Thai Fast Food Image Classification Using Deep Learning

Narit Hnoohom and Sumeth Yuenyong
Image, Information and Intelligence Laboratory,
Department of Computer Engineering,
Faculty of Engineering,
Mahidol University, Thailand
narit.hno@mahidol.ac.th, sumeth.yue@mahidol.ac.th

***Abstract*—This paper presents a prediction model for classifying Thai fast food images. The model uses a deep learning process that was trained on natural images (GoogLeNet dataset) and was fine-tuned to generate the predictive Thai fast food model. The researchers created a dataset, called the Thai Fast Food (TFF) dataset, which contained 3,960 images. The dataset was divided into eleven groups of food images comprised of omelet on rice, rice topped with stir-fried chicken and basil, barbecued red pork in sauce with rice, stewed pork leg on rice, Thai fried noodle, rice with curried chicken, steamed chicken with rice, shrimp-paste fried rice, fried noodle with pork in soy sauce and vegetables, wide rice noodles with vegetables and meat. The final group comprised dishes which are not members of the other ten groups listed (non-ten-types), but which exist among Thai fast food. The classification average accuracy on a separate test set shows that Thai fast food can be predicted at 88.33%.**

Keywords—Thai fast food images; deep learning; classification.

I. INTRODUCTION

Image recognition research had been advancing very rapidly, with new applications being proposed in various domains. One of these domain is the image recognition of foods, which is potentially very useful for monitoring of diets or estimating the amount of calories consumed. In this paper, the researchers propose a prediction model for classifying Thai fast food images based on a recently proposed state-of-the-art convolutional neural network for image recognition. We created our own training database that we call the Thai Fast Food (TFF) dataset, which contain approximately 4,000 images separated into eleven categories in which common Thai fast food dishes can be broadly grouped.

Studies on food recognition had been proposed in parallel with the recent rise of image recognition models. The work in [1] use the Overfeat [2] deep convolutional neural network to extract feature from images of food from the UEC-FOOD100 dataset (http://foodcam.mobi/dataset100.html) by removing the last softmax layer and using the output of the remaining network as features. They also combine these features with those extracted from hand-crafted color patches and RootHOG patches, coded into Fisher vector [3]. Two separate classifiers were used: SVM for the features extracted from the Overfeat network and AROW [4] for the hand-crafted features. The outputs of the classifiers were combined in the late fusion manner [5]. The report accuracies were 72.26% as the top-1 accuracy and 92.00% as the top-5 accuracy. The authors improved upon their own work in [6] by pre-training a deep convolutional neural network based on the AlexNet architecture [7] on the ImageNet2000 dataset and then fine-tuning on the UEC-FOOD100 and the UEC-FOOD256 datasets. They also combine the features extracted by the convolutional neural network with hand-crafted feature based on color patches and RootHOG patches like in their previous work. The reported accuracies were 77.35% top-1 and 94.58% top-5 for the UEC-FOOD100 dataset and 63.77% top-1 and 85.82% top-5 for the UEC-FOOD256 dataset.

The work in [8] used the inception deep convolutional network architecture [9] to address the problem of food classification. The authors also pre-trained the network on the ImageNet dataset and then fine-tune the network with UEC-FOOD100 food dataset. They reported accuracies of 76.3% top-1 and 94.6% top-5. However, unlike in [1, 6] they did not use any hand-crafted features. Similar approach was used in [10], but the authors used a newer version of the inception architecture call Inception V3 [11]. They also pre-train the network on the ImageNet dataset before fine-tuning on the food database. The reported accuracies for the UEC-FOOD100 were 81.45% top-1 and 97.25% top-5, for the UEC-FOOD256 they reported accuracies of 76.17% top-1 and 92.58% top-5.

In [12] the author focused on Thai foods in the THFOOD-50 dataset. They used an architecture proposed in their own work called Nu-InNet which is based on the inception architecture. The key difference between Nu-InNet and inception is that all large convolutions are replaced by 3×3 convolutions in order to bring the computational cost down enough to run well on a smartphone. Like in other works they pre-train their network on the ImageNet dataset and then fine-tune on the actual food database. They only reported top-1 accuracy which was 80.34% for the THFOOD-50 dataset.

This study explores the performance of TFF classification by fine-tuning the Google versions of Inception V3 trained on the ImageNet dataset. The TFF dataset consists of differing dishes, backgrounds, and locations. The initial analysis of the TFF images for these foods could then be conducted using the results as test images.

The remainder of the paper is organized as follows. First, a general introduction to the classification methods for food images is briefly described. In Section 2, a description of the methodology with the processes of classification is addressed. Experiment results and discussions are included in Section 3. Finally, the conclusions are presented in Section 4.

II. METHODOLOGY

The objective of the methodology is to classify the Thai fast food images. A block diagram of the methodology is shown in Figure 1. The method depicted consists of four processes: (1) Input image; (2) Pre-processing; (3) Deep learning; and (3) classification.

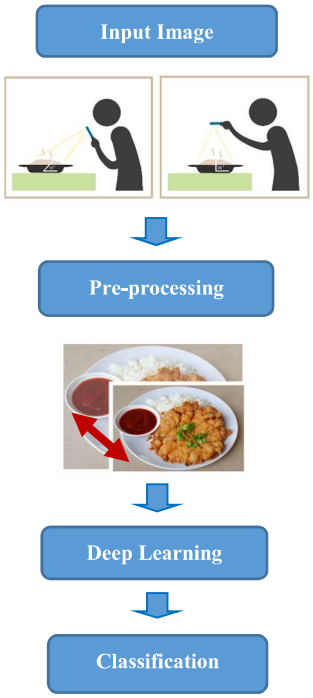

Fig. 1 Block diagram for the methodology.

A. Input Image

The TFF dataset was collected via smartphone by the researchers. We took 3,960 images of Thai fast food from different dishes, different backgrounds, and different locations. Figure 2 shows some examples of the food images.

The TFF dataset was divided into 11 groups, both for the training set and test set. The training set consists of 300 images of omelet on rice (Fig. 2 (a)), 300 images of rice topped with stir-fried chicken and basil (Fig. 2 (b)), 300 images of barbecued red pork in sauce with rice (Fig. 2 (c)), 300 images of stewed pork leg on rice (Fig. 2 (d)), 300 images of Thai fried noodle (Fig. 2 (e)), 300 images of rice with curried chicken (Fig. 2 (f)), 300 images of steamed chicken with rice (Fig. 2 (g)), 300 images of shrimp-paste fried rice (Fig. 2 (h)), 300 images of fried noodle with pork in soy sauce and vegetables (Fig. 2 (i)), 300 images of wide rice noodles with vegetables and meat (Fig. 2 (j)), and 300 images that do not belong to the other ten types (Fig. 2 (k)). For the test set, each group of food images is randomly selected from the TFF dataset. We reserved 660 images for testing, which were divided into the same 11 groups (60 images per group).

B. Pre-processing

In this step, the pre-processing prepared the TFF image by reducing the size from 1,334×1,000 to 256×256 pixels. This improved processing time while maintaining image detail. Moreover, the researchers used a histogram equalization to adjust the contrast of the TFF image for contrast enhancement.

C. Deep learning

From the perspective of researchers, the issue presents a eleven-class image classification problem, which differs from performing object detection to find the food since the exact position or size of the food in the image are not important. The most advanced image classification techniques today are reliant upon a deep learning approach. Deep learning requires a multi-layered neural network, while convolutional layers are necessary to support the recognition of images. The process of convolution involves sliding a mobile window across the image to determine the sum of the product under the window and the underlying image before down-sampling. The image passes through the layers, gradually becoming smaller, while the features, including lines and shapes, are extracted automatically. The classifier is the final layer, which is fully connected and works on the features which the convolution layers were able to extract. The eventual output generated by the network is a one-hot vector for which the output equals 1 for any class which the network can identify as the class of the input image. For any other class, the output will be zero.

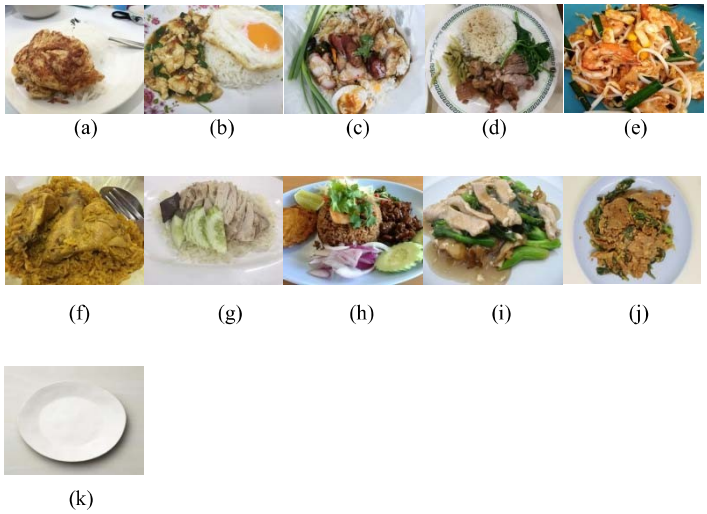

Fig. 2 The Thai fast foods used in the study.

It was not necessary to carry out pre-processing beyond the need to resize to 256×256 to ensure compatibility with the GoogLeNet neural network which was selected [13]. GoogLeNet is the latest state-of-the-art image recognition network, and comprises a number of convolution and max pooling layers in addition to several inception layers which serve as size 1 convolution with longer strides allowing image down-sampling with no loss of spatial information which would normally occur with max pooling. While GoogLeNet is too large

to be shown here in full, it involves the concatenation of "module" which can be seen in Figure 3. The concept is based on the distribution of the signal obtained from the previous layer to convolution of a different size. The result is then brought together to form a sparse structure. This operates in a manner similar to increasing the layer size, which is an established means of raising the representational capacity of a neural network while minimizing the growing computational load. The network is first initialized using weights which had undergone pre-training using the ImageNet database which comprises a vast number of natural images. In cases where limitations are imposed on the training data then fine-tuning can lead to better performance where the model has been trained using a greater dataset, even if the dataset is not similar to the target dataset. This approach is better than attempting to train a model from the starting point using a limited quantity of data, and it should be understood that for a model such as GoogLeNet which has a large number of parameters, 4,000 is considered a small number of images [14].

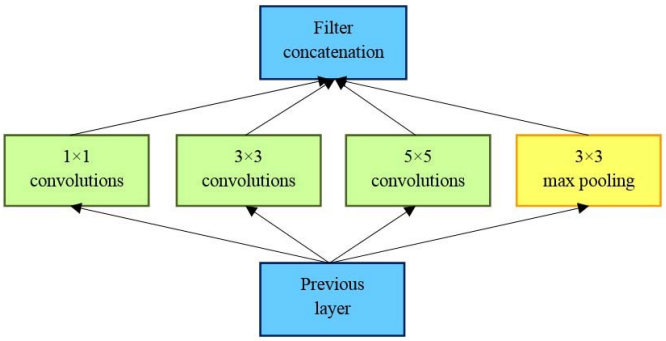

Fig. 3 Schematic of a GoogLeNet module.

D. Classification

The GoogLeNet framework [10] were used to classify the TFF images in this study. The model was fine-tuned using the Google versions of Inception V3 with a learning rate of 0.001 and a training step of 4,000. Once the training starts, GoogLeNet framework will plot training curves, loss curves and testing accuracies.

E. Evaluation

The accuracy values can be derived from the true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*) values which are given as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

III. Experimental Results

This section reports the experiment on the available TFF image datasets and evaluation results. The notebook computer used for the experiment has an i7-4700 CPU 2.4 GHz, 8 GB memory, and Microsoft Windows 8.1 Enterprise as the operating system. In order to classify the Thai fast food by using deep learning, the experimental results of classification on 3,300 images were used as training set, in which 300 images per group. For the test set, we reserved 660 images, which were divided into 60 images per group.

The result is summarized as a bar chart in Figure 4. It shows that out of 100% accuracy with a barbecued red pork in sauce with rice (Fig. 4 (d)), 100% accuracy with other non-ten-types exists on the TFF (Fig.4 (k)) were correctly classified as TFF present.

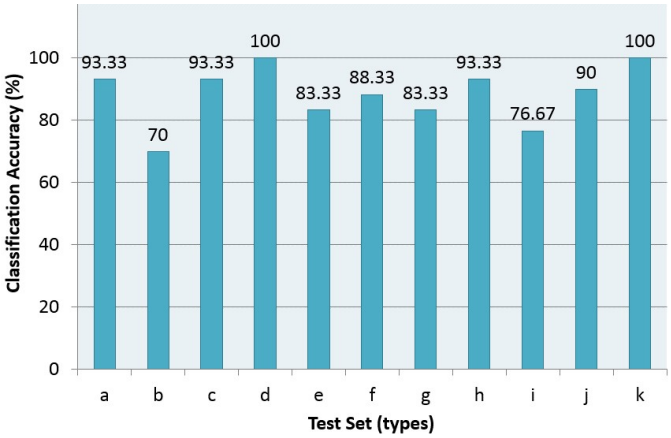

Fig. 4 Prediction results of Thai fast food.

For the negative class, the prediction model can classified 93.33% accuracy with an omelet on rice (Fig. 4 (a)), 70% with a rice topped with stir-fried chicken and basil (Fig. 4 (b)), 93.33% accuracy with barbecued red pork in sauce with rice (Fig. 4 (c)), 93.33% accuracy with a stewed pork leg on rice (Fig. 4 (e)), 83.33% accuracy with a Thai fried noodle (Fig. 4 (f)), 88.33% accuracy with a rice with curried chicken (Fig. 4 (g)), 83.33% accuracy with a steamed chicken with rice (Fig. 4 (h)), 93.33% accuracy with a shrimp-paste fried rice (Fig. 4 (h)), 76.67% accuracy fried noodle with pork in soy sauce and vegetables (Fig. 4 (i)), and 90% accuracy with a wide rice noodles with vegetables and meat (Fig. 4 (j)). From the data in the bar chart, the average of the accuracies of all 11 classes is 88.33%.

IV. Conclusion

This research study examined the use of deep learning algorithms in order to address the problem of TFF classification. The method employed TFF images obtained via smartphone to comprise the data which would serve to provide the researchers with a clear representation of TFF classes. The TFF dataset consisted of differing dishes, backgrounds, and locations. The initial analysis of the TFF images for these foods could then be conducted using the results as test images. By fine-tuning the Inception V3 trained on the ImageNet dataset, the researchers were able to achieve 88.33% classification average accuracy on the TFF dataset which contained 3,960 images. Further research should include a mobile application for real-time classification.

Acknowledgment

This work was supported by the Department of Computer Engineering, Faculty of Engineering, Mahidol University.

References

[1] Kawano, Y. and Yanai, K, "Food image recognition with deep convolutional features," Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, ACM, 2014.

[2] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus and Yann LeCun, "Overfeat: Integrated recognition, localizationand detection using convolutional networks," arXiv preprint arXiv:1312.6229, 2013.

[3] Perronnin, F., and Dance, C., "Fisher kernels on visual vocabularies for image categorization," IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07, 2007.

[4] Crammer, K., Kulesza, A. and Dredze, M., "Adaptive regularization of weight vectors," Advances in neural information processing systems, 2009.

[5] Hilal Ergun, Yusuf Caglar Akyuz, Mustafa Sert and Jianquan Liu, "Early and Late Level Fusion of Deep Convolutional Neural Networks for Visual Concept Recognition," International Journal of Semantic Computing, pp. 379-397, 2016.

[6] Yanai, K. and Kawano, Y., "Food image recognition using deep convolutional network with pre-training and fine-tuning," IEEE International Conference on Multimedia and Expo Workshops (ICMEW), IEEE, 2015.

[7] Krizhevsky, A., Sutskever, L., and Hinton, G., "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, 2012.

[8] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma, "Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment," International Conference on Smart Homes and Health Telematics, Springer, 2016.

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, 2015.

[10] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, and Stefano Cagnoni, "Food image recognition using very deep convolutional networks," Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, ACM, 2016.

[11] C Szegedy, V Vanhoucke, S Ioffe, J Shlens and Z Wojna, "Rethinking the inception architecture for computer vision," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016.

[12] Termritthikun, C., & Kanprachar, S, "Accuracy improvement of Thai food image recognition using deep convolutional neural networks," IEEE transactions on Electrical Engineering Congress (iEECON), IEEE, 2017.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Computer Vision and Pattern Recognition (CVPR), 2015.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.