Maltese Food Recognition Using Mask R-CNN

Mattei Vella, Alan Gatt

Institute of Information & Communication Technology Malta College or Arts, Science & Technology Corradino Hill

Paola PLA 9032

*{*mattei.vella.e22881, [alan.gatt*}*@mcast.edu.mt](mailto:alan.gatt%7d@mcast.edu.mt)

LITERATURE REVIEW

There have been many successful attempts at recognizing different types of food items using neural networks, although recent research has shown that deep learning techniques also managedto achieve good results. S. Mezgec and B. K. Seljak stated in their paper that, *“Deep learning achieved considerably better results in this field, as deep learning is a fully automatic machine learning approach, and is therefore more appropriate for food image recognition”*. Throughout their research, an architecture called *“NutriNet”* was created and used to recognize food items/beverages. The main difference in this architecture is that both segmentation and the classification of the image where being dealt with in the one single network. At the time of the release of this paper, both authors claimed that the provided solution achieved many results which better than others in the field which is always an intriguing factor when comparing any sort of result. For possible continuation on their research, the idea of approaching food weight estimation by using reference objects which could be anything which proves to be a of adequate size was also mentioned.

Fast R-CNN was also researched in 2017 when processing food images and detecting what each item is. *“In Fast R- CNN, region proposals are generated from a feature map of the entire image,so that CNN is used only once. Then, bounding boxes and class scores for each category are estimated from a region proposal finally. Thus, Fast R-CNN realizes high- speed and highly accurate detection.”* [[2].](#_bookmark12) Ege and Yanai stated that at the time, recognition of food photos have proven to work and have also been released, but not many are capable of detecting multiple dishes in the same image, it is why they opted to use Faster R-CNN. This algorithm makes use of RPN(Regional Proposal Networks) as an extra step before CNN to provide more accurate results with smaller executions times. RPN uses anchor boxes to generate proposals, these proposals can be referred to as the location of each

different object found in an image. In order to properly test the algorithm a dataset with multiple food dishes in one image was created with different school lunch items, these where also combined together with the *UEC FOOD-100* [[3]](#_bookmark13) which included multi-label food images with bounding box annotations. With regards to the classification of each food item, the extracted results from the proposed method exerted a *”mean average precision”* with an impressive result of 90.7%, but there is still room for improvement if this ought to be used on a daily basis and in real life scenarios. The paper also included calorie estimation of multiple food items on one lunch-room tray. A data set of calorie annotated images were used in order to train the model with the use of Chainer, a deep learning framework. By using Chainer, the researchers were able to output the calories directly from a photo by extracting each bounding box found in the image to separate each item, then adding up the total of each estimated calorie per bounding box to create the total calories in the image. The results for the calorie estimation still had a lot to be improved on, as most images had an approximation of 40% error.

Tiankaew, Chunpongthong and Mettanant also gave their take on food image processing when they developed an application which can recognize up to 13 different Thai food classes.These individuals opted to make use of a custom modified version of *“VGG19 CNN”*.*“We remove three fully connected layers from VGG19, and replace them with a fully connected layer with 1024 nodes, followed by a dropout layer, another fully connected layer with 1024 nodes, and the output layer with 13 nodes using softmax as its activation function, respectively.””* [[4].](#_bookmark14) An interesting approach which the research took was that instead of opting to gather the images themselves, a script was written which accesses Google images via the food class name in order to retrieve images of the food items. With this approach they managed to gather a good dataset of 7,632 images without having to manually take these images or pay to get them taken. The only manual work needed was then to re-size and clean the images. Whilst they did include an option to track the calories of each food item based on the images, it seems as if this being calculated via a static table. For example, the class *Fried Rice*, would be equal to 600 calories without taking into consideration the portion size and proportion of the item based off of a calibration object to be able to calculate the scale factor. Having said this, the authors did also mention that there can be improvements made with the predictions of calories.

In 2017 Matterport inc. released an open source code for their implementation of Mask R-CNN in which many people had contributed in implementing this method to various different topics [[5].](#_bookmark15) This repository was an extremely good introduction to what Mask R-CNN can provide in many different scenarios whilst keeping a high level of accuracy in most cases. Apart from the high levels of accuracy being generated from this algorithm, the most interesting aspect of it was that one would be able to not only analyze a bounding box around the object being analyzed in the image, but also provide the exact pixels where it is located. This extra layer of

pixel by pixel detection is what differs between Faster R-CNN and Mask R-CNN, it is why many people in the past have opted to call it, it’s predecessor. In 2018 K. He, G. Gkioxari,

1. Dollar and R. Girshick mentioned these differences in their paper on Mask-RCNN, whilst also stating their new and improved version would *“outperform all existing single- model attempts”* [[6].](#_bookmark16) Results also speak for themselves since they have also stated that they show top results in all three tracks of the COCO suite of challenges, including instance segmentation,bounding-box object detection, and person key point detection [[6].](#_bookmark16) Being able to have the extra layer mentioned in the algorithm provided a great possibility of calculating the weight of a food item’s with greater accuracy since food can’t always be shown in the exact shape and form. This was researched specifically by researchers in the Department of Computer Science found in Taiwan due to finding out many problems in the everyday diet of Taiwanese people [[7].](#_bookmark17) They made use of two main dataset’s, these being the *Food-256* and*VillaCafe* dataset’s which where made up of around 16 different food classes when combined together. Many different portions sizes where taken into consideration so that they had varied results whilst always keeping a fixed photograph angle and fixed image size as to reduce the number of variables present which could affect the final results. To analyze the relationship between the ratio of each food item with the actual portion in reality, linear regression was made use of to try and predict the actual weight based of the previous record. Using these techniques they have managed achieve results of above 96% in all metrics used with regards to the classification of the food classes and an average absolute error of 8.22 when estimating the weight of the food items from a total of 320 points.

During the conference on *Multimedia Information Process- ing and Retrieval* in 2019, researchers focused on Calorie Esti- mation from food images, had the opportunity to review their different attempts using varies different methodologies. Five main works where being mentioned, *CaloireCam, Region- Segmentation based food calorie estimation, AR DeepCalo- rieCam V2, DepthCalorieCam and RiceCalorieCam* [[8].](#_bookmark18) The researchers described ”CalorieCam” to be a mobile application capable of estimating the calories with the help of an object which can be determined upon setting up the application in order to make it easier for people to use items which they can carry around anywhere like a wallet or credit card. Once an estimate of how big the food item is in context to the personal belonging, clustering, segmentation and CNN based recognition is done on an engine found on the smartphone it’s self. They stated that currently it is taking around 0.2 seconds on an android phone to be able to compute the estimation. With the ”CalorieCam” it was always assumed by the researchers that there would only be one food item to recognize in the picture, which limited the application quite a bit, this is why they introduced CNN-based region proposals in the second work mentioned. In the *Segmentation Based Calorie Estimation* work, area ratios from multiple foods are taken into consideration in order to calculate calories for differentfood classes in one image without having to take more then a single photo. When discussing the 3rd piece of work which is the ”AR DeppCalorieCam”, the researchers stated that the previous works are very similar but in this case an inertial sensor which built into a smartphone is also used. This was used to be able to directly detect the size of the food item without having a reference object in the image it’s self. The framework from Apple ”ARKit” was used to do so. Getting rid of the reference object makes the software even more use- able on a day to day basis where life is very fast, and having to always use a specific object to detect the calories on a plate of food might be found tedious. Another feature which iPhone provided which helped to create another version of their previous works, was the multi-camera feature. These cameras could be used as stereo cameras, so by knowing the distance between each back camera they where able to calculate the the actual sizes of objects based on triangula- tion. The application was implemented on iPhone and called ”DepthCalorieCam”. When comparing ”CalorieCam”, ”AR CalorieCam” and ”DepthCalorieCam”, the results showed that ”DepthCalorieCam” was providing very small margin’s of error when compared to the previous two. Having said this, when ”CalorieCam” and ”AR CalorieCam” where compared, the results where quite different because in each of the 3 food classes chosen, which where ”pork with sauce”, ”fried chicken” and ”croquettes”, the results varied. In some cases ”CalorieCam” performed better and achieved more accurate results and other times it was the other way round. The Azumio Team released their original mobile application called *Calorie Mama AI: Diet Counter* which was one of the first ever fully functioning AI application available for anyone with a smartphone, whilst providing accurate tracking. This was first released in 2017 with the developers stating that the app is capable of learning continuously whilst providing real time estimations [[9].](#_bookmark19) Recently in July of 2021 a case study was launched to test Image recognition applications, particularly using Calorie Mama as their main study [[10].](#_bookmark20) The individuals taking part in the case study made use of two testing techniques, manual and automated testing in order to see how the application would perform. The main differences between the two techniques where that in automated testing developers wrote scripts to test the application and made use of tools to generate the necessary images for detection. When around 400 different cuisines where tested, the app resulted a 33% failure rate whilst giving them 43.75% failure rate when using automated testing. Even though the rates seem to be quite large .Unfortunately the algorithm being used is not shared with the general public so it is not possible to know what they are using to get such great results, but an application like this shows large possibilities that if one makes use of the correct algorithm an everyday usable application can be used. When it came to understanding what and how other data sets are being used for image processing, some key features from each set where identified. The Food 101 data set, was a very common choice for many people, it having fifteen- thousand plus Citations as can be analyzed from Table [I.](#_bookmark2)

This is because it is made up of a large number of images, food items and categories, making it a very good starting point for anyone who would want to classify a larger number of categories.

Seven out of the nine data sets also had multiple number of labels for each image, this means that if there were two food items in a single image, there would be a label representing both these items for the same image. Unlike the Food 101 [[11]](#_bookmark21) and Food 50 [[12],](#_bookmark22) the data sets only offered single labels per image which might not be adequate if one would like to test his concept on multiple items in one image.

A feature which only the UNIMIB2016 [[13]](#_bookmark23) data set offered was the ready-made data annotations. These annotations were created in a COCO format [[14]](#_bookmark24) where each food item had a Bounding Box and its Segmentation. This would be of great help because it would not require for and individual to annotate each image one by one, thus resulting in data set which would require less preparation. Below one can find a list of food image datasets which have been analyzed, while also giving some important information on each table in tabular form.

1) Food 101, [[11]](#_bookmark21)

2) Food 50, [[12]](#_bookmark22)

3) UEC FOOD-100, [[3]](#_bookmark13)

4) UNIMIB2016, [[13]](#_bookmark23)

* 1. AIFood, [[15]](#_bookmark25)
  2. VireoFood-172, [[16]](#_bookmark26) 7) Food524DB, [[17]](#_bookmark27)

8) Food475DB, [[18]](#_bookmark28)

9) UEC FOOD-256, [[19]](#_bookmark29)

TABLE I

DATA-SET COMPARISON.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Images** | **Categories** | **Citations** | **Annotations** | **Labels** |
| 1 | 101,000 | 101 | 15,800 | No | Single |
| 2 | 15,770 | 50 | N/A | No | Single |
| 3 | 9,060 | 100 | N/A | No | Multiple |
| 4 | 1,027 | 73 | N/A | Yes | Multiple |
| 5 | 372,095 | 24 | 287 | No | Multiple |
| 6 | 110,241 | 172 | N/A | No | Multiple |
| 7 | 247,636 | 524 | 33 | No | Multiple |
| 8 | 247,636 | 475 | 33 | No | Multiple |
| 9 | 31,397 | 256 | N/A | No | Multiple |

Shown in the below Table, one can find some results from different research papers, each paper making use of different methods and approaches to the image processing of food. From the first glance one can see that there are already a substantial amount of good results, thus giving the option to experiment with multiple techniques.

Since this research is partly focused around the recognition of food items in an image, a confusion matrix to describe the performance of this classification is extremely use-full to better visualize the predicted/actual results. This can be shown in Table [III.](#_bookmark4)

With the results from Table [III,](#_bookmark4) the evaluation metrics can be derived by exchanging the abbreviations in the equations below with the numeric values obtained from the matrix.

TABLE II

FOOD RECOGNITION RESULTS

**Study Data set Algorithm Accuracy**

|  |  |  |  |
| --- | --- | --- | --- |
| [[20]](#_bookmark30) | Thai Fast Food | Deep Learning | 88.33% |
| [[21]](#_bookmark31) | Custom-made data set | Multi-View Multikernel SVM | Multikernel SVM: 70-100%  Singlekernel SVM: 60-90% |
| [[22]](#_bookmark32) | FOOD-5K  FOOD-11 FOOD-101 | Deep Feature Extraction via CNN | 99.00%  88.08%  79.86% |
| [[23]](#_bookmark33) | UEC-FOOD100 | CNN | 60.90% |
| [[13]](#_bookmark23) | UNIMIB2015 | CNN | 99.03% |
| [[24]](#_bookmark34) | ETH Food-101 | Ensemble Net CNN | 73.50% |
| [[7]](#_bookmark17) | Food-256  Ville Cafe | Mask R-CNN | 99% |

TABLE III

FOOD ITEMS EVALUATION METRICS

**Predicted**

FoodItem Not-FoodItem

**Actual** FoodItem

TP

Not-FoodItem FN

FP TN

1. **Accuracy (ACC)** - A calculation used in order to see in general, how often the classifier is correct.

*TP* + *TN*

*Accuracy*(*ACC*) =

*P* + *N*

(1)

1. **Precision (PPV)** - When the classifier predicts a correct value, how correct is it?

*Precision*(*PPV* ) =

*TP*

*TP* + *FP*

(2)

1. **Recall (TPR)** - When the classifier correctly identifies a positive value, is the item actually positive?

*TP*

*Recall*(*TPR*) =

*P*

1. **F1 Score** - A measure of a the classifier’s accuracy.

2*TP*

*F* 1*Score* =

2*TP* + *FP* + *FN*

(3)

(4)

1. **True Negative Rate (TNR)** - When The object is

actually false, how often does the classifier predict no?

*TN*

*TNR* = (5)

*TN* + *FP*

REFERENCES

1. O. O. on Health Systems and M. Policies, “Country health profile 2019, state of health in the eu, oecd publishing, paris/european observatory on healthsystems and policies, brussels.” 2019.
2. T. Ege and K. Yanai, “Estimating food calories for multiple-dish food photos,” in *2017 4th IAPR Asian Conference on Pattern Recognition* *(ACPR)*, 2017, pp. 646–651.
3. Y. Matsuda, H. Hoashi, and K. Yanai, “Recognition of multiple-food images by detecting candidate regions,” in *Proc. of IEEE International* *Conference on Multimedia and Expo (ICME)*, 2012.
4. U. Tiankaew, P. Chunpongthong, and V. Mettanant, “A food photogra- phy app with image recognition for thai food,” in *2018 Seventh ICT* *International Student Project Conference (ICT-ISPC)*, 2018, pp. 1–6.
5. W. Abdulla, “Mask r-cnn for object detection and instance segmentation on keras and tensorflow,” [https://github.com/matterport/Mask RCNN,](https://github.com/matterport/Mask_RCNN) 2017.
6. K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask r-cnn,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 42,  no. 02, pp. 386–397, feb 2020.
7. M.-L. Chiang, C.-A. Wu, J.-K. Feng, C.-Y. Fang, and S.-W. Chen, “Food calorie and nutrition analysis system based on mask r-cnn,” in *2019 IEEE 5th International Conference on Computer and Communications* *(ICCC)*, 2019, pp. 1721–1728.
8. T. Ege, Y. Ando, R. Tanno, W. Shimoda, and K. Yanai, “Image-based estimation of real food size for accurate food calorie estimation,” in *2019 IEEE Conference on Multimedia Information Processing and Retrieval* *(MIPR)*, 2019, pp. 274–279.
9. A. Ltd., “Introducing calorie mama,” Mar 2017. [Online]. Available: <https://www.azumio.com/blog/azumio/calorie-mama-release>
10. C. Tao, D. Cao, H. Guo, and J. Gao, “A case study of testing an image recognition application,” 07 2021.
11. L. Bossard, M. Guillaumin, and L. Van Gool, “Food-101 – mining dis- criminative components with random forests,” in *European Conference* *on Computer Vision*, 2014.
12. C. Termritthikun and S. Kanprachar, “Nu-resnet: Deep residual networks for thai food image recognition,” *Journal of Telecommunication, Elec- tronic and Computer Engineering (JTEC)*, vol. 10, no. 1-4, pp. 29–33, 2018.
13. G. Ciocca, P. Napoletano, and R. Schettini, “Food recognition: a new dataset, experiments and results,” *IEEE Journal of Biomedical and* *Health Informatics*, vol. 21, no. 3, pp. 588–598, 2017.
14. T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays,

P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr, “Microsoft coco: Common objects in context,” 2015.

1. G. G. Lee, C.-W. Huang, J.-H. Chen, S.-Y. Chen, and H.-L. Chen, “Aifood: A large scale food images dataset for ingredient recognition,” in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 2019, pp. 802–805.
2. H. Hoashi, T. Joutou, and K. Yanai, “Image recognition of 85 food categories by feature fusion,” in *2010 IEEE International Symposium*  *on Multimedia*, 2010, pp. 296–301.
3. G. Ciocca, P. Napoletano, and R. Schettini, “Learning cnn-based features for retrieval of food images,” in *New Trends in Image Analysis and Processing – ICIAP 2017: ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11-15, 2017, Revised Selected Papers*, S. Battiato, G. M. Farinella, M. Leo, and G. Gallo, Eds. Springer International Publishing, 2017, pp. 426–434.
4. ——, “Cnn-based features for retrieval and classification of food images,” *Comput. Vis. Image Underst.*, vol. 176-177, pp. 70–77, 2018. [Online]. Available: <https://doi.org/10.1016/j.cviu.2018.09.001>
5. 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.
6. N. Hnoohom and S. Yuenyong, “Thai fast food image classification using deep learning,” in *2018 International ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineer-* *ing (ECTI-NCON)*, 2018, pp. 116–119.
7. H. He, F. Kong, and J. Tan, “Dietcam: Multiview food recognition using a multikernel svm,” *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 3, pp. 848–855, 2016.
8. A. engr, Y. Akbulut, and . Budak, “Food image classification with deep features,” in *2019 International Artificial Intelligence and Data* *Processing Symposium (IDAP)*, 2019, pp. 1–6.
9. W. Zhang, D. Zhao, W. Gong, Z. Li, Q. Lu, and S. Yang, “Food image recognition with convolutional neural networks,” in *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th*

*Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated* *Workshops (UIC-ATC-ScalCom)*, 2015, pp. 690–693.

1. P. Pandey, A. Deepthi, B. Mandal, and N. B. Puhan, “Foodnet: Recog- nizing foods using ensemble of deep networks,” *IEEE Signal Processing* *Letters*, vol. 24, no. 12, pp. 1758–1762, 2017.
2. J.-M. Noh, G.-R. Jang, K.-N. Ha, and J.-H. Park, “Data augmentation method for object detection in underwater environments,” in *2019 19th International Conference on Control, Automation and Systems (ICCAS)*, 2019, pp. 324–328.