|  |
| --- |
| **CS5785 Final Project Report: Image Classification with Ensemble Classifier** |

**Anonymous Authors**

Cornell Tech

111 8th Ave. #302, New York, NY

**Abstract**

We implemented different classification methods including: support vector machine, logistic regression and neural network and applied a voting ensemble to classify 3,000 images into 200 classes. We achieved an accuracy of 54% (top 5) on the kaggle competition leaderboard.

**1 Introduction**

Image recognition has been an area of intense study in machine learning. Currently the most successful model on this task is deep convolutional neural network. Trained on 1.2 million images, AlexNet achieved a prediction accuracy of 62.5% for a 1,000 class classification problem [1].

In the CS5785 final project, we are provided with three different feature sets: pre-trained features from convolutional neural network, handed-tuned features from bag-of-word SIFT descriptors and binary attributes, and we are expected to use the methods we learned in class to come up with a best classifier for this problem.

Image recognition are usually trained with millions of images such as ImageNet. One of the major challenges in our project is the scarcity of data. 3,000 training data with 200 classes left us with 15 positive training data for each class. A second challenge would be “the curse of dimensionality”. The provided feature sets have dimensions of: 4,096, 4,096 and 102. Even we use only one feature set, the number of dimensions still exceeds the number of training examples. This will make the training data even sparser. A third challenge is how to combine the three different data sets to improve performance.

In our ensemble model, we approached these challenges by: training a separate classifier for each feature set, make use of unlabeled data, and use voting ensemble to improve the performance.

**2 The Dataset**

**2.1 Labeled and unlabeled data sets**

3k labeled training data set: the training set for the competition includes 3,000 images belonging to 200 categories. There are exactly 15 images for each category.

10k unlabeled data set: in additional to the labeled training data, we are provided with

10,000 images with no label. However, each of the 10,000 images has five captions.

**2.2 Features**

For training and testing data sets, we are provided with three set of features: 4,096 pre-trained features from convolutional neural network (CNN features); 4,096 hand-tuned features from bag-of-word SIFT descriptors with spatial pyramid (BOW features); 102 binary attributes. For the additional 10k unlabeled data, only CNN features and BOW features are available. Some details of the feature sets are listed below in Table 1.

Table. 1 Features overview

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CNN features** | **BOW features** | **Attributes** |
| **Type** | Continuous | Continuous | Binary |
| **Ranges** | (-25.2 , 23.3) | (0 , 0.67) | (0 , 1) |
| **Dimension** | 4,096 | 4,096 | 102 |

CNN features: Since CNN features are pre-trained using convolutional neural network (AlexNet), the features of CNN should represent an abstraction of some patches of the original image. We tried to visually interpret the CNN features but we lack the information to make meaningful inference from the feature itself.

BOW features: as described in the competition, BOW features are hand-tuned features from bag-of-word Scale Invariant Feature Transform (SIFT) descriptors in a spatial pyramid. SIFT descriptors are basically features that trained to detect certain objects in an image, and these features are presented in a bag-of-word fashion.

Attribute features: the attribute features are binary and there are 102 different attributes. Each image has one or more attributes in it.

**3 Methods**

**3.1 Feature Preprocessing**

Since all features are in high dimensional space, we looked at the linear dependence of the feature spaces to look for possibility of dimension reduction. We generated screeplot (principal component vs. variance explained) for each of the feature set of the training data. Figure 1 shows the comparison:

**3.2 Training with SVM**

**3.3 Voting Ensemble**

**4 Results**

**5 Discussion**