An improved Faster R-CNN-Based Food Object Detection and Classification System.

NGUYEN Thanh Long, Yutaro IWAMOTO, *Yen-Wei CHEN.
Ritsumeikan University, Shiga, Japan.
*Corresponding author: chen@is.ritsumei.ac.jp

Abstract—The rise in peoples consciousness towards their daily eating habits has attracted significant attention from the field of automatic food analysis. Specifying in Japanese daily food items, which may become helpful for almost all people at every age in society. In this paper, we propose a mobile application that can take the picture of the meal of the user as an input, and output the corresponding categories. The system is based on an improved Faster R-CNN including center loss to increase the classification accuracy.

Index Terms—Convolutional neural networks(CNNs), Region Proposal Networks(RPN), center loss, detection.

I. INTRODUCTION

MODERN diseases such as obesity and diabetes are major killer in this era. The two leading causes to these problems are physical inactivity and poor dietary practice. For these reasons, the concern about food consumption has risen dramatically among people. Therefore, the need for an application that can help people keep track of both physical activities and nutrition habits are growing quickly, a field where automatic food analysis plays a significant role. Focusing on daily foods, food recognition algorithm could enable monitoring of daily food consumption for people.

From the aspect of computer vision, a few approaches have been introduced to deal with the issue, most of them using Convolutional Neural Networks(CNNs) [1], [2], [3], [4]. Numerous works consider the improvement of methods for food recognition, which is, being able to recognize the dish described in an image in which a single plate is shown. However, it is crucial to know that food has its fine-grained nature, meaning that especially in the field of food analysis the intra-class variability and inter-class similarity are making the problem of getting robust food recognition methods difficult.

Numerous works have been introduced for application purpose which is food consumption self-monitoring [4], [5], [6], in which the user takes photos of each meal and the system would eventually track any nutritional information associated.

In this work, we develop a mobile system that allows food detection and food recognition by designing an android application along with a server that can take input picture and output the corresponding food categories. The food detection and classification network are built based on and improve Faster R-CNN. The Faster R-CNN [7] is the state-of-the-are detection method. As an improvement of Faster R-CNN, we added a newly developed center loss [8] as an auxiliary loss on the classification layer of Fast R-CNN to improve the discriminative power of the deeply learned features.

II. RELATED WORK

There is a significant interest in undertaking visual food analysis study, particularly in its applicability to monitoring the users diet based on the inherent nutritional information contained in food pictures. In this field, researchers have paid interest in a few aspects related to automatic food analysis. The most basic aspect in this field is the binary food detection where the problem is to determine the presence or absence of food in an image. This problem is also called food/non-food classification or food detection [1]. First, come the combination of a BoF model and an SVM, but soon be outperformed by the introduction of CNNs [1].

In food analysis, once the image containing food are identified, food recognition is often the next step to apply. By using CNNs, they have achieved an outstanding result reaching approximately 90% in the dataset with around 100 different kinds of food [9]. A few food datasets such as UECFOOD-100 [10], UECFOOD-256 [11], Food-101 [12] have also been published to promote future advances in this field.

However, most approaches focused on detecting one food per image, but it is crucial to take to account that most people take pictures of their meal containing many food items in only one picture. That is where the introduction of object detection algorithm comes to the rescue, one of the state-of-the-art methods is Faster R-CNN. In this paper, we introduce an improved Faster R-CNN-based system that can take the food picture of the user and output the corresponding categories through the help of Faster R-CNN and center loss.
III. METHOD

This work proposes a system applying Faster R-CNN for food detection and classification. To increase the score of the classification task, we also applied center loss to improve the discriminative power of the deeply learned features.

A. Faster R-CNN Based Architecture

We developed a system for automatic food object detection and classification. The proposed system is based on an improved Faster R-CNN, in which we applied a center loss to improve the discriminative power of the deeply learned features.

Faster R-CNN framework is currently one of the best object detection approaches in the state of the art. It allows predicting the bounding boxes by using the Region Proposal Network (RPN) module, as the name revealed RPN is a network to propose regions. For instance, after getting the output feature map from a pre-trained model (VGG-16 [13]), if the input image has 600x800x3 dimensions, the output feature map would be 37x50x256 dimensions. After that, the RPN is connected to a Convolutional layer, and the output is connected to two 1x1 convolutional layers for classification and box-regression. At the same time, non-maximum suppression is applied to make sure that there is no overlapping for the proposed regions. Then we go to the second stage of Faster R-CNN, pooling is used for these proposed regions. After that, we flatten this layer with some fully connected layers and finally use a softmax function for classification and linear regression to fix the boxes location. As an improvement of Faster R-CNN, we added a newly developed center loss as an auxiliary loss on the classification layer of Fast R-CNN [14] to improve the discriminative power of the deeply learned features, which will be described in the next section.

B. Loss Function

In adding to the conventional softmax loss, we use a center loss on the classification layer of Fast R-CNN to improve the classification accuracy. We briefly review Center Loss, which is a newly developed loss function and has demonstrated impressive results in face recognition task. The basic idea of center loss is to encourage the network to learn the discriminative features that minimize the intra-class variations while enlarging the inter-class variations.

The formulation of center loss is

\[ L_c = \frac{1}{2} \sum_{i=1}^{m} \| x_i - c_{y_i} \|_2^2 \]  

where \( c_{y_i} \in \mathbb{R}^d \) denotes the \( y_i \)th class center of deep features. The formulation effectively characterizes the intra-class variations. Ideally, the \( c_{y_i} \) should be updated as the deep features changed. Our purpose is to minimize the intra-class variations.

Center loss is supposed to be optimized jointly with the softmax loss. It has been proved that the center loss is very efficient in minimizing the intra-class variations, and the softmax loss has some merits in maximizing the inter-class variations of the learned features. Hence, it is very reasonable to use a combination of center loss and softmax loss to pursuit the discriminative features.

C. Client-server system

We propose a mobile application to visualize the results as well as aiming to produce a complete product for real use. The system is based on a client-server architecture, which is illustrated in Figure 3.

Fig. 3. Mobile application with a client-server architecture

First, the user takes pictures of their meal and upload it to the server which is made using Firebase using android application. The processing server which is made using Python-Django framework then fetch the picture and put it in the Faster R-CNN trained model to predict detect the object and predict the corresponding class. The server then gives back the result to the database, and the android application will fetch the data from the database to visualize to the user.

IV. EXPERIMENT AND RESULT

In this section, we first described the dataset used to evaluate the proposed system, which is consisted of images taken everywhere. Then we will describe the results obtaining from the use of center loss and Faster R-CNN with that of only Faster R-CNN.

A. Dataset

Our dataset is mostly taken from UECFOOD-101 dataset and some of our own, which contains 2103 images for both training and validating. There are 12 classes in our dataset, which are: rice, miso, fries, fried chicken, grilled fish, croquette, omelet, ramen, tofu, spaghetti, tempura, and udon. All images are set to the size of 480x300 in RGB. For experimental purposes, the dataset has been split into a training set and test set: the former contains 1875 images, while the latter contains 228 images. The annotations included in the dataset contain, for each food item: the polygon defining its boundaries, the bounding box, and the food label.
B. Experimental results

As for the experiment setup, our Faster R-CNN was trained on our custom dataset as described above. The models were trained during 10000 iterations, and a learning rate of 1e-3 and batch size of 256, using Resnet-101 [15] based network.

<table>
<thead>
<tr>
<th></th>
<th>rice</th>
<th>miso</th>
<th>filet</th>
<th>grillel</th>
<th>croquettes</th>
<th>ramen</th>
<th>tofu</th>
<th>spaghetti</th>
<th>tempura</th>
<th>udon</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved Faster R-CNN</td>
<td>1.00</td>
<td>0.32</td>
<td>0.90</td>
<td>0.96</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.528</td>
</tr>
<tr>
<td>Faster R-CNN</td>
<td>0.96</td>
<td>0.86</td>
<td>0.82</td>
<td>0.2985</td>
<td>0.8401</td>
<td>0.005</td>
<td>0.005</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.532</td>
</tr>
</tbody>
</table>

TABLE I
Comparing the average precision of each class between using the normal Faster R-CNN[7] and Faster R-CNN (with center loss).

From Table I above, we can see that the center loss has improved the classification result comparing to using normal Faster R-CNN. The overall mean average accuracy (mAP) rises 4.28% from 77.74% to 82.02%.

As for the application, we input the image through the mobile app, the picture will be sent to the server, then the server will use Faster R-CNN output weight to predict the food items from the input picture. The result is shown in Figure 4 below.

![Fig. 4](image_url)

Fig. 4. The result is shown in the mobile application. The left side image is the input of the user, the right side is the result shown in the mobile application.

V. Conclusion

In this paper, we presented a client-server model-based mobile system for automatic food detection and food classification. We proposed an improved Faster R-CNN, in which we added a newly developed center loss to improve the classification accuracy of the normal Faster R-CNN. The overall mAP rises by 4.28% from 77.74% to 82.02%. The system can be used in the form of daily life applications for almost everyone who has mobile devices. In future work, we are going to increase the dataset for real applications.

REFERENCES
