ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse129952020.pdf
https://doi.org/10.30534/ijatcse/2020/129952020

Semantic Parsing for Automatic Retail Food Image Recognition

Sunusi Bala Abdullahi[1], Kanikar Muangchoo[2*]
[1]Department of Physics, Faculty of Physical Sciences, Bayero University Kano, Nigeria, sbabdullahi@ieee.org
[2]Faculty of Science and Technology, Rajamangala University of Technology Phra Nakhon (RMUTP), 1381, Pracharat 1 Road, Wongsawang, Bang Sue, Bangkok 10800, Thailand, kanikar.m@rmutp.ac.th

ABSTRACT

Explosion of the world of computer vision, pave the way to visual recognition which is now extended to visual food and metadata recognition. Physical activities disruption due to Covid-19 pandemic rapidly increases the online food order. Online Customers relies on available good quality food images and metadata information, to make decision of the type of food to order. Recently developed deep learning networks outperform the classical approaches for visual food recognition, however these networks thirsts for large datasets to achieve best performance. Majority population of present online customers order food to small-scale restaurants, these restaurants deals with small datasets, thus restricted them to take advantage of modern tools and to participate in the billion-digital business. In this work, we modified the existing deep-CNN architecture to fit the small-scale restaurant dataset and trained on an end-to-end DeepLab v3+ initialized from ResNet weight. Our proposed novel architecture is designed by exploiting the output of multiscale contextual information of CNN encoder and fed in the low-level features of our constructed Resnet-18 as the backbone network, and finally fine-tuned with simple filters, and bilinear interpolation on the order factor by 4. This approach reduces the serious overfitting of the deep-CNN. The metadata recognition was done using enhanced-OCR, where the segmented image was analyzed at high-level layers. The accuracy of our method is reported using IoU and BF score. The numerical validation of the method is carried out on ETH-food-101 dataset and it demonstrates compliance to the state-of-the-art performance.

Key words: CNN, DeepLab v3+, Food Recognition, Resnet-18, small-scale restaurant, Semantic Segmentation.

1. INTRODUCTION

The rapid increase by internet of everything and the explosion world of computer vision, shapes the standard restaurants to maintain an online presence. However, COVID-19 pandemic disruption to physical activities, leaves the public with only one option to order food online. This substantially increases the global annual sale average of food online ordering to more than 20% and the figure from U.S. 35 billion is predicted to reach U.S. 365 billion by year 2030 [Deloitte, Forbes, 2020]. This great benefit only impacts the standard restaurants with few local restaurants, whereas the small-scale restaurants are diminishing and counting loss. The survival of small-scale restaurants depends on the extent they have migrate to the online marketing. Appearance and customization of food images are the key factors to facilitate customer's decision. Customers search for where and what to eat [1], customers examine the quantity of per-meal nutrition from the food metadata image to control obesity [2]. Customers and tourists are finding it very difficult to order online food without the available good quality food images and metadata information. Therefore, customers need good quality food image.

Computer vision system, as a novel promising technology can spontaneously using intelligent techniques visualize and recognize food images. Various promising literatures tries to address the problems of food recognition at large datasets, but food image recognition using small datasets from small-scale restaurants remains limited.

Presently, relevant technocrats and researchers extensively worry and research on online food retailing algorithms, which have wide range of application platforms and potential commercial value [3], however these algorithms led to inaccurate and geometrical food information deformation, on the retailers websites, leaving customers to adopt various manual-intensive ways for acquiring information [4]. the current Covid-19 pandemic and the digital divide, in addition to the setting procedures promulgated by many countries to stop the spread of covid-19 and to confirm the social distancing, this method of online food ordering is very ineffective, thus, there is need for improvement and small-scale restaurants lacks standard architecture for their operations. However, this limit their marketing strategy and restrict them to participate on the current virtual marketing, and to harvest the benefits of 20th century as well.

The relevant scholars try to solve the problem of food image recognition using hand crafted features [5], this method is prone to information lost and misclassification. Existing deep learning schemes such as [1] put forward relevant algorithms, to simultaneously model and recognizes food images, meal

names and restaurant identities, to overcome the anomaly of the binary correlation between food images and dish labels. This approach achieved recognition accuracy using an architecture coded PAMT-CNN. But validated on large-scale dataset, this approach cannot be extended to small-scale restaurant datasets. Reference [6] proposes an automatic CV to separate foreground and background from complex product image, overcome the handcrafted techniques that hinders the recognition accuracy. The method is limited to retail product photography images. Document [7] schemes deep architecture that analyze the food composition with specific focus on the vertical food characteristics, to overcome the dimensionality issues of deep architecture in food classification problem. This architecture is validated on large datasets, which restrict it application on small-scale datasets. Reference [8] takes into account distinctive spatial layout and common semantic patterns. This approach fused three aggregated features into unified representation to describe the food images and extract the semantics of food images with high statistical score. Deep architecture makes the aggregated features highly discriminative and invariable to geometrical deformation. However, the method gives deep representation to the food image and visual recognition, but the method could be restricted to the large dataset. A simple visual food representation model can be demonstrated by the three most basic traits, food image, dish name, and restaurant identity. Recently, number of researches studied the nutritional facts, which only accounts for the recognition of metadata information [4],[9],[10] but the approaches did not accounts both for food image and restaurant identity, as these provide the customers basis for online food order decisions. Our method will develop this strategy of multi-scale fusion.

Methods based on Convolutional Neural Networks (CNN) and Deep-CNN, partial handcrafted features, post-processing, as well as having difficulty to learn and segment the most important intelligence from the food images. It hinders the credible segmentation efficiency. Reference [11] reported that ResNet poorly performs compared to Ensemble learning due to small datasets. Therefore, we proposed a Semantic Segmentation, it labeled each pixel in the image [12] into various object classes, and the DeepLab v3+ model give the location of the encoded pixel from the image for segmentation. This to the best of our knowledge the first paper to address the problem of food recognition using semantic segmentation validated on small datasets.

DeepLabv3+ is developed and invented by the Google open source to solve the individual limitations of DeepLabv1 and DeepLabv2 respectively, for semantic image segmentation. DeepLabv3+ comprised of encoding and decoding phase, respectively. Encoding phase exploits engrained information from the food images by employing a convolutional neural network (CNN), considering decoding phase, reimagined the high dimensionality output by using information acquired through the encoding phase. However, DeepLab v3+ supports various network backbones, but for the purpose of this paper, we initialized this network using the weights of ResNet-18. ResNet-18 is reported [13], as productive tool for applications with limited processing resources, as well as performance for small datasets. This our work used the dataset of food images provided by the ETH food-101, we automatically recognized the food images along with the metadata information, to overcome the loss of essential information due to pre-segmentations and handcraft approach.

1.1 Proposed Architecture for Semantic Segmentation

As mentioned earlier our architecture is developed based on deep-CNN and initialized using ResNet-18 as the backbone network. The architecture comprised of multi-layers in encoder-decoder fashion [14]-[15]. Assuming *M* to be the convolutional layers if any layer *m* convolutes a nonlinear representation W_m for any given input I^{m-1}, it demonstrates an output layer of $I^{(m)}$. The CNN constructed the nonlinear representation as:

$$W = W_m \emptyset W_{m-1} \emptyset \dots\dots\dots \emptyset W_1 \quad (1)$$

$\emptyset$ show the series of birepresentations (bilinear).

$$A^m = K^m * I^{m-1} + G^m, \qquad G^m \in \mathbb{R}^{X_m \times Y_m \times Z_m} \quad (2)$$

$$I^m = W_m(I^{m-1}) = \rho(A^m), \quad 1 \le m \le M \quad (3)$$

Equation (2) demonstrate the activation function of the convolution A^m, with 4D tensors $K^m \in \mathbb{R}^{S_m \times T_m \times U_m \times Z_m}$, Where the first two dimensions denotes spatial dimensions, the third denotes the input feature dimension and the final dimension is the output feature. Furthermore, the equation has bias of convolutional layer G^m, and it set of real numbers are the 3D tensors. However, the first and second dimensions of the tensors; are the spatial dimensions, while the third dimension gives the feature. Whereas equation (3), illustrates the output of the $m-th$ convolutional layer. Furthermore, we can now arguably map the convolution between the two tensors (4), where the sum of 2D input features are convoluted with a standard stride of 1. Since semantic segmentation is the interest of this paper, in general input image $I \in \mathbb{R}^{X \times Y \times 3}$ through the learning process will map an output *E*. the parameters of this network can be put together in the vector $\vec{\varphi}$ [14].

$$(K * I)_{xyz} = \sum_{h=1}^{S} \sum_{i=1}^{T} \sum_{j=1}^{U} K_{hijz} I_{x+h-1,y+i-1,j,} \quad (4)$$

The output of the CNN network only contain set of segmented representation $E^c \in \mathbb{R}^{X \times Y} \ (1 \le c \le S)$, where *c* and *S* denotes the image class and set of segmented image respectively. Intuitively, we can develop our network using the general case:

$$I^m = E \in \mathbb{R}^{X \times Y \times S} \quad (5)$$

The output from the CNN network E^c for the $s-th$ segmentation map, maps the input image I to E^c using W^c, can be demonstrated as:

$$E^c = W^c(I) \; \forall \; (1 \le c \le S) \tag{6}$$

The network is trained using a set of V food images $I(h) \in \mathbb{R}^{X \times Y \times 3}$ which are automatically aligned with the labeled segmentation results (ground truth):

Corollary 1:
$E(h) \in \mathbb{R}^{X \times Y \times S}, (1 \le h \le V)$

$\forall \; E(h)$ at c image class, includes S ground truth segmentation representations:
Corollary 2:

$E^c \in \mathbb{R}^{X \times Y}$

The network training is done by making $\vec{\varphi}$ adaptive as illustrates in (7):

$$\vec{\varphi} = \underset{\varphi}{\operatorname{argmin}} \sum_{h=1}^{V} V_{cat}\,(W(I(h), E(h)) \tag{7}$$

The V_{cat} can be defined as the cross entropy between two segmentation maps [14] $S, T \in \mathbb{R}^{X \times Y \times S}$:

$$V_{cat}\,(T,S) = -\sum_{x=1}^{X}\sum_{y=1}^{Y}\sum_{c=1}^{C} S_{xyc} \log T_{xyc} \tag{8}$$

To overcome the problem of gradients vanishing in this deep architecture, and to take the advantage of having multi-scale contextual information, this our work proposed DeepLabv3+ as adopted in [16], [17],[15]. DeepLabv3+ is a special case of DeepLabv3 with encoder-decoder layers. This approach exploits the advantage of rich semantic information from the Deeplabv3 [16], in addition robust decoder module is fused to the approach to reconstruct the missing object boundaries because of the striding operations within the network backbone [15]. Atrous convolution is also added to control the density of the encoder features, which relies upon the computational resources.

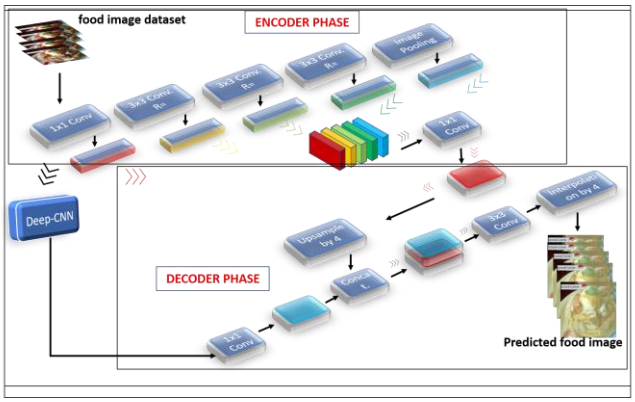


Figure 1: Architecture of the DeepLabv3+

2. METHODS

The food images are two-dimensional, therefore equation (6) can be customized; *c* location from the output representation *E*, where we applied convolutional filter *K* on top of the input representations *I*, the stride *R* at the input discover the sampling input, whereas at the output, stride at the last two blocks was set to 2 and 4 in atrous fashion respectively. The encoder and decoder design is adopted according to [15],[17]. As reported, atrous convolution at this our fashion overcome the computational complexity remarkably, and however it demonstrates superb performance of our architecture. In our decoding section, as shown in Figure 1 a simple and suitable decoder is chosen to fulfil our hypothesis. The decoder is deployed on top of the encoding output; where the 1x1 convolution is injected and achieved limited channels, the 3x3 convolution injected gives distinct segmentation. Equation (9) details the customization.
For details read [16], [17],[15].

$$E^c = \sum_{f} I(C + R.f)K(f) \tag{9}$$

However, feature map [1 x 1, 32] was used for Conv2 before striding in ResNet-18 residual block, thus reduces the channels of the low-level feature map from the encoding phase to 32, the decoding strategy is the same in [17], but only the values on the parameters are different. Equation (10) details the residual operations,

The high-level layer is now fused together with maximally stable extremal regions (MSER) layer. MSER approach is leverages as image blob detection. The advantage of using MSER as text detection approach is that the regions are exclusively considered by the intensity function of the region and the external boundary [18]. This property can be defined by mapping set of food images, V :
Corollary 3:
$V : L \subset \mathbb{Z}^2 \to K$

Must fulfil the following conditions:
K should be total, antisymmetric, and transitive binary relations $\leq$ exists. Adjacency relation of set L is conjugated as $R \subset L \times L$.
Corollary 4:
For achieving maximally stable extremal region, a sequence of ordered sets collection of extremal regions $O_n \subset O_{n+1}$
Can achieve stability $\forall \; o(n) = \; |O_{n+\Delta} \backslash O_{n-\Delta}|/|O_n|$ at local minimum n. With $n \in K$ and $n + \Delta \in K$. Where O_n, n, $|..|$, and $\Delta \in K$, denotes threshold region, intensity value, cardinality and method parameter.
The text regions are computed from the automatic segmented image in our trained network as shown in Figure 2. Corollary 3 and 4 are used to define the properties of regions on the segmented food image. We computed the region properties with the aid of bounding box (bbox), where best aspect ratio was chosen for the boxes. Food images are carefully analyzed

to detect only bounding boxes that contained text information, as well to remove all other boxes that contained food image or plates due to those areas contained dominant pixels than the text areas. Equations (10) and (11) compromised the process. Where (10) remove non-text regions along the horizontal planes, and (11) remove the non-text regions along the vertical planes.

$$O_x = \frac{\sum_n |O_{n+\Delta} \backslash O_{n-\Delta}| / |O_n|}{|O_n|} \tag{10}$$

$$O_y = \sum_{n=1}^{n} |O_{n+\Delta} \backslash O_{n-\Delta}| + \frac{\sum_n |O_{n+\Delta} \backslash O_{n-\Delta}| / |O_n|}{|O_n|} \tag{11}$$

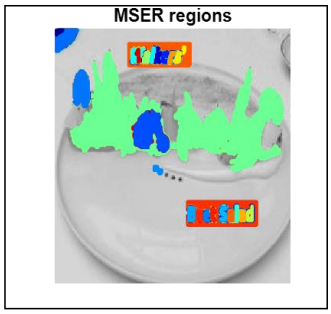


Figure 2: Segmented Food Image bounded using MSER

The Figure in 2, illustrates the output of the filtered image, where the non-text regions are automatically filtered from the frame using geometrical properties, however still the image suffers misclassification due to the presence of noise to pixels at regions. Figure 3 illustrates the result of expanded boxes in the image to automatically concentrate on the text boxes and ignore all other boxes.

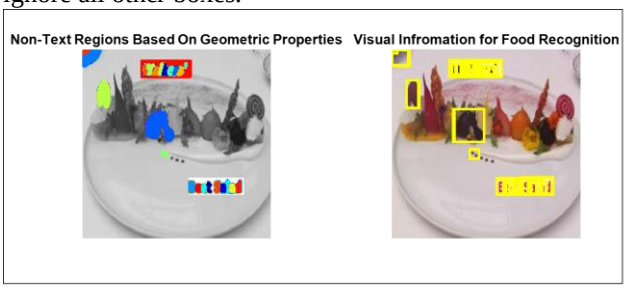


Figure 3a: Enhancement to remove non-text regions,
3b: Expansion of Bboxes for accuracy.

Therefore, we applied the overlap ratio in Figure 3b between the bounding boxes and it was set to zero, to demonstrate the best graph. The final output of this operation is illustrated in Figure 4. This was after the complex background was remove.

Figure 4: Final complete recognized food image

Furthermore, the accuracy of our text detection within the food dish was confirm using optical character recognition (OCR) within each considered box. Confidence score values was used as the evaluation metrics in our text recognition. Figure 5 below illustrates the output from the OCR.

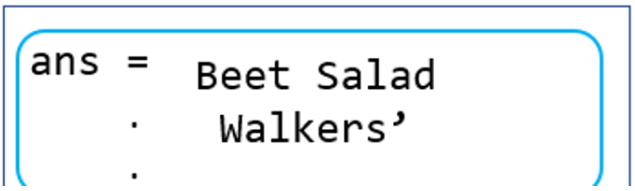


Figure 5: OCR text recognition.

The deepLabv3+ network is constructed using the MATLAB (R2019b) command "lgraph" and "resnet18" respectively, for semantic segmentation with atrous convolution. The proposed network model is actualized on the Food-101 dataset consists of food images from food spotting. The data consisted of 101,000 food images from 101 classes, obtained at http://www.vision.ee.ethz.ch/datasets/food-101/, https://www.research-collection.ethz.ch/ [23].

We randomly selected only 99 different food images and established three standard classes. Our selected data set was partitioned into three parts, 60% of the data were randomly selected for calibration phase *(imdsTrain*), 20% of the data for verification (*imdsVer*), and the remaining 20% for the validation (*imdsVal*) respectively. We modified the food images with food names, and restaurant identities illustrated in Figure 6. These modified food images in Figure 7, were automatically labeled using Matlab image labeler algorithm2. The new food images have unequal number of cognizance due to the meal cover more spaces on the image, however it has many pixels than the food name, and restaurant identity. Therefore, the imbalance is undesirable to learning process and poses challenges of biased. Class weighting is adopted to overcome the class imbalance shown in Figure 8. Datastore is created for the entire food images, details in algorithm3. The food images are augmented using "imageDataAugmenter" as it improves the accuracy of our architecture, we randomly selected left/right reflection and X/Y translation of +/- 10 pixels.

The image patches are cropped to 224 throughout the learning process. A simple and effective learning rate of 0.001 is used

from formula [17],[19]-[20], with drop factor of 0.3. mini batch size of 8 is employed, this is maintained for both the upsampling and downsampling respectively. Our architecture is optimized using stochastic gradient descent with momentum of technique of 0.9, where best hyper-parameters are selected. Therefore, performance of this architecture is utilized using Intersection Over Union (IoU) as the metrics of evaluation. The training of the multi-scale fusion deep-CNN is details in algorithm4.

Figure 6: Constructed food image

Figure 7: Constructed food label images.

Figure 8: Test on segmented image.

3. RESEARCH ENVIRONMENT

We have conducted our experiment on PC laptop with 4 GB of RAM equipped with NVIDIA GTX GEFORCE 1660, core i7 Intel core, 9th generation.

Algorithm 1 Create Folder

```
Root            //create MATLAB path
    Sub 1       //my folder
        Sub i. create +vision folder
            Sub a. +labeler folder
    Sub 2. Create directory with sub 1 + sub i.   //make directory
```

Return Folder

Algorithm 2 Image Labeling Algorithm

Input:

$J = \{1,2,3\ldots., N\}$ // the first j+1 image set all food image sequences.

Length // length of image sequence

Output:

K // labels acquired automatically in sequence display

For each $j \in [1, N+1]$ **do**

Create one label definition
then hit the app toolstrip
If $j <$ length+1 **then**
end if
Select algorithm

For checkLabelDefinition ()

do automate the session
if checkSetup true **then**
initialize + run automatically on each frame
else if return false
else
terminate
end if
end if

end for

Return K

Algorithm 3 Image Datastore

Input:

Food_images

J = {1,2, 3…., N} //the first j+1 image set all food image sequences.

Length //length of image sequence

Output:

Q //image datastore

Initialize the sequence and frames ***K*** produced by Algorithm 2

1. **if** N is the food image then return K [1, N+1] **do**
if N is list then {
Q = []
For each (n) in K:
add store (imds (N), n, n+1) to Q
Return imagedatastore Q with depth n
}
If N is double then {
Q = []
For each (ni,) in (K):
Add store (ni, n) to Q
Return Q

}

end if

end if

Return Q

Algorithm 4 DeepLabv3+ Algorithm

Inputs: A (max num classes, max num epochs, num filters in layers (1…n-layer), num strides for all layers), I (image size), J (atrous Conv2), H (Resnet18), N (image datastore, l, m, n), Q (pixel datastore, p, o, q), R (deep-CNN training options), V (food images).

Output: mIoU J // Predicted labeled food image

Y = Set A // Set Network blocks structure.

N = Partition Food Images (L, M, N, O, P, Q) // Analyze Network graph.

If Q is a datastore, then partition (N, Q)

else // partition the pixels of food images V into calibration, validation,
// and verification and store to Q as well.

For (each pixel datastore, E of Q), calibrate (E);

Partition (N, Q)

end if

end for

∂ = deeplabv3plusLayers (I, A, K) // Create DeepLab v3+
W = Call R // Call encoder and Define training options + Validation Patience

Repeat:

Compute (8) // Calculate the loss function.

P_1 = Compute (9) // realize output feature map.

∂_1 = Compute semantic representation at the higher layer and fine tune using (8)

X = imageDataAugmenter (Translation [I]) // Augmentation to food image at a/b.

∂_2 = pixelLabelImageDatastore (L, O, X) // training preparation.

∂_3 = Call H // Construct the backbone network.

P_2 = Compute (4) with a stride of 1. // This will compute low-level features
// to reduce the number of channels.

∂_4 = Compute parameter Ø from (2) and (3), and fine tune by factor of 4 // refine feature map
// and interpolated by factor of 4.

Compute $\vec{\varphi}$ as in (7) with (∂_1, ∂, W) // Network training.

Do Train $\vec{\varphi}$ = false;

If do train $\vec{\varphi}$

$(\partial_2, info) = \vec{\varphi}$

else

Compute (6) and update $\vec{\varphi}$ // Load Pretrained Network.

Net = data.net.

end

Pretrained network = full file (V, ∂)

end for

Initialize the parameter ∂_1 with bounding boxes.

Perform MSER (Q, N, props) // Measure the region properties using bounding box.

S = Metrics // Evaluation of semantic segmentation.

Compute (10) and (11) adaptively //determine the regions to be removed.

Perform MSER Stats (Q) // Using geometrical properties to remove non-text regions.

Perform bbox expansion (Q, N) // This will clip the bboxes around the image bounds.

Call graph (overlap ratio) // This will create graph.

Apply graph (Q, N).

OCR (Q, N) // This will show the final image detected.

Compute recognition confidence (D)

Compute (10) // Calculate the mean Intersection over union

Until: (8) and (10) convene

Return J // Compute the intersection over union.

3.1 Validation Phase

Evaluation Metrics

Intersection over Union (IoU):

$$IoU(F_i, F_l) = \frac{|F_i \cap F_l|}{|F_i \cap F_l|} \quad (12)$$

Where F_i the set of prediction pixels, and F_l the set of food ground truth pixels, Ո and U denotes intersection and union operations, respectively. |... | denotes calculating the number of pixels in the set. Moreover, boundary F-1 score (BF score) is also evaluated to add more confidence and reliability to our proposed algorithms as adopted in [21].

4. RESULTS AND DISCUSSION

The results of the semantic parsing using DeepLab v3+ on encoder-decoder fashion network are compared with several existing studies as Table 2. Calibration with a residual depth of 18 as specified in algorithm 4 is designed. Table 1 compare the performance of conventional method which recognized the visual food using three basic traits: meal image, restaurant identity, dish name, respectively. Our proposed architecture outperforms the existing PAMT-CNN, both in food recognition and text detection. Appendix demonstrate the ability of our proposed architecture, when validated on the selected dataset achieved an accuracy of 79.0%, without serious violations of overfitting. When the architecture is evaluated using equation (12) demonstrates mean IoU of 0.9250, weighted IoU of 0.8950, mean of boundary F-1 score (BF score) 0.8333 and mean accuracy of 0.9132, respectively. Whereas the MSER achieved character extraction confidence score of 0.9810.

Table 1: Performance Accuracy for the overall classification

Global Accuracy	Mean Accuracy	Mean IoU	Mean BF Score	Text Confidence Score
0.9302	0.9132	0.9250	0.8333	0.9810

Table 2: Performance of some existing works on the Food-101 dataset according to the Top-1 accuracy. While our approach is evaluated according to the Global accuracy. The best results are bolded.

Method	Dataset		Accuracy (%)
	Classes	No. of sample/class	
Ensemble Net	101	750	72.12
WISeR [7]	101	1000	90.27
PAMT-CNN [1]	100	240	74.87
CNN_5 [2]	100	100+	**97.12**
MSMVFA(ResNet-152) [8]	101	1000	90.37
Semantic Parsing	**3**	**99**	**93.02**

Table 3: Comparison of our proposed model with conventional method on the Food-101 dataset [20]. The best results are bolded.

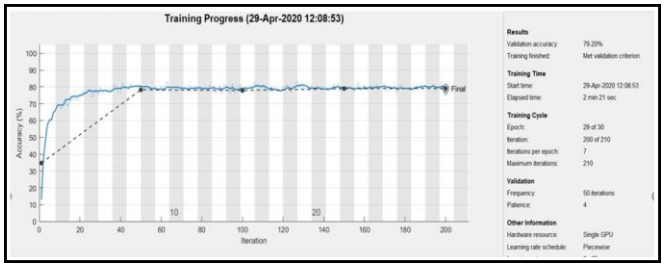
Method	Accuracy of the dish (%)	Accuracy of the Restaurant identity (%)	No. of samples per class	No. of Classes
PAMT-CNN	74.87	64.75	240	100
Semantic Parsing	**79.20**	**98.10**	**33**	**3**

5. CONCLUSION

This paper extends the application of deep-CNN semantic segmentation to the visual food retail recognition by introducing a simple and effective residual connection as the network backbone, to bridge the gap between the small-scale restaurants and the online food retailing business. The best training hyper-parameters were chosen. The major advantage of this our proposed architecture is the absolute control of output decisions of contextual features from the CNN layers, which reported impossible in the previous models [12]. Adopting ResNet-18 as the decoder module produce better performance to the small datasets, which was difficult to achieve by the traditional ResNet architectures. Finally, when tested on ETH food-101 dataset the model achieves 79.0% at 99 images for three classes, without further post-processing and manual processing. To exploit the full advantage of online visual food recognition and make it mobile system, we need to include the GPS coordinates of restaurant, is a development direction of future research.

APPENDIX

The following is the snapshot of the training progress of the data.



ACKNOWLEDGEMENT

The second author was financially supported by Rajamangala University of Technology Phra Nakhon (RMUTP) Research Scholarship.

REFERENCES

[1] H. Wang, W. Min, X. Li, and S. Jiang, "**Where and what to eat: Simultaneous restaurant and dish recognition from food image**," 2016. doi: 10.1007/978-3-319-48890-5_51.

[2] J. Teng, D. Zhang, D. J. Lee, and Y. Chou, "**Recognition of Chinese food using convolutional neural network,"** *Multimed. Tools Appl.*, vol. 78, no. 9, pp. 11155–11172, 2019, doi: 10.1007/s11042-018-6695-9.

[3] S. Kim, E. Park, and D. Lamb, "**Extraordinary or ordinary? Food tourism motivations of Japanese domestic noodle tourists.**," *Tour. Manag. Perspect.*, vol. 29, no. 2019, pp. 176–186, 2019.

[4] V. Gundimeda, R. S. Murali, R. Joseph, and N. N. T. Babu, "**An Automated Computer Vision System for Extraction of Retail Food Product Metadata Venugopal,**" in *First International Conference on Artificial Intelligence and Cognitive Computing*, 2019, pp. 199–216, doi: 10.1007/978-981-13-1580-0.

[5] L. Bossard, M. Guillaumin, and L. Van Gool, "**Food-101 - Mining discriminative components with random forests**," in *Proceeding of European Conference on Computer Vision*, 2014, vol. 8694, no. 6, pp. 446–461, doi: 10.1007/978-3-319-10599-4_29.

[6] R. Joseph, N. T. N. Babu, R. S. Murali, and V. Gundimeda, "**Automatic Retail Product Image Enhancement and Background Removal,**" in *First International Conference on Artificial Intelligence and Cognitive Computing*, 2019, pp. 1–15, doi: 10.1007/978-981-13-1580-0.

[7] N. Martinel, G. L. Foresti, and C. Micheloni, "**Wide-Slice Residual Networks for Food Recognition,**" in *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, vol. 1, pp. 567–576, doi: 10.1109/WACV.2018.00068.

[8] S. Jiang, W. Min, L. Liu, and Z. Luo, "**Multi-Scale Multi-View Deep Feature Aggregation for Food Recognition**," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 265–276, 2020.

[9] J. Chen and C. W. Ngo, "**Deep-based ingredient recognition for cooking recipe retrieval**," in *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*, 2016, pp. 32–41, doi: 10.1145/2964284.2964315.

[10] A. Meyers *et al.*, "**Im2calories: Towards an automated mobile vision food diary.** Proceedings of the IEEE International Conference on Computer Vision," in *International Conference on Computer Vision, Iccv*, 2016, no. 2, pp. 1233–1241.

[11] P. Pandey, A. Deepthi, B. Mandal, and N. B. Puhan, **"FoodNet: Recognizing Foods Using Ensemble of Deep Networks**," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1758–1762, 2017, doi: 10.1109/LSP.2017.2758862.

[12] C. Liang-Chieh, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "**Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation,**" in *Proceedings of the European Conference on Computer Vision, ECCV 2018*, 2018, pp. 801–818.

[13] M. Shu, "**Deep learning for image classification on very small datasets using transfer learning,**" MSc Dissertation, Department of Electrical and Computer Science, Iowa State University, 2019.

[14] L. Mauch, C. Wang, and B. Yang, "**Subset selection for visualization of relevant image fractions for deep learning based semantic image segmentation**," *J. Franklin Inst.*, vol. 355, no. 4, pp. 1931–1944, 2018, doi: 10.1016/j.jfranklin.2017.08.001.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "**Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation**," in *European Conference on Computer Vision, ECCV*, 2018, vol. 84, no. 1, pp. 801–818.

[16] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "**Feedforward semantic segmentation with zoom-out features,**" in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 1, pp. 3376–3385, doi: 10.1109/CVPR.2015.7298959.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "**Rethinking Atrous Convolution for Semantic Image Segmentation,**" *Conference on Computer Vision and Pattern Recognition*, 2017. http://arxiv.org/abs/1706.05587 (accessed Aug. 12, 2020).

[18] I.-S. Oh, J. Lee, and A. Majumder, "**Multi-scale Image Segmentation Using MSER**," in *15th International Conference on Computer Analysis of Images and Patterns*, 2013, vol. 47, no. 2, pp. 201–208, doi: 10.1007/s10851-013-0451-6.

[19] D. Devani, A. Suryadibrata, and J.C. Young, **"Implementation of Siamese Conolutional Neural Network from Cell Images for Malaria Disease Identification**," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no.4, pp. 4195–4200, 2020, https://doi.org/10.30534/ijatcse/2020/04942020.

[20] I. Kich, E. Ameur, Y. Taouil and A. Benhfid, **"Image Steganography by Deep CNN Auto-Encoder Networks**," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no.4, pp. 4707–4716, 2020, https://doi.org/10.30534/ijatcse/2020/75942020.

[21] R. Thomas, and J.E. Judith, **"A Novel Ensemble Method for Detecting Outliers in Categorical Data**," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no.4, pp. 4947–4953, 2020, https://doi.org/10.30534/ijatcse/2020/108942020

[22] Deloite, **Create a crisis growth plan**: Lessons from the Pandemic on How to Breakdown organizational silos and optimize workforce potential. S. Hatfield. https://www.forbes.com/companies/deloitte/#757696

767e1f, access August, 2020.
[23] http://www.vision.ee.ethz.ch/datasets/food-101/, https://www.research-collection.ethz.ch/