Food Recognition Paper Notes

* UEC-FOOD101, UEC-FOOD256, UEC-FoodPix have bounding boxes already labeled on the data. Is it worth it to use these datasets for object detection?
* **Aguilar et al. (15) introduced the usage of deep learning in food image recognition, by applying YOLOv2, DarkNet-19 on the UNIMIB2016 dataset to eventually obtain a precision of 0.841.**
  + How feasible is it to use Yolo Darknet? This was YOLOv2 as well, can we achieve better results with YOLOv5?
* Freitas et al. (18) reported results from various experiments with different deep learning architectures (including Mask R-CNN, DeepLab V3, SegNet, ENet) on a proprietary dataset of Brazilian food items. They demonstrate that Mask R-CNN outperforms the rest of the approaches, with a mAP of 0.87, whereas the rest of the methods scored a mAP of <0.79.
  + Mask R-CNN still outperforms all deep learning architectures, reinforcing what I presented last week with the three algorithms.
* The food images are categorized in 273 different classes with at least 35 annotations per class.
  + How small of a dataset can we get away with in our project? In this paper, they had 35 annotations at minimum, so does this mean that we can use fewer annotations per class than normally expected?
* What model/algorithm did they use for image segmentation? If we had it, would it let us make much more accurate and efficient labeling?
* Hybrid Task Cascade (22) and Mask R-CNN (16) were the two most popular architectures.
  + About Mask R-CNN: “It results in producing many noisy predictions.”
* “The success in the segmentation task, and the difficulty in the classification task indicates the need for much larger training dataset sizes, especially in case of the class distributions across 273 categories.”
  + Classification will always require much more data than segmentation/detection.

What is Hybrid Task Cascade?

* Uses both masks and bounding boxes to train this type of object detection model

​​<https://ieeexplore.ieee.org/document/8917599>

General Notes

* Interesting method of data augmentation: “As the resolution of the training images varies greatly, random JPEG compression was applied.”
  + Since we cannot expect the same quality of images as we received in this first batch, it may be sensible to apply this type of data augmentation as well to train the machine on lower resolution images.
  + “On the assumption that the only way to distinguish between such classes as peppermint tea and herbal tea was through subtle differences in coloring, we kept color-related augmentations to a minimum.”
    - We may be unable to do color augmentation, as they bring up an important point on the subtle effects of color.
* Mean Average Recall points to the percentage of True Positives. Precision is the probability of the predicted bounding boxes matching actual ground truth boxes, also referred to as the positive predictive value.
* The “Backbone” of a Model is just the feature extractor network. (https://stackoverflow.com/questions/59868132/what-does-backbone-mean-in-a-neural-network#:~:text=Backbone%20is%20a%20term%20used,models%20to%20generate%20segmented%20masks.)