University of Nottingham
UK | CHINA | MALAYSIA

Food Recipe Generator: an APP with Recipe1M+ to recognise dishes recipes from images

Interim Report

Yichen Lu

Supervised by Andrew French

School of Computer Science
University of Nottingham

December 2022

Abstract

According to Facebook, the number of food pictures on Instagram has exceeded 300 million. However, access to how to cook food is still limited through cooking websites or related tutorials. This has led to the creation of several searching technics to enrich the cooking method behind the food pictures. The main purpose of the project is the implementation of an optimised food recognition method, which is divided into two processes: model analysis and practical application. The interim report will aim to explain the food image recognition models, comparing them and talk about the future plan. In the report, I will discuss four models including the traditional indexing methods and their optimisation ideas. The comparation shows the evolution of the essential logic chains and applicability of the newly proposed ideas. It paves the way for the further application development.

Contents

1. Introduction

1.1 Background and Motivation

Food, specifically speaking, ingredients not only will have a profound impact on human's life including in-take nutrition and physical health, but also play a significant role in defining people's identity, social status, and culture well-beings. [1] As the famous French gourmet Brillat-Savarin said, "Tell me what you eat, and I will tell you who you are." As the steady development of living condition, people are looking for more healthy nutrition, which is why human have to acquire a deeper insight into the ingredient. Thus, there is always a research focus on diet field, especially the problem of food choice.

There have been already substantial researches in a large scale related to food, including diet preference [2], nutrition detective [3], eating assumption [4], production safety [5] as well as food culture [6]. In result of food-related research covers several scientific areas, the whole investigation is like fragments and lack of a systematic set. Therefore, in 2019, Min with other researchers [7] proposed an advanced framework of calculating food. They pointed out that the food calculation is not limited to goals of sensation, recognition, detection and recommendation, but also serves to medicine, biology, agriculture, food industry, nutritional health and other fields. In which, one of the basic and essential tasks is food image recognition.

Considered as a field in computer version, food image recognition is a significant branch of the Fine-Grained Image Classification (FGIC) [8-11], which contains important research value. Recently, range of mobile devices such as phones and cameras, and the wearables like DJI Action 2 Power Combo [12] have been springing up all across the daily life. Moreover, rapid development of artificial intelligence technology empowered the food recognition as a promising implementation. Taking the Amazon Fresh Go Store [13] as an example, through recognizing the food categories, ingredients and relative attributes, the store could realize the automatic payment and settlement. Besides, the system inside also possibly utilizes such features to analyse food nutrients and assess user eating habits to ensure users daily healthy in-take. For instance, FoodAI [14], as a new food image recognition via deep learning for smart food logging, has been deployed as an API service and is one of the components powering Healthy 365, a mobile app developed by Singapore's Heath Promotion Board. Even further, food image recognition will achieve food recommendations and food indexing in social networks. In terms of design, food image recognition gradually becomes the hot spot in a large scale.

Food image recognition belongs to the Fine-Grained Image Classification, which means to identify different subclasses in a group of same class objects, such as different kinds of birds or vehicles. Its main task is to categorize the food in the given image, or to index different ingredients and cuisines. The exploration of its technology can be traced to 1977, when Parrish et al. [15] first conducted visual-based fruit and vegetable identification study. This

was followed by food logging system developed by Kitamura's team [16] to proffer diet suggestion as analysing the ingredient and calorie in food. In 2014, Bossard et al. [17] released the first large-scale Western cuisine image dataset "Food-101" and earlier used deep learning for food image recognition as a pioneer. With the rapid development of deep learning technologies and steady increase of extensive food image datasets, the related research sprang up. Throughout the history of food image processing, it inspires an idea which tends to combine functions like food categories recognition and nutrition analysis, and then implement them in a practical application.

Thanks to the development of social media, sharing food is becoming very popular. For many real-world applications, people are eager to learn about the underlying recipes for a food product. Relatively in daily life, people usually face the challenge that how to cook different kinds of dishes every day by using staple provided ingredients. It will intrigue people that if there is an application which is able to generate a set of recipes when users uploaded the pictures of ingredients or ready-made food, whose function can satisfy them both requirements at the same time.

1.2 Aim and Objectives

It will obviously solve the problem that people usually do not know how to cook even if they have the ingredients. According to the design, the program will eventually generate appreciated recipe based on what food in image. The main aim of the project is to build a system capable of accurately recognising ingredient at uploaded images. The output of which can in turn be used to create both proper diet recipes and total nutrition intake calculated with its menu.

The aim of the first half of this project is to compare and improve selection of different food recipes recognition models based on analysis of their output precision and time consumed. Next, a website will be designed as an implementation of the model. Considered with the breadth of recipes and the uncertainty of ingredients, the project is proposed to focus on specific fields of recipes and cuisines which can avoid the unnecessary web crawling technology and be measured against existing industry approaches to obtain an optimised performance. Up to now, during the progress, the main form of such implementation gradually has evolved into three steps: training, analysing and finally application.
In the first semester, I have successfully analysed the differences among four main recognition models which will explored in the following report (Section 2: Methodology) and choose one of them as an optimised model for further implementation.

1. Investigate more approaches for food image recognition models, enriching both input and output not limited to food pictures and recipes.
2. Design and implement a website with both front and back ends. Vue will be utilized to develop the interface and the system in back-end will hold the main framework of Springboot and Mybatis to get and post data with MySQL. If there is time, the application written by Vue could even expands into Android platform.
3. Implement the main functions of food recognition, recipe generation, and food nutrition

calculation in website. For further expansion, it could separate the features into two directions including generating recipes directly based on dishes pictures and food raw materials.

1.3 Related work

The original project was initially undertaken by a Research group at the Universitat Politecnica de Catalunya, Massachusetts Institute of Technology and Qatar Computing Research Institute. It provides the largest dataset named Recipe1M+ for the traditional indexing method. The dataset is a new large-scale, structured corpus of over one million cooking recipes and 13 million food images. As the largest publicly available collection of recipe data, Recipe1M+ affords the ability to train high-capacity models on aligned, multimodal data.

Generating recipes from images requires the understanding of both the constituent ingredients and the process of preparation (e.g., slicing, mixing with other ingredients, etc.). Traditional methods treat this problem as a retrieval task, retrieving recipes from a fixed dataset based on a similarity calculation between the input images and the dataset images [18]. Obviously, the traditional approach fails in cases where the dataset is missing a certain food preparation method.

In the report, I provide an advanced method to overcome this data limitation. It is to treat the picture-to-recipe problem as a conditional generation task. Rather than getting the recipe directly from the picture, it would be better to first predict the ingredients of the food and then generate a food preparation method based on the image and the ingredients. This would allow for some additional information to be obtained using the intermediate process of images and ingredients, which makes the prediction explainable and accurate.

2. Methodology

As mentioned in the related work part, the food image recognition model can be divided into two main Levels, the simple one is that the image generates recipes and another one is finding the ingredients from features of the food images and then generate the recipe based on them. The basic advantage of the former one could be high efficiency because it can directly link the food to the recipe which has the highly similar food image. However, to compare the food image needs tons of image dataset and tons of time to train the model. Thus, here comes the latter one. The latter one, the reversed cooking, presents its advantages that only needs to classify whether the ingredients exist, which not only reduces required database capacity, but also raise the accuracy of the organization of recipes. The results on user experiments show that cooking food in this way has a higher success rate than traditional search methods.

2.1 Image to Recipe

2.1.1 Similar Image Search

The original project was initially undertaken by a Research group at the Universitat Politecnica de Catalunya, Massachusetts Institute of Technology and Qatar Computing Research Institute. The ordinary way is to calculate the similarity between the image uploaded by the user and images pre-stored in the database. It divides the recipe into two major components: its ingredients and cooking instructions. The pre-trained embedding vector obtained by the word2vec algorithm was used with a bi-directional LSTM (since the ingredient list is an unordered set, a bi-directional LSTM model was chosen, which considers both forward and inverse ordering) [21], where the LSTM performs logistic regression on each word in the ingredient text. At the same time, a two-stage LSTM model is designed to encode the sequence of the sequences since the simple model can not contain a too long sequence of instructions as a whole. Each cooking step is first represented as a vector, and then an LSTM is trained with a sequence of these vectors to obtain a vector characterising all the steps. For the food image representation, the im2recipe model uses two deep convolutional networks, VGG-16 and Resnet-50, removes the last 'softmax' classification layer and connects the rest to a joint embedding model.

As shown in the diagram below (Figure 1), the recipe model consists of two encoders: one for the ingredients and the other for the cooking instructions. The outputs connecting the two encoders are embedded in a shared space of recipes-images. The image representations are also mapped into the same space by a simple linear transformation.

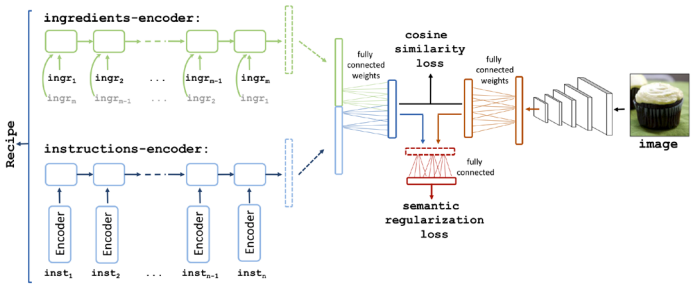


Figure 1: im2recipe model

The effectiveness of this model depends a lot on the storage capacity of the database and the computing power of the computer. Thus, in this work, we can address data limitations by introducing the large-scale Recipe1M dataset which contains one million structured cooking recipes and their images. Using these data, we can train a neural network to learn a joint embedding of recipes and images that can be used as a classification model.

2.1.2 Improved Similar Image Search

The project is divided into two main processes: data import and recipe retrieval. When importing data, the recipes are converted into a vector using the model im2recipe, and then stores the vector and the corresponding recipe information in MySQL, as shown below. When implementing recipe retrieval, firstly, the model im2recipe is used to convert the physical image into a vector, then the image vector is used to retrieve a vector of similar recipes and finally the recipe information corresponding to the recipe vector is found in MySQL.

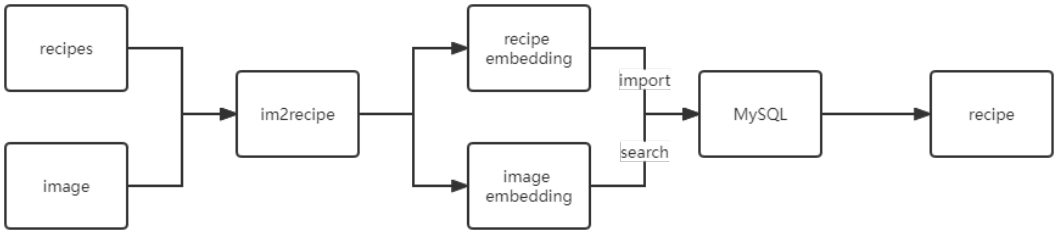


Figure 2: Improved im2recipe Model

2.2 Image To Ingredient To Recipe

2.2.1 Individual Attempts:

One way to overcome this data limitation is to treat the picture-to-recipe problem as a conditional generation task. Rather than getting the recipe directly from the picture, it is believed to be better to first predict the ingredients of the food and then generate a food preparation method based on the image and the ingredients. This would allow some additional information to be obtained using the intermediate process of images and ingredients [20]

This model was designed by me, which was based on the idea of recognising the ingredients and then generate a recipe. In the plan, I firstly would use the popular model named 'hugging face' as a classifier and use that to locate each food in a plate. The next step is using the NLTK module as a natural language processor to associate each ingredient together as a completed recipe. In fact, the classification model works not well because there are quite large scales of recipes and cuisines in the collected data and most of ingredients have less difference between each other than an apple and an egg. Therefore, the plan was abandoned.

2.2.2 Reversed Cooking

The model consists of two main components. Firstly, the researchers pre-train an image encoder and an ingredients decoder to extract the visual features of the input image to predict the ingredients. An ingredient encoder and an instruction decoder are then trained to generate food names and cooking procedures based on the image features of the input image and the predicted ingredients. As the structure of the model shown below, the input to the model is a picture of the food, the output is a sequence of cooking methods, and the intermediate step is to generate a list of ingredients based on the image.

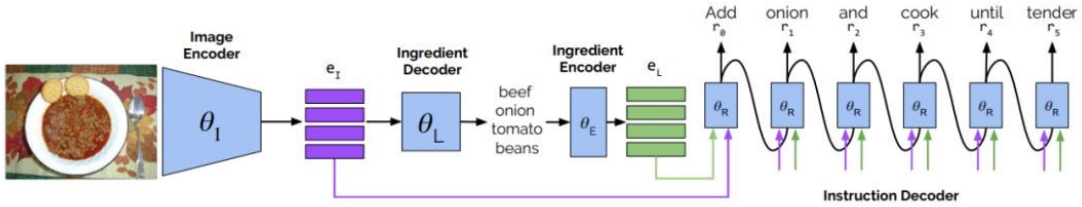
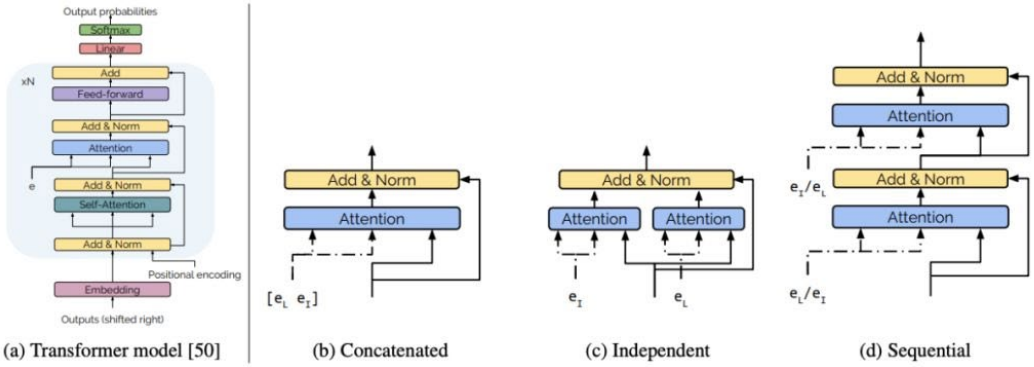


Figure 2: **Recipe generation model.** We extract image features $\mathbf{e}_I$ with the image encoder, parametrized by θ_I. Ingredients are predicted by θ_L, and encoded into ingredient embeddings $\mathbf{e}_L$ with θ_e. The cooking instruction decoder, parametrized by θ_R generates a recipe title and a sequence of cooking steps by attending to image embeddings $\mathbf{e}_I$, ingredient embeddings $\mathbf{e}_L$, and previously predicted words $(r_0, ..., r_{t-1})$.

During the text generating, the RNN is completely abandoned but to use the powerful Transformer instead. Each Transformer block consists of two attention layers and a linear layer. In order to combine image and ingredient information for cooking process generation and to achieve multimodal attention, the attention layer has been changed: the first attention layer performs self-attention on the previous output, which is consistent with the original Transformer. The second attention layer is changed to conditional attention in order to fine-grained extraction of the first layer of self-attention. This level of attention is subject to two constraints: image features and ingredients embeddings. In order to perform multimodal fusion simultaneously, three fusion structures were tried in the paper [20], which is shown in the diagram below as (b), (c), (d). The final experiment found that the Concatenated approach worked best.



(a) Transformer model [50] (b) Concatenated (c) Independent (d) Sequential

Figure 3: **Attention strategies for the instruction decoder.** In our experiments, we replace the attention module in the transformer (a), with three different attention modules (b-d) for cooking instruction generation using multiple conditions.

3. Design

3.2 Image Processing System

Since we are not able to ensure the quality of the image user uploaded, it needs to be measured with a threshold. However, we won't have a non-distorted image to calculate the quality of the distorted image. The only input obtained is the image whose quality is to be measured, there is no image at all that can be used as a reference. In this case, we need a reference-free IQA metric called the Reference-free Image Spatial Quality Assessor (BRISQUE). BRISQUE will assign the image a mean quality score [22] between 0 (best) and 100 (worst). It stands for blind or no reference image spatial quality evaluator, a reference-free spatial domain image quality evaluation algorithm. The overall principle of the algorithm is to extract mean subtracted contrast normalized (MSCN) coefficients from an image, fit the MSCN coefficients to an asymmetric generalized Gaussian distribution (AGGD), extract the features of the fitted Gaussian distribution and input them to a support vector machine SVM for regression to obtain the image quality evaluation results.

```
>>> import imquality.brisque as brisque
>>> import PIL.Image

>>> path = 'path/to/image'
>>> img = PIL.Image.open(path)
>>> brisque.score(img)
4.9541572815704455
```

Figure 3: BRISQUE function

3.2 Food Recipe Recognition System

The optimised algorithm should be the last one, reversed cooking. However, the main program language in the back end will be set to Java, but the reversed cooking algorithm is written by python language. To call the algorithm to compile in a java environment, it can be utilized with Runtime.getRuntime() function provided by Java, which will achieved in the next level.

3.3 Web Application

Vue will be utilized to develop the interface and the system in back-end will hold the main framework of Springboot and Mybatis to get and post data with MySQL. The main language is Java, and it can be implemented with the model easily. According to the plan, users are able to upload a clear image of food and the system will measure the image and recognise every ingredient inside, providing the related recipes about how to cook this dishes, listing the nutrition of the food containing.

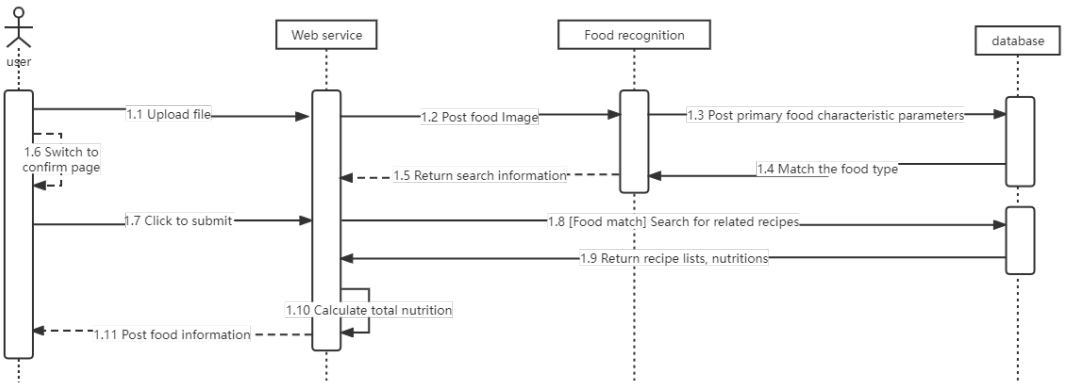


Figure 4: Operation sequence 2

Progress

4.1 Project Management

The waterfall methodology has been used for the management of the project, with a Gantt chart being utilized to decompose the project into a sequence of ordered tasks. The project is currently on track, with all the tasks required up until this point in time completed. Considering a full schedule with the heavy academic pressure in the first semester, such as quite a lot of language tests (TOEFL, GRE), applications for graduate study, and other classes in Year Four University, I should have more available time in the next semester.

Therefore, as more time has been spent on the project, an improved vision of the future tasks has been gained. This has enabled existing tasks to be decomposed further into more descriptive components, which better describe the breakdown of the project and enable a more accurate time frame to be set for each of them.

1. Throughout the next term, it is planed to search for more available models to acquire highly accurate recipes. Until getting started to implement the optimised model into the website, the model of reversed cooking could be continuingly improved.
2. I could learn further about the image pre-processing and figure out how it works with the BRISQUE evaluator. While implementing the BRISQUE, I will think of more approaches to extend the interface of the web application which allows it to receive a wider range of photos, such as a plate of food or just a few ingredients.
3. More functions could be implemented on the website. such as calculating the nutrition from the food recipes and planning the daily intake. Besides, I could add a recipe logging system to keep recording the daily intake, and send warning to the user who exceeds the standard amount of nutrients.

A full list of changes is reflected in the newly updated chart for the project, as shown in Figure 5.

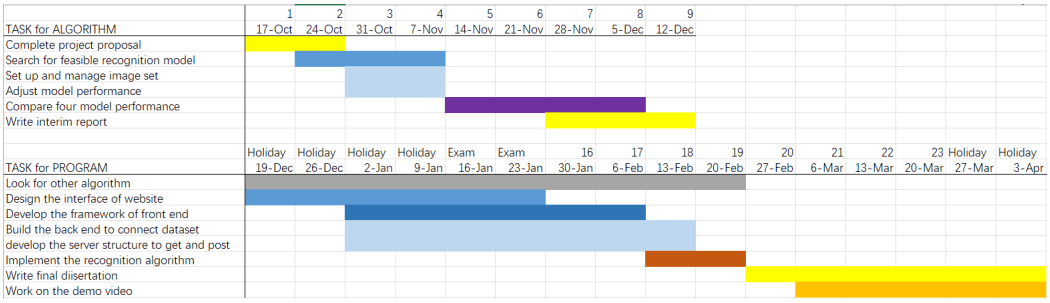


Figure 5: Gantt chart

4.2 Contributions and reflections

During the project, I have searched quite a lot of useful papers that strongly support me to write the report. Since find a better way to generate the food recipe which is firstly recognising

the ingredients and generating the recipe based on them, I never hesitated to try implementing the model to my local computer and manage the program to run properly. Specifically, in the project I avoided all possible ways to collect user's data and made full use of the public resources online. I ensure the original project of Recipe1M+ that code, data and models are publicly available, and adhere to the computer laws, social, ethical and professional issues (LSEPI).

Moreover, the website will be designed to read aloud the identification content for the disables who can not see the screen, while extending some classes for a variety of people from different cultural backgrounds.

Considering about my project, it showcases a sound understanding of data protection law and the appropriate use of new technologies, which can be seen in the data collected for the intended use or analysis, because the individual or group of individuals has consented.

Bibliography

[1] Khanna S K. Food and culture: A reader, by Carole Counihanand Penny Van Esterik. Ecology of Food and Nutrition, 2009, 48(2): 157-159

[2] Nestle M, Wing R, Birch L, et al. Behavioral and social influences on food choice. Nutrition Reviews, 2009, 56(5): 50-64

[3] Sørensen L B, Møller P, Flint A, et al. Effect of sensory perception of foods on appetite and food intake: a review of studies on humans. International Journal of Obesity, 2003, 27(10): 1152-1166

[4] Pauly D. A simple method for estimating the food consumption of fish populations from growth data and food conversion experiments. National Oceanic and Atmospheric Administration, 1986, 84(4): 827-840

[5] Chen Z, Tao Y. Food safety inspection using "from presence to classification" object-detection model. Pattern Recognition, 2001, 34(12): 2331-2338

[6] Harris, Marvin. Comment on Vayda's review of good to eat: Riddles of food and culture. Human Ecology, 1987, 15(4): 511-517

[7] Min Weiqing, Jiang Shuqiang, Liu Linhu, Rui Yong, Jain Ramesh C Jain. A survey on food computing. ACM Computing Surveys, 2019, 52(5): 1-36

[8] Qiu Jianing, Frank Po Wen Lo, Yingnan Sun, Siyao Wang, Benny Lo. Mining discriminative food regions for accurate food recognition//Proceedings of the British Machine Vision Conference. Cardiff, UK, 2019: 158

[9] Zhang X, Zhou F, Lin Y, et al. Embedding label structures for fine-grained feature representation//Proceedings of the Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1114-1123

[10] Zhou F, Lin Y. Fine-grained image classification by exploring bipartite-graph labels//Proceedings of the Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1124-1133

[11] Hou Saihui, Feng Yushan, Wang Zilei. VegFru: A domain-specific dataset for fine-grained visual categorization//Proceedings of the International Conference on Computer Vision. Venice, Italy, 2017: 541-549

[12] DJI Action 2 Dual-Screen Combo - DJI Mobile Online Store (Netherlands). (n.d.). DJI Store. Retrieved October 24, 2022, from https://m.dji.com/nl/product/dji-action-2?vid=107631
[13] Wankhede K, Wukkadada B, Nadar V. Just walk-out technology and its challenges: A case of Amazon Go[C]//2018 International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2018: 254-257.
[14] Sahoo D, Hao W, Ke S, et al. FoodAI: Food image recognition via deep learning for smart food logging[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 2260-2268.
[15] E Parrish, A K Goksel. Pictorial pattern recognition applied to fruit harvesting. Transactions of the ASAE, 1977, 20(5): 822-827
[16] Keigo Kitamura, Toshihiko Yamasaki, Kiyoharu Aizawa. Food log by analyzing food images//Proceedings of the 16th International Conference on Multimedia. Vancouver, Canada, 2008: 999-1000
[17] Lukas Bossard, Matthieu Guillaumin, Luc Van Gool. Food-101 - mining discriminative components with random forests//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 446-461
[18] Salvador A, Hynes N, Aytar Y, et al. Learning cross-modal embeddings for cooking recipes and food images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3020-3028.
[19] Wang H, Sahoo D, Liu C, et al. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11572-11581.
[20] Salvador A, Drozdzal M, Giró-i-Nieto X, et al. Inverse cooking: Recipe generation from food images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10453-10462.
[21] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.