Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 183 (2021) 616–623

Procedia Computer Science

www.elsevier.com/locate/procedia

10th International Conference of Information and Communication Technology (ICICT-2020)

Automatic image annotation based on an improved nearest neighbor technique with tag semantic extension model

Wei Wei[a], Qiong Wu[a], Deng Chen[a,b,*], Yanduo Zhang[a], Wei Liu[a], Gonghao Duan[a], Xu Luo[a]

[a]*Hubei Provincial Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan, China*
[b]*Lingyun technology group co. LTD, Wuhan, China*

Abstract

Nearest Neighbor method (KNN) is a typical method to solve the problem of automatic image annotation (AIA). However, traditional AIA methods based on KNN only consider the relationships among images and labels. In this paper, we propose an improved KNN image annotation method based on a tag semantic extension model (TSEM). Our approach uses the convolutional neural network (CNN) to extract image features and predicts image tags automatically via nearest features. Different from existing work, the proposed method considers correlations among images, correlations between images and labels and those among labels. Additionally, a label quantity prediction (LQP) model is proposed to predict the number of tags, which further improves the tag prediction accuracy. Comparison experiments were performed on three typical image datasets Corel5k, ESP game and laprtc12. Experimental results show that the average *F1* of our model is 0.427, which outperforms the state-of-the-art KNN image annotation methods.

Keywords: Automatic image annotation; convolutional neural network; tag semantic extension model; label quantity prediction;

1. Introduction

Automatic image annotation (AIA) is one of the fundamental problems in computer vision and machine learning. For a given image, the AIA methods aim to predict a set of textual labels that describe the semantics of the image

* Corresponding author. Tel.: +86-15392917835;
E-mail address: dchen@wit.edu.cn

10.1016/j.procs.2021.02.105

[1]. The principle of AIA is to propagate labels among images based on similarities between image visual features and semantic contents [2]. Compared with manual image annotation, AIA has higher efficiency and precision.

AIA techniques have been studied extensively. Typical methods include convolutional neural network (CNN) based AIA methods [3, 6], support vector machine (SVM) based AIA methods [7, 9] and k-nearest neighbor (KNN) based AIA methods [10, 16]. The KNN based AIA methods attract much attention from researchers for their prominent efficiency and precision. For example, Verma, Y. et al. [11] proposed a two-pass k-nearest neighbor (2PKNN) algorithm for image annotation. First, it obtains the neighborhood set of an image. Then, it uses the weighted similarity of the image to predict image labels. A distinguishing characteristic of this approach is that it can address the problem of sparse labels. Ma, Y., et al [16] (SEM) used the CNN to extract image features other than hand-crafted features, and then predicted image tags based on contributions of neighborhood images. The feature extraction method based on CNN avoids the incompleteness of manual features and is able to achieve more accurate label prediction results. However, existing AIA methods based on KNN only take the image-image correlation and image-label correlation into account. Less attention has been paid to correlations among labels. In addition, the existing works are unable to predict the exact label quantity of images, which limits the accuracy of AIA methods.

In this paper, we propose an improved AIA method based on KNN. We use CNN to extract image features, combined with tag semantic extension model and tag quantity prediction model, our method can effectively improve the accuracy of tag prediction.

The main contributions of our work are as follows:

- An improved KNN image annotation method based on CNN features is proposed. Compared with hand-crafted image features, CNN deep features may be more comprehensive and robust for image annotation.
- A tag semantic model is proposed to extend traditional image annotation technique based on KNN, which is beneficial for improving the accuracy of AIA by considering semantic correlations among textual labels.
- Experiments are conducted based on three typical datasets: corel5k, ESP game and laprtc12 and promising results are achieved.

2. Our Technique

The framework of our AIA method is shown in Fig. 1. First, we extract image features through the pre-trained CNN. Then the label prediction based on KNN is used to obtain the neighborhoods and the candidate label set of unlabeled images. After that, we extend the candidate tag set using WordNet based-tag correlation to obtain the semantic associated label set. Simultaneously, to further improve the accuracy of label prediction, we propose a label quantity prediction model based on image feature similarity to predict the quantity of labels. Finally, combined with the prediction of the number of tags, we select tags from the candidate label set and semantic associated label set as the final label prediction result for image annotation.

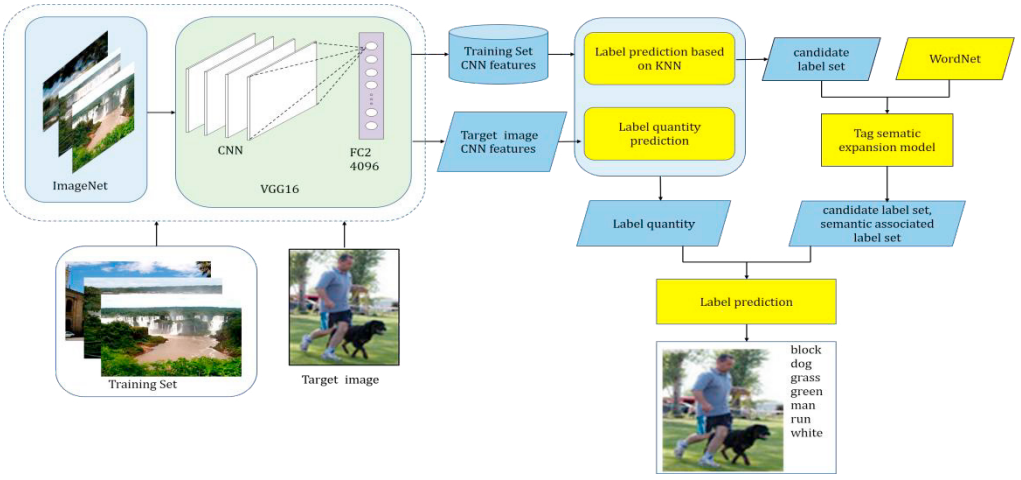

Fig. 1. The flowchart of the proposed framework.

2.1. Label prediction based on image neighborhood

In this paper, we use the pre-trained model VGG-16 [17] as the feature extraction network, and take the output of the second fully connected layer as image features. For the image dataset $I = \{I_i | i = 1,...,n\}$, where n is the number of image samples. $F = \{F_i | F_i \in R^l\}$ is the l-dimensional vector set for image set I. $W = \{W_i | w_i^1, w_i^2, ..., w_i^{k_i}\}$ is the label set of image set I, where k_i is the number of labels for image I_i. For image I', the AIA methods based on KNN calculate the neighborhood image set $N(I',k)$ of I' as follows:

$$\forall I^x \in N(I',k), I^y \notin N(I',k), d(F',F^x) \leq d(F',F^y) \quad (1)$$

where $F' \in R^l$, $F^x \in R^l$, $F^y \in R^l$ are the feature vectors of images I', I^x, I^y, respectively, $d(\cdot,\cdot)$ is the distance function of image feature vectors. According to the neighbourhood image set $N(I',k)$ and the label set W of the image set I, the candidate label set $cl(I')$ of image I' can be obtained as follows:

$$cl(I') = \bigcup_{I^x \in N(I',k)} W_x \quad (2)$$

where W_x is the label set of I^x. For each label $w \in cl(I')$, the AIA methods based on KNN calculate the correlation weight between w and I', and select top-k labels as the prediction result of I'.

The feature vector description and distance function are crucial for AIA methods based on KNN. In this paper, we use VGG-16 to learn a 4096-dimensional feature vector of an image and use Euclidean distance to measure the distance between image features. Given a pair of images feature vectors $F_i, F_j \in R^{4096}, i,j = 1,...,n$, the distance between the F_i and F_j is calculated as follows:

$$d(F_i, F_j) = \sqrt{\sum_{l=1}^{4096} \left(F_i^l - F_j^l\right)^2} \quad (3)$$

For $w \in cl(I')$, we use the method proposed by Ma [16] to calculate the correlation probability $p(w|I')$ between w and image I'. This method evaluates the contribution of each neighborhood image to the image I', and then determines which labels are more relevant to the image based on the contribution value. The calculation method is as follows:

$$p(w|I') = \sum_{I^z \in N(I',k)} smr(F',F^z) \cdot \delta(w, I^z) \quad (4)$$

$$smr(F',F^z) = \frac{1}{1+exp(\theta \cdot d(F',F^z))} \tag{5}$$

$$\delta(w, I^z) = \begin{cases} 0 & w \notin W_z \\ 1 & w \in W_z \end{cases} \tag{6}$$

where $smr(F',F^z)$ is used to calculate the similarity between the image feature vectors of I' and its neighborhood image I^z. $\delta(w,I^z) = \{0,1\}$ is the 0-1 function, W_z is the label set corresponding to image I^z, when the label set W_z has label w, the value of $\delta(w,I^z)$ is 1, otherwise the value of $\delta(w,I^z)$ is 0. θ is the image feature similarity factor.

2.2. Tag semantic extension model

The existing KNN-based AIA methods are mainly based on the visual characteristics of image for image annotation, while ignoring the semantic relationship between labels. For example, if an image has label "*clouds*", then "*sky*" may also appear, because the tags "*sky*" and "*clouds*" have strong semantic correlation. Based on the above analysis, we propose a tag semantic extension model. This model takes the above candidate tag set as input, calculates the semantic correlation between tags according to WordNet, and outputs an extended candidate tag set.

Suppose I' is the unlabeled image and $cl(I')$ is the candidate label set of I', for $w \in cl(I')$, the tag semantic extension model uses the JNC[18] method which proposed by Jiang et al. to obtain the semantic extension label set $S(w)$ of w from WordNet and add it to the extended label set $cle(I')$, namely:

$$cle(I') = \bigcup_{w \in cl(I')} S(w) \tag{7}$$

The JNC method uses the maximum similarity value between the classes of words to determine the correlation between tags. Then, for $w \in cl(I')$, the method of calculating the related tag of label w according to JNC is as follows:

$$S(w) = \{w_i | \varphi(sim(w_i,w)) \geq \partial, w_j \in cl(I')\} \tag{8}$$

$$sim(w_i,w) = \max_{c_i \in sen(w_i), c_j \in sen(w)} [sim(c_i,c_j)] \tag{9}$$

$$sim(c_i,c_j) = \max_{c \in Sup(c_i,c_j)} (-log\, p(c)) \tag{10}$$

where $w_i \in W$, ∂ is the relevance threshold to prevent the introduction of labels with too small relevance. $\varphi(\cdot)$ is sigmoid function. $sim(\cdot)$ is used to calculate the correlation between tags. $sen(w)$ represents a set of word meanings of label w, which is represented as the set of ancestor nodes of w in WordNet. $Sup(c_i,c_j)$ represents a group of concepts that contains both c_i and c_j, which is represented as a set of common ancestor nodes of c_i,c_j in WordNet. Where $p(c) = freq(c)/N$ and $freq(c)$ represents the total number of occurrences of semantic concept c in the label set W.

We get all the associated tags according to formula 8. For $w_i \in cle(I')$, we cannot calculate $p(w_i|F')$ directly because the w_i is unrelated to image I'. But w_i is the associated tag of w, so we use the correlation degree between w_i and w and the labeling probability $p(w|F')$ of w to approximate $p(w_i|F')$ in this paper, it is specifically expressed as:

$$p(w_i|I') = p(w|I') \cdot \varphi(sim(w_i,w)) \tag{11}$$

2.3. Label quantity prediction

For further improving the accuracy of image annotation, we propose a prediction model for the quantity of image tags. Given an image I', the label quantity prediction model calculates the feature similarity $smr(F',F^z)$ between the

image I' and I^z, where $I^z \in N(I',k)$, then predict the quantity of tags of image I' according to $smr(F',F^z)$. We sort the neighborhood images from largest to smallest according to $smr(F',F^z)$, and the top-m are used for label quantity prediction. The specific formula for label quantity prediction of image I' is (12):

$$a = \frac{\sum_{L=1}^{m} x_L \times smr(F', F^L)}{\sum_{L=1}^{m} smr(F', F^L)}, (\forall I^L \in N(I',k), I^y \notin N(I',k), smr(F',F^L) \geq smr(F',F^y)) \tag{12}$$

where x_L denotes the actual number of labels of image I^L. $F' \in R^l$, $F^L \in R^l$, $F^y \in R^l$ are the feature vectors of images I', I^L, I^y, respectively.

After we get all the probability $p(w|I')$ of labels in candidate label set $cl(I')$ and the probability $p(w_i|I')$ of labels in extended label set $cle(I')$ and the number a of labels for image I', we select the largest a probability from $p(w|I') \cup p(w_i|I')$ and their corresponding labels to group into a W', the W' is the final annotation result of image I'.

On the basis of the AIA methods based on KNN, we propose a tag semantic extension model based on WordNet and a label quantity prediction model to improve the accuracy of AIA. The improved algorithm of AIA based on tag semantic extension model is proposed as Algorithm 1.

Algorithm 1 Label prediction based on Tag Semantic Extension Model

Input: Image Set *I*, the feature set *F* of Image Set, label set *W*

Output: Prediction results *W'* of AIA

for an untagged image feature F' **do**
  for image feature F_i **in** F **do**
    $d(F', F_i) = \sqrt{\sum_{l=1}^{4096} \left(F'^l - F_i^l\right)^2}$
  end for
  sort $d(F', F_i)$ **in** descending order
  get the neighbourhood feature group $N(I',k)$ according to the top-k $d(F', F_i)$
  get the tags $cl(I')$ of each image in $N(I',k)$
  for w **in** $cl(I')$ **do**
    for F^z **in** $N(I',k)$ **do**
      if w belongs to the image whose feature is F^z then
        $\mathrm{smr}(F', F^z) = \frac{1}{1+\exp(\theta \cdot dis(F', F^z))}$
        $p(w|I') = p(w|I') + \mathrm{smr}(F', F^z)$
    end for
  end for
  sort $\mathrm{smr}(F', F^Z)$ in descending order
  $\mathrm{a} = \frac{\sum_{L=1}^{m} x_L \times smr(F', F^L)}{\sum_{L=1}^{m} smr(F', F^L)}$
  for w **in** $cl(I')$ **do**
    for w_i **in** W
      $w_i \leftarrow \varphi(sim(w_i, w)) \geq \partial$
      add w_i **to** $cle(I')$
      $p(w_i|I') = p(w|I') \cdot \varphi(sim(w_i, w))$
    end for
  end for
  $p(w'_i|I') = p(w|I') \cup p(w_i|I')$
  sort $p(w'_i|I')$ in ascending order
  select the top-a $p(w'_i|I')$ and finding the corresponding words w'_i
  add w'_i to W'
end for
Output Prediction results W'

3. Experiments and results

3.1. Datasets and settings

We use three most widely used image datasets Core5k, ESP game and Iaprtc12 [19] as the benchmark dataset. Among them, the Corel5k which contains 5000 images and the dictionary contains 260 words. The images are divided into 4500 training and 500 test examples. The ESP game has a total of 20770 pictures and 268 words. There are 18689 images in the training set and 2081 images in the test set. The Iaprtc12 has 19627 images and the images are divided into 17665 training examples and 1962 test examples. The dictionary contains 291 words. In this paper we evaluate our model with precision(P) and recall(R) and $F1$-score($F1$).

In this paper, θ is the image feature similarity factor and a better result can be obtained when θ set as 20 according to experiment in literature [16]. The parameter ∂ is the threshold of correlation degree. We set its value range is 0.4 to 0.8 when we test our model, and the model effect is best when it set as 0.6. For the parameter m, we set its value to 2, 3, 4, 5, 6, 7, 8 and test it on three datasets respectively and determine its optimal value by comparing $F1$ score. The result can be seen from the Figure 2 that the $F1$ score is highest when the value of m is 2 on Corel5k and ESP-game, and the $F1$ score is highest when the value of m is 3 on Iaprtc12. Integrating the influence of m on three different datasets, we take the average and set m to 2.

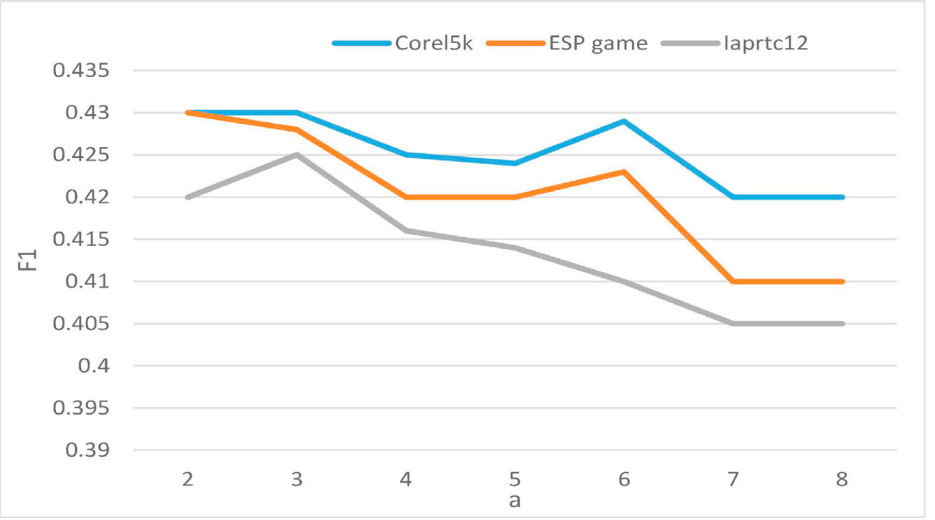

Fig. 2. $F1$ scores of our model (TSEM) on Corel5k, ESP game, Iaprtc12 datasets varying the parameter m.

3.2. Experiments results

We set up experiments on three datasets to verify the model proposed in this paper. In the Experiment, TSEM was combined with LQP, and the average of accuracy, recall and $F1$ score were used as evaluation indexes to compare with the existing image annotation methods JEC, Tag prop, SEM and Weight-KNN. It can be seen from Table 1 that on the three datasets, our model performs best on $F1$ score, which shows that our model is superior to the traditional AIA methods based on KNN. On Corel5k, the accuracy of our model is 21.6% higher than SEM which is the previous model with best accuracy, but the recall rate is decreased. On ESP game and Iaprtc12, our model has the best performance in $F1$ score, and the recall rate is increased by 7.1% and 7.7% respectively compared with SEM. Although the accuracy rate is not the highest, the difference is small. The TSEM mainly adds candidate tags to the model based on the correlation between tags, but the correlation between tags in different data sets is different. The improvement of our model by TSEM is also affected by this. At the same time, we calculate the average $F1$ score on the three datasets to further judge our model. The average $F1$ of our model is 0.427, which is

75.6%, 18.6%, 4.1% and 77.9% higher than JEC, Tag Prop, SEM, and Weight-KNN respectively. This further proves the effectiveness of our model.

Table 1. Experimental results of our model (TSEM) with competitive models on three datasets

Dataset	Corel5k			ESP game			Iaprtc12		
	P	R	F	P	R	F	P	R	F
MBRM	0.24	0.25	0.25	0.18	0.19	0.19	0.24	0.23	0.24
CRM	0.19	0.19	0.17	-	-	-	-	-	-
JEC	0.27	0.32	0.29	0.24	0.19	0.21	0.29	0.19	0.23
Tag Prop	0.33	0.42	0.37	0.39	0.27	0.32	**0.45**	0.34	0.39
SEM	0.37	**0.52**	**0.43**	0.38	0.42	0.40	0.41	0.39	0.40
FFSS	0.27	0.33	0.30	0.21	0.23	0.22	0.29	0.29	0.29
Weight-KNN	0.22	0.15	0.18	**0.46**	0.22	0.30	0.42	0.17	0.24
TSEM(Ours)	0.38	0.46	0.42	0.44	0.40	0.42	0.39	0.42	0.41
TSEM+LQP(Ours)	**0.45**	0.40	**0.43**	0.44	**0.43**	**0.43**	0.42	**0.42**	**0.41**

4. Conclusions

In this work, we give up the traditional method of hand-crafted features but use CNN to extract image features to improve the accuracy of AIA. Moreover, based on the experience of the traditional AIA models based on KNN, we propose a tag semantic extension model based on WordNet and a tag quantity prediction model to achieve accurate image annotation. We applied this model to three data sets, Corel5k, ESP game and Iaprtc12. The experiment compares our model with the traditional AIA methods based on KNN. The experiment result shows that our model has a certain improvement in accuracy, recall and *F1* score compared with the traditional model on the three data sets. In the future, we will mainly study how to use label relevance for large-scale deep learning-based image annotation networks.

Acknowledgement

This work was supported in part by Hubei Provincial Educational Science Planning Project (No.2019GA090), Hubei Province Technology Innovation Project (No.2019AAA045), Hubei Provincial Undergraduate Training Programs for Innovation and Entrepreneurship (S201910490051) and Survey Project from China Vocational Education Association of Hubei Province (No. HBZJ2020016).

References

1. Dutta, A., Y. Verma and C.V. Jawahar, Automatic image annotation: the quirks and what works. Multimedia Tools and Applications, 2018. 77(24): p. 31991-32011.
2. Cheng, Q., et al., A survey and analysis on automatic image annotation. Pattern Recognition, 2018. 79: p. 242-259.
3. Niu, Y., et al., Multi-modal multi-scale deep learning for large-scale image annotation. IEEE Transactions on Image Processing, 2019. 28(4): p. 1720-1731.
4. Zhang, W., H. Hu and H. Hu, Training Visual-Semantic Embedding Network for Boosting Automatic Image Annotation. Neural Processing Letters, 2018. 48(3): p. 1503-1519.
5. Kashani, M.M. and S.H. Amiri. Leveraging deep learning representation for search-based image annotation. 2017. Shiraz, Iran: Institute of Electrical and Electronics Engineers Inc.
6. Murthy, V.N., S. Maji and R. Manmatha. Automatic image annotation using deep learning representations. 2015. Shanghai, China: Association for Computing Machinery, Inc.
7. Hao, Z., H. Ge and L. Wang, Visual attention mechanism and support vector machine based automatic image annotation. PLOS ONE, 2018. 13(e020697111).
8. Amaral, I.F., et al. Hierarchical medical image annotation using SVM-based approaches. 2010: Institute of Electrical and Electronics Engineers Inc.

9. Cusano, C., G. Ciocca and R. Schettini. Image annotation using svm. 2004. San Jose, CA, United states: SPIE.
10. Ma, Y., et al., A weighted KNN-based automatic image annotation method. 2019.
11. Verma, Y. and C.V. Jawahar, Image Annotation by Propagating Labels from Semantic Neighbourhoods. International Journal of Computer Vision, 2017. 121(1): p. 126-148.
12. Jin, C. and S. Jin, Image distance metric learning based on neighborhood sets for automatic image annotation. Journal of Visual Communication and Image Representation, 2016. 34: p. 167-175.
13. Johnson, J., L. Ballan and L. Fei-Fei. Love thy neighbors: Image annotation by exploiting image metadata. 2015. Santiago, Chile: Institute of Electrical and Electronics Engineers Inc.
14. Yu, Y., W. Pedrycz and D. Miao, Neighborhood rough sets based multi-label classification for automatic image annotation. International Journal of Approximate Reasoning, 2013. 54(9): p. 1373-1387.
15. Zhang, M. and Z. Zhou, ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 2007. 40(7): p. 2038-2048.
16. Ma, Y., et al., CNN-feature based automatic image annotation method. MULTIMEDIA TOOLS AND APPLICATIONS, 2019. 78(3): p. 3767-3780.
17. Simonyan, K. and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2015. San Diego, CA, United states: International Conference on Learning Representations, ICLR.
18. Jiang J J, Conrath D W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy[J]. rocling, 1997:11512--0.
19. Dutta A, Verma Y, Jawahar C V. Automatic image annotation: the quirks and what works[J]. Multimedia Tools and Applications. 2018, 77(24): 31991-32011.