Abstract

Representing procedure text such as recipe for cross-modal retrieval is inherently a difficult problem, not mentioning to generate image from recipe for visualization. This paper studies a new version of GAN, named Recipe Retrieval Generative Adversarial Network (R^2GAN), to explore the feasibility of generating image from procedure text for retrieval problem. The motivation of using GAN is twofold: learning compatible cross-modal features in an adversarial way, and explanation of search results by showing the images generated from recipes. The novelty of R^2GAN comes from architecture design, specifically a GAN with one generator and dual discriminators is used, which makes the generation of image from recipe a feasible idea. Furthermore, empowered by the generated images, a two-level ranking loss in both embedding and image spaces are considered. These add-ons not only result in excellent retrieval performance, but also generate close-to-realistic food images useful for explaining ranking of recipes. On recipe1M dataset, R^2GAN demonstrates high scalability to data size, outperforms all the existing approaches, and generates images intuitive for human to interpret the search results.

1. Introduction

Food is fundamental to health and social participation. Due to abundant food images and recipes available online, food computing for healthcare has recently captured numerous research attentions [34, 22]. Managing to retrieve the recipe of food intake, for example, can assist the estimation of nutrition consumption and hence benefit food logging [22,5]. The past efforts on food computing range from food categorization [19, 20, 21], food attribution recognition [3,4,23], zero-shot recipe retrieval [3] to food perception [35, 27] and recommendation [9,8,39].

This paper studies food-to-recipe and recipe-to-food retrieval, which is a typical problem of cross-modal retrieval [38] but peculiar to the domain of food computing. Specifically, recipe is a text article describing preparation of food material and procedure of cooking. A typical recipe consists of three sections: title, ingredients, and cooking instructions, which may or may not align with the visual appearance of a cooked dish. For instance, some ingredients (e.g., sugar, salt) are not visible in dish. Furthermore, cooking instruction more often implies the cause-and-effect of cooking rather than visually depicting the dish appearance. The nature of problem conflicts with the assumption made by the existing cross-modal retrieval, which trains model using text narration that explicitly refers to visual content [31, 32, 18]. Modeling lengthy procedure text such as recipe can thus be a new challenge for cross-modal retrieval.

In the literature, the problem of food-to-recipe retrieval is addressed by either classification [3, 4] or cross-modal learning [35, 2]. Classification-based approaches annotate rich food attributes (e.g., ingredients, cooking and cutting methods) in food images and then match these attributes against words extracted from recipes for retrieval [4]. A major drawback is the significant efforts required in labeling of food attributes, which are not only cost expensive and labour intensive. Cross-modal learning smartly alleviates this requirement, by training latent space that can accommodate both image and text modalities for similarity measurement. The labeling efforts are significantly reduced by requiring only recipe-image pairs, which are easy to collect, than to painstakingly annotate visual food attributes [4]. To model text description in recipe, neural networks of different complexities have been investigated in [35, 5] to learn embeddings for different sections of a recipe. Although efficient, cross-modal learning is inherently an unexplainable model compared to classification-based approaches, which are able to list out the matched attributes as evidences to re-
count the retrieval result.

This paper addresses the limitation of cross-modal learning for recipe retrieval. Specifically, a novel deep architecture is designed to interpret cross-modal matching, by synthesizing thumbnail images from recipes to assist the browsing of search results. The machine-generated thumbnails represent how a system perceives the effect of cooking and visually provides cue to explain the ranking of a recipe. Figure 1 shows the examples of thumbnails generated from recipes. As observed, these thumbnails (right) are not only similar to the examples (middle) generated from image embedding, but also the original images (left).

The proposed architecture is built upon cross-modal embedding [35] and generative adversarial network (GAN) [10]. Note that GAN has not yet been studied for this problem. Due to the use of GAN for Recipe Retrieval, we name the proposed model as $R^2GAN$. As recipes are rich of procedure descriptions, conventional GAN with one generator and one discriminator turns out to be ineffective. As a consequence, $R^2GAN$ is designed to have two discriminators, with one to guess between real and fake images as in common practice, and the other to predict the source of embedding, i.e., whether a fake image is generated from image or recipe embedding. Leveraging on the images generated from different modalities, a novel two-level rank loss function is designed to consider losses in both embedding and image spaces. The overall design of $R^2GAN$ is to encompass a rich set of functions to quantify cross-modal embedding, image reconstruction, food semantics and adversarial losses. With these, $R^2GAN$ is capable of learning compatible embeddings for image-to-recipe similarity measure, and performing recipe-to-image generation to explain the rationale of similarity.

The main contribution of this paper is exploration of GAN for cross-modal recipe retrieval. Despite the wide use of GAN in various problem domains [30][40][37][41], GAN surprisingly remains not attempted for recipe retrieval. Using GAN, this paper novelly utilizes image generation to visualize what is preserved in a recipe embedding for the explanation of search results. To the best of our knowledge, the proposed $R^2GAN$ with one generator and two discriminators is a relatively new idea. Although the design of dual discriminators has been recently investigated by D2GAN [26], the purpose is to address the issue of mode collapse by combining Kullback-Leibler (KL) and reverse KL divergences into a unified objective function in optimization, which is completely different from this paper. $R^2GAN$ aims for cross-modal learning and its dual discriminators, in contrast to D2GAN, are designed to be functionally different aiming to learn compatible embeddings and explainable thumbnails jointly.

2. Related Works

The core problem of cross-modal retrieval is to measure the similarity between two modalities. Learning common feature subspace is currently the main stream of research [38]. The approaches range from canonical correlation analysis (CCA) [31][29], which learns subspace to maximize correlation between modalities, to the most recent stacked cross attention model [17], which discovers the full latent alignment to capture fine-grained relationship across modalities. This section focuses on works relevant to food computing.

2.1. Recipe and Food Retrieval

Stacked attention model was first studied in [6] for image-to-recipe retrieval. By representing ingredients extracted from recipe as a binary vector, the model attends to image regions with salient ingredients for learning common latent space. This work, nevertheless, explores only ingredients and cannot disambiguate recipes with the same ingredients list but different cooking procedures. Joint neural embedding (JNE) addresses this problem by proposing bi-directional LSTM to embed the sparse list of ingredients and a hierarchical LSTM to encode the lengthy and complex descriptions of cooking procedure [35]. In addition, regularization with semantic loss, specifically to enforce the learnt embedding to predict food category, is found to be crucial in feature learning. The recent work in [5] improves JNE by introducing title encoder and multi-level attention modeling of cooking instructions from word-level to sentence-level. The new model is capable of assigning lower weights to visually insignificant words, such as “classic” and “home-made”, resulting in better retrieval accuracy. Built upon JNE [35], AdaMine recently proposed in [2] surpasses the performances of [35][5] with large margin, by proposing a double-triplet learning scheme and an adaptive strategy for informative triplet mining. The adaptive strategy is effective in alleviating the problem of gradient diminishing, and hence is also adopted by $R^2GAN$.

Classification-based approaches are also studied for this problem. In [5], ingredients are multi-labeled on food im-
ages to match recipes for retrieval. As only a limited number of 353 ingredients is trained for recognition, the idea of zero-shot recipe retrieval is introduced to retrieve recipes with ingredients unknown to a training model. The problem is addressed by constructing a large graph with both known and unknown ingredients as nodes. The graph models the co-occurrence relationship among ingredients, and conditional random field (CRF) is employed to propagate the prediction scores from known to unknown ingredients for recipe retrieval. This approach, nevertheless, is effective when only a small number of unknown ingredients is considered in the graph. The approach is later extended in [4] by predicting cooking and cutting attributes in addition to ingredients when matching with keywords extracted from recipes. Comparing to cross-modal retrieval, classification-based model is explainable as attributes are explicitly evaluated to quantify the final similarity score. However, training classification models to sufficiently cover a wide variety of food attributes for retrieval is practically intractable.

2.2. Cross-modal GAN

GAN has been applied for generating food images [13], but not in the context of cross-modal learning. In [13], conditioned on food category and ingredients respectively, CGAN [24] is employed to synthesize novel dish images. However, recipes information, including cooking style and process, has not yet been explored.

GAN has captured a lot of research attentions [1] [25] [41] [40] [15]. Although GAN has not been studied for recipe retrieval, cross-modal GAN is not a new idea. Examples include ACMR [37], GXN [11] and CM-GANS [28], with the common goal of learning embedding features for cross-modal retrieval. Different from most GANs, ACMR [37] does not have generator to reconstruct image. Instead, features are generated from images or text captions for the discriminator to guess the source of modality, which is similar to the second discriminator of $R^2GAN$. GXN [11] has two pairs of generator-discriminator, where a generator synthesizes examples of different modalities for discriminator to guess between real and fake samples. CM-GANS [28], different from ACMR and GXN, also has two pairs of generator-discriminator for image-to-image and text-to-text generation. Similar to ACMR, cross modal learning is enabled by having a discriminator to predict the modality of an embedded feature. Having two pairs of generator-discriminator is not considered in $R^2GAN$ because generating procedure description from image is practically implausible. Instead, the design of pairing one generator with dual discriminators is adopted. Different from ACMR and CM-GANS, the second discriminator of $R^2GAN$ makes prediction of modality source on the generated images rather than embeddings. The design enables $R^2GAN$ to encapsulate a rich set of loss functions as well as using two-level ranking losses for effective learning of compatible features.

3. $R^2GAN$

3.1. Preliminaries

Problem Formulation. The goal of image-to-recipe retrieval is to search for relevant recipes that textually describe the preparation of a dish given a food image as query. Similar but in the reverse direction, recipe-to-image retrieval is to rank food images according to the likelihood of being cooked based on a given recipe. Denote $P = \{p_i = (r_i, v_i)\}_{i=1}^{N}$ as a set of $N$ recipe-image pairs, where $r_i \in R$ is a recipe and $v_i \in V$ is its food image. The notations $R$ and $V$ denote the collections of recipes and images respectively. A pair $p_i$ may be assigned a semantic label $c_i \in C$, where $C \subseteq \mathbb{R}^k$ represents the set of $k$ food categories such as waffle, spaghetti bolognese and chicken quesadilla, which correspond to the predefined food groups of recipes. It is worth noting that each image belongs to a unique recipe, while each recipe is allowed to contain more than one image. Furthermore, the state of an image is assumed “after cooking”, meaning that an image captures only a fully prepared dish.

Due to the domain gap between recipe and image, the extracted raw features from both domains cannot be matched for similarity measurement. Similar in spirit as [35] [2], this paper aims to learn a common latent subspace as enable cross-modal comparison between recipe and food image. Specifically, a mapping function $\Psi(R, V) \rightarrow (E_R, E_V)$ needs to be learnt. Given $n$ recipe-image pairs, the function $\Psi$ produces both recipe embeddings $E_R$ and image embeddings $E_V$, where $E_R \in \mathbb{R}^{n \times d}$, $E_V \in \mathbb{R}^{n \times d}$, and $d$ is the dimension of the learnt embedding.

Generative Adversarial Network. The vanilla GAN [10] is composed of a generator $G$ and a discriminator $D$ which can be trained simultaneously in an adversarial way. The generator $G$ is trained to capture the real data distribution $p_{data}$ and generate fake images to fool discriminator $D$. On the other hand, the discriminator $D$ is trained to distinguish between real and fake images. Specifically, $G$ and $D$ play a minmax game to optimize the following objective function:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} \log D(x) + \mathbb{E}_{z \sim p_z(z)} \log (1 - D(G(z)))$$

(1)

where $x$ is the real image with a data distribution $p_{data}$, and $z$ is a noise with a prior distribution $p_z$.

3.2. Model Architecture

Figure depicts the model architecture of our $R^2GAN$. The architecture is composed of two modules for recipe and
image embeddings, and two modules for learning of GAN and semantic classification. The architecture is learned in an end-to-end fashion.

**Recipe Embedding Learning.** This module follows the work of [35], which employs a bi-directional LSTM and a hierarchical LSTM for representation learning of ingredients and cooking instructions respectively. The learnt representations are concatenated and fed into a fully connected layer for learning of recipe embedding.

**Image Embedding Learning.** Similar as other works in cross-modal recipe retrieval [35, 2, 5], the state-of-the-art ResNet-50 model is employed to extract image feature. We remove the last softmax classifier layer of ResNet-50 and initialize the rest layers with parameters pretrained in ImageNet ILSVRC12 dataset [33]. The resulting feature is further mapped by a fully connected layer to produce an image embedding in the same dimension as a recipe embedding.

**GAN Learning.** This module is specifically designed to learn compatible and explainable embeddings for image-recipe pairs. We redesigned vanilla GAN with one generator and two discriminators for cross-modal feature learning. As shown in Figure 2, the generator \( G \) is trained to be capable of reconstructing image from either recipe or image embedding. The reconstructed images from recipe and image embeddings are denoted as \( v_f^{\text{r}} \) and \( v_f^{\text{i}} \) respectively, where the subscript \( f \) represents a fake or reconstructed image and the superscript indicates the recipe or image source.

The first discriminator \( D_1 \), similar to traditional GAN, is to distinguish between real and fake images, i.e., \( v_{\text{real}} \) and \( v_f^{\text{i}} \). The second discriminator \( D_2 \), in contrast, is to differentiate between \( v_f^{\text{r}} \) and \( v_f^{\text{i}} \) to tell the source of modality. The intuition of having \( D_2 \) is to nudge the distribution of \( v_f^{\text{r}} \) to be as similar or compatible as \( v_f^{\text{i}} \) which is learnt from the original image \( v_{\text{real}} \). The generator \( G \) plays a special role in transforming textual recipe embeddings to images that are difficult for \( D_2 \) to predict the source. This min-max game played by GAN learning module novelty provides feedback to make the learnt recipe embedding self-explainable, specifically by having \( G \) to recount the visual appearance of an embedding for \( D_2 \) to make judgement. Note that this procedure naturally simulates an interpretable cross-modal retrieval, by showing user \( v_f^{\text{i}} \) as an explanation of how a recipe is visually interpreted and ranked by a system. In short, by having two discriminators, \( R^2GAN \) effectively enforces \( v_f^{\text{i}} \) to learn from real food image \( v_{\text{real}} \) and then \( v_f^{\text{r}} \) from \( v_f^{\text{i}} \), until reaching a state where the reconstructed images from a different modality share similar or even a same distribution with the original image.

**Semantic Learning.** \( R^2GAN \) also takes advantage of high-level semantics (i.e., food categories) to assist the learning of recipe and image embeddings. Intuitively, both modalities should exhibit the same semantic interpretation when projected to the same common subspace.
3.3. Objective Formulation

Two-level Ranking Loss. Similar to other cross-modal retrieval methods [17, 35], triplet ranking loss is employed. Different from these works, nevertheless, $R^2$GAN considers two-level of losses due to embedding and reconstruction. Let $E$ represent an embedding, $v$ as a reconstructed image, and the subscripts $q, p, n$ refer to query, positive and negative candidates respectively. We use a large-margin based ranking loss function which can be formalized as follows:

$$L_{rank} = \max\{d(E_q, E_p) - d(E_q, E_n) + \alpha_1, 0\} + \mu \max\{d(v_q, v_p) - d(v_q, v_n) + \alpha_2, 0\},$$

where $d(\cdot, \cdot)$ is a distance function measuring the similarity between a given pair of query and candidate, for example, $(E_q, E_p)$ as a positive embedding pair and $(v_q, v_p)$ as the corresponding image pair. Note that the elements of a pair belong to different modalities. The parameters $\alpha_1$ and $\alpha_2$ are margins, and $\mu$ is a trade-off hyperparameter.

The two-level ranking loss enhances the robustness of learning, through enforcing the distances between positive pairs to be always smaller than negative pairs, not only in the embedding space but also the reconstructed image space. We use cosine similarity as distance function for embedding space as [35, 2], and pixel-wise Euclidean distance for image space.

Adversarial Loss. The three parts of $R^2$GAN, i.e., $G$, $D_1$, $D_2$, are optimized alternatively by adversarial training. Due to use of two discriminators, the losses produced by $D_1$ and $D_2$ are averaged as the training loss of $G$. Therefore, the GAN module losses are as follows:

$$L_{D_1} = \mathbb{E}_{x \sim \text{image}} [\log D_1(x)] + \mathbb{E}_{E_V \sim \text{image}} [\log (1 - D_1(G(E_V)))],$$

$$L_{D_2} = \mathbb{E}_{E_V \sim \text{image}} [\log D_2(G(E_V))] + \mathbb{E}_{E_R \sim \text{recipe}} [\log (1 - D_2(G(E_R)))],$$

$$L_G = \frac{1}{2} [\mathbb{E}_{E_V \sim \text{image}} [\log (1 - D_1(G(E_V)))] + \mathbb{E}_{E_R \sim \text{recipe}} [\log (1 - D_2(G(E_R)))],$$

where $E_R$ and $E_V$ denote embeddings of recipe and image respectively.

Reconstruction Loss, which also considers two-level of losses in feature and image levels, is introduced to encourage the reconstructed images to retain as much as information of the original image. The reconstruction loss is defined as follows:

$$L_{recon} = \frac{1}{2} \left( \|\Phi(v_{\text{real}}) - \Phi(v_f)\|^2_2 + \|\Phi(v_f) - \Phi(v_R)\|^2_2 + \beta(\|v_{\text{real}} - v_f\|^2_2 + \|v_f - v_R\|^2_2) \right),$$

where $\Phi(\cdot)$ is a feature extractor for the input image, $v_{\text{real}}$ stands for real food image, and the images $v_f$ and $v_R$ are reconstructed from image and recipe embeddings respectively. Following the practice in [7], the output before last layer of the discriminator is used as $\Phi(\cdot)$. The term $\|\Phi(v_1) - \Phi(v_2)\|^2_2$ refers to feature-level loss and the term $\|v_1 - v_2\|^2_2$ refers to the image-level loss, with both using Euclidean distance. The parameter $\beta$ controls the relative importance between feature and image losses.

Semantic Loss is characterized by cross-entropy loss as following:

$$L_{sem} = -\log \frac{\exp(E_c)}{\sum_i \exp(E_c_i)},$$

where $E_c$ denotes either a recipe or image embedding category.

Overall Loss. The four modules of $R^2$GAN are learnt end-to-end. However, the parameters of modules are optimized separately using different loss functions. The full loss, defined as following, is used to update the parameters of embedding and semantic modules:

$$L_{full} = L_{rank} + \gamma L_{recon} + \lambda L_{sem},$$

where $\gamma$ and $\lambda$ are trade-off hyperparameters.

On the other hand, the parameters of two discriminators are updated by $L_{D_1}$ and $L_{D_2}$, while the parameters of generator $G$ are updated by incorporating adversarial and reconstruction losses as following:

$$L_{G_{full}} = L_G + \delta L_{recon},$$

where $\delta$ balances the relative importance of the two parts.

4. Experiments

4.1. Experiment Settings

Dataset. Recipe 1M [35] is the only large-scale food dataset with English recipes and images publicly available. The raw dataset contains more than 1 million recipes and almost 900,000 images. The experiments are conducted on the pre-processed recipe-image pairs provided by [35], which have totally 340,922 pairs with 70% for training, 15% for validation and 15% for testing. Each pair is assigned to one of the 1,048 semantic food categories compiled by [35].

Evaluation Metrics. Median rank (MedR) and recall rate at top K (R@K) are used to evaluate retrieval accuracy. MedR refers to the median rank position of true positives for all the testing queries. R@K measures the fraction of

1 An alternative way of computing $\Phi(\cdot)$ is by using VGG network [14]. However, there is no obvious performance difference between these two approaches in our in-house experiment.
Methods | image-to-recipe | recipe-to-image |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1K</strong></td>
<td><strong>10K</strong></td>
<td><strong>1K</strong></td>
</tr>
<tr>
<td>Random</td>
<td>Random</td>
<td>AdaMine</td>
</tr>
<tr>
<td>MedR</td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>500</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>15.7</td>
<td>14.0</td>
<td>32.0</td>
</tr>
<tr>
<td>5.2</td>
<td>24.0</td>
<td>51.0</td>
</tr>
<tr>
<td>4.6</td>
<td>25.6</td>
<td>53.7</td>
</tr>
<tr>
<td>2.5</td>
<td>36.4</td>
<td>66.2</td>
</tr>
<tr>
<td><strong>R^2GAN</strong></td>
<td><strong>R^2GAN</strong></td>
<td><strong>R^2GAN</strong></td>
</tr>
<tr>
<td>2.0</td>
<td>39.1</td>
<td>71.0</td>
</tr>
<tr>
<td>39.8</td>
<td>7.2</td>
<td>19.2</td>
</tr>
<tr>
<td>24.8</td>
<td>9.0</td>
<td>24.0</td>
</tr>
<tr>
<td>39.2</td>
<td>7.0</td>
<td>19.4</td>
</tr>
<tr>
<td>13.9</td>
<td>13.5</td>
<td>33.5</td>
</tr>
<tr>
<td>15.7</td>
<td>14.0</td>
<td>32.0</td>
</tr>
<tr>
<td>5.2</td>
<td>24.0</td>
<td>51.0</td>
</tr>
<tr>
<td>4.6</td>
<td>25.6</td>
<td>53.7</td>
</tr>
<tr>
<td>2.5</td>
<td>36.4</td>
<td>66.2</td>
</tr>
</tbody>
</table>
| **Table 1.** Cross-modal retrieval performance comparison in terms of MedR (median rank) and R@K (recall@K). A lower MedR and a higher R@K indicate a better model. The symbol “-” means that the results are not available in the original paper.
Figure 3. Scalability test between $R^2$GAN and AdaMine for image-to-recipe retrieval.

on Recipe1M, $R^2$GAN manages to boost MedR by almost three ranking positions in both image-to-recipe and recipe-to-image retrieval in 10K setting. Observed from the similar thumbnails generated from image and recipe embeddings, we attribute the improvement to the peculiar design of the GAN learning module which enforces the embedding module to learn more compatible features.

**Scalability.** To investigate the robustness $R^2$GAN against large dataset beyond 10K, we further compare its MedR performance against AdaMine. For image-to-recipe retrieval, as shown in Figure 3, the gap between $R^2$GAN and AdaMine becomes obvious and larger with the increase of subset size. On the 50K dataset, which is almost equivalent to the original size of testing set provided by [35], $R^2$GAN manages to rank the true positive by 11.4 positions ahead of AdaMine on average, which is statistically significant. Similar results are also obtained for recipe-to-image search, where $R^2$GAN ranks true positives by 14 positions ahead of AdaMine on average, which is statistically significant. As claimed in JNE [35] and ATTEN [5] that food semantics play an important role, we also study the performance of two other variants without semantic classification (i.e., $R^2$GAN-Semantic in Figure 5(c)) and with only semantic classification (i.e., Semantic only in Figure 5(d)). Additionally, we also compare to a variant, $R^2$GAN-, which employs conventional one-level ranking loss without image-level ranking loss. In other words, Equation 2 is modified as follows:

$$L_{rank} = \max\{d(E_q, E_p) - d(E_q, E_n) + \alpha_1, 0\}, \quad (10)$$

Table 2 lists the results of ablation study. First of all, the baseline GAN already outperforms all the previous models including AdaMine on this dataset. However, GAN*, which uses a variant of $D_2$, exhibits worse performance than GAN which is without $D_2$. The result is not surprising because reconstruction of image from recipe is highly difficult. Directly learning to imitate real image can re-

<table>
<thead>
<tr>
<th>Query Image</th>
<th>Ground Truth</th>
<th>Retrieved Recipe Title</th>
<th>Ranking</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christmas Pudding Granola</td>
<td>Christmas Pudding Granola</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pumpkin Spice Latte Granola</td>
<td>Peanut Butter and Nutella Popcorn</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saffron Cherry Oat Muffins</td>
<td>Saffron Cherry Oat Muffins</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Christmas Pudding Granola</td>
<td>Christmas Pudding Granola</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pumpkin Spice Latte Granola</td>
<td>Peanut Butter and Nutella Popcorn</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saffron Cherry Oat Muffins</td>
<td>Saffron Cherry Oat Muffins</td>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4. Examples showing the interpretability of $R^2$GAN. By judging from the generated images (last column) from recipes, one can easily guess the ground-truth recipes of query images.
Table 2. Ablation study. Results are reported in terms of MedR with different subset sizes.

<table>
<thead>
<tr>
<th>Methods</th>
<th>image-to-recipe</th>
<th>recipe-to-image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10K 20K 30K 40K 50K</td>
<td>10K 20K 30K 40K 50K</td>
</tr>
<tr>
<td>Semantic only</td>
<td>16.0 30.6 45.7 60.8 75.7</td>
<td>15.1 28.6 42.8 56.8 70.9</td>
</tr>
<tr>
<td>R^2GAN-Semantic</td>
<td>19.3 37.8 55.9 74.1 92.9</td>
<td>18.1 35.6 52.7 69.8 87.0</td>
</tr>
<tr>
<td>GAN</td>
<td>15.8 30.7 45.7 60.3 75.2</td>
<td>14.2 28.1 41.9 55.4 69.0</td>
</tr>
<tr>
<td>GAN*</td>
<td>19.3 37.9 56.1 74.2 92.9</td>
<td>17.2 34.0 50.5 67.1 83.4</td>
</tr>
<tr>
<td>R^2GAN</td>
<td>14.6 28.4 42.0 55.2 69.0</td>
<td>13.2 25.2 37.5 49.9 61.9</td>
</tr>
<tr>
<td>R^2GAN-Semantic</td>
<td>13.9 26.8 39.9 52.7 66.0</td>
<td>12.6 24.2 35.7 47.4 59.0</td>
</tr>
</tbody>
</table>

Figure 5. Variants of architectures derived from \( R^2GAN \) for ablation study.

Figure 6. Comparison of images generated by \( R^2GAN \), GAN* and GAN. The last column shows the thumbnails reconstructed from image embedding \( v_f^I \) and recipe embedding \( v_f^R \).

Result in overfitting harmful to the overall end-to-end learning. Instead, indirectly learning as in \( R^2GAN \) to imitate fake image generated from image embedding, which is inherently an easier task, appears to be more effective. The result listed in Table 2 also aligns with [35, 3] where semantic loss plays a critical role. Semantic-only, which is without GAN, performs better than its counterpart \( R^2GAN - \) Semantic, which is with GAN only but without semantics. The proposed \( R^2GAN \) successfully compromises both information, i.e., semantics and GAN, and shows the consistently best performances across subsets of different sizes from 10K to 50K. Comparing two-level versus one-level ranking loss, \( R^2GAN \) also shows incremental improvement over \( R^2GAN - \) consistently across all the subsets. Figure 6 compares the images generated from image and recipe embeddings by different GANs. \( R^2GAN \) manages to generate thumbnails substantially more realistic than other variants and are apparently more similar to the original images.

5. Conclusion

We have presented a new network architecture based on GAN for cross-modal recipe retrieval, which attains the new state-of-the-art performance on Recipe1M dataset. \( R^2GAN \), particularly, exhibits robustness against large-size dataset and is more scalable compared to other models. Through the experiments, we attribute the improvement to the design of architecture which makes the learning of embedding compatible across text and visual modalities. This can be evidenced from the high similarity in food images despite being generated from different modalities. These generated images also greatly facilitate the self-explaining of search results. Using more advanced GANs [1, 25] and generating higher resolution images [40] may further improve performance and enhance search result interpretation. Through ablation studies, we show that the design of dual discriminators plays an important role in boosting the retrieval performance. Finally, despite that the two-level ranking loss boosts performance by a relatively small margin, the improvement is consistently noticed across different sizes of subsets. While encouraging, \( R^2GAN \) currently considers only image generation from recipe and not vice versa. With the release of new dataset, such as [12] which includes processing images for every step of cooking instructions, potentially recipe-from-image is a mission-possible task which worth further investigation.

Acknowledgement

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


