Computers in Biology and Medicine 95 (2018) 217–233

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Combining deep residual neural network features with supervised machine learning algorithms to classify diverse food image datasets

Patrick McAllister[a], Huiru Zheng[a,*], Raymond Bond[a], Anne Moorhead[b]

[a] Ulster University, Jordanstown Campus, School of Computing, Northern Ireland, United Kingdom
[b] Ulster University, Jordanstown Campus, School of Communication and Media, Northern Ireland, United Kingdom

ARTICLE INFO

Keywords:
Obesity
Food logging
Deep learning
Convolutional neural networks
Feature extraction

ABSTRACT

Obesity is increasing worldwide and can cause many chronic conditions such as type-2 diabetes, heart disease, sleep apnea, and some cancers. Monitoring dietary intake through food logging is a key method to maintain a healthy lifestyle to prevent and manage obesity. Computer vision methods have been applied to food logging to automate image classification for monitoring dietary intake. In this work we applied pretrained ResNet-152 and GoogleNet convolutional neural networks (CNNs), initially trained using ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset with MatConvNet package, to extract features from food image datasets; Food 5K, Food-11, RawFooT-DB, and Food-101. Deep features were extracted from CNNs and used to train machine learning classifiers including artificial neural network (ANN), support vector machine (SVM), Random Forest, and Naive Bayes. Results show that using ResNet-152 deep features with SVM with RBF kernel can accurately detect food items with 99.4% accuracy using Food-5K validation food image dataset and 98.8% with Food-5K evaluation dataset using ANN, SVM-RBF, and Random Forest classifiers. Trained with ResNet-152 features, ANN can achieve 91.34%, 99.28% when applied to Food-11 and RawFooT-DB food image datasets respectively and SVM with RBF kernel can achieve 64.98% with Food-101 image dataset. From this research it is clear that using deep CNN features can be used efficiently for diverse food item image classification. The work presented in this research shows that pretrained ResNet-152 features provide sufficient generalisation power when applied to a range of food image classification tasks.

1. Introduction

Obesity is a global concern and is a serious health condition that can cause diseases such as heart disease, type-2 diabetes, and some cancers [1,60]. The increase of obesity has also been reported as a major burden on health care institutions through direct and indirect costs [56]. One of the major ways that obesity can be managed is through dietary management methods such as food logging and other methods [3]. Food logging is an activity in which the user document their energy intake to monitor their diet. Other methods may include the use of an exercise log book to document physical activities and the duration. Previously, users documented their intake using a food diary however many users now use smartphone applications to document their energy intake. The increase in smartphone usage has also led to the increase of well-being applications that are able to facilitate food logging. Many of these applications incorporate a simple diary entry, and/or connect to an online database/API to search for nutritional content for each of the users entries. Other novel methods include allowing the user to photograph the food items to determine calorie values. Using images has the potential to remove much of the complexity from traditional food logging to make it convenient for the user to document food intake to promote dietary management. Many studies have been completed in researching the use of computer vision methods to classify photographs of food to promote food logging [4–6]. This interactive approach to food logging through using a camera within a smart-device may promote the use of food logging which is an important method to maintain weight loss. The remainder of this paper is structured as follows: Section 2 presents related work in how this problem has been tackled in previous research. Section 3 discusses the aim, objectives, and contributions of this work. Section 4 describes the methods used in this work and the use of Convolutional Neural Networks (CNNs) for feature extraction. Experiment results are presented in Section 5 followed by a discussion in Section 6.

* Corresponding author.
E-mail addresses: mcallister-p2@ulster.ac.uk (P. McAllister), h.zheng@ulster.ac.uk (H. Zheng), r.bond@ulster.ac.uk (R. Bond), a.moorhead@ulster.ac.uk (A. Moorhead).

https://doi.org/10.1016/j.compbiomed.2018.02.008
Received 25 May 2017; Received in revised form 11 February 2018; Accepted 12 February 2018

Section 7 highlights study limitations and areas for future work.

2. Related work

Food logging is a beneficial method to aid dietary management and recent novel methods have utilised meal photographs for food logging. A review [41] was completed to highlight a variety of computer vision methods that have been applied in food image recognition to promote dietary management. Key messages from this review are that there is a need for real food intake monitoring and one of the main challenges for diet monitoring using wearable sensors is practicability when used in a different environments and how automatic dietary monitoring is important to document nutritional intake habits to prevent chronic conditions.

Food image recognition is a difficult task due to the amount of variation within food types. Food items in images are usually accompanied with other food items as well as other unrelated non-food items. The high variation of colour, shape, size, and texture in food items means that one method of image feature extraction and classification may not adapt to other foods and therefore a feature combination approach may be needed. Conventional ways to classify images utilise the use of hand-crafted feature extraction, e.g. global or local feature extraction using Speed-Up-Robust Features (SURF) [38] or texture features such as local binary patterns (LBP) [39] and texture filter banks (e.g. STF) [7]. Feature engineering is used to determine what type of features and parameters are best used to successfully classify certain food types and food categories and much work has been completed in this area. In Ref. [5] a bag-of-features model was proposed that used a combination of scale invariant feature transform (SIFT) features along with hue-saturation-value (HSV) colour features and a linear SVM to classify images into 11 categories with 78% accuracy. Other works also utilise a combination approach using SIFT and SPIN features and achieve high accuracy in classifying high level food groups (89% accuracy in classifying sandwiches and 91.7% in classifying chicken) using Pittsburgh Fast-Food Image Dataset (PFID). However, PFID dataset is an image dataset that was developed in a controlled laboratory environment, further works could be completed in applying this feature combination approach to similar image categories photographed in real-world environments [6]. Other works use feature selection methods to determine optimal features [8] for food image classification. As well as using traditional feature extraction methods, CNN methods have become increasingly popular for image classification and this can be attributed to ImageNet Image Large-Scale Visual Recognition Challenge (ImageNet ILSVRC) as it allows users to compete against each other in achieving a high classification accuracy and the winners in recent years have used convolutional neural networks (CNNs). Great emphasis has been placed on using CNNs for image classification and this is evident in a surge of recent research in this area relating to the fine-tuning CNN [11], deep feature extraction [12], and also training CNNs from scratch [11].

2.1. Detecting food in images using CNN

CNN has been utilised for food image detection. This problem can be condensed down to a simple binary classification problem (food/non-food). The purpose of food image detection process is to first determine if food is present within an image or video. In regards to a food image recognition pipeline, this would be the first stage in food image recognition framework i.e. determining if the image contains food or not. In Ref. [13] GoogleNet pretrained model was fine-tuned using Food-5K dataset. The training process in Ref. [13] utilised a subset of Food-5K data using 1000 iterations. The learning rate was changed to 0.01 and the learning rate policy was polynomial. Results from Ref. [13] achieved 99.2% accuracy in determining food/non-food classes. Other research also utilised CNNs for food detection [14] and used 6-fold cross validation with different hyper-parameters to determine optimal settings and experiments achieved 93.8% in food/non-food detection. It is clear from research that CNNs can be used effectively for food detection in images.

2.2. Predicting food type in images using CNN

Extensive research has been carried out in utilising CNN for food item recognition. The food item recognition process would take place after the food detection phase in which the actual food item is then predicted within food image. In Ref. [15] CNNs were utilised to extract features from convolutional layers in order to classify food items and food groups, experiments achieved 70.13% for 61 class dataset and 94.01% for 7 class datasets.These experiments used AlexNet deep features with a SVM classifier applied to PFID dataset [15]. In Ref. [16] the aim of the work was to compare conventional feature extraction methods with CNN extraction methods utilising UECFood-100 dataset. Results from Ref. [16] achieved 72.6% accuracy for top-1 accuracy and 92% for top-5 accuracy. Also in Ref. [14], as well as performing food/non-food experiments, food group classification was performed. A CNN was developed and was trained using extracted segmented patches of food items [14]. The food items used in this work were based around 7 food major types. The patches were then fed into a CNN using 4 convolutional layers with different variations of filter sizes and using 5 × 5 kernels to process the patches. Results in Ref. [14] achieved 73.70% accuracy using 6-fold cross validation. These studies confirm that CNNs provide an efficient method for food image recognition to provide for accurate food logging to promote dietary management.

2.3. CNN deep feature extraction methods for food detection/food item classification

Recent research has focused have used deep features extracted from pretrained CNN architectures to train machine learning classifiers for food image classification. Some research have opted for deep feature extraction opposing to fine-tuning pretrained CNN or training from scratch because less computational power and time is needed or datasets that are used are small. Well-known CNN architectures (e.g. AlexNet, VGG-16, GoogleNet) have been used for deep feature extraction in classifying food images to automate food logging. This section discusses research that use deep feature extraction to detect food in images and classify food items in images for automated food logging. A comparative review was carried out on analysing the performance of a number of pretrained CNN architectures [43]. This review used VGG-S, Network in Network (NIN), and AlexNet for deep feature extraction to train food detection models. A food/non-food image dataset was collated and deep features were extracted from the models to train machine learning classifiers (one-class SVM classifier and binary classifier). Results showed that binary SVM classifiers trained with deep features achieved 84.95% for AlexNet, 92.47% for VGG-S, and Network In Network model achieving 90.82%. It is worth noting that UNICT-FD889 dataset used for deep feature extraction in Ref. [43] contains minimal noise as the images are focused on the food item, therefore this may contribute to high accuracy results. Further work could be completed in utilising a larger food image dataset consisting of images from different environments and also using different machine learning classifiers for further comparison.

Other research explored the effect of training machine learning classifiers using deep features extracted from different layers using a pretrained AlexNet architecture [10,15]. Authors used AlexNet model to extract deep features from various layers deep in the architecture (FC6, FC7, and FC8 layers). The food image dataset used in Ref. [15] was PFID. Two experiments were presented in Ref. [15]; classifying high-level food categories by organising PFID dataset into 7 category dataset and also classifying individual categories in PFID (61 classes). Results showed that the highest accuracy for the 61 class dataset was 70.13% using deep features extracted from layer FC6 in AlexNet. For the 7 class dataset, the highest accuracy achieved for deep features was 94.01% using layer from FC6. The contribution in Ref. [15] supports the same findings in Ref. [43] suggesting that deep feature extraction provides high accuracies in classifying small grouped food image datasets (related food items) as well as datasets with specific different food types. Results also suggest that