Final Project: Dog and Cat Classification

Shengbin Wu

05/01/2017

1 Introduction

In 2007, Microsoft present Asirra, A CAPTCHA(Completely Automated Public Turing test to tell Computers and Humans Apart) that distinguish human and computer by asking users to identify cats from dogs in an image set, *Figure 1*. Asirra was based on the assumption that machines can not classify these images automatically, while human can successfully accomplish it within a short time.[1] The easiest way

Figure 1: Asirra

of attacking this problem is by random guessing, image recognition technique provide better guesses than by chance.The reason why it is difficult to classify these images is the diversity of the database, including the background, position, angle, pose and lighting. Several years ago, experts in an information poll point out that a classifier with accuracy higher than 60% is hard to obtain without major advance in computer vision. A 60% classifier improves the guessing probability of a 12-image HIP(Human Interactive Proof) from 1/4096 to 1/459.[2]

With the development of computer vision and deep learning, Asirra is no longer considered a safe system from attack. Pattern classification technique can be adopted to attach Arissa with considerable accuracy. In the project, a classifier of SVM(Support Vector Machine) and a classifier of CNN(Convolutional Neural Network)will be explored to experiment the accuracy of automatically classifying the image set of cats and dogs.

2 Background

Philippe Golle(2008) from Palo Alto Research Center attack this problem by providing a classifier with hgiher than 80% accuracy. The classifier Golle provided is combined by two SVM classifiers, one was trained on color features of the images, the other was trained on texture features of the images. The classifier is supposed to solved this problem with 10.2% probability. The classifier demonstrated the accuracy of 82.7% accuracy by using 15,760 color features and and 5,000 texture features.
In 2012, Omkar M Parkhil,et al solve this problem by training their classifier with only the shape features which were extracted by a Discriminatively Trained Part-Based Mode s[3] and got an accuracy of 92.9%. In their paper, they can not only classify which family(Dog vs Cat) an image belongs to, but also can classify which breed it belongs to.[4]
In 2014, Bang Liu, Yan Liu et al use SVM(Support Vector Machine) to solve a related problem in a Kaggle competition. They used features extracted by SIFT(Scale Invariance Feature Extraction) to train their classifier but only got an accuracy of 67.6%.However, they got an accuracy of 94% by training their SVM classifier with the features leaned by CNN(Convolutional Neural Network). [5]

3 Method

3.1 Algorithms

1. **Support Vector Machine**
   Support Vector Machine(SVM) is a discriminative classifier formally defined by a separating hyperplane, *Figure 2*. SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. In other words, SVM learns the hyperplane that maximize the distance of both nearest points.SVM is applied to maximize the margin of the training data, but also applied as a learning algorithm to prioritize the accuracy of classification. An SVM model is a non-probability binary linear classifier because it assigns new examples to one category or the other.

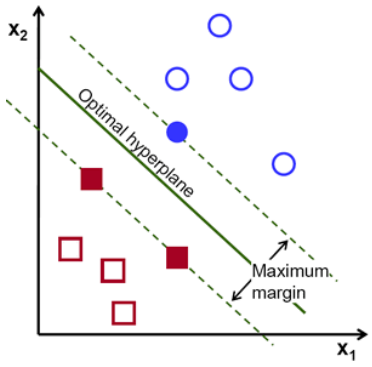

Figure 2: SVM

2. **Bag of Word**
   The Bag of Word(BOW) Model represent an object as a bag of visual words, which is one of the most common method for object representation. This method is base on the idea that extracted

the key points from image and then quantize them into one of the visual words, *Figure 3*. The key points are "interesting" local patches which can be extracted by densely sampling or by intrest point detector. Then an image is represented by the histogram of the visual words, *Figure 4*. For this purpose, the visual words(dictionary) are generated by a clustering algorithm, generally the K means algorithm.[6]

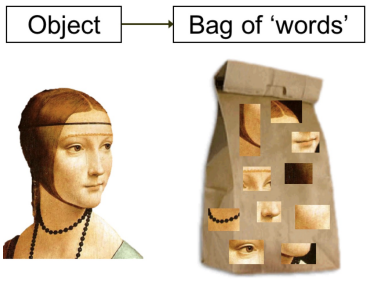

Figure 3: Bag of Words

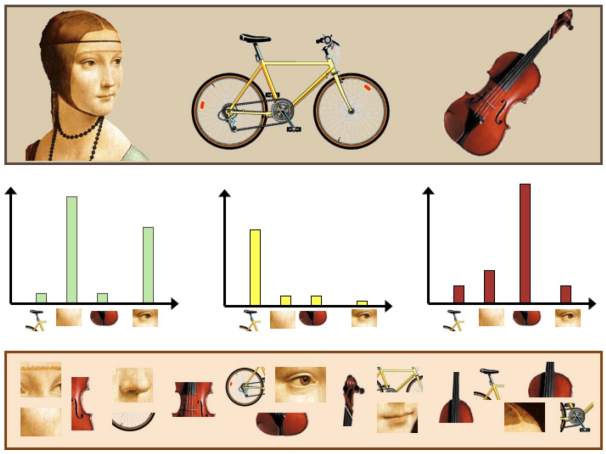

Figure 4: Object Representated by BOW

3. **Convolutional Neural Network**

   A Convolutional Neural Network (CNN) is comprised of one or more convolutional layers (often with a subsampling step) and then followed by one or more fully connected layers as in a standard multilayer neural network, *Figure 5*. Compared with traditional neural network, convolutional neural network contains not only input layer, hidden layer and output layer, but also contains convolutional layers and pooling layers. The architecture of a CNN is designed to take advantage of the 2D structure of an input image (or other 2D input such as a speech signal). [7]

   The convolutional layers are consist of feature maps, which are 2 dimensional weighted matrix nodes used to detect features in the images. Different features maps are different kernels which are used to detect different features. Once a feature is found, the exact location is not as important as its rough location relative to other features. For image recognition, there are always more than one features map in a single layer.

   The pooling layers are usually used after the convolutional layers. The function of pooling layers is to simplify the information in the output from the convolutional layer. To be specific, the pooling

layer takes the output from each feature map in the convolutional layer and do sub sampling to get a more condense feature map.

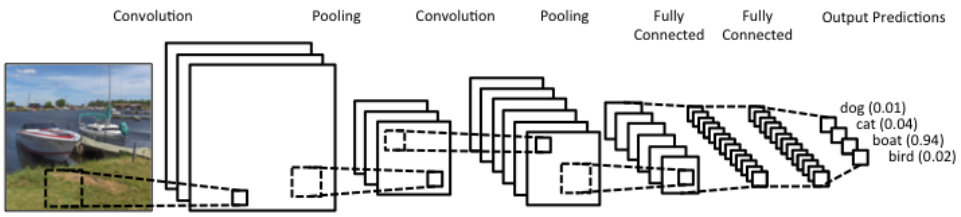

Figure 5: A simple CNN

3.2 Method Selected

1. Method One

   For the first method, we train a SVM classifier to classify the images by families(family dog or family cat). SIFT will be applied to extracted the features from each image in the training set. Then the bag of word model will be applied to create a dictionary for these extracted features. This dictionary is formed by a clustering algorithm K-means. One cluster of features is viewed as a visual word in this dictionary. After the dictionary is created, images are represented by frequency vectors which represent the proportion of features belong to a visual word. Then, a SVM classifier is trained based on these frequency features.

2. Method Two

   For the second method, we used CNN model to extracted features from the images. For this project, keras application are used. These applications are deep learning models with pre-trained weights. They can be used for prediction, feature extraction and fine tuning. The ResNet50, InceptionV3 and Xception model are used for extracting features from the images. Then these features will be used to trained a CNN model.

4 Experiment

4.1 Data Set

The data set for this project is the provided by Microsoft Research and Kaggle. The training set consists of 25,000 images with half cat images and half dog images. The training set contains 12,500 images without labels. The size of these images are about 350×350. After downloading the image set from Kaggle, the images are separated into two folder, one for cat images, one for dog images. For dogs, the corresponding label is 1, for cats, the corresponding label is 0. The sample images in the data set are shown in the *Figure 6*.

4.2 Method One

- **Process**

  For traditional image classification problems, features extracted by human are chosen to train clas-

Figure 6: Images in the data set

sifiers. Then feature descriptors are use to represent the images. SIFT, HoG, RGB and HSV are the common features that are used to represent images. For this project, SIFT are used to extracted the features and compute the feature descriptors. Because we know that the shapes of images of cat and dog are different with each other. Features extracted by SIFT will play an important role in the classification of the image set.

After implementing the SIFT algorithms to extract features, we get the features descriptors of all the images in the training set. K-means clustering algorithms is applied to generate a dictionary of visual words for the features descriptors in the training set. All the images are then represented by by frequency vectors which represent the proportion of features belong to a visual word. Based on the frequency vectors generated by the BOW method, a SVM classifier was trained to make classification. The visualize process of this model is shown in the *Figure 7*.

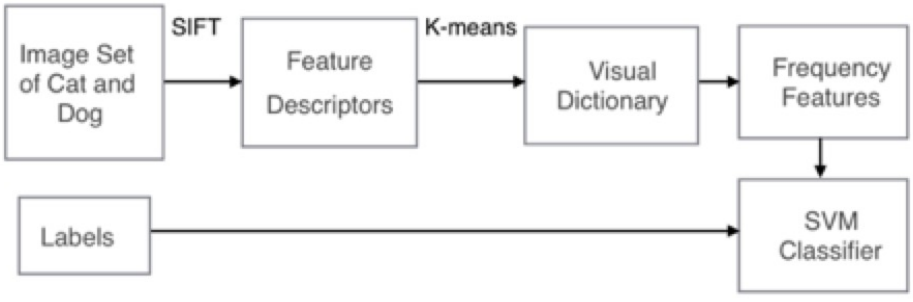

Figure 7: Process of Method One

In method one, 20,000 images with half cats and half dog in the training set are used to train the SVM model, while the remain 5,000 images are used for validation.

Source code for this method: traditional pattern classification implementation

- **Result and Analysis**

Table 1: Accuracy after Feature Reduction

Feature Reduction Method	Principle Component Analysis	Recursive Feature Elimination
Accuracy Score	65.50%	65.50%

For this method, it took about 16 minutes to apply SIFT to get the feature descriptors from the training set. When K means algorithm was conducted to generate the visual words dictionary, it took far more longer to get the dictionary even the size of images are compressed to 256×256. It takes 45 minutes to generate the dictionary by K means.

As the training set contains 20,000 images, it is very time consuming to use method one to make the classification. K fold validation is not used when doing validation. For this method, we only validate the result once in the validation set.

The accuracy of this method is about 62% when the images were compressed to 128×128, which is quite dissatisfactory for a binary classification problem. While when the images were compressed to 256×256, the accuracy only increase to 65.46%.

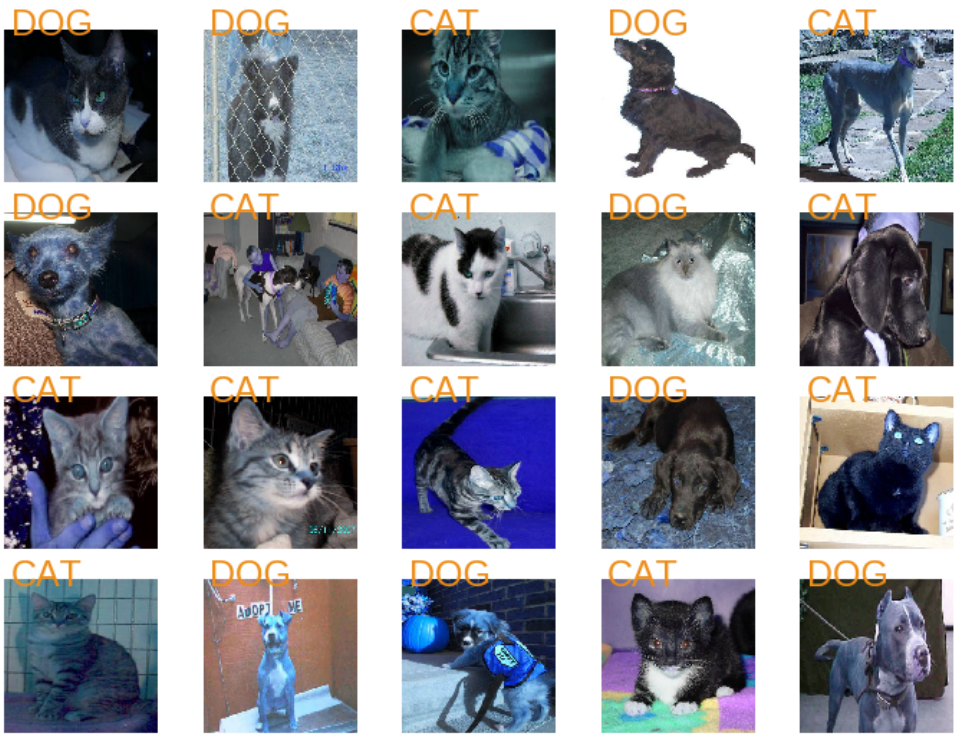

Figure 8: Results Sample of Method One

After the first attempt of classification, feature reduction method, Principle Component Analysis and Recursive Feature Elimination were applied. The accuracy score are shown in *table 1*. It seems that the accuracy do not improve by applying feature reduction.

The most important part of this method is the feature extraction process. The result of poor performance may be caused by features extracted from the images, as the explaination below.

- First, the background of this images are diverse and complex, while image segmentation was not applied to segment target animals from their backgrounds, because segmenting the target animals from their diverse backgrounds automatically is a considerable difficult problem.

The features obtained from the background are only useless noise and they even can perform negative influence on this classification problem.

- Second, the images was compressed, which will lost a lot of shape information from the original image. The features in a low resolution image will have less features than a high resolution one. In addition, the feature matching of the images will also be affected.
- Third, the shape features of cat and dogs are similar, they both have one head, one tail, two ears and four legs, and most of these features are similar. Thus, the features extracted may not contribute much to make the classification.

4.3 Method Two

- **Process**

  In Method Two, image features will be extracted by a Convolutional Neural Network model. After that, we can simply use dropout to classify the validation set and test set. Compared with the feature extraction by SIFT, convolutional neural network learn features from the images.

  Several pre-trained model in keras application are used in this method to learn the features from the cat and dog image set. The ResNet50, InceptionV3 and Xception models are chosen to learn the features,which are object detection model in image recognition provided by keras. The weights of these models are pre-trained on ImageNet.

  In order to improve the performance of the classification model, these three models are used together to learn the features from the images. These three models build up a huge network. If a fully connected layer is added directly after this huge network to train the classification model, the computation cost will be extremely large. Thus, the features extraction and classifier training are conducted separately. The pre-trained models are used to extract features. And then, these features are used to train the classifier.[8]

  For the trainning process, a simple neural network was used as the classification model, this model includes an input layer, a hidden layer with dropout rate 0.5, and an output layer with sigmoid as activation function. The features number learned by each model is 2048, and then 6144 features was learned from the feature extraction process. The number of nodes in the input layer is 6144. For the hidden layer, there is also 6144 nodes, but they are not fully connected because dropout was applied in this layer. For the output layer, there is only 1 node because it is a binary classification problem.

  With fine features learning by pre-trained models, a simple model can make a good classification. The visualized process of this model is shown is *Figure 9*.

  Source code for this method: convolutional neural network implementation

- **Result and Analysis**

  When the classifier being trained, 20,000 images were used to train the model, while 5,000 images were used to validate the model. The accuracy in the validation set is as high as 99.3%, which is far more higher than the accuracy in method one. The result of cross validation for this method is shown is *Figure 9*. In addition, the cost time of feature extraction for each model(ResNet50, InceptionV3 and Xception) is less than 20 minutes, which is also more efficient than the method

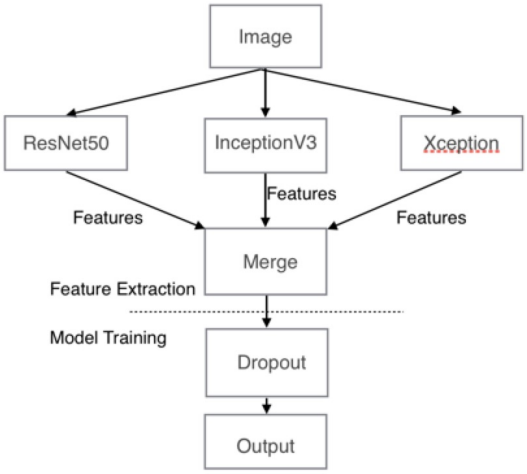

Figure 9: Model Visualization

one. Moreover, it only take less than 10 minutes to train the classifier.

```
In [9]: model.fit(X_train, y_train, batch_size=128, nb_epoch=8, validation_split=0.2)

        Train on 20000 samples, validate on 5000 samples
        Epoch 1/8
        20000/20000 [==============================] - 2s - loss: 0.1067 - acc: 0.9664 - val_loss: 0.0310
         - val_acc: 0.9916
        Epoch 2/8
        20000/20000 [==============================] - 1s - loss: 0.0297 - acc: 0.9909 - val_loss: 0.0242
         - val_acc: 0.9922
        Epoch 3/8
        20000/20000 [==============================] - 1s - loss: 0.0223 - acc: 0.9925 - val_loss: 0.0229
         - val_acc: 0.9924
        Epoch 4/8
        20000/20000 [==============================] - 0s - loss: 0.0193 - acc: 0.9940 - val_loss: 0.0240
         - val_acc: 0.9920
        Epoch 5/8
        20000/20000 [==============================] - 1s - loss: 0.0174 - acc: 0.9937 - val_loss: 0.0232
         - val_acc: 0.9922
        Epoch 6/8
        20000/20000 [==============================] - 1s - loss: 0.0159 - acc: 0.9938 - val_loss: 0.0224
         - val_acc: 0.9928
        Epoch 7/8
        20000/20000 [==============================] - 1s - loss: 0.0151 - acc: 0.9947 - val_loss: 0.0230
         - val_acc: 0.9924
        Epoch 8/8
        20000/20000 [==============================] - 0s - loss: 0.0138 - acc: 0.9949 - val_loss: 0.0221
         - val_acc: 0.9932
Out[9]: <keras.callbacks.History at 0x7f5cd825b850>
```

Figure 10: Result of Method two in Validation Set

5 Conclusion

5.1 conclusion

For the project, two methods were explored to classify the images of cats and dogs.

The first method is the traditional pattern recognition and classification method, which use computer vision technique, including SIFT and BOW method, to extract image features. And then these features were used to train a SVM model for making classification. In this method, feature reduction was also applied to eliminate useless feature. However, the best accuracy score in this method is only 65%.

In the second method, CNN models were explored to learn the features from the images in the training set. In order to improve the accuracy of classification, three models were used together to learn the features.

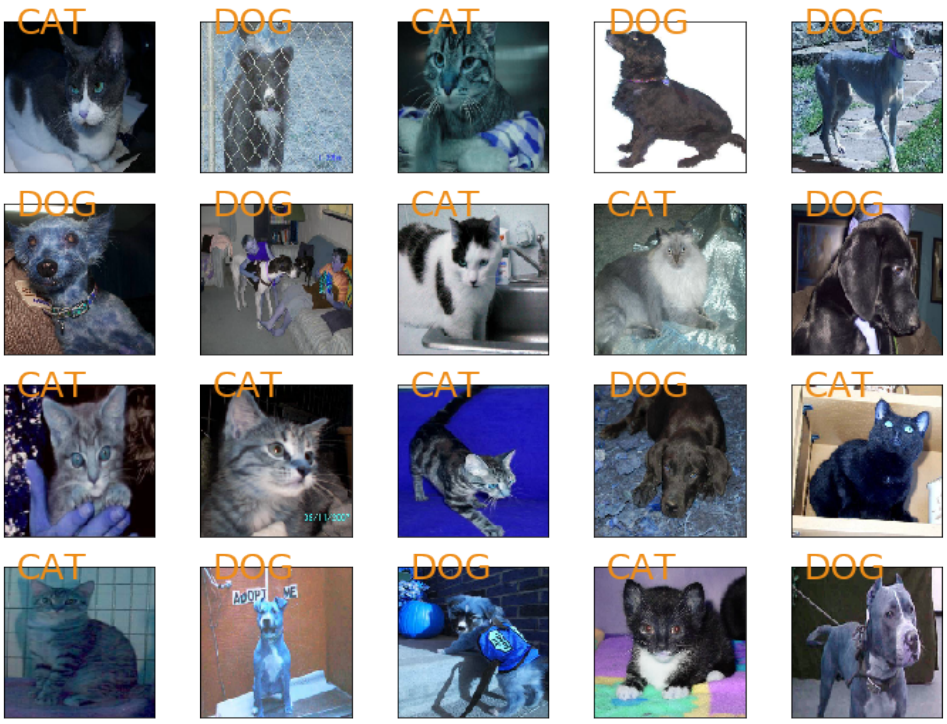

Figure 11: Results Sample of Method Two

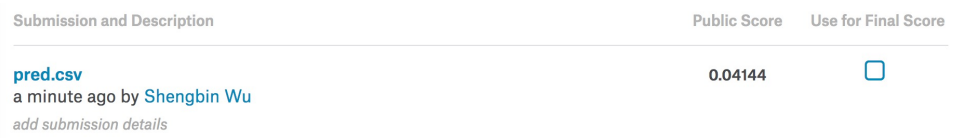

Figure 12: Score in Kaggle

Then, these features were used in to dropout neural network to train a model for classification. Fot these method, the accuracy is far more higher than the first method.

It is obvious that the deep convolutional neural network method is better than traditional pattern classification in this problem. First, extracting features from images by traditional computer vision method require a lot of work, including image segmentation, feature detection, feature clustering, and visual word generation, which are all considerable time consuming. Image segmentation in this problem is especially difficult because the backgrounds in the images are quit complex. That is the reason why image segmentation is not conducted in the method one. On the other hand, for the convolutional method, we can use pre-trained model to detect object in the images, without considering the background and location of the features, which is more efficient than method one.

5.2 Future Work

In the future, image segmentation should be applied to improve the accuracy of method one. In addition, color features and texture feature should also be taken into consideration to generate more powerful feature in the feature extraction process. Similar, image segmentation may be used for method two to eliminate useless information in the background. And other model and parameters should also be explored to try to get a better performance.

Acknowledgement

The convolutional network method implementation in python is refer to Peiwen Yang's post in a China online community *Zhihu*.

References

[1] Philippe Golle. Machine learning attacks against the asirra captcha. In *Proceedings of the 15th ACM conference on Computer and communications security*, pages 535–542. ACM, 2008.

[2] Jeremy Elson, John R Douceur, Jon Howell, and Jared Saul. Asirra: a captcha that exploits interest-aligned manual image categorization. In *ACM Conference on Computer and Communications Security*, volume 7, pages 366–374. Citeseer, 2007.

[3] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[4] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3498–3505. IEEE, 2012.

[5] Bang Liu, Yan Liu, and Kai Zhou. Image classification for dogs and cats.

[6] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.

[7] Michael A Nielsen. Neural networks and deep learning. *URL: http://neuralnetworksanddeeplearning. com/.(visited: 01.11. 2014)*, 2015.

[8] Francois Chollet. Building powerful image classification models using very little data. https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html, 2016. [Online Tutorial].