Journal of Theoretical and Applied Information Technology
31st December 2023. Vol.101. No 24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PROPOSED ENHANCED FEATURE EXTRACTION FOR MULTI-FOOD DETECTION METHOD

SUHAILA ABUOWAIDA[1], ESRAA ELSOUD[1], ADAI AL-MOMANI[1], MOHAMMAD ARABIAT[1], HAMZA ABU OWIDA[2], NAWAF ALSHDAIFAT[3], HUAH YONG CHAN [4]

[1]Department of Computer Science Faculty of Information Technology , Zarqa University, Zarqa 13100, Jordan Email:sabuoweuda@zu.edu.jo
[2]Department of Medical Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Zip-code (Postal Address):19328, Amman, Jordan.
[3] Department of Computer Science Prince Hussein Bin Abdullah, Faculty of Information Technology, Al al-Bayt University, Mafraq, 25113, Jordan.
[4]School of Computer Sciences, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia

ABSTRACT

This research presents a comprehensive system that utilizes computer vision and deep learning techniques to develop the detection of multiple food methods. Despite the incorporation of deep learning techniques, the effectiveness of the existing detection method for different food products is unsatisfactory due to the utilization of ResNet-101 for feature extraction. The features maps of the ResNet-101 exhibit a size reduction or may vanish entirely following the down-sampling process. The ResNet-101 blocks may have been subject to varying degrees of repetition, with certain blocks receiving a restricted number of repeats and others being excessively repeated. There is an ongoing need to develop the rate of detection in the field of food recognition. The procedure under consideration consists of a series of primary steps. The optimization of the ResNet-101 block entails the careful selection of an appropriate number of repetitions. A supplementary convolutional layer is suggested. The results produced from this approach were later compared to the outcomes reached by the latest algorithms in object detection, specifically Mask R-CNN and CASCADE R-CNN. The evaluated algorithm exhibits exceptional performance in accuracy AP over multiple thresholds. The numbers regularly exceed the relevant criteria of three commonly utilized techniques. The thresholds demonstrate higher magnitudes than the thresholds of three frequently utilized methods.
Keywords: *Deep Learning, Object Detection, Resnet-101, Features Maps, Down-Sampling Process.*

1. . INTRODUCTION

The importance of eating cannot be overstated about individuals' nutritional well-being and overall quality of life [1]. The efficient identification and categorization of food photos depicting daily meals enables individuals to acquire useful data and proficiently evaluate and condense their requirements [2]. For detection, individuals not medically diagnosed with specific dietary restrictions might strive to maintain a balanced nutrition intake [3,4]. Conversely, individuals diagnosed with diabetes are advised to abstain from consuming dishes that are rich in sugar [5]. Furthermore, healthcare professionals can assess a patient's prior dietary patterns and subsequently provide appropriate and rational dietary guidance [6].

The detection of food images has consistently been a subject of interest and importance. In a previous study [7], researchers introduced statistical techniques for determining the features of dish photographs to facilitate food identification. In the study conducted by the researcher referenced as [8], the random forest algorithm was employed to extract shallow features to classify image. The classification of dish photographs was examined in [9], focusing on using texture Anti-Textons characteristics. In these instances, the method accuracy and ability performance is subpar.

The utilization of convolutional neural networks for image classification has been prominent in computer vision, owing to the advancements in deep learning [10–12]. Kagaya et al. [13] presented the Alex-Net

model to classify food photos. [14] demonstrated the utilization of a more intricate inception model to classify photos of food. [15] employed the Google model to categorize photos into two distinct classes: food images and non-food images. The advancement of object detection approach [16,17] has presented increased demands for picture recognition. The primary algorithms that are widely used can be classified into two distinct types. The proposed approach is founded on the convolutional neural network known as Region Proposal, which encompasses a series of models such as R-CNN [18], Fast R-CNN [19], Faster R-CNN [20], and Mask R-CNN [21]. The R-CNN methods are characterized by a two-step approach, wherein they produce candidate boxes for the target and subsequently forecast the detection outcomes. Another type of method is the single-stage approach, shown by YoLo [22–24] and SSD [25]. These algorithms rely solely on convolutional neural networks (CNNs) to infer various objects' categories and spatial coordinates immediately.

In general, it has been observed that the two-stage approach exhibits higher detection quality than the single-stage method. Notably, the Mask-RCNN method demonstrated superior performance in the COCO Challenge, surpassing the current single-model entries in all tasks [26].

However, the Mask-RCNN algorithm incurs significant computational costs, mostly attributable to using the ResNet-101 architecture as the backbone model architecture. The ResNet-101 model has various concerns: Some ResNet-101 blocks could have been subjected to a restricted number of repeats, and others may have been overly reproduced. There is an ongoing need to develop the detection rate for food detection. Furthermore, it is seen that the feature maps of the ResNet-101 undergo a decrease in size or may even disappear completely as a result of the down-sampling procedure.

The ResNet-101 model has been developed with the addition of two layers. The optimization of the ResNet-101 block necessitates a meticulous selection of the optimal number of iterations. It is recommended to incorporate a convolutional layer at each hierarchical level to optimize the model's accuracy. The inherent problems of this endeavor hindered the work of segmenting food items. The primary goal of this research article is to develop a theoretical framework for food detection to address the challenges above. Therefore, a novel approach, multiple food detection, was proposed, consisting of two steps. The initial phase develops the process of extracting food features by enhancing RestNet-101. This leads to the development of a developed ResNet-101 backbone, which effectively addresses the issue of low detection rates. The subsequent phase entails the integration of the Region Proposal Network (RPN) to facilitate the localization of diverse food items. The primary objective of the research is to investigate the process of multiple food detection. The utilization of the method involves the implementation of the Region of Interest (RoI), for instance, segmentation, hence facilitating the extraction of distinct attributes about each instance.

The assessment of the multiple food detection encompasses two metrics, such as the average intersection over union (IoU) score with different thresholds, also known as average precision (AP).

2 . PROPOSED METHOD

The present study provides the spotlight on the process of multiple food detection. Specifically, it addresses the identification of tasks associated with numerous food items and the localization and detection of each food item from the provided image. Several factors, including the presence of many food items and their differences in color and size, make these tasks challenging. Consequently, we need to extract richer semantic and more abstract features to describe many food images and obtain the best results.

CNN, an abbreviation for Convolutional Neural Network, is a deep learning framework that leverages various inputs, including audio, images, text, and video. CNNs have demonstrated a significant level of precision across several areas, such as object detection [26-29]. Many CNN designs, such as AlexNet, VGG, and ResNet-101, are commonly acknowledged as backbone networks. The ResNet-101 design has been considered a viable method for attaining developed precision in object detection [21]. The ResNet-101 model provides superior performance by utilizing a deep neural network architecture that integrates a sequence of blocks specifically engineered to address the issue of gradient vanishing. The accomplishment of this is facilitated with the

incorporation of expedited connections.

Nevertheless, ResNet-101 exhibits many concerns: Some ResNet-101 blocks may have been subjected to a restricted number of repeats, while others may have been overly repeated. There is an ongoing need to develop the detection rate for food identification. The ResNet-101 design has been improved to include three distinct phases. The optimization of the ResNet-101 block necessitates the meticulous selection of the optimal number of repetitions. It is recommended to incorporate a convolutional layer at each level to optimize the model's performance. The proposed improvement to the ResNet101 backbone entails modifying the initial convolution layer by decreasing the filter size and reordering the layers that need more iterations.

Furthermore, a convolutional layer is incorporated at each stage to optimize the feature selection process, extracting many distinct and precise feature maps from the lower-level layers. Figure 2.1 illustrates the improved version of the ResNet-101 backbone. Therefore, the current work has devised an innovative methodology for identifying several food items, resulting in improved precision. Consequently, this strategy effectively addresses the concerns above.

The current study addresses the problem of a restricted number of repeated blocks, which hinders the attainment of developed precision and decreases durations for training and testing. The frequency of iterations is determined by conducting experiments on the ResNet-101 backbone, specifically in the context of food data. The experiment involves assessing the precision of the ResNet block. If the newly obtained accuracy surpasses the previously attained accuracy, the ResNet block is iteratively replicated until a superior level of accuracy is attained. The occurrence of repeats ceases whenever the newly attained level of accuracy falls below the previously achieved level of accuracy. The process of optimization entails the determination of the optimal number of repetitions. The above technique has been conducted to ascertain the blocks that need either further or fewer iterations. Following this, the outcome of each step is combined with an extra convolutional layer to augment the precision of the recommended ResNet-101 backbone.

Furthermore, the proposed methodology outlined in the study entails the selection of suitable repetitions for each block inside the backbone architecture. Additionally, it contains the reduction of filter size during the first stage to augment the extraction of intricate features from the input image. Furthermore, the integration of a convolution layer is applied at each hierarchical level. Incorporating a convolution layer at each stage is a widely adopted technique in the implementation of depth-wise separable convolutions. The objective of this strategy is to disassemble the convolutional layer in relation to the depth axis, which leads to improved accuracy [29]. In addition, the suggested backbone architecture integrates a 1x1 convolution operation to effectively reduce the dimensionality of the input without compromising any inherent features [29].

This characteristic demonstrates notable benefits throughout the process of summation, as seen in Figure 2.1, whereby each stage is outfitted with a unique channel. The suggested backbone architecture receives an image as input and produces a feature map of a predetermined size across many layers. The objective of this strategy is to improve the effectiveness of the selected method by combining many local feature maps and extracting both shallow layer feature maps and strong semantic features, as seen in Figure 2.1.

The decrease in spatial resolution of the robust semantics features can be attributed to the down sampling procedure. As a result, the suggested backbone integrates the bilinear up sampling technique in order to augment the resilience of the features, amalgamating them with the feature map obtained from the preceding steps. The objective of this merger is to provide a comprehensive understanding of semantics and to create a feature map that possesses the utmost level of detail.

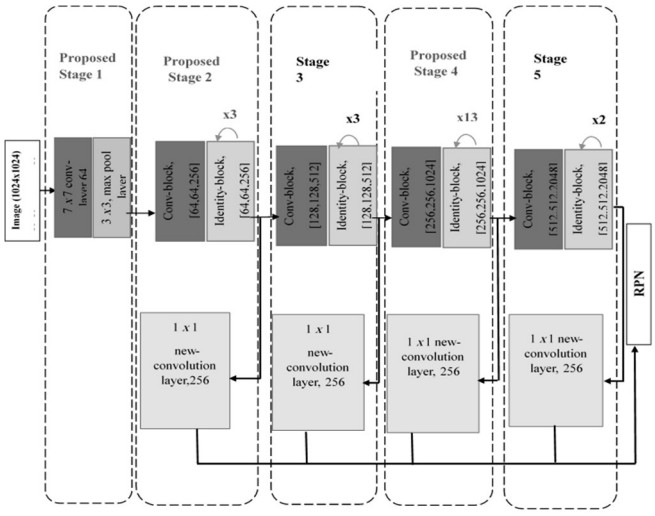

Figure 2.1. Proposed Method Of Multiple Food Detection

3. . EXPERIMENTS, RESULTS AND DISCUSSION

In this part, a series of tests were done to evaluate the performance of the proposed technique in detecting numerous food images. This section commences with a description of the evaluation metrics and experimental conditions. Subsequently, we present the experimental data and analyze multiple food image detection.

3.1 Datasets

The experiments are conducted on multiple food detection dataset [30], which includes 27,000.
80% images for training, 20% for validation

3.2 Implementation Details

To assess the efficacy of the suggested methodology, a comparative analysis is conducted between the proposed approach and two established models, namely the standard Mask R-CNN and CASCADE R-CNN. The assessment metrics for detection include using Average Precision (AP) as evaluation indicators. AP is a reliable measure for assessing the degree of similarity between the actual target and the predicted target. Regarding the consumption of the model, it mainly pertains to the size of the model and the duration required for training. The experiments were performed using the GPU model Tesla V100 with a memory capacity of 16 GB, as well as the virtual central processing units (VCPUs) with eight cores and a memory capacity of 16 GB. The implementation used Tensorflow 2.0 and Python 3.6.

The dimensions of the anchor are defined as (128, 256). A value of 512 is assigned to the aspect ratio, defined as (0.5, 1, 2). Stochastic gradient descent (SGD) optimizer is used in the training. The learning rate was configured to 0.001, the momentum was set at 0.9, and the training process consisted of 50,000 epochs.

3.3 Result Of Resnet-101 Backbone

The ResNet-101 backbone was carefully designed, taking into consideration the optimal amount of duplicates for each convolution block. Additionally, using convolutional layers at certain levels has been found to improve the accuracy of food detection significantly [31].
The approach was built by using food visuals as a foundation. Furthermore, using convolutional layers at a certain level has been seen to improve the precision of food detection significantly.

The procedure entails the collection of local feature maps from each step and extracting the ultimate feature maps. The final feature map is acquired by aggregating the local feature maps from each phase through the summation process. Subsequently, the feature map is employed as the input provided to the subsequent component inside the multi-food detection methodology. The RPN operates by the principles of Mask R-CNN. The performance of the ResNet-101 model that was constructed is depicted in Figure 3.1. Experiments were carried out at every level of ResNet-101, and the ideal number of repetitions of the ResNet-101 block was identified during the training process [32]. Subsequently, further convolution layers were included in some stages.

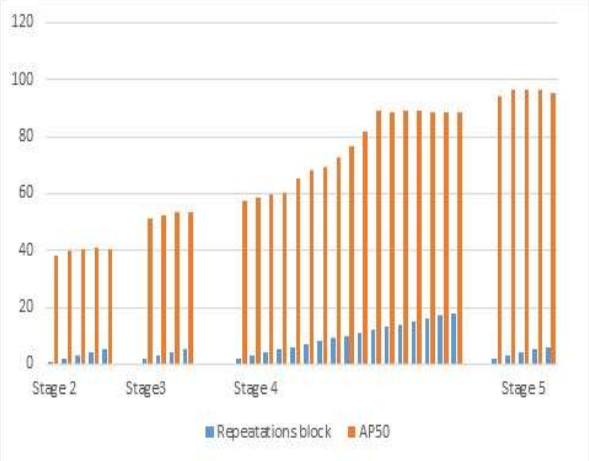

Figure 3.1. Result Of Resnet-101 Backbone

3.4 Evaluation results of the enhanced multiple food detection method with different state-of-the-art methods

A comparison was conducted between the performance of the proposed method with the state-of-the-art techniques, namely Mask, R-CNN, and CASCADE R-CNN; these methods were then trained and tested with various thresholds

amounting to 0.5, 0.75, 0.9. The comparison is shown in Table 1. The outcome indicates that the proposed method exhibits a superior solution.

Table 1. Multiple Food Object Performance With Different Thresholds (0.5, 0.75)

Methods	AP50	AP75	AP90
CASCADE R-CNN	94.41	83.71	71.02
Mask R-CNN	91.02	76.82	69.87
Proposed method	95.88	85.22	72.24

The proposed enhanced feature extraction for the multiple-food detection method produces the best result due to the developed ResNet-101 backbone by selecting duplicates of blocks. Furthermore, by adding the convolution layer.

The performance of the enhanced feature extraction for multiple-food detection method compared with another state-of-the-art method. Based on the result, the proposed enhanced feature extraction for multiple food detection methods has a better solution.

4. CONCLUSION

In this research, a method for multiple food image detection of the improved framework of Mask R-CNN was presented. To fulfill the objective of the study proposed a new method; the experiment results on multiple food datasets demonstrate that proposed work can significantly improve the detection efficiency of the multiple food images.

Hence, this study has successfully developed an enhanced ResNet-101 backbone for multiple food detection. The proposed framework shows superior performance in accuracy AP across various thresholds. The thresholds utilized in this study exhibited greater values than the state of art methods. Namely Mask R-CNN and CASCADE R-CNN.

4 ACKNOWLEDGMENT

This research is funded by the Deanship of Research and Graduate Studies in Zarqa University /Jordan. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved their research.

REFERENCES:

[1] A. E. Mesas, M. Muñoz-Pareja, E. López-García, and F. Rodríguez-Artalejo, "Selected eating behaviours and excess body weight: a systematic review," Obesity Reviews, vol. 13, no. 2, pp. 106–135, 2012.

[2] M. B. E. Livingstone and A. E. Black, "Markers of the validity of reported energy intake," The Journal of Nutrition, vol. 133, no. 3, pp. 895S–920S, 2003.

[3] K. Aizawa, Y. Maruyama, H. Li, and C. Morikawa, "Food balance estimation by using personal dietary tendencies in a multimedia food log," IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 2176–2185, 2013.

[4] S. Mezgec and B. Koroušić Seljak, "NutriNet: a deep learning food and drink image recognition system for dietary assessment," Nutrients, vol. 9, no. 7, p. 657, 2017.

[5] K. Takahashi, K. Doman, Y. Kawanishi et al., "Estimation of the attractiveness of food photography focusing on main ingredients," in Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in Conjunction with the 2017 International Joint Conference on Artificial Intelligence, pp. 1–6, Melbourne, Australia, August 2017.

[6] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, "Deepfood: deep learning-based food image recognition for computer-aided dietary assessment," in Proceedings of the International Conference on Smart Homes and Health Telematics, pp. 37–48, Wuhan, China, May 2016.

[7] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2249–2256, San Francisco, CA, USA, June 2010.

[8] R. Xu, L. Herranz, S. Jiang, S. Wang, X. Song, and R. Jain, "Geolocalized modeling for dish recognition," IEEE Transactions on Multimedia, vol. 17, no. 8, pp. 1187–1199, 2015.

[9] G. M. Farinella, D. Allegra, M. Moltisanti, F. Stanco, and S. Battiato, "Retrieval and classification of food images," Computers in Biology and Medicine, vol. 77, pp. 23–

39, 2016.

[10] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, "Multi-temporal ultra dense memory network for video super-resolution," IEEE Transactions on Circuits and Systems for Video Technology.

[11] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: a generative adversarial network for infrared and visible image fusion," Information Fusion, vol. 48, pp. 11–26, 2019.

[12] J. Ma, X. Wang, and J. Jiang, "Image superresolution via dense discriminative network," IEEE Transactions on Industrial Electronics, vol. 67, no. 7, pp. 5687–5695, 2020.

[13] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in Proceedings of the 22nd ACM International Conference on Multimedia-MM'14, pp. 1085–1088, Mountain View, CA, USA, June 2014.

[14] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, "Food image recognition using very deep convolutional networks," in Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management-MADiMa '16, pp. 41–49, Amsterdam, Netherlands, October 2016.

[15] A. Singla, L. Yuan, and T. Ebrahimi, "Food/non-food image classification and food categorization using pre-trained googlenet model," in Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management-MADiMa '16, pp. 3–11, Amsterdam, Netherlands, October 2016.

[16] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm detector: a novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 7, pp. 5146–5158, 2019.

[17] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: an efficient framework for geospatial object detection," IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 2, pp. 302–306, 2019.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, Columbus, OH, USA, June 2014.

[19] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448, Santiago, Chile, December 2015.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in Proceedings of the Advances in Neural Information Processing Systems, pp. 91–99, Montreal, Canada, December 2015.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969, Venice, Italy, October 2017.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788, Las Vegas, NV, USA, June 2016.

[23] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271, Honolulu, HI, USA, July 2017.

[24] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, http://arxiv.org/abs/1804.02767.

[25] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in Computer Vision—ECCV 2016, pp. 21–37, Springer, Cham, Switzerland, 2016.

[26] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in Computer Vision—ECCV 2014, pp. 740–755, Springer, Cham, Switzerland, 2014.

[27] Abuowaida, S. F. A., Chan, H. Y., Alshdaifat, N. F. F., & Abualigah, L. (2021). A novel instance segmentation algorithm based on improved deep learning algorithm for multi-object images. Jordanian J Comput Inf Technol

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(JJCIT), 7(01), 10-5455.

[28] S. F. Abuowaida and H. Y. Chan, "Improved deep learning architecture for depth estimation from single image," Jordanian Journal of Computers and Information Technology, vol. 6, no. 4, 2020.

[29] Alshdaifat, N., Osman, M. A., & Talib, A. Z. (2022). An improved multi-object instance segmentation based on deep learning. Kuwait Journal of Science, 49(2).

[30] S. F. A. Abuowaida and H. Y. Chan, "Improved deep learning framework for multi food instance segmentations," International Journal, vol. 9, no. 4, 2020.

[31] Mohammed Alghaili,Zhiyong Li,Ahmed AlBdairi,Malasy Katiyalath, "Generating Embedding Features Using Deep Learning for Ethnics Recognition", The International Arab Journal of Information Technology (IAJIT) ,Volume 20, Number 04, pp. 669 - 677, July 2023, doi: 10.34028/iajit/20/4/13.

[32] Nouh Alhindawi, "IoT Based Technique for Network Packet Analyzer", The International Arab Journal of Information Technology (IAJIT) ,Volume 20, Number 04, pp. 678 - 685, July 2023, doi: 10.34028/iajit/20/4/14.