Kaggle Challenge

By
Dibyendu Roy Chaudhuri
MT19034

Kaggle Rank: 1 **Public score:** 0.48 **Private score:** 0.47428

Final Features

I have read few CBIR(Content-Based Image Retrieval) papers and found that Color Structure Descriptor (CSD) (64 features), Scalable Color Descriptor (SCD) (128 features), Color Layout Descriptor (CLD) (120 features), Dominant Color Descriptor (DCD) (total 58 features out of which 9 is taken), Homogeneous Texture Descriptor (HTD) (62 features), Edge Histogram Descriptor (EHD) (80 features) are goods features to classify an image. Those features are also known as MPEG features. So first I have extracted these features as a pre-processing step. The number of features in the final model is 454/463(with DCD). These features are very good as it contains shape, colour, texture and edges information of an image.

Final Classifier

I have used mainly two classifier- random forest and perceptron. I have also tried with SVM but SVM did not work well with my image features. Later I used bagging with random forest or perceptron which actually gives me 9-10 % more validation accuracy than using it directly. Maximum validation accuracy, I have got using RF with bagging is 42% and 49% in the case of the perceptron. I have used a maximum of 3 hidden layers(150, 200, 100) in the perceptron model. While validation, training data is splitted into 80:20 ratio. I have used smote for oversampling which gave me 1.5-2 % extra validation score.

Drive link: https://drive.google.com/open?id=1_pthJrtCIzJZuiPuLkOF1D33vt0seGK9

Execution Instruction

1. Please download the “Dataset_Kaggle” file first. Train and test features are stored in this pickle file. There is no need of “SML_Train.csv” file as tag information is already associated with train features in “Dataset_Kaggle” file.
2. Please download the model weight, associated with code.
3. You can re-train the model or generate the output file using pre-stored weight.
4. Generally, It takes around 1-1.5 hours to train the model depending upon the code.
5. You also comment on the validation part to save your time. It will save 30 min from total execution time.

Experiments & Challenges

1. At first, I did not extract any features and just convert the image into grayscale and run the classifier on it. I have got around 22% using SVM. But other classifier did not give that much accuracy. I thought RBF kernel makes that data linearly separable in higher space. That’s why accuracy is high.
2. Then I applied PCA of 150 features. Accuracy increased slightly.
3. Then I use a combination of hog and histogram features to train the classifier. But SVM classifier’s accuracy dropped but other classifiers accuracy increased.
4. I also tried to classify using SIFT and LBP features. Although I did not get any exciting improvement and dropped the plan.
5. I also tried using histogram but accuracy decreased to 23% using SVM.
6. I also tried to inject some noise into the images using gaussian. Accuracy increased by 2 %. I thought that it prevents to overfit the classifiers. But up to then, I did not get validation accuracy more than 26 %.
7. I have read some CBIR paper and came to know that CSD, SCD, CLD, DCD, HTD, EHD are good features for classification. But SVM did not work well using those features. But I have got 28% using Random forest and 30 % using the perceptron model.
8. Firstly, I did not use HTD as an image’s size has to be 128*128 to generate HTD features. Using RF-based bagging classifier, I got 38% validation accuracy.
9. Then I tried the perceptron model using the same features and got validation accuracy of around 45%.

10. Finally, I converted every image into 128*128 and generate HTD features and combine those with features space. I also did oversampling using smote. I got the validation accuracy of 47.5-49%.
11. As I am using bagging so due to bootstrap dataset, accuracy for both validation and final is differing with times. So lastly I interference from validation, K-fold, also train accuracy that this bootstrap is good enough and submit in the Kaggle.
12. The maximum public score of my model is 0.48
13. My private score is 0.47428

Kaggle Leaderboard

Public score	Private score
0.48	0.47428
0.47555	0.46761
0.46222	0.46761
0.44222	0.45714
0.40000	0.39714