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. 6, Issue 5, May 2018

Feature Learning and Collaborative Tagging Using Non-Personalized Social Media Images and Tags

Priyanka S. Padwal, Dr. Neeta A. Deshpande

Department of Computer Engineering, D.Y. Patil College of Engineering, Akurdi, Pune, India

ABSTRACT: In recent years the rush forwarded in popularity of social networks, users have changed from submissive receiver of information to active user and inventor of content. Feature representation for visual content is the key aspect in the growth of much primary function like the way of explanation for particular text or diagram. There exists almost inexhaustible multimedia content and their equivalent user behaviors such as “comments,” “like,” and “views.” We call them collective intelligence because they are contributed by collective social efforts from many users.
We describes an innovative feature learning standard depending on collective intelligence that is simple to acquire across the world. Proposed method classify images using CNN technique for deep learning classification. In which we apply single neural network to the full image. This network split the image into various components and deduce the bounding boxes and possibility for each constituents. We are carrying out extensive trial to show the efficiency of the this feature learning prototype in generating superior features for various tasks, such as retrieval and labeling, of an images. We concentrated on learning from the collection of user-generated images with captions because it is the most common social media and its intelligence can be exploited for many fundamental applications like image search and classification by using Flickr30K dataset which contains images and sentences.

KEYWORDS: Collaborative Tagging, Feature learning model, Flickr30K, Convolutional Neural Networks (CNN), Semantic relations.

I. INTRODUCTION

An improvement in multimedia applications is done because of the advances in feature illustration used for multimedia contents. From many previous years we observes evolution of visual features from the color histogram to SIFT interest points and to the recent deep learning features, that helps to move large varieties of multimedia applications from academic prototypes into industrial products. Normally the learning-based features by deep neural networks can outperform most hand engineered features and therefore free us to focus on designing algorithms and end applications. Attribute representation for multimedia content is the key concept to the advancement in many primary multimedia chore. The present upgrading in deep learning furnishes means towards this task, they are limited in application to domains where high-quality and large-scale training data are somewhat difficult to obtain.
The advancement in multimedia applications is basically due to the increased growth of feature representations for multimedia content.
To find users first choice and modeling, a well accepted tagging is considered. Collaborative tagging is mostly used, which allows users to contribute the function with keywords to tag photos, bookmarks, text, online content and other content. This approach is useful because people have more sources to find information.

Features of Deep learning are used to do better than hand-engineered features on many vision benchmarks, there is an inadequate research on how to further fine-tune the features for specific domains, especially for the social multimedia in the wild.. One possible cause might be the sparsity and noise in social media, which makes the problem very challenging.

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. 6, Issue 5, May 2018

A. Problem Formulation

We focus on learning from the collection of user-generated images with captions because its intelligence can be utilized for many fundamental applications like image/video search and classification.

To solve mentioned problem we propose a method of feature learning using social images and tags classify images using CNN technique for deep learning classification. The fresh content can constantly uploaded by the user with particular label for the images which are uploaded. It has been noted that users can interpret images, mostly according to their content. The tags given to the images are correlated to the image perspective and not randomly observable content. The label or tag may contain capturing device metadata, explanation of personal feelings, opinion and judgment of the user who assigned them. Even though it might contain completely wrong or confusing description.

B. Objectives

•Creates the image and word pair to maintain semantic relations.

•Seeks a latent embedding vector space for both images and words, where the similarity between them preserves how the images and words related to each other.

•Embedding space helps to learn a generic visual-semantic map that underpins our feature learning paradigm.
•Collaborative tagging and searching.

•Provides fast response time as it uses tagged and personal history search.

II. RELATED WORK

To acquire collective intelligence from huge multimedia data, which helps to better understand the semantic relations between images and words. A deep learning method is proposed [1] to learn image word embeddings from a million image-caption pairs of dataset. Then, images are embedded so that image features are extract .

Based on users inclination and modeling, the popular social activity of classification is taken into consideration. Attributes of an image have the vital role in a many image recognition problems. An improved algorithm for computing approximate PageRank vectors, [2] which permit us to find such a set in time proportional to its size.

Elia Bruni, Nam-Khanh Tran et al. recommend a model which can be easily modified in response with alteration to merge text and image on the basis of sharing information. Where these models remove contents of data completely from text, which is an very less compared to the wide variations that ground human semantic information [3].

"NEIL" is an artificial [4] semi-supervised learning (SSL) system which makes use of the large visual data to take out ordinary interrelation and followed by utilizing these relationships to label visual instances of existing categories without human intervention. It is an effort to build up the world's largest visual structured knowledge based on minimum human effort, that reflects the factual content of the images on the Internet, and that probably valuable to many computer based application.

The dataset used in [5] gives an estimation of previous image explanation and multilabel image categorization, on the basis of visual and text attributes. The web image dataset produced by NUSs Lab which includes: 2,69,648 images and the related tags from Flickr, with a total of 5,018 exclusive tags; and Six attributes extracted from these images.

The author investigates deep convolution features along with tasks like scene identification, domain alteration detection challenges. They also uses Term matching retrieval [6] by checking unreliability of document term as an arithmetical problem. These arithmetical procedures are used for the estimation latent structure free from obscuring "noise".

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. 6, Issue 5, May 2018

DeCAF is an implementation which is available for anyone to use [7]. A deep convolution creation features, together with linked network attributes to allow researchers to conduct tests with deep representations transversely to the collection of visual learning model.

Various instance learning patterns [8] are used to train visual detectors for terms that frequently come in heading part together with numerous parts of speech. So that we get the output in the form of maximum-entropy language representation. The language models find out from a set of 400,000 image interpretation to express the usage of words.

Xue Geng et al . gives deep model for feature demonstration with users and images. By altering the diverse networks into uniform low-dimensional illustration. The fast online algorithm [10] that scaled up to large networks in a parallel way is also developed. The outputs of these experiments display the feature learning approach significantly.

Stacked Auxiliary Embedding [11] that conveys knowledge from inadequate images to achieve better the accuracy in retrieval of image. They also describes CNN features work improved than traditional visual features for their task, with very economic dimensionality. This adds to the growing list of recent results in the vision community showing the effectiveness of pre-trained CNN activations as a common representation for recognition.

Authors tackle with the difficulty of recovering objects depending on open-vocabulary natural language queries [12] with an expression telling a detailed object, consider one example like corn flakes box, to discover the perfect match in the set of images. Where Dist-Belief is used for computing clusters with large number of machines to train large form.

The author brings a new standard of collection [13] for sentence-based image illustration and search, images containing 8,000 images such that every image matching with five different tags that give the overall representation of the significant feature and events. Micah Hodoshet al. [14] introduce a new benchmark collection for sentence-based image description and search, consisting of 8,000 images which offers plain explanation of the significant features.

III. PROPOSED SYSTEM

Tuning a good deep Convolutional Neural Networks (CNN) for feature learning is tricky. CNN itself is a technique of classifying images as a part of deep learning. In which we apply single neural network to the full image. It split the image in the distinct parts and forecast bounding boxes and the possibility for the distinct part. Every box is have some values by the forecasting the probability.

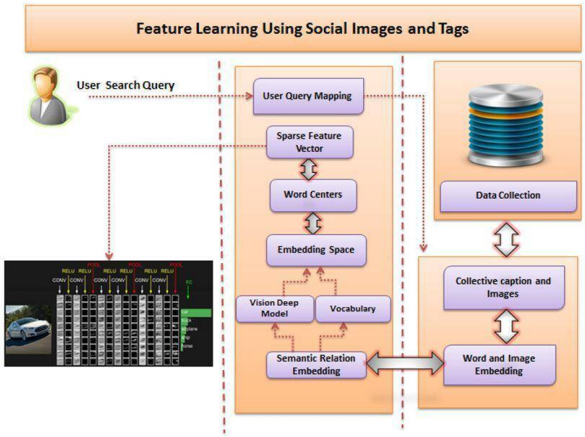

Fig. 1. Structural Design Diagram of Proposed Idea

The Fig.1 gives the architecture of the proposed scheme. The user can continuously upload new content and also provide tags for the images that are uploaded. It has been noted that users do tend to explain images, mainly according to their content. There exist tags that are associated with the image context and not to the directly observable content. Examples include tags that might be capturing device metadata, descriptions of personal feelings or thoughts of the user

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. 6, Issue 5, May 2018

who assigned them, non-existent words, ambiguity, and even completely irrelevant or misleading information. Users provide textual queries and the system responds with lists of images with relevant tags. To look into the user's priority and modeling, the social activity of tagging is measured. Collaborative tagging has become an approved way of distributing and arranging the resources, providing a large number of user-generated annotations.

Firstly, User Search for a Query, then the query is mapped using collective caption and images. Sparse Feature Vector is used for separating the categorized data into the classes and returns the highest predictive value related to the input query. Model is trained which creates the image and word pair to maintain semantic relations. Where the image and the tags are to be embedded as we can use them as resultant features. It seeks a latent embedding vector space. The images and words correspondence to preserves how they are associated with each other. This embedding space helps to learn a generic visual-semantic map that underpins our feature learning paradigm.

Convolutional layers transforming the result to the subsequent layer and it keep a track of response by every neuron. Every convolutional neuron routes data only for its accessible field.

We are using feed forward neural network that have been previously used to find out attribute as well as categorize data. Sparse Feature Vector is used for separating the categorized data into the classes and returns the highest predictive value related to the input query.

IV. SYSTEM IMPLEMENTATION

A. Algorithm Used

Traditional feature learning methods rely on semantic labels of images as supervision. They usually assume that the tags are evenly exclusive and thus do not pointing out towards the complication of labels. The learned features endow explicit semantic relations with words.

We also develop a novel cross-modal feature that can both represent visual and textual contents. CNN itself is a technique of classifying images as a part of deep learning. In which we apply single neural network to the full image.

1) Allows input of volume size W1*H1*D1
2) involve four hyper parameters:

a) Number of filters K

b) Their spatial extent F

c) The stride S

d) The amount of zero padding P

3) Generate a volume of size W2*H2*D2 where:
a) W2=(W1-F+2P)/S+1

b) H2=(H1-F+2P)/S+1

With constraint allocation, it bring in F*F*D1 weights per filter, for a total of (F*F*D1)*K weights and K biases.

4) In the output volume, the d-th depth slice (of size W2*H2) is the result of executing a suitable convolution of the d-th filter over the input volume with a stride of S, and then equalize by d-th bias.

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. 6, Issue 5, May 2018

5) A common setting of the hyper parameters is F=5,S=1,P=0

Still, there are general presentation signs and regulations of rule of thumb that stimulate these hyper parameters.

B. Dataset Used

We obtained the images from Flicker dataset. We gather the data like airplane images from Flickr along with their relations are collected. To get better correlation between the Flickr data and INRIA dataset results, we first search in Flickr using the queries which are mentioned in dataset to get the related images. For Collecting, training and testing purpose data is collected from following websites:
1.https://api.qwant.com/api/search/images?count=50&o set
=1&q=apple
2.http://lear.inrialpes.fr/ jegou/data.php

We use INRIA and Flickr Datasets to collect images and
related tags. Two binary file formats are used:

1) .siftgeo format:
Descriptors in this format are stored in raw form which includes the region information provided by the software of Krystian.This file format uses the file length to find the number of descriptors so there is no need of any header. A descriptor takes 168 bytes.

2) .fvecs format:
This one is used to store centroids. As for the .siftgeo format,this format also does not require any header.Centroids are stored in raw format. Each centroid takes 516 bytes.

V. EXPERIMENTAL RESULT

We obtained the images from Flicker dataset. The dataset includes:

- The images.

- The set of words which are taken out from these images, with the equivalent extractor and descriptor, for a different dataset (Flickr60K).

- Two sets of clusters used to quantize the descriptors. These have been obtained from Flickr60K.

We gather the data like images, users, and groups from Flickr along with their relations. To get a better correlation between the Flickr data and INRIA dataset results, we first search in Flickr using the queries which are mentioned in the dataset to get the related images.

The CNN splits the whole image into the different parts and forecast the bounding boxes and possibilities for each part.

We focus on learning from the set of user-generated images with captions because it is the most common social media and its intelligence can be exploited for many fundamental applications like image/video search and classification.

The CNN technique is used for classifying images. These Convolutional Neural Networks divide the figure into parts that defining characteristics of each part and predict restricted boxes and possibility for every region. It also performs the piece of work such as Image recovery, classification, and regular sentence generation. This system gives simple graphical user interface, permitted the user to search related images based on the tag.

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. 6, Issue 5, May 2018

TABLE I: .DATASET DISTRIBUTION FOR DIFFERENT TAG

Sr. No	Dataset	Images
1	Aeroplane	2000
2	Butterfly	3000
3	Flower	2000
4	Watch	2500
5	StarFish	3500

Table I shows distribution of different tags of images with different categories. In the proposed system, we used flicker dataset for social images. The dataset we collected is set of popular tags including airplane, butterfly, flower, watch, starfish images. These tags are used to get related images. Fig.2 shows graphical representation of dataset distribution which we have used for our experiments

Here we implement scheme in a real world comparison. The comparison is based on comparable security levels for texts. From the Table II.

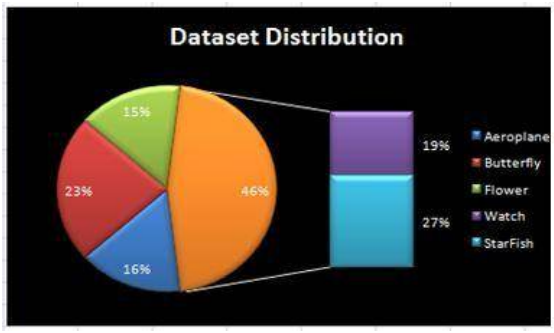

Fig 2. Graphical representation of dataset distribution

TABLE II: COMPARISON OF DIFFERENT METHODS

Sr. No	Algorithms	Tested On	Accuracy	% Accuracy
1	SVM Classifier	200	124	0.62
2	Random Forest	200	120	0.6
3	CNN(Our Approach)	200	182	0.91

From the Table II we come to the following conclusion:

1. CNN gives better performance than that of other two like SVM and RF.

2. As the as the number of training model increases the complexity is also increasing. So it is computationally more expensive, compared to CNN.

3. The RF algorithm works well for large dataset. It is necessary infrastructure to train them in a realistic time.

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. 6, Issue 5, May 2018

4. For a huge amount of user-generated annotation and better accuracy, we are using CNN.

Also we have measured and compare accuracy of different algorithms. Fig.3 shows graphical comparison of CNN Classifier with existing Classifiers such as SVM and Random Forest. Where accuracy tested on CNN classifier is higher than other two.

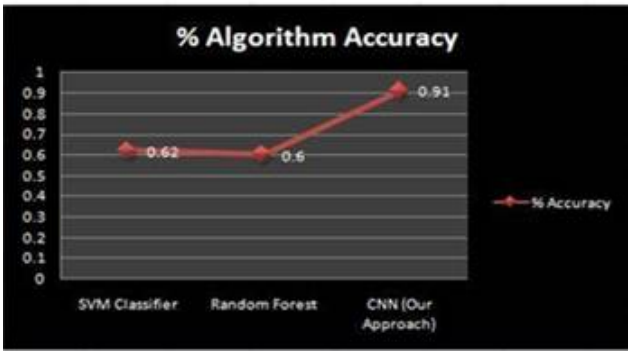

Fig .3 Average Accuracy of Proposed Classifier with Existing Classifiers

VI. CONCLUSION

Proposed Feature learning model describes social images with appropriate tags. We achieve collective intelligence from various communicative media which help out to analyze semantic connection in between images and words. To discover user choices and modeling, tagging is considered. Collaborative tagging allows users to contribute content with keywords to label the images and other content. Which helps users to find out more resources of information.

To conforms the efficiency of learned feature we will conduct various experiments through image applications. This model provides easy, and less expensive image extraction method. So, less time is required to retrieve the related images. More than one related outcomes occurs with single search.

However, millions of personal and non-personal images are uploaded to social media everyday. Thus, our future work would be to extend the Named Entity Tagging for non-personalized images.

REFERENCES

[1] Hanwang Zhang and Xindi Shang, Huanbo Luan Meng Wang Tat- Seng Chua Learning from Collective Intelligence: Feature Learning Using Social Images and Tags. In ACM Trans. Multimedia Comput. Commun. Appl., Vol. 13, No. 1, Article 1, Publication date: November 2016
[2] Reid Andersen, Fan Chung, and Kevin Lang. 2006. Local graph partitioning using pager-ank vectors. In 2006 47th Annual IEEE Symposium on Foundations of Computer Science(FOCS06). IEEE, 475486.
[3] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. J. Artif. Intell. Res. 49, 147 (2014).
[4] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. Neil: Extracting visual knowledge from web data. In Proceedings of the IEEE International Conference on Computer Vision. 14091416.
[5] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng.2009. NUS-WIDE: A real-world web image database from national university of Singapore.In Proceedings of the ACM International Conference on Image and Video Retrieval. ACM 48.
[6] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41, 6 (1990).
[7] Je_ Donahue, Yangqing Jia, Oriol Vinyals, Judy Ho_man, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A deep convolutional activation feature for generic visual recognition. In International Conference on Machine Learning. 647655.
[8] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar,Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, and others. 2015a. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 14731482.
[9] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Je_ Dean, Tomas Mikolov, and others. 2013. Devise: A deep visual-semantic embedding model. In NIPS.29
[10] Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning image and user features for recommendation in social

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. 6, Issue 5, May 2018

networks. In Proceedings of the IEEE International Conference on Computer Vision. 42744282

[11] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In European Conference on Computer Vision. 529545.

[12] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V. Le, and others. 2012. Large scale distributed deep networks. In Advances in Neural Information Processing Systems. 12231231.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[14] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. J. Artif. Intell. Res. 47 (2013), 853899.30