Simultaneous Estimation of Food Categories and Calories with Multi-task CNN

Takumi Ege and Keiji Yanai
Department of Informatics, The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 JAPAN
{ege-t,yanai}@mm.inf.uec.ac.jp

Abstract

In this paper, we propose simultaneous estimation of food categories and calories for food photos. Since there exists strong correlation between food categories and calories in general, we expect that simultaneous training of both brings performance boosting compared to independent single training. To this end, we use a multi-task CNN. In the experiments, we collected calorie-annotated recipe data from the online cooking recipe sites, and trained multi-task and single-task CNNs. As results, the multi-task CNN achieved the better performance on both food category estimation and food calorie estimation than single-task CNNs.

1 Introduction

In recent years, because of a rise in healthy thinking on eating, many mobile applications for recording everyday meals have been released so far. Some of them employ food image recognition which can estimate not only food names but also food calories. However, in most of the cases, the estimated calories are just associated with the estimated food categories, or the relative size compared to the standard size of each food category which are usually indicated by a user manually. Currently, no applications which can estimate food calories automatically exist. Although most of the image recognition tasks including food category recognition have been almost solved due to great progress of CNN-based image recognition methods, fully-automatic food calorie estimation from a food photo has still remained an unsolved problem. We think that food calorie estimation not only helps people's health a lots, but also is promising as a new problem of image recognition studies.

Regarding food calorie estimation, two approaches have been proposed so far. One is to estimate calories based on the estimated food category and its size or volume, which is a quite standard approach. One of the representative works of this approach is Im2Calories by Google [1], which estimates food categories, ingredients, volumes of each of the dishes included in a given food photo, and finally outputs food calories by calculation based on the estimated volume and the calorie density corresponding to the estimated food category.

The other approach is to estimate calories from food photos directly without estimating food categories and sizes. The works adopting this approaches is a few. One of the representative works is the work by Miyazaki et al. [2]. The biggest difficulty on direct calorie estimation is creating datasets which contains calorie-annotated food images. They hired dietitians to annotate calories on 6512 food photos which uploaded to the commercial food logging service, FoodLog [1]. Unfortunately, their dataset was not released, because they used the food photos picked up from the master image database of the commercial service. In their work, they adopted image-search based calorie estimation, in which they searched the calorie-annotated food photo database for the top k similar images based on conventional hand-crafted features such as SURF-based BoF and color histograms and estimated food calories by averaging the food calories of the top k food photos. Since their method ignored information on food categories, their method was applicable for any kinds of foods. However, the number of food images was not enough for the search-based method, and the employed image features was too simple. As results, they failed to estimate food calories with high accuracy.

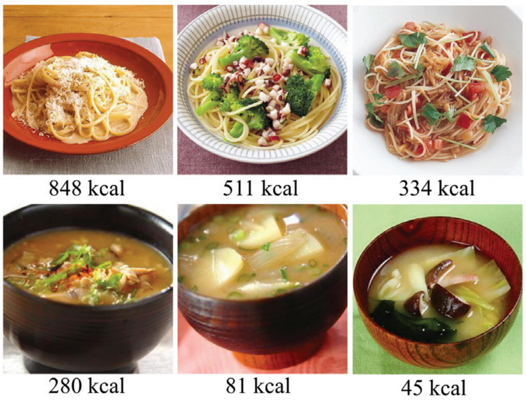


Figure 1. The differences of food calorie values within the same food categories, "Miso soup" and "Spaghetti". "Spaghetti" in the top row and "Miso soup" in the bottom row are shown, respectively.

In this work, basically we adopt the latter approach which is direct food calorie estimation. To resolve the problems clarified by Miyazaki et al.'s work, we (1) use the state-of-the-art Convolutional Neural Network (CNN), (2) take into account food categories for estimating food calories, and (3) build a calorie-annotated food photo database by collecting food images from online cooking recipe sites. Note that our dataset contains not the photos of large-serving or small-serving foods but only the photos of normal-serving foods, since the photos in our dataset were collected from the recipe site. In our work, we focus on estimating the differences of food calories depending on ingredients, seasonings, topping or how-to-cook within the same food categories as shown in Figure 1 rather than estimating the amounts of served foods. We think estimating calories from the appearance is important in the task of food calorie estimation, which cannot be resolved by

[1] http://www.foodlog.jp/

only food category classification.

In addition, we focus on estimating food calories of single dishes. We assume that a given food image contains only one dish as shown in Figure 1. Dish detection is easily possible by using the state-of-the-art CNN-based object detection methods such as Faster R-CNN [3]. We will combine food dish detection and the work presented in this paper as a future work.

To this end, we propose simultaneous estimation of food categories and calories from food photos using multi-task CNN [4]. Since there exists strong correlation between food categories and calories especially regarding ingredients and cooking methods, we expect that simultaneous training of both brings performance boosting compared to independent single training. In fact, Chen and Ngo [5] proposed using a multi-task CNN to estimate food categories and food ingredients at the same time, and proved that simultaneous estimation boosted estimation performance on both tasks. Inspired by this work, we introduce a multi-task CNN for food calorie estimation.

To summarize our contributions in this paper, we use multi-task CNN for for the task of food calorie estimation with joint estimation of food categories, and create a food photo database by collecting food images from online cooking recipe sites, in which each of the food photos is annotated with both food category and food calorie.

2 Multi-task Learning with CNN

In recent years, food image recognition by a CNN has been actively studied, and it has been proved that a CNN improved the performance of food category recognition significantly. In food categorization, a CNN accepts a single food image as input data and outputs class probabilities. Contrary to this, in case of food calorie estimation, a CNN accepts an image as well and outputs the value of food calorie intake directly. If it is possible to learn a single CNN in such different tasks simultaneously, the knowledge gained in both tasks is expected to be shared, which will makes good effects on the performance of both tasks. To learn multiple tasks simultaneously, multi-task CNN has been proposed so far [4]. In the original work, it was applied to the face attribute detection task. Recently, it was applied to food category and ingredient estimation by Chen and Ngo [5]. They showed that simultaneous estimation boosted estimation performance on both tasks. Inspired by this work, we introduce a multi-task CNN for the task of simultaneous estimation of food category and calorie.

2.1 Implementation

The architecture of our multi-task CNN is based on VGG16 [6] as shown in Figure 2. We do not use dropout but use batch normalization for the *fc6* layer and the *fc7* layer. The *fc6* layer is shared by both tasks, and the *fc7* layer is branched to each task, so that each task has the *fc7* layer and the output layer independently. The food calorie estimation task has the *fc7c* layer with 4096 dimension and an output layer composed of one unit which outputs food calorie. The food categorization task has the *fc7f* layer with 4096 dimension and an output layer composed of units corresponding to each category. Figure 2 shows the overview of our multi-task CNN. Initially, each layer contained in the original VGG16 is pre-trained by ImageNet dataset. We define *L1* as the loss function of the food calorie estimation task and *L2* as that of the food categorization task, and denote N as the total number of learning data. The overall loss function L is as follows:

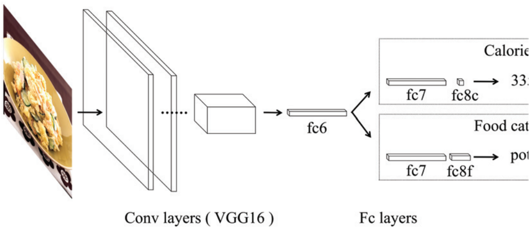


Figure 2. Overview of our Multi-task CNN.

$$L = -\frac{1}{N}\sum_{i=0}^{n}(\mathit{L1} + \lambda_1 \mathit{L2}) \tag{1}$$

The food calorie estimation task is treated as a regression problem. Generally, in the regression problem, a mean square error is used as the loss function, but in this paper we use loss function as shown below. We denote $\mathit{L1}_{ab}$ as an absolute error and $\mathit{L1}_{re}$ as an relative error, *L1* is defined as following:

$$\mathit{L1} = \mathit{L1}_{re} + \lambda_2 \mathit{L1}_{ab} \tag{2}$$

The absolute error is the absolute value of the difference between the estimated value and the ground-truth, and the relative error is the ratio of the absolute error to ground-truth. This loss function improves learning speed and performance. Let y_i as the estimated value of an image x_i and g_i as the ground-truth, $\mathit{L1}_{ab}$ and $\mathit{L1}_{re}$ are defined as following:

$$\mathit{L1}_{ab} = |y_i - g_i| \tag{3}$$

$$\mathit{L1}_{re} = \frac{|y_i - g_i|}{g_i} \tag{4}$$

Because the food categorization task is a multi-class classification problem, a soft-max function is used for the output layer of the food categorization task, and *L2* is calculated by the standard soft-max cross entropy.

3 Dataset

It is extremely costly to collect pair data of food images and food calories. For this reason, we focus on using commercial cooking recipe sites on the Web. For the experiments of this work, we collect these data from six websites shown in Table 1. Each recipe presented in these sites contains an ingredient list, descriptions on cooking procedure, food images, and the value of a food calorie. These websites do not provided the method for general users to post recipe information. All of the recipe sites except for “Ajinomoto” clearly indicate that professionals such as chefs provided recipe information which was expected to be reliable. In all the site, the food calorie written in the recipe pages indicated the amount of foods for serving for one person.

Since we use food categories on multi-task learning, it is necessary to collect food category information as well. In this paper, we limited collecting recipes on

Table 1. The commercial cooking recipe sites we used for data collection and the number of the collected recipes.

Website	Number of samples
Ajinomoto [1]	864
Kikkoman [2]	333
e-Recipe [3]	1190
Lettuce Club [4]	1329
Kyou no Ryouri [5]	534
Orange Page [6]	627

Table 2. The statistics on the built dataset.

Category	Number of samples
Pilaf	167
Curry	201
Fried rice	506
Fried noodle	326
Spaghetti	621
Gratin	400
Miso soup	869
Stew	205
Simmered meat and potatoes	183
Hamburg steak	379
Cold tofu	130
Sushi bowl	129
Omelet with fried rice	141
Potato salad	226
Mixed rice	394

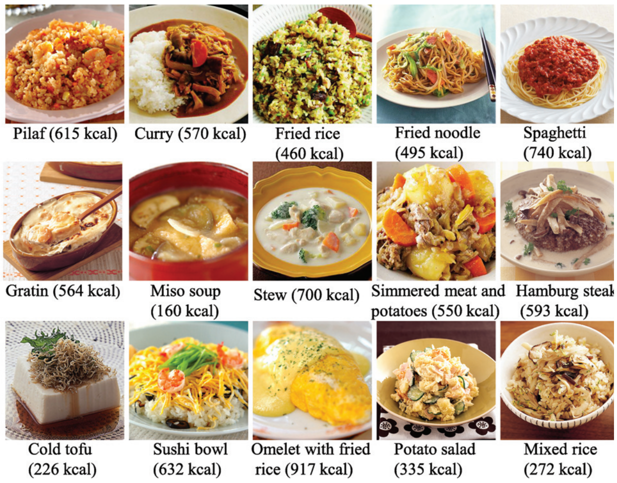
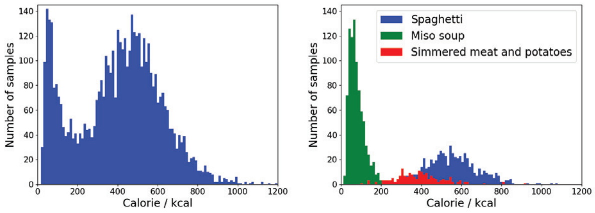


Figure 3. Examples of the collected food photos with categories and calories.



(a) All the 15 foods.

(b) "Spaghetti", "Miso soup" and "Simmered meat and potatoes".

Figure 4. The distribution of the calorie values of the collected recipes.

the representative 15 categories included in the UEC Food-100 [7]. The UEC Food-100 contains food photos of 100 kinds of Japanese foods without food calorie information. We collect recipe data which has food calorie information for one person. Then, we manually excluded the images with low resolution or multiple labels and images taken during cooking. In the end, a total of 4877 images were collected on 15 categories. The minimum food calorie is 18 kcal of "Miso soup", while the maximum value is 1199 kcal of "Curry".

Table 2 shows the number of recipes in each category, and Figure 3 shows some examples in the built dataset. Figure 4(a) shows the calorie distribution of all the collected recipes, and Figure 4(b) shows that of "Spaghetti", "Miso soup" and "Simmered meat and potatoes". As shown in Figure 4(b), the values of food calories tend to gather within a certain range depending on food categories. In some food categories, the range of food calories within the same food category are relatively broad, it is necessary to reflect the visual differences as shown in Figure 1.

4 Experiments

In the experiments, we estimated both food calories and food categories with a multi-task CNN. We made experiments with single-task CNNs for comparison as well. We used 70% of the food photos in the built dataset for training, 10% for validation during training, and the rest 20% for performance evaluation. For optimization of the CNNs, we used Momentum SGD with the learning rate, 0.001, and the momentum value, 0.9. The size of mini-batch was 8. We set λ_1 and λ_2 as 0.003 and 0.3 so as to balance the value scales of three loss functions, $L1_{ab}$, $L1_{re}$ and $L2$.

Table 3 shows the results on estimation of both calories and categories by single-task CNNs and a multi-task CNN. Regarding the calories, we shows the average of the absolute error representing the differences between the estimated values and the ground-truth, and the relative error representing the ratio between both. On the other hand, regarding food categories, we show the top-1 and top-5 classification accuracy. Table 3 indicates that by introducing the multi-task CNN, 11.59kcal and 2.22% were reduced on the absolute error and the relative error, and 2.05% and 0.61% were increased on the top-1 and top-5 accuracy.

For comparison to the existing work on direct food calorie estimation by Miyazaki et al. [2], we show the ratio of the estimated calories within 20% error and 40% error and the correlation coefficient which were used in [2] as the evaluation statistics in Table 4. Note that this comparison is only for reference and not formal, since the detailed conditions and the datasets are different. The dataset used in [2] contains multiple-dish food photos and has no limitation on food categories, both of which are different from our dataset. Compared with our results, our result by multi-task CNNs outperformed the result of [2] on the inclusion ratio with 20% ralative error and the correlation coefficient. Especially on the correlation, our method achieved 0.81 and 0.78 by multi-task and single-task, respectively, while their method was 0.32. From these results, our method outperformed their method greatly. The possible main reason is introducing CNN-based methods instead of hand-crafted feature based image search.

Figure 5 shows the relation between the ground truth values and the estimated calorie values by the single-task CNNs, while Figure 6 shows the relation by the multi-task CNN. The number of the estimated values which were below the -40% line was reduced.

Figure 7 and Figure 8 shows four examples of suc-

[1] http://park.ajinomoto.co.jp/
[2] https://www.kikkoman.co.jp/homecook/
[3] http://erecipe.woman.excite.co.jp/
[4] http://www.lettuceclub.net/recipe/
[5] http://www.kyounoryouri.jp/
[6] http://www.orangepage.net/

Table 3. The results on calorie and category estimation.

	Calorie regression		Categorization	
	abs. err.(kcal)	rel. err.(%)	Top-1(%)	Top-5(%)
Single-task	105.73	30.15	80.43	96.93
Multi-task	94.14	27.93	82.48	97.54

Table 4. Evaluation of calorie estimation.

	±20% err.(%)	±40% err.(%)	correlation
Baseline	35	79	0.32
Single-task	43	76	0.78
Multi-task	48	80	0.81

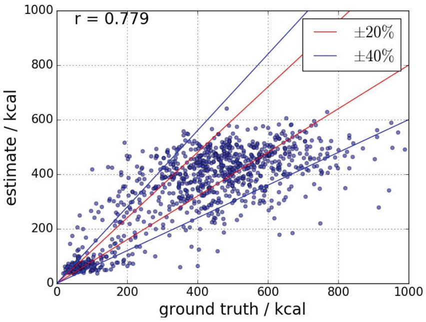


Figure 5. The relation between the ground truth values and the estimated calorie values by single-task CNNs. The correlation coefficient was 0.78.

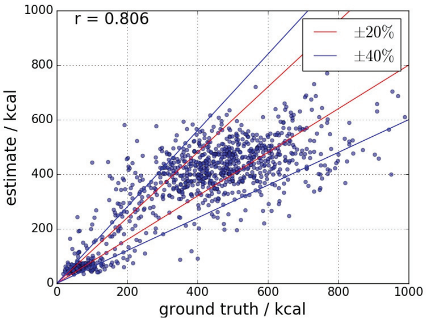


Figure 6. The relation between the ground truth values and the estimated calorie values by a multi-task CNN. The correlation coefficient was 0.81.

cessfully estimated results and wrongly estimated results, respectively. We found many results for relatively high-calorie foods with underestimated calorie values as erroneous results. We guess that the possible reason is that the number of samples with higher calories were not enough.

We examined the classification accuracy on the top-200 samples with the larger and the smaller error on calorie estimation. As results, the average classification accuracy on the top-200 samples with the larger error was 71%, while the average classification accuracy on the top-200 samples with the smaller error was 86%. This implies that the food photos which were difficult to classify their categories tends to be difficult to estimate their food calories as well.

5 Conclusions

In this paper, we proposed a new method for direct food calorie estimation using multi-task CNN with joint learning of food categories. In the experiments, we used a 15-category calorie-annotated food photo dataset created by Web image mining. By training food calories as well as food categories at the same time by multi-task CNN, the performance of both estimation was boosted compared to the result by independent single task CNNs. In addition, we saved the costs to build a calorie-annotated food photo dataset by collecting data from cooking recipe sites on the Web.

As future work, we plan to use the other recipe information such as food ingredients and how-to-cook for the training of the multi-task CNN.

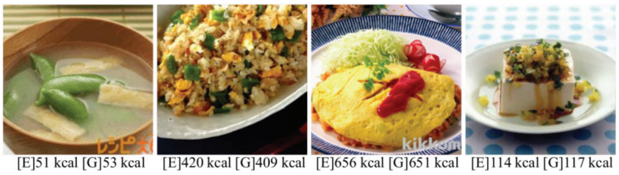


Figure 7. Examples on successful estimation of food calories. [E] and [G] represents the estimated calorie values and the ground-truth values, respectively.

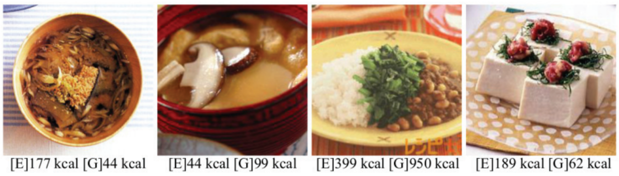


Figure 8. Examples on failure estimation of food calories.

Acknowledgment This work was supported by JSPS KAKENHI Grant Number 15H05915 (Grant-in-Aid for Scientific Research on Innovative Areas “Innovative SHITSUKAN Science and Technology”).

References

[1] A. Meyers and N. Johnston et. al. Im2Calories: towards an automated mobile vision food diary. In *Proc. of International Conference on Computer Vision*, pages 1183–1195, 2015.

[2] T. Miyazaki, G. Chaminda, D. Silva, and K. Aizawa. Image ‐ based calorie content estimation for dietary assessment. In *Workshop on Multimedia for Cooking and Eating Activities*, pages 363–368, 2011.

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, pages 91–99, 2015.

[4] H. A. Abrar, W. Gang, L. Jiwen, and J. Kui. Multi-task CNN model for attribute prediction. *IEEE Transactions on Multimedia*, pages 1949–1959, 2015.

[5] J. J. Chen and C. W. Ngo. Deep-based ingredient recognition for cooking recipe retrival. In *Proc. of ACM Multimedia*, 2016.

[6] K. Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *arXiv preprint arXiv:1409.1556*, 2014.

[7] Y. Matsuda and K. Yanai. Multiple-food recognition considering co-occurrence employing manifold ranking. In *Proc. of IAPR International Conference on Pattern Recognition (ICPR)*, 2012.