CSED703B Vision and Language Assignment 2

Gyunam Park
20172050
IME, POSTECH
gnpark@postech.ac.kr

1. Introduction

PASCAL Visual Object Classes Challenge (PASCAL VOC) is one of the most popular competitions researchers in computer vision are thriving to outperform others with breakthrough ideas. The goal of this challenge is to recognize objects in real-world scenes. Datasets are provided with the labels so that we can apply supervised learning methods to achieve goal. There are twenty objects including Person, Animal(e.g. bird, cat, etc), Vehicle(e.g. aeroplane, bicycle, and so on), and Indoor(e.g. bottle, chair, dining table et al.) There are mainly three kinds of competitions such as classification, detection, and segmentation. In this assignment, we focus on classification where we predict presence and absence of objects(i.e. 20 classes).

2. Dataset

A basic classification problem aims to classify single object from an image. On the other hand, in PASCAL VOC, images with multiple objects are provided with corresponding labels, which makes the problem much trickier. We can use training data and validation data as building a classification model. With validation data, we can evaluate our model before actually testing our model with test data and improve it accordingly. After completing modeling, we can evaluate our model with test data, aided by software written in MATLAB. For modeling purpose, 9,963 images are offered, 50% as training data and 50% as validation data.

3. Transfer Learning

Convolution Neural Network commonly used for visual recognition require huge amount of training data and naturally resources to process it. For instance, ImageNet ILSVRC model utilized 1.2 million images over multiple GPUs. In the context, transfer learning is conceived to transfer the weights of trained model to a new model. Using transfer learning is normal for researchers who achieve high performance in visual tasks. Razavian et al(2014) showed that deploying the weights learned in ImageNet ILSVRC achieves near state-of-the-art performance in various computer vision problems.

In this assignment, I used Inception V3 for transfer learning. Inception is CNN variant which is popular for winning in ImageNet 2014. Conventional CNN models applied same-sized convolution filter in sliding manner. On the other hand, Inception utilized various filters with different sizes in order to capture characteristics of images more efficiently. Inception V3 is upgraded version of it and introduced smaller sized convolution filters. Based on this powerful image-processing model, we fine-tuned the weights of it via backpropagation and replaced and retrained the last part of ConvNet(i.e. classifier).

4. Model

As mentioned before, Inception V3 is adopted as base model for transfer learning. The final fully connected layer of Inception V3 is excluded for customizing classifer for our PASCAL VOC classification task. Alternatively, new fully connected layer with 'ReLU' activation function and sigmoid layer with number of nodes equal to number of classes are employed. The sigmoid layer is utilized instead of softmax layer, which is commonly used for categorical classification tasks, because our goal in this task is to classify multi-label images. If we use the softmax layer as usual, we ended up with poor performance because the results are calculated to be summed into 1, which is undesirable in our case.

In addition to replacing the final layer of existing Inception V3, we also fine-tuned the model by updating the weights according to our dataset. For training, we adopted Adam optimizer with learning rate 0.001 and sets the batch size to 32. For every epoch, the validation loss and accuracy was recorded and the model was saved only if there were improvements between epochs. The optimal number of epochs we examined was 10.

Data augmentation is an effective method to increase the dataset size and regularize the model for achieving generalizability. The basic idea of data augmentation is to transform the dataset. There are several typical transformations and I adopted some of them to improve classification performance. These are what I used in this task: rotation, width shift, height shift, shearing, zooming, fliping.

5. Result

For every class, we generated prediction score, which is called confidence here. It is used to compute precision, recall, and average precision. Table shows the average precision of each class.

Table 1. Evaluation Result

Category	Average Precision
aeroplane	0.793
bicycle	0.790
bird	0.778
boat	0.770
bottle	0.412
bus	0.659
car	0.829
cat	0.840
chair	0.490
cow	0.590
dining table	0.450
dog	0.780
horse	0.795
motorbike	0.783
person	0.873
potted plant	0.503
sheep	0.650
sofa	0.539
train	0.870
tv monitor	0.681

Person has the highest average precision of 0.873 and bottle achieved lowest average precision of 0.412. Some categories were classified relatively well with precision value around 0.8.(e.g. aeroplane, bicycle, bird, boat, car, cay, dog, horse, motorbike, person, train). On the other hand, Inner object(i.e. bottle, chair, dining table and potted plant) was difficult to classify with the model I developed with average precision around 0.5.

Figure 1. shows the precision-recall curve of each category. Chair, horse, person, and sofa shows the balanced relationship between person and sofa. On the other hand, bird, car, and cat shows highly unbalanced precision-recall relationship.

6. Discussion

The reason I figured out for the low performance on bottle is that it is presented with other objects frequently. It seems that the model has difficulty of updating the weights when the bottle was present. Another reason for it might be in the process of preparing training data. I inputed a single-labeled image into the model. In other words, same image was offered several times with different labels since the image has several labels.

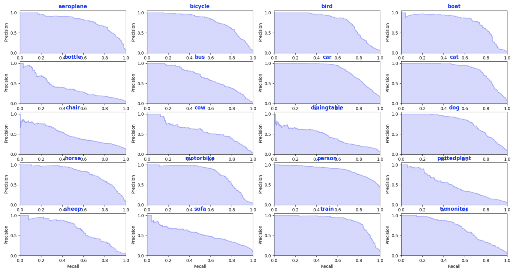

Figure 1. Precision-recall curves of categories

For the future work, I can also apply ensemble learning method to improve performance of classification. Further I can turn the hyper-parameters in various manner. Other techniques such as early-stopping, batch normalization, and dropout can also contribute to the better performance. Finally, for data augmentation, I should figure out which is proper and works better than others.