Cat or Dog Images Classification: BOVW Method vs. CNN

Yubin Fan
yubfan@pdx.edu

ABSTRACT

Bag of Visual Words (BOVW) is a more efficient and resource-saving image classification method than CNN. This paper confirms this statement by comparing these two approaches on Cat or Dog image classification problem.

1. INTRODUCTION

Distinguishing between cat and dog images is a classic computer vision problem and a CNN approach could achieve a quite high accuracy. However, training a CNN requires a lot of time, hardware resources and large quantity of training samples. Another classification algorithm is Bag of Visual Words (BOVW). SIFT features could be described as text descriptors by a vocabulary of visual words. Then these "words" could be sent to some machine learning classifier (e.g., a support vector machine) for classification. Classifiers like SVM need small size of training data and less hardware resources to achieve a good performance.

Intuitively, BOWL is a more efficient and resource-saving classification method than CNN. This paper will confirm this hypothesis by comparing these two algorithms on Cat or Dog Images Classification problem.

This paper is structured as follows: In section 1 we overview the main structure of CNN model. In section 2 we briefly review the Bag-of-Visual-Words approach for image classification. In section 3 we introduce multiple classifiers of BOVW. In our experiments, we determinate a optimal BOVW model firstly; then we compare BOVW and CNN. Finally, section 6 concludes this paper and gives a brief outlook on future work.

2. THE CNN MODEL

Convolutional Neural Network (CNN) is a deep learning model suited for analyzing visual imagery. CNN processes images using matrixes of weights called filters (features) that detect specific attributes such as vertical edges, horizontal edges, etc. Moreover, as the image progresses through each layer, the filters are able to recognize more complex attributes. A CNN model is mainly structured by four components: convolution layer, pooling layer and fully connected layer.

Convolutional computation is actually a filter. The purpose of convolution layer is to receive a feature map. Feature detection is based on 'scanning' the input with the filter of a given size and applying matrix computations in order to derive a feature map. In classic CNN models, Rectified Linear Unit (ReLU) activation function is used after convolution that returns 0 for every negative value in the input image while it returns the same value for every positive value.

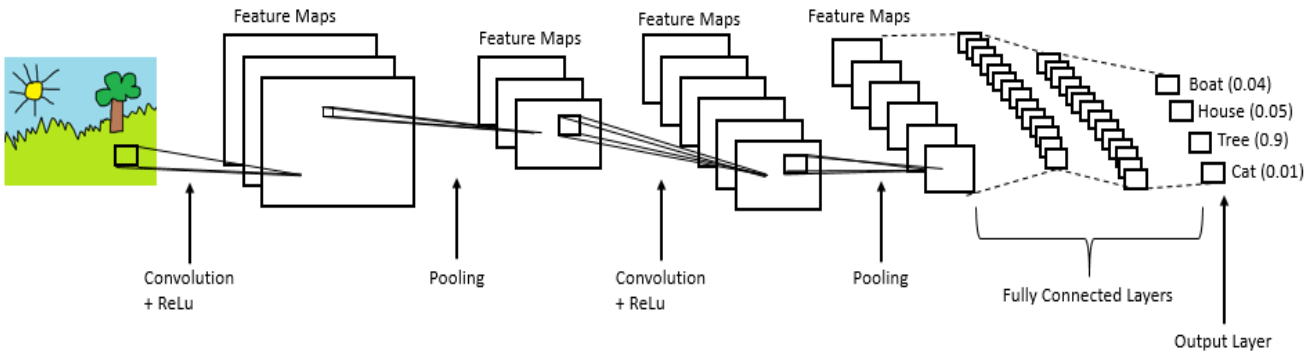

Figure 1: A classic model of convolutional neural network

The goal of this layer is to provide spatial variance, which simply means that the system will be capable of recognizing an object as an object even when its appearance varies in some way. Pooling layer will perform a downsampling operation along the spatial dimensions (width, height), resulting in output such as [16x16x12] for pooling_size=(2, 2).

In a fully connected layer, we flatten the output of the last convolution layer and connect every node of the current layer with the other node of the next layer. Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks and work in a similar way [1].

CNN is a powerful imagery classifier. Its accuracy on Dog & Cat problem is higher than 85% []. However, CNN has many disadvantages. Firstly, it requires large quantity of training samples. In my model, it requires more than 25,000 training images to achieve 85% accuracy. It also need long training time and huge memory expense because of its complex network structure and matrix computation. To lower the training time, CNN might require graphic card computation. In the next section, I introduce a more efficient and resource-saving classification method that is BOVW.

3. THE BOVW MODEL

The Bag-of-Visual-Words (BOVW) approach extends an idea from text retrieval to visual classification [2]. Similarly, an image can be described as a frequency distribution of visual words, independent of their spatial position in the image plane. By vector quantization of these features a discrete vocabulary is created. Local features from novel images are assigned to the closest word in the vocabulary and by counting the number of local features per vocabulary word a BOVW vector is extracted per image. Then we can send all the vectors of training and testing images to a classifier.

I extract local features from several images using Scale Invariant Feature Transform (SIFT) that aggregates 8 gradient orientations at each of 4*4 patches surrounding the sampling point to a $4*4*8 = 128$ dimensional feature vector.

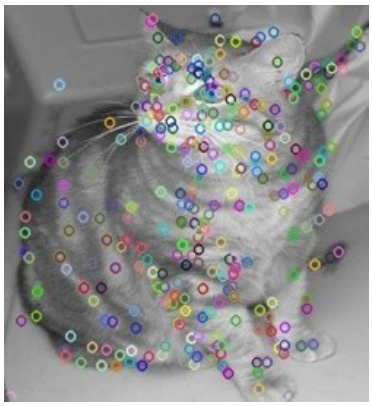

Figure 2: SIFT points of a cat image

Then I quantize the feature space. Make this operation via clustering algorithms such as K-means. K-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The center points, that we get from the clustering algorithm, are our visual words.

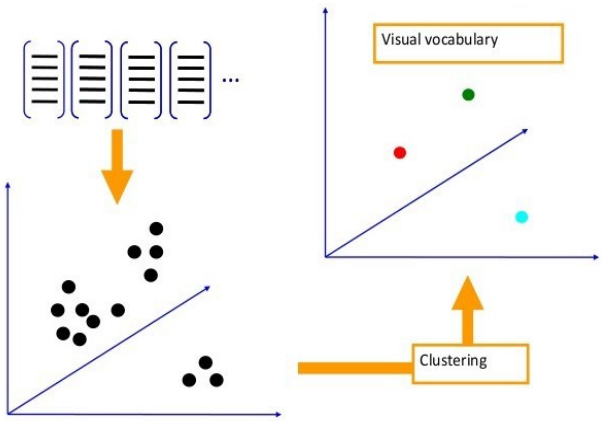

Figure 3: Find visual words using K-means

Next I extract local features and compare these features with visual words to create histograms for each image both for the testing and training dataset.

At last we can send both training and testing histograms into a classifier. There are multiple classi-

fiers that can be applied in BOVW. In the next section, I will briefly describe these classifiers.

4. BOVW CLASSIFIERS

Since raw images have been quantized as vectors by BOWL. Many general machine-learning classifiers can be applied in the last step of BOWL. Following approaches shows efficiency and good performance on binary classification. Cat or Dog problem is a binary classification.

4.1 K Nearest Neighbors

K Nearest Neighbors classification is an example of instance-based learning [3]: instead of attempting to construct an internal model it simply stores instances of the training data (i.e. the BOVW vectors of all training images). The idea behind nearest neighbor methods is to retrieve the *k* training images closest in distance to a new image and predict the label from these training examples based on computation of a simple majority vote. In other words, the category of an image is set to the category that has the most representatives among the k nearest training images. The distance metric used can be any metric measure, however, standard Euclidean distance is the most common choice.

4.2 Logistic Regression

The general assumption behind logistic regression is that the probability of a category label y_p being assigned to an image represented by its BOVW vector x can be written as a logistic sigmoid acting on a linear function of x so that:

$$p(y_p|x)=\sigma(w^T x) \quad (1)$$

with $p(y_n|x)=1-p(y_p|x)$ and $\sigma()$ is the logistic sigmoid function.

4.3 Support Vector Machines (SVM)

SVM is a classifying model that finds hyperplane that maximizes the margin between the positive and negative examples [4]. The choice of the kernel function is crucial for good classification results.

In the beginning, we have linear kernels:

$$K_{linear}(x,y)=x^T y \quad (2)$$

Polynomial kernel function:

$$K_{poly}=(\sum_{i=1}^{n} x_i y_i)^2 \quad (3)$$

And non-linear kernel functions like rbf (radial basis function) which requires additional computation:

$$K_{rbf}=\exp(-\frac{1}{\gamma}\|x-y\|^2) \quad (4)$$

I will evaluate these classifiers in the next section .

5. EXPERIMENTS

At first, I unify the format of input images and resize the row images in 250*250 pixels. Preprocessing is necessary to decrease noises during classifying.

Before comparing BOVW and CNN methods, I must determine the BOVW model (i.e., how many visual words and which kind of classifier to use).

5.1 Determinate BOVW Model

Visual words, that are the resulting clusters of K-means, describe a picture. The image could be fully described with small number of visual words. However, too many words would cause overfitting while classifying. Thus I test different vocabulary sizes with 1,000 training images, 1,000 testing images and SVW-rbf as the classifier. Table 1 shows that around 100 visual words get optimal accuracy.

Vocabulary Size	50	100	300	1000	2000
Accuracy (%)	62.2	69.5	68.1	67.2	66.0

Table 1: Evaluation of the size of vocabulary

Next I test different classifiers with 1,000 training images, 1,000 testing images and 100 visual words. Table 2 shows SVM-rbf is the optimal classifier on Cat or Dog problem. This paper [5] also

states that SVM-rbf works well on image classification.

Classifier	KNN	SVM-linear	SVM-poly	SVM-rbf	Logistic Regression
Accuracy (%)	63.4	68.2	68.4	69.5	66.0

Table 2: Evaluation of BOVW classifiers

Finally, I determinate the BOVW model with 100 Visual words and SVM-rbf classifier.

5.2 BOVW vs. CNN

I build the CNN model by repeating convolution and pooling layers three times, then fully connect the feature map. Table 3 shows the comparing results between BOVW and CNN. In the next section I will evaluate these two methods.

6. CONCLUSION & FUTURE WORK

Table 3 demonstrates that BOVW performances better than CNN with small training sample size. Suppose there are only 500 training images, BOVW could achieve 6% more accuracy than CNN! Thus, BOVW is a more efficient tool than CNN.

With 1,000 training images, BOVW takes 11 minutes to training the model, while CNN takes more than 40 minutes (on the same laptop with Intel i7 processors and neither uses graphic card). Thus BOVW is more resources saving than CNN.

However, there also leaving a fatal flaw of BOVW: the accuracy is not high. Its accuracy vibrates at around 70% and is hard be improved by enlarging the training data.

To improve its accuracy, I plan to add another classifier to BOVW. It is a shape-fitting model that is an average shape model for each class by edge detecting. Each testing image should fit the model before goes to BOVW. And the final result the sum of weighted results of these two classifiers. This is the future work of my project.

7. REFERENCES

[1] S. Lawrence, C. Gilesand, A. Tsoi and A. Back. Face Recognition: A Convolutional Neural-Network Approach. In *IEEE Transactions on Neural Networks*, Vol 8, page 98-112. IEEE,1997

[2] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, number Iccv, pages 1470–1477. IEEE, 2003.

[3] J. Bentley. Multidimensional binary search trees used for associative searching, 1975.

[4] C. Cortes and V. Vapnik. Support Vector Machines in *Machine Learning*. pages 273–297. 1995.

[5] C. Hentschel and H. Sack. Does one size really fit all? Evaluating classifiers in Bag-of-Visual-Words classification. In *i-KNOW'14*, September, pages 16-19. ACM, 2014

Size of training data	500	1000	2000	3000	4000	20000
BOVW Accuracy (%)	65.7	69.5	71.6	70.9	72.9	72.5
CNN Accuracy (%)	59.7	62.2	67.0	67.9	70.4	85.8

Table 3: Comparison between BOVW and CNN