11-2018

Cross-modal recipe retrieval with stacked attention model

Jing-Jing CHEN

Lei PANG

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Citation

1

Multimed Tools Appl (2018) 77:29457–29473
https://doi.org/10.1007/s11042-018-5964-y

Cross-modal recipe retrieval with stacked attention model

Jing-Jing Chen[1] · Lei Pang[1] · Chong-Wah Ngo[1]

Received: 9 May 2017 / Revised: 2 March 2018 / Accepted: 3 April 2018 /
Published online: 17 April 2018

Abstract Taking a picture of delicious food and sharing it in social media has been a popular trend. The ability to recommend recipes along will benefit users who want to cook a particular dish, and the feature is yet to be available. The challenge of recipe retrieval, nevertheless, comes from two aspects. First, the current technology in food recognition can only scale up to few hundreds of categories, which are yet to be practical for recognizing tens of thousands of food categories. Second, even one food category can have variants of recipes that differ in ingredient composition. Finding the best-match recipe requires knowledge of ingredients, which is a fine-grained recognition problem. In this paper, we consider the problem from the viewpoint of cross-modality analysis. Given a large number of image and recipe pairs acquired from the Internet, a joint space is learnt to locally capture the ingredient correspondence between images and recipes. As learning happens at the regional level for image and ingredient level for recipe, the model has the ability to generalize recognition to unseen food categories. Furthermore, the embedded multi-modal ingredient feature sheds light on the retrieval of best-match recipes. On an in-house dataset, our model can double the retrieval performance of DeViSE, a popular cross-modality model but not considering region information during learning.

Keywords Recipe retrieval · Cross-modal retrieval · Multi-modality embedding

✉ Jing-Jing Chen
jingjchen9-c@my.cityu.edu.hk

Lei Pang
leipang3-c@my.cityu.edu.hk

Chong-Wah NGO
cscwngo@cityu.edu.hk

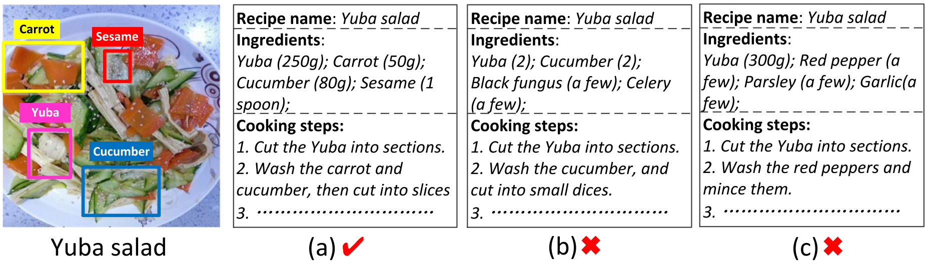
[1] City University of Hong Kong, Kowloon Tong, Hong Kong

1 Introduction

Food recognition is an important research topic since it serves as the key technology for automatic dietary assessment services. Generally, this task is considered as a challenging problem due to diverse appearances of food as a result of non-rigid deformation and composition of ingredients. In recent years, food recognition has started to capture more attention [3, 4, 16, 18] partly due to the success of deep learning technologies. With deep learning technologies, the accuracy of food recognition can be as high as 80% on the benchmark datasets such as Food101 [4], FoodCam-256 [13] and VIREO Food-172 [5]. The success gives light to the development of techniques for automatic dietary food tracking [1, 14, 18] and nutrition estimation [32], which has long been recognized as a challenge not only in multimedia [1, 28] but also in health and nutritional science [20].

Nevertheless, the existing efforts are mostly devoted to recognizing a pre-defined set of food categories, ranging from 100 to 256 categories [4, 5, 13, 16]. Extending to large-scale recognition, for example tens of thousands food categories, remains an area yet to be researched. In this paper, we pose food recognition as a problem of recipe retrieval. Specifically, given a food picture, of whether the category has been seen during the training process, the aim is to retrieve a recipe for the food. The advantages of having recipe, rather than the name of food category, as output are numerous. Sharing food pictures in social media has been a trend. The ability to recommend recipes along will benefit users who want to cook a particular dish, and the feature is yet to be available. In addition, recipe provides rich information, such as cooking methods, ingredients and their quantities, which can facilitate the estimation of food balance and nutrition facts. The challenge of recipe retrieval, nevertheless, comes from the fact that there could be many recipes named under the same categories, each of which differs in the composition of ingredients. Figure 1 shows an example, where recommending the right recipe for "Yuba Salad" indeed requires also fine-grained recognition of ingredient composition.

This paper explores the recent advances in cross-modality learning for addressing the aforementioned problems. Specifically, given food pictures and their associated recipes, our aim is to learn a model that captures their correspondence by learning a joint embedding space for visual-and-text translation. We exploit and revise a deep model, stacked attention network (SAN) [31], originally proposed for visual question-answering for our purpose. The model learns the correspondence through assigning heavier weights to the attended regions relevant to the ingredients extracted from recipes. Notice that



Recipe name: *Yuba salad*
Ingredients:
Yuba (250g); Carrot (50g); Cucumber (80g); Sesame (1 spoon);
Cooking steps:
1. Cut the Yuba into sections.
2. Wash the carrot and cucumber, then cut into slices
3. ....

Recipe name: *Yuba salad*
Ingredients:
Yuba (2); Cucumber (2); Black fungus (a few); Celery (a few);
Cooking steps:
1. Cut the Yuba into sections.
2. Wash the cucumber, and cut into small dices.
3. ....

Recipe name: *Yuba salad*
Ingredients:
Yuba (300g); Red pepper (a few); Parsley (a few); Garlic(a few);
Cooking steps:
1. Cut the Yuba into sections.
2. Wash the red peppers and mince them.
3. ....

Fig. 1 Although recipe (**a**), (**b**) and (**c**) are all about "Yuba salad", only recipe (**a**) uses the exactly same ingredients as the dish picture. Retrieving best-match recipe requires fine-grained analysis of ingredient composition

ingredient-irrelevant but context-relevant regions may also be useful for recipe retrieval, for example, the “casserole” regions would be effective in identifying dishes like “casserole rice noodles” or “casserole tofu”. Similar case also happens for ingredients (e.g., water) appear in dish but are not written in recipe. Therefore, directly ignoring these regions would decrease retrieval performance. Thanks to attention mechanism, our model will not completely ignore context-relevant regions but assign weights that are usually lower than ingredient regions, so long as the contextual information is useful in reducing training error. For the task of recipe retrieval, fortunately the learning does not require much effort in labeling training examples. There are already millions of food-recipe pairs uploaded by professional and amateur chefs on various cooking websites, which can be freely leveraged for training. We demonstrate that using these online resources, a fairly decent model can be trained for recipe retrieval with minimal labeling effort. As input to SAN includes ingredients, the model has higher generalization ability in recognizing food categories unseen during training, as long as all or most ingredients are known. Furthermore, as ingredient composition is considered in SAN, the chance of retrieving the best-match recipes is also enhanced. To this end, the contribution of this paper lies in addressing of food recognition as a recipe retrieval problem. Under this umbrella, the problem is turned into cross-modality feature learning, which can integrally model three inter-related problems: scalable food recognition, fine-grained ingredient recognition and best-match recipe retrieval. The preliminary version of this paper is published in [6]. This paper provides more empirical insights and discussion of the proposed network, as followings:

- Empirical comparison to a recently published work [5].
- Visualization of attention maps in comparison to Pool5 feature maps.
- Generalization of the network to unseen food categories.
- Analysis of success and failure examples in best-match recipe retrieval.

The remaining of this paper is organized as follows. Section 2 describes the related works. Section 3 elaborates our architecture for region selection and joint embedding feature learning. Section 4 presents experimental results and finally Section 5 concludes this paper.

2 Related work

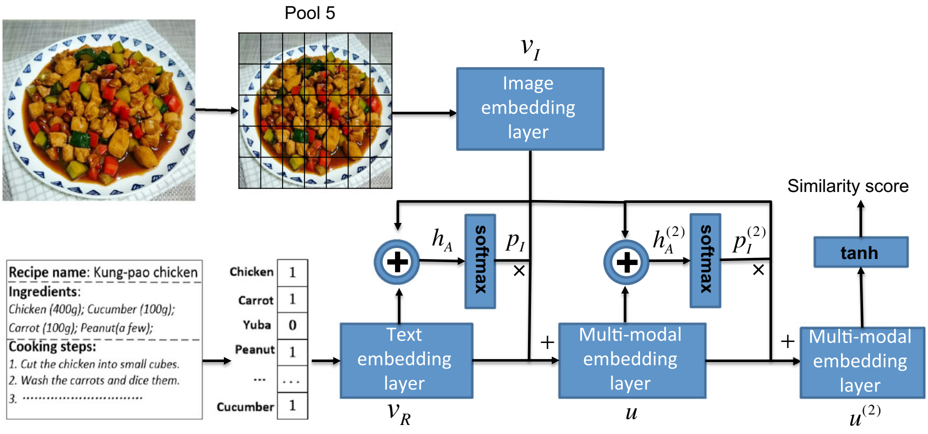
Analysis of recipes has been studied from different perspectives, including retrieval [5, 23, 26, 27], classification [17, 25] and recommendation [15]. Most of the approaches employ text-based analysis based upon information extracted from recipes. Examples include extraction of ingredients as features for cuisine classification [25] and taste estimation [17]. More sophisticated approaches model recipes as cooking graphs [27, 29] such that graph-based matching can be employed for similarity ranking of recipes. The graph, either manually or semi-automatically constructed from a recipe, represents the workflow for cooking and cutting procedures of ingredients. In [27], multi-modality information is explored, by late fusion of cooking graphs and low-level features extracted from food pictures, for example-based recipe retrieval. Few works have also studied cross-modality retrieval [5, 15, 26]. In [15], recognition of raw ingredients is studied for cooking recipe recommendation. Compared to prepared food where ingredients are mixed or even occlude each other, raw ingredients are easier to recognize. In [26], classifier-based approach is adopted for visual-to-text retrieval. Specifically, the category of food picture is first recognized, followed by retrieval of recipes under a category. As classifiers are trained from

UPMC Food-101 dataset [4], retrieval is only limited to 101 food categories. The issues in scalability and finding best-match recipes are not addressed. The recent work in [5] explores ingredient recognition for recipe retrieval. Using ingredient network as external knowledge, the approach is able to retrieve recipes even for unseen food categories. Different from [5], [23] aims to find a joint embedding of recipes and images for image-recipe retrieval task. More specifically, the joint space is learnt upon recipe and whole image which ignores regional features critical for fine-grained recognition. Different from [23], this paper aims to learn a joint space on regional level rather than image-level.

Cross-modality analysis has been actively researched for multimedia retrieval [8, 12, 21]. Frequently employed algorithms include canonical correlation analysis (CCA) [11] and partial least squares (PLS) [22], which find a pair of linear transformation to maximize the correlation between data from two modalities. CCA, in particular, has been extended to three-view CCA [10], semantic correlation matching (SCM) [21], deep CCA [2] and end-to-end deep CCA [30] for cross-modality analysis. Among variants of model, deep visual semantic embedding (DeViSE) [8] is generally used and usually exhibits satisfactory performance. These models, nevertheless, consider image-level features, such as fc7 extracted from deep convolutional network (DCNN), and usually ignore regional features critical for fine-grained recognition. One of the exceptions is the deep fragment embedding (DFE) proposed in [12], which aligns image objects and sentence fragments while learning the visual-text joint feature. However, the model is not applicable here for requiring of R-CNN [9] for object region detection. In the food domain, there is yet to have an algorithm for robust segmentation of ingredients, which can be fed into DFE for learning.

3 Stacked Attention Network (SAN)

Figure 2 illustrates the SAN model, with visual and text features respectively extracted from image and recipe as input. The model learns a joint space that boosts the similarity between images and their corresponding recipes. Different from [31], where the output layer is for classification, we modify SAN so as to maximize the similarity for image-recipe pairs.



Recipe name: Kung-pao chicken
Ingredients:
Chicken (400g); Cucumber (100g);
Carrot (100g); Peanut(a few);
Cooking steps:
1. Cut the chicken into small cubes.
2. Wash the carrots and dice them.
3. ....

Fig. 2 SAN model inspired from [31] for joint visual-text space learning and attention localization

3.1 Image embedding feature

The input visual feature is the last pooling layer of DCNN – Pool5 – that retains the spatial information of the original image. The dimension of Pool5 feature is 512×14×14, corresponding to 14×14 or 196 spatial grids of an image. Each grid is represented as a vector of 512 dimensions. Denote $\boldsymbol{F}_I$ as the Pool5 feature and is composed of regions $f_i, i \in [0{,}195]$. Each region f_i is transformed to a new vector or embedding feature as follows:

$$\boldsymbol{V}_I = \tanh(\boldsymbol{W}_I \boldsymbol{F}_I + \boldsymbol{b}_I), \tag{1}$$

where $\boldsymbol{V}_I \in \mathbb{R}^{d\times m}$ is the transformed feature matrix, with d as the dimension of new vector and $m = 196$ is the number of grids or regions. The embedding feature of f_i is indexed by i-th column of $\boldsymbol{V}_I$, denoted as $\boldsymbol{v}_i$. The transformation is performed region-wise, $\boldsymbol{W}_I \in \mathbb{R}^{d\times 512}$ is the transformation matrix and $\boldsymbol{b}_I \in \mathbb{R}^d$ is the bias term.

3.2 Recipe embedding feature

A recipe is represented as a binary vector of ingredients, denoted as $\boldsymbol{r} \in \mathbb{R}^t$. The dimension of the vector is t corresponding to the size of ingredient vocabulary. Each entry in $\boldsymbol{r}$ indicates the presence (1) or absence (0) of a particular ingredient in a recipe. As Pool5 feature, the vector is embedded into a new space as follows:

$$\boldsymbol{v}_R = \tanh(\boldsymbol{W}_R \boldsymbol{r} + \boldsymbol{b}_R), \tag{2}$$

where $\boldsymbol{W}_R \in \mathbb{R}^{d\times t}$ is the embedding matrix and $\boldsymbol{b}_R \in \mathbb{R}^d$ is the bias vector. Note that, for joint learning, the embedding features of recipe ($\boldsymbol{v}_R \in \mathbb{R}^d$) and Pool5 region ($i$-th column of $\boldsymbol{V}_I$) have the same dimension.

3.3 Joint embedding feature

The attention layer is to learn the joint feature by trying to locate the visual food regions that correspond to ingredients. There are two transformation matrices, $\boldsymbol{W}_{I,A} \in \mathbb{R}^{k\times d}$ for image I and $\boldsymbol{W}_{R,A} \in \mathbb{R}^{k\times d}$ for recipe R, mimicking the attention localization, formulated as follows:

$$\boldsymbol{H}_A = \tanh(\boldsymbol{W}_{I,A}\boldsymbol{V}_I \oplus (\boldsymbol{W}_{R,A}\boldsymbol{v}_R + \boldsymbol{b}_A)), \tag{3}$$

$$\boldsymbol{p}_I = \mathrm{softmax}(\boldsymbol{W}_P \boldsymbol{H}_A + \boldsymbol{b}_P), \tag{4}$$

where $\boldsymbol{H}_A \in \mathbb{R}^{k\times m}$, $\boldsymbol{p}_I \in \mathbb{R}^m$, $\boldsymbol{W}_P \in \mathbb{R}^{1\times k}$. We denote by $\oplus$ the addition of a matrix and a vector that performed by adding each column of matrix by the vector. Note that $\boldsymbol{p}_I$ aims to capture the attention, or more precisely relevance, of image regions to a recipe. The significance of a region f_i is indicated by the value in the corresponding element $p_i \in \boldsymbol{p}_I$.

The joint visual-text feature is basically generated by adding the embedding features $\boldsymbol{V}_I$ and $\boldsymbol{v}_R$. To incorporate attention value, regions $\boldsymbol{v}_i$ are linearly weighted and summed (equation-5) before the addition operation with $\boldsymbol{v}_R$ (equation-6), as follows:

$$\tilde{\boldsymbol{v}_I} = \sum_{i=1}^{m} p_i \boldsymbol{v}_i, \tag{5}$$

$$\boldsymbol{u} = \tilde{\boldsymbol{v}_I} + \boldsymbol{v}_R, \tag{6}$$

where $\tilde{\boldsymbol{v}_I} \in \mathbb{R}^d$, and $\boldsymbol{u} \in \mathbb{R}^d$ represents the joint embedding feature.

As suggested in [31], progressive learning by stacking multiple attention layers can boost the performance, but will heavily increase the training cost. We consider two-layer SAN, by feeding the output of first attention layer, $\boldsymbol{u}^{(1)}$, into the second layer to generate new joint embedding feature $\boldsymbol{u}^{(2)}$ as follows:

$$\boldsymbol{H}_A^{(2)} = \tanh\left(\boldsymbol{W}_{I,A}^{(2)}\boldsymbol{V}_I \oplus \left(\boldsymbol{W}_{R,A}^{(2)}\boldsymbol{u} + \boldsymbol{b}_A^{(2)}\right)\right), \tag{7}$$

$$\boldsymbol{p}_I^{(2)} = \text{softmax}\left(\boldsymbol{W}_P^{(2)}\boldsymbol{H}_A^{(2)} + \boldsymbol{b}_P^{(2)}\right), \tag{8}$$

$$\tilde{\boldsymbol{v}_I^{(2)}} = \sum_i p_i^{(2)}\boldsymbol{v}_i, \tag{9}$$

$$\boldsymbol{u}^{(2)} = \tilde{\boldsymbol{v}_I^{(2)}} + \boldsymbol{u}. \tag{10}$$

As $\boldsymbol{p}_I^{(2)}$ indicates the region relevancy, the attention map can be visualized by back projecting the attention value p_i to its corresponding region f_i, followed by upsampling to the original image size with bicubic interpolation.

3.4 Objective function

To this end, the similarity between food image and recipe is generated as follows:

$$S\langle \boldsymbol{V}_I, \boldsymbol{v}_R\rangle = \tanh(\boldsymbol{W}_{u,s}\boldsymbol{u}^{(2)} + b_s), \tag{11}$$

where $\boldsymbol{W}_{u,s} \in \mathbb{R}^{1\times d}$ and $b_s \in \mathbb{R}$ is the bias. $S\langle \boldsymbol{V}_I, \boldsymbol{v}_R\rangle$ outputs a score indicating the association between the embedding features of image and recipe. The learning is based on the following rank-based loss function with a large margin form as the objective function:

$$\mathcal{L}(\boldsymbol{W}, D_{trn}) = \sum_{(\boldsymbol{V}_I,\boldsymbol{v}_R^+,\boldsymbol{v}_R^-)\in D_{trn}} \max(0, \triangle + S\langle \boldsymbol{V}_I, \boldsymbol{v}_R^-\rangle - S\langle \boldsymbol{V}_I, \boldsymbol{v}_R^+\rangle). \tag{12}$$

The training set, D_{trn}, consists of triples in the form of $(\boldsymbol{V}_I, \boldsymbol{v}_R^+, \boldsymbol{v}_R^-)$, where $\boldsymbol{v}_R^+$ $(\boldsymbol{v}_R^-)$ is true (false) recipe for food V_I. The matrix $\boldsymbol{W}$ represents the network parameters, and $\triangle \in (0, 1)$ controls the margin in training and is cross-validated.

4 Experiments

4.1 Settings and evaluation

Here we detail the parameter setting of SAN. The dimension of the embedding feature is set to $d = 500$ for both Pool5 regional and recipe features, while the dimension of h_A is $k = 1,024$ for equations 3 and 7. Through cross-validation, the hyperparameter $\triangle$ for the loss function is set as 0.2. SAN is trained using stochastic gradient descent with momentum set as 0.9 and the initial learning rate as 1. The size of mini-batch is 50 and the training stops after 10 epochs. To prevent overfitting, dropout [24] is used. The Pool5 feature can be extracted from any DCNN models. We employ the multi-task VGG released by [5], which reported the best performances on two large food datasets, VIREO Food-172 [5] and UEC Food-100 [16]. The model, as shown in Fig. 3, has two pathways, one for classifying 172 food categories while another for labeling 353 ingredients. For a fair comparison, all the compared approaches in the experiment are using multi-task VGG features, either Pool5 or deep ingredient feature (fc7), as shown in Fig. 3.

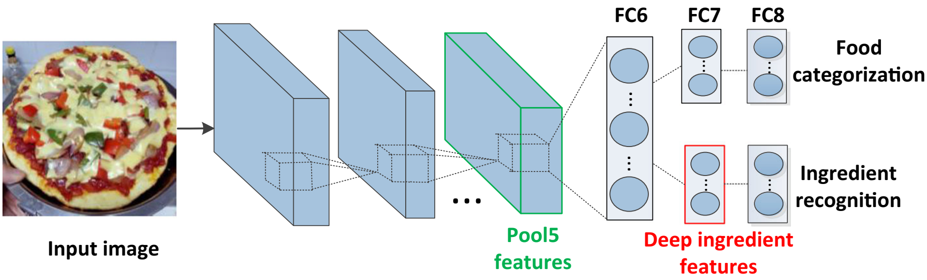


Fig. 3 Multi-task VGG model in [5] offering Pool5 and deep ingredient features for cross-modal joint space learning

As the task is to find the best possible recipe given a food picture, the following two measures are employed for performance evaluation:

- Mean reciprocal rank (MRR): MRR measures the reciprocal of rank position where the ground truth recipe is returned, averaged over all the queries. This measure assesses the ability of the system to return the correct recipe at the top of the ranking. The value of MRR is within the range of [0, 1]. A higher score indicates a better performance.
- Recall at Top-K (R@K): R@K computes the fraction of times that a correct recipe is found within the top-K retrieved candidates. R@K provides an intuitive sense of how quickly the best recipe can be located by investigating a subset of the retrieved items. As MRR, a higher score also indicates a better performance.

4.2 Dataset

The dataset is composed of 61,139 image-recipe pairs crawled from the "Go Cooking"[1] websites. Each pair consists of a recipe and a picture of resolution 448 × 448. The dataset covers different kinds of food, like Chinese dishes, snacks, dessert, cookies and Chinese-style western food, as shown in Fig. 4. Each recipe includes the list of ingredients and cooking procedure. As the recipes were uploaded by amateurs, the naming of ingredients is not always consistent. For example, "carrot" is sometimes called as "carotte". We manually rectified the inconsistency and compiled a list of 5,990 ingredients, both visible and non-visible (e.g., "honey"), from these recipes. The list, represented as a binary vector indicating the presence or absence of particular ingredients in a recipe, serves as input to the SAN model. Note that in some cases the cooking and cutting methods are directly embedded into the name of ingredient, for example, "tofu" and "tofu piece", "egg" and "steamed egg".

The dataset is split into three sets: 54,139 pairs for training, 2,000 pairs for cross-validation, and 5,000 pairs for testing. Furthermore, we selected 1,000 images from the testing set as queries to search against the 5,000 recipes. The queries are sampled in such a way that there are around 45% of them (446 queries) belonging to food categories unknown to SAN and multi-task VGG models. In addition, around 85% of the queries have more than one relevant recipe. We recruite a homemaker, who has cooking experience, to manually pick the relevant recipes for each of the 1,000 queries. The homemaker is instructed to label relevant recipes based on title similarity in recipes, titles that are named differently because

[1] https://www.xiachufang.com

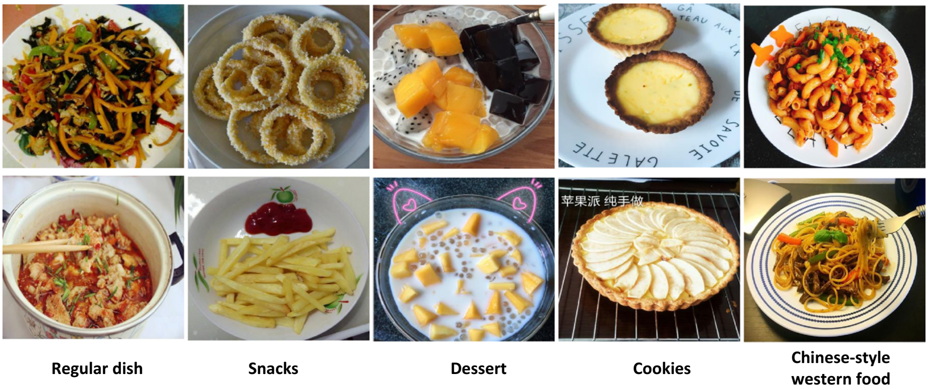


Fig. 4 Examples of dishes in the dataset

of geography regions or sharing almost the same cooking procedure with similar key ingredients. For example, the dish “sauteed tofu in hot and spicy sauce” is sometimes called as “mapo tofu” in the restaurant menu. In the extreme case, some queries have more than 60 relevant recipes. On average each query has 9 number of relevant recipes. Note that the testing queries are designed in these ways so as to verify the two major claims in this paper, i.e., the degree in which the learnt model can generalize to unseen food categories (Section 4.4) and the capability in finding the best-matched recipe (Section 4.5).

4.3 Performance comparison

We compared SAN to both shallow and deep models for cross-modal retrieval as following. The inputs to these models are the deep ingredient feature (fc7) of the multi-task VGG model and the ingredient vector of 5,990 dimensions. The Pool5 feature is not used due to its high dimensionality (14×14×512). As reported in [7], simply concatenating the features from 14×14 grids performs worse than fc7 in visual recognition.

- Canonical Correlation Analysis (CCA) [11]: CCA is a classic way of learning latent subspace between two views or features by maximizing the correlation between them. Two linear mapping functions are learnt for projecting features into subspace.
- Partial Least Squares (PLS) [22]: Similar to CCA, PLS learns two linear mapping functions between two views. Instead of using cosine similarity as in CCA, PLS uses dot product as the function for measuring correlation.
- DeViSE [8]: DeViSE is a deep model with two pathways which respectively learn the embedded features of recipe-image pairs to maximize their similarities. Note that, instead of directly using word2vec as in [8], the embedded feature of ingredients is learnt from the training set of our dataset. This is simply because word2vec is learnt from documents such as news corpus [19] and lacks specificity in capturing information peculiar to ingredients. Different from SAN, DeViSE is not designed for attention region localization.
- DeViSE++: We purposely includ a variant of DeViSE, which takes the hand-cropped regions of food as input to the deep model. The cropping highlights the target food region and basically removes the background or irrelevant part of food pictures. The aim of using DeViSE++ is to gate the potential improvement over DeViSE when only food

Table 1 MRR and R@K for recipe retrieval

Method	MRR	R@1	R@5	R@10	R@20	R@40	R@60	R@80	R@100
CCA	0.055	0.023	0.079	0.123	0.182	0.262	0.329	0.371	0.413
PLS	0.032	0.009	0.039	0.073	0.129	0.219	0.284	0.338	0.398
DeViSE	0.049	0.016	0.060	0.108	0.182	0.300	0.391	0.456	0.524
DeViSE++	0.050	0.016	0.059	0.105	0.174	0.307	0.404	0.471	0.531
Multi task [5]	0.097	**0.051**	0.128	0.184	0.251	0.324	0.372	0.408	0.438
SAN	**0.115**	0.048	**0.161**	**0.249**	**0.364**	**0.508**	**0.601**	**0.671**	**0.730**

The best performance is highlighted in bold font

region is considered, and more importantly, to justify the merit of SAN in identifying appropriate attention region in comparison to the hand-cropped region.

- Multi task [5]: In Multi task [5] model, ingredient recognition is formulated as a problem of multi-task learning and the learnt semantic labels as well as the external knowledge of the contextual relations among ingredients are utilized for recipe retrieval.

Table 1 lists the results of different approaches. Deep models basically outperform shallow models in terms of recall at the depth of 20 and beyond. In contrast to PLS, which does not perform score normalization, CCA manages to outperform DeViSE in terms of MRR and R@K for $K < 20$. Among all these approaches, the proposed model SAN consistently exhibits the best performance in terms of MRR. Compared to DeViSE and Multi task, SAN achieves a relative improvement of 130% and 18% in MRR, respectively. In terms of R@K, SAN performs significantly better than DeViSE and doubles its performance at R@20, which is fairly impressive. Compared with Multi task model, SAN also performs much better when $K > 5$, and the performance gap becomes larger when the depth increases.

To further provide insights, Fig. 5 visualizes the attention maps learnt from SAN while comparing to Pool5 feature maps. From the figure, it is obvious that the learnt attention model can locate the ingredient regions more accurately than Pool5 feature maps.

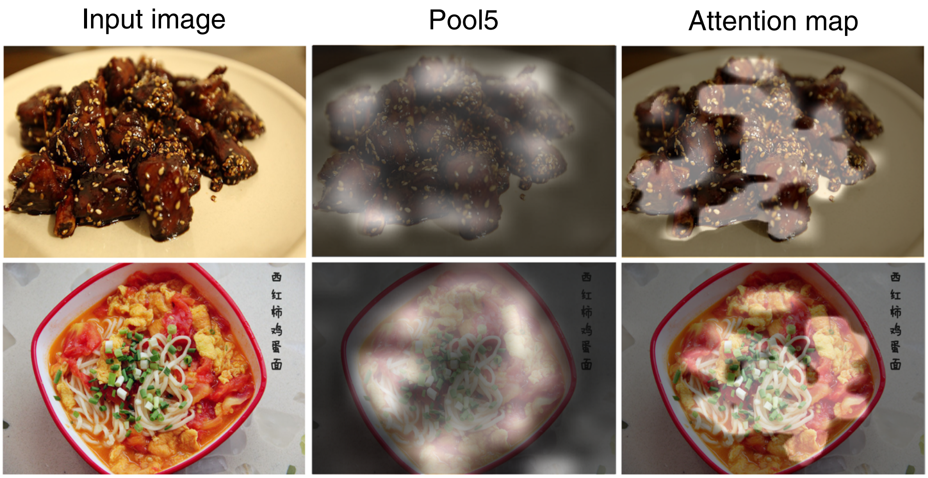


Fig. 5 Visualizing attention maps, the learnt attention regions are highlighted in white

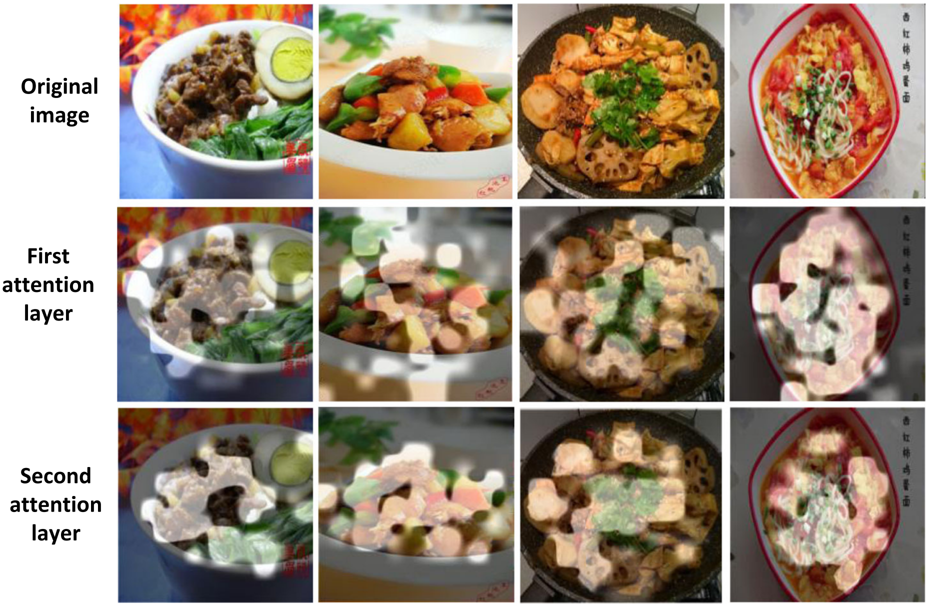


Fig. 6 Visualization of two attention layers

To observe the difference between the two learnt attention layers, we visualize the attention maps $\boldsymbol{p}_I$ and $\boldsymbol{p}_I^{(2)}$. Four examples are shown in Fig. 6. From the figure, we can observe that the second attention layer reduces the noises in the first layer and hence is more accurate in localizing ingredient regions.

Despite the encouraging performance by SAN, the value of R@1 is only around 0.05. Figure 7 shows some successful and near-miss examples. The first two pictures show query images where all visible ingredients are clearly seen. SAN manages to retrieve the ground-truth recipe at top-1 rank in such cases. In the third example, SAN ranks “grilled salmon” higher than “fried salmon” as the current model does not consider cooking attributes. In addition, SAN overlooks the beef and peanuts which are mixed and partially occluded by salmon, while confused by the ingredients of similar appearance, i.e., caviar and red pepper, bean sprout and basil. The last query image shows an example of how non-visible ingredients, flour in this example, affect the ranking. The flour is used to make the dish into round shape, and this knowledge does not seem to be learnt by SAN.

Another result worth noticing is that there is no performance difference between DeViSE and DeViSE++. While DeViSE is not designed for attention localization, the model seems to have the ability to exclude irrelevant background regions from recognition. To provide further insights, Fig. 8 shows some examples visualizing the attention regions highlighted by SAN and in contrast to hand-crafted regions. In the first example, the region attended by SAN is about the same as the region manually cropped. In this case, DeViSE+ and SAN use to have similar performance. The next two examples highlight the superiority of SAN in excluding soup and foil as attention regions, which cannot be not easily done by simple region cropping. SAN significantly outperforms DeViSE in such examples. Finally, the last example shows a typical case that SAN only highlights part of dishes as attention. While there is no direct explanation of why certain food regions are ignored by SAN for joint space

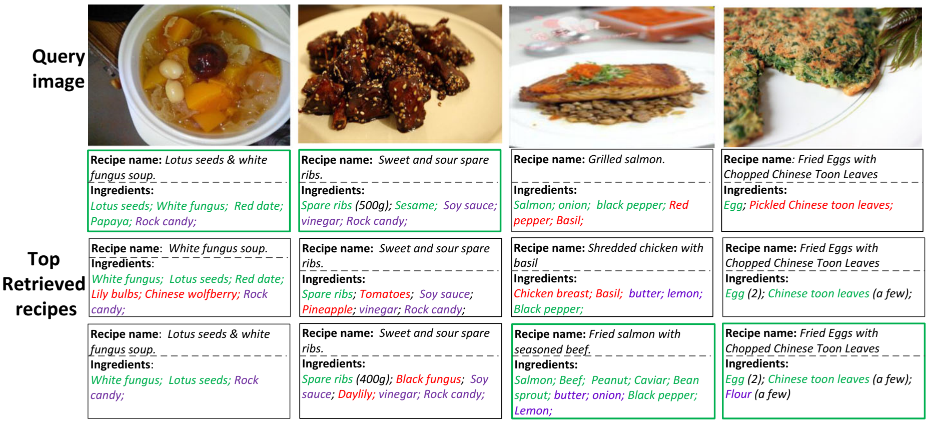

Query image
Top Retrieved recipes
Recipe name: Lotus seeds & white fungus soup.
Ingredients: Lotus seeds; White fungus; Red date; Papaya; Rock candy;
Recipe name: Sweet and sour spare ribs.
Ingredients: Spare ribs (500g); Sesame; Soy sauce; vinegar; Rock candy;
Recipe name: Grilled salmon.
Ingredients: Salmon; onion; black pepper; Red pepper; Basil;
Recipe name: Fried Eggs with Chopped Chinese Toon Leaves
Ingredients: Egg; Pickled Chinese toon leaves;
Recipe name: White fungus soup.
Ingredients: White fungus; Lotus seeds; Red date; Lily bulbs; Chinese wolfberry; Rock candy;
Recipe name: Sweet and sour spare ribs.
Ingredients: Spare ribs; Tomatoes; Soy sauce; Pineapple; vinegar; Rock candy;
Recipe name: Shredded chicken with basil
Ingredients: Chicken breast; Basil; butter; lemon; Black pepper;
Recipe name: Fried Eggs with Chopped Chinese Toon Leaves
Ingredients: Egg (2); Chinese toon leaves (a few);
Recipe name: Lotus seeds & white fungus soup.
Ingredients: White fungus; Lotus seeds; Rock candy;
Recipe name: Sweet and sour spare ribs.
Ingredients: Spare ribs (400g); Black fungus; Soy sauce; Daylily; vinegar; Rock candy;
Recipe name: Fried salmon with seasoned beef.
Ingredients: Salmon; Beef; Peanut; Caviar; Bean sprout; butter; onion; Black pepper; Lemon;
Recipe name: Fried Eggs with Chopped Chinese Toon Leaves
Ingredients: Egg (2); Chinese toon leaves (a few); Flour (a few)

Fig. 7 Examples of top-3 retrieved recipes (ranked from top to bottom). Ground-truth recipe is marked in green. The ingredients in different colors have different meanings: green – true positive, purple – true positive but non-visible in dish, red – false positive

learning, it seems that SAN has the ability to exclude regions that are vague and hard to be recognized even by human.

4.4 Finding the best matches recipes

Recalled that around 85% of query images have more than one relevant recipe. This section examines the ability of SAN in identifying the best (or ground-truth) recipe from the testing set composed of 5,000 recipes. Figure 9 shows the performance of best match recipe retrieval compares with relevant recipe retrieval. For recall@top5, the performance of relevant recipe retrieval improves when the number of relevant recipe increases while the trend is opposite for best-match recipe retrieval.

To provide insights, we select the queries that retrieve at least one relevant recipe (excluding ground-truth recipe) within the top-5 position for analysis. The purpose is to show

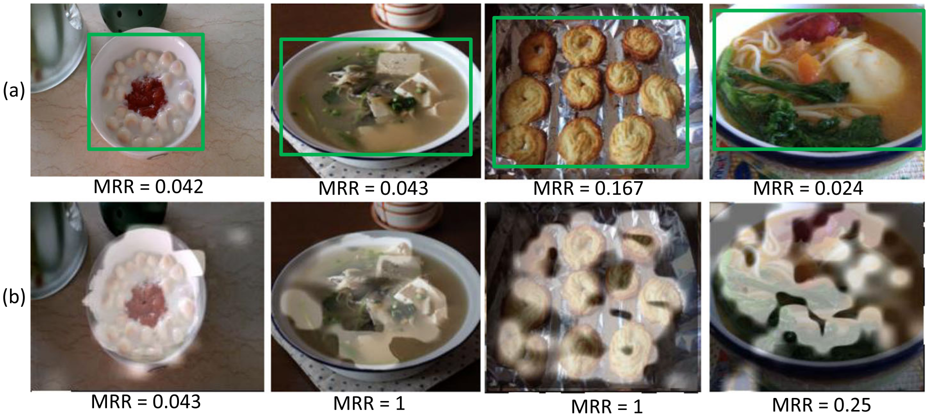


Fig. 8 **a** Examples contrasting the manually cropped region (green bounding box), **b** the learnt attention region (masked in white) by SAN

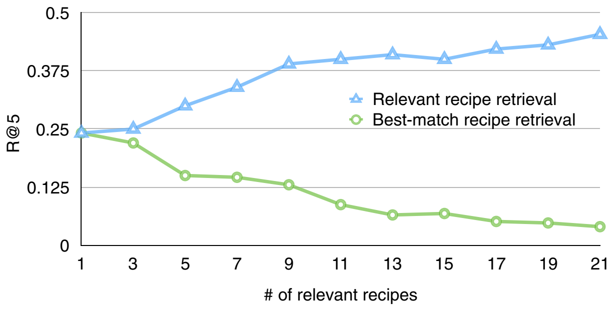


Fig. 9 Performance of best match recipe retrieval and relevant recipe retrieval

how the performance of best-match recipe retrieval is impacted by the increasing number of relevant recipes. We divide the selected queries into seven groups with the intention to make the number of queries in each group as even as possible. Note that, as the numbers of recipes distribute in a long-tail like manner, the recipe numbers in each group are uneven. Table 2 lists the performance for each group. As can be seen from the table, the difficulty of finding best-match is proportional to the number of relevant recipes. Compared to DeViSE, SAN generally shows better performance for R@1. As the number of recipes increases, they tie in performance. Nevertheless, while looking deeper into the list, SAN consistently outperforms DeViSE in terms of R@5 and R@10. Two main reasons that ground truth recipes are not ranked higher are due to occluded ingredients and the use of different non-visible ingredients. Two such examples are shown in the last two pictures of Fig. 7.

4.5 Generalization to unknown categories

Figure 10 further shows the performance of SAN to unseen categories. As expected, the performance is not as good as that for the food categories known to SAN and multi-task VGG. Figure 11 shows both success and failure examples of recipe retrieval. Basically, when the ingredients of unknown food categories are previously seen and can be correctly identified, SAN performs satisfactorily. In contrast, when some ingredients, especially key ingredients, are unknown, the model will likely fail in retrieving relevant recipes. In the first example, the ingredients are correctly recognized despite that the dish belongs to unseen

Table 2 Performance comparison between SAN and DeViSE in retrieving best-match recipes

		R@1		R@5		R@10	
Recipe #	Query #	SAN	DeViSE	SAN	DeViSE	SAN	DeViSE
2-3	33	0.21	0.15	0.67	0.48	0.82	0.76
4-7	66	0.18	0.17	0.56	0.53	0.70	0.67
8-11	54	0.17	0.15	0.54	0.30	0.60	0.50
11-15	38	0.13	0.08	0.47	0.39	0.63	0.55
16-30	48	0.06	0.06	0.46	0.39	0.62	0.52
31-61	25	0.08	0.08	0.28	0.26	0.44	0.44

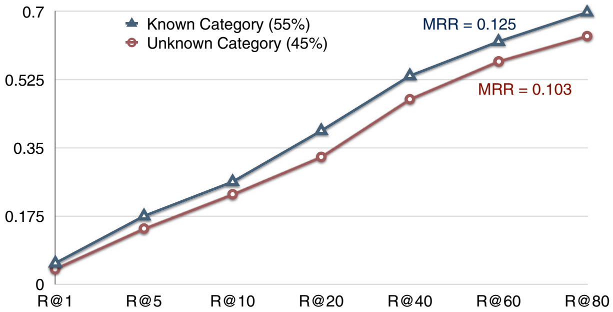
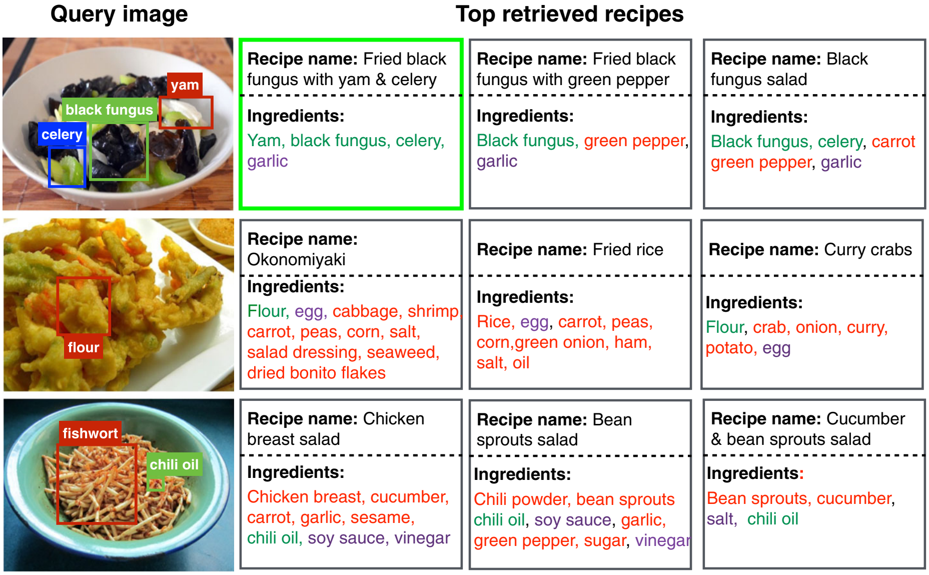


Fig. 10 Generalization of SAN to unseen food categories

food categories. As results, our model is able to rank the best match recipe at the top-1 place. However, when the ingredient is covered by flour (second example), the model is unlikely to recognize the ingredients and hence fails to retrieve the correct recipes at top ranks. Finally, when the dish contains unseen key ingredients, for example, “fishwort” in the third example, our model will fail.

We further compare the generalization ability of our model with DeViSE and Multi task [5]. The retrieval performances are evaluated on 446 queries that come from unknown food categories. As can be seen from the Fig. 12, our model enjoys higher generalization ability and the performance gap becomes larger when the depth of recall increase. The better generalization ability of our model verifies the advantages of cross-modal learning on region-level with stacked attention networks.



Recipe name: Fried black fungus with yam & celery
Ingredients: Yam, black fungus, celery, garlic

Recipe name: Fried black fungus with green pepper
Ingredients: Black fungus, green pepper, garlic

Recipe name: Black fungus salad
Ingredients: Black fungus, celery, carrot green pepper, garlic

Recipe name: Okonomiyaki
Ingredients: Flour, egg, cabbage, shrimp carrot, peas, corn, salt, salad dressing, seaweed, dried bonito flakes

Recipe name: Fried rice
Ingredients: Rice, egg, carrot, peas, corn,green onion, ham, salt, oil

Recipe name: Curry crabs
Ingredients: Flour, crab, onion, curry, potato, egg

Recipe name: Chicken breast salad
Ingredients: Chicken breast, cucumber, carrot, garlic, sesame, chili oil, soy sauce, vinegar

Recipe name: Bean sprouts salad
Ingredients: Chili powder, bean sprouts chili oil, soy sauce, garlic, green pepper, sugar, vinegar

Recipe name: Cucumber & bean sprouts salad
Ingredients: Bean sprouts, cucumber, salt, chili oil

Fig. 11 Examples of top-3 retrieved recipes for unknown food categories. Ground-truth recipe is marked in green. The ingredients in different colors have different meanings: green – true positive, purple – true positive but non-visible in dish, red – false positive

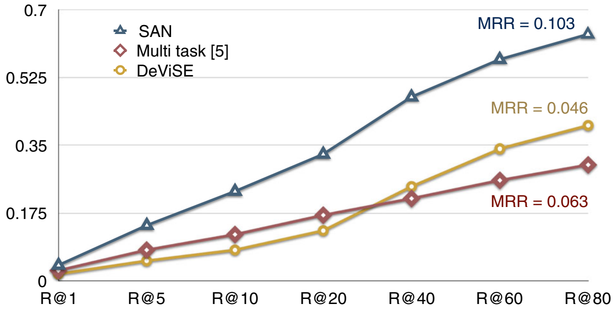


Fig. 12 Comparison of generalization ability among different methods

5 Conclusion

We have presented a deep model for learning the commonality between image and text at the fine-grained ingredient level. The power of model comes from the ability to infer attended regions relevant to the ingredients extracted from recipes. This peculiarity enables retrieval of best-match recipes even for unseen food category. The merit of our approach is that it requires much less labeling efforts compared to learning individual ingredient classifiers. The experimental results basically verify our claims that the model can deal with unknown food categories to the extent that at least key ingredients are seen during training. In addition, SAN exhibits consistently better performance than DeViSE, showing the advantage of fine-grained ingredient analysis at the regional level for best-match recipe retrieval.

While the current model does not consider food category information, it is expected that such information will boost performance especially when there are errors in ingredient localization and attention modeling. How to incorporate food category information into the current model is worth further investigation. Finally, our current model can be extended to explicitly model cutting and cooking attributes in cross-modal learning, which could address some limitations identified in the experiments. In addition, as the attention layers couple both visual and text features, the embedding features cannot be offline indexed and have to be generated on-the-fly when the query image is given. This poses limitation on retrieval speed for online application, which is an issue needs to be further researched.

Acknowledgements This work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11203517).

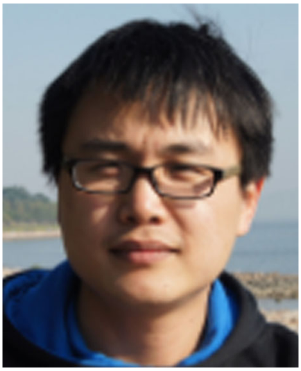
References

1. Aizawa K, Ogawa M (2015) Foodlog: multimedia tool for healthcare applications. IEEE Multimed 22(2):4–8
2. Andrew G, Arora R, Bilmes JA, Livescu K (2013) Deep canonical correlation analysis. In: Proceedings of international conference on machine learning, pp 1247–1255
3. Beijbom O, Joshi N, Morris D, Saponas S, Khullar S (2015) Menu-match: restaurant-specific food logging from images. In: Proceedings of IEEE workshop on applications of computer and vision, pp 844–851
4. Bossard L, Guillaumin M, Van Gool L (2014) Food-101–mining discriminative components with random forests. In: Proceedings of european conference on computer vision, pp 446–461

5. Chen J, Ngo CW (2016) Deep-based ingredient recognition for cooking recipe retrieval. In: Proceedings of ACM international conference on multimedia
6. Chen J, Pang L, Ngo CW (2017) Cross-modal recipe retrieval: how to cook this dish? In: Proceedings of international conference on multimedia modeling. Springer, pp 588–600
7. Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) Decaf: a deep convolutional activation feature for generic visual recognition. In: Proceedings of international conference on machine learning, pp 647–655
8. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Mikolov T et al (2013) Devise: a deep visual-semantic embedding model. In: Proceedings of neural information processing systems, pp 2121–2129
9. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 580–587
10. Gong Y, Ke Q, Isard M, Lazebnik S (2014) A multi-view embedding space for modeling internet images, tags, and their semantics. Int J Comput Vis 106(2):210–233
11. Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: an overview with application to learning methods. Neural Comput 16(12):2639–2664
12. Karpathy A, Joulin A, Li FF (2014) Deep fragment embeddings for bidirectional image sentence mapping. In: Proceedings of neural information processing systems, pp 1889–1897
13. Kawano Y, Yanai K (2014) Foodcam-256: a large-scale real-time mobile food recognitionsystem employing high-dimensional features and compression of classifier weights. In: Proceedings of ACM international conference on multimedia, pp 761–762
14. Kitamura K, Yamasaki T, Aizawa K (2008) Food log by analyzing food images. In: Proceedings of ACM international conference on multimedia, pp 999–1000
15. Maruyama T, Kawano Y, Yanai K (2012) Real-time mobile recipe recommendation system using food ingredient recognition. In: Proceedings of ACM international workshop on interactive multimedia on mobile and portable devices, pp 27–34
16. Matsuda Y, Hoashi H, Yanai K (2012) Recognition of multiple-food images by detecting candidate regions. In: Proceedings of international conference on multimedia and expo
17. Matsunaga H, Doman K, Hirayama T, Ide I, Deguchi D, Murase H (2015) Tastes and textures estimation of foods based on the analysis of its ingredients list and image. In: New trends in image analysis and processing–ICIAP 2015 workshops, pp 326–333
18. Meyers A, Johnston N, Rathod V, Korattikara A, Gorban A, Silberman N, Guadarrama S, Papandreou G, Huang J, Murphy KP (2015) Im2calories: towards an automated mobile vision food diary. In: Proceedings of IEEE international conference on computer vision, pp 1233–1241
19. Mikolov T, Dean J (2013) Distributed representations of words and phrases and their compositionality
20. Probst Y, Nguyen DT, Rollo M, Li W (2015) mhealth diet and nutrition guidance. mHealth
21. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GR, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of ACM international conference on multimedia, pp 251–260
22. Rosipal R, Krämer N. (2006) Overview and recent advances in partial least squares. In: Subspace, latent structure and feature selection. Springer, pp 34–51
23. Salvador A, Hynes N, Aytar Y, Marin J, Ofli F, Weber I, Torralba A (2017) Learning cross-modal embeddings for cooking recipes and food images. In: Proceedings of IEEE conference on computer vision and pattern recognition
24. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
25. Su H, Lin TW, Li CT, Shan MK, Chang J (2014) Automatic recipe cuisine classification by ingredients. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: adjunct publication, pp 565–570
26. Wang X, Kumar D, Thome N, Cord M, Precioso F (2015) Recipe recognition with large multimodal food dataset. In: Proceedings of international conference on multimedia and expo workshop, pp 1–6
27. Xie H, Yu L, Li Q (2010) A hybrid semantic item model for recipe search by example. In: 2010 IEEE international symposium on Proceedings of multimedia (ISM), pp 254–259
28. Xu R, Herranz L, Jiang S, Wang S, Song X, Jain R (2015) Geolocalized modeling for dish recognition. IEEE Trans Multimed 17(8):1187–1199
29. Yamakata Y, Imahori S, Maeta H, Mori S (2016) A method for extracting major workflow composed of ingredients, tools and actions from cooking procedural text. In: 8Th workshop on multimedia for cooking and eating activities
30. Yan F, Mikolajczyk K (2015) Deep correlation for matching images and text. In: Proceedings of international conference on machine learning, pp 3441–3450

31. Yang Z, He X, Gao J, Deng L, Smola A (2015) Stacked attention networks for image question answering. arXiv:1511.02274
32. Zhang W, Yu Q, Siddiquie B, Divakaran A, Sawhney H (2015) Snap-n-eat: food recognition and nutrition estimation on a smartphone. J Diabetes Sci Technol 9(3):525–533

Jing-Jing Chen received the B.Sc. degree from Wuhan University of Technology, Wuhan, China, in 2011, the M.Sc. degree from the Tianjin University, Tianjin, China, in 2014. She is currently working toward the Ph.D. degree in computer science at the City University of Hong Kong, Hong Kong. She is currently with VIREO Group, City University of Hong Kong. Her research interest lies in diet tracking and nutrition estimation based on multi-modal processing of food images, including food recognition, crossmodal recipe retrival.



Lei Pang received the B. Eng degree from Nankai University, Tianjin, China, in 2010, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2015. He is currently a staff researcher with iFlight Limited Company. His research interest lies in multimedia content analysis, including Web video face naming, multimedia question answering, and emotion prediction on Web videos.

Chong-Wah Ngo received the B.Sc. and M.Sc. degrees in computer engineering from Nanyang Technological University, Singapore, and the Ph.D. in computer science from the Hong Kong University of Science and Technology (HKUST), Hong Kong. He is currently a Professor with the department of Computer Science, City University of Hong Kong, Hong Kong. Before joining the City University of Hong Kong, he was a Postdoctoral Scholar with the Beckman Institute, University of Illinois at Urbana- Champaign (UIUC), Urbanna, IL, USA. He was also a Visiting Researcher with Microsoft Research Asia, Beijing,China. His research interests include large-scale multimedia information retrieval, video computing, multimedia mining, and visualization. Prof. Ngo was the Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (2011-2014). He was the Conference Co-Chair of the ACM International Conference on Multimedia Retrieval 2015 and the Pacific Rim Conference on Multimedia 2014. He also served as Program Co- Chair of ACM Multimedia Modeling 2012 and ICMR 2012. He was the Chairman of ACM (Hong Kong Chapter) from 2008 to 2009.