Automatic Image Annotation via Combining Low-level Colour Feature with Features Learned from Convolutional Neural Networks

Yi Lin*, Honggang Zhang

ABSTRACT

In this paper, a feature combination approach to annotate and retrieve images is proposed. In addition to using low-level colour features from original images, we extract features learned from convolutional neural networks (CNNs). We find these two sets are complementary to each other in conducting automatic image annotation (AIA). For both single-label CIFAR-10 and multi-label COREL-5K AIA tasks, the CNN-learned features perform slightly better than the low-level image features. Finally, when combining the two feature sets as inputs into the deep neural network-based AIA systems, we obtain the best performance in both cases.

Key Words: Automatic Image Annotation, Deep Learning, Convolutional Neural Networks, Feature Combination

DOI Number: 10.14704/nq.2018.16.6.1612 **NeuroQuantology 2018; 16(6):679-685**

Introduction

Automatic image annotation (AIA) is an important problem in computer vision (*Barnard et al.*, 2003). As images often contain complex and different kinds of content information, how to query, retrieve, and organize image information quickly and effectively becomes a crucial issue. It is a difficult task because of the lack of correspondence between keywords and images (Gupta *et al.*, 2012). AIA is a labelling problem wherein the task is to predict multiple textual labels for an unseen image describing its contents or visual appearance (Makadia *et al.*, 2008; Murthy *et al.*, 2014). Compared with a simple label image retrieval, a multi-label solution yields more information. It also matches relevant images faster and with more accuracy. This paper addresses the image multi-label annotation problem based on the deep learning model.

Previous work on AIA has developed several techniques to address these issues, such as the cross-media relevance model (CMRM) (Jeon *et al.*, 2003), multiple Bernoulli relevance model (MBRM) (Feng *et al.*, 2004), continuous-space relevance model (CRM) (Guillaumin *et al.*, 2009), maximum entropy (ME) (Garneiro, 2007), Markov random field (MRF) (Globerson and Roweis, 2006), and conditional random fields (CRF) (He, 2004).

However, inspired by the theory of the cranial nerve, deep neural networks have started to become widely used in the field of computer vision, natural language processing, and so on. In 1943, Warren McCulloch and Pitts proposed and presented the concept of the artificial neural network and the mathematical model of artificial neurons, which is considered as the foundation for the theory of neurons in the field of biology and physiology. The milestone in artificial neural network research is the invention of the Backpropagation algorithm (BP) (Rumelhart *et al.*, 1986).

Corresponding author: Yi Lin
Address: School of information and communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
e-mail ✉ linyibupt@outlook.com
Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
Received: 4 March 2018; **Accepted:** 27 April 2018

The artificial neural network in such aspects as structure principle and function features is closer to the human brain. It can adapt to the environment itself, Summarize laws, perform some operations, identification, or process control. It was not until 2012 that the artificial neural network became popular again due to the success of using the deep convolutional neural network in image classification (Krizhevksy, 2012; Simonyan and Zisserman, 2014).

Recent advances in the deep neural network have brought new intuitions for the AIA problem. The convolutional neural network (CNN) instinctively extracts hidden features of images. So in this paper, we propose the new approach of combining the CNN's visual features with conventional features. Specifically, we present a new deep neural network architecture that adds colour features to the first fully connected layer of the CNN for multi-label image annotation. We test the proposed approach on CIFAR-10 and COREL-5k and compare it with existing work. The feature vector combination realized complementary and achieved a better performance than existing work on the CIFAR-10 and COREL-5k datasets.

Related Work

A number of models use associations between terms and images for image annotation and retrieval. Generative models can randomly generate observational data, especially if certain implicit parameters are near-given. It can further classify the theme model and the mixed model. The theme model annotates the image as a sample of a particular mixed subject, each of which is the visual feature of the image and the distribution of the label. Prominent examples include the cross-media relevance model (Jeon *et al.*, 2003) (CMRM). It considers the task as a translation, but borrows ideas from cross-lingual information retrieval, and thus allows for both image annotation and retrieval. The discriminative model learns a separate classifier for each tag and uses the classifier to predict the test image and to determine whether to mark an image with a tag. Unlike the generation model, the discriminative model does not take into account the joint distribution between the visual features of the image and the label. However, as with the generation model, the discriminative model also selects the visual features of the image in advance, and does not analyze the differences between the features. The Translation Model (Duygulu *et al.*, 2002) uses a classic machine translation technique to translate from a vocabulary of terms to a vocabulary of blobs. The Nearest Neighbour model is becoming more and more popular with the growth in training data. The work in (Makadia *et al.*, 2008) introduces a transfer mechanism of the nearest neighbour tag, which models the image annotation as a retrieval problem. Nearest neighbours depend on the average of the distances calculated by visual features, also known as joint equal contributions.

For an image to be tested, the label is passed through the neighbour. The basic colour and texture of the visual features are used for comparison and testing; the regularization of feature selection is also based on the similarity of the label to consider. However, it does not increase sparseness and does not significantly improve accuracy. The MBRM (Feng *et al.*, 2004), the CRM (Guillaumin *et al.*, 2009), which is a continuous version of the CMRM model that performs favourably, maximum entropy (ME) (Garneiro, 2007), the Correspondence LDA Model (Gao *et al.*, 2006) (CLDA) allows annotation and retrieval. It is based on latent Dirichlet allocation (Grangier and Bengio, 2008) and is a generative model that assumes a low dimension. An image is divided into a set of sites or elements (i.e. grids or regions) and the visual features extracted from the grids are used for representing images. For the CRM and MBRM models, the Gaussian densities are applied to modelling distributions of visual features, while others apply k-means clustering to tokenize the image elements.

In recent years, as the deep neural networks continue to develop, some scholars have started to study how to use them in computer vision. The architecture of deep neural networks to some extent mimics the activity in layers of neurons in the neocortex. Take the neurons in the primary visual cortex as an example. Figure 1 illustrates the organization of the primary visual cortex, where the grey matter in the primary visual cortex is divided into six layers, namely I, II, III, IV, V, VI which comprise different types of neurons (Lyes, 2012). In each of these small visual cortices, which are not composed entirely of neurons, there are still different hierarchies. At this level, the neural network in the human brain is DNN+ CNN + RNN plus the pulse as the encoding. It's more like DNN in the layer, which is very similar to CNN, and it's going to be RNN in time.

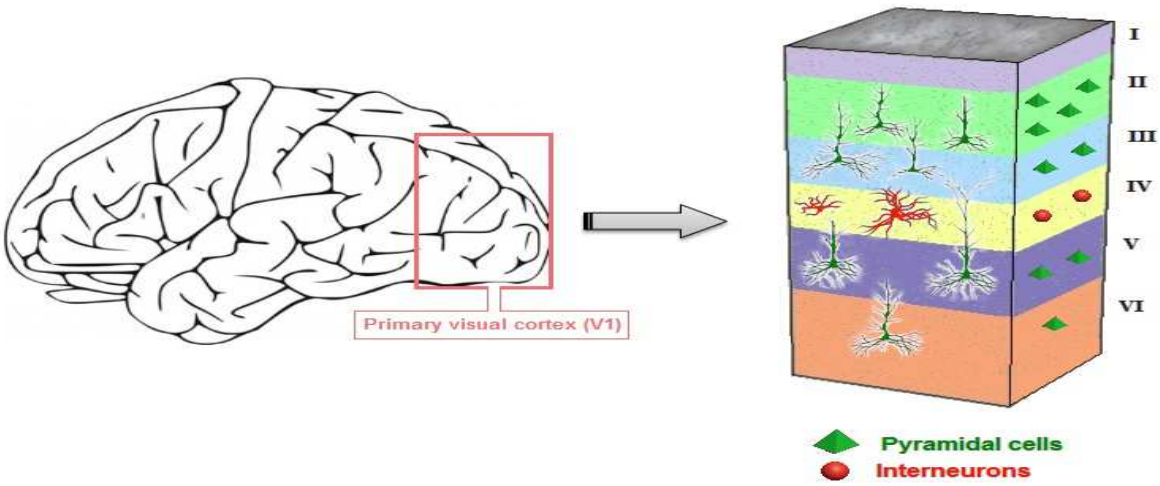

Figure 1. Organization of the primary visual cortex

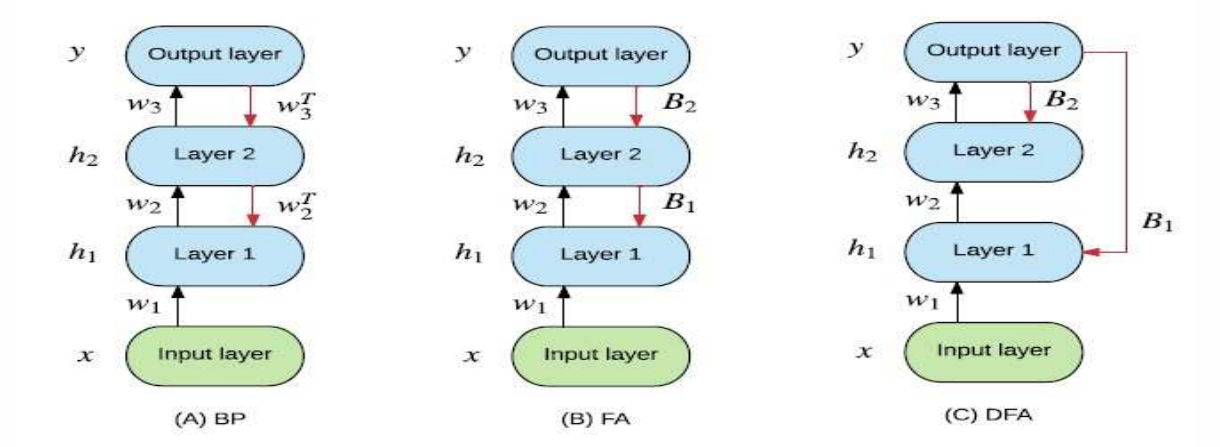

Figure 2. Training methods

The training mode of the deep neural network is mainly reverse propagation, which is propagated from the output layer to the first layer, and the errors are continuously corrected in each layer, as shown in Figure 2. But there is no such thing as a reverse transmission mechanism in the brain. The simplest explanation is that the neuron signals are directional, and there is no chance to return the signal to the next level.

The development of the deep neural network which was inspired by the cranial nerve is now used in various fields such as speech recognition and natural language processing. Especially in the field of image recognition, the visual system of the human eye can be thought of as a combination of three parts; the three areas of the links are the optic nerve and optic radiation. They are neuronal axons, similar to the interlayer links in deep learning neural networks.

In 2012, Hinton *et al*., used the multi-layer convolutional neural network for image classification in the extensive large-scale database ImageNet for the current image recognition, and achieved very good recognition results. There has been a lot of research on the convolutional neural network regarding structure, performance and other aspects of improvement. However, most of the research to solve the problem of image classification does not take into account the situation of multi-label images. This paper presents an automatic image annotation solution based on the deep neural network model by comparing the performance characteristics of the current deep learning model of the proposed AIA method.

Automatic Image Annotation

Here we provide details about how to fuse different features in the new deep neural network architecture for image annotation. We first give an overview of the architecture of the deep neural network, followed by details about the feature extraction method for conventional colour features and the CNN visual features.

Architecture Overview

Figure 3 illustrates the architecture of the deep neural network for image annotation:

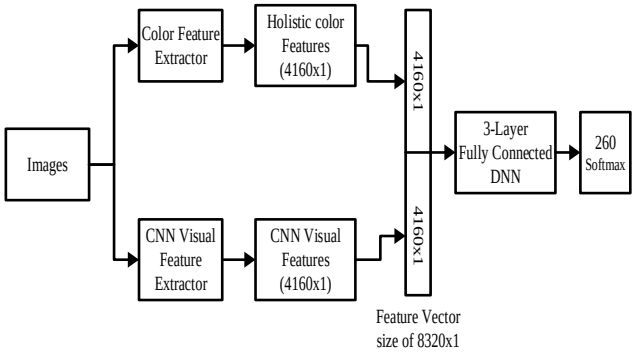

Figure 3. Architecture of the image annotator

The training images are simultaneously fed to two feature extractors. The Colour Feature Extractor (CFE) borrows the idea of "Bag of Words"(BOW) and outputs the holistic colour features of an image. The CNN Visual Feature Extractor (VFE) leverages the inherent capability of feature extraction of the convolutional layers and outputs visual features. The two kinds of features are then combined and fed into a three-layer, fully connected deep neural network, which outputs the distribution of the labels with respect to the image. By training this neural network, we obtain a multi-label image annotator with both low-level colour features and high-level features from CNN.

Details about the two extractors and the three-layer, fully connected DNN are discussed in the following section.

Colour Feature Extractor

The design of the Colour Feature Extractor (CFE) is inspired by the Bag of Words (BoW) in document representation. In document processing, a text document is modelled as a sequence of words in a lexicon. The BoW solution ignores the order of words and uses a high-dimensional vector to represent the document. This vector will encapsulate the statistics of terms that occur in the document, such as the co-occurrence of semantic and syntactic relations. The vector is considered as a feature of the document and can be used in document categorization.

Similarly, the CFE builds a visual lexicon and represents an image using a set of visual words. More specifically, we focus on colour features, so we consider the colour feature of a block of an image as a visual. By encapsulating the statistics feature and category of the visual words of an image, we represent an image's colour feature as a high-dimensional vector. The details of the operation of the CFE are explained as follows.

Firstly, we divide an image into a set of blocks of 6x6 pixels. The colour histogram of each block is then calculated. The colour feature of a block is denoted as a C-dimensional vector where C is the number of colours in the colour space. In our case, we use eight-bit colour images so C is 256. The Colour Feature Vector (CFV) of the block in the ith row and jth column, Bi,j is denoted as Eq.(1).

$$CFV_{i,j} = \{c_1, c_2, \ldots, c_C\} \in \mathbb{R}^C \quad (1)$$

Where c_i represents the percentage of the pixels of colour i in the block Bi,j. For a concrete example as illustrated by Figure 4, the colour feature vector of block B1,1 is calculated as CFV1,1= {0.8,0.15,0.05, 0, 0, ...}, assuming that c1, c2 and c3 denote the percentage of white, red and blue pixels and 0.8, 0.15 and 0.05 are the corresponding value.

Secondly, based on all the colour feature vectors of all the blocks in all the images in the training set, we build a visual lexicon with K visual words. The lexicon is built via clustering all the CFVs. We use the k-means algorithm to cluster all CFVs in the training set. Therefore, if a CFV is categorized into a cluster i, the corresponding block Bi,j is represented by a visual word wi. As a result, the image with m×n blocks is transformed as a document with m×n visual words. Again, looking at the example in Figure 4, the visual word corresponding to each block is marked as the underlined symbol at the upper right corner of the block. If two blocks have the same visual word representation, e.g., B1,2 and B2,2, it is because their CFVs have been categorized into the same cluster, i.e., w2.

Based on the operation, we extract the colour feature of the image as extracting the statistic feature of the document. The colour feature of an image contains two parts:

1) We consider the normalized frequency of a visual word occurrence in an image. This will produce a K dimensional vector where the i th component of the vector is the frequency that the word w_i appears in the image. For the example in

Figure 4, the visual word occurrence frequency vector of the image is V= {1/9,2/9,1/9,1/9,1/9,2/9,1/9,1/9,1/9,0,0, ...} K.

2) Besides the visual word occurrence frequency, we are also concerned with the words: co-occurrence. The operation to calculate the visual word co-occurrence is similar to the words: bigram co-occurrence calculation in document processing. For any pair of two visual words (wi, wj), we count the pair's occurrence in adjacent blocks. For the example in Figure 4, the block $B_{2,2}$ is adjacent to block B1,1, B1,2, B1,3, B2,1, B2,3, B3,1, B3,2, B3,3. When calculating at block $B_{2,2}$, the co-occurrences of pairs (w2,w1), (w2,w2), (w2,w3), (w2,w4), (w2,w6), (w2,w7) are added by one and the co-occurrence of (w2,w5) is added by 2. By sliding the calculating window through all the blocks in the image, we get a K×K matrix WK×K whose component wi,j is the normalized co-occurrence frequency of the word pair (wi, wj).

Finally, we stretch the co-occurrence matrix W into a K2×1 dimensional vector and combine it with the occurrence vector V. We get a K+K2 dimensional vector that we use as the colour feature vector for an image. In our particular case, we set K equal to 64. Therefore, the Colour Feature Extractor outputs a 4160×1 dimensional vector as the holistic colour feature of an image which will be further fused with the CNN feature for multi-label image annotation.

$B_{1,1}$ $w1$ $B_{1,2}$ $w2$ $B_{1,3}$ $w3$
0.15
0.05
$B_{2,1}$ $w4$ $B_{2,2}$ $w2$ $B_{2,3}$ $w5$
$B_{3,1}$ $w6$ $B_{3,2}$ $w7$ $B_{3,3}$ $w5$

Figure 4. Illustration of colour feature calculation of an image

CNN Visual Feature Extractor

The convolutional neural network(CNN) is an efficient model in image classification. As an image passes through each convolutional layer, the intermediate output images naturally reveal the image's high-level abstract features. Therefore, rather than using the CNN directly as a multi-label image classifier, we use the CNN architecture as a high-level visual feature extractor. Then along with the colour feature from CFE, the CNN feature is fed to the fully connected deep neural network. Note that the colour feature from CFE is a constant vector for an image and the CNN visual feature is updated repeatedly while training the neural network.

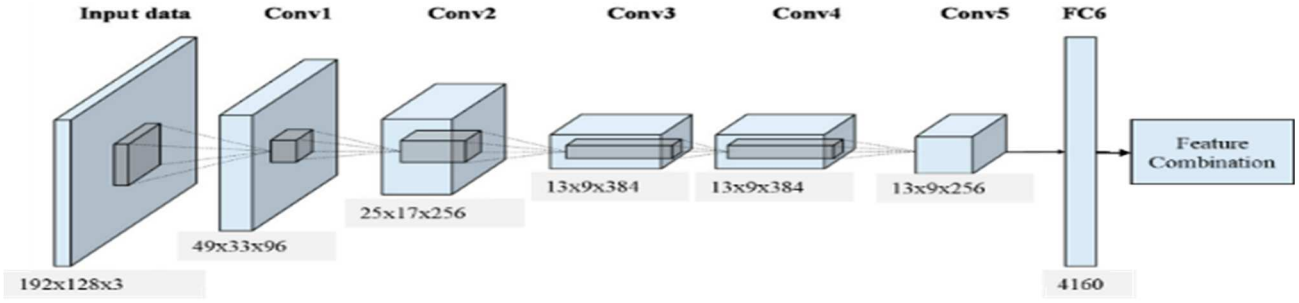

Figure 5. Architecture of the CNN visual feature extractor

Table 1. Parameter setting for CNN visual feature extractor

Layer	Description and Setting
Input Layer	32x32 colour image for CIFAR-10, 192*128*3 Pixel Image for COREL-5K Dataset
1st Conv Layer	96 filters of 11 * 11 * 3 - Stride 4 - Response Normalization - Max-Pooling - ReLU
2nd Conv Layer	256 filters of 5 * 5 * 48 - Response Normalization - Max-Pooling - ReLU
3rd Conv Layer	384 filters of 3 * 3 * 256 - ReLU
4th Conv Layer	384 filters of 3 * 3 * 192 - Max-Pooling - ReLU
5th Conv Layer	256 filters of 3 * 3 * 192 - ReLU
1st Fully Connected Layer	4160 neurons - ReLU
Feature Combination	4160-d Colour Feature Vector with 4160-d CNN Visual Feature Vector: 8320-d Feature Vector
2nd Fully Connected Layer	64 neurons - ReLU
3rd Fully Connected Layer	8320 neurons - ReLU
Output Layer	260 Softmax for COREL-5K, 10 Softmax for CIFAR-10

Training the Image Annotator

We train our models using stochastic gradient descent. The colour feature vector is pre-calculated before being fed to the DNN and keeps constant during the learning process. The CNN visual feature vector and all weights are updated in every training epoch. The weights in each layer are initialized from a zero-mean Gaussian distribution with standard deviation 0.01. The initial neuron biases in the second, fourth, and fifth convolutional layers, as well as in the fully connected hidden layers, are set at the constant 1. We trained the network on Keras with two NVIDIA GTX 980 4GB GPUs.

Experiment and Evaluation Analysis

We evaluate the performance of our image annotator with two datasets. We use CIFAR-10 to evaluate the annotator's performance for single-label classification and use COREL-5k for multi-label annotation. All shown results are measured by the averages of precision (mP), recall (mR), and F1 (mF) over all keywords. The details of the results are discussed in the following section.

Single-Label AIA on CIFAR-10 Dataset

The CIFAR-10 dataset consists of 60,000 32x32 colour images in 10 classes, with 6000 images per class. We use 50,000 images for training and 10,000 images for the test. Figure 6 gives the results of our annotator for single-label image classification in a single-label dataset CIFAR-10.

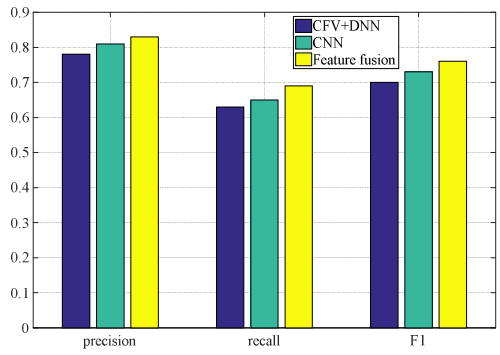

Figure 6. CIFAR-10 results

The results show that our feature combination solution outperforms the stand-alone CNN solution and the DNN with HSV colour feature input solution.

Multi-Label AIA on COREL-5k Dataset

For multi-label image annotation, we test our solution on the COREL-5K dataset. This dataset was first used in J. Jeon's paper (Jeon *et al.*, 2003). Since then, it has become an important benchmark for keyword-based image retrieval and image annotation. It contains around 5000 images manually annotated with 1 to 5 keywords. The vocabulary contains 260 words. We use a subset of 500 images for testing, and the rest is used for training. A comparison is made with stand-alone CNN solution, DNN solution, DNN with softmax regression and several state-of-the-art statistic solutions, including CMRM (Jeon *et al.*, 2003), MBRM (Feng *et al.*, 2004), SML (Guillaumin *et al.*, 2009), LASSO, JEC (Makadia *et al.*, 2008), TGLM (Liu *et al.*, 2009), TagProp (Guillaumin *et al.*, 2009), Group Sparsity (Zhang *et al.*, 2010), and CCD (SVRMKL+KPCA) (Nakayama, 2011).

Table 2. Comparison with state-of-the-art statistic solutions on COLER-5k

	Precision	Recall	F1-Score
CMRM	0.16	0.19	0.17
MBRM	0.24	0.25	0.24
SML	0.23	0.29	0.26
LASSO	0.24	0.29	0.26
JEC	0.27	0.32	0.29
TGLM	0.25	0.29	0.27
TagProp	0.33	0.42	0.37
GS	0.30	0.33	0.31
CCD	0.36	0.41	038
DNN	0.34	0.31	0.33
DNN+softmax regression	0.37	0.34	0.36
CNN	0.38	0.34	0.36
FF-CNN	0.41	0.37	0.39

As can be seen from Table 2, the solution approach proposed in this paper improves the accuracy and recall rate compared to existing methods. We also find that the combination of the CNN visual feature with the artificial extraction feature shows a certain complementary and can achieve better results.

Annotation Examples in COREL-5k

In Figure 7, we qualitatively assess our feature fusion image annotator through some examples. For each image shown below, its identity number (ID), ground truth (Truth), and annotation results are listed in that order. Different from other annotation algorithms that use fixed-length labels to annotate images, our proposed algorithm obtained sets of annotation results of variable sizes according to the confidences of keyword models.

We found that some of the image labels were mistakenly annotated and the semantics ambiguous, such as" lake"and "sea" in Figure 7(a).

However, some labels annotated by our FF-CNN are considered rational although these labels are not marked in the ground truth reference, such as "grass" in Figure 7(b) and "water" and "sky" in Figure 7(c). These extra labels may downgrade the quantitative benchmarks. They do reveal some extra information of the image, such as background. So in other words, our FF-CNN multi-label annotator may perform even better than the quantitative benchmark in Table 2.

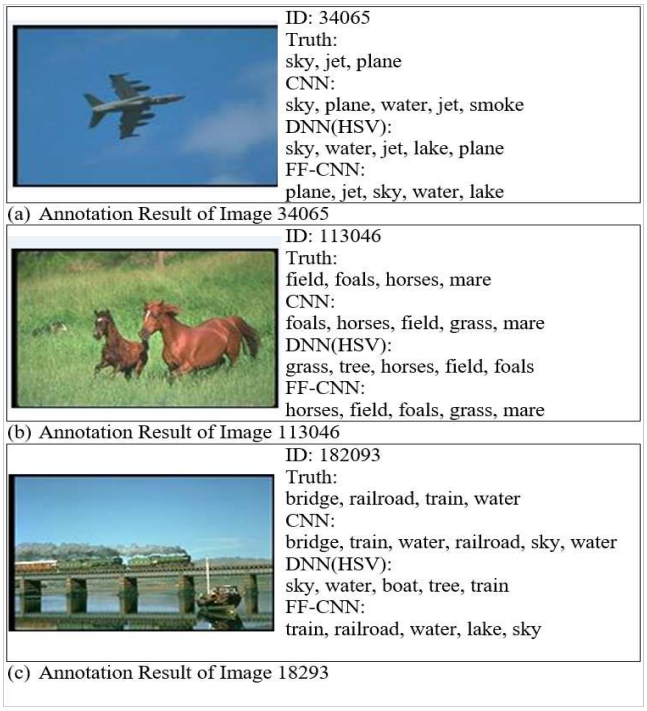

Figure 7. Qualitative assessment with some image examples

Conclusions

In this paper, we propose a new approach to multi-label image annotation which combines the CNN visual feature with the conventional colour feature. We i) first extract the colour feature from the original images, build a visual lexicon, and use a bigram to present a co-occurrence relation, ii) then input images into the CNN through five convolutional layers and let the pooling layer attain a visual feature, and iii) finally combine the two features at the first hidden layer of the DNN system. Our experimental result is compared with state-of-the-art solutions. The proposed framework has verified the feasibility of our work on the CIFAR-10 dataset and achieved better results on the COREL-5k dataset over baseline solutions.

References

Barnard K, Duygulu P, Forsyth D, Freitas ND, Blei DM, Jordan MI. Matching words and pictures. Journal of Machine Learning Research 2003; 3(2): 1107–35.

Duygulu P, Barnard K, Freitas JFGD, Forsyth DA. Object recognition as machine translation, Learning a lexicon for a fixed image vocabulary. European Conference on Computer Vision 2002; 2353(6): 97-112.

Globerson A, Roweis ST. Metric learning by collapsing classes. Advances in Neural Information Processing Systems 2006: 451-58.

Grangier D, Bengio S. A discriminative kernel based approach to rank images from text queries. IEEE transactions on Pattern Analysis and Machine Intelligence 2008; 30(8): 1371-84.

Guillaumin M, Mensink T, Verbeek J, Schmid C. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. IEEE International Conference on Computer Vision In ICCV 2009; 30(2), 309-16.

Gupta A, Verma Y, Jawahar CV. Choosing linguistics over vision to describe images. AAAI Conference on Artificial Intelligence 2012; 5(1): 606-12.

He X, Zemel RS. Multiscale conditional random fields for image labeling, IEEE Computer Society Conference on Computer Vision & Pattern Recognition 2004: 695-703.

Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval 2003; 2003: 119-26

Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 2012; 60(2): 1097-05.

Liu J, Li M, Liu Q, Lu H, Ma S. Image antation via graph learning. Pattern recognition 2009; 42(2): 218-28.

Makadia A, Pavlovic V, Kumar S. A new baseline for image antation. European Conference on Computer Vision 2008; 5304: 316-29.

McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics 1943; 5(4): 115-33.

Murthy VN, Can EF, Manmatha R. A hybrid model for automatic image annotation. International Conference on Multimedia Retrieval 2014; 2014: 369-76.

Nakayama H. Linear distance metric Learning for large-scale generic image recognition. PhD thesis, The University of Tokyo Japan, 2011.

Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986; 323(6088): 533-38.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.

Zhang S, Huang J, Huang Y. Automatic image antation using group sparsity. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010; 119(5): 3312-19.