# Deep Residual Convolutional Networks for Food Image Segmentation

Chintushig Ochirsukh
*School of Sciences, University of Massachusetts Lowell*

## Abstract

***The objective of this study is to review the state of the art in segmenting instances of food from images with computer vision and machine learning techniques, and proposing the use of residual convolutional networks for more accurate segmentation results. A few categories of food will be studied against various methods, and based on the results we hope to gain insight on the performance of various segmentation algorithms with an evaluation metric of Intersection over Union (IoU).***

## I. INTRODUCTION

In the past few years, there have been a number of advances in Instance Segmentation - a task involving classifying every pixel in an image as belonging to a particular instance of a class. Unlike Semantic Segmentation, Instance Segmentation requires multiple instances of the same class to be differentiated accordingly. The challenging nature of this task is represented in the number of submissions in the COCO image segmentation challenge and the Cityscapes challenge leaderboards. In COCO, there were only 5 submissions for instance segmentation as of Mar. 2018, with 31 submissions for Object Detection[1]. In Cityscapes, there were 11 submissions for Instance Segmentation, yet 58 submissions for Semantic Segmentation.

Additionally, due to the lengthy training time of machine learning models for Semantic Segmentation, many of the applications of image segmentation for food images often resort to computer vision techniques without Machine Learning. Much of the existing work for dietary assessment fall into one of three computer vision techniques: contrast based thresholding, Sobel boundary extraction, or a deformable parts model [2]. All three of these operate on the color domain, but due to the similar color and textures of food and non-food items, it becomes difficult to achieve high accuracy using only color based methods. Among the deep learning methods for food image segmentation, there has been increasing research in the use of Region-based Convolution networks (R-CNN)[3], but training the network would take a week on a modern 8-core GPU. To mitigate this, we propose using a method of Semantic Segmentation named Masked R-CNN, originally developed at Facebook AI for general purpose Semantic Segmentation [4], for the purpose of segmenting only food images.

## II. LITERATURE REVIEW

Within Instance Segmentation, there are a number of methods which are currently used to segment foods. One of the primitive methods used presently is purely color-based Otsu's thresholding proposed by Mery et Al [5]. In this process, the original image is converted into a contrast intensity image. Then, a threshold is determined to separate the foreground and the background, and finally the variance and standard deviation of this image is used in separating instances of various foods. Another study involving the use of thresholding was that of Kang et al [6] in which a polynomial equation was formed to express the food color distribution, but both methods above performed poorly when presented with food instances containing multiple colors. Both of these methods operate on the color domain, and it has been hypothesized by [2] that they often perform poorly due to the similarity in color of the foods and the vessels they are placed on.

### A. Brief History of Convolutional Networks

A more recent and accurate method of segmenting food within images is the use of Convolutional Networks. These networks were inspired by the research of Hubel and Wiesel's experiments on the Primary Visual Cortex (V1) of a Cat and its columnar architecture. They soon identified and were awarded a Nobel Prize for understanding that the cells with similar functions

were grouped together and built on previous ”layers” of cells to perform higher level recognition. In 2012, Alex Krizhevsky et al developed a deep convolutional network and won the ImageNet competition [7]. However, the task of Image Classification was slightly different from the task of Image Segmentation, but served the backbone for a number of subsequent architectures.

Inspired by the work of Alex et al. at the University of Toronto, a team at UC Berkeley led by Jitendra Malik developed a region-based convolution network for the task of Semantic Segmentation [8]. In this work, an image is fed to the network, and a set of bounding boxes are chosen at different points in the image (using a process described Selective Search) and each bounding box is fed to an AlexNet (the Convolutional Network proposed by by Alex et Al.) to classify the type of the object. The process of proposing these bounding boxes is named the **Region Proposal**, and is important in our methodology shown in section III.

Unfortunately, RCNN requires running the AlexNet from beginning to end once for each region of interest (RoI), which requires a training time of several days or weeks on a modern GPU. To overcome this, Ross Girschick - the author of RCNN - developed a new method named Fast R-CNN by realizing that the computation of running the entire alex net does not need to be calculated for each bounding box. Rather, we can feed the entire image through the AlexNet and pass the features of the last layer (FC7 features) to a pooling layer, then two separate layers for classification and regression. This whole process required just one forward pass of a single network.

Although R-CNN was quite fast, the process of Selective Search for generating region proposals was a bottleneck in the speed, so a new method composed of Shaoqing Ren et al at Microsoft Research created Faster-RCNN [9]. This method realized that the FC7 features extracted in the last layer of the Convolution Network can be used to propose Regions of Interest (RoI) instead of running a Selective Search on the whole image again. To do this, Faster R-CNN adds another Convolutional Layer on top of the FC7 features, this is called the Region Proposal Network (RPN). Then, a final pooling layer, named RoiPool, will expand the regions in the feature map back into the regions in the original image (this process is necessary because the conv layers reduce dimensionality, and a cell in the RPN no longer corresponds to a pixel in an image).

Extending this work, we propose a more recent method of Instance Segmentation named Mask R-CNN that performs comparatively well among other food segmentation methods suggested in this paper.

# III. About Mask-RCNN

So far, RCNN and Faster R-CNN have only been used for classification and bounding box regression of objects, but we need a more novel approach to be able to classify each pixel in an image as belonging to a class. To do this, Mask-RCNN was proposed by Kaiming He et al. at Facebook AI [4].

## A. Backbone CNN Features

Unlike Faster RCNN, the backbone Convolutional Features which are fed to the Region Proposal Network (RPN) are no longer trained with an AlexNet, but rather a 101-layer deep residual network proposed by Microsoft Research [10] named ResNet-101. The insight in using Residual Networks is that with deeper-layer network architectures, the training and test error were often greater. This raised the suspicion that back-prop and SGD were not optimized for multi-layer networks. To overcome this, a skip connection was added between every other layer, allowing gradients to flow backward faster. Further insight on this can be found in [10].

## B. Mask Layer

Their work begins by starting with the Faster R-CNN network, and adding a branch after the Region Proposal Network (RPN) to detect whether or not each pixel in that RoI is part of an object. This branch is just a fully convolutional layer on top of the Region Proposal Network (RPN).

## C. RoiAlign (based on RoiPool)

Next, rather than applying RoIPool, Kaiming’s team realized that RoiPooling was slightly misaligned and Image Segmentation required pixel-level specifity unlike bounding box regression. When presented with a feature map of 25x25 for an original image of 128x128 and we need to choose 15 pixels from the original image, RoiPooling was simply scaling the feature map by 2.93 (15 * 25 / 128). However, to avoid this rounding, RoiAlign uses Bilinear Interpolation to estimate the color and texture in pixel 2.93 of the feature map. This seemingly small change made a significant boost in pixel-level accuracy.

## IV. My Work

### A. Training

To further analyze the effect of Mask R-CNN on food image segmentation, we trained the entire Mask R-CNN architecture described in [4] on a NVidia Volta V100 GPU for 10 hours constrained to only 9 categories of food: 'apple', 'sandwich', 'orange', 'broccoli', 'carrot', 'hot dog', 'pizza', 'donut', and 'cake'. During this process, we only trained the Mask Layer (Subsection B), and did not re-train ResNet-101 backbone (as training this would take weeks on this GPU). Since we used the pre-trained weights for the ResNet backbone, the region proposals and bounding boxes will point to non-food items, but the mask layer only segments food items. All 9 of the categories are included in the COCO image dataset.

### B. Additional Tools

## V. RESULTS

Overall, it is difficult to empirically compare the Mask-RCNN image segmentation method proposed in this paper with that of computer vision techniques due to the lack of publicly available datasets of food

## VI. DISCUSSION

## VII. CONCLUSION

## VIII. ACKNOWLEDGEMENTS

## References


[1] CodaLab. (2017) COCO Detection Challenge. [Online]. Available: https://competitions.codalab.org/competitions/5181

[2] Hsin-Chen Chen, Wenyan Jia, Xin Sun, Zhaoxin Li, Yuecheng Li, John D. Fernstrom, Lora E. Burke, Thomas Baranowski, and Mingui Sun, "Saliency-aware food image segmentation for personal dietary assessment using a wearable computer," *Meas Sci Technol*, vol. 26, 2015.

[3] Yanchao Liang, Jianhua Li, "Deep Learning-Based Food Calorie Estimation Method in Dietary Assessment," *School of Information Science and Engineering, East China University of Science and Technology, China*, 2017.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick, "Mask R-CNN," *Facebook AI Research*, 2017.

[5] Ying-Wen Chang and Yen-Yu Chen, "An Improve Scheme of Segmenting Colour Food Image by Robust Algorithm," *Workshop on Combinatorial Mathematics and Computation Theory*, vol. 23, 2005.

[6] Kang SP, Sabarez HT., "Simple colour image segmentation of bicolour food products for quality measurement," *Journal of Food Engineering*, vol. 25, 2009.

[7] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *ImageNet*, 2012.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *ImageNet*, 2014.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Towards real time object detection with Region Proposal Networks," *Microsoft Research*, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep residual learning for image recognition," *Microsoft Research*, 2015.