Where and What to Eat: Simultaneous Restaurant and Dish Recognition from Food Image

Huayang Wang, Weiqing Min, Xiangyang Li, and Shuqiang Jiang(✉)

Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
{huayang.wang,weiqing.min,xiangyang.li}@vipl.ict.ac.cn, sqjiang@ict.ac.cn

Abstract. This paper considers the problem of simultaneous restaurant and dish recognition from food images. Since the restaurants are known because of their some special dishes (e.g., the dish "hamburger" in the restaurant "KFC"), the dish semantics from the food image provides partial evidence for the restaurant identity. Therefore, instead of exploiting the binary correlation between food images and dish labels by existing work, we model food images, their dish names and restaurant information jointly, which is expected to enable novel applications, such as food image based restaurant visualization and recommendation. For solution, we propose a model, namely Partially Asymmetric Multi-Task Convolutional Neural Network (PAMT-CNN), which includes the dish pathway and the restaurant pathway to learn the dish semantics and the restaurant identity, respectively. Considering the dependence of the restaurant identity on the dish semantics, PAMT-CNN is capable of learning the restaurant's identity under the guidance of the dish pathway using partially asymmetric shared network architecture. To evaluate our model, we construct one food image dataset with 24,690 food images, 100 classes of restaurants and 100 classes of dishes. The evaluation results on this dataset have validated the effectiveness of the proposed approach.

Keywords: Dish recognition · Restaurant recognition · Multi-Task CNN

1 Introduction

Food tourism is to explore the food as the purpose of the tourism and has become one important part of tourism[1]. In food tourism, where (i.e., the restaurant) and what (i.e., the dish) to eat is a basic need among tourists. In addition, the visual appearance of meals is one of the most important factors in assisting people's food choices [5,17]. Therefore, effectively modeling the visual information and multi-attribute information (e.g., the dish and restaurant information) plays an important role in novel applications like food image based restaurant visualization and recommendation [6,9]. The proliferation of online food image sharing

[1] https://en.wikipedia.org/wiki/Culinary_tourism#cite_note-lucy-long-2.

© Springer International Publishing AG 2016
E. Chen et al. (Eds.): PCM 2016, Part I, LNCS 9916, pp. 520–528, 2016.
DOI: 10.1007/978-3-319-48890-5_51

websites (e.g., Yelp and Dianping[2]) has provided rich food data, such as the food images and different attribute information (e.g., the dish labels and the restaurant information) for this research.

Existing methods mainly focus on visual information for food recognition. For example, Yang *et al.* [18] proposed a visual representation for food items that calculates pairwise statistics between local features. Such approach is bound to standardized meals. Lukas *et al.* [4] mined discriminative parts of food images using random forests for dish recognition. Compared with these shallow models, Kagaya *et al.* [10] applied the CNN for food detection and recognition. In addition, some work [1,3,7,16] developed restaurant-specific dish recognition. Based on the food image recognition, Meyers *et al.* [14] further proposed a system which can recognize the contents of the meal from one image, and then predicted its nutritional contents. However, little work has investigated the problem of modeling the correlation among visual content and rich attribute information, especially for simultaneous restaurant and dish recognition from food images.

It is non-trivial to model the visual content, the dish label and the restaurant identity for simultaneous restaurant and dish recognition. Firstly, although the context information such as GPS information enables the restaurant recognition, we do not always obtain this kind of information, especially in the food-related websites. Secondly, simply recognizing the restaurant identity and dishes separately is not reasonable. Each restaurant has its own special dishes with similar visual patterns, which are different from other restaurants, and thus the dish semantics provides the evidence for recognizing the restaurant identity. For example, if we have recognized the dish "hamburger", we can infer that the food image is from the restaurant "KFC" or "McDonald" with larger probability. Therefore, how to design a model to consider the dependence of the restaurant identity on the dish semantics is challenging.

In order to address these problems, after constructing a food image dataset from Dianping including the food images, the dish labels and the restaurant identity, we propose a model called Partially Asymmetric Multi-Task Convolutional Neural Network (PAMT-CNN) to capture both the dish semantics and the restaurant identity (Fig. 1). Considering the dependence of the restaurant

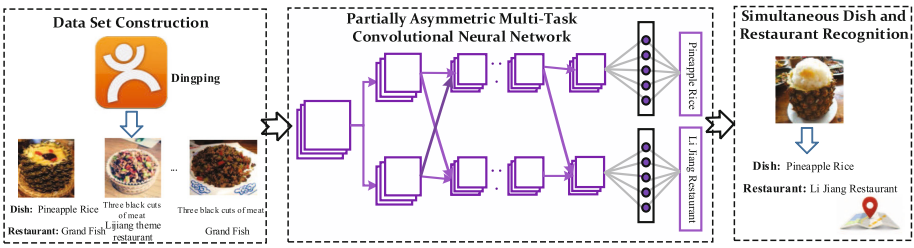

Fig. 1. The proposed framework

[2] The largest venue review website in China, similar to Yelp.

identity on the dish semantics, PAMT-CNN makes the lower layer from the restaurant pathway guided by the dish layer to constitute partially asymmetric network architecture. Based on the PAMT-CNN, we can recognize both the dish and restaurant identity from one food image. The right of Fig. 1 shows an example: PAMT-CNN can not only recognize the dish "Pineapple Rice", but also the restaurant identity "Li Jiang Theme Restaurant".

2 PAMT-CNN

2.1 Network Architecture

The designed PAMT-CNN is illustrated in Fig. 2. Two pathways are designed for classification. "Conv", "Pool" and "Fc" represent the convolutional layer, the pooling layer, and the fully connected layer respectively. In addition, the dish pathway is considered as the first pathway and the restaurant pathway as the second pathway. For instance, "Conv5_2" indicates that it is the fifth convolutional layer in the second pathway.

Two key ideas are exploited in PAMT-CNN. (1) The dish semantics and the restaurant identity describe different aspects of the food image and thus need different learning pathways from the raw data to the supervised information. In order to realize it, PAMT-CNN consists of the dish pathway and the restaurant pathway to learn the dish semantics and the restaurant identity, respectively. (2) As discussed before, the restaurant identity should also be influenced by the dish semantics of the food image. To realize this, PAMT-CNN adopts the partially asymmetric network architecture to make the higher layer of the restaurant pathway determined by both the lower layer of the restaurant pathway and

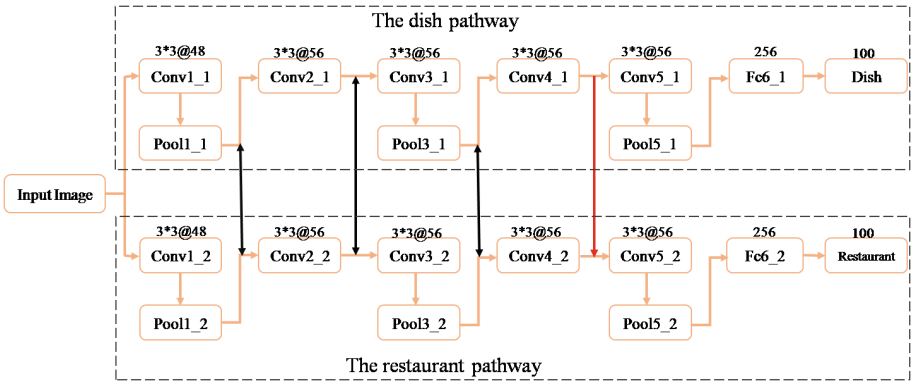

Fig. 2. An illustration of PAMT-CNN. The top imaginary line box is designed for dish classification task and the bottom one is for restaurant classification task. The black double arrow line between two layers denotes that these two layers interact with each other. The red arrow line indicates that the fifth layer from the restaurant pathway is guided by the fourth layer from the dish pathway. (Color figure online)

the dish pathway [8]. Specifically, the fifth convolutional layer of the restaurant pathway is also guided by the fourth convolutional layer of the dish pathway. This is because the higher layers learn not only generic patterns, but also semantically meaningful dish features in the dish pathway [13].

Now we describe the proposed model in details. In the dish pathway, there are five convolutional layers and two fully connected layers, where the first, third, and fifth convolutional layers are followed by the pooling layer. The size of convolution kernels in the convolutional layers are illustrated over each layer in Fig. 2. Both the kernel size and stride size in the pooling layer are three. In the forward propagation stage, the transformation function from $(l-1)^{th}$ layer to the l^{th} can be formulated as two different formulas:

$$x_l^1 = \sigma(W_{l-1}^{1,1} x_{l-1}^1 + b_{l-1}^{1,1} + W_{l-1}^{1,2} x_{l-1}^1 + b_{l-1}^{1,2}), 1 < l \leq 4 \tag{1}$$

where $\sigma(*)$ is the activation function, $\sigma(x) = x$ when $x > 0$ and 0 otherwise. $W_{l-1}^{1,i}(i = 1, 2)$ refers to the weights from $l-1$ layer of the i^{th} pathway to the l layer of the first pathway, and $b_{l-1}^{1,i}$ is the corresponding bias.

$$x_l^2 = \sigma(W_{l-1}^1 x_{l-1}^1 + b_{l-1}^1), 4 < l \leq 7 \tag{2}$$

where W_{l-1}^1 refers to the weights from the $(l-1)^{th}$ layer to the l^{th} layer, and b_{l-1}^1 is the bias. The final output $\tilde{d}$ is defined as:

$$\tilde{d} = softmax(\sigma(W_7^1 x_7^1 + b_7^1)) \tag{3}$$

where $softmax(*)$ is a softmax function.

Similar to the dish pathway, we use the same type of layers and model parameters in the restaurant pathway, but the connections have a little difference. In addition to the last two layers, all the other layers in the restaurant pathway are connected to the previous layers and they are formulated as follows:

$$x_l^2 = \sigma(W_{l-1}^{2,1} x_{l-1}^2 + b_{l-1}^{2,1} + W_{l-1}^{2,2} x_{l-1}^2 + b_{l-1}^{2,2}), 1 < l \leq 5 \tag{4}$$

The final output $\tilde{r}$ is defined as:

$$\tilde{r} = softmax(\sigma(W_7^2 x_7^2 + b_7^2)) \tag{5}$$

2.2 Training

For each dish image, we expect the output $\tilde{d_i}$ and $\tilde{r_i}$ to get close to the goal classification vector d_i and r_i respectively. In our dataset, both the dish and restaurant label have 100 classes. Therefore, d_i and r_i are a 100-D vector with one-shot representation. The loss function for dish classification can be defined as:

$$L_1 = \frac{1}{2}\sum_{i=1}^{N}\left(d_i - \tilde{d_i}\right)^2 + \frac{\lambda}{2}\sum_{l=1}^{3}\left(||W_l^{11}||_F^2 + ||W_l^{12}||_F^2\right) + \frac{\mu}{2}\sum_{l=4}^{7}||W_l^1||_F^2 \tag{6}$$

where $\tilde{d}_i$ is the output in Eq. 3 for image I_i; λ and μ are the factors to balance the loss and regularization to previous over-fitting.

For the restaurant pathway, we define the loss function for restaurant classification as L_2, which can also be computed according to the loss function of regression:

$$L_2 = \frac{1}{2}\sum_{i=1}^{N}(r_i - \tilde{r_i})^2 + \frac{\lambda}{2}\sum_{l=1}^{4}\left(||W_l^{21}||_F^2 + ||W_l^{22}||_F^2\right) + \frac{\mu}{2}\sum_{l=5}^{7}||W_l^2||_F^2 \quad (7)$$

We select a batch of dish images in the dataset and use L_1 to update the weights W_1 in the dish pathway, and use L_2 to update the weights W_2 in the restaurant pathway. We repeat the two operations until the errors converge.

For one image I_i, we can obtain the dish and restaurant labels simultaneously based on the trained PAMT-CNN.

3 Experiment

3.1 Datasets and Implementation Details

Since none of existing food datasets has the restaurant information, we need to build our food dataset. Specifically, we crawled the food images from Dianping with the dish name and the restaurant information. For preprocessing, we firstly discarded the dish with less than 20 images, and then removed the restaurants with less than 3 dishes and the dishes contained by less than 3 restaurants. The number of the resulted dataset is 24,690 with 100 classes of dishes and 100 classes of restaurants. Each dish image in our dataset belongs to both one of the hundred dishes and one of the hundred restaurants. Figure 3 shows the number of images per restaurant and dish, respectively. For space consideration, we use Arabic numerals to denote restaurants and the dishes. Figure 4 shows some examples from our food dataset [15].

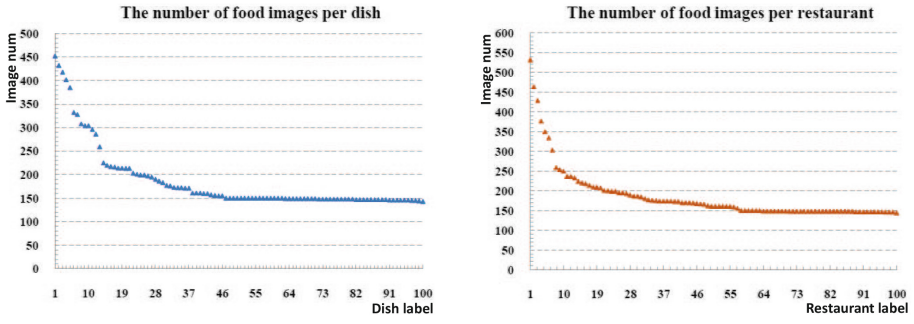

Fig. 3. The distribution of food images per dish and restaurant. For space consideration, we use Arabic numerals to denote restaurants and dishes.

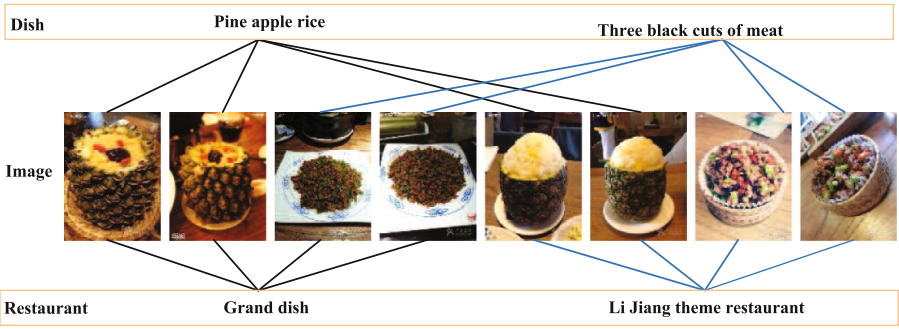

Fig. 4. Some examples of our food dataset. There are two kinds of dishes: Pine apple rice and Three black cuts of meat and two classes of restaurants: Grand dish and Li Jiang theme restaurant

For parameter settings, following AlexNet [11], each image is resized to 256×256, and five 224×224 crops are cropped from the center and the four corners of the resized image. The dropout rate is 0.5 in the last two fully connection layers in each task. The learning rate starts from 0.01 for all layers and is divided by 10, when the error rate stops reducing. It is trained on a single GeForce GTX 780 Ti with 3 GB memory.

3.2 Dish and Restaurant Recognition

To assess our models on the food image dataset, the data set is split into three subsets: 18,626 cases as the training set, 1,122 cases as the validation set and 4,942 cases as the testing set. We adopt classification accuracy as the evaluation metric and choose the following baselines for comparison: (1) CNN-ST: CNN-SingleTask. This baseline trains the dish pathway and the restaurant pathway separately in a single-task way. (2) amtCNN [12]: Each layer of the restaurant pathway is guided by the layer from the dish pathway. (3) amtCNN-Inv: Each layer of the dish pathway is guided by the layer from the restaurant pathway. (4) CNN-MT [2]: This baseline adopts the traditional multi-task deep architecture and each layer between two pathways influences each other. (5) PAMT-CNN-4D: The first fourth layers between two pathways interact with each other while the last remaining layers are separate. (6) PAMT-CNN-4D-2S: The first fourth layers between two pathways interact with each other and the following two layers from the restaurant pathway are guided by the layer from the dish pathway. The last layer between two pathways are separate. (7) PAMT-CNN-4D-3S: The first fourth layers between two pathways interact with each other and the remaining layers from the restaurant layer are guided by the layers from the dish pathway.

The experiment results are shown in Table 1. From these comparison results, we can see that: (1) The interaction between two kinds of information contributes to both dish recognition and restaurant recognition. Therefore, the performance of all the multi-task deep model methods is better than CNN-

Table 1. The performance of all the algorithms in terms of accuracy.

Method	Accuracy on the dish	Accuracy on the restaurant
CNN-ST	70.01 %	56.16 %
amtCNN [12]	70.65 %	59.89 %
amtCNN-Inv	73.72 %	56.21 %
CNN-MT [2]	72.97 %	63.06 %
PAMT-CNN-4D	73.25 %	61.41 %
PAMT-CNN-4D-2S	72.78 %	63.48 %
PAMT-CNN-4D-3S	70.92 %	63.19 %
PAMT-CNN	**74.87 %**	**64.75 %**

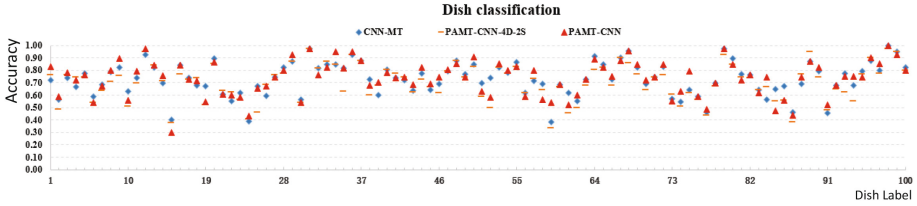

Fig. 5. Comparison of different algorithms over 100 classes of dishes in terms of accuracy. For space consideration, we use Arabic numerals to denote the dishes.

ST. (2) amtCNN in restaurant recognition and amtCNN-Inv in dish recognition has nearly 3 % improvement. These results demonstrate the effectiveness of the guidance from the other pathway. (3) Our method PAMT-CNN outperforms all other baselines. PAMT-CNN in dish recognition and restaurant recognition both has 1 % improvement compared with the best baselines. The reasons are as follows: Firstly, the lower fourth layers of CNN mostly learns generic features [13]. Therefore, interaction with these lower layers between two pathways enables more robust lower layer features. Secondly, the higher layers from the dish pathway learn semantically meaningful dish features. Since the restaurant identity depends on the dish semantics. The guidance from fourth layer of the dish pathway leads to more robust restaurant-oriented features. In addition, the performance of PAMT-CNN is better than PAMT-CNN-3D-2S and PAMT-CNN-3D-3S. The possible reason is that the guidance of higher-level features from the sixth and seventh layer of the dish pathway leads to larger weight of dish-oriented features and thus affect of the performance in restaurant recognition. The detailed comparisons among better baselines CNN-MT, PAMT-CNN-3D-2S and our method PAMT-CNN over each individual class for two tasks are illustrated in Figs. 5 and 6, respectively.

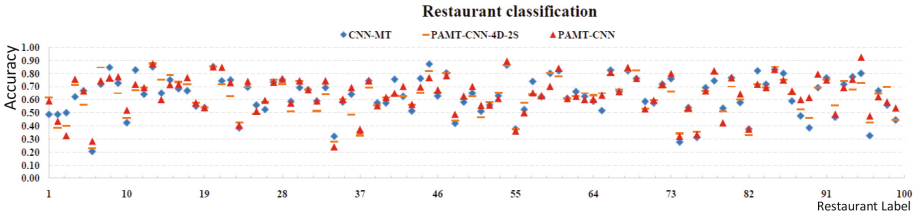

Fig. 6. Comparison of different algorithms over 100 classes of restaurants in terms of accuracy. For space consideration, we use Arabic numerals to denote the restaurants.

4 Conclusion

In this paper, we have proposed a Partially Asymmetric Multi-Task Convolutional Neural Network (PAMT-CNN) model to address the problem of simultaneous dish and restaurant recognition from food images. PSMT-CNN makes the restaurant's identity learned from both the restaurant pathway and the dish pathway to constitute the partially asymmetric shared network architecture. The experiments have justified the effectiveness of our PAMT-CNN. In the future, we plan to investigate the following research directions: (1) We can utilize the correlation among the food images, the dish label and restaurant information learned from the model to conduct novel applications, such as dish and restaurant topic visualization. (2) We will add the user dimension for personalized restaurant recommendation.

Acknowledgements. This work was supported in part by the National Basic Research 973 Program of China under Grant No. 2012CB316400, the National Natural Science Foundation of China under Grant Nos. 61532018 and 61322212, the National High Technology Research and Development 863 Program of China under Grant No. 2014AA015202, China Postdoctoral Science Foundation under Grant No. 2016M590135, Beijing Science And Technology Project under Grant No. D161100001816001. This work is also funded by Lenovo Outstanding Young Scientists Program (LOYS).

References

1. Beijbom, O., Joshi, N., Morris, D., Saponas, S., Khullar, S.: Menu-match: restaurant-specific food logging from images. In: Applications of Computer Vision, pp. 844–851 (2015)
2. Bengio, Y.: Learning deep architectures for ai. Mach. Learn. **2**(1), 1–127 (2009)
3. Bettadapura, V., Thomaz, E., Parnami, A., Abowd, G.D., Essa, I.: Leveraging context to support automated food recognition in restaurants. In: Applications of Computer Vision, pp. 580–587 (2015)
4. Bossard, L., Guillaumin, M., Gool, L.: Food-101 – mining discriminative components with random forests. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 446–461. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10599-4_29

5. Cordeiro, F., Bales, E., Cherry, E., Fogarty, J.: Rethinking the mobile food journal: exploring opportunities for lightweight photo-based capture. In: Proceedings ACM, pp. 3207–3216 (2015)
6. Ge, M., Ricci, F., Massimo, D.: Health-aware food recommender system. In: ACM, pp. 333–334 (2015)
7. Herranz, L., Xu, R., Jiang, S.: A probabilistic model for food image recognition in restaurants. In: ICME (2015)
8. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM MM, pp. 675–678 (2014)
9. Kadowaki, T., Yamakata, Y., Tanaka, K.: Situation-based food recommendation for yielding good results. In: ICMEW, pp. 1–6 (2015)
10. Kawano, Y., Yanai, K.: Foodcam-256: a large-scale real-time mobile food recognitionsystem employing high-dimensional features and compression of classifier weights. In: Proceedings ACM MM, pp. 761–762 (2014)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in NIPS, pp. 1097–1105 (2012)
12. Liu, S., Cui, P., Zhu, W., Yang, S.: Learning socially embedded visual representation from scratch. In: Proceedings of ACM MM, pp. 109–118 (2015)
13. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10590-1_53
14. Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., Murphy, K.P.: Im2calories: towards an automated mobile vision food diary. In: Proceedings of ICCV, pp. 1233–1241 (2015)
15. Wang, S., Jiang, S.: Instre: a new benchmark for instance-level object retrieval and recognition. ACM Trans. Multimedia Comput. Commun. Appl. **11**(3), 1–21 (2015)
16. Xu, R., Herranz, L., Jiang, S., Wang, S., Song, X., Jain, R.: Geolocalized modeling for dish recognition. IEEE TMM **17**(8), 1187–1199 (2015)
17. Yang, L., Cui, Y., Zhang, F., Pollak, J.P., Belongie, S., Estrin, D.: Plateclick: bootstrapping food preferences through an adaptive visual interface. In: Proceedings of KSEM, pp. 183–192 (2015)
18. Yang, S., Chen, M., Pomerleau, D., Sukthankar, R.: Food recognition using statistics of pairwise local features. In: CVPR, pp. 2249–2256 (2010)