I	R_I	F_I	MAP
KNN[4]	32.6	19.3	24.3	42.9	53.4	47.6	49.6
KNN[4] with VGG16	35.1	30.5	32.6	45.8	57.2	50.9	52.2
WARP[8]	31.7	35.6	33.5	48.6	60.5	53.9	54.9
CNN-LSTM[42]	36.7	38.2	37.4	49.2	61.0	54.5	57.3
CNN-RNN[36]	40.5	30.4	34.7	49.9	61.7	55.2	56.1
ResNet-50[10]	39.6	43.8	41.6	51.5	63.2	56.8	59.6
ResNet50+DCNN[30]	42.1	44.9	43.5	52.6	65.4	58.3	61.9
RIA[14]	39.2	41.6	40.4	51.3	62.9	56.5	59.5
LSEP[21]	**48.4**	42.7	45.4	**59.3**	60.8	60.0	60.6
Ours w/o CE w/o GRRE	34.2	31.6	32.8	45.3	55.9	50.1	47.5
Ours w/o GRRE	35.4	37.7	36.5	48.3	60.1	53.6	52.8
Ours w/o CE	43.4	45.0	44.2	52.7	65.6	58.4	61.5
Ours w/o EI	42.8	44.6	43.7	49.9	63.4	55.8	58.3
Ours w/o GRRE w. RNN	41.5	43.8	42.6	51.4	65.7	57.7	60.9
Ours	44.2	**49.3**	**46.6**	53.9	**68.7**	**60.4**	**63.5**

Table 1: Performance evaluation on NUS-WIDE dataset for $k = 3$.

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

Method	P_L	R_L	F_L	P_I	R_I	F_I	MAP
CNN-Softmax[36]	59.0	57.0	58.0	60.2	62.1	61.1	47.4
WARP[8]	59.3	52.5	55.7	59.8	61.4	60.7	49.2
CNN w/o RNN[36]	65.3	54.5	59.3	68.5	61.3	65.7	57.2
CNN-LSTM [42]	66.7	54.2	59.8	68.1	65.0	66.5	62.3
CNN-RNN[36]	66.0	55.6	60.4	69.2	66.4	67.8	61.2
WeakCNN[26]	-	-	-	-	-	-	62.8
RIA[14]	63.2	56.6	59.7	70.4	65.3	67.9	63.1
ResNet-50[10]	66.5	55.8	60.7	71.9	66.3	70.0	64.3
LSEP[21]	**73.5**	56.4	63.8	**76.3**	61.8	68.3	62.9
Ours	71.9	**59.6**	**65.2**	74.3	**69.7**	**71.8**	**66.7**

Table 2: Performance evaluation on MS-COCO dataset.

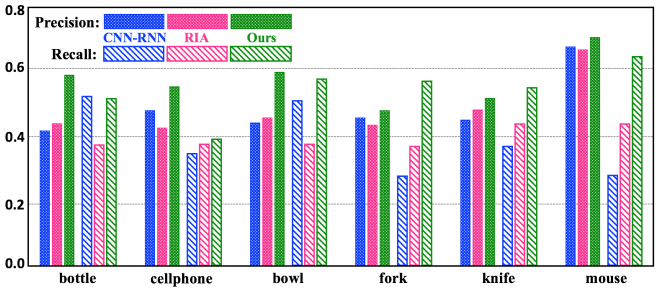

Figure 5: Comparison performances on small objects between CNN-RNN [36], RIA [14] and Ours.

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_L) and recall (R_L) 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.

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

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

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 benefits 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 la bels. 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 $\mathbf{V}_t$

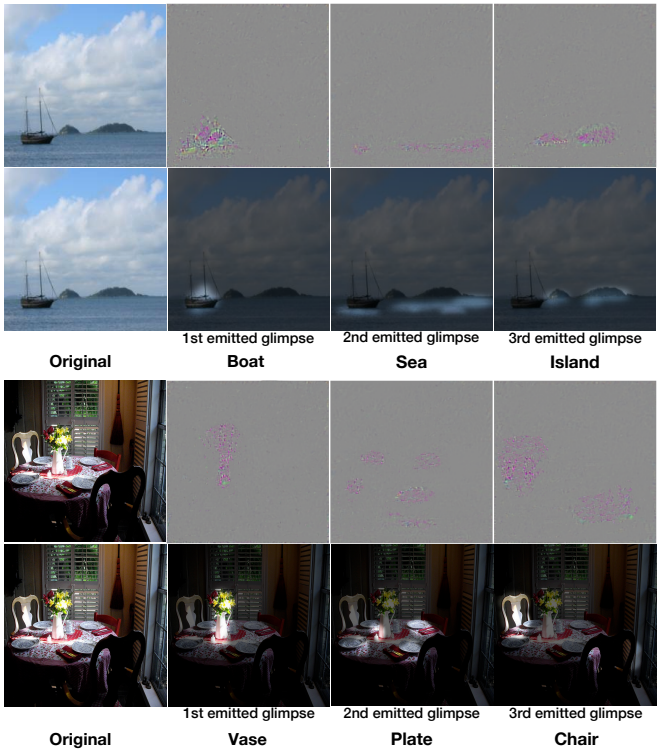

Figure 6: Visualization of regions with high activations at different emitted glimpses. The first column is the original image, and the rest are the high activation maps and corresponding highlighted multi-label images.

(consistent with Eq.2) at different emitted glimpses, and then apply the selected feature maps to Deconvolutional network [47] to generate the new images, which are in the same size of original images and intuitively show the emphasis at different emitted glimpses. 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, could benefit from the specific visual features.

Besides, in this supplementary materials, we show more experimental results of our proposed end-to-end attentive recurrent neural network for multi-label image classification under only image-level supervision, including the good examples, bad examples, the visually small objects, and even some objects which are not labelled by the ground truth.

All the above comparison experiments show that our model achieves better performances compared with state-of-the-art methods. Extensive in-depth analyses and visualizations are performed to demonstrate the effectiveness of the components of our model.

5 CONCLUSION

This paper proposes an 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, 61532018, 61332016, 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

[1] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. 2014. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755* (2014).
[2] Xiao Cai, Feiping Nie, Weidong Cai, and Heng Huang. 2013. New graph structured sparsity model for multi-label image annotations. In *ICCV*. 801–808.
[3] Xiaochun Cao, Hua Zhang, Xiaojie Guo, Si Liu, and Dan Meng. 2015. SLED: Semantic Label Embedding Dictionary Representation for Multilabel Image Annotation. *TIP* 24, 9 (2015), 2746–2759.
[4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *ACM international conference on image and video retrieval*. 48.
[5] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. 2011. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*.
[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE, 248–255.
[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*. 580–587.
[8] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. 2013. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894* (2013).
[9] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. 2009. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*. IEEE, 309–316.
[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* (2015).
[11] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. 2016. Learning structured inference neural networks with label relations. In *CVPR*. 2960–2968.
[12] Qinghao Hu, Jiaxiang Wu, Jian Cheng, Lifang Wu, and Hanqing Lu. 2017. Pseudo Label based Unsupervised Deep Discriminative Hashing for Image Retrieval. In *ACM Multimedia*.
[13] Yunho Jeon and Junmo Kim. 2017. Active Convolution: Learning the Shape of Convolution for Image Classification. In *CVPR*.
[14] Jiren Jin and Hideki Nakayama. 2016. Annotation order matters: Recurrent image annotator for arbitrary length image tagging. *arXiv preprint arXiv:1604.05225* (2016).
[15] Mahdi M Kalayeh, Haroon Idrees, and Mubarak Shah. 2014. Nmf-knn: Image annotation using weighted multi-view non-negative matrix factorization. In *CVPR*. IEEE, 184–191.
[16] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[17] Piotr Koniusz, Fei Yan, Philippe-Henri Gosselin, and Krystian Mikolajczyk. 2017. Higher-Order Occurrence Pooling for Bags-of-Words: Visual Concept Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 2 (2017), 313–327.

[18] Maksim Lapin, Matthias Hein, and Bernt Schiele. 2018. Analysis and Optimization of Loss Functions for Multiclass, Top-k, and Multilabel Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 7 (2018), 1533–1554.

[19] Qiang Li, Maoying Qiao, Wei Bian, and Dacheng Tao. 2016. Conditional graphical lasso for multi-label image classification. In *CVPR*. 2977–2986.

[20] Yunsheng Li, Mandar Dixit, and Nuno Vasconcelos. 2017. Deep Scene Image Classification With the MFAFVNet. In *ICCV*.

[21] Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving Pairwise Ranking for Multi-label Image Classification. In *CVPR*.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer, 740–755.

[23] Volodymyr Mnih, Nicolas Heess, Alex Graves, and koray kavukcuoglu. 2014. Recurrent Models of Visual Attention. In *Advances in Neural Information Processing Systems 27*. 2204–2212.

[24] Venkatesh N Murthy, Subhransu Maji, and R Manmatha. 2015. Automatic image annotation using deep learning representations. In *ICMR*. ACM, 603–606.

[25] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*. 1717–1724.

[26] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. 2015. Is object localization for free? – Weakly-supervised learning with convolutional neural networks. In *CVPR*.

[27] Duangmanee Putthividhy, Hagai T Attias, and Srikantan S Nagarajan. 2010. Topic regression multi-modal latent dirichlet allocation for image annotation. In *CVPR*. IEEE, 3408–3415.

[28] Ronald A Rensink. 2000. The dynamic representation of scenes. *Visual cognition* 7, 1-3 (2000), 17–42.

[29] Robin Senge, Juan José Del Coz, and Eyke Hüllermeier. 2014. On the problem of error propagation in classifier chains for multi-label classification. In *Data Analysis, Machine Learning and Knowledge Discovery*. Springer, 163–170.

[30] Weiwei Shi, Yihong Gong, Xiaoyu Tao, and Nanning Zheng. 2017. Training DCNN by Combining Max-Margin, Max-Correlation Objectives, and Correntropy Loss for Multilabel Image Classification. *IEEE TNNLS* (2017).

[31] Fuming Sun, Jinhui Tang, Haojie Li, Guo-Jun Qi, and Thomas S Huang. 2014. Multi-label image categorization with sparse factor representation. *TIP* 23, 3 (2014), 1028–1037.

[32] Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. 1999. Policy gradient methods for reinforcement learning with function approximation.. In *NIPS*, Vol. 99. 1057–1063.

[33] Tiberio Uricchio, Marco Bertini, Lorenzo Seidenari, and Alberto Bimbo. 2015. Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging. In *ICCV Workshops*. 9–15.

[34] Yashaswi Verma and CV Jawahar. 2012. Image annotation using metric learning in semantic neighbourhoods. In *ECCV*. Springer, 836–849.

[35] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 319–326.

[36] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. CNN-RNN: A Unified Framework for Multi-label Image Classification. *arXiv preprint arXiv:1604.04573* (2016).

[37] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. 2014. CNN: Single-label to multi-label. *arXiv preprint arXiv:1406.5726* (2014).

[38] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.

[39] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*. 842–850.

[40] Pengtao Xie, Ruslan Salakhutdinov, Luntian Mou, and Eric P. Xing. 2017. Deep Determinantal Point Process for Large-Scale Multi-Label Classification. In *ICCV*.

[41] Xiangyang Xue, Wei Zhang, Jie Zhang, Bin Wu, Jianping Fan, and Yao Lu. 2011. Correlative multi-label multi-instance image annotation. In *ICCV*. IEEE, 651–658.

[42] Geng Yan, Yang Wang, and Zicheng Liao. 2016. LSTM for Image Annotation with Relative Visual Importance. In *BMVC*.

[43] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. 2016. Exploit bounding box annotations for multi-label object recognition. In *CVPR*. 280–288.

[44] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *CVPR*. 21–29.

[45] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. *arXiv preprint arXiv:1603.03925* (2016).

[46] Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).

[47] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. 2010. Deconvolutional networks. In *CVPR*. IEEE, 2528–2535.

[48] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. 2017. Learning Spatial Regularization with Image-level Supervisions for Multi-label Image Classification. In *CVPR*.