MULTI-FOOD ITEM DETECTION

Matthew Morgan

Oregon State University
COMPUTER SCIENCE

Introduction

Keeping track of dietary habits traditionally requires a written diary of the foods consumed at every meal. To simplify this process, we imagine an automated food computer vision system that would save people time and convenience by autonomously logging meals. In this project, we lay the foundation for this vision. We aggregate a dataset of annotated multi-label food images, then train a PyTorch model to recognize and localize 18 different foods with 91% accuracy. Our research creates a baseline for future work that we hope can be used to track nutritional intake in hospitals, schools, and correctional facilities.

Dataset

We assemble an American cafeteria food dataset of annotated meal images comprised from publicly available datasets [1] and [2]. Dataset [1] consists of 100 primarily Japanese food categories, from which we select 12 foods. Dataset [2] consists of 73 primarily Italian foods, which we also select 12 foods from. We then merge 6 of the classes from both datasets, as shown in Table 1. The result is a dataset of 3839 images with n=18 classes. From these images, we perform a 70/20/10 percent stratified split into train, test, and validation sets, respectively.

Table 1: Classes merged from [1] and [2]

Food Name	Count from [1]	Count from [2]	Total Count
pizza	134	89	223
bread roll	107	479	586
french fries	153	64	217
spinach	102	110	212
white rice	620	57	677
mixed salad	342	46	388

Methods

To detect food items, we implement the state-of-the-art YOLOR model [3] which detects objects by class and localizes them with bounding boxes. We specify the model's input and output image size as 1280x1280 pixels. Then, we train for 300 epochs with a batch size of 8 images, using the default YOLOR-P6 hyperparameters and 0.20 as the IoU training threshold.

Results

Our model detects 18 food items with 0.917 overall mAP@0.5 IoU* and 0.721 overall mAP@0.95 IoU on held-out test data. Per-class mAP@.5 and the count of each item in the overall dataset is displayed below.

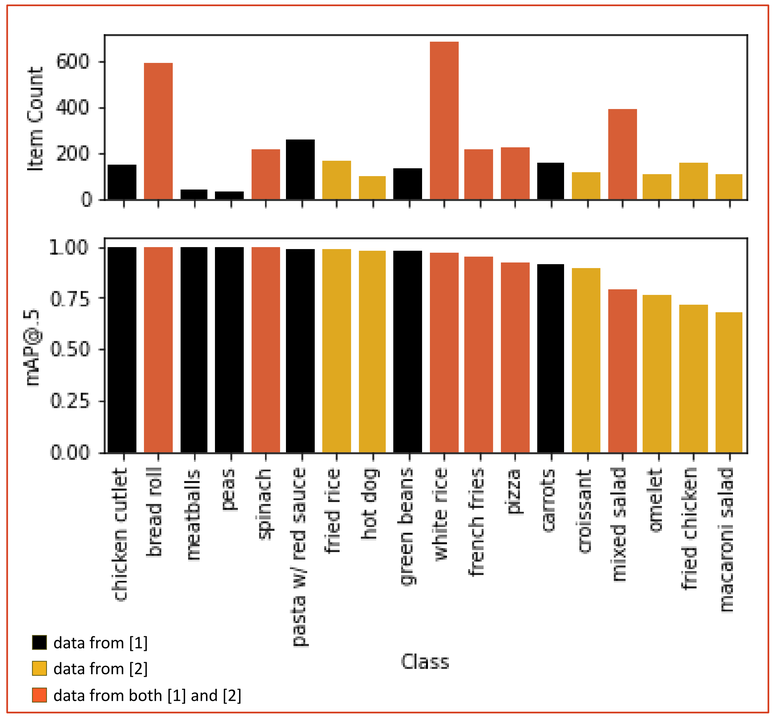

Figure 1: Per-class mAP@0.5 IoU

Discussion

Our results demonstrate aggregating images from disparate datasets as an effective method for developing a food detection model with high accuracy. While there are a few publicly available international food datasets, we hope our model will provide a baseline for producing a more thorough annotated American food image dataset. Towards that end, we plan to conduct a small data collection study in a university cafeteria in late April 2022.

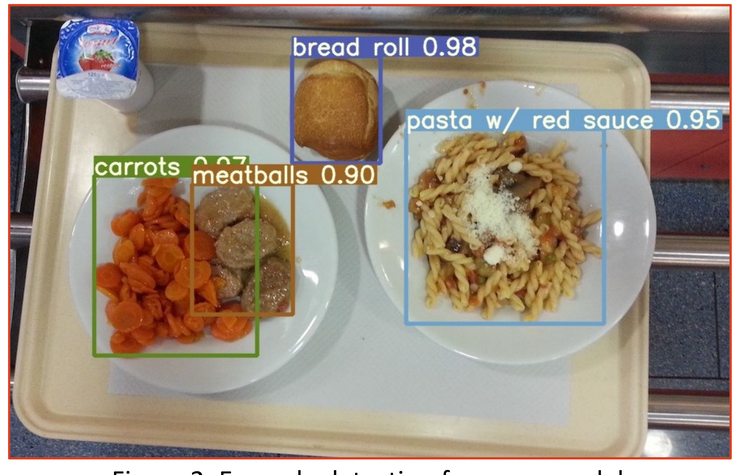

Figure 2: Example detection from our model

References

[1] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of Multiple-Food Images by Detecting Candidate Regions," 2012.

[2] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition: a new dataset, experiments and results," IEEE J. Biomed. Health Inform., vol. 21, no. 3, pp. 588–598, 2017.

[3] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You Only Learn One Representation: Unified Network for Multiple Tasks," ArXiv Prepr. ArXiv210504206, 2021.

$^{*}\ mAP@0.5 = mean\ average\ precision\ where\ true\ vs. pred.\ {}^{intersection}/_{union} > 0.5$