J. Vis. Commun. Image R. 33 (2015) 368–377

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Image automatic annotation via multi-view deep representation ☆

Yang Yang, Wensheng Zhang *, Yuan Xie

Institute of Automation, University of Chinese Academy of Sciences, China

ARTICLE INFO

Article history:
Received 16 February 2015
Accepted 9 October 2015
Available online 22 October 2015

Keywords:
Image annotation
Stacked auto-encoder
Imbalance learning
Multi-view learning
Image features
Semantic gap
Deep learning
Multi-labeling

ABSTRACT

The performance of text-based image retrieval is highly dependent on the tedious and inefficient manual work. For the purpose of realizing image keywords generated automatically, extensive work has been done in the area of image annotation. However, how to treat image diverse keywords and choose appropriate features are still two difficult problems. To address this challenge, we propose the multi-view stacked auto-encoder (MVSAE) framework to establish the correlations between the low-level visual features and high-level semantic information. In this paper, a new method, which incorporates the keyword frequencies and log-entropy, is presented to address the imbalanced distribution of keywords. In order to utilize the complementarities among diverse visual descriptors, we tactfully apply multi-view learning to search for the label-specific features. Thereafter, the image keywords are finally produced by appropriate features. Conducting extensive experiments on three popular data sets, we demonstrate that our proposed framework can achieve effective and favorable performance for image annotation.

1. Introduction

With the remarkable improvement of the information technic, the proliferation of digital images on the Internet posed a great challenge for large-scale image management. In order to organize, query and scan so large-scale images, image retrieval has been established. Text-based image retrieval (TBIR) [1], a typical image retrieval system, allows a user to present his/her information need as textual query and find the relevant images based on the match between the textual query and the manual annotations of images. Carefully chosen keywords can improve image retrieval accuracy, but the tagging process is known to be tedious and labor intensive [2]. Due to the tagged image databases containing rich information about the semantic meanings of the images, many image annotation algorithms exploit the exist tagged images to automatically annotate new images or add extra keywords to images with a few existing keywords.

The annotation process can be considered as a multi-label classification problem, and the existing methods can be grouped into two categories: generative models and discrimination models. Generative models try to learn the joint probability distribution between semantic concepts and image visual features, thus the image annotation can be achieved by using probabilistic inference [3–5,9,6–8]. Meanwhile, the discrimination models take the image annotation task as a supervised learning problem. Many typical supervised learning models have been introduced into this task, such as hidden Markov model (HMM) [10], supervised multi-class labeling (SML) [11], and support vector machine (SVM) [12,13]. Compared with the previous work, recent research efforts have taken more attention on extending the keyword prior knowledge and using hybrid model [14,18,19,17,15,16]. Recently, deep learning, as a novel machine learning algorithm, has achieved wonderful performance in the field of image understanding. This algorithm is also introduced to image annotation problem [25,28].

Although these algorithms have got wonderful results, the results are still unsatisfied. As a matter of fact, the annotation performance is limited in two major aspects:

(1) In real situations, image keywords from the Internet are extremely diverse with non-uniform distribution. Table 1 shows the image keyword distributions in three data sets, and we can see that the number of different keywords used for label images are seriously imbalanced. In this situation, the learning model trends to produce high accuracy for major keywords and poor performance for minority keywords. The average performance for all keywords of the model is limited by the minority keywords.
(2) Images are often described by diverse features which own different recognition capacities for different objects. For instance, in Fig. 1 "sky" and "sea" are likely identified by color, while "airplane" and "bird" are likely identified by shape. To capture information of the image as more as possible, we can concatenate different features into a long

☆ This paper has been recommended for acceptance by M.T. Sun.
* Corresponding author.

http://dx.doi.org/10.1016/j.jvcir.2015.10.006
1047-3203/

Table 1
Keyword distribution on the Corel-5 K, ESP game and IAPRTC-12.

Models	Keywords	Images	Avg. images/keyword	Max. images/keyword	Min. images/keyword
Corel5K	260	5000	66	1120	2
ESP game	268	20,770	364	5059	20
IAPRTC-12	291	19,627	386	5534	50

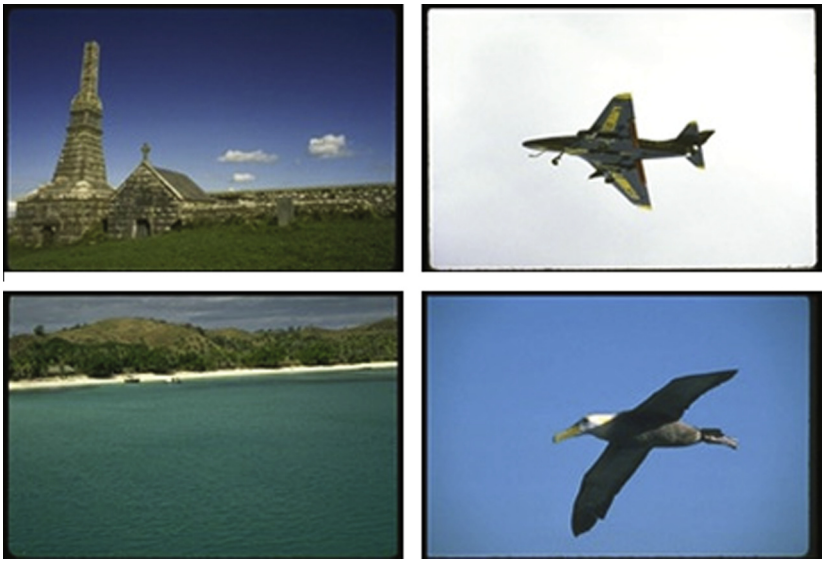

Fig. 1. Image examples that can be easily annotated by human.

vector. However, this concatenation causes overfitting in the case of a small size training sample and is not physically meaningful because each view has a specific statistical property. Therefore, how to utilize the different image feature properties is important in tagging process.

Stacked auto-encoder (SAE) [24], a typical deep learning structure, can learn extremely complicated relationships between image features and semantic information. Thus SAE is a suitable machine learning model for image understanding task. However, Compare with traditional image classification, image annotation usually has several keywords for each image. We modify the SAE model by applying sigmoid function as the SAE predictor and using ranking algorithm for generating image keywords. Further, the label number in image classification is usually limited, while the keywords for image annotation is always abundant. Due to this reason, image keywords can be seen as image text information. In this paper, we propose an iteration algorithm to retrain SAE twice.

To directly address the aforementioned limitations, we propose the multi-view stacked auto-encoder (MVSAE) with imbalanced learning. One aim of this paper is to present a simple method which can avoid low frequency keyword misclassification. Our method proposes to weigh different keywords depending on their frequency and adopt log-entropy to modify the object of SAE. For effectively utilize properties of different features, algorithm in [18] proposed a single label-specific classifier for each keyword. We tactfully incorporate the same idea into SAE and design a multi-view learning algorithm. The multi-view learning algorithm evaluates the annotation results on each keyword from different features and chooses the feature which has the best performance as the label-specific feature. As a result, each keyword can be produced by the most appropriate features. We carefully implement MVSAE for image annotation and conduct experiments on three popular image annotation data sets [14]. The experimental results demonstrate the effectiveness of MVSAE by comparing with the baseline algorithms.

The contributions of this paper are threefold:

(1) A novel SAE framework with sigmoid predictor is constructed for image annotation problem. Further, we propose a new iteration algorithm which combines the image visual features and text information as model inputs.
(2) The imbalance learning method with log-entropy weights is proposed for solving the problem of image keywords with imbalanced distribution.
(3) The multi-view SAE (MVSAE) is proposed to utilize the different descriptor properties.

The rest of this paper is organized as follows. Section 2 introduces a related work of image annotation. Section 3 describes the details of our proposed MVSAE framework. Section 4 compares our framework with the existing models and analyzes the results. Section 5 concludes the paper.

2. Related work

2.1. Advantage of deep learning

Deep neural networks, containing multiple nonlinear hidden layers, can learn very complicated relationships between their inputs and outputs. However, the nonlinear mapping between the inputs and outputs makes network be prone to local optimum and difficult to converge by back-propagation algorithm [20]. In order to overcome the learning problem of deep neural network,