Few-Shot and Zero-Shot Semantic Segmentation for Food Images

Yuma Honbu Keiji Yanai
{honbu-y,yanai}@mm.inf.uec.ac.jp
Department of Informatics, The University of Electro-Communications
Chofu-shi, Tokyo, Japan

ABSTRACT

With the popularity of health management applications, awareness of dietary management is increasing. When calculating the number of calories in a dish, discriminating between food regions is an important factor. However, when using deep learning, a large amount of data is required for training, and it is impractical to collect data for countless food categories. In recent years, a method called few-shot segmentation has been studied to learn a semantic segmentation model using a small amount of training data. In this study, we propose a few-shot and zero-shot segmentation model which targets food images to overcome the insufficient amount of food training data and show the effectiveness of the proposed model on a semantic segmentation task for new food classes. In the proposed model, we employed the word embedding pretrained with a large-scale recipe text dataset, which results in better accuracy than the previous methods.

CCS CONCEPTS

• **Computing methodologies** → **Image segmentation**.

KEYWORDS

food image, food calorie estimation, food region segmentation, UEC-FoodPix Complete

ACM Reference Format:
Yuma Honbu Keiji Yanai. 2021. Few-Shot and Zero-Shot Semantic Segmentation for Food Images. In *Proceedings of the 13th International Workshop on Multimedia for Cooking and Eating Activities (CEA '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3463947.3469234

1 INTRODUCTION

In recent segmentation tasks, CNN-based models have significantly advanced the performance of segmentation. One of the existing food segmentation dataset, UEC-FoodPix Complete [9], consists of 102 categories and 10,000 images in total. However, regarding the number of food categories, it is not enough large for real situations, since the number of real food categories is much more than 100 in general. Few-shot segmentation, which has been studied recently, aims to correctly segment untrained categories by using information from a few support set images when a large number of training images for the target domain class are available. In the case of zero-shot, since no support set is given, we have to use the similarity to the target classes based on not visual information but textual information. Since there are no overlapped categories between the training and testing data in Few-shot and Zero-shot Segmentation, both the tasks are more difficult than the usual segmentation tasks. Therefore, training methods and handling of the support set are important for these tasks.

In this study, we demonstrate effectiveness of the proposed zero-shot segmentation method for food domain, where the distributional change between training and testing categories is considered to be small. We also propose a new few-shot semantic segmentation model that incorporates the word embedding technique used in the zero-shot task.

CEA '21, August 21–24, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8532-9/21/08…$15.00
https://doi.org/10.1145/3463947.3469234

2 RELATED WORK

2.1 Problem Setting

Few-shot segmentation is a task of region segmentation using a small number of training data. There is no overlap between the training and testing categories. Therefore, query images of unknown categories, pairs of support images and region mask images of the unknown categories as well as query images of the same categories are given as support set as inputs for testing. There also exists a task called Zero-shot Segmentation, in which unknown categories are regionally segmented using word embedded features or similarity between pre-trained features.

2.2 Few-shot Segmentation

The Prior Guided Feature Enrichment Network (PFENet) [10] proposed by Tian *et al.*. generates a learning-independent prior mask that estimates the target region by the similarity between the support and query high-level features, and gives the spatial location information of the target object. As a method to enhance the spatial resolution, they proposed a Feature Enrichment Module (FEM), which convolves global feature maps from local features in order. A zero-shot model is also proposed in the work of PFENet [10], in which the Masked Average Pooling (MAP) vector averaged over the target region of the support image used in the FEM is replaced by a word vector transformed by a 1×1 convolution layer and a ReLU function. In this paper, we use PFENet as a baseline because the UEC-FoodPix Complete [9] images we use contain multiple foods, and the spatial location of the target region is important. The details of the FEM architecture are shown in Figure 2.

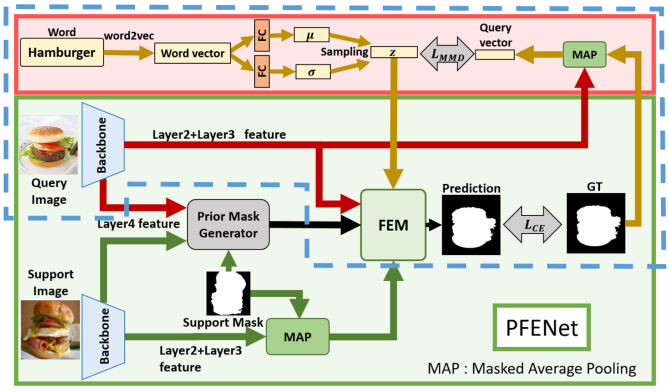

Figure 1: Proposed network for one-shot task (Blue dotted line: proposed Zero-shot model).

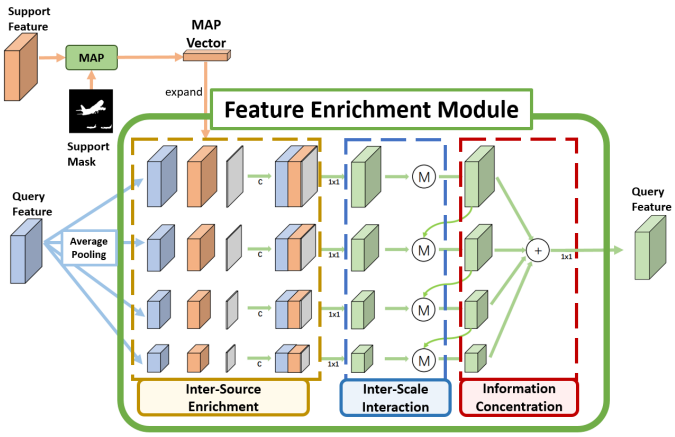

Figure 2: Feature Enrichment Module. (cited from [10])

2.3 Zero-shot Segmentation

For the zero-shot task, there are some methods to convert the embedded words into visual features. Bucher *et al.* [2] use Generative Moment Matching Networks (GMMN) [7], which consists of a full concatenation layer and Leaky ReLU to transform the noise concatenated embedded word vectors to visual features. Then, they use the Maximum Mean Discrepancies (MMD) reconstruction loss to facilitate the transformed features into visual features. Kato *et al.* [5] use Variational Autoencoder (VAE) [6] to transform them as latent spatial features, concatenate them with visual feature features, and convolve them into visual features. In this paper, we use the same transformation method and loss function.

3 METHOD

The distribution shifts between training and testing categories are usually large for common objects in appearance and shape between categories. Therefore, it is difficult to predict accurate region segmentation with only a few support sets. In this study, we focus on food domain images the distribution shift of which are expected to be less among categories compared to general object images such as Pascal VOC, since food images are often similar to each other in appearance and shape even between different categories. In addition to the support set, we proposed a new Few-shot model that can discriminate between the target region and unrelated region by adding the features of word embedding to enhance the information about the verification category. We also propose a zero-shot model using word embedding. A detailed diagram of the network is shown in Figure 1.

3.1 Word Embedding

We use word2vec [8] as a word embedding model. The word2vec model trained on the Wiki dataset is not appropriate for the food domain, because some food names do not exist in the Wiki data and the Wiki does not necessarily contain enough words related to foods. Therefore, we trained a word2vec model for the co-occurrence of ingredients by using about 160,000 recipe texts in the Cookpad dataset. By using the cooking ingredient words, dishes that use the same ingredients are embedded nearby in the representation space. Since we believe that the appearance of a dish depends on the seasonings and the color of the ingredients, obtained embeddings of ingredient words are effective when the training and test data have the same domain parts. We propose a method to incorporate "Recipe work2vec" into few-shot and zero-shot models for the food domain.

3.2 Few-shot model

The architectural details of the proposed method for few-shot are shown in Figure 1. In this study, we use the network of PFENet [10] as a base network, and the backbone is ResNet50 [4] pre-trained on Food-101 [1] as a feature extractor. The output feature map size of layer-2, 3, 4 is 1/8 of the original size. Similar to PFENet, Prior Mask is generated using the output of the high-dimensional features by layer-4. As for the query features, concatenated layer-2 and 3 features compress them into 256 dimensions using 1×1 Convolution and ReLU function. The support features are also compressed and masked average pooling into a support vector. These prior masks, support vectors, and query features are used in the Feature Enrichment Module (FEM).

The difference between ours and PFENet is the addition of the word processing branch in the orange area of Figure 1. The vector embedded in word2vec [8] is transformed into a word vector as a latent variable using VAE [6], and learned to reconstruct the MAP vector of the query by loss using MMD [7]. The word vectors are then used in the Inter-Source Enrichment part of the FEM where the pooled query features are convolved with the MAP vector and Prior Mask, and the query features are convolved with the word vector and Prior Mask separately. Then two features are concatenated and convolved into one feature.

3.3 Zero-shot model

In addition to the few-shot method, we also propose a zero-shot model, which is a method that does not use a support set. The details of the architecture are shown in the dotted blue area of Figure 1. The difference between the few-shot method and the zero-shot method is that the support set is used or not. Since the Prior Mask is also not generated, the Inter-Source Enrichment of FEM has a structure in which the query features and the generated

word vectors are concatenated and processed. The difference from the zero-shot model proposed in Tian *et al.*'s PFENet [10] is that the word vectors transformed by 1×1 convolution and ReLU are directly concatenated to the query features by FEM.

3.4 Loss function

As a loss function in this study, we use Maximum Mean Discrepancy (MMD) [7], which was used in ZS3Net [2], and the proposed model is trained so that the sampled word vector x_G generates the MAP query feature vector y_Q. MMD quantifies the difference between two distributions (x, y) by using the mean squared error $L_{MMD^2} = \|m_x - m_y\|_H^2$ between the two distributions mapped onto the Hilbert space(m_x, m_y) with the Gaussian kernel. Li [7] *et al.* have shown that by using the square root of L_{MMD^2} as the loss function, it is possible to learn to approach zero efficiently when the value becomes small. In this study, the square root of L_{MMD^2} was used as well. The overall loss is expressed as $L_{total.} = L_{PFENet} + L_{MMD}$ using the cross-entropy loss L_{PFENet}.

4 EXPERIMENTS

4.1 Dataset

As food image segmentation dataset, we use UEC-FoodPix Complete [9] which consists of 100 kinds of food categories with 10,000 pixel-wise-annotated images. As few-/zero-shot semantic segmentation, we define UECFoodPix-25^i by dividing UEC-FoodPix Complete [9] into four-folds . We also use Pascal-5^i, which are commonly used in the previous works as a general object dataset.

4.2 Experimental Details

To verify the effectiveness of the proposed methods for the few-shot model (wPFE) and the zero-shot model (zPFE), we conducted (1) quantitative analysis with each dataset and (2) validation of the embedded words. In (1), we made experiments with Pascal-5^i and UECFoodPix-25^i for the few-shot models. In the experiment with Pascal-5^i, we used ImageNet pre-trained backbone. In the experiment (2), we used the zPFE model for comparison. we used 40 million sentences from the wiki as the training data for word2vec [8], and the number of recipes from CookPad (10,000 and 160,000) for the zero-shot experiments. In addition, we conducted experiments using the PFENet to compare reconstruction methods that convert embedded words into visual features. The methods compared were: +Generator(GMMN)[7]+MMD, +VAE[6]+MMD (our method), and the method used directly in FEM without reconstructing the embedded words in the one-shot model of PFENet ("None").

We used mIoU (mean Intersection over Union) as the evaluation metric. The training was done with three folds and testing was done with one fold. The support set was randomly selected, and the average score of five trials was used. The results of experiments (1) are shown in Table 1, the results of experiment (2) in Table 2, the results of the qualitative analysis in Figure 3.

4.3 Experimental results

For experiment (1), the results show that the proposed method is more effective than the baseline for all folds when using the UECFoodPix-25^i dataset. On the other hand, the experiments using Pascal-5^i resulted in lower accuracy for both zero-shot and one-shot. Compared to the word2vec [8] model trained on ingredient words in the recipe dataset, the model trained on wiki sentences is expected to have embedded in the latent space that does not reflect visual information, as can be seen in Figure 6. This was possible to have affected the reconstructed vectors, resulting in a decreased

Table 1: Experimental results of one-shot and zero-shot segmentation.

	UECFoodPix-25^i					Pascal-5^i				
	One-shot		Zero-shot			One-shot		Zero-shot		
	wPFE (Ours)	PFENet [10]	**zPFE (Ours)**	PFENet [10]	Kato [5]	**wPFE (Ours)**	PFENet [10]	**zPFE (Ours)**	PFENet [10]	Kato [5]
Fold0	**0.847**	0.832	**0.808**	0.781	0.738	0.599	**0.617**	**0.524**	0.522	0.420
Fold1	**0.865**	0.855	**0.857**	0.822	0.767	0.681	**0.695**	0.637	**0.690**	0.583
Fold2	**0.818**	0.807	**0.788**	0.766	0.715	0.523	**0.554**	0.465	**0.524**	0.450
Fold3	**0.842**	0.832	**0.811**	0.776	0.722	0.541	**0.563**	0.442	**0.467**	0.364
Mean	**0.843**	0.832	**0.816**	0.786	0.736	0.586	**0.607**	0.517	**0.551**	0.454

Table 2: Results of zero-shot segmentation with the different word2vec models (left). Results of one-shot segmentation with the different reconstruction methods (right).

	CookPad(10K)	CookPad(160K)	wiki(40M)		+GMMN [7]+MMD	+VAE [6]+MMD	None
Fold0	0.798	**0.808**	0.807	Fold0	0.831	**0.847**	0.828
Fold1	0.823	**0.857**	0.827	Fold1	0.846	**0.865**	0.859
Fold2	0.780	**0.788**	0.772	Fold2	0.813	**0.818**	0.814
Fold3	0.783	**0.811**	0.801	Fold3	**0.846**	0.842	0.826
Mean	0.796	**0.816**	0.802	Mean	0.834	**0.843**	0.832

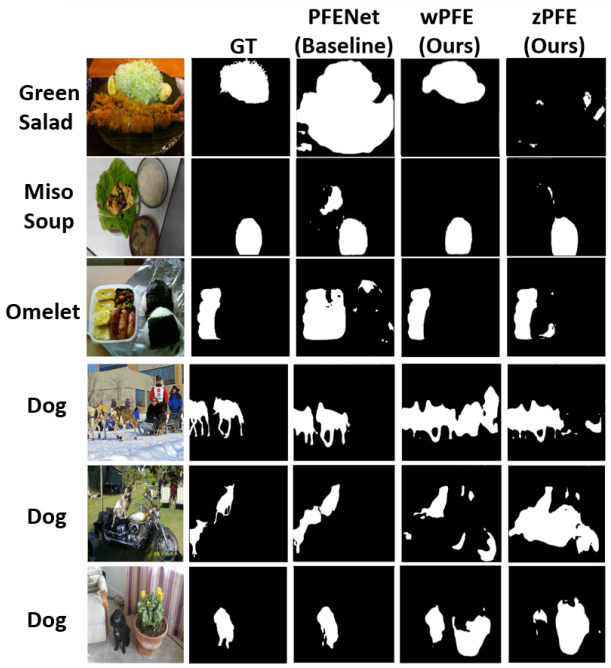

Figure 3: Results of qualitative experiments with UECFoodPix-25^i (Green Salad, Miso soup, Omelet) and Pascal-5^i (Dog).

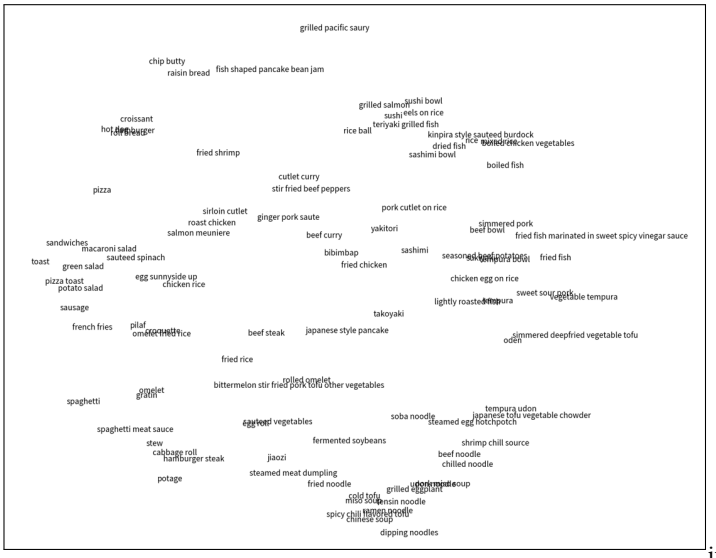

Figure 4: Categorical scattering plot of UECFoodPix-25^i. (Please enlarge the figure in a PDF viewer.)

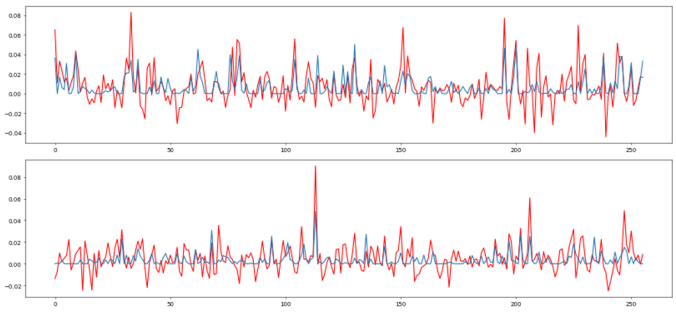

Figure 5: Reconstructed vectors. Red is the reconstructed features, blue is the MAP vector averaged over the features in the target region of the query. (Top: sweet-and-sour pork, Bottom: kenchin-soup)

accuracy. In addition, the categories of Pascal VOC have a very small number of similar categories compared to the food dataset. This is also expected to be one of the reasons why it is difficult to improve accuracy.

For experiment (2), the results in Table 2 show that the model using 160,000 recipes for training word2vec [8] and VAE [6] for reconstruction is effective. Since the words used in training of word2vec are ingredient words extracted from recipes, the same ingredients of dishes are distributed close to each other as seen in Figure 4, and the distribution of dishes using ingredients that affect the appearance of the dishes are also close to each other. Therefore, the word vectors reflect the visual features, and the new category of words generates vectors that combine features of dishes similar to the target region of the training category. As can be seen in Figure 5, this reconstruction method facilitates the conversion of the target region into visual features. Also, as shown in the left of Table 2, increasing the number of recipes improves the accuracy by allowing more visual information to be handled, which results

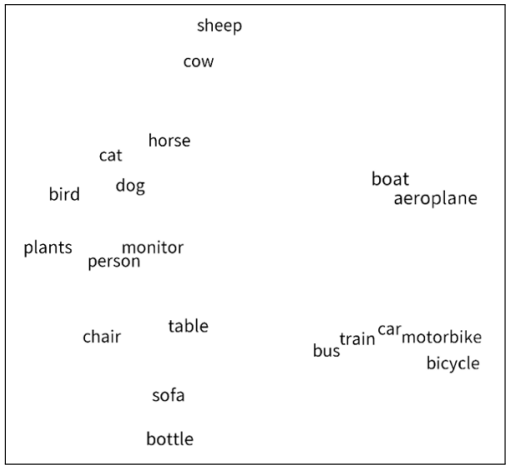

Figure 6: Categorical scattering plot of Pascal-5^i

in more discriminative region segmentation than the baseline, as shown in Table 1 and Figure 3.

5 CONCLUSION

In this study, we proposed a new method based on PFENet with the addition of word embedding. The experiments show that the proposed method outperformed the conventional method in both few-shot and zero-shot segmentation on the food dataset, which demonstrated the effectiveness of the proposed method on the food dataset.

Future works include the construction of a general-purpose model that can be adapted to general objects and the use of BERT [3] as a better word embedding method.

Acknowledgments In this study, we use the CookPad dataset, which is available at the NII Data Repository.

REFERENCES

[1] L. Bossard, M. Guillaumin, and L. Van Gool. 2014. Food-101 Mining Discriminative Components with Random Forest. In *Proc.of European Conferenceon Computer Vision (ECCV).*

[2] M. Bucher, T. Vu, M. Cord, and P. Pérez. 2019. Zero-Shot Semantic Segmentation. In *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc.of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).*

[4] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Deep Residual Learning for Image Recognition. In *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[5] N. Kato, T. Yamasaki, and K. Aizawa. 2019. Zero-Shot Semantic Segmentation via Variational Mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops.*

[6] P. Kingma and M. Welling. 2014. Auto-Encoding Variational Bayes. In *Proc.of International Conference on Machine Learning (ICML).*

[7] Y. Li, K. Swersky, and R. Zemel. 2015. Generative Moment Matching Networks. In *Proc.of Proceedings of International Conference on Machine Learning (ICML).*

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc.of International Conference on Learning Representations (ICLR).*

[9] K. Okamoto and K Yanai. 2021. UEC-FoodPix Complete: A Large-scale Food Image Segmentation Dataset. In *Proc.of the ICPR Workshop on Multimedia Assisted Dietary Management (MADIMA).*

[10] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia. 2020. Prior Guided Feature Enrichment Network for Few-Shot Segmentation. In *Proc.of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).*