Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Food Constituent Estimation for Lifestyle Disease Prevention by Multi-Task CNN

Sulfayanti F. Situju, Hironori Takimoto, Suzuka Sato, Hitoshi Yamauchi, Akihiro Kanagawa & Armin Lawi

To cite this article: Sulfayanti F. Situju, Hironori Takimoto, Suzuka Sato, Hitoshi Yamauchi, Akihiro Kanagawa & Armin Lawi (2019) Food Constituent Estimation for Lifestyle Disease Prevention by Multi-Task CNN, Applied Artificial Intelligence, 33:8, 732-746, DOI: 10.1080/08839514.2019.1602318

To link to this article: https://doi.org/10.1080/08839514.2019.1602318

Published online: 23 Apr 2019.

APPLIED ARTIFICIAL INTELLIGENCE
2019, VOL. 33, NO. 8, 732–746
https://doi.org/10.1080/08839514.2019.1602318

Food Constituent Estimation for Lifestyle Disease Prevention by Multi-Task CNN

Sulfayanti F. Situju[a], Hironori Takimoto[b], Suzuka Sato[b], Hitoshi Yamauchi[b], Akihiro Kanagawa[b], and Armin Lawi[c]

[a]Graduate School of Computer Science and Systems Engineering, Okayama Prefectural University, Okayama, Japan; [b]Faculty of Computer Science and Systems Engineering, Okayama Prefectural University, Okayama, Japan; [c]Department of Computer Science, Hasanuddin University, Makassar, Indonesia

ABSTRACT
Unbalanced nutrition due to an unhealthy diet may increase the risk of developing lifestyle diseases. Many mobile applications have been released to record everyday meals for the health-conscious to enable them to improve their dietary habits. Most of these applications only base their food classification on an image of the food, requiring the user to manually input information about the ingredients such as the calories and salinity. To address this problem, food ingredient estimation from food images has been attracting increasing attention. Automatic ingredient estimation could possibly strongly alleviate the process of food-intake estimation and dietary assessment. In this paper, we propose an automatic food ingredient estimation method from food images by using multi-task CNN. We focus on classification of the food category and estimation of the calorie content and salinity for lifestyle disease prevention. Two-stage transfer learning using a large number of food category recognition image databases is applied to train our multi-task CNN for improved ingredient estimation. We experimentally analyze the relationship between the food category and salinity by using multi-task CNN.

Introduction

The progress in distribution systems and food production technology in recent years has enabled people to easily obtain their preferred food at any time. Consumption of retort food and opportunities for dining out are increasing as more people lead solitary lives and the dual income ratio increases. Maintaining such eating habits increases the risk of lifestyle diseases owing to biased nutritional balance. Lifestyle-related diseases include hypertension, dyslipidemia, diabetes, and others, but, because they have few subjective symptoms, these diseases progress silently and unnoticed to damage the brain, heart, blood vessels, and so on.

CONTACT Hironori Takimoto ✉ takimoto@c.oka-pu.ac.jp

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uaai.

Several mobile applications have been developed to record everyday foods with the aim of increasing the extent to which people are conscious of developing eating habits for good health (Caloriemama 2018; Foodlog 2018; MyFitnessPal, 2018; MyNetDiary 2018). These applications are especially useful for dietary assessment and planning. Automatic food image recognition methods have been widely proposed to enhance the capability of an application. Automatic food image classification could potentially alleviate the process of food-intake estimation and dietary assessment. In most of the cases, we can easily estimate food materials and ingredients based on the food category classified from the food image captured by a mobile device such as a smartphone. However, the estimated standard values are only standard guidelines for each category because food materials and ingredients are not unique to the captured food.

On the other hand, obesity is one of the leading causes of numerous diseases, such as heart disease, hypertension, and diabetes. The accurate estimation of the calorie content of food is effective to prevent obesity. Although humans would find it difficult to precisely estimate the calorie content from a food image, it is possible to roughly estimate whether the calorie content of the food in an image is high based on prior knowledge. Previous studies proposed methods capable of directly estimating the calorie content of food from an image in an attempt to implement this human ability on a mobile device.

Although salt is an essential nutrient in the human body, it is only needed in relatively small quantities. Health risks such as high blood pressure increase as the salinity intake increases. It is clear that estimating the salinity of food from an image of the food is a problem humans find more difficult than calorie estimation. Since estimating the salt content from food images is a difficult problem, no studies have focused on the automatic estimation of food salinity from food images thus far.

In this paper, we propose an automatic method for food ingredient estimation from a food image by using a multi-task Convolutional Neural Network (CNN) (Abrar e al., 2015). With the aim of achieving lifestyle disease prevention, we focus on the recognition of the food category and the estimation of calories and salinity. CNN based on deep learning has been demonstrated to achieve excellent results for image classification and object detection. We realize the effective estimation of calories and salinity using multi-task learning with food category classification, by defining both calorie and salinity estimation as a regression problem. The underlying assumption for multi-task learning algorithms is that different tasks are related to each other. Although a previous study reported that a food category is closely related to its calorie content, research to clarify the relationship between the food category and salinity has not yet been reported. In this paper, we experimentally demonstrate the relationship between the food category and salinity by using multi-task CNN.

Then, the Xception model (Chollet 2016), which has achieved a high recognition rate in image classification tasks using Image-Net, is used as basis for the proposed architecture. The benchmark dataset we constructed for the evaluation of the proposed method is an original image dataset created by using publicly available images from several recipe-gathering websites. To improve the estimation accuracy, we proposed two-stage transfer learning using a large number of food category classification image databases. CNN is considered to require a large number of training images to achieve comparable or superior performance to the conventional local-feature-based methods. We proposed the two-stage transfer learning using a large number of food image databases for food category classification because it is difficult to collect many food images of which the calorie content and salinity are known.

Related Work

Food image recognition methods and food image databases have been widely developed because food image recognition enables nutrient estimation and health-care analysis corresponding to people‘s eating habits. Existing methods are divided into two categories: the handcrafted feature-based and deep learning-based approaches.

First, we describe the handcrafted feature-based approach. Many researchers have attempted to solve the problem of food image classification by using simple low-level feature extraction and coding methods. Joutou proposed an MKL-based feature fusion method, which adaptively integrates various kinds of image features such as color, texture, and BoF for food image recognition (Joutou and Yanai 2009). Yang proposed a method to analyze the ingredient relations in the food image by computing pairwise statistics between the local features (Yang et al. 2010). The probability with which a pixel belongs to each food ingredient category is calculated by using the semantic text on forests (Shotton, Johnson, and Cipolla 2008). The color information and calculated probability are employed to extract statistics of pairwise local features, to form a multidimensional histogram. Finally, the food image is classified by applying the obtained histogram to a multiclass SVM using a χ^2 kernel. Rahmana et al. proposed a method for generating scale and/or rotation invariant global texture features using the output of Gabor filter banks (Rahmana et al., 2012). They demonstrated that the proposed texture feature provides sufficient food classification accuracy in texture rich data. Kong et al. proposed "DietCam," which is an automatic camera-phone-based multi-view food classifier as part of a food intake assessment system (Kong and Tan 2012). DietCam detects food ingredients by using a deformable part-based model and a texture verification model. A food category is classified by using the detected ingredients and a multi-view multikernel SVM. He et al.

proposed an image segmentation and classification method to detect the food regions in an image, and to classify the food category (He et al. 2013). In addition, this method estimated the weight of food to extract the nutrient content from a food image using a shape template for foods with regular shapes and area-based weight estimation for foods with irregular shapes. Aizawa et al. developed a food-image detection and food-balance estimation method (Aizawa et al. 2013). First, food-image detection classifies the image into “food” or “nonfood” by employing supervised learning based on multiple image features and SVM. Then, food-balance estimation is used to estimate the food balance of the meal shown in each photograph by using the global color, circle, BoF, and block features and proposed improving the performance with the personal likelihood. Anthimopoulos et al. propose a BoF-based system for food image classification. First, a visual dictionary of 10,000 visual words is created from dense local features based on a scale-invariant feature transform on the HSV color space (Anthimopoulos et al. 2014). The food images are then classified by SVM.

Considering all of the previously described methods, it seems the best approach is to use a complex combination of a large number of image features. Some food types have a high intra-class and low inter-class variance since foods are typically deformable objects. Handcrafted feature-based approaches have achieved low classification accuracy because of these characteristics of food images.

On the other hand, the visual recognition paradigm changed rapidly after the appearance of the ImageNet dataset, demonstrating the power of data-driven feature learning. During the past years, CNN has become the most effective architecture to perform visual recognition. Hence, many studies using the deep-learning-based approach for food recognition have been reported (Christodoulidis, Anthimopoulos, and Mougiakakou 2015; Hassannejad et al. 2016; Kagaya, Aizawa, and Ogawa 2014; Kawano and Yanai 2014; Liu et al. 2016; Singla, Yuan, and Ebrahimi 2016; Yanai and Kawano 2015). These researchers demonstrated that it is possible to classify with higher precision than existing handcrafted feature-based approaches by conducting experiments using food image datasets. Myers et al. reported a deep-learning-based approach for calorie estimation (Meyers et al. 2015), and proposed the Im2Calories system for food recognition. The system made extensive use of CNN. They used the architecture of GoogLeNet (Szegedy e al., 2015) and fine-tuned the pre-trained model on Food101 (Food-101 dataset 2018). Pouladzadeh et al. proposed a smartphone-based system to estimate the calories contained in images of food taken by the user (Pouladzadeh et al. 2016). These works require reference information to estimate the quantity of food on the plate. In addition, they proposed a mobile food recognition system to estimate the calories and nutrition value of a meal by recognizing multiple food items in the same meal from a captured image (Pouladzadeh and Shirmohammadi 2017).

Abrar et al. proposed a joint multi-task learning algorithm to effectively estimate several image attributes using CNN (Abrar et al. 2015). In this multi-task CNN, effective learning of CNN is achieved because attributes in the same group are prompted to share more knowledge whereas attributes in different groups generally compete with each other. By using multi-task CNN, Chen et al. achieved food category and ingredient estimation from food images (Chen and Ngo 2016). Their paper reported that simultaneous estimation boosted the estimation performance on both tasks. Ege and Yanai proposed a simultaneous estimation method of food categories and calories with multi-task CNN (Ege and Yanai 2017). Although the accuracies of both tasks were improved compared with single-task CNN, the improvement of this work is slight because the number of datasets to train the multi-task CNN is insufficient.

Although many studies on category identification and calorie estimation have been published, the estimation of salinity from a food image has not yet been reported. In addition, a large training dataset is required to derive the performance of multi-task CNN even though it is effective to use multi-task CNN with category classification to estimate food ingredients. However, it is extremely costly to collect ingredient-annotated data of food images.

Materials and Methods

Overview

The aim of our research is to more accurately estimate the food category and ingredients from a food image by multi-task CNN. Excess calorie and salt intake poses strong health risks such as heart disease, hypertension, and high blood pressure. This led us to focus on the estimation of the calorie content and salinity of food.

The architecture of our multi-task CNN is shown in Figure 1. The proposed architecture is based on the Xception model. After feature extraction by Xception, our model branches into three tasks: food category

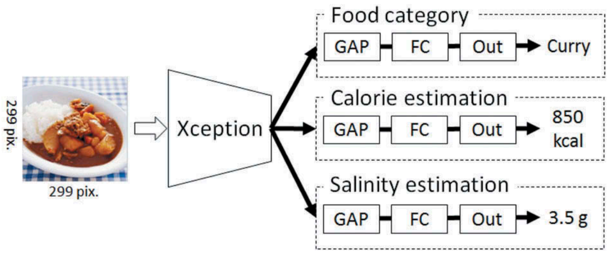


Figure 1. Architecture of proposed multi-task CNN.

classification, calorie estimation, and salinity estimation. In Figure 1, GAP, FC, and OUT denote the global average pooling layer, the fully connected layer, and the output layer, respectively.

A food ingredient-annotated food image dataset including both calorie content and salinity does not yet exist because other researchers have not focused on salinity estimation. In this study, we built a food image dataset annotated by calories and salinity. Effective and efficient training using multi-task CNN is achieved by continuously fine-tuning the multi-task CNN by using a small-scale ingredient-annotated dataset and a middle-scale dataset with only the categories annotated, as a two-stage fine-tuning procedure.

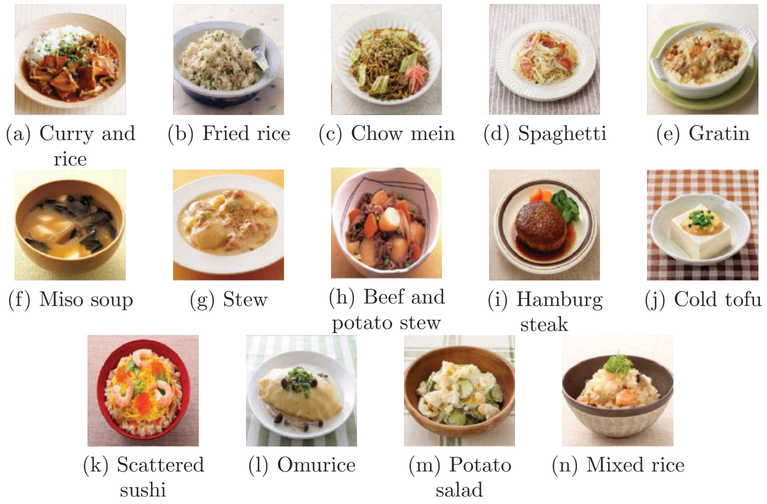
The contribution of the paper is summarized as follows,

- We built a food image dataset in which calories and salinity are annotated.
- We propose a multi-task CNN based on Xception architecture to classify the food category, and to estimate food ingredients.
- We propose two-stage transfer learning using an image database for category classification to achieve highly accurate classification and estimation by using a small dataset.
- We experimentally analyze the relationship between the food category and salinity by using the multi-task CNN.

Dataset

A large food image dataset annotated by calories and salinity has not been distributed thus far. We collected a large number of food images annotated with both types of information from six commercial cooking recipe sites on the Web (Ajinomoto Park, 2018; Lettuce Club News 2018; Kewpie 2018; Kikkoman Homecook 2018; Mizkan 2018; Orangepage net 2018). The recipe information published on these sites was provided by experts such as cooks and cooking researchers. Although the calorie and the salinity information provided on these recipe sites are for one person, the amounts specified for some food images are for multiple people. These food images for multiple people are excluded from the dataset since we focus on estimating the calorie and salinity content from a food image for one person.

The representative 14 categories, which are shown in Figure 2, included in the UEC Food-100 (UEC FOOD 100 2018), are collected. The UEC Food-100 contains food images of 100 kinds of Japanese foods. The 14 selected categories are included in the food category considered in the research by Ege and Yanai. (Ege and Yanai 2017).



(a) Curry and rice (b) Fried rice (c) Chow mein (d) Spaghetti (e) Gratin

(f) Miso soup (g) Stew (h) Beef and potato stew (i) Hamburg steak (j) Cold tofu

(k) Scattered sushi (l) Omurice (m) Potato salad (n) Mixed rice

Figure 2. Examples of food images for each collected category.

We excluded the low-resolution images or those with multiple labels from the collected images. All images were resized to 299 × 299 pixels. As a result of the aforementioned processing, a total of 3,051 images were collected on 14 categories. Details of the food image dataset we built are provided in Table 1.

Architecture of Our Multi-Task CNN

The proposed architecture, which is shown in Figure 1, is based on the Xception model. Xception is a CNN architecture inspired by Inception, where Inception modules have been replaced with depth-wise separable convolutions. This architecture significantly outperforms Inception V3 on a larger image classification dataset comprising 350 million images and 17,000 classes. Details of the Xception architecture are shown in Figure 3. In the figure, the global average-pooling layer, the optional fully connected layer, and the logistic regression at Exit flow are excluded from the original Xception architecture. This architecture is used as the common layer for all tasks.

The network branches to each task from the global average-pooling layer. Each task has a global average pooling layer, a fully connected layer with a dropout, and an output layer, respectively. The branched networks are adjusted by specializing in different tasks, namely the classification and regression tasks. The food classification task has a fully connected layer with 512 dimensions and an output layer corresponding to each food category. The calorie and salinity estimation task comprises a fully connected layer with 512 dimensions and an output layer composed of one unit, respectively.

Table 1. Details of collected food images for multi-task CNN.

Category	# of images	Calorie (kcal)		Salinity (g)	
		Ave.	S.D.	Ave.	S.D.
Curry and rice	214	531.14	220.57	2.630	1.045
Fried rice	217	469.83	165.32	2.310	0.995
Chow mein	140	552.62	118.86	3.202	0.921
Spaghetti	565	573.50	116.98	2.656	0.868
Gratin	264	397.99	160.43	1.985	0.836
Miso soup	373	84.65	52.20	1.946	0.643
Stew	136	379.53	119.70	2.331	0.914
Beef and potato stew	154	378.24	126.35	2.357	1.026
Hamburg steak	226	395.57	108.60	2.254	0.823
Cold tofu	114	141.56	57.30	1.235	0.594
Scattered sushi	107	534.91	129.93	2.961	1.122
Omurice	105	683.63	132.44	2.917	0.776
Potato salad	210	230.40	93.49	1.205	0.583
Mixed rice	226	409.65	97.20	1.814	0.751
Whole	3051	416.30	211.90	2.280	0.999

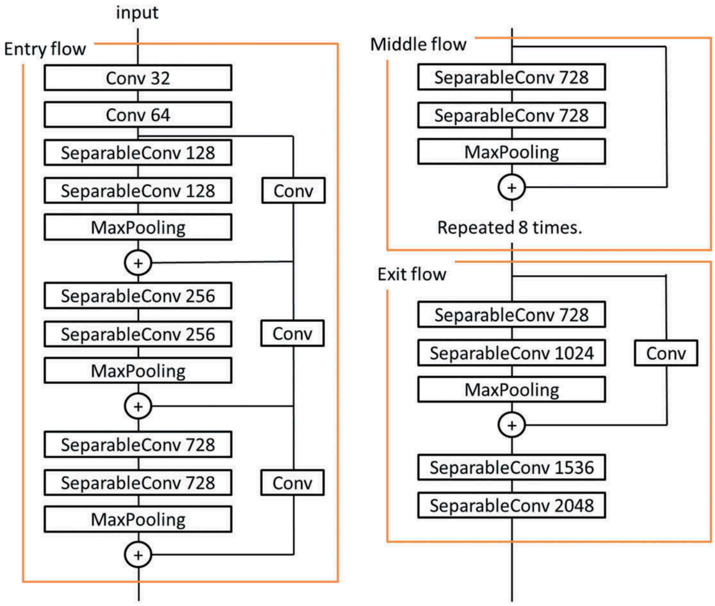


Figure 3. Xception architecture in our multi-task CNN.

The proposed multi-task CNN is trained based on the following loss function L .

$$L = \frac{1}{N}\sum_{i=0}^{N}(w_{cat}L^i_{cat} + w_{cal}L^i_{cal} + w_{sal}L^i_{sal}) \tag{1}$$

where L_{cat} , L_{cal} , and L_{sal} are the loss functions of the food classification, food calorie estimation, and salinity estimation tasks, respectively. Further,

w_α ($\alpha = \{cat, cal, sal\}$) are the weight coefficients for each loss function so as to balance the value scales of the three loss functions. N is the number of image samples.

A soft-max function is used for the output layer of the food classification task since this task is a multi-class classification problem. L_{cat} is defined based on the standard soft-max cross entropy.

$$L_{cat} = -\sum_{i=0}^{M} t_i \log y_i, \tag{2}$$

where t_i represents the ground-truth of the i th unit, which is binary, y_i is the output of the i th unit, and M is the number of food categories.

The food calorie task and salinity estimation task are treated as regression problems. Although the mean square error is generally used as the loss function in a regression problem, we employ the loss function by combining different errors, as proposed by Ege and Yanai ?.

$$L_{cal} = w_{cal,re}E_{cal,re} + w_{cal,ab}E_{cal,ab} \tag{3}$$

$$L_{sal} = w_{sal,re}E_{sal,re} + w_{sal,ab}E_{sal,ab} \tag{4}$$

where each w is the weight coefficient so as to balance the value scales of these errors. Each of the errors is defined by:

$$E_{cal,ab} = |y_{cal} - g_{cal}| \tag{5}$$

$$E_{cal,re} = |y_{cal} - g_{cal}|/g_{cal} \tag{6}$$

$$E_{sal,ab} = |y_{sal} - g_{sal}| \tag{7}$$

$$E_{sal,re} = |y_{sal} - g_{sal}|/g_{sal} \tag{8}$$

where y and g of each error are the estimated value and the ground-truth, respectively. The absolute error $E_{*,ab}$ is the absolute value of the difference between the estimated value and the ground-truth. The relative error $E_{*,re}$ is the ratio of the absolute error to the ground-truth.

Two-Stage Fine-Tuning

When applied to areas where large-scale data are much more difficult to gather, CNN has still proven effective through the use of transfer learning

(Oquab et al. 2014). Pre-trained CNNs are used as weight initializers for fine-tuning to the new target task. In general, the ImageNet dataset, which contains 50,000,000 images with 1,000 categories, is used for pre-training the CNN. ImageNet is not necessarily strongly related to the food ingredient estimation task in this paper because it is a dataset for generic object recognition. When fine-tuning a multi-task CNN for food category classification and food ingredient estimation using the pre-trained model with ImageNet, it is clear from the report of the previous study that the accuracy improvement is slight if the number of images per class is not large.

We address this issue by focusing on two-stage fine-tuning. First, we prepare a category-annotated food image dataset. The number of images in this dataset is larger than the dataset consisting of collected images with both ingredients annotated for the main task, but is smaller than ImageNet. Then, CNN pre-trained by ImageNet is again fine-tuned for food category classification by using only the category-annotated food image dataset. Finally, this CNN is again fine-tuned for food category classification and food ingredient estimation by using the main task dataset, which contains 3,051 images, as a two-stage fine-tuning process.

We collected category-annotated food images from the Rakuten recipe dataset, which is part of the Rakuten dataset Rakuten18. This dataset has 14 categories, which is the same as the dataset for the main task. The collected image limit for each category is 3,000. We excluded low-resolution images or those with multiple labels from the collected images. All images were resized to 299 × 299 pixels. As a result of the above processing, a total of 28,359 images were collected in 14 categories. The number of collected images for each category is shown in Table 2.

Experiments

Experimental Setup

In this study, we constructed two food image datasets, an ingredient-annotated dataset, and a category-only-annotated dataset for two-stage fine-tuning. In the ingredient-annotated dataset, 90% of the images in each

Table 2. Details of collected food images for two-stage fine-tuning.

Category	# of images	Category	# of images
Curry and rice	2787	Fried rice	2999
Chow mein	2883	Spaghetti	3000
Gratin	2095	Miso soup	3000
Stew	1619	Beef and potato stew	736
Hamburg steak	2500	Cold tofu	536
Scattered sushi	612	Omurice	908
Potato salad	1684	Mixed rice	3000

category was used for training the CNN, and the remaining images were used for testing. In the category-only-annotated dataset, 80% of the images in each category was used to train the CNN, and the remainder was used for testing.

We evaluated the performance of the proposed multi-task CNN with two-stage fine-tuning. We demonstrated the effectiveness of multi-task learning and two-stage fine-tuning, respectively, by comparing three methods: a single-task CNN, multi-task CNN without two-stage fine-tuning, and multi-task CNN with two-stage fine-tuning. The original Xception architecture is used for the single-task CNN. The single-task CNN is optimized for three tasks: food image classification, calorie estimation, and salinity estimation.

The initial learning rate is set to 0.001 and gradually decreases to 0.0001. All of the models were trained by Adagrad in 100 epochs. The dropout rate for each task is set to 0.5. The batch size is set to 16. We experimentally set the weight for each loss function to $w_{cat} = 1.0, w_{cal} = 0.4$, and $w_{sal} = 0.9$, respectively. We also set the weight for each loss function of each task to $w_{cal,re} = 1.0, w_{cal,ab} = 0.3, w_{sal,re} = 1.0$, and $w_{sal,ab} = 0.5$, respectively.

The classification accuracy was used as the criterion for evaluating the food image classification. The absolute error E_{ab} , relative error E_{re} , and correlation coefficients were also used as criteria for calorie and salinity estimation.

Results and Discussion

The results of the three tasks using all the methods are shown in Table 3. First, we discuss the results of category classification. The classification accuracy of the proposed multi-task CNN is higher than that of single-task CNN. Some miss-classification occurred between two categories in single-task CNN: “Fried rice” and “Mixed rice”, “Chow mein” and “Spaghetti”. It is generally difficult to classify these categories because each of these categories mostly contains rice and noodles. However, miss-classification between “Fried rice” and “Mixed rice” is improved in multi-task CNN with and without two-stage fine-tuning.

Table 3. Comparison of single-task CNN, multi-task CNN, and multi-task CNN with two-stage fine-tuning.

	Calorie			Salinity			Category
Methods	E_{ab} (kcal)	E_{re} (%)	CC	E_{ab} (g)	E_{re} (%)	CC	Accuracy (%)
Single-task CNN	100.2	41.7	0.80	0.75	37.2	0.40	90.0
Multi-task CNN	94.6	36.6	0.82	0.76	36.8	0.43	91.0
Multi-task CNN with two-stage fine tuning	**89.6**	**31.2**	**0.84**	**0.74**	**36.1**	**0.45**	**92.6**

The results of ingredients estimation is discussed next. The improvement seen in the calorie estimation task was significantly high. On the other hand, the salinity estimation with the multi-task CNN with two-stage fine-tuning showed slight improvement relative to the single-task CNN. The advantage of multi-task CNN is to effectively and efficiently learn the common features that contribute to each task if tasks are strongly correlated. In addition, the CNN is fine-tuned to enable the extraction of useful features for category classification in two-stage fine-tuning using a large number of images for food category classification. Therefore, the difference in improvement between calorie and salinity estimation is assumed to be attributable to the correlation difference between the food category classification task and the estimation of each ingredient. As a result, we consider the correlation between category classification and salinity estimation to be weaker than calorie estimation, from the point of view of an image feature.

On the other hand, in Table 1, the ratio of intra-class variance to interclass variance of the salt content is larger than that of the calories. Seemingly, the task of estimating the salinity is more difficult compared to that of estimating the calories. Even in such a situation, both the errors and correlation coefficient of the salinity estimation were improved in our methods. From these results, we demonstrated that the proposed multi-task CNN with two-stage fine-tuning correctly classifies the food category and estimates the calories and salinity because the multi-task CNN is superior to the single-task CNN, and the multi-task CNN with two-stage fine-tuning outperforms the multi-task CNN without two-stage fine-tuning.

Detailed results of the calorie and salinity estimation for each category are listed in tables 4 and 5. In the calorie estimation, the error in categories with a large mean and large intra-class variance is large. In contrast, it was difficult

Table 4. Comparison of calories estimation.

Category	Single-task		Multi-task		Multi-task w/2-F.T.	
	E_{ab} (kcal)	E_{re} (%)	E_{ab} (kcal)	E_{re} (%)	E_{ab} (kcal)	E_{re} (%)
Curry and rice	113.8	20.1	103.3	17.6	95.6	17.3
Fried rice	73.8	13.1	81.9	15.2	80.1	14.3
Chow mein	76.6	12.7	81.8	13.9	68.2	11.4
Spaghetti	98.7	16.2	97.2	16.0	86.3	14.0
Gratin	132.6	60.0	110.9	51.1	123.7	54.1
Miso soup	37.5	104.9	31.3	81.2	24.8	55.6
Stew	140.0	47.8	134.6	41.0	126.0	33.1
Beef and potato stew	94.8	32.0	79.5	28.6	104.8	33.6
Hamburg steak	101.8	31.6	117.3	34.6	81.6	24.5
Cold tofu	67.7	62.4	61.5	56.3	68.3	52.6
Scattered sushi	152.4	23.6	139.8	21.4	112.4	16.6
Omurice	172.4	25.5	119.3	20.3	179.1	27.5
Potato salad	128.0	84.7	112.6	73.0	92.1	60.4
Mixed rice	109.5	22.6	113.1	23.7	121.0	24.4

Table 5. Comparison of salinity estimation.

Category	Single-task		Multi-task		Multi-task w/2-F.T.	
	E_{ab} (g)	E_{re} (%)	E_{ab} (g)	E_{re} (%)	E_{ab} (g)	E_{re} (%)
Curry and rice	0.64	28.5	0.51	24.5	0.40	18.5
Fried rice	0.91	51.9	1.10	64.9	1.04	56.8
Chow mein	0.86	22.8	0.75	21.1	0.94	28.2
Spaghetti	0.77	32.5	0.80	34.6	0.74	31.0
Gratin	0.70	43.7	0.61	32.2	0.60	33.2
Miso soup	0.54	35.7	0.50	34.8	0.57	38.1
Stew	0.71	31.5	0.78	38.7	0.81	36.8
Beef and potato stew	0.92	59.4	1.26	67.9	1.03	62.7
Hamburg steak	0.74	37.0	0.86	39.4	0.88	41.2
Cold tofu	0.66	57.5	0.66	49.5	0.52	40.6
Scattered sushi	1.42	33.5	1.15	25.0	1.30	28.8
Omurice	0.87	25.4	0.72	26.4	0.77	26.9
Potato salad	0.45	40.3	0.33	26.1	0.50	42.5
Mixed rice	0.90	32.7	1.03	39.6	0.80	30.1

to find some correlation in the salinity estimation results. However, we confirmed the effectiveness of the proposed method since each absolute estimated error is smaller than the standard deviation in each category.

Conclusions

In this paper, we proposed a method to estimate food ingredients from a food image by multi-task CNN. As a benchmark dataset for evaluation of the proposed method, we constructed a new image dataset using publicly available images from several recipe-gathering websites. We proposed two-stage transfer learning using a large number of food category recognition image databases to improve the estimation accuracy because it is difficult to collect a large number of food images of which the calorie content and salinity are known. We demonstrated the effectiveness of multi-task learning with food category classification to estimate the calorie and salinity content. We experimentally confirmed the existence of a relationship between the food category and salinity.

Funding

This research was supported in part by a Grant-in-Aid for Scientific Research (C) grant from the Japan Society for the Promotion of Science (Grant No. 17K01573).

References

Abrar, H. A., W. Gang, L. Jiwen, and J. Kui. 2015. Multi-task CNN Model for Attribute Prediction. *IEEE Transactions on Multi-Scale Computing Systems* 17 (11):1949–59. doi:10.1109/TMM.2015.2477680.

Aizawa, K., Y. Maruyama, H. Li, and C. Morikawa. 2013. Food Balance Estimation by using Personal Dietary Tendencies in a Multimedia Food Log. *IEEE Transactions on Multi-Scale Computing Systems* 15 (8):2176–85. doi:10.1109/TMM.2013.2271474.

Ajinomoto Park, https://park.ajinomoto.co.jp/(accessed February 15, 2018).

Anthimopoulos, M., L. Gianola, L. Scarnato, P. Diem, and S. Mougiakakou. 2014. A Food Recognition System for Diabetic Patients Based on an Optimized Bag-of-Features Model. *IEEE Journal of Biomedical and Health Informatics* 18 (4):1261–71. doi:10.1109/JBHI.2014.2308928.

Caloriemama, http://www.caloriemama.ai/(accessed February 1, 2018).

Chen, J. J., and C. W. Ngo. 2016. "Deep-based Ingredient Recognition for Cooking Recipe Retrieval", Proc. of ACM Int. Conf. Multi.: 32–41, Amsterdam, The Netherlands.

Chollet, F. 2016. "Xception: Deep Learning with Depthwise Separable Convolutions", In arXiv preprint arXiv:1610.02357.

Christodoulidis, S., M. M. Anthimopoulos, and S. G. Mougiakakou. 2015. "Food Recognition for Dietary Assessment using Deep Convolutional Neural Networks", Proc. of the Int. Conf. on Image Analysis and Processing 2015: 458–65, Niagara Falls, Canada.

Ege, T., and K. Yanai. 2017. "Simultaneous Estimation of Food Categories and Calories with Multi-task CNN", Proc of 2017 Fifteenth IAPR Int. Conf. on Machine Vision App. (MVA2017): 198–201, Nagoya, Japan.

Food-101 dataset, https://www.vision.ee.ethz.ch/datasets_extra/food-101/(accessed February 15, 2018).

Foodlog, http://www.foodlog.jp/(accessed February 1, 2018).

Hassannejad, H., G. Matrella, P. Ciampolini, I. D. Munari, M. Mordonini, and S. Cagnoni. 2016. "Food Image Recognition using Very Deep Convolutional Networks", Proc. of the Int. Workshop on Multi. Assisted Dietary Manag.: 41–49, Amsterdam, The Netherlands.

He, Y., C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp. 2013. "Food Image Analysis: Segmentation, Identification and Weight Estimation", Proc. of the 2013 IEEE Int. Conf. on Multi. and Expo (ICME): 1–6, California, USA.

Homecook, K., https://www.kikkoman.co.jp/homecook/(accessed February 15, 2018).

Joutou, T., and K. Yanai. 2009. "A Food Image Recognition System with Multiple Kernel Learning", Proc. of the Int. Conf. on Image Processing(ICIP2009): 285–88, Cairo, Egypt.

Kagaya, H., K. Aizawa, and M. Ogawa. 2014. "Food Detection and Recognition using Convolutional Neural Network", Proc. of the ACM Multimedia Conf: 1055–88, Florida, USA.

Kawano, Y., and K. Yanai. 2014. "Food Image Recognition with Deep Convolutional Features", Proc. of ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing (UbiComp): 589–93, Seattle, USA.

Kewpie, https://www.kewpie.co.jp/recipes/(accessed February 15, 2018).

Kong, F., and J. Tan. 2012. "Dietcam: automatic dietary assessment with mobile camera phones", Pervasive and Mobile Comp.: 8 (4):147–63.

Lettuce Club News, https://www.lettuceclub.net/recipe/(accessed February 15, 2018).

Liu, C., Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma. 2016. "DeepFood: deep learning-based food image recognition for computer-aided dietary assessment", Proc. of the Int. Conf. On smart homes and health Telematics: 37–48, Wuhan, China.

Meyers, A., N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and P. K. Murphy. 2015. "Im2calories: towards an automated mobile vision food diary", Proc. of IEEE Int. Conf. on Comp. Vision (ICCV2015): 1233–41, Santiago, Chile.

Mizkan, http://www3.mizkan.co.jp/sapari/menu/cook/search/(accessed February 15, 2018).

MyFitnessPal.com, http://www.myfitnesspal.com/(accessed February 1, 2018).

MyNetDiary, http://www.mynetdiary.com/(accessed February 1, 2018).

Oquab, M., L. Bottou, I. Laptev, and J. Sivic. 2014. "Learning and transferring mid-level image representations using convolutional neural networks", Proc. of IEEE Conf. on Comp. Vision and Pattern Recog. (CVPR2014): 1717–24, Ohio, USA.

Orangepage net, http://www.orangepage.net/(accessed February 15, 2018).

Pouladzadeh, P., P. Kuhad, S. V. B. Peddi, A. Yassine, and S. Shirmohammadi. 2016. "Food calorie measurement using deep learning neural network", Proc. of IEEE Int. Instrumentation and Measurement Tech. Conf. (I2MTC): 1–6, Taipei, Taiwan.

Pouladzadeh, P., and S. Shirmohammadi. 2017. "Mobile multi-food recognition using deep learning". *ACM Transactions on Multimedia Computing, Communications, and Applications* 13 (3s): Article 36, 1–21.

Rahmana, M. H., M. R. Pickering, D. Kerr, C. J. Boushey, and E. J. Delp, "A new texture feature for improved food recognition accuracy in a mobile phone based dietary assessment system", Proc. of the 2012 IEEE Int. Conf. on Multi. and Expo Workshops (ICMEW): 418–23, Melbourne, Australia.

Rakuten recipe, https://www.nii.ac.jp/dsc/idr/rakuten/rakuten.html (accessed February 15, 2018).

Shotton, J., M. Johnson, and R. Cipolla. 2008. "Semantic text on forests for image categorization and segmentation", Proc. of the Conf. on Comp. Vision and Pattern Recog. (CVPR2008): 1–8, Alaska, USA.

Singla, A., L. Yuan, and T. Ebrahimi. 2016. "Food/non-food image classification and food categorization using pre-trained GoogLeNet model", Proc. of the Int. Workshop on Multi. Assisted Dietary Manag.: 3–11, Amsterdam, The Netherlands.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going Deeper with Convolutions. *Conference on Computer Vision and Pattern Recognition* 1–9.

UEC FOOD 100, http://foodcam.mobi/dataset100.html (accessed February 15, 2018).

Yanai, K., and Y. Kawano. 2015. "Food image recognition using deep convolutional network with pre-training and fine-tuning", Proc. of the 2015 IEEE Int. Conf. on Multi. and Expo Workshops: 1–6, Torino, Italy.

Yang, S., M. Chen, D. Pomerlau, and R. Sukthankar, "Food recognition using statistics of pairwise local features" Proc. of the Conf. on Comp. Vision and Pattern Recog. 2010 (CVPR2010): 2249–56, California, USA.