Recipe Retrieval with Visual Query of Ingredients

Yen-Chieh Lien
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
ylien@cs.umass.edu

Hamed Zamani
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
zamani@cs.umass.edu

W. Bruce Croft
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
croft@cs.umass.edu

ABSTRACT

Recipe retrieval is a representative and useful application of cross-modal information retrieval. Recent studies have proposed frameworks for retrieving images of cuisines given *textual* ingredient lists and instructions. However, the textual form of ingredients easily causes information loss or inaccurate description, especially for novices of cookery who are often the main users of recipe retrieval systems. In this paper, we revisit the task of recipe retrieval by taking *images* of ingredients as input queries, and retrieving cuisine images by incorporating visual information of ingredients through a deep convolutional neural network. We build an image-to-image recipe retrieval system to validate the effect of ingredient image queries. We further combine the proposed solution with a state-of-the-art cross-modal recipe retrieval model to improve the overall performance of the recipe retrieval task.

1 INTRODUCTION

Recipe retrieval is one of the representative and useful domain-specific information retrieval applications. It has been shown that people who like cooking enjoy a recipe retrieval system that suggests them some cuisines that they can make based on the ingredients they have available [7]. A good recipe often contains *textual* and *visual* information of ingredients, cooking steps, and the final dish. However, the current state-of-the-art models for the task are based on cross-modal neural retrieval models, e.g., [7]. In more detail, they assume that recipes are text and include the ingredient names and cooking instructions. This textual information together with dish images are used to train a cross-modal recipe retrieval model. However, text-only queries may not be accurate for the recipe retrieval task. For example, an ingredient could have multiple varieties of breeds (i.e. Grape Tomato and Big Boy Tomato) and different breeds have their own special textures and flavors, so two breeds of the same ingredient may be suitable for two different styles of dishes. A user who is new to culinary art would have trouble to notice the difference and have a chance of making wrong or ambiguous queries. Therefore, textual queries may easily cause information loss or misunderstanding. On the other hand, using pictures as queries could avoid this problem. In addition, taking a picture of remaining ingredients and searching for feasible dishes is desired by many users. Therefore, we believe that using visual queries is practical, informative, and accurate for the recipe retrieval task.

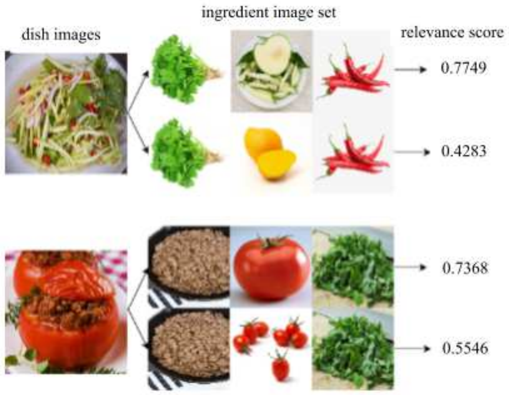

Figure 1: Two examples that show replacing an ingredient with a different breed of type of the same ingredient, the system should produce lower relevance score.

To clarify our perspective, we provide demonstrations of two cases in Figure 1. The chosen examples are “Risotto and Beef Stuffed Tomatoes” and “Thai Green Mango Salad”. Both dishes have ingredients that could have other breeds or forms. If we swap the crucial ingredients, which are the big boy tomato and the raw (green) mango, respectively with a different breed or type of the same ingredient, grape tomato and yellow mango, a good system should give the correct ingredient lists significantly larger similarity with the true positive dishes than the swapped ingredient lists. For these cases, visual queries are reliable because they all look different, but an inaccurate textual query like “tomato” may fail to match the right dish.

In this paper, we propose a framework to exploit ingredient images for dish image retrieval. In other words, the user can take pictures of available ingredients in their home and the system can retrieve dishes that can be cooked using the available ingredients. We learn representations of ingredients and dishes using convolutional neural networks (CNNs) and compute their matching score based on the similarity of the learned representations. The architecture has an analogy to query-document matching model, which is a common neural retrieval framework for ad-hoc text retrieval. The set of ingredient images are treated as a set of query terms and the images of all dishes are regarded as a collection of documents. We apply the architecture for image queries and documents by replacing text representation, including word embedding and recurrent

SIGIR '20, July 25–30, 2020, Xi'an, China

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

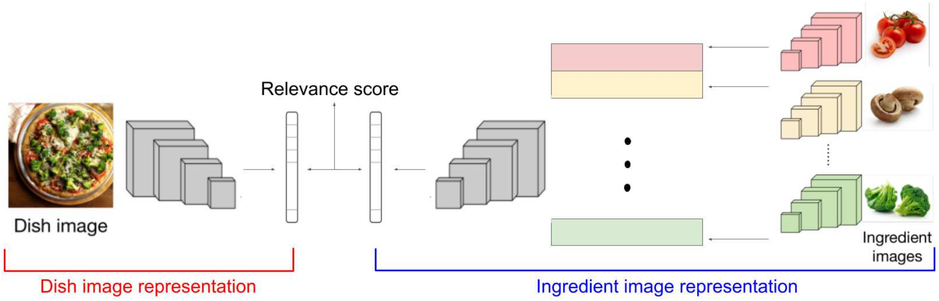

Figure 2: The proposed image-to-image architecture.

network, with a CNN-based network, which is more common for visual information.

We evaluate our method on Recipe1M [7], which is the largest public dataset for recipe retrieval. Since previous work has focused on cross-modal retrieval for this task, Recipe1M does not contain ingredient images. Therefore, we first collect ingredient images through the Google search engine, and consider them as input.[1] Our experiments demonstrate the effectiveness of the proposed image-to-image recipe retrieval model. We also improve a state-of-the-art cross-modal recipe retrieval model by combining it with the proposed image-to-image model.

In summary, our contributions include: (1) conducting the first study on image-to-image recipe retrieval, (2) proposing a neural model for the task, (3) expanding the Recipe1M dataset by including ingredient images, and (4) advancing the state-of-the-art result for the cross-modal recipe retrieval task.

2 RELATED WORK

As more food-related datasets have emerged, different analyses of recipes have been studied including classification [6, 8], recommendation [5] and retrieval [3, 9, 10]. Text-based analysis upon the parts of recipes is favored in most of those studies. Previous works include extraction of ingredients as features for taste/texture estimation [6] and for cuisine classification [8]. Graph-based analysis, which structures recipes into cooking graphs [10, 11] to represent the workflow of cooking procedures of ingredients, has been used for similarity ranking of recipes. In [10], multi-modality information was explored by late fusion of cooking graphs and low-level features extracted from food pictures for example-based recipe retrieval. Moreover, cross-modality retrieval has been studied in [3, 7, 9]. Wang et al. [9] have adopted a classifier-based approach for visual-to-text retrieval. The category of the dish is first predicted then the recipe is searched under that category. However, the classifiers were trained from UPMC Food-101 dataset[1], so the retrieval is limited due to the dataset only having 101 food categories. Salvador et al.[7] learn recipe representation by encoding ingredients and cooking instructions using recurrent neural networks and align it with the dish image representations based on cosine similarity measure. In [2], a similar architecture is used but the attention mechanism is also applied to the ingredients. Apart from [2, 7], our work differs from the recipe side because we focus on ingredient images as the information in the recipe.

3 METHODOLOGY

In this section, we introduce our model which incorporates ingredient images into recipe retrieval. First, we formulate necessary mathematical notations. Given a query $Q = \{q_1, q_2, ..., q_n\}$, where q_i is an image of a ingredient, the goal is to compute a relevance score $s(Q, d)$ for all dish image d such that $s(Q, d_i) > s(Q, d_j)$ if a dish image d_i is more matching to the query Q than a dish image d_j. This is a common IR task, but different from general ad-hoc retrieval, queries and documents are all images. Thus, traditional retrieval models for text are not feasible. Due to the success of neural models in the field of computer vision, we compute $s(Q, d)$ by the neural retrieval architecture.

Our model architecture is shown in Figure 2. The whole model can be split into two part: dish image representation and ingredient image representation. Through them, query and document can be transferred into two numerical embeddings, then we can compute the relevance score between them. Next, we will introduce the detail of the proposed framework.

3.1 Image to Embedding

The input data include a single dish image and a set of ingredient images. Both of them need to be converted into a distributed vector for further operations. For the transformation, we adopt the state-of-the-art deep convolutional network, ResNet-50 [4], as the architecture. We preprocess the image by resizing and center cropping to unify the size of input images to 256x256. Then, we do random cropping on each training image to generalize our model, but just do one more center cropping for testing. After preprocessing, they will be unified to 224x224 images. After that, we normalize each channel of the input RGB image with mean and standard deviation. Finally, We utilize the Resnet-50 as a feature extractor and extract the last embedding before the final classification as the representation of each image. The dimension of 1D feature vectors is 2,048.

3.2 Task-oriented representation

Although ResNet-50 is a strong model for image classification, it is not designed and trained for matching ingredient and dish images. To exploit the extracted feature better, we transform them from

[1] To foster the research in this area, we release our dataset for research purposes. Link is removed due to double-blind review.

the feature space of ResNet-50 into the other task-oriented spaces respectively for computing relevance scores.

Specifically, we apply a fully-connected layer to the dish image feature vector, resulting in a final dish embedding with its dimension empirically set as 1,024. For the ingredient feature vectors, which have more than one ingredients in general, we shape the representation of ingredient set as a $N_{ingr} \times 2048$ matrix, where N_{ingr} is the maximum number of ingredients. For the matrix, we apply a convolution layer. The size of the convolution kernel is set to $(3, 2048)$. and the number of output channel is 1024. After convolution, we further append a max pooling layer and a fully-connected layer to transform the result into a 1-D 1024-dimension vector.

For implementation, we fix N_{ingr} as 10. If the recipe has more than 10 ingredients, we randomly sample 10 from the list. If the recipe has less than 10 ingredients, we use zero vector to represent no ingredient.

After getting ingredient embedding and dish embedding, we feed a concatenation of two embeddings into fully-connected layers to compute the score, as several neural query-document matching models did. In contrast to the cosine similarity adopted by [7], the range of value is not constrained, and the computing of the score is also optimized through training.

3.3 Optimization

For optimization, we choose pairwise loss as the training objective. For a data instance (Q, d_1, d_2), where d_1 is a matching recipe and d_2 is not, we aim to minimize pairwise hinge loss $max(0, (s(Q, d_1) - s(Q, d_2)) + 1)$. For optimizer, Adam is adopted to optimize the parameters. In the experiment, the learning rate is set to 0.0001. For regularization, We employ dropout for regularization. In particular, we apply dropout layers with dropout probability from $[0.0, 0.2, 0.5]$ on all fully-connected layers except those generating final embedding and score.

We describe our model based on the scenario of retrieving dish images for a visual query of ingredients. Regarding ingredient images as information in the recipe, it is analogous to **recipe2im**, which is a task of retrieving dish images for a query of recipe information, in the recipe retrieval application. By exchanging the role of ingredient image set and dish images, we can also optimize our model for **im2recipe**, the task of the opposite direction, but we focus on **recipe2im** in this work.

3.4 Working with Cross-Modal Retrieval

Ingredient image representation can be regarded as one modality of a recipe. Thus, it is able to co-work with the current state-of-the-art cross-modal model [7]. In addition to the framework in Figure 2, we also evaluate the cross-modal framework with ingredient images, dish images, and textual information. For implementation, we append the representation of the ingredient image set to the concatenation of ingredient list encoding and textual instruction encoding. Finally, the concatenation of three encodings will be transformed into one joint recipe representation using the architecture of [7].

Table 1: Performance of the proposed method and the baseline models, in terms of MedR, R@1, R@5, and R@10.

Model	MedR	R@1	R@5	R@10
Random	500	0.0010	0.005	0.0100
CCA-1	37.0	0.0700	0.2000	0.2900
CCA-2	24.8	0.0900	0.2400	0.3500
JNE	6.6	0.2011	0.4674	0.5000
JNE+SR	5.35	0.2304	0.5063	0.6365
img2img (Ours)	29.95	0.0681	0.2031	0.3030
img2img+JNE+SR (Ours)	**5.1**	**0.2388**	**0.5131**	**0.6412**

4 EXPERIMENTS

In this section, we first introduce our dataset and the baselines. We further review our experimental setup. We finally report and discuss the obtained results.

4.1 Dataset

The dataset we use is Recipe1M [7], which consists of 1,029,720 structured cooking recipes and 887,536 associated images collected from popular cooking websites like "Allrecipes" and "Fine Cooking". The average number of ingredients per receipt is 9.3. The recipes without images will be ignored in the experiment and 33% of the dataset contains at least one or more images. Because the original dataset does not include the image of ingredients, we make use of Google image search to collect our data. For each ingredient, we use the name in the ingredient list of Recipe1M as the query for Google image search, and crawl top-5 results returned by Google. When we need to use a image for one ingredient, we randomly sample one from these 5 images as the representative. We don't filter data by their quality or relevance. Although it may include some bad data, it can be regarded as noise in training data.

4.2 Baselines

We compare our approach with several baselines and all of compared methods consider ingredient list and instruction as a recipe for **recipe2im** task. Canonical correlation analysis (**CCA**) learns two linear projections for mapping text and image features to a common space that maximizes their feature correlation. We directly refer the statistics of **CCA** with pretrained text embedding (**CCA-1**) and embedding trained on the recipe corpus (**CCA-2**) in [7]. Joint neural embedding (**JNE**) [7] is a cross-modal learning framework to align dish image representation and joint recipe representation which ingredient list and instruction are encoded in. **JNE+SR** [7] is **JNE** with semantic regularization. By adding a food category classifier for two kinds of representation, the model is penalized if it fails to categorize. For the above two, we use the authors' Pytorch code[2] to retrain the model with the default setting. By comparing these baselines, We have two proposed methods to be evaluated. **Img2Img** is the model with the architecture in Figure 2, and **Img+JNE+SR** is JNE+SR with our ingredient image representation as one modality of recipe.

[2] https://github.com/torralba-lab/im2recipe-Pytorch

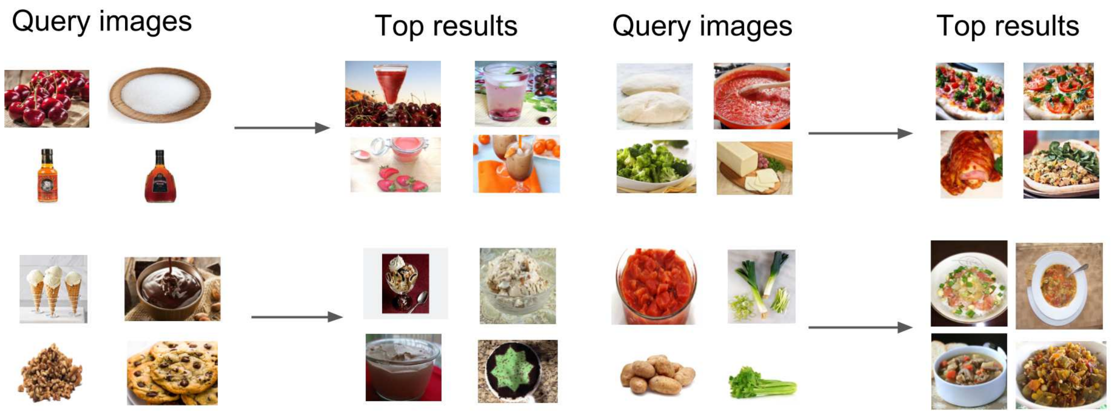

Figure 3: Case Study

4.3 Evaluation

As mentioned, we focus on the case regarding ingredient images as queries, so we limit the scope of evaluation to **recipe2im** task. Median retrieval rank (MedR) and recall at top K (R@K) are adopted for performance evaluation. MedR is the median rank position where the right recipe is returned. Thus, lower Median Rank score indicates higher performance. Because each recipe only has one matching dish image and vice versa, recall@K can be regarded as the fraction of times that a correct recipe is found within the top-K retrieved candidates. As previous work [2, 7] did, the proposed architecture would be evaluated on 1000 randomly sampled recipes from the test set.

5 RESULTS AND DISCUSSION

The result is shown in Table 1. **img+JNE+SR** achieves the best performance among all methods. Compared to **JNE+SR**, expanding recipes by ingredient images can improve the quality of dish image retrieval. It shows that ingredient images successfully include helpful information not in the textual recipes for **recipe2im** tasks. For **img2img**, it is not as good as other baselines, though it has comparable performance with **CCA-1**. It is not surprising because the textual instruction, which **img2img** ignores, contain a lot of information about dishes. According to the ablation studies in [2], instruction is the most powerful part as a query for **recipe2im** tasks. Besides, comparable performance with **CCA-1** also shows the potential of the image-to-image process in recipe retrieval.

To understand what information our model learns, we pick some successful testing cases **img2img** performs well on, and show the top results in Figure 3 (these cases successfully include the matching recipe in top 4 dishes). From the four examples in the figure, we can see that our model can capture the visual property of input ingredients and retrieve matching and similar dish images for these cases. For example, the top right case shows that the model understands pizza dough, tomato sauce, broccoli and cheese from the images and returns good results. Although its overall performance does not reach the state-of-the-art level, it can be observed that visual information is successfully analyzed by the model for positive samples.

6 CONCLUSION AND FUTURE WORK

In this paper, we studied the novel task of image-to-image recipe retrieval. We expanded the Recipe1M dataset by collecting ingredient images according to its list. To exploit the visual information of ingredient images, we proposed a CNN-based framework to encode multiple images into a high-dimensional representation, and build an image-to-image retrieval process to find matching dishes. Furthermore, we combined our representation into a state-of-the-art cross-modal framework which led to significant improvements. For future work, we plan to expand our work to other cross-modal text-to-image retrieval tasks. In the domain of recipe retrieval, we intend to incorporate images for cooking style and plating methods into the model, since they may provide good anchor points for findings dishes with similar style. Finally, the photos directly taken by users are different from those returned by the search engine. We plan to perform user studies to collect a more realistic test collection and study the in-situ impact of the proposed solutions on user experience.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *ECCV '14*. Springer, 446–461.
[2] Jingjing Chen, Chong-Wah Ngo, Fuli Feng, and Tat-Seng Chua. 2018. Deep Understanding of Cooking Procedure for Cross-modal Recipe Retrieval. In *MM '18*. 1020–1028.
[3] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *MM '16*. ACM, 32–41.
[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015).

[5] Takuma Maruyama, Yoshiyuki Kawano, and Keiji Yanai. 2012. Real-time mobile recipe recommendation system using food ingredient recognition. In *IMMPD '12*. ACM, 27–34.

[6] Hiroki Matsunaga, Keisuke Doman, Takatsugu Hirayama, Ichiro Ide, Daisuke Deguchi, and Hiroshi Murase. 2015. Tastes and textures estimation of foods based on the analysis of its ingredients list and image. In *ICIAP '15*. Springer, 326–333.

[7] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marín, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. In *CVPR '17*. 3068–3076.

[8] Han Su, Ting-Wei Lin, Cheng-Te Li, Man-Kwan Shan, and Janet Chang. 2014. Automatic recipe cuisine classification by ingredients. In *UbiComp '14 Adjunct*. ACM, 565–570.

[9] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. 2015. Recipe recognition with large multimodal food dataset. In *ICMEW '15*. IEEE, 1–6.

[10] Haoran Xie, Lijuan Yu, and Qing Li. 2010. A hybrid semantic item model for recipe search by example. In *IEEE International Symposium on Multimedia*. IEEE, 254–259.

[11] Yoko Yamakata, Shinji Imahori, Hirokuni Maeta, and Shinsuke Mori. 2016. A method for extracting major workflow composed of ingredients, tools, and actions from cooking procedural text. In *ICMEW '16*. IEEE, 1–6.