ISSN 1870-4069

A First Approach to Food Composition Estimation as an Image Classification Problem

Jorge Alberto Cabrero-Dávila, Humberto Pineda-Ivo

Benemérita Universidad Autónoma de Puebla,
Departamento de Ciencias Computacionales,
México

{jorgedavila33, ivopinedatorres}@gmail.com

Summary. Food related degenerative diseases levels have increased dramatically in the last decades. Food and diet guidelines promoted by governments and health agencies haven't had the desired effect on reversing these trends. One of the principal explanations is that guidelines are hard to follow and are unclear. Deep Learning algorithms can help to make easier for people to follow dietary guidelines using technology available to them. In this work, a first approach is proposed to solve an early stage of the entire scope of the problem which is the estimation of calories and nutrients in food based on their image.

Keywords: Food classification, image classification, deep learning.

1 Introduction

Calorie counting is one of the two major components in almost all issued health guidelines. Research has shown that restricting calories helps reduce weight and increase health markers such as blood pressure, insulin resistance, cholesterol levels, liver health, etc. [1]. The other major component is nutrient density of food, it is well known that the body needs certain nutrients (vitamins, minerals, and fiber) and macro-nutrients (proteins, carbs, and fats) to be healthy and that these nutrients and macro-nutrients are only available in food and can't be obtain elsewhere [2].

Following the major components of guidelines; a diet, is composed by a recommended number of calories based on age, gender, current weight, and activity. Also, by several levels of micro and macro nutrients, measured in grams, milligrams or micro-grams depending on the nutrient.

For example, a 25-year-old male with average activity of 30 minutes a day could need a diet of 2500 calories, 1 gram of protein per kilo of body weight, 30 grams of carbs, 10 milligrams of zinc, 30 milligrams of manganese, etc.

With that target in mind there are countless ways to achieve those levels and it is very easy to overpass those limits calorie-wise and be in the low side nutrient-wise. To stick to the recommended, the person should first know the composition of a lot of food,

Table 1. Datasets used.

	Images	Classes	Original Size	Size after treatment
Food 101	101,000	101	512 x 384	224 x 224
ECUSTD	2,987	20	816 x 551	224 x 224

Fig. 1. ECUSTD Data base examples. From left to right from top to bottom: apple, banana, bread, bun, donut, egg, lemon, mango, pear, kiwi, plum, sachiman, orange, mix. Source: compiled by the authors based on ECUSTD Database.

Fig. 2. Food 101 Data base examples. From left to right from top to bottom: baklava, beef tartare, Caesar salad, ceviche, deviled eggs, filet mignon, fried rice, hamburger, pizza. Source: compiled by the authors based on Food 101 Database.

and secondly, they should know exactly the amount they're eating, both aspects are difficult and impractical in a modern lifestyle.

If only there was a way to easily recognize, access and measure food composition without weight it or search for it would mean a major step for achieving better public health around the globe. In this work we believe that computer vision paired with deep learning algorithms can achieve just that by just taking a photo with a cell phone or the like of the food that it's about to be eaten.

2 State of the Art

Currently there aren't many works on the subject in question; however, tech giant Google have already started to research and develop in this matter with its leading scientist Kevin Murphy. They already have developed Im2calories an app that can count calories from a low-resolution photo taken with a cellphone of a dish [3], it only works in 30 percent of the cases, but they are improving it with people inputs since 2015. The model is not open to the public but is based on a convolutional neural network.

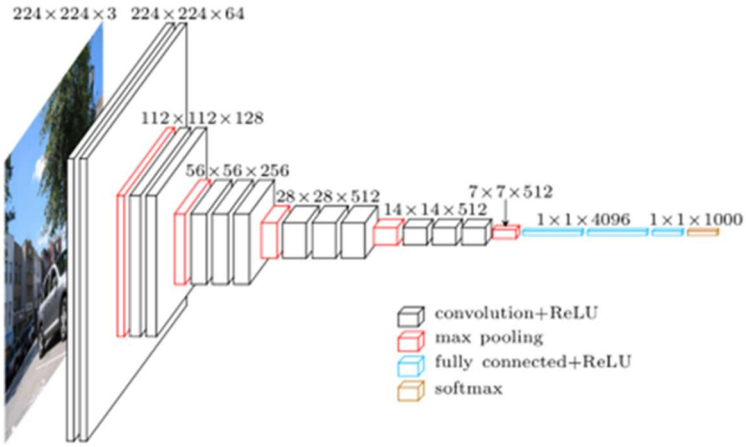

Fig 3. VGG16 Architecture [11].

Table 2. GoogLeNet Architecture [10].

Operator	Input
Conv1	3 x 224 x 224
Pool1	64 x 112 x 112
Conv2	64 x 56 x 56
Pool2	192 x 56 x 56
Inception3a	192 x 28 x 28
Inception3b	256 x 28 x 28
Pool3	480 x 28 x 28
Inception4a	480 x 14 x 14
Inception4b	512 x 14 x 14
Inception4c	512 x 14 x 14
Inception4d	512 x 14 x 14
Inception5e	512 x 14 x 14
Pool4	832 x 14 x 14
Inception5a	832 x 7 x 7
Inception5b	832 x 7 x 7
Pool5	1024 x 7 x 7
Fc	1024 x 1 x 1
prob	1000

Yanchao Liang and Jianhua Li from the East China University of Science and Technology in 2017 [4], created the ECUSTD database composed of 2987 images of 20 categories of food, also they introduced a CNN architecture for image classification and segmentation of food to estimate food calories. In their work results varied greatly, they predicted the weight of the food following a process of object recognition, image segmentation and image classification.

The best accuracies were for simple foods such as an orange with a mean error of 0.7%, plum with -2.1% and lemon with mean error of -2.8%. Other simple foods such

as apple, grape and banana had large mean errors: -18.7%, 33.5% and 28.8% respectively.

Simon Mezgec and Barbara Seljak from the Information and Communication Technologies department of the University of Slovenia developed in 2017 the NutriNet a CNN architecture that they claim to have 94 percent accuracy in some food datasets (classification only) [5].

Xioweng Wu et. al. from the Living Analytics Research Center of the University of Singapore, compiled a database (FoodSeg103) of 9490 images with 154 classes and segmentation masks [6].

Additionally, they proposed a multimodal pre-training approach called ReLeM which equips the model with semantic knowledge. They performed three segmentation experiments: Dilated Convolution based, Feature Pyramid based, and Vision Transformed based. The base and the models they proposed are intended to function as a point of reference for future work.

We can conclude from the works cited above and some others not mentioned that the problem of food recognition is challenging due to the nature of food items. Foods are deformable objects, which makes the process of defining their structure difficult. Furthermore, some foods types can have high intra-class and low inter-class variance. It seems that a deep learning approach can be very promising in the field of food image recognition.

None of the works reviewed so far have tried to estimate nutritional value on top of calorie estimation so it seems that there is a lot of room for improvement and the problem cannot be considered solved yet.

3 Methodology

The objective of this work is to develop a first approach for the complex problem of food composition estimation based on computer vision and deep learning. The problem is complex, and it requires two main treatments that need to be combined. The first one is pure image classification [7] (classifying foods based on the entire image they appear on), the second is object recognition and segmentation (segmenting different food items in the same image) [8].

In this work we focus on image classification only, we believe that the correct identification of food type is the first step for an accurate food composition estimation later.

Given the promising results of other works and the fact that convolutional neural networks are used to solve complex problems [9], we decided to test well known CNN's architectures in food data sets in order to get a grasp on food image classification.

We mainly used two datasets: The Food 101 dataset formed by 101,000 images of 101 food classes gather by Bossard et al. for their work in Mining Discriminative components with random forests. The dishes presented in this dataset can be as simple as an apple pie or complex as a Caesar salad.

Table 3. MobileNet Architecture [12].

Operator	Input
Conv / s2	3 x 224 x 224
Conv dw / s1	32 x 112 x 112
Conv / s1	32 x 112 x 112
Conv dw / s2	64 x 112 x 112
Conv / s1	64 x 56 x 56
Conv dw / s1	128 x 56 x 56
Conv / s1	128 x 56 x 56
Conv dw / s2	128 x 56 x 56
Conv / s1	128 x 28 x 28
Conv dw / s1	256 x 28 x 28
Conv / s1	256 x 28 x 28
Conv dw / s2	256 x 28 x 28
Conv / s1	256 x 14 x 14
5 x (Conv dw /s1, Conv / s1)	512 x 14 x 14
Conv dw / s2	512 x 14 x 14
Conv / s1	512 x 7 x 7
Conv dw / s2	1024 x 7 x 7
Conv / s1	1024 x 7 x 7
Avg Pool / s1	1024 x 7 x 7
FC / s1	1024 x 1 x 1
Softmax / s1	1000

The other dataset we used is the ECUSTD dataset gather by Yanchao Liang and Jianhua Li from the East China University of Science and Technology. This dataset is formed by 2987 images of 20 classes of food, the classes are simple as apple, banana, grapes, mango, doughnut, etc. There is only one class labeled as mix which include two foods from two different classes. Images in both datasets were treated to fit the size 224 x 224 pixels which is the size used in the most popular CNN architectures built for the ImageNet Challenge.

The treatment used to resize the images was a bi-cubic interpolation over a 4x4 pixel neighborhood. The language used was Python with cv2 library. We choose 3 architectures to test food classification: MobileNet, GoogLeNet and VGG16. The decision behind testing GoogLeNet seems obvious; the model has the lowest count of parameters of the modern architectures paired with great accuracy obtained in the ImageNet challenge. On the other hand, we choose the VGG16 model due to the simplicity and elegance of the architecture combined with its low error.

AlexNet and ZFNet were left aside due to the high number of parameters in the model together with a higher error-rate. ResNet model was left aside too, this time the decision was since the model has 152 layers, and the size of the images is difficult to fit to the model therefore we should have used the ResNet18 model instead which has lower accuracy than GoogLeNet and VGG16.

GoogLeNet and VGG16 were the 1st and 2nd places respectively in the ImageNet Large Scale Visual Recognition Challenge in 2014. GoogLeNet has 22 layers and VGG 16 has 19 layers. Both architectures are big and slow to train, but fast to use in

Table 4. Comparison of ImageNet champions and runner ups. Source: Compiled by the authors based on the results of the ImageNet Challenge.

Year	CNN	Developed by	Place	Top-5 Error Rate	No. of parameters
1998	LeNet	Yann Le Cunn et al			60 thousand
2012	AlexNet	Alex Krizhenvsky et al.	1st	15.3%	60 million
2013	ZFNet	Matthew Zeiler et al	1st	14.8%	60 million
2014	GoogLeNet	Google	1st	6.67%	4 million
2014	VGG16	Simonyan and Zisserman	2nd	7.3%	138 million
2015	ResNet	Kaiming He	1st	3.6%	60 million

predictions once the training is over. Table 4 shows a comparison of several champions of the ImageNet challenge and their characteristics.

MobileNet was also chosen to test the accuracy of the classification task. The MobileNet it's part of a family of MobileNets designed also by Google to effectively maximize accuracy while being mindful of the restricted resources for an on-device or embedded application.

MobileNets are small, low-latency, low-power models parameterized to meet the resource constraints of a variety of use cases. They can be built upon for classification, detection, embeddings, and segmentation.

MobileNet architecture uses depth-wise convolutions (Conv dw), where each filter channel is used only at one input channel.

We made some changes in the fully connected layers of all three architectures before training them with the food images datasets. For VGG16 we changed the number of neurons in the fully connected layers, originally the architecture has 3 fully connected layers of 4096 neurons each and a final output of 1000. We changed the number of neurons to 512 for the 3 final layers and the final output was changed to 20 for the ECUSTD training and to 101 for the Food 101 training.

For the GoogLeNet and MobileNet we changed the number of neurons in the final output layer in order to match the number of categories in both databases. The 3 architectures were trained on both datasets, for the ECUSTD dataset we used two techniques for the training: transfer learning and full learning.

Given that the Food 101 data set is considerably larger it was only tested with transfer learning. The transfer learning technique applied in this work was to freeze the convolutional layers of the network with the weights trained with the ImageNet dataset and only train the fully connected layers for 100 epochs. The full training technique trains all layers including the convolution layers with the selected dataset for 100 epochs.

Table 5. Results – ECUSTD (in Colab).

Model	Transfer Learning	Epochs	Optimizer	Validation	Acc.	Time
GoogLeNet	True	100	Adam	5 folds	99.56	1 h
GoogLeNet	True	100	SGD	5 folds	99.43	45 m
GoogLeNet	False	100	Adam	5 folds	99.6	1.7 h
VGG16	True	100	Adam	5 folds	10.4	1.25 h
VGG16	False	100	Adam	5 folds	5.01	2 h
MobileNet	True	100	Adam	5 folds	99.76	40 m
MobileNet	False	100	Adam	5 folds	99.63	1.5 h

Table 6. Results – Food 101 (in Buap Ser).

Model	Transf Learning	Epochs	Optimizer	Validation	Acc.	Time
GoogL Net	True	100	Adam	5 folds	63.08	12 h
VGG16	True	100	Adam	5 folds	18.14	24 h
Mobileet	True	100	Adam	5 folds	66.3	14 h

The Loss function used in all experiments was Cross Entropy Loss Function, the Adam Step optimizer was use in most of the experiments except for one where Stochastic Gradient Descent was used. The validation method used was a 5-fold cross validation.The experiments were carried on using python's library Pytorch and run in Google's colaboratory and BUAP's server.

4 Results

Tables 5 and 6 shows the results of the different experiments performed. From the results in tables 5 and 6 it can be seen that for the ECUSTD dataset a very good accuracy can be obtained with transfer learning in two of the three architectures with relative low training times. VGG16 had the lowest accuracy with both transfer learning and without it, it's important to point out that VGG16 is the model with more parameters to optimize and that's why the training times are the largest in general.

MobileNet and GoogLeNet didn't improved the accuracy a lot by making a full training without transfer learning. For the Food 101 data set the highest accuracy achieved was obtained with GoogLeNet, second and very close was the MobileNet. VGG16 as with the other dataset failed to do a decent classification.

We can conclude that food classification using transfer learning and well-known CNN architectures can be satisfactory achieved in cases where the foods classes are simple and isolated in an image, for example, images of fruits.

In the other hand, when complex dishes are taken into consideration the recognition drops dramatically, however, a 60% accuracy for complex dishes without any kind of prior segmentation or object detection reflects a positive result, because it means that with proper segmentation and detection algorithms, accuracy can improve a lot.

5 Conclusions

Further experiments need to be carried out with the VGG16 architecture, because the default experiments achieved a very low accuracy, changes in the fully connected

layers, epochs, optimizer, and kernel sizes can be tested in order to see if the accuracy can increase. The results of the experiments confirm our assumption that this approach can be a first step for a full food composition process.

Next steps could include object detection and segmentation algorithms paired with the image classification methods mentioned above. Furthermore, even though a lot of improvements have been done in Deep Learning with CNN algorithms, for food composition estimation exists a real challenge to construct a model that can classify food correctly, one of the major setbacks is that there are not many datasets on the subject, many of them are small or they represent small classes of food.

Also, many food dishes have irregular shapes and wide color ranges, and that represents a very complex problem to solve even today with the most robust of architectures. Future work can focus too on obtaining datasets of regional foods, as we believe also that regional foods images can increase the accuracy in any model trying to classify dishes of a given country diet.

References

1. Varady, K. A.: Intermittent versus daily calorie restriction: which diet regimen is more effective for weight loss? Obesity reviews, vol. 12, no. 7, pp. 593–601 (2011)
2. Sauberlich, H. E.: Bioavailability of vitamins. Progress in food and nutrition science, vol. 9, no. 1–2, pp.1–33 (1985)
3. Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Murphy, K. P.: Im2calories: Towards an automated mobile vision food diary. In: Proceedings of the IEEE International Conference on Computer Vision pp. 1233–1241 (2015)
4. Liang, Y., Li, J.: Computer vision-based food calorie estimation: Dataset, method, and experiment (2017) arXiv preprint arXiv:1705.07632
5. Mezgec, S., Koroušić Seljak, B.: NutriNet: A deep learning food and drink image recognition system for dietary assessment. Nutrients, vol. 9, no. 7, pp.657 (2017)
6. Wu, X., Fu, X., Liu, Y., Lim, E. P., Hoi, S. C., Sun, Q.: A large-scale benchmark for food image segmentation. pp. 506–515 (2021) doi: 10.1145/3474085.3 475201
7. Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25 (NIPS 2012)
8. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs (2014) doi: 10.48550/arXiv.1412.7062
9. Patel, S.: A Comprehensive analysis of convolutional neural network models. In: Proceedings of International Journal of Advanced Science and Technology, vol. 29, no. 4. pp. 771–777 (2020)
10. Kölsch, A., Afzal, M. Z., Liwicki, M.: Multilevel context representation for improving object recognition. In: Proceedings of 14th IAPR International Conference on Document Analysis and Recognition, pp. 10–15 (2017) doi: 10.1 109/ICDAR.2017.322
11. Qassim, H., Verma, A., Feinzimer, D.: Compressed residual-VGG16 CNN model for big data places image recognition. In: IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), pp. 169–175 (2018)
12. Qiya, N., Yunlai, T., Lin, C.: Design of gesture recognition system based on deep learning. In: Proceeding of Journal of Physics: Conference Series, vol. 1168, no. 3 (2019) doi: 10.1088/1742-6596/1168/3/032082