6-2019

Learning cross-modal embeddings with adversarial networks for cooking recipes and food images

Hao WANG
Singapore Management University, hwang@smu.edu.sg

Doyen SAHOO
Singapore Management University, doyens@smu.edu.sg

Chenghao LIU
Singapore Management University, chliu@smu.edu.sg

Ee-peng LIM
Singapore Management University, eplim@smu.edu.sg

Steven C. H. HOI
Singapore Management University, CHHOI@smu.edu.sg

Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images

Wang Hao[†,*] Doyen Sahoo[†,*] Chenghao Liu[†] Ee-peng Lim[†] Steven C. H. Hoi[†,‡]

[†]Singapore Management University [‡]Salesforce Research Asia
{hwang, doyens, chliu, eplim, chhoi}@smu.edu.sg

Abstract

Food computing is playing an increasingly important role in human daily life, and has found tremendous applications in guiding human behavior towards smart food consumption and healthy lifestyle. An important task under the food-computing umbrella is retrieval, which is particularly helpful for health related applications, where we are interested in retrieving important information about food (e.g., ingredients, nutrition, etc.). In this paper, we investigate an open research task of cross-modal retrieval between cooking recipes and food images, and propose a novel framework Adversarial Cross-Modal Embedding (ACME) to resolve the cross-modal retrieval task in food domains. Specifically, the goal is to learn a common embedding feature space between the two modalities, in which our approach consists of several novel ideas: (i) learning by using a new triplet loss scheme together with an effective sampling strategy, (ii) imposing modality alignment using an adversarial learning strategy, and (iii) imposing cross-modal translation consistency such that the embedding of one modality is able to recover some important information of corresponding instances in the other modality. ACME achieves the state-of-the-art performance on the benchmark Recipe1M dataset, validating the efficacy of the proposed technique.

1. Introduction

With the rapid development of social networks, Internet of Things (IoT), and smart-phones equipped with cameras, there has been an increasing trend towards sharing food images, recipes, cooking videos and food diaries. For example, the social media platform "All Recipes" [1] allows chefs to share their created recipes and relevant food images. Their followers or fans follow the cooking instructions, upload their pictures for reproducing the same dishes and share their experiences for peer comments. As a result, the community has access to rich, heterogeneous sources of information on food. In recent years, food-computing [32] has become a popular research topic due to its far-reaching impact on human life, health and well being. Analyzing food data could support a lot of human-centric applications in medicine, biology, gastronomy, and agronomy [32]. One of the important tasks under the food-computing umbrella is Food Retrieval, i.e., we are interested in retrieving relevant information about a specific food. For example, given a food image, we are interested in knowing its recipe, nutrition content, or calorie information. In this paper, we investigate the problem of cross-modal retrieval between cooking recipes and food images, where our goal is to find an effective latent space to map recipes to their corresponding food images.

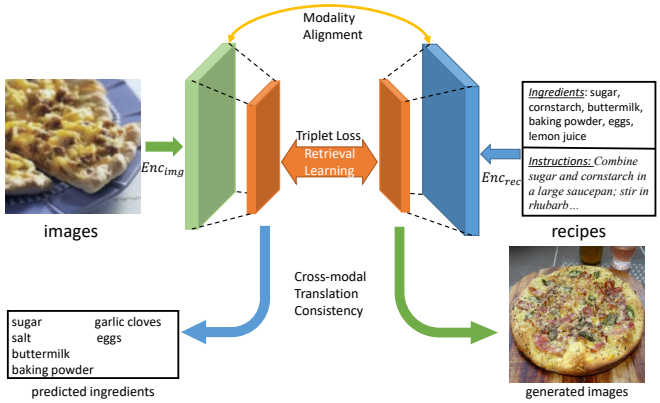

Figure 1: Example of illustrating the idea of Adversarial Cross-Model Embedding (ACME), where the embeddings of cooking recipes and food images are aligned. These embeddings are useful for several health applications from the perspective of understanding characteristics about food, nutrition and calorie intake.

The idea of cross-modal retrieval in the food domain is to align matching pairs in a common space, so that given

*denotes equal contribution
[1]https://www.allrecipes.com

a recipe, the appropriate relevant images can be retrieved, or given a food image, the corresponding recipe can be retrieved. Recent efforts addressing the problem for large-scale cross-modal retrieval between cooking recipes and food images [40, 6, 9] use CNNs [18] for encoding the food images, and LSTM [21] to encode recipes, and align the feature vector using common retrieval losses (e.g., pairwise cosine loss and triplet loss). Despite achieving promising performance, there are several critical shortcomings: (i) there can be high variation in images corresponding to one recipe, making it difficult for a triplet loss using a naive sampling strategy to converge quickly; (ii) They do not consider aligning the feature distributions of the two heterogeneous modalities: cooking recipes and food images, which can have very different distributions; and (iii) Cross-Modal mapping using Deep Neural Networks can lead to possible loss of information in the embedding process, and for the existing approaches it is unclear if the embedding of one modality is able to capture relevant semantic information in the other modality.

To address the above limitations, we propose a novel end-to-end framework named **A**dversarial **C**ross-**M**odal **E**mbedding (ACME). Specifically, we propose an improved triplet loss scheme empowered with hard sample mining, which considerably improves the performance and convergence over a traditional triplet loss. We propose to align the embeddings from the two modalities using an adversarial loss, in an attempt to making the feature distribution of the cooking recipes and food images indistinguishable. We also enforce the cross-modal translation consistency by recovering the relevant information of one modality from the feature embedding of another modality, i.e., we train the model to generate an appropriate image given a recipe embedding, and to predict the ingredients given a food image embedding.

Figure 1 shows an overview of ACME framework, which is end-to-end trainable, and significantly outperforms state-of-the-art baselines for cross-modal retrieval on Recipe1M [40] dataset. We conduct extensive ablation studies to evaluate performance of various components of ACME. Finally, we do qualitative analysis of the proposed method, and visualize the retrieval results. The code is publicly available[2].

2. Related Work

2.1. Food Computing

Food Computing has evolved as a popular research topic in recent years [32, 2, 1]. With rapid growth of IoT and the infiltration of social media into our daily lives, people regularly share food and image recipes online. This has given rise to using this rich source of heterogeneous information for several important tasks, and an immense opportunity to analyze food related information. This has far reaching impact to our daily lives, behaviour, health [35, 13] and culture[39]. Such analysis could have implications in medicine[15], biology[5], gastronomy [33], agronomy [20] etc.

There are many ways to analyze food, and many tasks that can be addressed using food computing. One commonly studied task is recognition of food from images. This problem has been studied for years, where large datasets are collected, and prediction models are trained on them. Early approaches used kernel-based models [25], while more recent approaches have exploited deep CNNs [27, 45, 30]. Another task that has received substantial attention is food recommendation. This is a more difficult task than traditional recommendation systems, as a variety of contextual heterogeneous information needs to be integrated to make recommendations. There have been several efforts in literature for food recommendation, including chatbot-based[14], and dietary recommendation for diabetics[38].

Our focus is on another common task: Food Retrieval. We aim to retrieve relevant information about a food item given a query. For example, it can be a difficult task to estimate calorie and nutrition from a food image, but if we could retrieve the recipe and ingredients from an image, the task of nutrition and calorie estimation becomes much simpler. There are several types of retrieval within food computing: image-to-image retrieval [10], recipe-to-recipe retrieval [7], and cross-modal image-to-recipe and recipe-to-image retrieval[40]. Our work is in the domain of cross-modal retrieval between cooking recipes and food images.

2.2. Cross-Modal Retrieval

The goal of cross-modal retrieval is to retrieve relevant instances from a different modality, e.g. retrieving an image using text. The main challenge lies in the media gap [36], which means features from different modalities are inconsistent, making it difficult to measure the similarity.

To solve this issue, many efforts focus on using pairs to learn a similarity or distance metric to correlate cross-modal data [40, 42, 11, 44]. Apart from metric learning methods, some alignment ideas are also used for cross-modal retrieval: like global alignment [23, 4, 3] and local alignment [26, 24, 34]. Canonical Correlation Analysis (CCA) [23] utilizes global alignment to allow the mapping of different modalities which are semantically similar by maximizing the correlation between cross-modal (similar) pairs. In [26], local alignment is used to embed the images and sentences into a common space. In order to enhance both global and local alignment, [24] learns a multi-modal embedding by optimizing pairwise ranking. Another line of work uses adversarial loss [16] which has often been used for alignment in domain adaptation[22] and cross-modal retrieval [43].

Existing work for large-scale cross-modal retrieval for cooking recipes and food images include JE [40], which uses a pairwise cosine loss to learn a joint embedding between the two modalities. This method was extended to

[2] https://github.com/LARC-CMU-SMU/ACME

use a simple triplet loss by AdaMine [6]. Another approach tried to improve the embedding of the recipes using hierarchical attention [9]. Unlike these efforts, our work uses a superior sampling strategy for an improved triplet loss, imposes cross-modal alignment, and enforces cross-modal translation consistency towards achieving more effective and robust cross-modal embedding.

3. Adversarial Cross-Modal Embedding

We now present our proposed Adversarial Cross-Modal Embedding (ACME) between recipes and food images.

3.1. Overview

Given a set of image-recipe pairs $(\mathbf{i}^t, \mathbf{r}^t)$ for $t = 1, \ldots, T$, where a food image $\mathbf{i}^t \in \mathbf{I}$ and a recipe $\mathbf{r}^t \in \mathbf{R}$ (where $\mathbf{I}$ and $\mathbf{R}$ correspond to the image and recipe domains respectively), our goal is to learn embedding functions $\mathbf{E}_V : \mathbf{I} \rightarrow \mathbb{R}^d$ and $\mathbf{E}_R : \mathbf{R} \rightarrow \mathbb{R}^d$ which encode the image and the recipe into a $d-$dimensional visual vector and recipe vector, respectively. The embedding functions should be learned so that the embedding of a food image and its corresponding recipe should be close to each other (for $\mathbf{i}^{t1}$ and $\mathbf{r}^{t2}$ where $t1 = t2$), and should be distant from embedding of other images or recipes (for $\mathbf{i}^{t1}$ and $\mathbf{r}^{t2}$ where $t1 \neq t2$). Such an embedding is well suited for retrieval tasks. Note that in our work, the paired instances may have a many-to-one relationship from images to recipes, i.e., we may have many images for a given recipe. Accordingly, we want the embeddings of all the images for a given recipe to be close to the embedding of this recipe, and distant from the embedding of other recipes.

A simple way to learn this embedding is to use a pairwise cosine loss at the level of feature representation, and use back-propagation to learn the embedding functions [40]. As this loss function may not be ideal for retrieval tasks, triplet-loss was also considered [6]. However, we hypothesize that both of these approaches suffer from several critical limitations: (i) They give equal importance to all the samples during the optimization of the triplet loss, which may significantly hinder the convergence and generalization, as there could be a high variance among many images corresponding to the same recipe; (ii) Being from heterogeneous modalities, the feature distributions can be very dissimilar, and the existing approaches do not try to align these distributions; and (iii) The embedding functions possibly lose important information during the embedding process, as a result, the embedding of one modality may not effectively capture semantic information in the other modality.

To address these issues, we propose a novel end-to-end framework for **A**dversarial **C**ross-**M**odal **E**mbedding (ACME) between Cooking Recipes and Food Images. To address the first challenge, we propose a new retrieval learning component that leverages a hard sample mining [19] strategy to improve model training and performance of the existing triplet loss. To address the modality-distribution alignment, we use an adversarial loss [16] to ensure that features of the embedding functions across different modalities follow the same distribution. Finally, we introduce a novel cross-modal translation consistency component, wherein food images are generated using the embedded recipe features, and the ingredients of a recipe are predicted using the image features. The entire pipeline can be trained in an end-to-end manner with a joint optimization objective. The overall proposed architecture is shown in Figure 2.

More formally, during the feed-forward flow through the pipeline, the food images $\mathbf{i}$ and recipes $\mathbf{r}$ are encoded using a CNN (parameterized by $\mathbf{E}_V$) and an LSTM (parameterized by $\mathbf{E}_R$), respectively. The CNN gives us high-level visual features $V_m \in \mathbb{R}^d$, and the LSTM gives us high-level recipe features $R_m \in \mathbb{R}^d$. These high-level features are then passed through a fully-connected layer (parameterized by $\mathbf{FC}$), where both modalities share the same weight [37], with the purpose of correlating the common representation of each modality, and give us the final representation V and R for the visual features and recipe features, respectively. The recipe features R are then used to generate food images, and the visual features V are used to predict the ingredients in the particular instance. This framework is then optimized over three objectives: to achieve a feature representation that is good at retrieval tasks; to obtain a feature representation that aligns the distribution of the two modalities in order to make them modality-invariant; and the features achieve the cross-modal translation consistency. The overall objective of the proposed ACME is given as:

$$\mathcal{L} = \mathcal{L}_{Ret} + \lambda_1 \mathcal{L}_{MA} + \lambda_2 \mathcal{L}_{Trans}, \tag{1}$$

where λ_1 and λ_2 are trade-off parameters. The retrieval learning component $\mathcal{L}_{Ret}(V, R)$ receives the two high-level feature vectors: $V \in \mathbb{R}^d$ for the image and $R \in \mathbb{R}^d$ for the recipes, and computes the retrieval loss. The modality alignment component $\mathcal{L}_{MA}(V_m, R_m)$ operates on the penultimate layer features $V_m \in \mathbb{R}^d$ and $R_m \in \mathbb{R}^d$, and aims to achieve modality-invariance using an adversarial loss to align the two distributions. The cross-modal translation consistency component $\mathcal{L}_{Trans}$ is further divided into two sub components: recipe2image (generates food image from R) and image2recipe (predicts ingredients based on V). recipe2image and image2recipe are optimized using an adversarial loss and classification loss respectively. At the end of the training procedure, the feature representations V and R are used for retrieval tasks.

3.2. Cross-Modal Retrieval Learning

After images and recipes are passed through the encoder functions and the weight sharing layers, visual representations V and recipe representations R are obtained. Our goal is to have feature representation V^{t1} and R^{t2} be similar for a

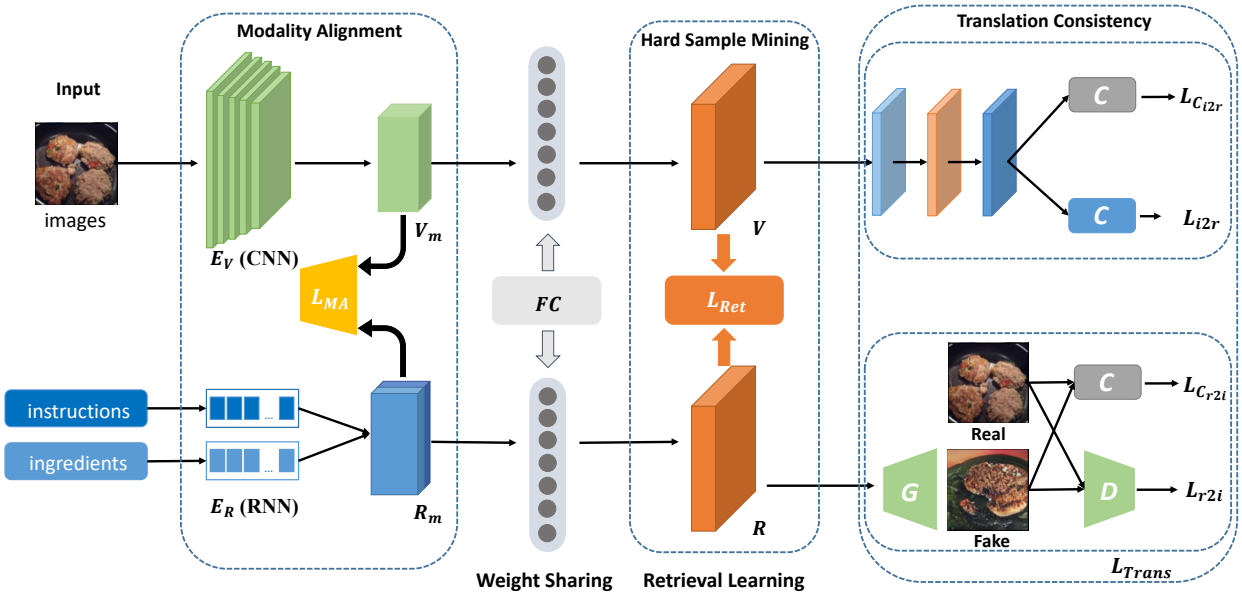

Figure 2: **ACME: Adversarial Cross-Modal Embedding:** Our proposed framework to achieve effective cross-modal embedding. The food images and recipes are encoded using a CNN and LSTM respectively to give feature embeddings. The feature embeddings are aligned using an adversarial loss $\mathcal{L}_{MA}$ to achieve Modality Alignment. Both image and text embeddings are then passed through a shared FC layer to give the final embedding. This embedding is learned by optimizing $\mathcal{L}_{Ret}$, which uses triplet loss with hard sample mining. This embedding is also optimized to achieve cross-modal translation consistency ($\mathcal{L}_{Trans}$), where the recipe embedding is used to generate a corresponding real food image, and the the food image embedding is used to predict the ingredients in the food item.

given pair ($t1 = t2$) and dissimilar for $t1 \neq t2$. This is done via the usage of the triplet loss [41]. A triplet comprises one feature embedding as an anchor point in one modality, and a positive and negative feature embedding from another modality. The positive instance corresponds to the one we want it to be similar to the anchor point, and the negative instance should be dissimilar to the anchor point. In our case, we have two types of triplets: one where the visual feature V behaves as the anchor, and another where the recipe feature R behaves as the anchor. The objective is then given as:

$$\begin{aligned}\mathcal{L}_{Ret} = &\sum_{V} [d(V_a, R_p) - d(V_a, R_n) + \alpha]_+ \\ &+ \sum_{R} [d(R_a, V_p) - d(R_a, V_n) + \alpha]_+ ,\end{aligned} \tag{2}$$

Our goal is to minimize this objective as:

$$\min_{\mathbf{E}_V, \mathbf{E}_R, \mathbf{FC}} \mathcal{L}_{Ret}$$

Here $d(\bullet)$ is the Euclidean distance, subscripts a, p and n refer to anchor, positive and negative samples, respectively, and α is the margin of error. To improve the convergence of learning, we integrate the hard sample mining strategy into the process of learning with triplet loss. In particular, unlike the traditional approach that simply treats all instances equally important for the triplet construction for a given anchor, the proposed approach gives preference to the most distant positive instance and the closest negative instance during the training procedure.

3.3. Modality Alignment

A major challenge in using encoded features from different modalities is that the distributions of the encoded features can be very different, resulting in poor generalization and slower convergence. A more desirable solution is to align the distribution of the encoded features. We aim to align the distributions of the penultimate layer features V_m and R_m. To do so, we use an adversarial loss, where we try to achieve a feature representation such that a discriminator D_M cannot distinguish whether the feature representation was obtained from the image or the recipe. We empirically adopt WGAN-GP [17] in our experiment. The objective $\mathcal{L}_{MA}$ is given as:

$$\begin{aligned}\mathcal{L}_{MA} = &\mathbb{E}_{\mathbf{i}\sim p_{image}}[\log D_M(\mathbf{E}_V(\mathbf{i})] + \\ &\mathbb{E}_{\mathbf{r}\sim p_{recipe}}[\log(1 - D_M(\mathbf{E}_R(\mathbf{r})))],\end{aligned} \tag{3}$$

and solved by a min-max optimization as:

$$\min_{\mathbf{E}_V, \mathbf{E}_R} \max_{D_M} \mathcal{L}_{MA}$$

3.4. Translation Consistency

In order to have better generalization, we want the learned embedding of one modality to be able to recover the corresponding information in the other modality, leading to better

semantic alignment. This would enforce a cross-modal translation consistency, ensuring that the learned feature representation preserve information across modalities. Specifically, we aim to use recipe features R to generate a food image, and visual features V to predict the ingredients of the recipe. The objective is given as:

$$\mathcal{L}_{Trans} = \mathcal{L}_{r2i} + \mathcal{L}_{i2r}$$

Below we present the details of each of these losses, which help achieve the cross-modal translation consistency.

3.4.1 Recipe2Image

For the recipe2image generation, we have a two-fold goal: the generated images must be realistic, and their food-class identity is preserved. Taking the recipe encoding as input, we use a generator G_{r2i} to generate an image. A discriminator D_{r2i} is then used to distinguish whether the generated image is real or fake. At the same time, a Food-Classifier C_{r2i} is used to predict which food category the generated image belongs to. During training, the discriminator D_{r2i} and classifier C_{r2i} receive both real images and the generated fake images as input. Specifically, we borrow the idea of StackGAN [46] to guarantee the quality of generated food images, and the generator G_{r2i} is conditioned on the recipe feature $R = \mathbf{FC}(\mathbf{E}_R(\mathbf{r}))$. The adversarial objective is formulated as:

$$\begin{aligned}\mathcal{L}_{G_{r2i}} =& \mathbb{E}_{\mathbf{i}\sim p_{image}}[\log D_{r2i}(\mathbf{i})] + \\ & \mathbb{E}_{\mathbf{r}\sim p_{recipe}}[\log(1 - D_{r2i}(G_{r2i}(\mathbf{FC}(\mathbf{E}_R(\mathbf{r})))))]\end{aligned} \tag{4}$$

While the adversarial loss is able to generate realistic images, it does not guarantee translation consistency. Therefore, we use a food-category classifier which encourages the generator to use the recipe features to generate food items in the appropriate corresponding food-category. The objective is a cross-entropy loss denoted by $\mathcal{L}_{C_{r2i}}$. Combining these two objectives, our optimization is formulated as:

$$\min_{G_{r2i}, C_{r2i}, \mathbf{E}_R, \mathbf{FC}} \max_{D_{r2i}} \mathcal{L}_{r2i} = \mathcal{L}_{G_{r2i}} + \mathcal{L}_{C_{r2i}}$$

3.4.2 Image2Recipe

To achieve the image2recipe translation, we aim to recover the ingredients in the food image. This is done by applying a multi-label classifier on the visual features V, which will predict the ingredients in the food image. The multi-label objective is denoted as $\mathcal{L}_{G_{i2r}}$ (as it is in some sense generating the ingredients from a given image). Like in the case of recipe2image, we also maintain translation consistency by ensuring the features can be classified into the correct food category, using cross-entropy loss $\mathcal{L}_{C_{i2r}}$. The optimization is formulated as:

$$\min_{G_{i2r}, C_{i2r}, \mathbf{E}_I, \mathbf{FC}} \mathcal{L}_{i2r} = \mathcal{L}_{G_{i2r}} + \mathcal{L}_{C_{i2r}}$$

4. Experiments

4.1. Dataset and Evaluation Metrics

We evaluate the effectiveness of our proposed method on Recipe1M dataset [40], one of the largest collection of public cooking recipe data along with food images. It comprises over 1m cooking recipes and 800k food images. The authors assigned a class label to each recipe based on the recipe titles. This resulted in 1,047 food-classes (which provide relevant information for the cross-modal translation consistency). The authors also identified 4,102 frequently occurring unique ingredients. We adopt the the original data splits [40] using 238,999 image-recipe pairs for training, 51,119 pairs for validation and 51,303 pairs for testing.

We evaluate the model using the same metrics as prior work [40, 6]. Specifically, we first sample 10 different subsets of 1,000 pairs (1k setup), and 10 different subsets of 10,000 (10k setup) pairs in the testing set. Then, we consider each item in a modality as a query (for example, an image), and we rank instances in the other modality (example recipes) according to the L2 distance between the embedding of query and that of candidate. Using the L2 distance for retrieval, we evaluate the performance using standard metrics in cross-modal retrieval task. For each test-subset we sampled before (1k and 10k), we compute the median retrieval rank (MedR). We also evaluate the Recall Percentage at top K (R@K), i.e., the percentage of queries for which the matching item is ranked among the top K results.

4.2. Implementation Details

Our image encoder $\mathbf{E}_V$ is initialized as a ResNet-50 [18] pretrained on ImageNet [12], and gives a 1024-dimensional feature vector. A recipe comprises a set of instructions and a set of ingredients. The recipe encoder $\mathbf{E}_R$ is a hierarchical LSTM [21] to encode the instructions (where the word-level embedding is obtained from a pretrained skip-thought algorithm [29]), and a bidirectional-embedding to encode the ingredients (where the ingredient embeddings are obtained from word2vec) [31]. Both embeddings are then concatenated and go through a fully-connected layer to give a 1024-dimensional feature vector. During image generation, from recipe2image, our generator G_{r2i} generated images of size $128 \times 128 \times 3$. We set λ_1 and λ_2 in Eq. (1) to be 0.005 and 0.002 respectively. The model was trained using Adam [28] with the batch size of 64 in all our experiments. Initial learning rate is set as 0.0001, and momentum is set as 0.999.

4.3. Baselines

We compare against several state-of-the-art baselines:

CCA [23]: Canonical Correlation Analysis is one of the most widely-used models for learning a common embedding from different feature spaces. CCA learns two linear projections for mapping text and image features to a common

Table 1: **Main Results.** Evaluation of performance of ACME compared against the baselines. The models are evaluated on the basis of MedR (lower is better), and R@K (higher is better).

Size of test-set		Image to recipe retrieval				Recipe to image retrieval			
	Methods	**medR** ↓	**R@1** ↑	**R@5** ↑	**R@10** ↑	**medR** ↓	**R@1** ↑	**R@5** ↑	**R@10** ↑
1k	CCA [23]	15.7	14.0	32.0	43.0	24.8	9.0	24.0	35.0
	SAN [8]	16.1	12.5	31.1	42.3	-	-	-	-
	JE [40]	5.2	25.6	51.0	65.0	5.1	25.0	52.0	65.0
	AM [9]	4.6	25.6	53.7	66.9	4.6	25.7	53.9	67.1
	AdaMine [6]	1.0	39.8	69.0	77.4	1.0	40.2	68.1	78.7
	ACME (ours)	**1.0**	**51.8**	**80.2**	**87.5**	**1.0**	**52.8**	**80.2**	**87.6**
10k	JE [40]	41.9	-	-	-	39.2	-	-	-
	AM [9]	39.8	7.2	19.2	27.6	38.1	7.0	19.4	27.8
	AdaMine [6]	13.2	14.9	35.3	45.2	12.2	14.8	34.6	46.1
	ACME (ours)	**6.7**	**22.9**	**46.8**	**57.9**	**6.0**	**24.4**	**47.9**	**59.0**

space that maximizes feature correlation.

SAN [8]: Stacked Attention Network (SAN) considers ingredients only (and ignores recipe instructions), and learns the feature space between ingredients and image features through a two-layer deep attention mechanism.

JE [40]: uses pairwise cosine loss to find a joint embedding between different modalities. They attach a classifier to the embedding and predict the food category (1,047 classes).

AM [9]: In addition to triplet loss, AM uses an attention mechanism over the recipe, which is applied at different levels (title, ingredients and instructions). Compared with JE, AM uses additional information from food title.

AdaMine [6]: uses a double triplet loss where the triplet loss is applied to both learning a joint embedding and classification of the embedding into appropriate food category. They also integrate the adaptive learning schema which performs adaptive mining for significant triplets.

4.4. Quantitative Results

4.4.1 Main Results

We show the performance of ACME for cross-modal retrieval against the baselines in Table 1. In retrieval tasks on both 1k, and 10k test datasets, our proposed method ACME outperforms all the baselines across all metrics. On the 1k test dataset, ACME achieves a perfect median rank of 1.0 in both image to recipe and recipe to image retrieval tasks, matching the performance of the current state-of-the-art best method. When we shift to the larger 10k test dataset, where retrieval becomes much more difficult, ACME achieves a Median Rank of less than 6.7 and 6.0 respectively for the two retrieval tasks, which is about 50% lower than the current state-of-the-art method. The performance of ACME for Recall@K is significantly superior to all the baselines by a substantial margin. On the whole, the performance of ACME is shown to be very promising, beating all the state-of-the-art methods across all the metrics consistently.

4.4.2 Ablation Studies

We also conduct extensive ablation studies to evaluate the contributions of each of our proposed components in the ACME framework. Specifically, we first evaluate the performance of most basic version of the model with a basic triplet loss **(TL)**. We incrementally add more components: first, we evaluate the gains from introducing Hard-sample Mining **(HM)**. Based on this model, we then add the Modality Alignment component **(MA)**. Based on the resultant model, we then add the cross-modal translation components to measure their usefulness. In particular we want to see the effect of each cross-modal translation **R2I** and **I2R**, and their combined effect. We also measure the effect of the full model with and without **MA**, and finally combine all the components in the ACME framework. We evaluate these components on image-to-recipe retrieval, and the results are shown in Table 2. In general, we observe that every of the proposed component adds positive value to the embedding model by improving the performance, and all of them work in a collaborative manner to give an overall improved performance.

4.5. Qualitative Results

Here, we visualize some results of ACME. We use the trained ACME model, and perform the two retrieval tasks (Recipe to Image and Image to Recipe) on the full 50k test dataset. Note that this is much harder than the 1k dataset presented in the previous section, and retrieving the correct instance (out of 50k possibilities) is difficult. Our main goal is to show, that despite this difficulty, the top retrieved items appear to be (qualitatively) good matches for the query. We first show the query instance in one modality, and the

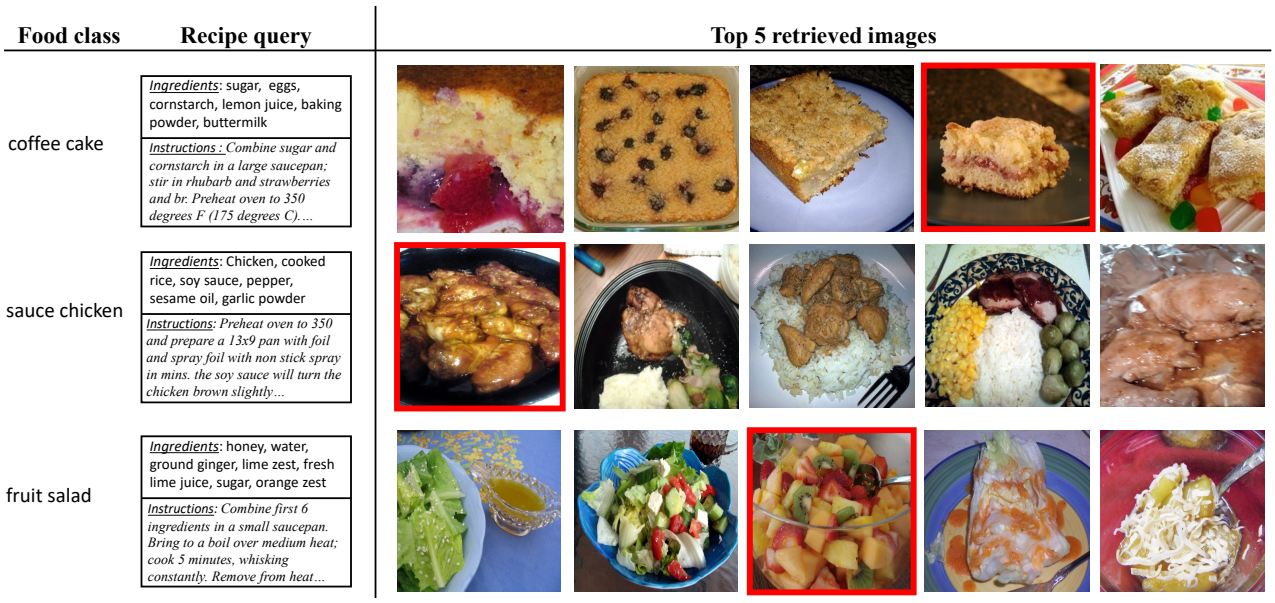

Figure 3: Analysis of Recipe to Image Retrieval on the full 50k test dataset. The first column denotes the food category, and the second column is the query recipe. The top 5 results retrieved by ACME are shown. The best match (i.e., the ground truth) is boxed in red. In most cases, the top retrieved images display similar concepts as the ground truth.

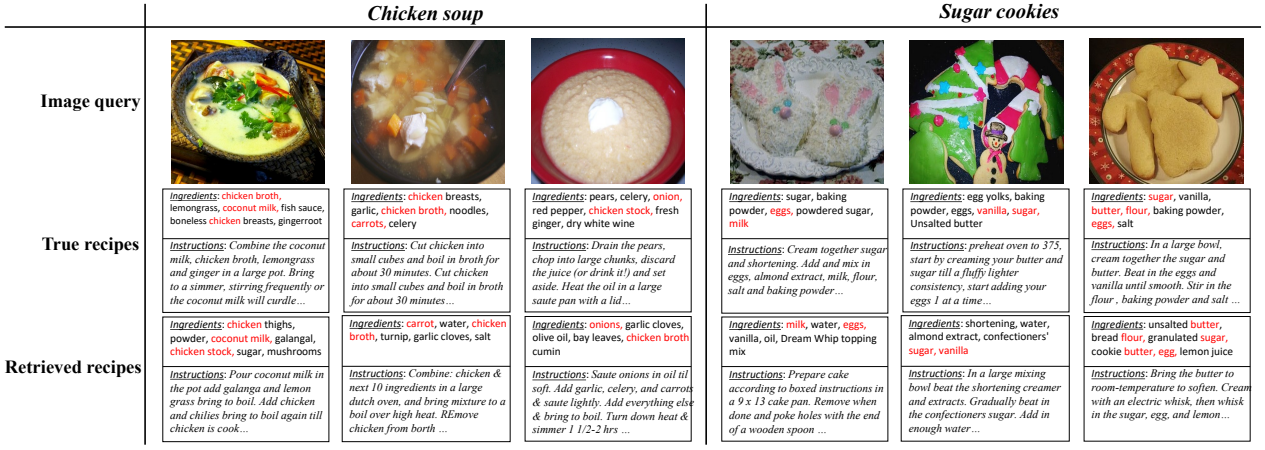

Figure 4: Analysis of Image to Recipe Retrieval on the full 50k test dataset. We consider two food-categories: *chicken soup* and *sugar cookies*. We show 3 different image-queries in each of these categories. Below each image is the true recipe, and below that is the top retrieved recipe by ACME. We can observe that our retrieved recipe has several common ingredients with the true recipe.

corresponding ground truth in the other modality. We then retrieve the top results in the second modality, which could show the comparison with the true instance. Even though, often we are not able to obtain the ground truth, ACME retrieves instances that are very similar to this ground truth.

4.5.1 Recipe to Image Retrieval Results

We show some examples of recipe to image retrieval in Figure 3. We consider 3 food-classes: *coffee cake*, *sauce chicken*, and *fruit salad*, and pick up a recipe to query from each of these classes. ACME is then used to rank the images, and then we look at the top 5 retrieved results. In all cases, the retrieved images are visually very similar to the ground truth image. For example, the images retrieved in case of *coffee cake*, all resemble cakes; images retrieved for *sauce chicken*, all have some form of cooked chicken, couple of them with rice (the ground truth does not have rice, but the ingredients in the query recipe use cooked rice) ; images retrieved for *fruit salad* appear to have fruits in all of them.

Table 2: **Ablation Studies**. Evaluation of benefits of different components of the ACME framework. The models are evaluated on the basis of MedR (lower is better), and R@K (higher is better).

Component	MedR	R@1	R@5	R@10
TL	4.1	25.9	56.4	70.1
TL+HM	2.0	47.5	76.2	85.1
TL+HM+MA	1.9	48.0	77.3	85.5
TL+HM+MA+R2I	1.4	50.0	77.8	85.7
TL+HM+MA+I2R	1.8	49.3	78.4	86.1
TL+HM+R2I+I2R	1.6	49.5	77.9	85.5
All	**1.0**	**51.8**	**80.2**	**87.5**

This suggests that the common space learned by our structure has done a good job of capturing semantic information across the two modalities.

4.5.2 Image to Recipe Retrieval Results

We show the image to recipe retrieval results in Figure 4. We consider two food categories: *chicken soup*, and *sugar cookies*, and use 3 images in each category as a query image. We retrieve the top recipe for each image query, and compare it with the true recipe corresponding to that image. In the case of *chicken soup*, it is often hard to view the chicken in the food image (as it maybe submerged in the soup), yet the ingredients in the recipe retrieved based on the ACME embedding contains chicken. We also observe several common ingredients between the true recipe and the retrieved recipe for the *sugar cookies* images. Such a performance can have several interesting applications for food-computing. For example, a user can take a food image, retrieve the recipe, and estimate the nutrients and the calories for that food image.

4.5.3 Cross-Modal Translation Consistency

An interesting by-product of the proposed ACME framework is the cross-modal translation. These translation components were primarily used to add constraints to obtain a better semantic embedding for our primary retrieval task. We do not claim that these components are robust and generate very good translation (as the food categories are noisy, generating real images is difficult, classification of food into 1,047 categories and multi-label classification of ingredients into over 4,102 categories is very difficult). However, the cross-modal translation can sometimes obtain very interesting results.

For example, in Figure 5 we see a recipe2image translation. Given a recipe query, we retrieve a real food image from the dataset, and compare it with the image generated by the r2i component of ACME. For two instances of *ice cream*, we can see that the generated images show incredible resemblance to the real food images. Similarly, we look at the ingredient prediction from image to see the results of the i2r component in Figure 6. The query image is a *pork chop*. We can see several correctly predicted ingredients for the query image. In fact, it is wholly plausible that using the ingredients predicted, it would be possible to cook the dish to obtain a very similar image as the query image.

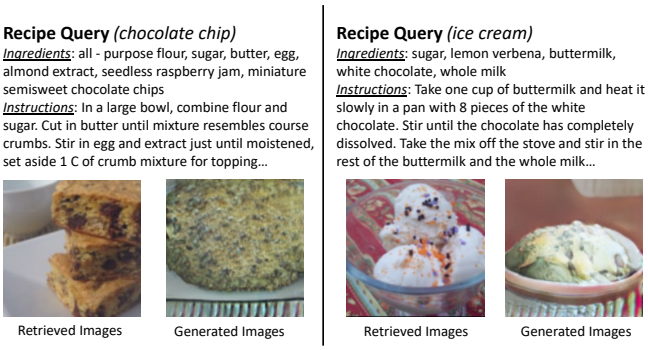

Figure 5: Visualization of the retrieved food images and generated images (generated by r2i translation component) for a query recipe.

True Ingredients:
Tomatoes, cayenne pepper, pork chop, garlic clove, salt
Retrieved Ingredients:
boneless sirloin pork chops, black pepper, chili sauce, garlic powder, salt
Predicted Ingredients:
browning sauce, pork rib chops, cracked pepper, dried parsley flakes, garlic scapes

Figure 6: Visualization of the retrieved ingredients and predicted results (predicted by i2r translation component) for query image *pork chops*.

5. Conclusion

In this paper, we have proposed an end-to-end framework **ACME** for learning a joint embedding between cooking recipes and food images, where we are the first to use adversarial networks to guide the learning procedure. Specifically, we proposed the usage of hard sample mining, used an adversarial loss to do modality alignment, and introduced a concept of cross-modal translation consistency, where we use the recipe embedding to generate an appropriate food image, and use the food image embedding to recover the ingredients in the food. We conducted extensive experiments, and ablation studies, and achieved state-of-the-art results in the Recipe1M dataset for cross-modal retrieval.

Acknowledgments

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.

References

[1] Palakorn Achananuparp, Ee-Peng Lim, and Vibhanshu Abhishek. Does journaling encourage healthier choices?: Analyzing healthy eating behaviors of food journalers. In *Proceedings of the 2018 International Conference on Digital Health*, pages 35–44. ACM, 2018.

[2] Palakorn Achananuparp, Ee-Peng Lim, Vibhanshu Abhishek, and Tianjiao Yun. Eat & tell: A randomized trial of random-loss incentive to increase dietary self-tracking compliance. In *Proceedings of the 2018 International Conference on Digital Health*, pages 45–54. ACM, 2018.

[3] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[4] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

[5] Carl A Batt. Food pathogen detection. *Science*, 316(5831):1579–1580, 2007.

[6] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *ACM SIGIR*, 2018.

[7] Minsuk Chang, Léonore V Guillain, Hyeungshik Jung, Vivian M Hare, Juho Kim, and Maneesh Agrawala. Recipescape: An interactive tool for analyzing cooking instructions at scale. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 451. ACM, 2018.

[8] Jingjing Chen, Lei Pang, and Chong-Wah Ngo. Cross-modal recipe retrieval: How to cook this dish? In *International Conference on Multimedia Modeling*, pages 588–600. Springer, 2017.

[9] Jing-Jing Chen, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua. Deep understanding of cooking procedure for cross-modal recipe retrieval. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1020–1028. ACM, 2018.

[10] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Learning cnn-based features for retrieval of food images. In *International Conference on Image Analysis and Processing*, pages 426–434. Springer, 2017.

[11] Antonia Creswell and Anil Anthony Bharath. Adversarial training for sketch retrieval. In *European Conference on Computer Vision*, pages 798–809. Springer, 2016.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

[13] Takumi Ege and Keiji Yanai. Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 367–375. ACM, 2017.

[14] Ahmed Fadhil. Can a chatbot determine my diet?: Addressing challenges of chatbot application for meal recommendation. *arXiv preprint arXiv:1802.09100*, 2018.

[15] Giovanni Maria Farinella, Dario Allegra, Marco Moltisanti, Filippo Stanco, and Sebastiano Battiato. Retrieval and classification of food images. *Computers in biology and medicine*, 77:23–39, 2016.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[20] José Luis Hernández-Hernández, Mario Hernández-Hernández, Severino Feliciano-Morales, Valentín Álvarez-Hilario, and Israel Herrera-Miranda. Search for optimum color space for the recognition of oranges in agricultural fields. In *International Conference on Technologies and Innovation*, pages 296–307. Springer, 2017.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[22] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *ICML*, 2018.

[23] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[24] Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, and Yueting Zhuang. Deep compositional cross-modal learning to rank via local-global alignment. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 69–78. ACM, 2015.

[25] Taichi Joutou and Keiji Yanai. A food image recognition system with multiple kernel learning. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 285–288. IEEE, 2009.

[26] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[27] Yoshiyuki Kawano and Keiji Yanai. Food image recognition with deep convolutional features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 589–593. ACM, 2014.

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

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

[30] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma. Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. In *International Conference on Smart Homes and Health Telematics*, pages 37–48. Springer, 2016.

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[32] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *arXiv preprint arXiv:1808.07202*, 2018.

[33] Ole G Mouritsen, Lars Duelund, Ghislaine Calleja, and Michael Bom Frøst. Flavour of fermented fish, insect, game, and pea sauces: garum revisited. *International journal of gastronomy and food science*, 9:16–28, 2017.

[34] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1899–1907. IEEE, 2017.

[35] Ferda Ofli, Yusuf Aytar, Ingmar Weber, Raggi al Hammouri, and Antonio Torralba. Is saki# delicious?: The food perception gap on instagram and its relation to health. In *Proceedings of the 26th International Conference on World Wide Web*, pages 509–518. International World Wide Web Conferences Steering Committee, 2017.

[36] Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2372–2385, 2018.

[37] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. Cm-gans: Cross-modal generative adversarial networks for common representation learning. *arXiv preprint arXiv:1710.05106*, 2017.

[38] Faisal Rehman, Osman Khalid, Kashif Bilal, Sajjad A Madani, et al. Diet-right: A smart food recommendation system. *KSII Transactions on Internet & Information Systems*, 11(6), 2017.

[39] Sina Sajadmanesh, Sina Jafarzadeh, Seyed Ali Ossia, Hamid R Rabiee, Hamed Haddadi, Yelena Mejova, Mirco Musolesi, Emiliano De Cristofaro, and Gianluca Stringhini. Kissing cuisines: Exploring worldwide culinary habits on the web. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1013–1021. International World Wide Web Conferences Steering Committee, 2017.

[40] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. *Training*, 720(619-508):2, 2017.

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[42] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1):207–218, 2014.

[43] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 154–162. ACM, 2017.

[44] Wei Wu, Jun Xu, and Hang Li. Learning similarity function between objects in heterogeneous spaces. *Microsoft Research Technique Report*, 2010.

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

[46] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the ICCV*, pages 5907–5915, 2017.