심층 합성 곱 신경망을 이용한 음식 위치 탐지 및 인식 기법에 관한 연구

로힛 타쿠르, 강주성, 한현준, 이흥노*
광주과학기술원 전기전자컴퓨터공학부

e-mail: trohit920@gist.ac.kr, heungno@gist.ac.kr*

Concurrent Food Localization and Recognition using Deep Convolution Neural Network

Rohit Thakur, Jusung Kang, Hyunjun Han and Heung-No Lee*

School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology
Gwangju, South Korea

Abstract: **The analysis of food quality is considered as an important task because of the people's nutrition habits that affect the global population. For this, it is very important to keep track of our eating habits by making automatic nutrition diaries. Keeping this in mind, we demonstrate how a deep learning technique is used to the tasks of localizing and recognizing various types of foods present in given images. Firstly, the Grad-CAM algorithm (a class-discriminative localization technique) along with data processing is used for generating discriminative image regions which helps in obtaining object proposals for detection purpose. Secondly, a classifier for object recognition is employed over it. Further, we compare our result with best state of art work and show our proposed method to be well equipped in terms of precision and accuracy. Through experiments we have verified that proposed approach is a convincing solution for concurrent food localization and recognition.**

Keywords — Convolutional Neural Network; Food Recognition; Grad-CAM; Guided Backpropagation;

I. INTRODUCTION

Computer vision and multimedia community have recently shown great interest in food image analysis and automatic nutrition diaries. Developing such a recording tool may help users to record their eating habits which is usually a manual exercise using textual description, but at the same time it is time consuming and inefficient. So this fact leads to development of many approaches to monitoring diet by making use of information technology [1]. The widespread use of mobile devices and digital cameras can be seen as a potential source for collecting photographs of meals before eating. This together with advances in computer vision enables the automatic construction of nutrition diaries [2]. We can automatically infer user's eating pattern by developing food detection and food recognition algorithms which can be applied for both mobile and lifelogging images. A recent approach that could ease the automatic construction of diet assessment diaries is based on Deep Learning, more precisely Convolutional Neural Networks (CNN) which gained attention by winning ImageNet competition (ILSVRC) and outperforming by far the competition [3]. These networks are able to handle local feature extraction, feature coding, complex spatial patterns and learning from images.

In this paper we propose to use a novel and fast approach for localization and recognition of various foods in given images using CNNs. We use the Grad-CAM [4] technique which is capable of generating discriminative image regions by class-specific gradient information and produces heat map activations corresponding to different classes. A data processing technique called Guided Backpropagation (GB) [5] which gives a high resolution and fine grained image details achieved through pixel-space gradient visualizations is further added with Grad-Cam. We obtain the precise bounding boxes manually according to the coarse localization map which we get from the fusion of Grad-Cam and Guided Backpropagation. This shows a great improvement in localization result as compare to other methods. Apart from this, we use a food recognition algorithm which learns by itself to recognize food after fine-tuning and pre-training of networks. We compare our result with others and obtain high accuracy and recall.

In literature, a similar work is in [6] where authors introduced a food activation map for determining object proposals and then recognize each of the food present in each bounding box which causes overlapping of food related objects and also require changing network architecture every single time.

II. PROPOSED METHODOLOGY

Our novel method specialized for food detection detects all discriminative features of different image regions with one forward pass by back propagating weights from final layer to convolutional layer and leads to high precision and recall. Grad-CAM algorithm localizes discriminative image regions to generate heat maps which are weighted sum of class specific visual patterns at different spatial positions.

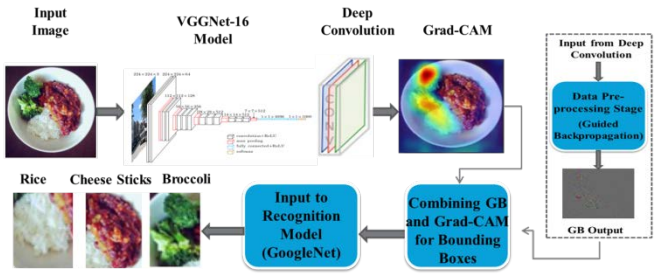

Fig. 1: Network Structure

Food Localization using Grad-CAM: By visualizing regions which are important for predictions Grad-CAM helps detecting food related objects in given images. Mathematically to compute class-discriminative localization map [4] $L^c_{Grad-Cam} \in R^{u \times v}$ of width '*u*' and height '*v*' for any class '*c*', we calculate $\frac{\partial y^c}{\partial A^k}$ which is gradient of the score for class '*c*' with respect to feature maps 'A^k' of convolutional layer. While back propagating the gradients through global pooling we get the importance weights α_k^c as, $\alpha_k^c = \frac{1}{Z}\sum_i \sum \frac{\partial y^c}{\partial A_{ij}^k}$. This weight α_k^c captures the important feature map '*k*' for a target class '*c*' which is food here. Finally, $L^c_{Grad-Cam} = \mathrm{Re}\,LU(\sum_k \alpha_k^c A^k)$ gives the coarse heat map of food as same size as of convolutional layer. Additional data processing technique called Guided Backpropagation obtains high resolution and fine grained food image regions which when combined together with Grad-Cam by point wise multiplication resulted in segments of high quality pixels around single target. So we draw the boxes around the single largest segment obtained to get precise bounding boxes. These boxes will serve as input to recognition network later. This is general approach and can be used for all kind of CNN networks.

Grad-CAM's Fine-tuning for obtaining activation map:

We use VGG-16 [7] pre-trained model from Caffe Model Zoo [8] with no change in architecture structure. The model is only trained for classification on class labels and no bounding box annotations are used for the purpose of localization while in [6] GoogleNet structure have been modified to add Global Average Pooling layer before final Softmax layer.

Food Recognition by Transfer Learning:

Food Recognition is the final step of our network which classifies each of the detected regions as a type of food. We fine-tuned the Google Net pre-trained on ILSVRC. We perform an additional fine tuning before the final test of performance so that our network can extract valuable information from extensive food datasets by parameters adaptation before the final fine tuning. At the end we fine-tuned the network or targeted domain data to analyze our network performance.

III. EXPERIMENTAL RESULTS

We will describe about our working environment i.e. data set used, pre-training performed; algorithm used for generating object proposals, experimental setup and obtained results.

Datasets

UECFood256 [9]: The dataset consists of 256 different classes of intercontinental dishes with at least 100 samples each.

Food101 [10]: This dataset is basically for food recognition which consists of 101 different types of foods from around the world with 1000 samples per class.

UECFood100 [11]: This dataset consists of 100 kinds of Japanese food which are used for recognition purpose.

ILSVRC 2013 [12]: It contains more than 400,000 images and 1,000 classes for training and validation (with a subset of classes related to food).

Fine Tuning and Pre-training:

Here we will discuss all the pre-processing done at each step. Before training on specific food datasets our model Inception V3 is pre-trained on ILSVRC 2013 dataset for 48 hours during 200000 iterations.

Food Recognition Training:

We used the Food101 Dataset as the first dataset for fine-tuning the food recognition network pre-trained on ILSVRC. We split the datasets into 80/20% split where 80% for training and 20% for validation, respectively. The test set provided was used for testing. On the second fine-tuning, the same preprocessing was applied on both UECFood256 and UECFood100. Parameters used for fine tuning are learning rate of 0.001; momentum is 0.9 and epsilon value is0.1. A decay of 0.1 was applied with batch size of 30 images.

After fine tuning, adjusting parameters and number of iterations we achieve the best Top-1 and Top-5 Validation and Test Accuracy as shown in table1 which are better than other state-of-art methods.

Table 1

		Validation Accuracy		Test Accuracy	
Datasets	Pre-training	Top-1	Top-5	Top-1	Top-5
Food101	ILSVRC	76.33%	90.88%	79.84%	92.25%
UECFood 256	ILSVRC Food101	70.18%	82.58%	72.54%	86.81%
UECFood 101	ILSVRC UECFood256	78%	94.06%	85.79%	97.62%

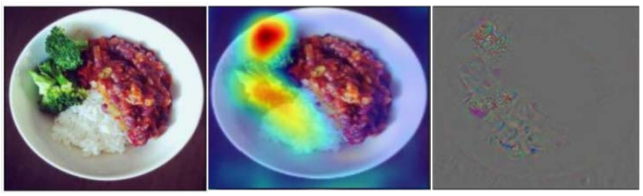

Fig2: Showing results of Grad-CAM and GB

Figure 2 shows the result of using Grad-CAM algorithm and performing additional data processing for localizing different food instances in a single image. The high discriminative image regions obtained leads to get precise bounding boxes and when labeled accurately, these images serve as input to the Recognition network which is the main cause for increase in the performance.

Comparison Results:

For comparison purpose we define the Intersection of Union (IOU) metrics which is how precise is the predicted bounding box (bb) with respect to ground truth (GT) annotation.

IoU (bb) = $\frac{GT \cap bb}{GT \cup bb}$. A bounding box is considered valid when its IoU $\geq$ 0.5. The other evaluation metrics used are: Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$ and Accuracy = $\frac{TP}{TP+FP+FN}$, where the True Positives (TP) are the bounding boxes correctly localized, the False Positives (FP) are the predicted bounding boxes that do not exist in the ground truth, and the False Negatives (FN) are the ground truth samples that are lost by the model.

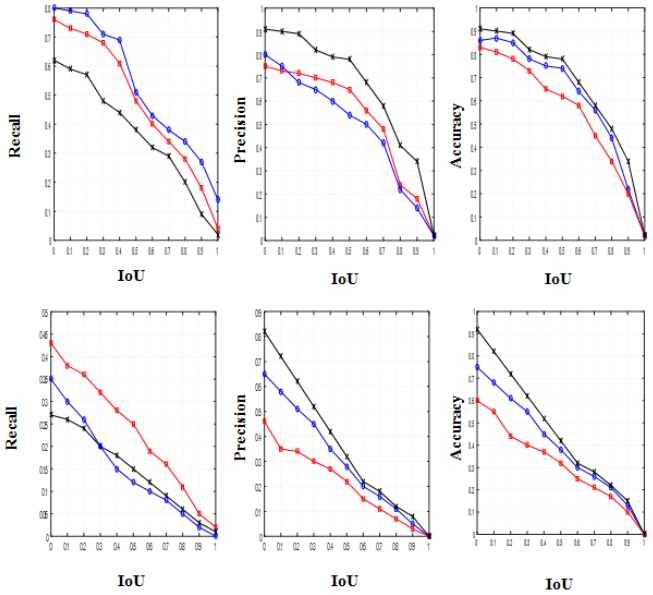

Fig. 3: Curves of Recall vs IoU (left), Precision vs IoU (centre) and Accuracy vs IoU (right) on the test sets of UECFood256 (top) and Food101 (bottom). Our method is shown in Black, Faster R-CNN [13] in Red and FAM [6] in Blue.

The comparison shows that proposed method performs well in terms of Accuracy and Precision while giving tough competition in Recall. This is clear because our purpose here is to localize as well as recognize food objects in given images while Faster-RCNN is generic object localizer and FAM [6] doesn't have additional data processing so lacks a little in terms of precision and accuracy. In terms of Recall, we can leave all the details of every object as our aim is efficiently localize objects not to find whatever appearing in images. Figure 4 shows some of the obtained results.

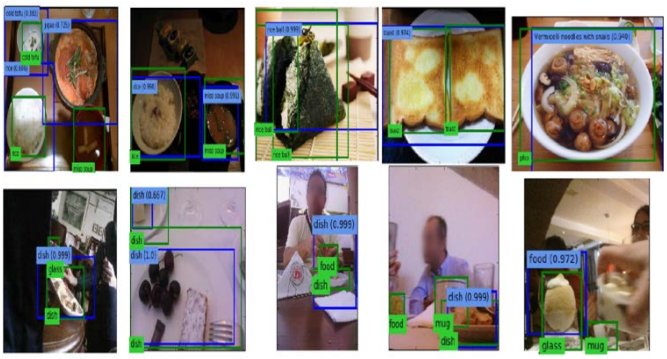

Fig. 4: Examples of localization and recognition. Ground truth is shown in green and our method in blue (recognition score between parenthesis).

IV. CONCLUSION

We propose a methodology of concurrent Food Localization and Recognition. Our approach with Grad-CAM and data processing achieves good results than other baseline object localizers. This approach can be used for both conventional and egocentric images.

V. ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government(MSIP) (NRF-2015R1A2A1A05001826) and Korea Aerospace Research Institute (KARI).

REFERENCES

1. F. E. Thompson, et.al."Need for technological innovation in dietary assessment," *J.Amer. Diet. Assoc.*, vol. 110, no. 1, pp. 48–51, 2010.
2. Kiyoharu Aizawa, Yuto Maruyama, He Li, and Chamin Morikawa. "Food balance estimation by using personal dietary tendencies in a multimedia food log." Multimedia, IEEE Transactions on, 15(8):2176–2185, 2013.
3. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, pages 1097–1105, 2012.
4. Ramprasaath R. Selvaraju, et.al."Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization" in arXiv:1610.02391v2 [cs.CV] 30 Dec 2016

5. J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. “Striving for Simplicity: The All Convolutional Net”. CoRR, abs/1412.6806, 2014. 2, 3, 19
6. Bolaños, M. and Radeva, P. (2016). “Simultaneous food localization and recognition.” In arXiv:1604.07953, 2016.
7. K. Simonyan and A. Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In ICLR, 2015.
8. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. “Caffe: Convolutional Architecture for Fast Feature Embedding”. In ACM MM, 2014
9. Y. Kawano and K. Yanai. “Automatic expansion of a food image dataset leveraging existing categories with domain adaptation.” In Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV), 2014.
10. Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. “Food-101–mining discriminative components with random forests.” In Computer Vision–ECCV 2014, pages 446–461. Springer, 2014.
11. Yoshiyuki Kawano and Keiji Yanai. “Food image recognition with deep convolutional features.” In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pages 589–593. ACM, 2014.
12. Olga Russakovsky et.at.” ImageNet Large Scale Visual Recognition Challenge”. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015.
13. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks.” In Advances in Neural Information Processing Systems, pages 91–99, 2015.