Zero-Annotation Plate Segmentation Using a Food Category Classifier and a Food/Non-Food Classifier

Wataru Shimoda and Keiji Yanai
The University of Electro Communications, Tokyo, Japan

Abstract

To refine a weakly-supervised segmentation method in the food image domain, we propose a novel method to infer plate regions without any pixel-wise annotation. We synthesize segmentation masks for training a plate segmentation model by difference between two types of visualization. In this paper, we train two types of classifiers: a food category classifier and a food/non-food classifier. These two classifiers can highlight food regions by visualization for the classifiers. Our finding is that there is a difference between visualizations of food regions of two types of classifiers and the difference corresponds to strong co-occurence with foods, that is a plate of food. We demonstrate the proposed approach can capture a region of a plate of foods and we also utilize the plate segmentation results for boosting the weakly-supervised food segmentation accuracy.

1. Introduction

In this paper, we propose a novel method to synthesize segmentation masks for food plate areas without pixel-wise annotation and we utilize the plate segmentation results for improvement of the weakly supervised food segmentation. In Fig.1, we show illustration for the idea of the approach to synthesize segmentation masks for food plate areas without pixel-wise annotation.

In order to deduce plate areas without pixel-wise annotation, in this study, we train not only food class classifiers but also food/non-food classifiers. In the recognition of food/non-food, plate areas will respond because those have strong co-occurrence with foods. On the other hand, in the recognition result of the food category, plates are included in most of food images, so the contribution to the recognition of the food category is not large. That is, in the visualization of the food class classifier and the food/non-food classifier is different and the difference may have correlation with plate areas. In this study, the difference between the visualization results of the dish area in these two classifiers is used to infer the dish area without the area level annotation.

2. Method

In this study, we propose a weakly supervised segmentation method for food images. In order to achieve it, we also train a food-plate segmentation model without pixel-wise annotation. We make use of the inference results of the plate areas and apply it to improve an accuracy of weakly supervised segmentation.

In this research, we synthesize segmentation masks for learning a segmentation model that infers plate areas of food images. In order to generate the plate areas, we use visualization results of a food class classifier and a food/non-food classifier. We assume that $v_O = CAM(x; \theta_{cl,O}) \in \mathbb{R}^{C \times H \times W}$ is a visualization result of the C-class food classifier for input image x generated by Class Activation Mapping (CAM) [4]. In the similar manner, the visualization results of the food/non-food classifier are represented by $v_F = CAM(x; \theta_{cl,F}) \in \mathbb{R}^{2 \times H \times W}$, where $\theta_{cl,O}$ and $\theta_{cl,F}$ are parameters for the classifiers. v_F will have large scores in food regions of images. On the other hand, the visualization result v_O of the food category classifier will have large responses in regions that are important for class identification. Both the visualization results are expected to respond to the area of the food regions of images. However, there is a difference between these visualizations. In particular, while the visualization of the food/non-food re-

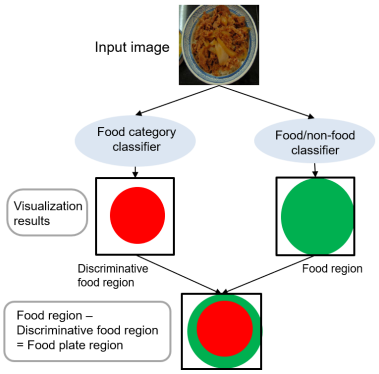

Figure 1. An illustration of the proposed approach for synthesizing plate segmentation masks using visualization.

turn clear responses in regions of strong co-occurrence with the food images, the visualization of the food category will not respond in the regions of strong co-occurrence with the food images because the objects which have strong co-occurrence with foods are included in most of food images and the objects are useless for recognition of food categories. In this paper, we assume that the regions that have strong co-occurrence with foods are plates of the food and we synthesize plate segmentation masks by utilizing the difference between the visualization results of the food/non-food classifier and the food category classifier.

We denote the steps of synthesizing of segmentation masks for plate areas from the visualization results. First, we obtain binary segmentation masks m_F whose pixels represent belonging to foods or non-food objects from v_F. Secondly, we obtain segmentation masks m_y for category labels y assigned to images from visualization results. If m_F and m_y are able to extract correctly, the difference in the masks should be objects have strong co-occurrence with food. We define the segmentation masks synthesized by the above processing as m_P. Here, we define a set of pixels of m_P of plate regions as S_P. This set can be represented by $S_P = S_F^{fg} - S_y^{fg}$, where S_y^{fg} is a set of the foreground of the discriminative food region and S_F^{fg} is a set of the fore ground of the whole food region. We train the plate segmentation model by synthesized background, plate area and food area ternary masks m_P. The learning loss of the model as follows:

$$\mathcal{L}_{plate} = -\frac{1}{\sum_{k=(0,1,2)} |S_k|} \sum_{k=(0,1,2)} \sum_{u \in S_k} \log(h_u^k(x;\theta_P)), \tag{1}$$

where h_u^k is conditional probability of observing any label k at any location u. S_k is a set of pixels for a class k of the mask m_P, where $S_0 = S_F^{bg}, S_1 = S_y^{fg}$ and $S_2 = S_P$.

We apply CRF to the probability map of the plate segmentation model and used the CRF applied results as the final plate segmentation results.

3. Experiments

In the experiments, we used the UEC-FOOD100 dataset [2]. The UEC-FOOD100 dataset [2] consist of 100 class food categories and each category includes 100 images. It should be noted that each food item has an annotated bounding box, but there are no annotation for pixel-wise segmentation masks. Therefore, we add new semantic segmentation annotation to 10% of UEC-FOOD100 dataset. We used this pixel-wise annotation for only evaluation.

Fig.2 shows examples of the food and plate area inferred by the plate segmentation model. We observed that the proposed method can infer various types of plate areas, instead of simply relying on color features or estimating only circular objects. These results are excellent, considering that we did not use any pixel-wise annotation for training of the segmentation model.

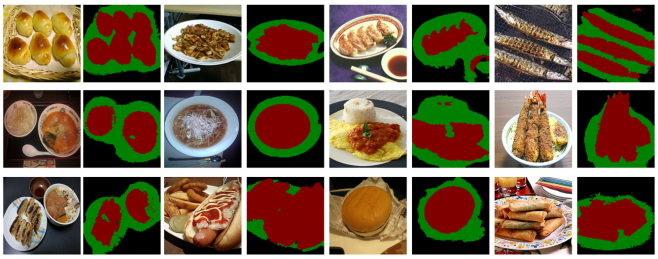

Figure 2. Examples of the plate segmentation model.

Table 1. Comparison with existing methods.

Method	mIoU	Pix acc
Baseline method [3]	50.2	77.5
BB annotation + grabcut [1]	51.1	81.9
Proposed	52.3	80.4

Table 1 shows the comparison between the proposed method and the existing method. The first baseline method is the method [3] used by the proposed method as a base framework. BB annotation + grabcut is a method that utilizes bounding box annotation proposed in [1]. This approach with usage of bounding boxes is expected to have a great advantage over the approach using only class labels, and can be considered a powerful baseline. Astonishingly, although the proposed method is a method using only class labels, it achieves accuracy close to the method using the bounding box. In addition, regarding the mIoU, the proposed approach using the inference result of the plate area by the proposed method achieved the higher accuracy than the methods using not only the baseline method but also the bounding box, which is an impressive result.

4. Conclusions

In this paper, we proposed a method to synthesize segmentation masks for food plate by visualization. Actually, we used a food category classifier and a food/non-food classifier for visualization and extracted strong co-occurrence regions with foods from the difference between the visualization results of the two types of the classifiers. In addition, we demonstrated that we can boost weakly supervised food segmentation by inference results of a plate segmentation model trained by the synthesized masks.

References

[1] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017. 2

[2] Y. Matsuda, H. Hoashi, and K. Yanai. Recognition of multiple-food images by detecting candidate regions. In *ICME*, pages 1554–1564, 2012. 2

[3] W. Shimoda and K. Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *ICCV*, 2019. 2

[4] B. Zhou, A. Khosla, A. Lapedriza, and A. Oliva, A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1