INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2022.16.23 Volume 16, 2022

The Image Annotation Refinement in Embedding Feature Space based on Mutual Information

Wei Li, Haiyu Song, Hongda Zhang,Houjie Li, Pengjie Wang
School of Computer Science and Engineering, Dalian Minzu University,
Dalian 116600,
China

Received: July 2, 2021. Revised: December 3, 2021. Accepted: January 8, 2022. Published: January 10, 2022.

Abstract—The ever-increasing size of images has made automatic image annotation one of the most important tasks in the fields of machine learning and computer vision. Despite continuous efforts in inventing new annotation algorithms and new models, results of the state-of-the-art image annotation methods are often unsatisfactory. In this paper, to further improve annotation refinement performance, a novel approach based on weighted mutual information to automatically refine the original annotations of images is proposed. Unlike the traditional refinement model using only visual feature, the proposed model use semantic embedding to properly map labels and visual features to a meaningful semantic space. To accurately measure the relevance between the particular image and its original annotations, the proposed model utilize all available information including image-to-image, label-to-label and image-to-label. Experimental results conducted on three typical datasets show not only the validity of the refinement, but also the superiority of the proposed algorithm over existing ones. The improvement largely benefits from our proposed mutual information method and utilizing all available information.

Keywords—Annotation refinement; mutual information; semantic embedding.

I. Introduction

With the development of internet and digital imaging technologies, more and more people like to share photos on social networks (e.g. YouTube, Facebook, Twitter, and Wechat) and the number of digital images are proliferating faster than the expectation. Moreover, the user used to query in a manner like natural language, instead of depending on low-level visual feature or example image. To solve these problems in image retrieval, image annotation is proposed to label the image with keywords, which helps in the intelligent retrieval of relevant images through simple query representation with the keywords of the image[1]. The assignment of keywords can be performed manually or automatically. Because of the disadvantages of manual image annotation in objectivity and subjectivity, Automatic Image Annotation (AIA) is the main tendency. AIA can automatically assign semantic keywords according to the visual information of images, so that images can be retrieved by semantic keywords, and images can be organized and managed by traditional relation database[2].

In recent two decades, researches on AIA have made a great extent of development. The typical methods include CRM[3], MBRM[4], JEC[5], 2PKNN[6], TagProp[7], D^2IA[8], etc. Semantic gap is one of the most challenges of AIA [6]. Many AIA methods focus on bridging the semantic gap, however, the issue has not been resolved thoroughly. Although some of the methods have achieved better performance in the ideal image databases created by experts, such as JEC, MBRM, 2PKNN, and D^2IA, they are not suitable for the realistic social images from content-sharing websites and social networks. In contrast to traditionally well-annotated image databases by experts, user-provided tags from those databases usually are subjective, incomplete, containing noisy words. Despite continuous efforts in inventing new annotation methods, it would be advantageous to develop a dedicated approach that could refine imprecise annotations.

Jin is the earliest researcher of image annotation refinement, who proposed the concept of image annotation improvement on ACM MM 2005 and gave the method WNM based on WordNet [9]. For a query image, an existing image annotation method is first employed to obtain a set of candidate annotations. Then, the candidate annotations are re-ranked and only the top ones are reserved as the final annotations. After the refinement of image annotation, the image annotation results are improved because the noise words are removed and the initial annotations are optimized. However, the existing methods on image annotation refinement only focus on label-to-label correlation without noting the quantitative correlation on image-to-label and image-to-image. Furthermore, image-to-label correlation and

image-to-image correlation can reflect more information in the image.

In this paper, we propose a novel image annotation refinement method based on multi-modal mutual information in embedding feature space to improve the AIA results. The Mutual Information is used to measure relevance of candidates [10], which is determined based on the order of the words' confidence values in initial annotation results. To achieve much better refinement performance, we proposed the multi-modal mutual information to measure the relation of label-to-label, image-to-label, and image-to-image. We fully utilize available all modal information in our proposed method. Compared with the existing annotation refinement mothod, our method can accurately describe the relationship between non-annotated image and candidate annotation words and images. As a consequence, our proposed method can get a better performance, which largely depends on our utilizing all modal information and our proposed weighted mutual information method.

II. Related Works

A. Discriminative models

Discriminative model-based AIA models view image annotation as a multi-label classification problems. A separate classifier is trained for each label using the visual features of the image and the trained classifier predicts particular labels for a test image. The image annotation method based on discriminative models are presented in [11-15]. Most of the discriminative models are based on support vector machine (SVM) or its variants [11,12,16,17]. The multi-class SVM is used in [11,15] to classify the images in one of the predefined classes. Some different kernels are used in [16] to find a certain type of visual properties of the image and to approximate the underlying visual similarity relationships between images more precisely. SVM is used as a classifier in [13,18] and the discriminative models are applied extensively for the medical image annotation. The SML model [19] is one of the models to treat AIA as a multi-classification problem and learns class-specific distributions for each label. The SVM-DMBRM model [20] makes some improvements in classification based on previous studies and presents a hybrid model to take full advantages of the merits of both generative and discirminative models for AIA. While the above methods try to solve the issue of AIA, the multi-label classification approaches cannot extend to a large number of categories since a binary classifier has to be built for each category.

B. Generative modes

The generative model aims at learning a joint distribution over visual and contextual features so that the learned model can predict the conditional probability of labels given the image features[21]. The model captures dependency between visual features and associated labels accurately. The generative models are usually based on topic model, relevance model and mixture model. Typical topic models include the latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA), latent Dirichlet allocation (LDA) and PLSA-WORDS. The relevance models include the translation model (TM) [1], across media relevance model (CMRM) [2], continuous space relevance model (CRM) [3], and multiple Bernoulli relevance models (MBRM) [4]. The representative mixture models are based on finite mixture model (FMM) [22], expectation maximization (EM) [23], and Gaussian mixture model (GMM) [24]. The generative models have made remarkable contribution to the development on AIA. However, they cannot guarantee optimization of the label prediction and cause a high computing demand because of the complexity of the algorithm.

C. Nearest Neighbor based models

In recent decades, Nearest Neighbor based models are widely used in AIA because of their simplicity and effectiveness. The nearest neighbor based models primarily focus on selecting the similar neighbors and then propagating the labels to the test image [47][48]. The similar neighbors can be defined by the image-to-image similarity (visual similarity) or image-to-label similarity or both. The Joint Equal Contribution (JEC) model [25] is one of the most classical nearest neighbor modes. It creates a family of very simple and intuitive baseline methods for AIA. The JEC model utilizes global low-level image features and a simple combination of basic distance measures to find nearest neighbors of a given image. Keywords are then assigned using a greedy label transfer mechanism, which selects keywords from the nearest neighbors based on co-occurrence and frequency factors. Others typical image annotation works using nearest neighbor based models include TagProp [7], 2-pass K-nearest neighbor (2PKNN) [6].

TagProp model is a tag propagation model based on the weighted nearest neighbor model where the weights of the neighbor are assigned based on its ranking or distance. TagProp model transfers labels by taking a weighted combination of the label presence and absence of neighbors. Moreover, it introduces word-specific discriminant models, which boost the probability for rare tags and decrease the probabilities for frequent ones concurrently to overcome the class-imbalance problem.

2PKNN model represents a classical solution to solve problems related to class-imbalance and weak-labeling [6]. It identifies all related semantic neighbors for each label by selecting k similar images in the vocabulary. The 2PKNN uses the two types of similarity in two passes. In the first pass, image-to-label similarity is used, and image-to-image similarity is used in the second pass. Since the success of 2PKNN in solving label -imbalance problem, it is still one of the most influential AIA

approaches.

D. Deep learning based methods

The most recent decade, deep learning based methods are presented to solve AIA task and have shown great performance in many computer vision tasks by extracting effective feature vectors from images[26-30]. The approaches of AIA based on deep learning can be classified into two categories. First is the end-to-end category, in which the deep learning networks perform AIA by the means of multi-label multi-class classification. Most approaches in this case focus on modifying the output layer or activation function based on fundamental deep learning architecture with a large image training dataset. Second is based on feature extracting category, in which the function of deep learning models is just to extract feature vector from image. Most deep learning networks used in AIA are based on convolution neural network (CNN) [29-32], and the commonly used feature extraction deep learning networks are AlexNet, VGGNet, and ResNet[30-33].

Jia, the creator of the Caffe, proposes the first solution based on CNN named CNN+WARP model [29]. The loss function of the model is defined as a multi-label variant of the WARP with the top-k annotation accuracy optimized by a stochastic sampling approach, which promotes image annotation performances. The CCA-KNN model [30] is based on the Canonical Correlation Analysis (CCA) framework that helps in modeling both visual features and textual features of the data. It was shown that CNN features were advantageous over 15 handcrafted features in the existing models, including JEC, 2PKNN, and SVM-DMBRM. The CNN-RNN framework [35] utilizes recurrent neural networks (RNN) to capture high-order label relationships at a moderate level of computational complexity. In this framework, the CNN and RNN are jointly utilized to derive image representation and the correlation between the adjacent labels, based on which the final outputs, such as label probability, are computed. The CNN-RNN architecture are also adopted to social image understanding. The D^2IA approach is different from the aforementioned annotation methods based on CNN model, which is based on generative adversarial network (GAN) model. The D^2IA creates semantically relevant, yet distinct and diverse labels [36].

Deep learning models are extensively being used for various computer vision tasks and shown a breakthrough performance, which mainly contributes to end-to-end feature extraction through convolution neural networks, but the application of deep learning for AIA is still in its early stage [1]. The deep learning based AIA is a quite new but promising direction for AIA [2].

E. Image Annotation Refinement

Jin is the first researcher on annotation refinement in ACM MM2005. He proposed WordNet-based Method (WMN) and used semantic similarity between words in semantic network as a measure of vocabulary relevance. The relevance between candidate tagging words and images depends on the semantic similarity between the words and other candidate words. By eliminating the less relevant candidate tagging words and retaining only the more relevant words in the initial results, the purpose of improving the initial tagging results is achieved. The correlation measure between candidate tagging words and the image to be labeled is as Equation (1).

$$p(w_i \mid I_q) = \alpha \sum_{j=1}^{N} p(w_i \mid w_j) \tag{1}$$

Where α is normalization constant, and $p(w_i \mid w_j)$ is the semantic similarity between w_i and w_j in semantic network.

WordNet semantic network only provides the qualitative evaluation of the correlation between different concepts without giving the quantitative measurement method. In order to evaluate the semantic similarity between words quantitatively, some methods, such as Lin, LCH and JNC, were proposed to measure the semantic correlation of vocabulary based on semantic distance or vocabulary sharing information. However, these methods can not achieve better performance [9], and the non scalability of vocabulary also limits the application of such improved methods based on dictionary annotation. For image annotation refinement, although the effect of WNM is not satisfactory, it inspires the direction for the researches of related fields.

Lin proposed a novel approach IARM based on recommendation model for automatic image annotation [39]. They first select some related images with tags from training dataset according to their visual similarity. Then, they estimated the initial ratings for tags of the training images based on tag ranking method and constructed a rating matrix. They also constructed a trust matrix based on visual similarity with a k-NN strategy. The recommendation model was built on the two matrices to rank candidate tags for the target image. Their experimental results indicated their effectiveness.

A content-based image annotation refinement (CIAR) algorithm is proposed to re-rank the candidate annotations [40]. It leverages both corpus information and the content feature of a query image. Experimental results on a typical Corel5k dataset show not only the validity of the refinement, but also the superiority of the proposed algorithm over existing methods. An algorithm using Random Walk with Restarts (RWR) is proposed to leverage both the corpus information and the original confidence information of the annotations. Experimental results on both non-Web images of Corel5k dataset and

Web images of photo forum sites demonstrate the effectiveness of the proposed method.

III. Measurement of Vocabulary Relevance Based on Weighted Mutual Information

A. The Mutual Information

The measurement of candidate tagging vocabulary relevance is the core work of image annotation refinement. In probability theory, the mutual information of random variables X and Y measures the degree of their interdependence[41], which can be defined as Equation (2).

$$I(X;Y)=\sum_{y\in Y}\sum_{x\in X}Pr(x,y)\log\left(\frac{Pr(x,y)}{Pr_X(x)Pr_Y(y)}\right) \quad (2)$$

Where $Pr_{(x,y)}$ is the joint probability distribution of X and Y, and $Pr_x(x)$ and $\mathrm{Pr_y(y)}$ are the marginal distribution of X and Y. The sketch map of mutual information can be represented as Figure 1, which can be represented as Equation (3) also.

$$H(X;Y)=H(X)-H(X\mid Y) \quad (3)$$

Where $H(X)$ is the entropy of X and $H(X\mid Y)$ is conditional entropy.

We represent the mutual information of two words according to Equation (2) as Equation (4)[10].

$$I(w_i;w_j)=Pr(w_i,w_j)\log\left(\frac{Pr(w_i,w_j)}{Pr(w_i)Pr(w_j)}\right) \quad (4)$$

Where w_i and w_j are two different words, $Pr(w_i)$ represents the probability of w_i, which can be calculated as Equation (5).

$$Pr(w_i)=\frac{count_{w_i\in J}(J)}{N} \quad (5)$$

Where $count_{w_i\in J}(J)$ is the number of event J containing w_i, and N is the total number of samples.

B.the Mutual Information base on the embedding feature space

In probability theory, the relevance of arbitrary two phenomenons, for example w_i and w_j, can be calculated according to Equation (4) which is the mutual information of w_i and w_j in the events occurred. The measurement of relevance of w_i and w_j can be used to predict the relation of w_i and w_j in subsequent events, however, it ignores the hypothesis of mutually independent and identically distributed.

In the view of images, the relevance of two words, w_i and w_j, can be measured through the mutual information measurement of the two words, and then the model is generated which is used to predict the relationship of the two words in a test image. However, the hypothesis of relevance of two words neglects the co-occurrence between training images and testing images. Whether the probability of words existing in the image or the relevance of two words in the image depends on the visual content of the image fundamentally . As in Equation (4), $Pr(w_i)$ represents the probability of containing w_i in glossary and $Pr(w_i,w_j)$ represents the probability of containing w_i and w_j in a image simultaneously. But, the method simply counts the frequency of words in the training dataset and applies the probability relationship to the test image, which ignores the relationship between the training image and the test image completely.

The existing image annotation refinement methods based on the coexistence relationship of words often pay close attention to only one-modal information, such as text vocabulary, but ignore other modal information in image dataset, such as visual information. There are different similarities between the training image and the test image containing the same word. The training image should have a greater relevance with the visually similar image to be labeled, which has the greater impact on the labeling results of the test image. Therefore, we propose the mutual information based on the embedding feature space, which can describe the relevance of any two words w_x and w_y in the test image I_i more accurately by comprehensively considering the label-to-label relationship and label-to-image relationship.

To express the multi-modal mutual information based on the embedding feature space, we calculate the visual similarity of J_j and I_i as Equation (6). Where I_i is the test image and J_j is a neighbour image visually similar with I_i in the training dataset.

$$sim_{multi-modal}(J_j,I_i)=\exp(-\alpha\times Dist(J_j^{CCA},I_i^{CCA})) \quad (6)$$

Where J_j^{CCA} and I_i^{CCA} are the new embedding features concated by new visual and new label features trained by canonical correlation analysis (CCA) [42]. To promote semantic level of image feature vector, we introduce CCA model into the calculation of the similarity of two images. The CCA model can map visual features and label features to a common meaningful semantic space.

The expansion of the weighted probability of w_x and the weighted joint probability of w_x and w_y are shown as Equation (7) and Equation (8) respectively.

$$Pr_{wt}(w_x)=\frac{\sum_{J_j\in \boldsymbol{DS}_x} sim_{multi-modal}(J_j^{CCA},I_i^{CCA})}{N} \quad (7)$$

$$Pr_{wt}(w_x, w_y) = \frac{\sum_{J_j \in DS_{xy}} sim_{multi-modal}(J_j^{CCA}, I_i^{CCA})}{N} \quad (8)$$

Where $Pr_{wt}(w_x)$ is the weighted probability of the word w_x and $Pr_{wt}(w_x, w_y)$ is the weighted joint probability of w_x and w_y. $Pr_{wt}(w_x)$ and $Pr_{wt}(w_x, w_y)$ are not the probability relation of text vocabularies through simple statistical methods, but they integrate the visual similarity between the label-to-image and the test image defined as Equation (6). $Pr_{wt}(w_x)$ is the probability of w_x appearing in the image, and $Pr_{wt}(w_x, w_y)$ is the probability of w_x and w_y appearing in the image together. Where N is the total number of images of the training dataset same as Equation (5), DS_x is the image subset containing the word w_x, and DS_{xy} is the image subset containing both the word w_x and the word w_y. Furthermore, the similarity between the image which the word belongs to and the image I_i to be labeled as the weight can also reflect the influence of the probability relationship about the words on I_i.

Finally, we can express the mutual information based on the embedding feature space as Equation (9).

$$I_{wt}(w_x; w_y; I_i) = Pr_{wt}(w_x, w_y) \log\left(\frac{Pr_{wt}(w_x, w_y)}{Pr_{wt}(w_x) Pr_{wt}(w_y)}\right) \quad (9)$$

Where $I_{wt}(w_x; w_y; I_i)$ is the mutual information of w_x,w_y based on the embedding feature space and I_i. The expression represents the image-to-image relation, image-to-label relation and label-to-label relation.

By calculating the mutual information based on visual and label features, we integrate the multi-modal feature information (image and label text) into the calculation of mutual information. The mutual information of our method could obtain the relationship between the information embedded in the multiple features space, which can better reflect the relationship between image and label to select more suitable annotations by using the features of neighbor image and related label.

C.The image annotation refinement in the embedding feature space base on the multi-modal mutual information

The multi-modal mutual information in the embedding feature space can measure the relevance between candidate words more accurately, and can be used in image annotation refinement. However, the relevance between the final candidate words and the test image should also be influenced by the confidence of candidate words in the initial annotation results. We propose the image Annotation Refinement in the Embedding feature space base on Mutual Information (AREMI) integrating image-to-label relationship (the initial confidence of candidate tagging words of the test image), label-to-label relationship (correlation between candidate tagging words), and image-to-image relationship (visual similarity between the test image and training image where the word belongs to) based on the measurement method of weighted mutual information. The process of AREMI algorithm is as follows. Firstly, we calculate the relevance between the test image I_t and the top M words with the highest confidence in the initial annotation results. Then, we select K candidate annotation words with the highest relevance as the final annotation results through annotation refining. The relevance between w_x and I_t is defined as Equation (10).

$$RL(I_t, w_x) = \sum_{i=1, i \neq x}^{M} I_{wt}(w_i; w_x; I_t) \delta_x \delta_i \quad (10)$$

Where M is number of selected words. δ_x and δ_i both are the weights which represent the confidence information of words w_x and w_i of the image I_t in initial annotation results which can be set as the value of the confidence or the confidence function. For example, in probability model, δ_x and δ_i are set as Equation (11).

$$\delta_x = Pr(w_x \mid I_t), \quad \delta_i = Pr(w_i \mid I_t) \quad (11)$$

The image annotation refinement algorithm in embedding feature space based on the mutual information is described as follows.

1) The confidence of each word in the test image is generated according to the initial annotation model. For example, the confidence of the word w_x in image I_t is denoted as $Pr(w_x|I_t)$;

2) In a test image, the M words with the highest confidence are selected as candidate tagging words.

3) According to Equation (9), the multi-modal mutual information based on embedding features between any two candidate tagging words in the test image is calculated.

4) According to Equation (10), the relevance between each candidate tagging word and the test image is calculated;

5) The k most relevant words in the test image are selected as the final annotation results.

As mentioned above, we apply CCA method to optimize the visual features and label features of the image respectively. We calculate the weighted mutual information between the features and select k labels with the highest mutual information value from N candidate

labels (more than 15) as the final annotation results.

Comparing with the existing improved model of image annotation, our AREMI algorithm has the following two main distinguishing characteristics:

1) We take the visual similarity of images as the weight of word frequency;

2) By fully mining the information contained in the image dataset, the original simple modal problem of the words relationship between text modes is extended to multiple modal problems of label-to-label, label-to-image, and image-to-image relationship.

IV. Experiment

A. Datasets

We conduct experiment on the Corel5k, ESP Game and IAPR TC-12 benchmark datasets, which have become the standard datasets in the field of image annotation.

The Corel5k dataset includes 5000 images (4500 as training images and 500 as test images). Each image in the dataset has one to five manually labeled words, and the total number of vocabularies in the dataset is 260. Each image in the dataset is either 192 × 128 or 128 × 192 pixels.

ESP Game dataset was published by von Ahn and Dabbish in 2004. The dataset consists of 18689 training images and 2081 test images. Each image is manually annotated with up to 15 labels, with 4.7 labels on average from a dictionary of 268 labels. The dataset images are annotated by game player using an online game. The two mutually unknown players are required to predict the same keyword(s) to score points for a randomly given image, which makes this dataset quite challenging and diverse.

IAPR TC-12 dataset was introduced by Grubinger for cross-lingual information retrieval in 2007. Each image is initially associated with a long description. The English nouns extracted from the descriptions by Makadia [4], [12] are treated as annotations. The dataset consists of 17665 training images and 1962 test images. Each images is 480×360 or 360×480 pixels. Each image is manually annotated up to 23 labels, with 5.7 labels on average from a dictionary of 291 labels. The dataset has been widely used for evaluating image annotation models.

B.Evaluation metrics

1) Per-label metrics

The per-label evaluation metrics have been widely used to evaluate image annotation approaches in the past two decades. The per-label evaluation metrics including precision, recall, and F1-measure, are considered as standard metrics in image annotation now. For each approach, we take the top five words as the final annotation.

For each label, per-label precision is defined as the number of images correctly predicted over the total number of images predicted with the label, and per-label recall is defined as the number of images correctly predicted over the total number of images having the label in its ground-truth. After averaging over all the labels in the vocabulary to get average per-label precision and average per-label recall as shown as Equation (12), respectively. Further, the per-label F1-measure can be computed with the two average metrics. F1-measure is the harmonic mean of precision and recall shown as Equation (12).

$$Precision_L = \frac{N_{correct}}{N_{predicted}} \tag{12}$$

$$Recall_L = \frac{N_{correct}}{N_{ground-truth}} \tag{13}$$

$$F1_L = \frac{2 \times Precison_L \times Recall_L}{Precison_L + Recall_L} \tag{14}$$

Where $N_{correct}$ is the number of images correctly annotated with a label w. $N_{predicted}$ is the number of images predicted with the same label w. $N_{ground-truth}$ is the number of images manually annotated with w. F1-measure combines $Precision_L$ and $Recall_L$, which indicates the integrated result. F1-measure is used for comprehensive performance evaluation by combing precision and recall.

2) Per-image metrics

In addition to per-label metrics, more and more researchers adopt per-image metrics to evaluate annotation performance[43-46] including precision, recall, and F1-measure. Recently, some researchers have pointed out that the per-label metrics are biased toward infrequent labels because making them correct could have a very significant impact on final accuracy [29][37]. Therefore, they propose per-image metrics to accurately evaluate annotation performance. The values of per-image metrics are averaged over all the images in the test dataset to get average per-image precision, average per-image recall, respectively. The definitions of per-image metrics are as follows.

$$Precision_I = \frac{N_{correct}}{N_{predicted}} \tag{15}$$

$$Recall_I = \frac{N_{correct}}{N_{ground\text{-}truth}} \quad (16)$$

$$F1_I = \frac{2 \times Precison_I \times Recall_I}{Precison_I + Recall_I} \quad (17)$$

Where $N_{correct}$ is the number of labels that are contained in the image and are correctly predicted the label by the annotation model. $N_{predicted}$ is the total number of labels that are predicted by the model. $N_{ground\text{-}truth}$ is the number of labels that is contained in the image. F1-measure combines $Precision_I$ and $Recall_I$, which indicates the integrated result.

3) Other metrics

We also consider other metrics as the evaluation of image annotation performance, including the $N+$ metric, the mean average precision (MAP).

The $N+$ metric counts how many label in the vocabulary are correctly predicted for at least one on test images.

The MAP is a widely used metric in the field of image retrieval [31,7,38]. It consists of per-label MAP (MAP_L) and per-image MAP (MAP_I), which take into consideration all labels for every image, and evaluate the full ranking. MAP_L measures image-ranking quality corresponding to labels, but MAP_I measures label-ranking quality corresponding to images. MAP is a traditional metric that measures the full ranking of images instead of only the top labels for each image [31]. Thus, MAP_L is less noisy and preferable to other per-label metrics. To evaluate the image annotation performance more comprehensively, we adopt MAP_L and MAP_I as supplementary evaluation metrics to evaluate image annotation approaches.

Furthermore, we apply the hybrid F1-measure (called H-F1) combining $F1_L$ and $F1_I$ with the harmonic mean [43].

C. Implementation details

We use overlapping rasterization method to divide an image into same size grids in feature extraction. Each grid is responsible to extract three statistics including color statistic (mean, variance), Gabor texture and SIFT feature. We construct 500, 500 and 1000 dimensional visual dictionaries to represent the image. Finally, each image is represented as a 2000 dimensional bag of words (BOW) histogram vector. Our method adopt the features obtained by CCA concating of BOW features and labels, and the labels of the test dataset are pre-annotated through content-based image retrieval (CBIR) [34] method.

D. Results and Comparison

The experiments are mainly performed using Matlab on a computer of Intel Corei7-9750H CPU with 2.6GHz and 16 GB RAM, running Windows 10 OS. For a fair comparison with the state-of-art approach in AIA, we carry our experiments on the same three benchmark datasets mentioned above (Corel5k, ESP Game and IAPR TC-12) with five labels predicted for each test image. We compare our method with two famous nearest-neighbour methods using per-label metrics, per-image metrics, and MAP, which are JEC and TagProp.

The experiment results on Corel5k, ESP Game, and IAPR TC-12 are summarized in Tables I-III, respectively. The purpose of our proposed annotation refinement is to optimize the annotation results based on other annotation method. Therefore, we put into effect our annotation refinement method based on JEC and TagProp, respectively.

TABLE I. PERFORMANCE EVALUATION ON COREL5K DATASET

Model	P_L	R_L	$F1_L$	N^+	P_I	R_I	$F1_I$	H-F1	MAP_L	MAP_I	Time cost（s）
JEC	28.64	27.88	28.25	124	31.22	44.19	36.59	31.98	28.84	39.66	6.14
AREMI(Based on JEC)	**37.83**	**37.10**	**37.46**	**157**	**36.07**	**50.72**	**42.16**	**39.67**	**34.51**	**46.47**	**3.39**
TagProp	17.06	22.47	19.39	108	32.55	45.99	38.12	25.70	22.78	42.38	32.36+0.44
AREMI(Based on TagProp)	**31.84**	**27.74**	**29.65**	**116**	**41.20**	**58.01**	**48.18**	**36.71**	**29.16**	**55.22**	**3.24**

TABLE II. PERFORMANCE EVALUATION ON ESP GAME DATASET

Model	P_L	R_L	$F1_L$	N^+	P_I	R_I	$F1_I$	H-F1	MAP_L	MAP_I	Time cost

											(s)
JEC	23.25	16.37	19.21	220	26.11	29.13	27.54	22.63	11.61	28.96	28.46
AREMI(Based on JEC)	**33.65**	**28.01**	**30.57**	**265**	**31.24**	**34.27**	**32.68**	**31.59**	**19.03**	**33.25**	36.14
TagProp	28.12	15.30	19.82	231	21.21	23.43	22.26	20.97	12.47	24.71	108.27
AREMI(Based on TagProp)	**42.31**	**18.92**	**26.15**	**227**	**34.06**	**35.34**	**34.69**	**29.82**	**15.10**	**27.41**	36.22

TABLE III. PERFORMANCE EVALUATION ON IAPR TC-12 DATASET

Model	P_L	R_L	$F1_L$	N^+	P_I	R_I	$F1_I$	H-F1	MAP_L	MAP_I	Time cost (s)
JEC	28.37	18.65	22.51	211	37.16	35.01	36.06	27.71	20.63	39.69	28.15
AREMI(Based on JEC)	**43.49**	**28.00**	**34.07**	**255**	**42.92**	**40.41**	**41.63**	**36.97**	**27.43**	**43.52**	35.05
TagProp	32.96	17.85	23.16	221	31.71	29.74	30.69	26.40	21.01	34.66	107.40
AREMI(Based on TagProp)	**46.07**	**20.74**	**28.61**	**215**	**43.40**	**40.90**	**42.11**	**34.07**	**23.89**	**44.08**	35.16

The first two lines of Table I-III are the initial evaluation values of JEC and TagProp. The third line in the tables is our evaluation values based on JEC, and the forth line in the tables is our evaluation values based on TagProp. From Tables I-III, we can see that our annotation refinement method significantly improves the annotation result of general annotation methods, such as JEC, TagProp. In addition, the time cost of our method is similar to JEC, and significantly smaller than TagProp.

Our experimental results show that each evaluation values of AREMI is better than the initial annotation methods. In essence, image annotation refinement method is to improve the correlation between the final annotation vocabulary and image visual information or semantic concepts by optimizing the existing candidate annotation results, so as to improve the quality of image annotation. However, most annotation refinement processes are often independent of the test image (the image to be labeled) currently, and only consider the relationship between text words. Our proposed AREMI integrates a multi-modal information (calculated by the formula in section III), such as label-to-image relationship, label-to-label relationship and image-to-image relationship, and takes the visual information of the test image as an important basis for annotation refinement in each step. The multi-modal mutual information reflects the multi-modal information between images and labels, so the annoation refinement method based on multi-modal mutual information can achieve good results under project approval status.

The experimental results of our method are based on the original results of other two approaches, which are further optimized according to the calculation of image-to-image, label-to-image and label-label mutual information in the embedding space. Therefore, the original annotation results are improved to a certain extent.

V. CONCLUSIONS AND FUTURE WORK

We propose an Annotation Refinement method in Embedding feature space based on multi-modal Mutual Information (AREMI). The experimental results show that the multi-modal mutual information can measure the correlation between words more accurately. It can be applied not only to AREMI method, but also to other complex annotation refinement models with better annotation refinement performance. The main reason that our method can achieve better performance than some famous models including JEC and TagProp can be summarized into three aspects: (1)Fully multi-modal information integrating label-to-label relationship, label-to-image relationship and image-to-image

relationship. (2)We proposed weighted mutual information method to measure label-to-label relationship, so it can accurately measure the relationship between the words to ensure the effect of annotation refinement. (3)The weighted proposed mutual information is much better than the visual similarities. The multi-modal mutual information base on the embedding feature space and AREMI method proposed in this paper can not only be used for general image annotation and annotation refinement, but also can be used for object detection and recognition in specific fields, or for object recognition or scene recognition in machine vision.

Reference

[1] P. K. Bhagat and P. Choudhary, ‘‘Image annotation: Then and now’’ Image Vis. Comput., vol. 80, pp. 1–23, Dec. 2018.
[2] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, ‘‘A survey and analysis on automatic image annotation,’’ Pattern Recognit., vol. 79, pp. 242–259, Jul. 2018.
[3] V. Lavrenko, R. Manmatha, J. Jeon, “A model for learning the semantics of pictures,” In: Advances in neural information processing systems, pp.553–560, 2003
[4] S. Feng, R. Manmatha, and V. Lavrenko, “Multiple bernoulli relevance models for image and video annotation,” in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, vol. 2. IEEE, 2004, pp.II–II.
[5] A. Makadia, V. Pavlovic, and S. Kumar, “Baselines for image annotation,”International Journal of Computer Vision, vol. 90, no. 1, pp.88–105, 2010.
[6] Y. Verma and C. Jawahar, “Image annotation by propagating labels from semantic neighbourhoods,” International Journal of Computer Vision, vol. 121, no. 1, pp.126–148, 2017.
[7] A. Dutta, Y. Verma, and C. V. Jawahar, “Automatic image annotation:The quirks and what works,” Multimedia Tools Appl., vol. 77, no. 24,pp. 31991–32011, Dec. 2018.
[8] B. Wu, W. Chen, P. Sun, W. Liu, B. Ghanem, and S. Lyu, “Tagging like humans: Diverse and distinct image annotation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp.7967–7975.
[9] Y. Jin, L. Khan, L. Wang and M. Awad. Image annotations by combining multiple evidence & wordnet[C]. In Proceedings of the 13th Annual ACM international Conference on Multimedia, 2005:706-715.
[10] Song H, Research on the Automatic Image Annotation and Annotation Refinement Algorithms[D], Changchun: Univ. Of Jilin, 2012.
[11] C. Cusano, G. Ciocca, R. Schettini, Image annotation using SVM, 5304–5304-9. Proc. SPIE 5304 (2003) https://doi.org/10.1117/12.526746.
[12] K.S. Goh, E.Y. Chang, B. Li, Using one-class and two-class SVMs for multiclass image annotation, IEEE Trans. Knowl. Data Eng. 17 (10) (2005) 1333–1346.
[13] A.Mueen,R.Zainuddin,M.S.Baba,Automaticmultilevelmedicalimageannotation and retrieval, J. Digit. Imaging 21 (2007) 290–295.
[14] D. Grangier, S. Bengio, A discriminative kernel-based approach to rank images from text queries, IEEE Trans. PatternAnal.Mach. Intell. 30 (8) (2008) 1371–1384. https://doi.org/10.1109/TPAMI.2007.70791.
[15] S. Lindstaedt, R. Mörzinger, R. Sorschag, V. Pammer, G. Thallinger, Automatic image annotation using visual content and folksonomies,Multimedia ToolsAppl. 42 (1) (2009) 97–113. https://doi.org/10.1007/s11042-008-0247-7.
[16] J. Fan, Y. Gao, H. Luo, Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation, IEEE Trans. Image Process. 17 (3) (2008) 407–426.
[17] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, T. Huang, Large-scale image classification: fast feature extraction and SVM training, CVPR 2011, 2011. pp. 1689–1696. https://doi.org/10.1109/CVPR.2011.5995477.
[18] T. Tommasi, F. Orabona, B. Caputo, Discriminative cue integration for medical image annotation, image CLEF 2007. Pattern Recogn. Lett. 29 (15) (2008) 1996–2002. https://doi.org/10.1016/j.patrec.2008.03.009.
[19] G. Carneiro, A.B. Chan, P.J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 394–410.
[20] V.N. Murthy, E.F. Can, R. Manmatha, A hybrid model for automatic image annotation, in: International Conference on Multimedia Retrieval, 2014, pp. 369–376.
[21] Y. Han, F. Wu, Q. Tian, Y. Zhuang, Image annotation by input-output structural grouping sparsity, IEEE Trans. Image Process. 21 (6) (2012) 3066–3079.
[22] J. Fan, Y. Gao, H. Luo, G. Xu, Automatic image annotation by using conceptsensitive salient objects for image content representation, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’04, ACM, New York, NY, USA, 2004, pp. 361–368. https://doi.org/10.1145/1008992.1009055.
[23] R. Zhang, Z. Zhang, M. Li, W.-Y. Ma, H.-J. Zhang, A probabilistic semantic model for image annotation and multimodal image retrieval, Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1, vol. 1, 2005. pp. 846–851. Vol. 1.
[24] C.Wang,S.Yan,L.Zhang,H.J.Zhang,Multi-labelspars ecoding forautomatic image annotation, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009. pp. 1643–1650.
[25] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation, in: European Conference on Computer Vision, 2008, pp. 316–329.
[26] M. Koskela and J. Laaksonen, ‘‘Convolutional network features for scene recognition,’’ in Proc. 22nd ACM Int. Conf. Multimedia, K. A. Hua, Y. Rui, R. Steinmetz, A. Hanjalic, A. Natsev, and W. Zhu, Eds. Orlando, FL, USA: ACM, 2014, pp. 1169–1172.
[27] K. Simonyan and A. Zisserman, ‘‘Very deep

convolutional networks for large-scale image recognition,'' in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), Y. Bengio and Y. LeCun, Eds.San Diego, CA, USA: arXiv,2015. [Online]. Available: http://arxiv.org/abs/1409.1556
[28] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, ''DeCAF: A deep convolutional activation feature for generic visual recognition,'' in Proc. Int. Conf. Mach. Learn. (ICML), vol. 32, Jun. 2014, pp. 647–655.
[29] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, ''Deep convolutional ranking for multilabel image annotation,'' in 2nd Int. Conf. Learn. Represent. (ICLR), Y. Bengio and Y. LeCun, Eds. Banff, AB, Canada: arXiv, 2014. [Online]. Available: http://arxiv.org/abs/1312.4894
[30] V. N. Murthy, S. Maji, and R. Manmatha, ''Automatic image annotation using deep learning representations,'' in Proc. 5th ACM Int. Conf. Multimedia Retr., A. G. Hauptmann, C. Ngo, X. Xue, Y. Jiang, C. Snoek, and N. Vasconcelos, Eds. Shanghai, China: ACM, 2015, pp. 603–606.
[31] J. Johnson, L. Ballan, and F. Li, ''Love thy neighbors: Image annotation by exploiting image metadata,'' in Proc. IEEE Int. Conf. Comput. Vis., Santiago, Chile, Dec. 2015, pp. 4624–4632.
[32] M. Zang, D. Wen, K. Wang, T. Liu, and W. Song, ''A novel topic feature for image scene classification,'' Neurocomputing, vol. 148, pp. 467–476, Jan. 2015.
[33] T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo, ''Automatic image annotation via label transfer in the semantic space,'' Pattern Recognit., vol. 71, pp. 144–157, Nov. 2017.
[34] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences, and trends of the new age, ACM Comput. Surv. 40 (2008) 5:1–5:60.
[35] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, Cnn-rnn: a unified framework for multi-label image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2285–2294.
[36] B. Wu, W. Chen, P. Sun, W. Liu, B. Ghanem, and S. Lyu, ''Tagging like humans: Diverse and distinct image annotation,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7967–7975. http://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Tagging_Like_Humans_CVPR_2018_ paper.html
[37] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multi-label image annotation," arXiv preprint arXiv:1312.4894, 2013.
[38] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval," ACM Computing Surveys (CSUR), vol. 49, no. 1, pp. 1–39, 2016.
[39] Z. Lin, G. Ding, J. Wang, Image Annotation Based on Recommendation Model[C]// International Acm Sigir Conference on Research & Development in Information Retrieval. ACM, 2011.
[40] Wang C , Jing F , Lei Z , et al. Content-Based Image Annotation Refinement[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2007.
[41] Maes F, Collignon A, Vandermeulen D et al. Multi-modality image registration by maximization of mutual information. In: Proc. IEEE Workshop Mathematical Methods in Biomedical Image Analysis . San Francisco , CA , 1996, 14-22.
[42] Z. Li, J. Tang, and T. Mei, ''Deep collaborative embedding for social image understanding,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 9, pp. 2070–2083, Sep. 2019.
[43] Y. Niu, Z. Lu, J.-R. Wen, T. Xiang, and S.-F. Chang, "Multi-modal multiscale deep learning for large-scale image annotation," IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1720–1731, 2018.
[44] H. Song, P. Wang, J. Yun, W. Li, B. Xue, and G. Wu, "A weighted topic model learned from local semantic space for automatic image annotation," IEEE Access, vol. 8, pp. 76 411–76 422, 2020.
[45] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "Semantic regularisation for recurrent image annotation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2872–2880.
[46] D. Putthividhy, H. T. Attias, and S. S. Nagarajan, "Topic regression multimodal latent dirichlet allocation for image annotation," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010, pp. 3408–3415.
[47] H. Song, P. Wang, J. Yun, et al. "A Weighted Topic Model Learned from Local Semantic Space for Automatic Image Annotation," IEEE Access, 2020, 99,pp: 1-1.
[48] H. Song, J. Yun, H. Li, et al. "An Efficient and Effective Model Based on Mean Positive Examples for Social Image Annotation," IEEE Access, 2020, 8,pp: 210695-210708.

Wei Li received the B.S. and M.S. degrees in computer software and theory from Jilin University, in 2002 and 2005, respectively. She received the Ph.D. degree in the Transportation College, Jilin University, China. She is also a Lecture with the School of Computer Science and Engineering, Dalian Minzu University, China. Her current research interests include image understanding, computer vision, and machine learning.

Haiyu Song received the B.S.,M.S.,and Ph.D. degrees in computer software and theory from Jilin University, in 1996, 2003, and 2012, respectively. He is currently a associate professor with the School of Computer Science and Engineering, Dalian Minzu University, China. His current research interests include image understanding, computer vision, and machine learning.

Hongda Zhang is currently pursuing the M.S. degree with the School of Information and Communication Engineering, Dalian Minzu University, China. His current research interests include computer vision, image processing, and machine learning.

Houjie Li Received the B.S. and M.S. degrees in communication and information system from Jilin University, in 2001, 2004, and the Ph.D. degree in signal and information processing from Dalian University of Technology, in 2019. He is currently an associate professor with the School of Information and Communication Engineering, Dalian Minzu University, China. His current research interests include Image Processing, computer vision, and machine learning.

Pengjie Wang received the B.S. and M.S. degrees in computer software and theory from Jilin University, in 2001, 2004, and the Ph.D. degree in computer application from Zhejiang University, in 2011. He is currently a professor with the School of Computer Science and Engineering, Dalian Minzu University, China. His current research interests include computer vision, virtual reality, and machine learning.

Author Contributions: Please, indicate the role and the contribution of each author:

Wei Li was responsible for investigating relevant literature, planning the system, and completing the simulation and optimization work.

Haiyu Song was responsible for the construction of the experimental model of the fourth part.

Hongda Zhang was responsible to implement the forth part of the experiment.

Houjie Li were responsible for completing the experimental part.

Pengjie Wang was responsible for completing the experimental part.