Neural Computing and Applications (2022) 34:7271–7282
https://doi.org/10.1007/s00521-021-06866-y

ORIGINAL ARTICLE

Adaptive image annotation: refining labels according to contents and relations

Fen Xiao[1] · Yuyu Chen[1] · Yiming Zhang[1] · Xue Gong[1] · Xieping Gao[1]

Received: 24 February 2021 / Accepted: 12 December 2021 / Published online: 30 January 2022

Abstract
Image annotation has been an active research in computer vision. Most of the prior research works focus on annotating images with fixed number of labels, while it is unreasonable to annotate all images with the same number of labels and do not take into consideration their contents. In this paper, we present an extensive survey on the recent works about image annotation with label-to-image semantic relevance and propose a general framework for image adaptive annotation. Compared to previous works on image annotation methods, the proposed framework is novel in the following aspects: (1) It predicts label numbers of each image according to its visual features, which is more reasonable and practical for real-world image annotation. (2) It models label-to-image relevance with similar images and related labels, which can generate abundant candidate labels. (3) It can progressively refine the image label sets, which ensures the selected label set to be truly representative and with few redundancies. Experimental results on two benchmark multi-label image annotation datasets demonstrate that the proposed model outperforms the prior state-of-the-art approaches.

Keywords Image annotation · Variable length labels · Diverse labels · Similar images

1 Introduction

As an important and useful research topic in multimedia and computer vision fields, automatic image annotation aims at describing the contents of images with a set of semantic labels, which has attracted lots of researchers' interest. It is not only helpful to understand the semantics of images, but also widely used in various applications of digital image processing, such as image retrieval [25, 50] and caption generation [5]. Automatic image annotation well bridges semantic gaps between low-level features used to represent images and high-level semantic labels used to reflect image contents. On the other hand, due to the complex semantic relationship between image content and labels, label noises, etc., it is still a challenging task [36, 52].

A great number of approaches have been developed for automatic image annotation. Some early search-based annotation methods [16, 29, 36] mainly focused on image-to-label relevance, which tagged an image with labels of its semantic neighborhoods. How to measure the similarity between images is a key problem and many techniques have been proposed to deal with relations between images [53] such as search paradigms and visual similarity. Recently, image annotation was considered as a multi-label learning and classification problem with the help of deep learning methods [27, 28, 51]. Each label was assumed to be an individual category, and models can output some of labels with a series of pre-trained classifiers. And because of its simplicity and efficiency, some label classification methods have been widely used for annotation.

Different from the mutually exclusive image categories in classification, the labels in the scene are highly related. And how to integrate the relationships and encourage the diversity between labels to improve model performance becomes a new trend [14, 18]. There are a variety of methods, such as directed graphs [10, 46] and Recurrent

✉ Xieping Gao
xpgao@xtu.edu.cn

Fen Xiao
xiaof@xtu.edu.cn

[1] Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Xiangtan 411105, Hunan, China

Neural Networks (RNNs) [41], that have been used to reveal the co-occurrence relations between labels and obtained satisfactory annotation results. In addition, some other intuitive semantic relationship within labels, such as hierarchical structure ("vehicle" includes "car"), synonymous relationship("people" and "person"), which can be used to refine label and output more descriptive label sets [26, 43, 45, 47].

Kulesza et al. [22] proposed a conditional determinant point process (DPP) model, which models the probability distribution of a fixed-size label subset and then integrates the global negative correlation between the elements in the subset to ensure the diversity among them. Therefore, in order to ensure labels diversity, some image annotation methods embed the label relationship in the DPP model [43, 45]. These works assumed that all images have the same number of labels and set a fixed label length k.

Many other works [28, 41, 44, 48] also evaluated their performance using top-k retrieval performance, where k is fixed at 3 or 5. This assumption gives convenience for comparing performance between different methods, but it does not take into account the difference among semantic complexity of each image [20]. It is well known that the number of ground-truth labels increases with the complexity of the image content. As shown in Fig. 1a and b, the labels statistics on two widely used image annotation data sets(IAPRTC-12 and ESP-Game) are quite- different. Figure 1c and d gives some examples of annotated results. Conventional annotation algorithms are not effective in case of imbalanced label distribution. We found that images with salient foregrounds are more likely to be encoded with dominant objects, such as "bottle" or "tree". While for some complicated images, humans usually would like to depict them with more labels. Therefore, it is more practical to tagging images with different numbers of labels according to their contents. Predicting the number of labels in images and tagging them with variable length labels can make a good balance between the diversity and accuracy of image annotation.

Considering the diversity of the label subset and labels refinement, we propose an adaptive image annotation (AIA) method which progressively refines labels according to image contents and label relations. The model consists of three components: Label Length Prediction, Label-to-Image Relevance and Diverse Subset Inference. First, we use a pre-trained Inception-ResNet-V2 [35] as a feature extractor and predict the number of labels according to the feature distribution. Then, we give a model to simulate the label-to-image relevance with visual similarity and semantic relationship. Given an image to be labeled, we can get its diverse candidate label set which consists of related labels from its similar images. Finally, we use the WordNet-based weighted semantic paths as prior knowledge to refine labels.

Our main contributions are threefold: (i) We design an Label Length Prediction module to predict the variable label length of the image in line with the semantic information of the image. The predicted number of labels is more consistent with the actual tagging task. (ii) We use similar images and related labels to calculate the correlation between images and labels, and get a highly correlated and rich candidate labels set. (iii) We propose a new annotation method that treats image annotation as a label subset selection problem, which combines the label

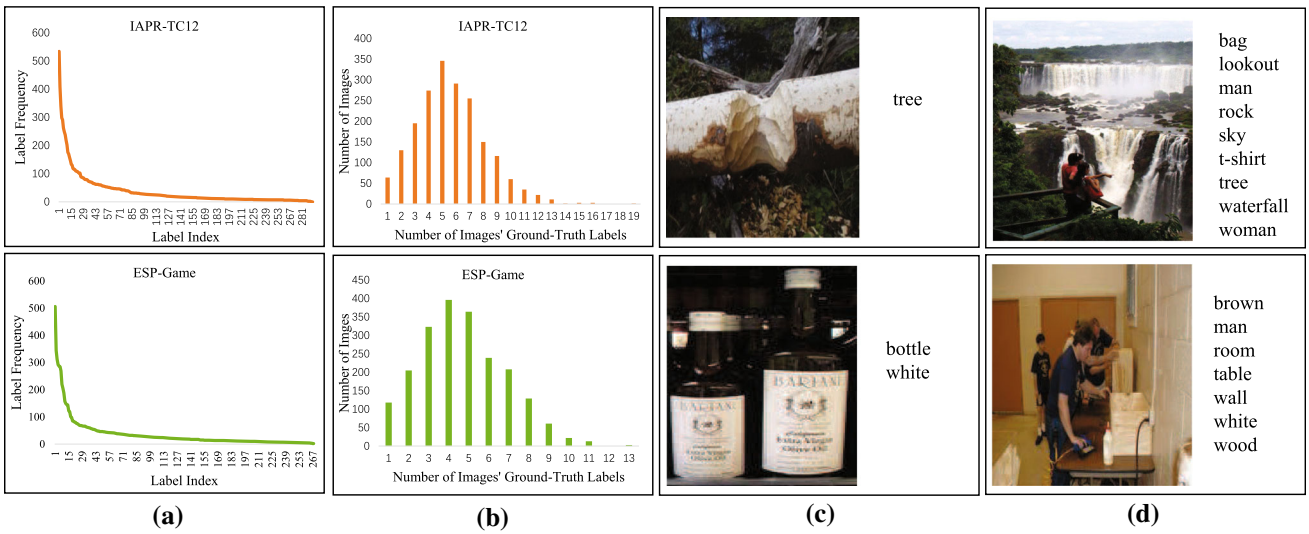

Fig. 1 The statistics of testing images on IAPRTC-12 and ESP-Game. **a** The number of times the label appears in the two datasets. **b** The number of images corresponding to different label lengths. **c** Examples of image annotation with salient foregrounds. **d** Examples of image annotations with complex content

semantic path and the sampling algorithm to select representative variable-length labels from the candidate labels. To evaluate the effectiveness of the proposed model, experimental results on ablation analysis, as well as performance, in comparison with several state-of-the-art methods on two benchmark datasets are reported.

2 Related work

2.1 Image annotation with image-to-label relevance

The primary goal of the image annotation methods with image-to-label relevance is to model the visual and textual relationship [1]. There are many methods that focus on image-to-label relevance to improve annotation accuracy, which can be broadly divided into generative models, nearest neighbor models and discriminative models [7].

Generative models aim to learn a joint distribution over image context features and semantic labels, and then predict the conditional probability of the label for the image to be labeled. Jeon et al. [19] proposed a cross media relevance model (CMRM), which clustered the segmented image blocks into blobs and learned the joint probability distribution between blobs and semantic labels to deduce a set of labels with the greatest correlation to the image. The model proposed in [11] was the joint probability distribution of labels and feature vectors from multiple image rectangular regions. They also used multiple Bernoulli model to estimate the predicted probability of each label. Foumani et al. [12] combined Convolutional Neural Network (CNN) and the Locality-constrained Linear Coding (LLC) to generate more informative visual words. To deal with label imbalance, they used a set of trained parameters to weight each label. The tr-mmLDA model [34] presented a topic-regression multi-modal Latent Dirichlet Allocation method, which incorporated a linear regression module to correlate two hidden topics of images and texts. It could capture correlations between image features and annotation texts.

Nearest neighbor models assume that visual similar images are much likely to share same labels, so the keywords of the nearest image can be assigned to the input test image. Makadia et al. [31] exploited the distances between unlabeled image and the labeled training images, and then assigned the labels of its nearest neighbors with a greedy algorithm. Li et al. [24] efficiently measured the label importance weight for each image by accumulating votes from visual neighbors' label distribution. In [15], the authors adopted logistic discriminant and weighted nearest neighbor algorithm to increase the importance of infrequent labels and suppress frequent labels. Verma et al. [39] proposed a 2-pass k-nearest neighbor algorithm with image-to-label and image-to-image similarities to optimize image annotation. Verma et al. [38] studied the diverse image annotation with missing labels (DIAML), which assumed that the training data to be partially labelled, and then annotate test images with a fixed number of labels that are simultaneously diverse, representative and maximally relevant. For DIAML task, they also proposed a new k-nearest neighbor (k-NN)-based algorithm, which used the Bernoulli process to indicate whether the label present and placed a Gaussian kernel over the image' feature map to calculate the similarity between images. Ma et al. [30] incorporated a multi-label linear discriminant classification method to assign different weights to image features, which acquired the k-nearest neighbors of the image more accurately. Wu et al. [42] used WordNet [9] to expand the candidate label set of unlabeled images obtained based on KNN, which pay more attention to the semantic correlations among textual labels.

Discriminative models treat each label as a class and learn an binary classifier for each label. Yu et al. [49] proposed the traditional low rank empirical risk minimization framework to learn the parameters of multi-label classifier by minimizing the empirical loss. In [4], the author trained two classifiers to evaluate image and label independently, and ensemble them into agreement via co-regularization in a joint loss function. With the development of deep learning, some annotation models use CNN to extract image feature with different convolution kernels and then use them to a classifier at the last layer. Niu et al. [32] proposed a multi-scale deep CNN model for fusing rich and discriminative features at different layers, which was effective for representing a wide range of visual concepts. Ke et al. [21] proposed an end-to-end feature pyramid annotation model. And a multi-label data enhancement method based on Wasserstein Generative Adversarial Network (WGAN) was proposed to reduce the over-fitting problem of small-scale datasets.

2.2 Image annotation with label relations

There are important and complex relationships between semantic labels, such as co-occurrence (e.g., "car" and "road") or semantic hierarchical relationship (e.g., "clothes" and "sweater"). Based on the above prior knowledge, label relationships can provide semantic clues for inferring label.

Some methods consider the co-occurrence relationship between labels, they either predict the labels in a sequential fashion or construct graph model based on label dependency. For example, Chen et al. [6] constructed a directed graph with the specified vertices and edges, where each node is a label and each edge weight is the label co-

occurrence probability between a pair of label. And, they used Graph Convolution Network (GCN) to map this label graphs to a set of inter-dependent label classifiers. Jin et al. [20] proposed a RIA model that forms image annotation problem as a sequence label generation, in which CNN was used to encode image as a visual feature vector and then RNN is utilized to decode the visual feature into a series of labels. It is the first work for very rare image annotations with arbitrary number of labels. However, the performance of RIA varies with the label order given in the training phase and rare-first order outperforms than dictionary order, random order, frequent-first order, and so on. However, it is not practical and difficult to rank labels in a order for a given sets of unlabeled images.

WordNet [9] is another method for describing semantic relationship between labels, which use multi-layer hierarchical structure to represent similar cluster, relations and sub-relations. The synonymous and hierarchical relationships included in WordNet can serve image annotation. Wu et al. [47] proposed a mixed graph to encode three kinds of label dependencies. It incorporates instance-level label similarity and class co-occurrence as undirected edges while semantic hierarchy is used as directed edges. This unified model performs well especially for dealing with noisy and missing labels. Lately, Wu et al. [45] proposed a diverse image annotation (DIA) model, which was the first work to encode the image-to-label correlation and the semantic relationship of labels in a DPP [23] sampling process. Furthermore, they constructed the hierarchical relationship of labels into weighted semantic paths to guide the model to sample the most diverse label subset from fixed-length labels set. Then, in [43], they optimized the DPP sampling process with a GAN to measure the relevance of image features and label sets. Chacko et al. [2] predicted the 5 category labels with the highest confidence based on CNN and then retrieved other semantic information about class labels among them through WordNet, such as hyponyms, hypernyms and their semantic similarity, which help in accurately tagging images.

3 Adaptive image annotation

Our AIA method mainly contains three modules: Label Length Prediction (LLP), Label-to-Image Relevance (LIR) and Diverse Subset Inference (DSI). We present more details of these three parts in the following subsections.

Given an image, we first utilize a pre-trained Inception-ResNet-V2 model to predict the number of labels according its contents. In order to make full use of the image-to-label relevance, we perform a content-based image retrieval and obtain plentiful candidate labels subsets which come from the labels of its similar image. We try to measure the similarity between images with the help of deep learning features. And then we embed image-to-labels relevance from the similar images and the related labels into the DPP model to optimize candidate labels. Finally, we utilize weighted semantic paths as semantic clue to sample more representative subset from candidate labels and obtain diverse labels as output. The flowchart of the proposed framework is shown in Fig. 2.

Notation Assume that $A = \{a_1, a_2, \ldots, a_c\}$ denotes the c possible annotation labels. The training image set is denoted as $X = [x_1, \ldots, x_n]$, where $x_j \in R^d$ denotes d dimensional feature vector of the j^{th} image, which extracted with some pre-trained deep learning models such as VGG and ResNet [45]. $y_j, \widetilde{y}_j$ represent the ground-truth and predicted label sets of the j^{th} image, respectively. And g_j represents the ground-truth number of the labels, $\widetilde{g}_j$ represents the predicted labels length, i.e. $g_j = |y_j|, \widetilde{g}_j = |\widetilde{y}_j|$.

3.1 Label length prediction

We notice that images with salient objects can be described with a small number of labels, while complex ones may require more labels to depict their confusing contents. The statistics about label distribution of two benchmark datasets ESP-Game and IAPRTC-12 are shown in Table 1. The total number of image labels in two datasets ESP-Game and IAPRTC-12 are 268 and 291, respectively. From the table, we can see that the maximum number of image label is 23. The standard deviation indicates a big number variation between images even in a same image dataset.

In order to establish the relationship between image content and label length, we formulate label length prediction as a regression problem. Therefore, we replace the final classification layer in Inception-ResNet-V2 [35] with three fully connected layers to solve the regression task. Given an image $x_j \in \Re^{H \times W \times 3}$, the modified Inception-ResNet-V2 is expected to extract the semantic information of the image. The output of the last fully connected layer specified as an integer value $\widetilde{g}_j$is provided as the predicted number of the image labels. We train the model to minimize Mean Squared Error (MSE), defined as follows:

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(\widetilde{g}_j - g_j\right)^2 \tag{1}$$

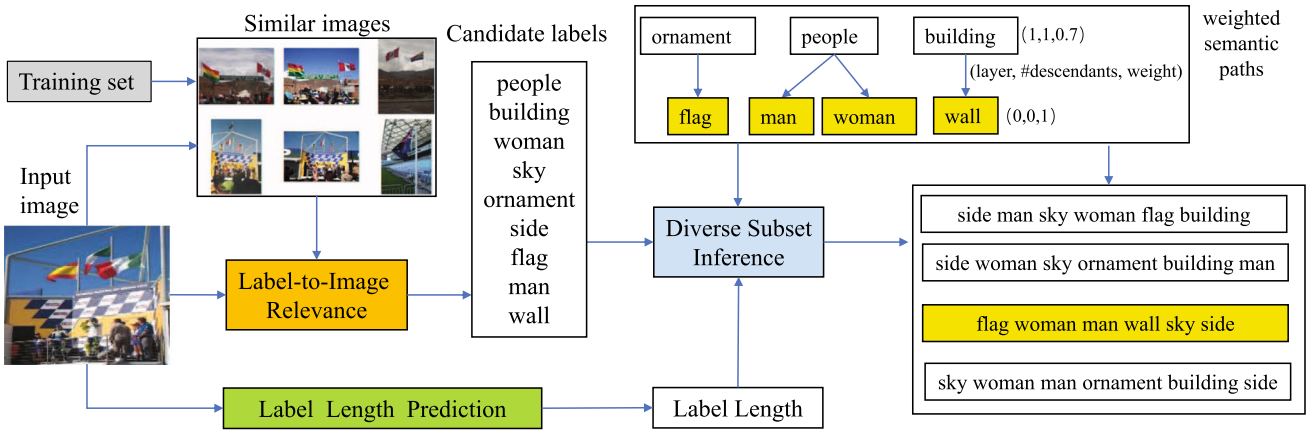

Fig. 2 Illustration of the proposed AIA method. It mainly contains Label Length Prediction (LLP), Label-to-Image Relevance (LIR) and Diverse Subset Inference (DSI). LLP is used to predict the number of labels for each image, LIR is used to generate candidate labels based on visual feature and similar images, and DSI is used to sample refined labels from the candidate labels

Table 1 Statistics of images and labels in two datasets

Dataset	ESP-Game	IAPRTC-12
Training images	18689	17495
Testing images	2081	1957
Labels	268	291
Labels per training image*	4.7/15/2.2	5.7/23/2.5
Labels per testing image*	4.6/13/2.2	5.6/19/2.5

* The form of "average/max/standard deviation"

where g_j is the ground-truth labels length, and n represents the number of images in the training set.

We adjust the size of all inputs images to 229×229, the learning rate is 1×10^{-4}, the decay is 1×10^{-6}, and batch size is 32. We also use the training sets of two datasets ESP-Game and IAPRTC-12 shown in Table 1 for training the LLP module, in which the label number for input image is set to the expected output.

3.2 Image-to-Label relevance

In this section, we exploit the relationships between image and label sets. Different from existed methods, the relevance in this section take both image visual features and label sets of similar images into consideration.

We consider the label associations between similar images firstly. Images are grouped together to c image subsets $M = [M_1, \ldots, M_c]$ according to their labels. Given an image, we build its semantic image set Z_M by selecting the most K similar images from each cluster (according to its labels) in M, Z_M consists of at most $c * k$ images. Here, we propose a deep learning method to measure the similarity between image pairs. The Euclidean distance of combined deep features extracted by the $F_{VGG-F}, F_{ResNet}, F_{VGG-6}, F_{VGG-7}$ is used to boost similarity accuracy. Then, we use parameter W to represent the correlation of image features to labels, this relevance $W = [w_1, \ldots, w_c] \in R^{d \times c}$ learned from training data by minimizing the negative log likelihood with ℓ_2 regularization [45]. It can be obtained the optimization problem as follows:

$$E(W) = -\frac{1}{n}\sum_{j=1}^{n} \log P_W(y_j|x_j) + \frac{\eta}{2}\sum_{i=1}^{c} \|w_i\|_2^2 \tag{2}$$

where P_W is the conditional DPP model, the sampling probability of the label subset is obtained through it. The parametric DPP is formulated as follows [23]:

$$P_W(y|x) = \frac{\det\left(L_y(x; W)\right)}{\det(L(x; W) + I)} \tag{3}$$

where I is an identity matrix, $\det(\cdot)$ means the determinant calculation. $L(x; W) \in R^{c \times c}$ is a positive semi-definite kernel matrix. The sub-matrix $L_y(x; W) = [L_{i,o}(x; W)_{a_i, a_o \in y}]$ is generated by selecting the rows and columns of $L(x; W)$ according to the label indexes in y:

$$L_{i,o}(x; W) = q_i(x) * S_A(i, o) * q_o(x), \forall i, o \in A \tag{4}$$

where $S_A(i, o)$ is the semantic similarity between label a_i and a_o which can be calculated by the following equation:

$$S_A(i,o) = \frac{1}{2} + \frac{\langle t_i, t_o \rangle}{\|t_i\|_2 \|t_o\|_2}, \forall i, o \in A \tag{5}$$

where $t_i, t_o \in R^{300}$ is the representation vector of labels a_i, a_o obtained by Glove [33]. Moreover, $\langle t_i, t_o \rangle$ denotes the inner product between t_i and t_o, $\|\cdot\|_2$ means the ℓ_2 norm of a vector. And in Eq.(4), $q_i(x)$ is used to measure the relevance between label a_i and input image x, which is defined as:

$$\begin{aligned} q_i(x) =& \mu \exp(0.5 w_i^T x) + \\ & (1-\mu) \sum\nolimits_M \exp(-D(x, Z_M)) \delta(a_i, Z_M)) \end{aligned} \tag{6}$$

where the former part of Eq. (6) mainly focuses on the relevance of image features and labels. The later part of Eq. (6) focuses on similar images share the same labels. $D(x, Z_M)$ is the Euclidean distance of features between image x and its similar image that in the subset Z_M. An indicator function $\delta(a_i, Z_M)$ takes a value 1 if similar image Z_M contains the label a_i and 0 otherwise. The parameter $\mu \in (0,1)$ trades off image's visual information and image's neighbor label information, and can be selected by doing cross-validation in training set. In fact, when $\mu = 1$, Eq. (6) is a special case given in [45].

For each image x_j , we can get the correlation information $q_i(x_j)$ between x_j and each label i with Eq. (6). We rank this vector and select the top $\lceil 1.5 * \widetilde{g}_j \rceil$ labels as the candidate labels set of image x_j, denoted as $\bar{y}_j$, in which $\widetilde{g}_j$ is the label length predicted by the LLP module.

3.3 Diverse subset inference

From the section 3.2, we can obtain abundant candidate labels $\bar{y}_j$ of each image. However, there may be some redundancy in these candidate labels. We should delete some labels and keep their integrity and descriptive at the same time. According to this, we explore the weight of each label in the coarse subset with the help of WordNet [9]. Given a particular subset, we can traverse the WordNet network to find semantic hierarchy ("pant" is a piece of "clothes") or synonyms ("people" and "person" have the same meaning) with related meanings.

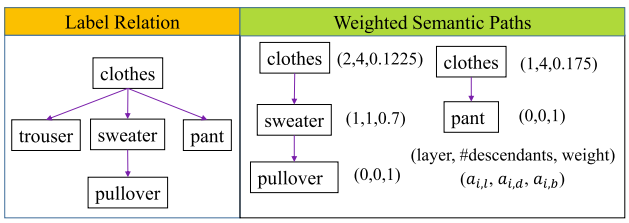

Fig. 3 Simple examples of weighted semantic paths

Weighted semantic paths are constructed based on the hierarchical relationship among all candidate labels [45]. Figure 3a shows a part of hierarchical relations between labels. Let ["clothes," "trouser," "sweater," "pant," "pullover"] denote the complete vocabulary of candidate labels. Their weights are closely related the semantic paths shown in Fig. 3b. Using WordNet, we can build directed paths such as ["clothes," → "sweater" → "pullover"] and ["clothes," → "pant"]. We encode each node i in a path with a triple set $(a_{i,l}, a_{i,d}, a_{i,b})$, which represent its layer, the number of descendants and weight, respectively. The weight of each label a_i in a given semantic path depends on the level of the label and the number of descendants labels of the label. For example, in semantic paths ["clothes," → "sweater" → "pullover"], "pullover" in bottom layers is a leaf node without descendant labels, and the weight of leaf node is set to 1. The weight of non-leaf node a_i in the semantic path is defined as:

$$a_{i,b} = \frac{\tau^{a_{i,l}}}{a_{i,d}} \tag{7}$$

where the factor $\tau = 0.7$ [45]. So, "clothes" in the left semantic path can be represented as (2,4,0.1225), which indicates that "clothes" is in the second layer, has four descendant labels in total, and its weight in this semantic path is $0.1225 = 0.7^2/4$.

For a given label set, there are many semantic paths, which are represented by $H_A = \{H_1, \ldots, H_r\}$. Although different paths may share labels, the weight of same label in different path may varies according to their level and descendants. For a given label, we add up all weights in the related semantic paths to get the final weight of the label. And then the weight of all labels can be denoted as $B = (a_{1,b}, \ldots, a_{i,b}, \ldots, a_{c,b})$.

3.4 AIA sampling algorithm

Algorithm 1 is a modified version of the standard k-DPP sampling algorithm [23], by embedding the LLP module in Line 1, Image-to-Label Relevance module in Line 2-3, and weighted semantic paths in Line 7-9. Given an unlabeled image, we use the LLP model to obtain the number of labels $\widetilde{g}_j$ and also use the extracted CNN features to retrieval similar images from the training set and named as Z_M. Then, we can get $1.5 * \widetilde{g}_j$ candidate labels with Eq. (6) according the image and label relevance. Subsequently, the k-DPP sampling algorithm and weight semantic paths are used to refine the coarse label set, as shown in Line 7-19. Line 9-11 ensures that two labels in the same semantic path will not be selected together. We run 10 samplings to obtain 10 different label subsets. Finally, we select the

subset with the largest sum of label weights in the 10 sampling process as the optimal label set $\widetilde{y}_j$.

algorithm 1 : Adaptive Image Annotation

Require: image x_j, semantic paths H_A, labels weights B;
Ensure: the label subset $\widetilde{y}_j$;
use the LLP model to predict label number $\widetilde{g}_j$ of x_j;
choose similar images Z_M by calculating the feature Euclidean distance between x_j and training set;
calculate and rank the probability of each label with the Eq. (6), and get the candidate label set $\bar{y}_j$ which consists of top $k_j = \lceil 1.5 * \widetilde{g}_j \rceil$ labels;
calculate the kernel matrix L with the Eq.(4), and build the sub-kernel matrix $L_{\bar{y}_j} = L(\bar{y}_j, \bar{y}_j)$;
compute eigenvalues λ_j of sub-matrix $L_{\bar{y}_j}$ and sort k_j eigenvalues by increasing order λ_j;
construct an matrix $e \in R^{(\widetilde{g}_j+1)\times(k_j+1)}$, in which $e_{0,i}$=1 for $i = 0, 1, \ldots, k_j$, $e_{l,0}$=0 for $l = 1, 2, \ldots, \widetilde{g}_j$, and $e_{l,i} = e_{l,i-1} + \lambda_{j,i} * e_{l-1,i-1}$ for $l > 0$ and $i > 0$;
for $t = 1, \ldots, 10$ **do**
  $\widetilde{y}_t = \emptyset, l = \widetilde{g}_j$;
  for $i = k_j, \ldots, 1$ **do**
    if $\widetilde{y}_{k_j}(i)$ is the same semantic path in H_A with any label in $\widetilde{y}_t$ **then**
      skip to next iteration;
    end if
    if $u \sim U[0,1] < \lambda_{j,i} \frac{e_{l-1,i-1}}{e_{l,i}}$ **then**
      $\widetilde{y}_t \leftarrow \widetilde{y}_t \cup \widetilde{y}_{k_j}(i), l \leftarrow l - 1$;
    end if
    if $l = 0$ **then** Break **end if**
  end for
  compute label subset weights $\tilde{B}\widetilde{y}_{t'} = \sum_j^{|\widetilde{y}_t|} B\widetilde{y}_t$;
end for
return $\widetilde{y}_j = \arg\max_{\widetilde{y}_{t'}} \tilde{B}\widetilde{y}_{t'}$;

4 Experiments

4.1 Experimental setting

Datasets. We use two benchmark image annotation datasets IAPRTC-12 and ESP-Game for experiment comparison. The IAPRTC-12 [13] dataset consists of 19627 images and there are 291 labels. The minimum and maximum number of image labels are 1 and 23 respectively. The ESP-Game [40] dataset includes a wide variety of images that consist of drawings, logos, personal photos, etc. There are 20770 images and the size of vocabulary is 268. The minimum label length of training images is 1 and the maximum is 15. For both datasets, we follow the training/test partition used in the normal way [37]. Some statistics of these datasets are shown in Table 2.

Comparison methods. We compare the annotation results of our method with five state-of-the-art image annotation methods including LEML [49], MLMG [47], DIA [45], D^2-GAN [43] and RIA [20]. The first four methods were proposed for predicting fixed number of label sets, and the last one is the baseline for variable length image annotation. The comparison experiments conducted according to the two conditions: Annotate image with the given the number of labels(3 or 5 as normal) and annotate with different number of labels. Because RIA method did not report results of fixed number tags, we do not compare with RIA in the first comparison experiment. We revise the first four method LEML, MLMG, DIA, D^2-GAN with our proposed LLP module and named as LEML-LLP, MLMG-LLP, DIA-LLP, D^2-GAN-LLP, respectively.

Table 2 The mean absolute error and accuracy of LLP

Datasets	Method	MAE↓	Accuracy(%)↑
ESP-Game	LLP	**1.53**	**21.09**
	5-tags	1.74	17.49
	3-tags	2.12	15.52
IAPRTC-12	LLP	**1.45**	**23.09**
	5-tags	1.97	17.68
	3-tags	2.90	9.96

Bold values indicate the best numerical result in the current column

RestNet152 [17], VGG16 [8] and VGG-F [3] are three widely used networks for image feature extraction [35]. DIA and D^2-GAN used pre-trained VGG-F model to extract a 4096-dimensional feature vector and got satisfactory results. Thus, for a fair comparison, we also use this model to extract the same dimensional features to measure image similarity and the corresponding method was named as AIA-Single. Furthermore, we investigate the performance of higher dimensional feature vector, which obtained with pre-trained VGG-F, VGG16 and RestNet152. We measure image similarity based on the combined features and keep the remaining steps same as AIA-Single, the corresponding method named as AIA-Mix.

Metrics. The standard metrics are precision, recall, and F1-measure, which are widely used in previous work. The corresponding semantic metrics based on weighted semantic paths, proposed by Wu et al. [45], can be used to evaluate relevancy and diversity more meaningful and precise. The semantic metrics are computed as follows:

$$SP_j = \frac{\left|H_{y_j} \bigcap H_{\widetilde{y}_j}\right|}{\left|H_{\widetilde{y}_j}\right|} \tag{8}$$

$$SR_j = \frac{\left|H_{y_j} \bigcap H_{\widetilde{y}_j}\right|}{\left|H_{y_j}\right|} \tag{9}$$

$$SF1_j = \frac{2(SP_j \times SR_j)}{(SP_j + SR_j)} \tag{10}$$

where H represents the semantic paths, $|.|$ represents the number of semantic paths. $|H_{y_j} \bigcap H_{\widetilde{y_j}}|$ is the weight of correct labels. *SP*, *SR* and *SF*1 are the average metrics value of all images in the dataset. The higher value indicates the better performance of the method.

4.2 Effective evaluation for LLP

We quantify the efficiency of LLP for labels number prediction with two metrics, Mean Absolute Error (MAE) and Accuracy. We choose the number of tags for each image g_j as a reference and use metrics to demonstrate the difference between fixed number(3 or 5) and predicted variable number. The former metric is the mean absolute error between the predicted quantities and ground-truth quantities, and the later metric is the proportion of correct predicted quantities in all testing images. The MAE and Accuracy are defined as follows:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|\widetilde{g_j} - g_j\right| \tag{11}$$

$$Accuracy = \frac{1}{n}\sum_{j=1}^{n}\delta\left(\widetilde{g_j}, g_j\right) \tag{12}$$

For fixed label numbers, we do not use LLP to predict the number of labels and assume the predicted quantity $\widetilde{g_j} = 3$ or $\widetilde{g_j} = 5$, and indicator function $\delta\left(\widetilde{g_j}, g_j\right)$ takes value 1 if $\widetilde{g_j} = g_j$, otherwise 0.

The MAE and Accuracy results of the LLP predicted label length and the traditional fixed label length are shown in Table 2. From statistics of image datasets, the average label number of testing images on the ESP-Game and IAPRTC-12 datasets are 4.6 and 5.6, respectively. It means that fixed label number 5 is closer to average number compare to 3-labels, and MAE and Accuracy values of 5-labels should be higher than those of 3-labels. Moreover, the MAE and Accuracy values with LLP are much better than 5-labels. We can make a conclusion that our quantity prediction is more in accordance with the ground-truth quantity than conventional fixed-length.

4.3 Comparison of models with fixed label length

According to most of the annotation methods proposed for a fixed number of labels k, we also adjust our method to let the model predict a given number of labels. Our method is compared with four image annotation methods that utilize label relationships, including LEML, MLMG, DIA, and D^2-GAN. It should be noted that DIA, D^2-GAN and our approach utilized diverse subset inference to ensure labels diversity. The main difference between DIA and AIA-Single is that DIA only used the first half of Eq. (6).

Table 3 Semantic metrics results (%) of methods on ESP-Game

Method	3 tags			5 tags		
	SP	*SR*	*SF*1	*SP*	*SR*	*SF*1
MLMG [47]	30.51	16.55	19.73	36.61	29.63	30.59
LEML [49]	45.16	23.61	28.31	**41.82**	33.87	34.58
DIA [45]	42.37	30.48	33.43	36.15	40.1	35.96
D^2-GAN [43]	42.96	32.34	34.93	35.04	41.50	36.06
AIA-Single	44.59	33.39	36.20	36.90	43.95	38.08
AIA-Mix	**47.84**	**36.34**	**39.09**	39.53	**47.47**	**40.87**

Bold values indicate the best numerical result in the current column

Table 4 Semantic metrics results (%) of methods on IAPRTC-12

Method	3 tags			5 tags		
	SP	*SR*	*SF*1	*SP*	*SR*	*SF*1
MLMG [47]	35.74	17.99	21.89	41.95	29.56	31.98
LEML [49]	43.03	19.54	24.86	47.27	29.76	33.67
DIA [45]	44.01	25.16	30.13	38.91	34.21	34.23
D^2-GAN [43]	43.57	26.22	31.04	37.31	35.35	34.41
AIA-Single	47.70	28.81	33.93	42.30	40.07	38.96
AIA-Mix	**53.12**	**32.11**	**37.85**	**47.03**	**44.88**	**43.42**

Bold values indicate the best numerical result in the current column

We present semantic metrics results of the proposed model on two benchmark datasets, which are shown in Tables 3 and 4. The first three lines are the reported results in [45], and the fourth line is reported results in [43]. It is evident that our AIA method exhibits superior performance for both 3 and 5 labels on two datasets. In the 5-tags evaluation, compared with D^2-GAN, *SF*1 score of AIA-Single increases 2.02% and 4.55% on ESP-Game and IAPRTC-12, respectively. And, compared with DIA, *SF*1 score of AIA-Single increases 2.12% and 4.73% on two datasets, respectively. The reason maybe the kernel matrix measures robust relevance between images and labels, and it can generate more correct and diverse candidate labels. Moreover, in the 5-tags evaluation, compared with AIA-Single, the *SF*1 scores of AIA-Mix on the two datasets are increased by 2.79% and 4.46%, respectively. This shows that multiple CNN-based features can boost AIA to achieve better performance than a single CNN-based feature. Also, our method has more improvements on IAPRTC-12 because the ESP-Game dataset is more challenging than IAPRTC-12.

4.4 Comparison of models with variable label length

Our method is first compared with the published results of the RIA method, which is a method for predicting variable length labels. Then, our method is compared with the LEML, MLMG, and DIA methods that use our label length prediction as a priori knowledge. For convenience, we renamed three methods to LEML-LLP, MLMG-LLP and DIA-LLP.

RIA is influenced by label sequence order and provided four results in different rules. For a fair comparison, we use the same precise, recall and F1-score metrics to measure our method, and the results are presented in Table 5. On ESP-Game dataset, AIA-Single-LLP outperforms RIA in frequent-first order and random order. On IAPRTC-12 dataset, AIA-Single-LLP outperforms RIA in frequent-first order, random order and rare-first order. Especially, AIA-Mix-LLP outperforms RIA in four orders, F1-Score of AIA-Mix-LLP increases 1.5% and 4.6% than the best result of RIA on ESP-Game and IAPRTC-12, respectively.

And then, our method is compared with LEML-LLP, MLMG-LLP and DIA-LLP, which use the result of label length prediction as a priori. The semantic metrics results of these methods are shown in Table 6. It can be seen that our method achieves better performance in multiple metrics. Compared with DIA-LLP, from semantic *SF*1 score, AIA-Single-LLP increases 1.95% and 4.75% on ESP-Game and IAPRTC-12, respectively. Compared with AIA-Single-LLP, *SF*1 score of AIA-Mix-LLP increases 2.82% and 4.84% on ESP-Game and IAPRTC-12, respectively. At the same time, we can compare the SF1 of AIA-mix in Tables 3–4 and SF1 of AIA-mix-LLP in Table 6 to infer the effectiveness of LLP in our method.

Moreover, we further count the number of images of methods at different *SF*1 scores, and results are shown in Fig. 4. When *SF*1 score is larger than 0.5, our method

Table 5 Precise, recall and F1-score Results (%) of variable number label annotation methods on ESP-Game and IAPRTC-12

Datasets	Method	*P*	*R*	*F*1
ESP-Game	RIA(frequent-first) [20]	34	23	24
	RIA(random) [20]	36	24	27
	RIA(dictionary) [20]	32	29	29
	RIA(rare-first) [20]	33	31	31
	AIA-Single-LLP	33.2	27.9	27.9
	AIA-Mix-LLP	**37.9**	**32.1**	**32.5**
IAPRTC-12	RIA(frequent-first) [20]	31	20	22
	RIA(random) [20]	33	25	28
	RIA(dictionary) [20]	32	28	29
	RIA(rare-first) [20]	35	34	34
	AIA-Single-LLP	36.0	30.6	30.7
	AIA-Mix-LLP	**45.8**	**37.7**	**38.6**

Bold values indicate the best numerical result in the current column

Table 6 Semantic metrics results(%) of methods with label number prediction on ESP-Game and IAPRTC-12

Datasets	Method	*SP*	*SR*	*SF*1
ESP-Game	LEML-LLP	31.87	36.34	32.41
	MLMG-LLP	34.21	37.94	34.34
	DIA-LLP	36.46	41.38	36.99
	AIA-Single-LLP	38.08	43.81	38.94
	AIA-Mix-LLP	**40.96**	**46.95**	**41.76**
IAPRTC-12	LEML-LLP	37.26	33.66	33.99
	MLMG-LLP	39.06	35.47	35.81
	DIA-LLP	39.18	35.85	36.00
	AIA-Single-LLP	42.93	41.53	40.75
	AIA-Mix-LLP	**48.36**	**46.15**	**45.59**

Bold values indicate the best numerical result in the current column

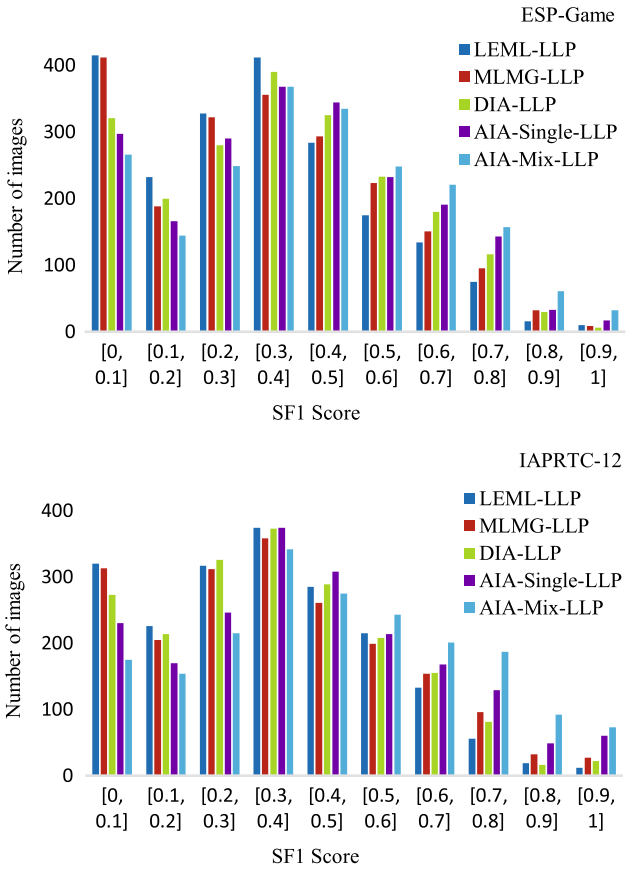

Fig. 4 The number of images in five methods at different semantic *SF*1 scores

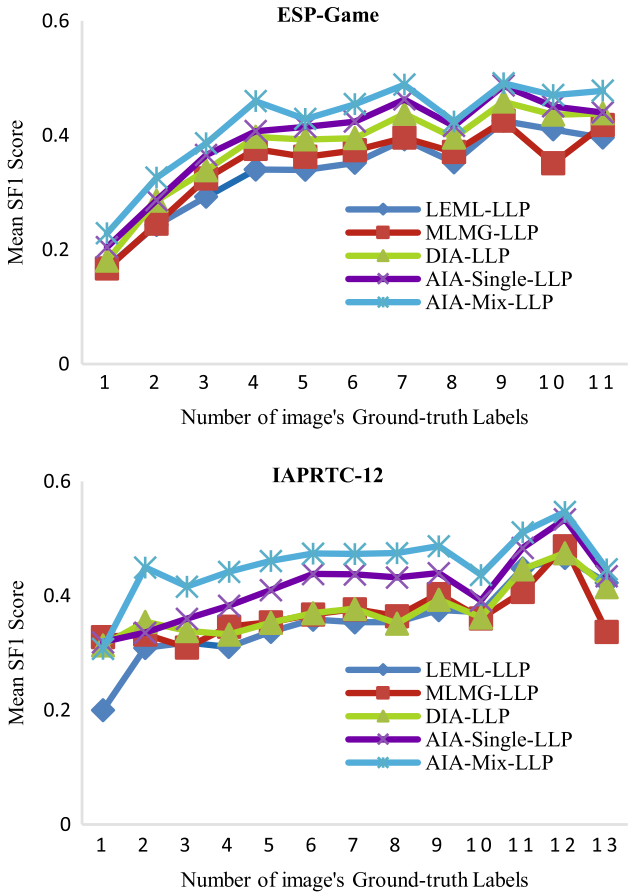

Fig. 5 The mean *SF*1 scores of five methods at different images' ground-truth labels

Fig. 6 Several image annotation results of four methods on IAPRTC-12. Labels in bold are consistent with ground-truth (GT) labels

contains more images compared to other methods. While *SF*1 score is lower than 0.5, our method contains less images.

Finally, we show mean *SF*1 scores of methods at different number of ground-truth labels in testing images, and the results are presented in Fig. 5. The overall trend is that as number of ground-truth labels increases, so does *SF*1 scores. Moreover, it can be seen that our method outperforms other methods when the predicted labels are variable.

Also, several image annotation results of four methods are shown in Fig. 6. Labels in bold are consistent with ground-truth labels. Look for the first image in Fig. 6, "a **statue** of a **man** on a **grey base**," MLMG predicted three labels are "adult," "people," and "person," they are related to "**man**." However, predicted labels of MLMG exist redundancy. DIA predicted three labels are "**statue**," "sky," and "**man**." And our method predicted four labels are "**statue**," "**base**," "front," and "**man**." Both DIA and our method predicted more diverse labels than MLMG, but our method predicted more correct labels than DIA, especially our method predicted four labels as same as the number of ground-truth label length.

5 Conclusion

In this paper, we proposed a variable length annotation model with diversity. Different from the conventional methods, our methods can generate label subsets with variable length according to the complexity of image semantics. The label quantity subproblem was solved by CNN architecture with regress layers. Furthermore, in order to avoid label redundancy, our method construct a robust label-to-image relevance to obtain the candidate labels, and then the optimal labels were decided from this candidate labels via weighted semantic paths. In order to evaluate the performance of our method, we first compare the proposed method with some state-of-the-art methods under the conventional fixed-length evaluation setting on two datasets. And then, our method is compared with the published results of RIA, which generated label sequence with variable length. Finally, we compared with several methods that use our label quantity as prior knowledge. From the experimental results on two datasets, our method outperforms several state-of-the-art methods.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 61771415, 61802328), Natural Science Foundation of Hunan province in China (Grant No.

2018JJ2405), Scientific Research Fund of Hunan Provincial Education Department (Grant No. 18K034).

Declarations

Conflict of interest All authors disclosed no relevant relationships. There are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Bhagat P, Choudhary P (2018) Image annotation: then and now. Image Vision Comput 80:1–23
2. Chacko JS (2018) Tulasi B Semantic image annotation using convolutional neural network and wordnet ontology. Int J Eng Technol 7(2.27):56–60
3. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531
4. Chen M, Zheng A, Weinberger K (2013) Fast image tagging. In: ICML, pp 1274–1282
5. Chen S, Jin Q, Wang P, Wu Q (2020) Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9962–9971
6. Chen ZM, Wei XS, Wang P, Guo Y (2019) Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5177–5186
7. Cheng Q, Zhang Q, Fu P, Tu C, Li S (2018) A survey and analysis on automatic image annotation. Pattern Recogn 79:242–259
8. Donahue J, Jia Y, Vinyals O, Hoffman J, Ning Z, Tzeng E, Darrell T (2014) Decaf: a deep convolutional activation feature for generic visual recognition. In: ICML, pp 647–655
9. Fellbaum C (1998) Wordnet: an electronic lexical database. Libr Q Inf Commun Policy 25(2):292–296
10. Feng L, Bhanu B (2016) Semantic concept co-occurrence patterns for image annotation and retrieval. IEEE Trans Pattern Anal Mach Intell 38(4):785–799
11. Feng SL, Manmatha R, Lavrenko V (2004) Multiple bernoulli relevance models for image and video annotation. In: CVPR, pp 1002–1009
12. Foumani SNM, Nickabadi A (2019) A probabilistic topic model using deep visual word representation for simultaneous image classification and annotation. J Visual Commun Image Represent 59:195–203
13. Grubinger M, Clough P, Muller H, Deselaers T (2006) The IAPR benchmark: a new evaluation resource for visual information systems. In: ICLRE, pp 13–23
14. Gu Y, Qian X, Li Q, Wang M, Hong R, Tian Q (2015) Image annotation by latent community detection and multikernel learning. IEEE Trans Image Process 24:3450–3463
15. Guillaumin M, Mensink T, Verbeek J, Schmid C (2009) Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV, pp 309–316
16. Guo QJ, Li N, Yang YB, Wu GS (2014) Image annotation by modeling supporting region graph. Appl Intell 40(3):389–403
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR, pp 770–778
18. Hu H, Zhou G, Deng Z, Liao Z, Mori G (2016) Learning structured inference neural networks with label relations. In: CVPR, pp 2960–2968
19. Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: ACM SIGIR, pp 119–126
20. Jin J, Nakayama H (2016) Annotation order matters: recurrent image annotator for arbitrary length image tagging. In: ICPR, pp 2452–2457
21. Ke X, Zou J, Niu Y (2019) End-to-end automatic image annotation based on deep CNN and multi-label data augmentation. IEEE Trans Multimed 21(8):2093–2106
22. Kulesza A, Taskar B (2011) k-dpps: Fixed-size determinantal point processes. In: ICML, pp 1193–1200
23. Kulesza A, Taskar B (2012) Determinantal point processes for machine learning. arXiv preprint arXiv:1207.6083
24. Li X, Snoek CGM, Worring M (2009) Learning social tag relevance by neighbor voting. IEEE Trans Multimed 11(7):1310–1322
25. Li X, Uricchio T, Ballan L, Bertini M, Snoek C, Bimbo A (2015) Socializing the semantic gap: a comparative survey on image tag assignment, refinement and retrieval. ACM Comput Surv 49(1):1–14
26. Liang X, Zhou H, Xing E (2018) Dynamic-structured semantic propagation network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 752–761
27. Lu D, Weng Q (2007) A survey of image classification methods and techniques for improving classification performance. Int J Remote Sens 28(5):823–870
28. Lyu F, Wu Q, Hu F, Wu Q, Tan M (2019) Attend and imagine: multi-label image classification with visual attention and recurrent neural networks. IEEE Trans Multimed 21(8):1971–1981
29. Ma Y, Liu Y, Xie Q, Li L (2019) CNN-feature based automatic image annotation method. Multimed Tools Appl 78(3):3767–3780
30. Ma Y, Xie Q, Liu Y, Xiong S (2019) A weighted kNN-based automatic image annotation method. Neural Comput Appl, 1–12
31. Makadia A, Pavlovic V, Kumar S (2008) A new baseline for image annotation. In: ECCV, pp 316–329
32. Niu Y, Lu Z, Wen JR, Xiang T, Chang SF (2018) Multi-modal multi-scale deep learning for large-scale image annotation. IEEE Trans Image Process 28(4):1720–1731
33. Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. In: EMNLP, pp 1532–1543
34. Putthividhy D, Attias HT, Nagarajan SS (2010) Topic regression multi-modal latent dirichlet allocation for image annotation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 3408–3415. IEEE
35. Szegedy C, Ioffe S, Vanhoucke V (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261
36. Tang C, Liu X, Wang P, Zhang C, Li M, Wang L (2019) Adaptive hypergraph embedded semi-supervised multi-label image annotation. IEEE Trans Multimed 21(11):2837–2849. https://doi.org/10.1109/TMM.2019.2909860
37. Tatler, Benjamin, W (2008) A new baseline for image annotation. In: ECCV, pp 316–329
38. Verma Y (2019) Diverse image annotation with missing labels. Pattern Recogn, 93, 470–484. https://doi.org/10.1016/j.patcog.2019.05.018. http://www.sciencedirect.com/science/article/pii/S0031320319301931
39. Verma Y, Jawahar CV (2016) Image annotation by propagating labels from semantic neighbourhoods. Int J Comput Vis, 1–23
40. von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: ACM SIGCHI, pp 319–326
41. Wang J, Yang Y, Mao J, Huang Z, Huang C, Xu W (2016) Cnn-rnn: A unified framework for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294

42. Wei W, Wu Q, Chen D, Zhang Y, Liu W, Duan G, Luo X (2021) Automatic image annotation based on an improved nearest neighbor technique with tag semantic extension model. Proc Comput Sci 183:616–623
43. Wu B, Chen W, Sun P, Liu W, Ghanem B, Lyu S (2018) Tagging like humans: Diverse and distinct image annotation. In: CVPR, pp 7967–7975
44. Wu B, Chen W, Sun P, Liu W, Ghanem B, Lyu S (2018) Tagging like humans: Diverse and distinct image annotation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 7967–7975. https://doi.org/10.1109/CVPR.2018.00831
45. Wu B, Jia F, Liu W, Ghanem B (2017) Diverse image annotation. In: CVPR, pp 6194–6202
46. Wu B, Jia F, Liu W, Ghanem B, Lyu S (2018) Multi-label learning with missing labels using mixed dependency graphs. Int J Comput Vis 126(8):875–896
47. Wu B, Lyu S, Ghanem B (2015) Ml-mg: Multi-label learning with missing labels using a mixed graph. In: ICCV, pp 4157–4165
48. Wu Y, Zhai H, Li M, Cui F, Wang L, Patil N (2019) Learning image convolutional representations and complete tags jointly. Neural Comput Appl 31(7):2593–2604
49. Yu H, Jain P, Kar P, Dhillon D (2014) Large-scale multi-label learning with missing labels. In: ICML, pp 593–601
50. Yuan BH, Liu GH (2020) Image retrieval based on gradient-structures histogram. Neural Comput Appl 32(15):11717–11727
51. Yuan C, Wu Y, Qin X, Qiao S, Pan Y, Huang P, Liu D, Han N (2019) An effective image classification method for shallow densely connected convolution networks through squeezing and splitting techniques. Appl Intell 49(10):3570–3586
52. Zhang J, He Z, Zhang J, Dai T (2019) Cograph regularized collective nonnegative matrix factorization for multilabel image annotation. IEEE Access 7:88338–88356. https://doi.org/10.1109/ACCESS.2019.2925891
53. Zhang J, Wu Q, Zhang J, Shen C, Lu J (2019) Mind your neighbours: Image annotation with metadata neighbourhood graph co-attention networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2956–2964