Personalized Knowledge Distillation-based Mobile Food Recognition

Zhao Heng[a], Sharmili Roy[a], Kim-Hui Yap[a], Alex Kot[a], Lingyu Duan[b]
[a]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798
[b]School of Electronics Engineering and Computer Science, Peking University, China 100080

Abstract— *Visual food recognition has received increasing attention in recent years due to its role in individual dietary monitoring and public health management. Existing food recognition solutions, (i) ignore individual's dietary preferences, thereby under-utilizing the available information, and (ii) do not focus on reducing the computational and memory requirements of such systems. We address these issues by proposing a new Personalized Knowledge Distillation (PKD) method for mobile food recognition which is: (i) personalized based on individual's dietary habits, and (ii) compact and lightweight making it more amenable for future deployment on mobile devices. PKD achieves high accuracy and low memory footprint using knowledge distillation and personalized learning. The proposed method outperforms comparative methods on benchmark datasets and newly constructed NTUIndianFood50 dataset.*

Keywords: Food recognition, knowledge distillation, personalized learning, compact network

1. Introduction

Unhealthy diet is a primary cause of various serious illnesses such as cardiovascular diseases, diabetes and obesity. Many human cancers are related to unhealthy diets. Monitoring dietary behavior, hence, is crucial to cultivate and maintain a healthy lifestyle. Large-scale, cumulative and analyzable dietary data, can be used to develop tools to perform continuous monitoring of people's health condition and dietary intakes. The conventional way of collecting dietary data via online or offline surveys and questionnaires are inefficient. This makes it difficult to monitor people's dietary behaviors and provide any real-time feedback information.

With recent progress in e-health, various fitness applications have been proposed on mobile platforms that assist users to maintain good dietary habits. Most of these applications, however, focus on recording and analyzing human physiological data. Two such applications are MyFitnessPal [1] and LoseIt! [2]. They ask users to manually log their diets for every meal taken. Manual data entry is tedious and time consuming, and market research shows that such applications cannot retain their users in the long run [3].

With the rapid enhancement of digital cameras on smart phones, more and more people tend to share their food via taking images on the Internet. As opposed to manual data entry, some approaches have proposed using the mobile camera to do diet logging instead, but such approaches can only log the meal without further analysis [4] or simply reply on expert nutritionists [5] or crowd sourcing [6] to analyze the images offline. As a result, such applications do not provide any immediate feedback to the user and the efficiency is greatly reduced.

2. Related work

Many vision based techniques have been proposed to automatically recognize food from pictures. These solutions employ traditional hand-crafted features such as texture, SIFT, HOG, bag-of-visual-words, pre-segmented image patches, etc. Kawano and Yanai [7] used HOG and color patches with the Fisher Vector coding as image features and tested on a 100-category food dataset. Zhang et al. [8] used generated image features via saliency detection and hierarchical segmentation to train a linear SVM classifier. Bossard et al. [9] proposed the popular benchmark food dataset Food-101 and applied random forest to mine discriminant superpixel-grouped parts in the food images. SVM was then used to classify these parts. On a diabetic dataset meant to help diabetic patients, Anthimopoulos et al [10] designed a food recognition system based on a bag-of-features model. Vision-based solutions work best in laboratory conditions where the picture backgrounds are clean and the food components are well demarcated. However, food images in real scenarios may be occluded or often mixed together, especially for Asian food such as Chinese and Indian cuisine. This decreases the reliability of using local image features to perform food recognition in real-life situations.

Deep convolutional neural networks (DCNNs) are currently the state-of-the-art technique for image recognition problems. DCNNs can estimate optimal feature representations from the data adaptively as against the traditional hand-crafted features. Recent research on DCNNs has shown that deep architectures can outperform traditional vision-based methods in food recognition problems. Pouladzadeh et al. [11] proposed a food recognition system based on extracted convolutional neural network (CNN) features which is able to recognize multi-food images by region mining. Tanno et al. [12] implemented a new mobile food recognition system based on DCNN, and Hassannejad et al. [13] fine-tuned Google's Inception module to perform classification on various food image datasets. Large intra-class variations in illumination and composition, similar visual appearance of dishes from different classes, occlusion etc. make food recognition a challenging problem to solve. Some solutions have proposed to use additional contextual information such as geographical location [14], dish ingredients [15], textual information [16] etc. to improve the classification accuracy.

Most existing food recognition systems are based on heavyweight convolutional networks such as VGGNet [15], ResNet [17] or InceptionNet [13] frameworks which have very high memory, computation, and energy footprints. A typical deep learning based solution to dietary logging would take the following approach. The end user would snap a picture of the food and the picture would be sent to a central server possibly with some contextual information such as geo-location. The classification would happen at the central server based on a heavyweight food classification engine and the results would be sent back to the local mobile device. There are two key limitations of such solutions. First, the classification engine at the server does not take into consideration the mobile user's personal food preferences, dietary habits, and dietary history. Second, since the classifier is not compact enough to fit on a mobile device, the users are required to send every picture to the server for analysis. This may not be practically feasible since the user may not have access to the Internet at all times or may not be willing to transfer pictures over the network due to privacy concerns.

In this paper, we address these two issues by proposing a personalized and lightweight food recognition engine that is (i) customized based on the end-user's dietary habits and preferences, and (ii) is small and efficient which makes it more amenable for future deployment on mobile devices. Typically, most people have preferences on what they like to eat. Often they eat their favorite dishes more frequently and visit their favorite restaurants time and again. This, we believe, presents a key information that can be leveraged to greatly improve the accuracy of the food classification engine. Our system is based on a compact neural network architecture that achieves high classification accuracy by leveraging on personalized dietary preferences and knowledge distillation from a big trainer neural network. We call this Personalized Knowledge Distillation (PKD) method and the resulting network, PKD-Net. Section 3 gives an overview of the PKD method. Section 4 discusses personalized learning and how we model an end-user's dietary preferences. Section 5 details the proposed methodology. Classification performance of PKD-Net in comparison to other comparative methods is evaluated in Section 6 and Section 7 concludes the paper.

3. System Architecture

Fig. 1 provides an overview of the proposed PKD framework. We train a generic heavyweight network (GHN) such as VGG16 or ResNet to perform food classification on the principal food dataset. The principal food dataset is typically a wide collection of food images crawled from the Internet image search engines such as Google and Yahoo. The knowledge of GHN is then distilled into a lightweight network based on compact architectures such as MobileNet [18] or SqueezeNet [19]. This compact knowledge-distilled network (CKDN) is then deployed to multiple end users.

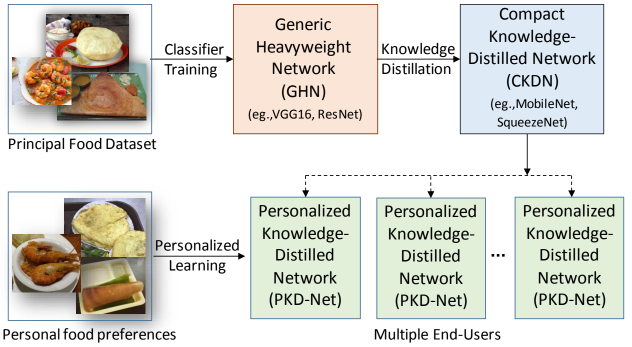
Principal Food Dataset
Classifier Training
Generic Heavyweight Network (GHN)
(eg.,VGG16, ResNet)
Knowledge Distillation
Compact Knowledge-Distilled Network (CKDN)
(eg.,MobileNet, SqueezeNet)
Personalized Learning
Personalized Knowledge-Distilled Network (PKD-Net)
Personalized Knowledge-Distilled Network (PKD-Net)
...
Personalized Knowledge-Distilled Network (PKD-Net)
Personal food preferences
Multiple End-Users

Fig. 1: Overview of the proposed PKD framework. A generic heavyweight network is trained on the principal food database. This network then distills its information to a compact network. The compact network is then personalized based on the end-user's dietary preferences.

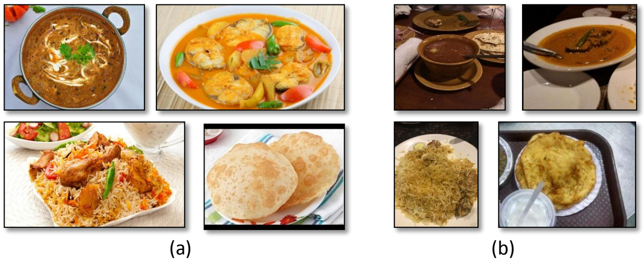

Fig. 2: Comparison of food images crawled from Google with those captured by mobile users in real restaurant settings. (a) Google images for four dishes namely, "Dal Makhani", "Fish Curry", "Bhatura", and "Chicken Biryani" are shown in clockwise direction, (b) the corresponding four dishes captured by mobile users in a restaurant.

As mobile users start using the CKDN model, they accumulate a collection of pictures from the dishes that they consume on a regular basis. This collection of pictures constitute the individual's dietary history or preference. We call this collection of food images an individual's Food Diary. As the Food Diary is populated and grows to a sufficient size, the images are uploaded onto a user account in the server. Using personalized learning on this Food Diary, CKDN is then customized to compute a personalized knowledge-distilled network (PKD-Net) that captures the end-user's dietary history.

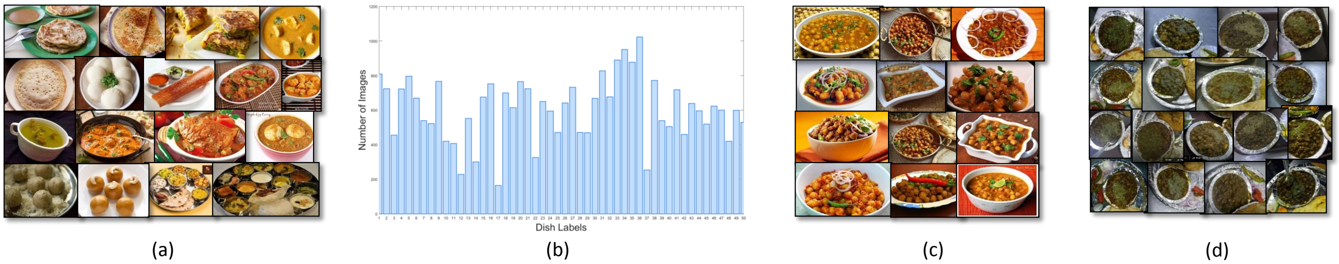

Fig. 3: Overview of the NTUIndianFood50 dataset. (a) Some selected dishes out of the 50 dish categories are shown here, (b) this plot shows the distribution of images across the 50 dishes, (c) some images for the dish named "Channa Masala" are shown in this figure, (d) this figure shows Food Diary images from the same dish "Channa Masala."

4. Personalized Learning

One of the two key aspects of PKD-Net is personalized learning. Most existing visual food classification solutions train and test their systems on a standard food database crawled from the Internet using search engines such as Google, Yahoo and Baidu. These resulting images typically have high resolution and clean background. Real life food images acquired using mobile devices in a restaurant or food joint, however, suffer from many artifacts such as low illumination of the indoor environment, clutter in the background, occlusion of the food or motion blurring, just to name a few. Fig. 2 compares some food images crawled from Google with corresponding images captured at real life restaurant scenarios where users have acquired pictures of the same dishes using mobile phones.

Hence, for personalized learning, a key challenge is to construct a realistic Food Diary that satisfies the following three conditions: (i) the images are captured using a mobile device, (ii) there is a relatively larger collection of images that represent the end-user's favorite dishes, and (iii) a significant collection of images should be captured at the restaurants the user visits frequently. None of the existing food databases such as Food-101 [9], Vireo Food-172 [15], or UEC Food-256 [20] can be directly used to model such a Food Diary.

We introduce a new dataset based on Indian cuisine in this paper called NTUIndianFood50. The two key features of NTUIndianFood50 are (i) the dataset contains/ models a realistic Food Diary constructed from user supplied mobile images of various Indian dishes, and (ii) the dataset introduces a new cuisine to the existing food databases. Indian dishes have diverse appearances and the composition of the dish varies largely depending on the regions where the dishes have been prepared. Automatic classification is challenging due to the wide variations in the way the ingredients are selected, chopped and mixed within the same dish. Many dishes are based on curries which make different dishes appear to be visually similar. Fig. 3 (a) shows a few representative dishes from this dataset.

4.1 NTUIndianFood50 dataset

To construct a new Indian food dataset, we crawled images for Indian dishes from Google and Yahoo image search engines. For each category, the dish name was given as the keyword to the search engines. The returned images were checked manually for removal of images that suffered major blurring, images that contained multiple dishes, and the images that did not represent the queried dish. Duplicate images and images with resolution lower than 256×256 were removed using automated scripts. The resulting dataset comprises of $30,378$ images with a mean of 607 images per dish. Fig. 3(b) shows the distribution of images among the 50 dishes. Due to limited space, only dish labels are used in the figure. A complete list of dish names and their corresponding labels are provided in the supplementary material. Fig. 3(c) shows images from a dish named "Channa Masala." We observe that the food dataset images crawled from Google and Yahoo are high quality with good illumination and clean background.

4.2 Food diary

To model the end-user's Food Diary, we select the most popular 24 dishes from the NTUIndianFood50 dataset. We then find the most popular Indian restaurants that serve these 24 dishes. Mobile users visiting these restaurants often post pictures of the food that they ate at www.zomato.com [21] which is a very popular restaurant search and review portal for Indian cuisine. The user posted pictures are typically captured using a mobile device. Some of the popular restaurants have more than $4,000$ review pictures in this portal. We find the selected Indian restaurants in www.zomato.com and search for user posted pictures of the popular 24 dishes. These user-supplied review pictures constitute the mobile-user's Food Diary. The current version of our Food Diary has $2,073$ pictures from 24 dishes with an average of 86 images per dish. As expected in real life situations, we observe that the user posted pictures are much poorer in illumination and quality and have much more clutter in the background when compared to the NTUIndianFood50 database images. Fig. 3(d) shows some examples of Food Diary images from a dish named “Channa Masala." Classification on Food Diary images is challenging and more realistic. In Section 6, we show that a classifier trained on the NTUIndianFood50 dataset does not achieve high classification accuracy when tested on images from the Food Diary, thereby motivating the need for classifier personalization.

5. Proposed Methodology

There are three major steps to compute the proposed PKD-Net: (i) the first step is to train a generic food classifier network GHN, (ii) the second step is to perform knowledge distillation from GHN to the compact CKDN, and (iii) the third step is to customize CKDN by personalized learning on the Food Diary.

5.1 Training GHN

We evaluated various heavyweight classification architectures on benchmark food datasets and the newly created NTUIndianFood50 dataset. Existing literature uses VGG16 as a food classification architecture [15] and we found that VGG16 achieves high accuracy for multiple food datasets. Hence, VGG16 was chosen as the GHN model. GHN was first pre-trained on the ImageNet database [22] for system parameter initialization. Then all the weights were optimized by training the classifier on the NTUIndianFood50 dataset. A cross entropy loss and stochastic gradient descent was used for the optimization.

5.2 Knowledge distillation

In order to design a recognition system that has low computational and memory requirements, we distill the knowledge of GHN to a compact network based on MobileNet-224 [18] architecture. MobileNets need only about 10 to 12 MB of storage as opposed to 510 MB required by a full weight VGG16. In addition, MobileNets use depth-wise separable convolutions to achieve 8 to 9 times less computation than standard convolutions with a small drop in accuracy [18]. To boost the accuracy of the MobileNet-224 model, we use knowledge distillation from the GHN classifier.

The goal of knowledge distillation is to transfer the knowledge of a large trainer network to a smaller trainee network such that the trainee network approximates the trainer network accurately. The trainer network, in our case, is GHN and the trainee network is the compact knowledge distilled network CKDN (Fig. 1). If we represent the trainer with $\phi_{GHN}(x)$ and the trainee with $\phi_{CKDN}(x)$, then the goal of knowledge distillation is to achieve $\forall_{x_i \in X} \phi_{GHN}(x_i) = \phi_{CKDN}(x_i)$ where X is the set of training images from the NTUIndianFood50 dataset and $|X| = N$.

For each input image, the output of GHN is a probability distribution over all the 50 dish categories. This probability is generated by the softmax layer from *logits*. Logits represent the output of the last fully connected layer. The dimension of the logits vector for both the trainer and the trainee network are equal to the number of categories. These logits are used to transfer the knowledge from $\phi_{GHN}(x)$ to $\phi_{CKDN}(x)$. For an input food image x_i, the logits vector generated by $\phi_{GHN}(x)$ can be denoted by $\boldsymbol{v_i}$ where the dimension of vector $\boldsymbol{v_i} = (v_i^1, v_i^2, \ldots v_i^C)$ is the number of dish categories $C = 50$. A generalized softmax layer converts the logits vector $\boldsymbol{v_i}$ to a probability distribution $\boldsymbol{q_i}$ as follows:

$$M_T(\boldsymbol{v_i}) = \boldsymbol{q_i}, \text{where } q_i^j = \frac{exp(\frac{v_i^j}{T})}{\sum_k exp(\frac{v_i^k}{T})}, \qquad (1)$$

where T is the temperature parameter. For traditional classification tasks, $T = 1$. Similarly, the trainee network $\phi_{CKDN}(x)$ generates a student logits vector $\boldsymbol{w_i}$ and the corresponding probability distribution $M_T(\boldsymbol{w_i})$. By modifying the temperature T, we can obtain a different probability distribution $\boldsymbol{v_i'}$, which is 'soft target labels' from the trainer $\phi_{GHN}(x)$. These soft labels usually contains richer information for object classification than the case when $T = 1$ and they can be used to perform knowledge transfer by providing the trainee with extra information from the trainer. Existing literature proposes to minimize Kullback-Leibler divergence between the soft probability distribution of the trainer and normal probability distribution of the trainee [25] :

$$L_{KD}(\phi_{GHN}, \phi_{CKDN}) = \frac{1}{N} \sum_{i=1}^{N} KL(M_T(\boldsymbol{v_i'}) || M_T(\boldsymbol{w_i})), \qquad (2)$$

where $KL(\boldsymbol{x}||\boldsymbol{y})$ represents the Kullback-Leibler divergence between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. When a set of image-label pairs $\{(x_i, \boldsymbol{l_i})\}$ are given, the trainee $\phi_{CKDN}(x)$ learns a standalone pure supervised classification task using the traditional cross-entropy loss which is defined as:

$$L_S(\phi_{CKDN}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{H}(\boldsymbol{l_i}, M_{T=1}(\boldsymbol{w_i})), \qquad (3)$$

where $\mathcal{H}$ is the entropy function. For transfer learning of $\phi_{CKDN}(x)$, we use a weighted combination of Kullback-Leibler loss (Equation 2) and the standalone cross-entropy loss (Equation 3) defined as follows:

$$\begin{aligned} L(\phi_{GHN}, \phi_{CKDN}) &= \alpha L_S(\phi_{CKDN}) \\ &\quad + (1-\alpha) L_{KD}(\phi_{GHN}, \phi_{CKDN}), \end{aligned} \qquad (4)$$

where α is the weighting factor. To make the distillation process as effective as possible, more emphasis should be given to $L_{KD}(\phi_{GHN}, \phi_{CKDN})$ while keeping the emphasis on $L_S(\phi_{CKDN})$ small. Hence, we set the weight α to 0.1. When the temperature T is high, the distillation process is equivalent to minimizing $\frac{1}{2}(\boldsymbol{w_i} - \boldsymbol{v_i'})^2$. At low temperature, however, most of the logits that are more negative than the average are neglected even though they may convey some useful information acquired by the trainer network [25]. As a result, we choose an intermediate value for the temperature,

Table 1: Comparison of classification accuracy, architecture and model size of the proposed CKDN model with existing prior solutions on the benchmark Food-101 dataset.

	CNN [9]	DCNN-Food [23]	DeepFood [24]	ResNet50 [17]	InceptionV3 [17]	**CKDN** (Ours)
Top-1 (%)	56.4	70.4	77.4	82.3	83.8	**84.0**
Architecture	AlexNet	Modified AlexNet	Modified GoogLeNet	ResNet-50	Inception-V3	MobileNet-224
Model Size	240 MB	425 MB	51 MB	98 MB	91 MB	**12** MB

Table 2: Comparison of classification accuracy, architecture and model size of the proposed CKDN model with existing prior solutions on the benchmark UECFood-256 dataset.

	DeepFoodCam [12]	DCNN-Food [23]	DeepFood [24]	InceptionV3 [17]	**CKDN** (Ours)
Top-1 (%)	63.64	67.57	63.8	76.7	**77.5**
Architecture	Modified AlexNet	Modified AlexNet	Modified GoogLeNet	Inception-V3	MobileNet-224
Model Size	240 MB	425 MB	51 MB	91 MB	**12** MB

$T = 4$, in order to allow the much smaller CKDN to capture most of the knowledge from GHN while ignoring the most negative logits.

The proposed system is implemented in the Caffe deep learning framework [26] on NVIDIA Titan Xp graphics processing unit (GPU). To realize the distillation process, we implemented a new Caffe layer which takes logits from the trainer ($\boldsymbol{v_i}$) to generate the 'soft' logits ($\boldsymbol{v_i'}$) with editable temperature parameter T. Then a new loss is computed using the logits from the trainee and the trainer as inputs as formulated in Equation 4.

5.3 Classifier personalization

The knowledge distilled CKDN classifier, $\phi_{CKDN}(x)$, captures the information of the principal food database, which in our case is the NTUIndianFood50 database. When tested on images from the Food Diary, however, the classification accuracy of $\phi_{CKDN}(x)$ is less than 75% as discussed in Section 6. The next step is to personalize $\phi_{CKDN}(x)$ for each mobile user based on the individual's Food Diary. This is done by optimizing the cross entropy loss function described in Equation 3 where now the image-label pairs $\{(x_i, \boldsymbol{l_i})\}$ come from the training images of the end-user's Food Diary and N is the size of this training set. Since each end user has a different set of images in the Food Diary, $\{(x_i, \boldsymbol{l_i})\}$, the resulting network parameters would be different and personal to each user. The resulting optimized network is the proposed personalized knowledge distilled network (PKD-Net).

6. Results

We present the experimental results in two parts. The first part shows that even without personalized learning, the proposed compact knowledge distilled CKDN model outperforms existing prior works on a benchmark food dataset. The second part demonstrates how personalized learning on end-user's Food Diary results in significant improvement of classification accuracy. All the experiments are carried out in NVIDIA Titan Xp GPU.

6.1 Evaluation of non-personalized CKDN model

In this section, we evaluate the proposed framework with existing works on two popular benchmark datasets Food-101 [9] and UECFood-256 [20]. Since existing food databases do not model personalized food diaries, for fair comparison, this set of experiments do not leverage on personalized learning. Hence, we compare our non-personalized CKDN model with prior methods in the literature. Table 1and Table 2 compare the top-1% classification accuracy, architecture, and model size of CKDN with other state-of-the-art deep learning based methods such as CNN [9], DeepFoodCam [12], DCNN-Food [23], DeepFood [24], ResNet50 [17], and Inception V3 [17] discussed in Section 2. These methods evaluate their models on Food-101 dataset and UECFood-256 dataset respectively. The model size in Table 1 and Table 2 represents the amount of disk space required to store the trained model. The proposed CKDN model clearly outperforms existing solutions on the both benchmark Food-101 and UECFood-256 datasets. In addition, the memory footprint of CKDN is much smaller than the other models which makes CKDN more favorable for future deployment on mobile devices.

6.2 Evaluation of classifier personalization

In this section, we demonstrate that personalized learning based on an end-user's Food Diary significantly improves the classification performance. As a 'baseline test', we first evaluate the classification accuracy of the non-personalized CKDN model on images from an end-user's Food Diary. The top-1 accuracy in this baseline test case was found to be 74.3% (Table 3). This clearly shows that the CKDN model does not perform well in classifying end-user's Food Diary images.

In the second set of experiments, we evaluate the classification accuracy of CKDN using a combination of test images from the end-user's Food Diary and from the NTUIndianFood50 dataset. This represents the scenario where the user regularly visits his favorite restaurants and eats his favorite dishes but time to time the user visits new restaurants and tries new dishes. To simulate this scenario we select 360 test images from the Food Diary and $1,000$ test images from the NTUIndianFood50 dataset. These test images were not used for training or personalization of the networks. We call this test case 'personalized test'. The classification accuracy of CKDN on personalized test is 82.6% which is higher than the baseline test because a significant number of test images come from the NTUIndianFood50 dataset, the same domain on which CKDN was trained.

Our third experiment demonstrates the utility of personalized learning. We evaluate PKD-Net, which is the personalized version of CKDN, on the personalized test case scenario described above. The classification accuracy of PKD-Net is 95.6%. Personalized learning, thus, improves the classification accuracy by 13% (Table 3).

Table 3: Comparison of classification accuracy between CKDN and PKD-Net.

Model	Test Case	Top-1 Accuracy (%)
CKDN	Baseline test	74.3
CKDN	Personalized test	82.6
PKD-Net	Personalized test	95.6

Table 4: Comparison of GHN and PKD-Net in terms of model size and number of parameters to optimize.

Model	Model Size	Number of parameters
GHN	510 MB	138.0 Million
PKD-Net	12 MB	4.2 Million

6.3 PKD-Net storage and inference time

Here we discuss the storage, inference time and time required for personalized learning for PKD-Net. Table 4 provides the model size and total number of network parameters for PKD-Net and compares these parameters with those of the GHN model on the server. PKD-Net provides more than 40 fold reductions in model size and the number of network parameters. Inference time of PKD-Net was found to be 23 milliseconds per image and the time taken for personalized learning on a Food Diary with 1512 images was 16 minutes.

7. Conclusion

In this paper, we propose a personalized knowledge distillation based network, PKD-Net, for mobile food recognition. The proposed PKD-Net recognizes dishes from food pictures captured by mobile phones and is based on a compact MobileNet architecture which makes it suitable for future deployment on mobile devices. PKD-Net achieves strong classification performance on a compact neural network architecture using two key techniques (i) personalized learning on end-user's dietary preferences and (ii) knowledge distillation from a deep trainer network. We introduce a new NTUIndianFood50 dataset that models end-user's food preferences via pictures of food taken using mobile devices in restaurant settings. Experiments on this newly constructed dataset and existing benchmark datasets demonstrate that the proposed PKD-Net outperforms many comparative methods in the literature.

Acknowledgement

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the Infocomm Media Development Authority, Singapore. This research is supported by the National Research Foundation, Prime Minister's Office, Singapore, under the NRF-NSFC grant NRF2016NRF-NSFC001-098. We gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan X Pascal GPU used for this research.

References

[1] MyFitnessPal, "Free calorie counter, diet & exercise journal," www.myfitnesspal.com.

[2] LoseIt!, "Weight loss that fits," www.loseit.com.

[3] Felicia Cordeiro, Daniel A Epstein, Edison Thomaz, Elizabeth Bales, Arvind K Jagannathan, Gregory D Abowd, and James Fogarty, "Barriers and negative nudges: Exploring challenges in food journaling," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 1159–1162.

[4] Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty, "Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 3207–3216.

[5] Corby K Martin, Hongmei Han, Sandra M Coulon, H Raymond Allen, Catherine M Champagne, and Stephen D Anton, "A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method," *British Journal of Nutrition*, vol. 101, no. 3, pp. 446–456, 2008.

[6] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z Gajos, "Platemate: crowdsourcing nutritional analysis from food photographs," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 1–12.

[7] Yoshiyuki Kawano and Keiji Yanai, "Foodcam: A real-time food recognition system on a smartphone," *Multimedia Tools and Applications*, vol. 74, no. 14, pp. 5263–5287, 2015.

[8] Weiyu Zhang, Qian Yu, Behjat Siddiquie, Ajay Divakaran, and Harpreet Sawhney, “"snap-n-eatâĂİ food recognition and nutrition estimation on a smartphone," *Journal of diabetes science and technology*, vol. 9, no. 3, pp. 525–533, 2015.

[9] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool, "Food-101–mining discriminative components with random forests," in *European Conference on Computer Vision*. Springer, 2014, pp. 446–461.

[10] Marios M Anthimopoulos, Lauro Gianola, Luca Scarnato, Peter Diem, and Stavroula G Mougiakakou, "A food recognition system for diabetic patients based on an optimized bag-of-features model," *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1261–1271, 2014.

[11] Parisa Pouladzadeh and Shervin Shirmohammadi, "Mobile multi-food recognition using deep learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 3s, pp. 36, 2017.

[12] Ryosuke Tanno, Koichi Okamoto, and Keiji Yanai, "Deepfoodcam: A dcnn-based real-time mobile food recognition system," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 2016, pp. 89–89.

[13] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni, "Food image recognition using very deep convolutional networks," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 2016, pp. 41–49.

[14] Michele Merler, Hui Wu, Rosario Uceda-Sosa, Quoc-Bao Nguyen, and John R Smith, "Snap, eat, repeat: a food recognition engine for dietary logging," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 2016, pp. 31–40.

[15] Jingjing Chen and Chong-Wah Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 32–41.

[16] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso, "Recipe recognition with large multimodal food dataset," in *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6.

[17] Eduardo Aguilar, Marc Bolaños, and Petia Radeva, "Food recognition using fusion of classifiers based on cnns," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 213–224.

[18] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[19] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[20] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.

[21] Zomato, "Indian restaurant search and discovery service," www.zomato.com.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[23] Keiji Yanai and Yoshiyuki Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6.

[24] 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," in *International Conference on Smart Homes and Health Telematics*. Springer, 2016, pp. 37–48.

[25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[26] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.