ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Image Tag Assignment and Refinement Using CNN

Harshal Kothule, Saurabh Thombare, Harshada Chaudhari, Jayesh Patil

Department of Computer Engineering, SITS, NARHE, Pune, India

ABSTRACT: Tag-based image search is one of the important method to find images contributed by social users in such social websites. How to make the top ranked result relevant and with diversity is challenging Tag-based image search. It is commonly used in social media than content based image retrieval and context and content based image retrieval. Social image tag refinement is to remove the noisy or irrelevant tags and add the relevant tags. Data are randomly partitioned into two groups in our proposed system, the learning data and the testing data. The learning data is utilized to learn the proposed model and evaluate the performance of image tag refinement. The testing data is for image tag assignment and images are randomly chosen as the learning data while the rest ones are used as the testing data. The testing images are utilized to validate the effectiveness of image tag assignment. In this system, we propose a model for social image understanding with the consideration of images relevance and diversity.

KEYWORDS: CNN, DNN, Image Tagging, Social Image Analysis, Tag-Based Image Retrieval.

I.INTRODUCTION

Machine Learning is an idea to learn from examples and experience, without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves[1**]**.Image recognition, in the context of machine vision, is the ability of software to identifyobjects, places, people, writing and actions in images[2]. For humans interpreting the visual world comes easy. When humans see something, there is an inherent understanding of what it is. In most cases, there is no need for a conscious study of the object in order to make sense of it. However, for computers, it is an extremely difficult task because they can only manipulate digits.Deep learning is a part of machine learning algorithms that are recently introduced to solve complex, high-level abstract and heterogeneous datasets, especially image and audio data. There are several types of deep learning architectures, which are deep neural network (DNN), convolutional Neural Network (CNN), deep belief networks (DBN) and convolutional deep belief networks (CDBN)[3].

In real-world applications, many photo sharing websites, such as Flickr and Facebook, have been becoming popular, which facilitate millions of users to upload, share and tag their images. It leads to the dramatic increase in the number of images associated with user-provided tags available. For example, it is reported in March 2013 that Flickr had more than 3.5 million new images uploaded daily [4]. It sheds new light on the problem of image understanding. Unfortunately, these tags are provided by amateur users and are imperfect, i.e., they are often incomplete or inaccurate in describing the visual content of images, which brings challenges to the tasks of image understanding such as tag-based image retrieval [4].

In this work, we focus on refining image tags to complement relevant tags and remove the irrelevanttags, and assigning tags to new images.Image annotation is traditionally treated as a machine learning problem, which always depends on a small-scale manually-labeled data. However, they fail to handle large-scale social images due to the weakly-supervised data. Different from the traditional image annotation, tag refinement is to remove irrelevant tagsfrom the initial tags associated with images.With the advent of mobile and communication technologies, smart phones and other image capturing applications are increasing day by day. Social media has affected in our daily lives. People are increasingly becoming moreinterested in posting their daily experience online and sharing their feelings with others. Flickr is one of the decent photo sharing website which contains more than 10 billion photographs from people in different situations. A picture provides wealth information about users' preference, insight and sentiment. This information could be widely used in several fields such as campaign prediction, stock price forecast and advertisement recommendation. However, these pictures may consist of irrelevant information or sometimes unclear points. Therefore based on this messy information, it is hard to identify feelings and correct concepts in the pictures.Consequently, image analysis has received an increasing attention in recent years [5].

In computer vision, object detection is addressed as one of the most challenging problems as it is prone to localization and classification error. The current best-performing detectors are based on the technique of finding region proposals in order to localize objects. Despite having very good performance, these techniques are computationally expensive due to having large number of proposed regions. In this paper, we develop a high-confidence region-based object detection framework that boosts up the classification performance with less computational burden. In order to formulateour framework, we consider a deep network that activates the semantically meaningful regions in order to localize objects. These activated regions are used as input to a convolutional neural network (CNN) to extract deep features. With these features, we train a set of class-specific binary classifiers to predict the object labels.

Feature extraction can be defined as the fact of reducing an algorithm input data when this data is considered too large or redundant for processing. The result of feature extraction process is called feature vector. The feature vector is supposed to include the most relevant information from the input data which will allow the use of the reduced information instead of the whole initial data without compromising the task process. Selecting suitable variables is a critical step for successfully implementing an image classification. The use of too many variables in a classification task may decrease the performance of classification. It is imperative to choice only the variables that are most valuable. Therefore, for image classification in machine learning, the input for feature extraction process is a build derived features from an initial set of measured data. These features are intended to be informative, non-redundant, facilitating the subsequent learning steps, and in some cases leading to better human interpretations [6].

II.RELATED WORK

In the multimedia and data mining communities, many researchers focus on the problem of social image analysis. Different traditional image annotation method that usually learn models from small-scale manually-labeled images, these methods exploit massive images associated with weakly-supervised user-provided tags. In this subsection, we present the related work about social image tag refinement and social image tag assignment. Social image tag refinement is to remove the noisy or irrelevant tags and add the relevant tags. In, the group information of images from Flickr is exploited with the assumption that the images within a batch are likely to have a common style. Zhu et al. proposed to decompose the image-tag matrix into a low rank 1matrix and a sparse matrix and considered the content consistency and tag correlation as regularization terms. The low rank matrix recovery is combined with maximum likelihood estimation to recover the missing tags and de-emphasize the noisy tags in [2]. Zhuang et al. discovered the relationships between images and tags by exploiting the textual and visual contentsof images to refine tags. Tag co-occurrence is used to find the related tags with the original tags in. In, tag refinement is performed using a topic model, i.e., regularized latent dirichlet allocation. In, a latent space is identified based on low rank approximation to link the visual features of images and tags. In, the relevance between images and tags consistent with the observed tags and the visual similarity is learned. However, most of them cannot directly handle new images out of the learning image set [7].

1.In this paper [8], It can naturally embed new images into the subspace using the learned deep architecture. Besides, to remove the noisy or redundant visual features, a sparse model is imposed on the transformation matrix of the first layer in the deep architecture. Finally, a unified optimization problem with a well-defined objective function is developed to formulate the proposed problem. Extensive experiments on real-world social image databases are conducted on the tasks of image tag refinement and assignment. Encouraging results are achieved with comparison to the state-of-the-art algorithms, which demonstrates the effectiveness of the proposed method. It can naturally embed new images into the subspace using the learned deep architecture. Besides, to remove the noisy or redundant visual features, a sparse model is imposed on the transformation matrix of the first layer in the deep architecture [8].

2.In this paper [9],The number of images associated with weakly supervised user-provided tags has increased dramatically in recent years. User-provided tags are incomplete, subjective and noisy. In this work, we focus on the problem of social image understanding, i.e., tag refinement, tag assignment and image retrieval. Different from previous work, we propose a novel Weakly-supervised Deep Matrix Factorization (WDMF) algorithm, which uncovers the latent image representations and tag representations embedded in the latent subspace by collaboratively exploring the weakly-supervised tagging information, the visual structure and the semantic structure. Due to the well-known semantic gap, the hidden representations of images are learned by a hierarchical model, which are progressively transformed from the visual feature space. It can naturally embed new images into the subspace using the learned deep architecture [9].

3. In this paper [10],Text sentiment analysis has gained a great value in social networks due to its popularity and simplicity. Image sentiment analysis has also attracted a lot of attention through recent years. It is apparent that these approaches, neither text sentiment nor image sentiment analyzes are by themselves sufficient to obtain an accurate performance. On the other hand, the combination of them has compounded the problem. Thus, this paper provides a way to utilize the strengths of these techniques to develop a sophisticated method, called Supervised Collective Matrix Factorization (SCMF). The visual feature and textual feature are represented by Alexnet deep learning network and Bag of Glove Vector (BoGV) respectively. The proposed approach takes label information into consideration during matrix factorization, which is inspired by the graph Laplacian work [10].

4. In this paper [11]**,** Semi-Non-Negative Matrix Factorization is a technique that learns a low-dimensional representation of a dataset that lends itself to a clustering interpretation. It is possible that the mapping between this new representation and our original data matrix contains rather complex hierarchical information with implicit lower-level hidden attributes, that classical one level clustering methodologies cannot interpret. In this work we propose a novel model, Deep Semi-NMF,that is able to learn such hidden representations that allow themselves to an interpretation of clustering according to different, unknown attributes of a given dataset. We also present a semi-supervised version of the algorithm, named Deep WSF, that allows the use of (partial) prior information for each of the known attributes of a dataset, that allows the model to be used on datasets with mixed attribute knowledge.

5. In this paper [12], we study the usefulness of various matrix factorization methods for learning features to be used for the specific acoustic scene classification (ASC) problem. A common way of addressing ASC has been to engineer features capable of capturing the specificities of acoustic environments. Instead, we show that better representations of the scenes can be automatically learned from time–frequencyrepresentations using matrix factorization techniques. Recognizing acoustic environments is one of the challenging tasks of the more general Computational Auditory Scene Analysis (CASA) [2] research field and is receiving an increasing interest in the machine listening community. We show that the unsupervised learning methods provide better representations of acoustic scenes than the best conventional hand-crafted features on both datasets [12].

III.METHODOLOGY USED

Convolutional Neural Networks:

Convolutional Neural Networks (CNN) are deep neural networks introduced for image classification purpose. The second one is image classification. The CNN architecture is based on layers where the output of a layer is the input of the next one. The global CNN architecture can be divided in two parts. The first one devoted to image vector representation is essentially composed of convolutional layers whereas the second part used for image classification is a fully connected layers.Each layer delivers image representation level. The CNN particularity is that for a local image location we will use the weights. Weights shared for the same input location form a filter. Convolutional part of a CNN is a succession of: (1) a convolution of the input with a set of filters for local feature extraction; (2) a non-linearity function such as the logistic function, for non-linear input data representation learning; and a pooling function, which groups feature statistics at nearby locations and therefore decrease the computational cost [13].

Convolutional neural networks (CNN) consists of one or more convolutional layers, alternating with subsampling layers and by the end of the network, optionally, a fully connected MLP. The convolutional layers are responsible for feature extraction and is called feature map andsometimes feature detection. After convolutional layer, it is often paired up with a pooling layer that will perform a pooling function based on the inputs. The pooling layer is also known as a subsampling layer, and it will alternate with a convolutional layer because it computes the statistics of the convolutional layer. The pooling layer will perform pooling functions and is called min-pooling, maxpooling layers or etc. according to its context of problem solving at the end of the series of alternation, a fully connected MLP will be added. It works as a classification module for the network. This layer will receive all neurons from its previous layers whether they are convolutional or pooling and connect them with its own neurons [14].

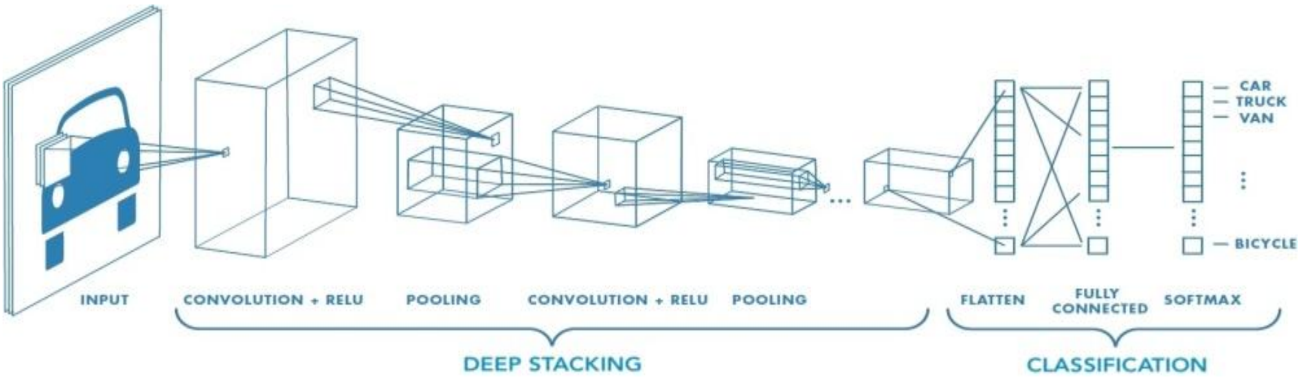


Fig.1 Convolutional Neural Networks

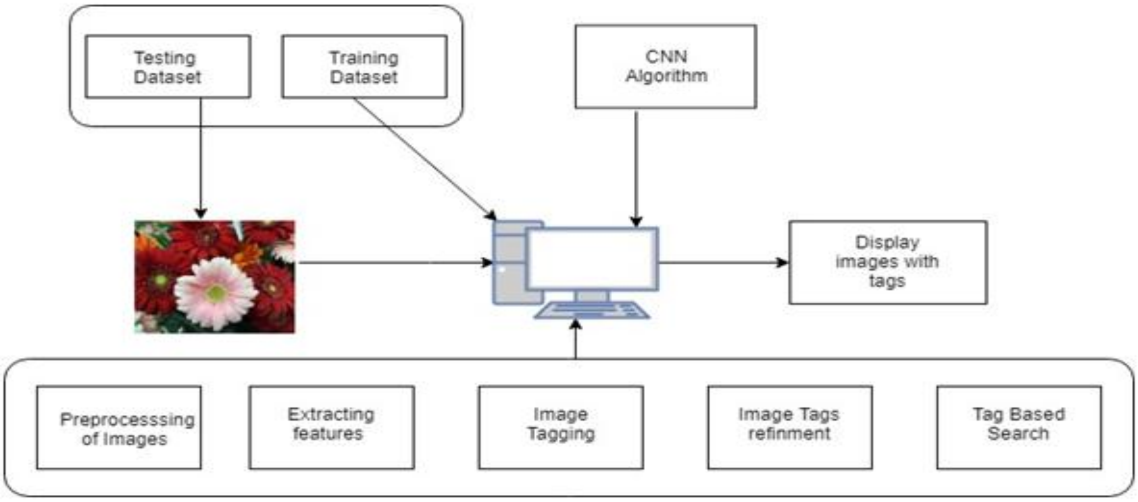


Fig 2. Architecture of proposed image tagging System

CNN image classifications takes an input image, process it and classify it under certain categories (Eg., Dog, Cat, Tiger,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Lion). Computers sees an input image as array of pixels and it depends on the image resolution. Based on the image resolution, it will see h x w x d(h = Height, w = Width, d = Dimension). Eg., An image of 6 x 6 x 3 array of matrix of RGB (3 refers to RGB values) and an image of 4 x 4 x 1 array of matrix of grayscale image.

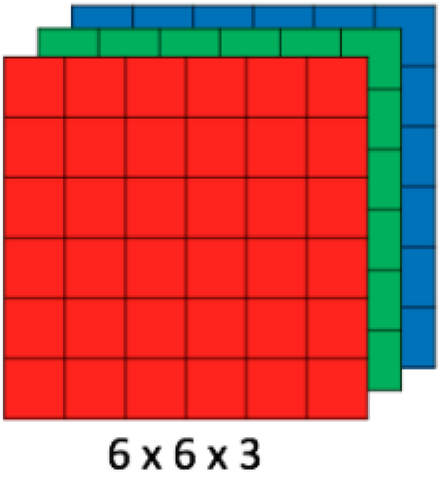
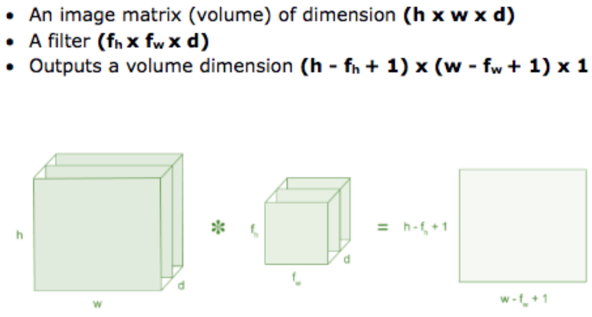


Figure 3 : Array of RGB Matrix

Convolution Layer

Convolution is the first layer to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel

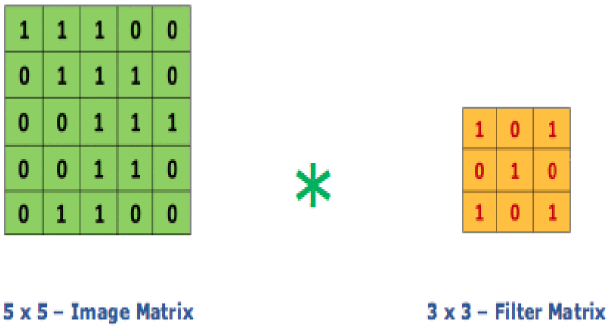


- An image matrix (volume) of dimension **(h x w x d)**
- A filter **(f_h x f_w x d)**
- Outputs a volume dimension **(h - f_h + 1) x (w - f_w + 1) x 1**

Figure 4: Image matrix multiplies kernel or filter matrix

Consider a 5 x 5 whose image pixel values are 0, 1 and filter matrix 3 x 3 as shown in below

DOI: 10.15680/IJIRCCE.2019. 0705061



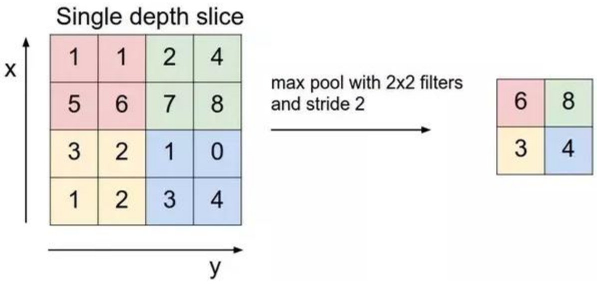
1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

1	0	1
0	1	0
1	0	1

Figure 5: Image matrix multiplies kernel or filter matrix

Pooling Layer

Pooling layers section would reduce the number of parameters when the images are too large. Spatial pooling also called subsampling or downsampling which reduces the dimensionality of each map.



1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

Figure 6: Image matrix multiplies kernel or filter matrix

ALGORITHM STRATEGY

Step 1: Start:

Step 2:Read image from testing dataset and pass it to the preprocessing:

Step 3: To initialize the neural network we create an object of the Sequential class. classifier = Sequential()

Step 4: To add the convolution layer, we call the add function with the classifier object and pass in Convolution2D with parameters**.**
classifier.add(Convolution2D(32, 3, 3, input_shape = (256, 256, 3), activation='relu'))

Step 5: In pooling, we reduce the size of the feature map. classifier.add(MaxPooling2D(pool_size=(2,2)))

Step 6: In flattening , all the pooled feature maps are taken and put into a single vector.classifier.add(Flatten())

Step 7: Full connection step is to use the vector we obtained above as the input for the neural network by using the Dense function in Keras.classifier.add(Dense(output_dim = 128, activation='relu'))

Step 8: We then compile the CNN using the compile function. This function expects three parameters: the optimizer, the loss function, and the metrics of performance. classifier.
compile(optimizer='adam',loss='binary_crossentropy',metrics=['accuracy'])

Step 9: ImageDataGenerator function works for flipping, rescaling, zooming, and shearing the images.
train_datagen=ImageDataGenerator(rescale=1./255, shear_range=0.2, zoom_range=0.2, horizontal_flip=True)

Step 10: The next step we need to do is create the training set.
training_set=train_datagen.flow_from_directory('training_set', target_size=(256, 256), batch_size=32, class_mode='binary')

Step 11: Now we use the predict method to predict which class the image belongs to.prediction=classifier.predict(test_image)

Step 12: End

IV.CONCLUSION

we propose a weakly supervised convolutional neural network for social image tag refinement and tag assignment method via the deep nonnegative low-rank model. The visual features and the high level tags are connected by the deep architecture. The tag refinement and the learning of parameters are jointly implemented, which makes the proposed method have good scalability. Extensive experiments are conducted on two widely used datasets and the experimental results show the advantages of the proposed method for tag refinement and assignment.To well handle the out-of-sample problem, the underlying image representations are assumed to be progressively transformed from the visual feature space. Besides, the proposed approach can deal with the noisy, incomplete or subjective tags and the noisy or redundant visual features. In future, we will focus on uncovering the latent structures of data and incorporating it into the proposed model in this work. How to extract representations from raw pixels based on the proposed model is also our future work.

REFERENCES

[1]https://www.expertsystem.com/machine-learning-definition/
[2]https://searchenterpriseai.techtarget.com/definition/image-recognition//
[3] Wang, T., W, D. J., Coates, A., & Ng, A. Y. (2012). End-to-end text recognition with convolutional neural networks. ICPR.
[4] Dieleman, S., mon Brakel, P., &Schrauwen, B. (2011). Audio-based music classification with a pretrained convolutional network. ISMIR.
[5] Le, D., & Provost, E. M. (2013). Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. IEEE.
[6] Hensman, P., &Masko, D. (2015). The impact of imbalanced training data for convolutional neural networks (Unpublished doctoral dissertation). KTH Royal Institute of Technology.
[7] G. Zhu, S.Yan and Y.Ma, "Image tag refinement towards low rank,content tag prior and error sparsity," in Proceedings of ACM *Int'i* conf. on multimedia ,2010,pp. 461-470.
[8]ZechaoLi ,Jinhui Tang "Deep Matrix Factorization for Social Image Tag Refinement and Assignment":
[9]Zechao Li, Jinhui Tang, "Weakly Supervised Deep Matrix Factorization for Social Image Understanding":
[10]Siqian Chen, JieYang ,Jia Feng , Yun Gu " Image Sentiment Analysis Using Supervised Collective Matrix Factorization" :
[11]George Trigeorgis, Konstantinos Bousmalis, StefanosZafeiriou, BjoernW.Schuller"Deep matrix factorization method for learning attribute representations" **:**
[12]Victor Bisot,RomainSerizel, Slim Essid ,Gaël Richard " Feature Learning With Matrix Factorization Applied to Acoustic Scene Classification":
[13]Yan, Y., Chen, M., Shyu, M.-L., & Chen, S.-C. (2015). Deep learning for imbalanced multimedia data classification. IEEE International Symposium on Multimedia.
[14]Liu, Y., Yu, X., Huang, J. X., & An, A. (2010). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Elsevier Ltd.
[15]www.google.com/search?q=medium.freecodecamp +cnn + diagram