10-2018

Deep understanding of cooking procedure for cross-modal recipe retrieval

Jingjing CHEN
City University of Hong Kong

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

Fu-Li FENG
National University of Singapore

Tat-Seng CHUA
National University of Singapore

Deep Understanding of Cooking Procedure for Cross-modal Recipe Retrieval

Jing-Jing Chen[1], Chong-Wah Ngo[1], Fu-Li Feng[2], Tat-Seng Chua[2]
[1]City University of Hong Kong, Hong Kong [2]National University of Singapore, Singapore
jingjchen9-c@my.cityu.edu.hk, cscwngo@cityu.edu.hk,
fulifeng93@gmail.com, chuats@comp.nus.edu.sg

ABSTRACT

Finding a right recipe that describes the cooking procedure for a dish from just one picture is inherently a difficult problem. Food preparation undergoes a complex process involving raw ingredients, utensils, cutting and cooking operations. This process gives clues to the multimedia presentation of a dish (e.g., taste, colour, shape). However, the description of the process is implicit, implying only the *cause* of dish presentation rather than the visual *effect* that can be vividly observed on a picture. Therefore, different from other cross-modal retrieval problems in the literature, recipe search requires the understanding of textually described procedure to predict its possible consequence on visual appearance. In this paper, we approach this problem from the perspective of attention modeling. Specifically, we model the attention of words and sentences in a recipe and align them with its image feature such that both text and visual features share high similarity in multi-dimensional space. Through a large food dataset, Recipe1M, we empirically demonstrate that understanding the cooking procedure can lead to improvement in a large margin compared to the existing methods which mostly consider only ingredient information. Furthermore, with attention modeling, we show that language-specific named-entity extraction becomes optional. The result gives light to the feasibility of performing cross-lingual cross-modal recipe retrieval with off-the-shelf machine translation engines.

KEYWORDS

Recipe retrieval; Cross-modal learning; Hierarchical attention

ACM Reference Format:
Jing-Jing Chen, Chong-Wah Ngo, Fu-Li Feng, Tat-Seng Chua. 2018. Deep Understanding of Cooking Procedure for Cross-modal Recipe Retrieval. In Proceedings of MM'18. ACM, NewYork, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240627

1 INTRODUCTION

Food intake tracking has recently captured numerous research attentions [1] [17] [28] for long-term impact of food consumption on health. The main pipeline of tracking is to take a picture of the dish, recognize its category and then search for relevant sources for nutrition and calories estimation [2] [27]. The sources are usually food labels and food composition tables (FCT) compiled by nutrition experts [23]. Nevertheless, in the free-living environment, dishes are often prepared in wild with no expert references for health index estimation. As ingredient recognition remains limited in scale [7], automatic enumeration of nutrition contents from ingredient composition inferred from food images is still far beyond the current technology.

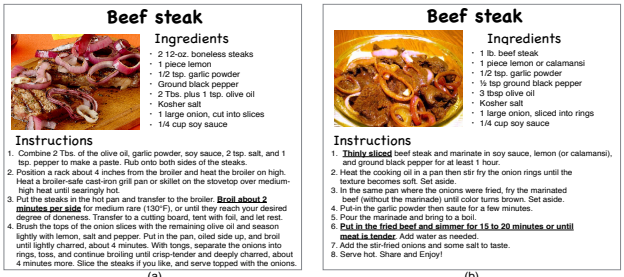

Figure 1: Understanding recipes is not easy even by the human. Both dishes have the same name and similar ingredients, but are prepared in different manners and result in different presentations. The differences (e.g., broil versus simmer) are underlined to highlight the cause-and-effect in cooking procedure.

The prevalence of sharing food images and recipes on the Internet [35], nevertheless, provides a new look to this problem. Specifically, there are social media platforms in both eastern and western countries, such as "Go Cooking" [1] and "All Recipes" [2], for master and amateur chefs to share their newly created recipes and food images. There are also followers or fans who follow the cooking instructions to reproduce the same dishes and upload their pictures to websites for peer comment. To date, these websites have accumulated over millions of recipes and images. These recipes are mostly listed with ingredients alongside with their quantities, supplying a new source of references for food intake tracking. Furthermore, cooking procedure, i.e., how ingredients are prepared and cooked (e.g., deep fried versus steam), provides another dimension of clues which is not listed in food label or FCT for health management. Hence, in principle, being able to link a food image to its right recipe available on the Internet will facilitate the evaluation of nutrition contents. Conversely, linking a recipe to its potential dish appearance can also encourage cooking at home.

1.1 Challenges

Image-to-recipe retrieval is essentially a cross-modal learning problem [10], which maps features of different modalities into the same

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

MM'18, , October 22–26, 2018, Seoul, Republic of Korea.

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240627

form for similarity assessment. A recipe uses to have three sections: title, ingredient, cooking procedure (see Figure 1). Title resembles phrase while ingredients can be regarded as keywords analog to traditional visual annotation problem [33], which explicitly list out the content of food image. Cooking instruction, on the other hand, is composed of a series of sentences detailing the food preparation and cooking process. Different from problems such as image captioning [36] and visual question-answering [3], the descriptions are not directly translatable to image content. Rather, the instruction at a step dictates the causality of food preparation which may not be relevant to final food presentation or even be visible in food image. For example, the instructions "position rack about 4 inches from the boiler" in Figure 1a and "put in the garlic powder then saute for a few minutes" in Figure 1b have insignificant outcome to the visual appearance of dishes. Furthermore, online recipes are user-generated and there are no rules governing the documentation of recipes. Sentences such as "Serve hot! Shared and enjoy!" (Figure 1b) are visually irrelevant, "slice the steaks if you like" (Figure 1a) presents visual uncertainty.

The purpose and format of recipe make the challenges of cross-modal retrieval different from other problem domains [3] [33] [36] in multimedia. As shown in Figure 1, both dishes have the same title and almost similar list of ingredients. However, the dish presentations exhibit different visual appearances beyond photometric changes due to differences in cooking processes. Precisely, the steak in Figure 1a is broiled while the steak in Figure 1b is fried and simmered. In addition, some ingredients are used in different stages for different purposes. For example, lemon in 1a is seasoned on onion slice, and lemon in Figure 1b is mixed with other sauces to marinate beef steak. These procedural descriptions do not directly link to visual content but have an implicit impact on final food presentation. Furthermore, the relationship of cooking and cutting actions to the visual content of food is not always one-to-one, but intertwines with types of ingredients and seasonings being added.

1.2 Contribution

The main contribution of this paper is encoding of a recipe into a vector for capturing cooking procedure that implies causality effect between ingredients and actions. Online recipes are written in free form with user-generated text and are difficult to be syntactically or semantically analyzed. Instead of modeling recipe as an action graph illustrating the flow of food preparation [20] [41], embedding recipe as a vector that captures word and sentence significances is more feasible with the rapid advancement of deep learning. In this paper, we propose a hierarchical attention mechanism based on [44] to model the complex word-to-word and sentence-to-sentence interactions in the recipe as a vector. The resulting vector representation is embedded in a form similar to the visual vector, allowing parameter tuning and data-driven search of weights to align the relevancy of words or sentence to visual content.

In the literature, there are only few approaches studying cross-modal retrieval in this specialized domain [7] [8] [10]. Most approaches extract partial information from recipes, mostly ingredients as text modality, to either feed into deep neural networks for learning cross-modal similarity [8] or match with the results of visual categorization [7]. Recently, some approaches also explore cooking instruction for retrieval, either by manual extraction of food attributes (cooking and cutting attributes) for classification [10] or auto encoding of instructions by two-stage long short-term memory units (LSTM) for embedding learning [30]. These approaches treat every word and sentence in a recipe equally when modeling joint visual-text relationship, overlooking the fact that some of these descriptions are not visually observable but rather as an implication of cause-and-effect consequence.

The novelty of our work originates from leveraging of attention mechanism to address the cause-and-effect consequence in the procedural description. Despite technically straightforward, this is the first attempt in the literature that investigates the extent which attention can deal with the causality effect while being able to demonstrate impressive performance on cross-modal recipe retrieval. In addition, we provide a unified way of dealing with three sections of information (i.e., title, ingredient, instruction) in the recipe. The work presented in this paper is more intuitive than [10] [30] in terms of problem formulation, and more generalized than [8] [30] in terms of level of information being considered in cross-modal retrieval.

2 RELATED WORK

Cross-modal learning is an active topic in multimedia. Classic examples include the employment of canonical correlation analysis (CCA) for semantic visual annotation [29]. Recent approaches mostly rely on deep learning, for examples, deep CCA [42], DeViSE [13], correspondence auto-encoder [12] and adversarial cross-modal retrieval [37]. These models are learnt from training examples that assume a direct correspondence between visual and text relationship, and cannot be straightforwardly extended for recipe processing and procedure-based cross-modal learning. In this section, we focus on presenting the existing research efforts in the food domain.

The current works are mostly devoted to food classification, specifically, to recognize food categories given pictures [24] [43] [11]. These works rely heavily on off-the-shelf deep models [32] [18] and have recently triggered construction of large datasets such as Cookpad [14] and Recipe1M [30]. These datasets differ in terms of cuisine, geography region and language. The state-of-the-art results for food recognition on medium size datasets, such as Food101 [6], FoodCam-256 [19], VireoFood-172 [7], can be higher than 80% of top-1 accuracy [16]. Several photo-based food logging mobile apps have also been developed, including DietLens [27], FoodLog [2] and Im2Calories [25]. The applicability of these efforts, nevertheless, is difficult to scale up for real scenario in dietary tracking due to limited number of food categories, mostly few hundreds as in DietLens [27], that can be recognized.

Contextualized retrieval-based methodology approaches the problem from a different view by searching for similar images with geographic constraint to infer food categories [5] [40]. However, food images are wildly diverse in terms of ingredient composition, visual appearance and ambiguity. Pure retrieval-based approaches based on visual features can hardly reach the performance of learning-based approaches. Recipe represents a rich source that supplements the visual content of food images. For example, in [7], ingredients

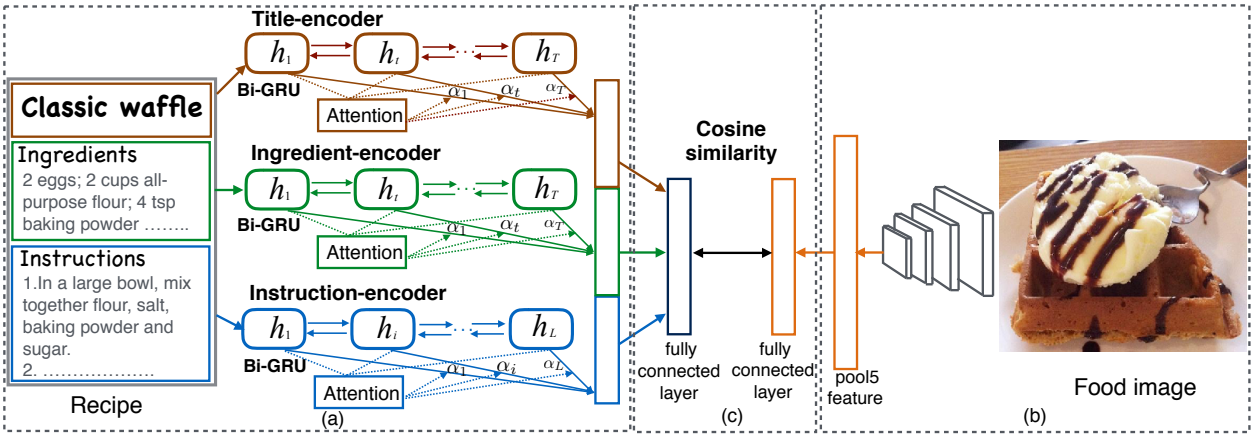

Figure 2: Framework overview: (a) recipe representation learning; (b) image feature learning; (c) joint-embedding space learning.

are recognized for recipe search and then the result is fused with image search to resolve visually ambiguity in food categorization.

The most related approach to our proposed work is [30]. Similar in spirit, [30] also learns recipe representation by encoding ingredients and cooking instructions using recurrent neural networks. Our work differs from [30] for incorporation of word-level and sentence-level attentions at three different levels of granularity (i.e., title, ingredient, instruction) for representation learning. The recent work in [10] extracts rich food attributes, including cutting and cooking attributes, from cooking instructions of recipes. The attributes are utilized for learning a multi-task convolutional network that is eventually applied for food annotation and recipe search. However, the attributes are manually extracted from recipes and then labeled by homemakers, which are both labor intensive and cost expensive. Apart from [10], to the best of our knowledge, there are no efforts yet studying the modeling of cooking instructions for cross-modal retrieval. Analysis of cooking procedure is investigated in other applications, for example knowledge representation [20] [31], recipe recommendation [34] and multimedia search [38], but not in the context of cross-modal learning. Very often the workflows of cooking are manually or semi-automatically [31] [38] created from recipes to serve these applications.

Cross-modal analysis in food domain is also studied by [8]. In [8], a stacked attention network is applied to simultaneously locate ingredient regions in the image and learn multi-modal embedding features. However, this approach considers only ingredient list and the attention mechanism is applied to weight image regions and not ingredients. Furthermore, the network is not appropriate for retrieval because the projections from visual and text to embedding space are coupled. In [7], an ingredient network is constructed from more than 65,000 recipes and modeled with conditional random field for cross-modal search. However, due to only 353 ingredients can be visually recognized, the effect of the network in boosting recipe search is limited. While most approaches including [7] [8] employ discriminative learning, the work in [26] studies generative learning using deep belief network for cuisine classification, food image retrieval and attribute inference. Different from our work, these approaches mainly model ingredients and auxiliary information (e.g., cuisine) in learning and ignore cooking procedure [7] [8] [26]. Inherently, they are incapable of disambiguating dishes that are different but using the same or similar ingredients (e.g., Figure 1).

3 METHODOLOGY

Figure 2 depicts the basic framework of our proposed attention mechanism. First, different modalities are input to both ends of the deep model for representation learning. Recipes, in particular, are split into three sections (title, ingredient, instruction) based on different levels of information granularities. These sections are encoded separately by attention mechanism into three representations, which are eventually concatenated as a recipe representation (2a). Together with image representation which is learnt through the convolutional network (2b), the proposed model learns to maximize the cosine similarity between textual recipes and their associated food images. The similarity learning is carried out through two representation transformations that aim to make recipe and image features as alike as possible (2c).

3.1 Recipe representation

Title encoder. Each recipe has a title as the name of the dish. As expected, the title uses to elicit dish peculiarity by capturing food uniqueness directly into the name. The characterization of food uniqueness is multi-perspective in nature, ranging from the taste, style (e.g., "old fashion", "home-made"), cuisine and geography region, ingredient and cooking method, to even cooking utensil. Examples include "peek potato and bacon casserole recipe", "caramelized beef skewers" and "home-made healthy granola bars". For title representation, the aim of the attention model is to assign higher weights to words that directly link to food content relative to contextually relevant terms about style and location.

Given a title with words $\boldsymbol{w}_t, t \in [0, T]$, we first embed individual word to a vector through a matrix $\boldsymbol{W}_e$, $\boldsymbol{x}_t = \boldsymbol{W}_e \boldsymbol{w}_t$. The title is treated as a sequence and a bidirectional gated recurrent unit (GRU) [4] is employed to encode the word sequence. The bidirectional GRU is composed of a forward $\overrightarrow{GRU}$ which reads title from $\boldsymbol{w}_1$ to $\boldsymbol{w}_T$ and a backward $\overleftarrow{GRU}$ which reads from $\boldsymbol{w}_T$ to $\boldsymbol{w}_1$, defined as

$$\boldsymbol{x}_t = \boldsymbol{W}_e \boldsymbol{w}_t, t \in [1, T], \quad (1)$$

$$\overrightarrow{\boldsymbol{h}_t} = \overrightarrow{GRU}(\boldsymbol{x}_t), t \in [1, T], \quad (2)$$

$$\overleftarrow{\boldsymbol{h}_t} = \overleftarrow{GRU}(\boldsymbol{x}_t), t \in [1, T]. \quad (3)$$

The representation of a word $\boldsymbol{w}_t$ can by obtained by concatenating the forward hidden state $\overrightarrow{\boldsymbol{h}_t}$ and backward hidden state $\overleftarrow{\boldsymbol{h}_t}$ as following

$$\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}]. \tag{4}$$

The attention mechanism further transforms word representation from $\boldsymbol{h}_t$ to $\boldsymbol{u}_t$ with a one-layer multi-layer perceptron (MLP). The contribution of a word is then rated by a weight α_t evaluated using softmax. Mathematically, we have

$$\boldsymbol{u}_t = tanh(\boldsymbol{W}_w \boldsymbol{h}_t + \boldsymbol{b}_w), \tag{5}$$

$$\alpha_t = \frac{exp(\boldsymbol{u}_t^\top \boldsymbol{u}_w)}{\sum_t (exp(\boldsymbol{u}_t^\top \boldsymbol{u}_w)}, \tag{6}$$

where $\boldsymbol{W}_w$ is the transformation matrix of MLP and $\boldsymbol{b}_w$ is its bias term. The weight α_t characterizes the similarity of word representation $\boldsymbol{u}_t$ and context vector $\boldsymbol{u}_w$ under softmax function. The context vector can be regarded as a reference object of $\boldsymbol{u}_t$ for cross-modal learning. For example, in attention-based visual question answering (VQA) [39], the context vector can be directly set as text features to calculate the attention weights on image regions. In our case, nevertheless, we do not wish to couple text and image features at this stage because otherwise the learnt features will have to be on-the-fly generated and cannot be indexed offline for retrieval. Instead, the context vector $\boldsymbol{u}_w$ is randomly initialized and updated subsequently during the learning process. Finally, the title representation $\boldsymbol{f}_{\text{title}}$ is generated by aggregation of weighted word representations as following

$$\boldsymbol{f}_{\text{title}} = \sum_t \alpha_t \boldsymbol{h}_t. \tag{7}$$

Ingredient encoder. A recipe usually has a section listing out ingredients, their quantities and optionally the corresponding cooking and cutting methods for food preparation. The ingredients include both visible items on the dish (e.g., onion, steak) and non-visible items (e.g., oil, salt). The aim of attention is to align the observations on recipe and food image such that ingredients, which are not visible or do no alter the outlook of a dish, will be assigned lower weights. The learning of ingredient representation, $\boldsymbol{f}_{\text{ingredient}}$, is similar to that of title representation. We first obtain the hidden representation of each ingredient (equations 1 to 4), and followed by quantifying the significance of an ingredient with a numerical weight (equations 5 to 6). The final representation is generated by weighted aggregation as in Equation 7.

Instruction encoder. Cooking instructions are composed of varying-length sentences written in free form. The descriptions are much denser than title and ingredient list for elaborating cooking steps in details. While rich in information, there might not be a direct correspondence between a sentence in cooking instruction and dish appearance. For example, the instruction "heat a 10-inch skillet over medium-high heat" has less effect than "lay two slices of bacon over the top" in the final food appearance. The importance should also not be directly impacted by sentence length. For example, the short sentence "bake for 1 hour" could change the dish outlook and should be assigned a higher weight. To this end, the attention mechanism aims to evaluate the relevancy between a sentence and food presentation, and meanwhile, the relevancy is also characterized by the importance of words in the sentence. This basically establishes a two-level hierarchy similar to [44] that propagates the contributions of words to sentence level and then sentences to dish appearance for forming recipe representation.

The same procedure as title and ingredient is adopted for word-level representation learning (equations 1 to 7) to generate sentence vectors, denoted as $\boldsymbol{s}_i, i \in [1, L]$, where L is the number of sentences in the instruction. The sentence-level representations are further aggregated into a vector using a similar procedure. Precisely, the bi-directional forward and backward GRUs followed by one-layer MLP are used to generate hidden representation $\boldsymbol{u}_i$ of $\boldsymbol{s}_i$ as following

$$\overrightarrow{\boldsymbol{h}_i} = \overrightarrow{GRU}(\boldsymbol{s}_i), i \in [1, L], \tag{8}$$

$$\overleftarrow{\boldsymbol{h}_i} = \overleftarrow{GRU}(\boldsymbol{s}_i), i \in [1, L], \tag{9}$$

$$\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h}_i}, \overleftarrow{\boldsymbol{h}_i}], \tag{10}$$

$$\boldsymbol{u}_i = tanh(\boldsymbol{W}_s \boldsymbol{h}_i + \boldsymbol{b}_s). \tag{11}$$

Denote $\boldsymbol{u}_s$ as the sentence-level context vector, the relevancy of a sentence is calculated as

$$\alpha_i = \frac{exp(\boldsymbol{u}_i^\top \boldsymbol{u}_s)}{\sum_i (exp(\boldsymbol{u}_i^\top \boldsymbol{u}_s)}, \tag{12}$$

where $\boldsymbol{W}_s$ is transformation matrix of MLP. $\boldsymbol{u}_s$ is the context vector. Similar to Equation 6, $\boldsymbol{u}_s$ is randomly initialized and progressively refined during training. The final representation is obtained through

$$\boldsymbol{f}_{\text{instruction}} = \sum_i \alpha_i \boldsymbol{h}_i. \tag{13}$$

Recipe representation. We adopt early fusion strategy to append the three levels of representations as following

$$\boldsymbol{f}_{\text{recipe}} = [\boldsymbol{f}_{\text{title}}, \boldsymbol{f}_{\text{ingredient}}, \boldsymbol{f}_{\text{instruction}}] \tag{14}$$

The dimensions of both $\boldsymbol{f}_{\text{title}}$ and $\boldsymbol{f}_{\text{ingredient}}$ are empirically set as 600. As instruction is dense in description, $\boldsymbol{f}_{\text{instruction}}$ is set as a 1,000 dimensional vector. No normalization is applied when concatenating the three vectors into recipe presentation.

3.2 Representation of images

The state-of-the-art deep convolutional network, ResNet-50 [18], is used for image feature extraction. As the network is not pre-trained on food images, we fine-tune ResNet-50 with UMPC Food-101 dataset [6], which contains 75,750 training images of 101 food categories. Different from [30], we do not integrate ResNet-50 with recipe representation for end-to-end feature learning. Instead, pool-5 features are extracted. The dimension of $\boldsymbol{f}_{\text{image}}$ is 2,048.

3.3 Joint embedding learning

The aim is to transform both recipe and image representations into vectors with an equal number of dimensions for similarity comparison. Two projections are learnt through transformation matrices $\boldsymbol{W}_R$ and $\boldsymbol{W}_v$, as following

$$\boldsymbol{\phi}_R = tanh(\boldsymbol{W}_R \boldsymbol{f}_{\text{recipe}} + \boldsymbol{b}_R), \tag{15}$$

$$\boldsymbol{\phi}_v = tanh(\boldsymbol{W}_v \boldsymbol{f}_{\text{image}} + \boldsymbol{b}_v), \tag{16}$$

where $\boldsymbol{\phi}_R$ and $\boldsymbol{\phi}_v$ are respectively the embedding features of recipe and image, and $\boldsymbol{b}_R$ and $\boldsymbol{b}_v$ are their bias terms. The feature dimension is empirically set as 1,024, which is the same with [30]. With this, cosine similarity is employed to evaluate the closeness between two transformed features. The learning goal is to ensure that a query can always score its true positive as higher as possible than negatives and thus the rank loss function with max margin is employed for the update of parameters. Since we target for both image-to-recipe and recipe-to-image retrieval, the input of loss function is composed of two triplets: $\langle \boldsymbol{\phi}_v, \boldsymbol{\phi}_R, \boldsymbol{\phi}_{R^-} \rangle$ and $\langle \boldsymbol{\phi}_R, \boldsymbol{\phi}_v, \boldsymbol{\phi}_{v^-} \rangle$. The first element of the triplet is either an image ($\boldsymbol{\phi}_v$) or recipe ($\boldsymbol{\phi}_R$) query, followed by a true positive and a negative example of a different modality as the second and third elements. Let the margin as $\delta \in (0, 1)$, the loss function is defined as

$$\begin{aligned} L =& max(0, \delta - cos(\boldsymbol{\phi}_v, \boldsymbol{\phi}_R) + cos(\boldsymbol{\phi}_v, \boldsymbol{\phi}_{R^-})) \\ &+ max(0, \delta - cos(\boldsymbol{\phi}_R, \boldsymbol{\phi}_v) + cos(\boldsymbol{\phi}_R, \boldsymbol{\phi}_{v^-})). \end{aligned} \tag{17}$$

Note that, in addition to the attention mechanism, the technical difference between this work and [30] are in four aspects. First, we do not adopt end-to-end image feature learning as in [30] for saving GPU memory and training time. Second, rank loss is employed. In our empirical study, rank loss is about three times faster in model convergence than the pairwise cosine similarity loss adopted by [30]. The number of epochs required by rank loss is 70, versus 220 epochs as required by cosine similarity loss for training. Third, [30] does not encode title information but instead utilizes titles as the constraint for regularization (see Section 4.5 for details). Finally, skip-thoughts [22] and LSTM are used in [30] to encode cooking instruction without attention modeling. In this work, we use GRU instead of LSTM as encoder because GRU is computationally more efficient than LSTM.

4 EXPERIMENT

4.1 Dataset

The experiments are conducted on Recipe1M[3], which is one of the largest datasets that contain recipes and images. The dataset is compiled from dozens of popular cooking websites such as "allrecipes" [4] and "fine cooking" [5]. We use the preprocessed version of the dataset provided by [30], in which 0.4% duplicate recipes and 2% duplicate images have been removed, for empirical studies. The dataset contains 1,029,720 recipes and 887,536 images, with around 70% of data being labeled as training and the remaining being split equally between validation and testing. The average number of ingredients and instructions per recipe are 9.3 and 10.5 respectively. All recipes are written in English and 33% of them are associated with at least one image. We treat a recipe and its associated image as a pair, and generate at most five pairs for recipes having more than one images. We do not use those recipes without images in our experiments.

[3]http://im2recipe.csail.mit.edu/dataset/
[4]https://www.allrecipes.com
[5]http://www.finecooking.com

Table 1: Contributions of different encoders and their combinations on 5K dataset.

	im2recipe				recipe2im			
	MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10
title	58.2	0.044	0.141	0.217	57.6	0.040	0.137	0.215
ingre.	71.0	0.045	0.135	0.202	70.1	0.042	0.133	0.202
inst.	33.9	0.070	0.202	0.294	33.2	0.066	0.201	0.295
title + ingre.	31.9	0.073	0.215	0.310	31.9	0.074	0.211	0.307
title + inst.	26.6	0.082	0.231	0.331	26.8	0.081	0.234	0.334
ingre. + inst.	30.0	0.079	0.223	0.316	29.0	0.075	0.220	0.316
all	20.0	0.104	0.274	0.382	19.1	0.101	0.272	0.382

4.2 Experiment setting

Implementation details. Adam optimizer [21] is employed for model training with learning rate set as 10^{-4}. The margin in Equation 17 is selected as 0.3 by validation and the mini batch size is set as 128. Per-batch online triplet sampling is employed during training. In each mini-batch, a recipe (image) is restricted to have exactly one ground-truth image (recipe). Furthermore, for each recipe (image), apart from its ground-truth image (recipe), the remaining images (recipes) are used as negatives for model training. The deep model is implemented on tensorflow platform. As end-to-end learning is only performed between recipe feature and joint embedding learning, the model can be trained on a single NVIDIA Tesla K40 GPU.

Evaluation metrics. We use median retrieval rank (MedR) and recall at top K (R@K) as in [30] for performance evaluation. MedR measures the median rank position among where true positives are returned. Therefore, a lower MedR score indicates higher performance. R@K, on the other hand, calculates the fraction of times that a correct recipe is found within the top-K retrieved candidates. Different from MedR, the performance is directly proportional to the score of R@K.

Testing. Same as [30], we report results for subsets of randomly selected recipe-image pairs from the test set. In a subset, every pair is issued alternately as image or recipe query to retrieve its counterpart, namely the image-to-recipe (im2recipe) or recipe-to-image (recipe2im) retrieval. To evaluate the scalability of retrieval, the subset sizes are respectively set to be 1K, 5K and 10K pairs. The experiments are repeated 10 times for each size of the subset and the mean results are reported.

4.3 Ablation studies

Table 1 lists the contributions of title, ingredient, instruction and their combinations towards performance improvement. On both im2recipe and recipe2im, instruction attains higher performance than title and ingredient alone in a large margin. The result clearly verifies the significance of cooking instructions, which embed processing of ingredients with rich procedural actions, in cross-modal retrieval. The title, which is often highlighted with the key ingredient and major cooking method, surprisingly outperforms ingredient. Title and ingredient, nevertheless, appear to be highly complementary, and the combination of them leads to improvement close to the performance of instruction alone. Meanwhile, combining instruction with either title or ingredient also results in improvement, and the best performance is achieved by concatenating all the three representations.

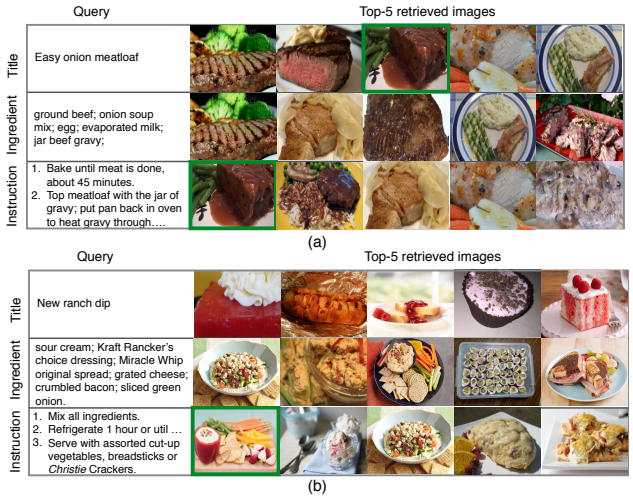

Figure 3: Retrieval results by title, ingredient or instruction. True positives are bounded in green box. The highly weighted sentences are listed in the instruction section.

Figure 3 shows two examples explaining the role of instruction in boosting performance. In Figure 3a, the title alone already ranks the true positive at the top-3 position. Instruction gives high weights to two sentences "bake until meat is done" and "top meatloaf with jar of gravy." As these sentences somewhat describe the interaction between the ingredients and the associated actions (e.g., bake, top), the true positive is ranked at the top-1 position. The ingredient, which misses the keyword "meatloaf", only manages to retrieve dishes with beef. The title "new ranch dip" in Figure 3b does not visually describe the content of dish and hence fails to retrieve any sensible images. Instruction encoder, by giving high weights to "refrigerate 1 hour" and "serve with assorted cut-up vegetables", is able to rank true positive at the top-1 position. Interestingly, most of the ingredients appear in the ingredient list are not mentioned in the cooking procedure. Instead, they are described by the sentence "mix all ingredients" which is ranked as the third highest sentence. Browsing the images retrieved by instruction in Figure 3b, most top-ranked images are with the effect of mixing ingredients and being refrigerated.

4.4 Effect of attention

We experiment the impact of attention modeling on cross-modal retrieval. Table 2 contrasts the performances on 5K datasets. Note that the results without attention are obtained by average sum of words and sentences. As seen in Table 2, attention modeling exhibits consistent improvement across different evaluation metrics and levels of comparison. MedR, for example, is averagely upgraded by two ranks for both image-to-recipe and recipe-to-image retrieval. Similar performance is also noted on the 1K dataset with MedR being boosted by one position.

Figure 4 shows two examples of image-to-recipe retrieval. In the first example, although the word "kalops" in the title is assigned lower weight, the true positive is still ranked at the top-1 position by attention modeling. This is mainly because sentences 4-7 in the cooking instruction, which characterize the unique way of cooking kalops, are assigned higher weights. Especially, the effects of the operations such as "simmer", "boil water" and "add bay leaves" are partially visible on the dish. Without attention modeling, "pressured cooked beef" will be ranked at the top instead. However, when the attention weights are not assigned properly, the result could be worse than without attention modeling as shown in the second example. The keyword "frozen", which characterizes the uniqueness of "mousse square", is not attended in both the title and cooking instruction. Instead, the sentence "remove dessert from freezer" is assigned the highest weight. In this case, although the top-5 retrieved images are all chocolate cakes, the true positive is not ranked at the top compared to the method without attention modeling.

4.5 Performance comparison

We compare our approach with canonical correlation analysis (CCA) [15], stacked attention network (SAN) [8], joint neural embedding (JNE) [30], and JNE with semantic regularization (JNE+SR) [30]. We do not compare to classification-based approaches such as [7] [10] because only a limited number of ingredients, cutting and cooking attributes can be recognized. CCA learns two linear projections for mapping text and image features to a common space that maximizes their feature correlation. The text feature is concatenated from word2vec ingredient vector and skip-thoughts instructor vector provided by [30]. SAN considers ingredient list only and learns the embedding space between ingredient and image features through a two-layer deep attention mechanism. JNE utilizes both ingredients and cooking instructions in joint space learning, but different from our approach, the attention mechanism and title encoder are not considered. JNE+SR is a variant of JNE by imposing a regularization term such that the learnt embedded features will be penalized if failing in performing food categorization. The number of food categories being exploited for SR is 1,047. The categories are semi-automatically compiled from Food-101 dataset [6] and the text mining result on recipe titles of Recipe1M dataset. As the categories are mostly mined from frequent bigrams of titles, we consider that JNE+SR also exploits titles, ingredients and instructions as our approach, except that titles are leveraged in a different stage of learning. We name our approach as attention and also implement attention+SR as a variant based on the 1,047 food categories shared by [30]. Besides, as JNE uses LSTM as encoders, we also implement our attention model with LSTM for comparison. Finally, note that different image features are used in these approaches: VGG pool-5 features [32] in SAN, ResNet-50 features [18] fine-tuned by Food-101 dataset, and RestNet-50 features fine-tuned by ImageNet ILSVRC 1000 dataset in JNE.

Table 3 lists the detailed performances. Note that we only compare CCA and SAN on 1K dataset. SAN is computationally slow and is not scalable to large dataset. In addition, SAN is designed for image-to-recipe retrieval only. As seen in the results, attention and JNE consistently outperform CCA and SAN across all evaluation metrics on 1K dataset for both im2recipe and recipe2im retrieval. SAN, although adopts attention mechanism, performs considerably worse. This is because SAN considers only ingredients and the image feature learning is based on VGG versus ResNet in other approaches. Our attention approach also outperforms JNE in MedR by raising the median rank for 2 positions, and in R@5 by more than 5.4% of absolute recall improvement. The performance is even

Table 2: Performance of attention modeling on 5K dataset. The signs "+" and "-" indicate the results with and without attention modeling respectively.

	im2recipe								recipe2im							
	MedR		R@1		R@5		R@10		MedR		R@1		R@5		R@10	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
title	58.2	61.5	0.044	0.042	0.141	0.139	0.217	0.211	57.6	58.7	0.040	0.039	0.137	0.134	0.215	0.209
ingredient	71.0	73.0	0.045	0.039	0.135	0.123	0.202	0.192	70.1	72.0	0.042	0.039	0.133	0.126	0.202	0.196
instruction	33.9	36.2	0.070	0.068	0.202	0.198	0.298	0.286	33.2	35.1	0.066	0.065	0.201	0.198	0.295	0.290
all	20.0	22.4	0.104	0.099	0.275	0.265	0.382	0.371	19.1	21.7	0.101	0.098	0.272	0.266	0.382	0.372

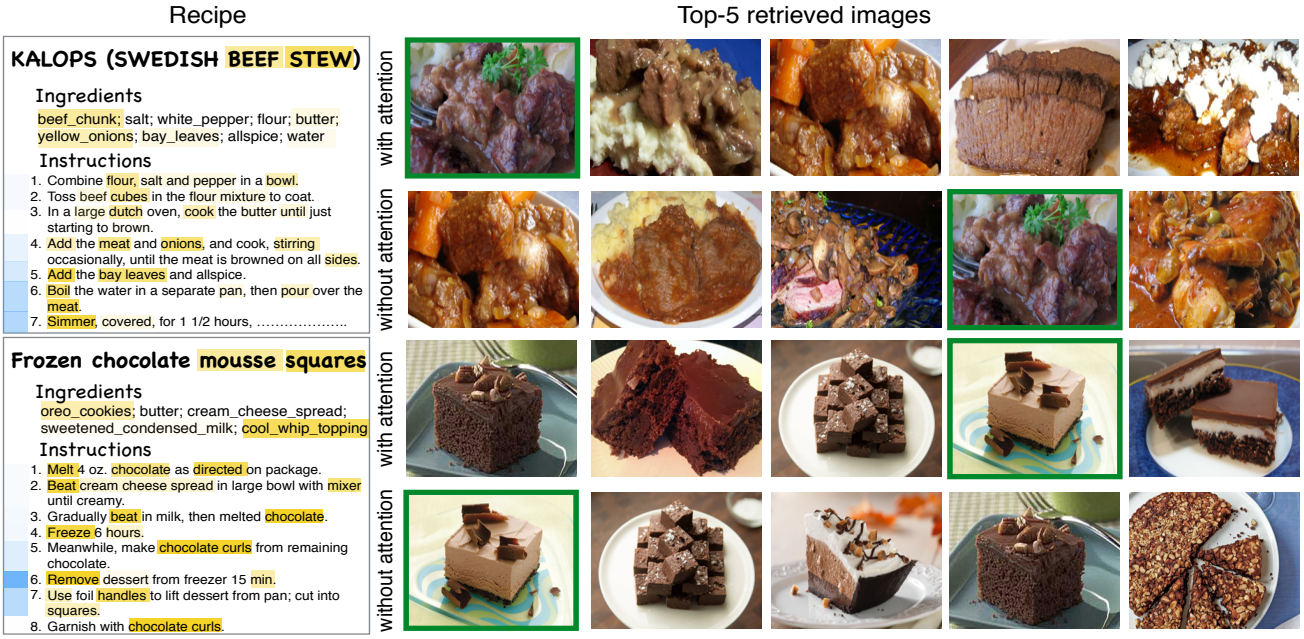

Figure 4: Results of image-to-recipe retrieval when attention weights are assigned properly (top) and incorrectly (bottom). The weights of words are highlighted by yellow pen, and the weights of sentences are indicated by blue bar. The intensity of colour indicates the degree of weight. True positives are bounded in green box.

slightly better than JNE+SR. Besides, using LSTM as encoder has similar performances with using GRU. When further enhancing our approach with attention+SR, however, only slight improvement is attainable. We speculate that the advantage of SR is limited on our approach because title information has been encoded as attended features for similarity learning. Further imposing food categorization performance, which is equivalent to learning to name food or recipe, in model training can only result in little gain in performance. On the other hand, as no end-to-end learning is conducted between SR and ResNet-50 image features, which could potentially increase training complexity, the improvement is also expected to be limited. Despite similar performance level as JNE+SR, our deep model is more intuitive than [30] because no ad-hoc compilation of food categorization by semi-automatic text mining is required.

As we move from 1K to 5K and 10K datasets, the performance gap between attention and JNE also gets larger, as indicated in Table 3. Our approach with attention manages to boost MedR by 10 and 20 ranks on 5K and 10K datasets, respectively, compared with JNE. When semantic regularization is employed, both approaches improve and attention+SR again outperforms JNE+SR with larger margin as data size increases.

4.6 Recipe preprocessing & cross-lingual retrieval

The recipes in Recipe1M dataset are contributed by Internet users and written in free-form. Thus, even extracting ingredient names out of recipes is considered not easy. In the previous experiments, we use the ingredients extracted by bi-directional LSTM as developed in [30] as input to our attention model. With this named-entity extraction technique, for example, *olive_oil* (instead of *olive* or *oil*) will be extracted from the sentence "1 tbsp of olive oil". Nevertheless, the extraction technique sometimes fails to extract ingredients from sentences such as "1 pack udon noodles" or "One 15 oz(240g) can chickpeas, drained and rinsed". Since attention model is capable of assigning weights to words and sentences, we speculate that the effect of noisy texts will be alleviated or even masked out during training. Therefore, instead of explicit preprocessing of recipes, we use raw recipes as input for model learning. In this experiment, we only remove numeric numbers from raw recipes to avoid the explosion of vocabulary size which will adversely affect learning effectiveness.

Table 4 shows the result that directly processing raw recipes can lead to further improvement than using the preprocessed recipes

Table 3: Performance comparison of our approach (attention) with various existing methods. The results of JNE and JNE+SR are quoted from [30]. The symbol '-' indicates that the result is not available in [30].

		im2recipe				recipe2im			
		MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10
1K	random	500	0.001	0.005	0.01	500	0.001	0.005	0.01
	CCA [30]	15.7	0.14	0.32	0.43	24.8	0.09	0.24	0.35
	SAN [8]	16.1	0.125	0.311	0.423	-	-	-	-
	JNE [30]	7.2	0.20	0.45	0.58	6.9	0.20	0.46	0.58
	JNE + SR [30]	5.2	0.24	0.51	0.65	5.1	0.25	0.52	0.65
	attention (LSTM)	4.8	0.253	0.530	0.665	4.8	0.255	0.536	0.665
	attention	4.8	0.254	0.532	0.663	4.7	0.256	0.534	0.667
	attention + SR.	**4.6**	**0.256**	**0.537**	**0.669**	**4.6**	**0.257**	**0.539**	**0.671**
5K	JNE [30]	31.5	-	-	-	29.8	-	-	-
	JNE + SR [30]	21.2	-	-	-	20.2	-	-	-
	attention	20.0	0.104	0.274	0.382	19.1	0.101	0.272	0.382
	attention + SR	**19.7**	**0.105**	**0.275**	**0.385**	**19.0**	**0.104**	**0.274**	**0.384**
10K	JNE [30]	62.8	-	-	-	58.8	-	-	-
	JNE + SR [30]	41.9	-	-	-	39.2	-	-	-
	attention	40.7	0.070	0.191	0.274	38.9	0.069	0.192	0.276
	attention + SR	**39.8**	**0.072**	**0.192**	**0.276**	**38.1**	**0.070**	**0.194**	**0.278**

Table 4: Results of parsing recipes without (i.e., raw recipe) and with (i.e., preprocessed recipe) named-entity extraction.

		im2recipe				recipe2im			
		MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10
1K	Raw recipe	4.4	0.259	0.546	0.671	4.2	0.262	0.551	0.677
	Preprocessed recipe	4.8	0.254	0.532	0.663	4.7	0.256	0.534	0.667
5K	Raw recipe	18.1	0.111	0.290	0.402	17.7	0.105	0.293	0.405
	Preprocessed recipe	20.0	0.104	0.274	0.382	19.1	0.101	0.272	0.382
10K	Raw recipe	37.2	0.072	0.202	0.290	35.3	0.069	0.203	0.294
	Preprocessed recipe	40.7	0.070	0.191	0.274	38.9	0.069	0.192	0.276

Table 5: Cross-lingual retrieval performance.

		MedR	R@1	R@5	R@10
Raw Recipe	Original	4.0	0.273	0.618	0.727
	Translated	8.0	0.218	0.455	0.564
Preprocessed Recipe	Original	4.0	0.291	0.545	0.673
	Translated	14.0	0.109	0.382	0.455

from [30]. The margin of improvement also gets larger with increase of data size. By attention modeling, our approach manages to recover some cases where ingredients are missed by named-entity extraction. In the example of "1 pack udon noodles", "udon" is assigned a relatively higher weight than other words, although our approach is incapable of extracting "udon noodles" as a phrase.

To further test the robustness of attention modeling on noisy text description, we conduct a simulation for cross-lingual recipe retrieval. The simulation is carried out by Google translating the English version recipes into recipes of different languages. We then reverse the process by translating the recipes in different languages back into English version for retrieval. During this process, the text description becomes noisy, for example, "in a large stockpot" becomes "in a big soup pot" and "stir-fried bee hoon" becomes "fry fried bees". Table 5 shows the result, where 55 English recipes are subsequently translated from English → Chinese → Japanese → English and then issued as queries for retrieval on the 1K dataset. As expected, the performance of using translated recipes is not as good as the original recipes. When directly processing the raw recipes, the top positives averagely drop by 4 ranks to 8th position in the retrieval list. The result is acceptable because a user can still locate the right recipe within the top-10 retrieved result. Applying named-entity extraction on the translated recipes, on the other hand, suffers larger rank degradation, where the MedR drops from 4th to 14th position. The result basically indicates the resilience of attention modeling in dealing with noisy text description.

5 CONCLUSIONS

We have presented a deep hierarchical attention model for the understanding of recipes. The model clearly shows the merit of leveraging cooking procedure for retrieval. More importantly, the advantage of attention modeling is evidenced in experiment – higher retrieval performance can be attained when weights are properly assigned to the sentences where their cooking effects are visible on images. Compared with [30], we also show that preprocessing of recipes with named-entity extraction is unnecessary, and indeed, directly processing raw recipes with attention leads to better performance. Currently, our work considers each section of recipes independently, which leads to inconsistency in weight assignment for the same words repeatedly appear in title, ingredient and instruction sections. In addition, co-attention modeling, i.e., assigning weights to both text and image regions, is not explored. Both issues will be the future directions of this work.

6 ACKNOWLEGEMENT

The work described in this paper was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11203517) and NExT++ project supported by National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative.

REFERENCES

[1] Kiyoharu Aizawa, Yuto Maruyama, He Li, and Chamin Morikawa. 2013. Food balance estimation by using personal dietary tendencies in a multimedia food log. *IEEE Transactions on multimedia* 15, 8 (2013), 2176–2185.

[2] Kiyoharu Aizawa and Makoto Ogawa. 2015. Foodlog: Multimedia tool for healthcare applications. *IEEE MultiMedia* 22, 2 (2015), 4–8.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[5] Oscar Beijbom, Neel Joshi, Dan Morris, Scott Saponas, and Siddharth Khullar. 2015. Menu-match: Restaurant-specific food logging from images. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, 844–851.

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*. Springer, 446–461.

[7] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 32–41.

[8] Jingjing Chen, Lei Pang, and Chong-Wah Ngo. 2017. Cross-Modal Recipe Retrieval: How to Cook this Dish?. In *International Conference on Multimedia Modeling*. Springer, 588–600.

[9] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 335–344.

[10] Jing-jing Chen, Chong-Wah Ngo, and Tat-Seng Chua. 2017. Cross-modal recipe retrieval with rich food attributes. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1771–1779.

[11] Wei-Ta Chu and Jia-Hsing Lin. 2017. Food image description based on deep-based joint food category, ingredient, and cooking method recognition. In *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*. IEEE, 109–114.

[12] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 7–16.

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, and others. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.

[14] Jun Harashima, Yuichiro Someya, and Yohei Kikuta. 2017. Cookpad Image Dataset: An Image Collection as Infrastructure for Food Research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1229–1232.

[15] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation* 16, 12 (2004), 2639–2664.

[16] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni. 2016. Food image recognition using very deep convolutional networks. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 41–49.

[17] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni. 2017. Automatic diet monitoring: a review of computer vision and wearable sensor-based methods. *International Journal of Food Sciences and Nutrition* 68, 6 (2017), 656–670.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[19] Yoshiyuki Kawano and Keiji Yanai. 2014. Foodcam-256: A large-scale real-time mobile food recognitionsystem employing high-dimensional features and compression of classifier weights. In *ACM MM*. 761–762.

[20] Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 982–992.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[22] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. 3294–3302.

[23] Corby K Martin, Theresa Nicklas, Bahadir Gunturk, John B Correa, H Raymond Allen, and Catherine Champagne. 2014. Measuring food intake with digital photography. *Journal of Human Nutrition and Dietetics* 27, s1 (2014), 72–81.

[24] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2016. Wide-Slice Residual Networks for Food Recognition. *arXiv preprint arXiv:1612.06543* (2016).

[25] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. 2015. Im2calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE International Conference on Computer Vision*. 1233–1241.

[26] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. 2016. Being a Super Cook: Joint Food Attributes and Multi-Modal Content Modeling for Recipe Retrieval and Exploration. *IEEE Transactions on Multimedia* (2016).

[27] Zhao-Yan Ming, Jingjing Chen, Yu Cao, Ciarán Forde, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Food Photo Recognition for Dietary Tracking: System and Experiment. In *International Conference on Multimedia Modeling*. Springer, 129–141.

[28] Nitish Nag, Vaibhav Pandey, and Ramesh Jain. 2017. Health Multimedia: Lifestyle Recommendations Based on Diverse Observations. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 99–106.

[29] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 251–260.

[30] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning Cross-modal Embeddings for Cooking Recipes and Food Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[31] Pol Schumacher, Mirjam Minor, Kirstin Walter, and Ralph Bergmann. 2012. Extraction of procedural knowledge from the web: A comparison of two workflow extraction approaches. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 739–747.

[32] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[33] Alan F Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, 321–330.

[34] Chun-Yuen Teng, Yu-Ru Lin, and Lada A Adamic. 2012. Recipe recommendation using ingredient networks. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 298–307.

[35] Christoph Trattner and David Elsweiler. 2017. Investigating the healthiness of internet-sourced recipes: implications for meal planning and recommender systems. In *Proceedings of the 26th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 489–498.

[36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 3156–3164.

[37] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial Cross-Modal Retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 154–162.

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

[39] Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*. Springer, 451–466.

[40] Ruihan Xu, Luis Herranz, Shuqiang Jiang, Shuang Wang, Xinhang Song, and Ramesh Jain. 2015. Geolocalized modeling for dish recognition. *IEEE transactions on multimedia* 17, 8 (2015), 1187–1199.

[41] 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 *International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–6.

[42] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 3441–3450.

[43] Keiji Yanai and Yoshiyuki Kawano. 2015. Food image recognition using deep convolutional network with pre-training and fine-tuning. In *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*. IEEE, 1–6.

[44] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.