MCEN: Bridging Cross-Modal Gap between Cooking Recipes and Dish Images with Latent Variable Model

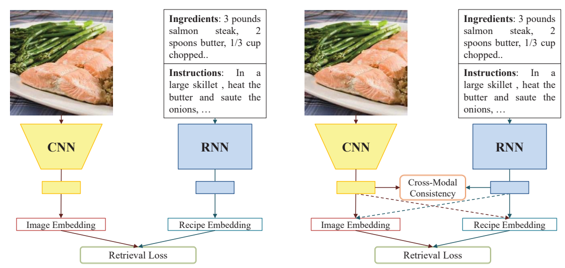
Han Fu[†‡] Rui Wu[†] Chenghao Liu[§] Jianling Sun[†‡*]
[†]Zhejiang University, Hangzhou, China
[‡]Alibaba-Zhejiang University Joint Institute of Frontier Technologies, China
[§]Singapore Management University, Singapore
{11821003, tactic, sunjl}@zju.edu.cn, twinsken@gmail.com

Abstract

Nowadays, driven by the increasing concern on diet and health, food computing has attracted enormous attention from both industry and research community. One of the most popular research topics in this domain is Food Retrieval, due to its profound influence on health-oriented applications. In this paper, we focus on the task of cross-modal retrieval between food images and cooking recipes. We present Modality-Consistent Embedding Network (MCEN) that learns modality-invariant representations by projecting images and texts to the same embedding space. To capture the latent alignments between modalities, we incorporate stochastic latent variables to explicitly exploit the interactions between textual and visual features. Importantly, our method learns the cross-modal alignments during training but computes embeddings of different modalities independently at inference time for the sake of efficiency. Extensive experimental results clearly demonstrate that the proposed MCEN outperforms all existing approaches on the benchmark Recipe1M dataset and requires less computational cost.

1. Introduction

Food is the paramount necessity of human life. As the saying goes, *we are what we eat*, food not only provides energy for life activities, but also plays a significant role in affecting human identity, social formation, history, and culture inheritance [19]. In our daily life, food is intricately linked to people's convention, lifestyle, health and social activities. Nowadays, with the development of Internet and mobile applications, sharing recipes and food images on social platforms has become a widespread trend [43]. Due to the massive amounts of data resource online, food computing has become a popular field, inciting numerous machine learning tasks such as ingredient recognition [38, 23], food image retrieval [54] and recipe recommendation [53, 49]. Among the research topics, Image-to-Recipe learning (im2recipe) is one of the most important problems due to its profound influence on health-oriented applications [40]. For instance, food-health analysis applications are required to predict detailed nutrition contents and calorie information from food images, and a recipe-retrieval system is a necessary solution on this scenario.



(a) Prior Work (b) Our Work

Figure 1. **A comparison between prior work (a) and the proposed MCEN (b).** We learn modality-consistent embeddings by capturing the interactions between images and recipes via latent variables. The dotted lines represents that the joint information is only used during training. At inference time, the embeddings are computed independently.

Im2recipe is a challenging task since it involves highly variant foods images and expatiatory textual recipes. A typical recipe consists of a list of ingredients and cooking instructions which may not directly align with the appearance of the corresponding food image. Typically, recent efforts have formulated im2recipe as a cross-modal retrieval problem [48, 37, 6, 62], to align matching recipe-image pairs in a shared latent space with retrieval learning approaches. Concretely, prior work builds two independent networks to encode textual recipes (ingredients and cooking instructions)

*Corresponding author: Jianling Sun.

and food images into embeddings respectively. And the retrieval loss object is learned to gather matching pairs and differentiate dissimilar items. Though existing methods are expressive and powerful, there remain two major concerns. 1) Current systems encode images and texts with two different networks independently. However, such independence brings barriers between modalities, resulting in obstacles to discover latent semantic alignments across modalities. Consequently, such approaches thus could suffer from polysemous instances [51]. 2) The recipe representations are obtained based on fixed pre-trained skip-thought vectors [28], leading to highly diversities between textual and image feature spaces.

To alleviate such limitations, we strive to take a step towards capturing joint information of different modalities and injecting the cross-modal alignments into the embedding learning processes on both sides. We introduce **M**odality-**C**onsistent **E**mbedding **N**etwork (MCEN) which learns joint cross-modal representations for textual recipes and dish images. The major idea is to exploit the interactions between visual and textual features explicitly and share the cross-modal information to the embedding spaces of both modalities with stochastic latent variable models. The stochastic variable is leveraged to capture the latent correlations between modalities during training, while the embeddings can still be calculated independently at test time for high efficiency and flexibility. Moreover, The randomness introduced by latent variables is also beneficial for handling polysemous instances where one recipe corresponds with multiple images.

In a nutshell, the **main contribution** of this work is threefold:

- We propose a novel cross-modal retrieval framework to obtain modality-consistent embeddings by explicitly capturing the correlations between recipes and food images with latent variables.
- We exploit the latent alignments during training with cross-modal attention mechanism and replace it with prior condition at inference time for efficiency.
- We propose a task-specific encoder for textual recipes based on hierarchical attentions, which cannot only adapt to the interaction with images, but also simplify and accelerate the training and inference procedure.

We conduct experiments on the challenging benchmark Recipe1M [48] and the results demonstrate that our model significantly outperforms all state-of-the-art approaches on the cross-modal recipe retrieval problem and requires less computational overhead.

2. Related Work

Computational Cooking. Food and cooking are essential parts of human life, which are closely relevant to health [53], social activities, bromatology, dietary therapy and culture [19], etc., profoundly affecting the quality of life. Therefore, research involving cooking recipes has drawn considerable attention. Food and cooking provide rich attributes on multiple channels, including both visual content (e.g., dish pictures) and texts (e.g., dish descriptions and cooking instructions). Current literature leverages the attributes in various ways. Typically, recent examples in computer visions are food classification and recognition [7, 34, 31, 61, 23], and retrieval of captions [14, 9], ingredients [8, 9] or recipe instructions [6, 48, 41, 42] according to dish images, while researchers from natural language processing community usually focus on such applications as recipe recommendation [53, 49], aligning instructions with video and speech [35], recipe texts generation from flow graph [44], workflow generation from recipe texts [58], cooking action tracking [4], recipe representation [36], checklist recipe generation [24] and recipe-based question answering [57, 36]. Moreover, there is also some work using machine learning approaches to connect health with food attributes, such as prediction of nutrient [29] or energy [39], and healthy recipe recommendation [13, 53, 60]. All these efforts contribute to the prosperity of food computation and understanding, bridging the gap between machine learning applications and people's daily life.

Recent introductions of large-scale food-related datasets have further accelerated the research improvements on food understanding. Considering the application purpose, the datasets can be categorized into two groups: food recognition [3, 38] and cross-modal recipe retrieval [48, 37, 41, 42, 7, 48]. We focus on recipe retrieval task in this paper, aiming at retrieval relevant cooking recipes with respect to the image query and vise versa. Typically, the datasets for retrieval generally incorporate both food images and other information such as ingredients, structured cooking instructions and flavor attributes. Among the datasets, Recipe1M [48] is the most well curated large-scale dataset with preprocessed English textual information and we evaluate the effectiveness of our method on it in this paper.

Text-Image Retrieval. Our work is related to current approaches on multi-modal retrieval task, where the key problem is to measure the similarity between a text and an image. The major challenge of this issue lies in the modality-gap, which means that the feature spaces of different modalities largely diverse from each other. Text-image retrieval is at meeting point between computer vision and natural language communities, attracting research attentions over decades [32]. Traditional approaches formulate this issue as either a language modeling task [27] or a correlation maximization problems [46, 18] using canonical correlation analysis (CCA) [21]. Recently, many efforts have been made to build end-to-end retrieval systems leveraging deep learning methods [52, 1, 16, 59, 45]. Another avenue is to improve the triplet loss with hard negative mining [50], such as [17, 55, 15].

Despite of the progress, the above approaches encode different modalities into independent feature spaces, suffering from modality gap between heterogenetic contents. To address this issue, recent works incorporate attention mechanism to capture the latent alignment relationships between words and different image regions [22, 30, 33, 56]. Though expressive, these methods require massive computational overhead during inference since the cross-modal attention scores between a query and each item in the reference set need to be calculated, limiting the scalability to large-scale retrieval scenario. In this paper, we leverage latent variables to incorporate cross-modal attention mechanism into retrieval tasks during training but maintain independent calculations for different modalities respectively at inference time.

Image-to-recipe is a newly proposed task and is formulated as a cross-modal learning task by recent efforts [8, 48], to retrieve the relevant recipes based on image queries. Following these settings, several inspiring methods have been introduced to improve the retrieval performance by using such techniques as additional textual feature [9], semantic information [6] and adversarial learning [62, 54].

3. Modality-Consistent Embedding Network

3.1. Overview

In this section we introduce the methodology of the proposed Modality-Consistent Embedding Network (MCEN).

Problem Formulation. The aim of the proposed framework is to measure the similarity between food images and the relevant textual recipes. Formally, denote $\{\mathbf{v}^i, \mathbf{r}^i\}_{i=1}^N$ as a set of N image-recipe pairs where an image $\mathbf{v}^i \in \mathbf{V}$ and a recipe $\mathbf{r}^i \in \mathbf{R}$. The notations $\mathbf{V}$ and $\mathbf{R}$ denote the visual and recipe spaces. It should be noted that one recipe corresponding to multiple images is allowed. A recipe $\mathbf{r}^i$ consists of a set of ingredients $\mathbf{X}^{ing,i}$ and a list of cooking instructions $\mathbf{X}^{ins,i}$. An image $\mathbf{v}^i$ contains the appearance of a completed dish. Importantly, the ingredients and cooking instructions of a recipe may not directly align with the appearance of the matching image, which brings additional heterogeneity challenge compared to traditional cross-modal retrieval tasks.

Considering the information gap between modalities, we set our target to learn the mapping functions from observed data to the embedding distributions as $\mathbf{V} \rightarrow \mathbf{E}^v$ and $\mathbf{R} \rightarrow \mathbf{E}^r$, where $\mathbf{E}^v \in \mathbb{R}^d$ and $\mathbf{E}^r \in \mathbb{R}^d$ denote the distributions of d-dimensional image embedding and recipe embedding respectively, so that a picture is closer to the corresponding recipe than any other image in the latent space.

Architecture. The architecture of MCEN is illustrated in Figure 2. The system consists of three major modules: a recipe encoder, an image encoder and an embedding learning component for modality-consistent space modeling. Through the training flow, the visual feature is extracted by feeding the food picture $\mathbf{v}^i$ to the CNN-based image encoder. Meanwhile, the high-level representations of instructions and ingredients are obtained by hierarchical attention-based RNN encoders. Then these representations are then fed to cross-modal attention components to exploit the interactions between images and texts. The cross-modal correlations are then leveraged to estimate the posterior distributions of embeddings with neural variational inference [26, 47]. With this method, we can discriminate training and inference process so as to reduce cross-modal computation at prediction time. To keep modality consistency, we align the distributions of latent representations by minimizing the KL-divergence of priors of different modalities. Finally, the latent representations sampled from the posterior distributions are passed to feed-forward layers to obtain the final embeddings of images and recipes respectively. The entire model is trained end-to-end with retrieval learning object.

The major novelty of MCEN comes from the incorporation of cross-modal correlation modeling with latent variables. MCEN captures the latent alignment relationships between images and texts during training while at inference time we do not require cross-modal attention since the posterior distribution is replaced by the prior during test. Though there exists prior work that focuses on modeling correlations between modalities [30, 33], these approaches come with high computation overhead since the alignment score between a query and each reference instance needs computing as many times as the size of reference set [51]. Conversely, MCEN obtains embeddings of different modalities independently during inference, which significantly reduces the computational overhead. Moreover, almost all prior methods require fixed pre-trained instruction vectors for recipes while parameters for image encoding are updated with respect to the retrieval object. The isomerism in training process leads to a diversity between feature spaces of images and recipes. In this work, the architecture of MCEN recipe encoder is quite different from prior systems and can be trained end-to-end from scratch.

3.2. Image Encoder

Given a food picture $\mathbf{v}$, the image encoder is responsible to extract the abstract features of the input. Different from previous methods, we use the output of the last residual block (res5c) of ResNet-50 [20] which consists of $7 \times 7 = 49$ columns of 2048 dimensional convolutional outputs, denoted by $\mathbf{H}^v = (\mathbf{h}_1^v, \mathbf{h}_2^v, \cdots, \mathbf{h}_{49}^v,)$. To obtain the representation for the image hidden states, we propose to use an attention layer, which estimates the importance of each hidden vector. Since a dish image may contain multi-

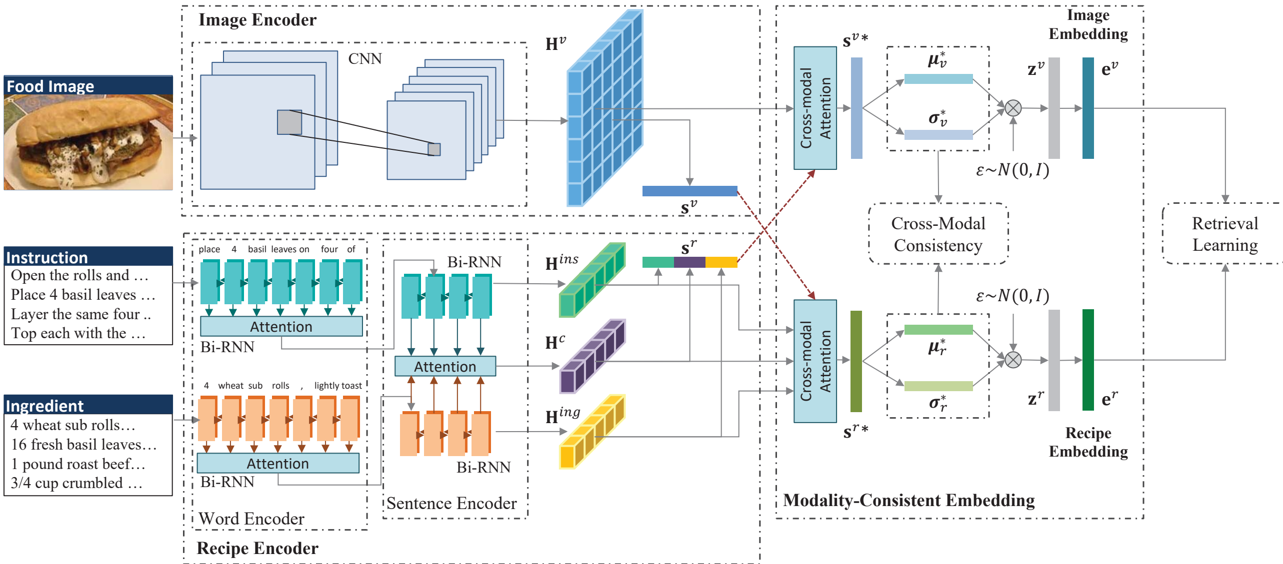


Figure 2. **The architecture and training flow of MCEN.** The red dotted lines denote that the cross-modal attention components only work during training and are omitted at testing time. The system is comprised of three major components: a recipe encoder, an image encoder, and a modality-consistent embedding component. The interaction between images and texts is captured with latent variables and shared by both latent spaces.

ple objects that are not relevant to the recipe (i.e., forks and flowers), the aim of attention model is to force the encoder to focus more on regions that may contribute to the retrieval object.

Formally, the image representation $\mathbf{s}^v$ is calculated with the weighted summation of convolutional states as:

$$\mathbf{s}^v = \sum_{i=1}^{49} \alpha_i^v \mathbf{h}_i^v, \quad (1)$$

where α_i^v is the attention score at position i, representing the importance of this region, calculated by:

$$\alpha_i^v = \mathrm{softmax}(\mathbf{v}_v^\top \tanh(\mathbf{W}_v \mathbf{q}_v + \mathbf{U}_v \mathbf{h}_i^v)), \quad (2)$$

where $\mathbf{W}_v$, $\mathbf{U}_v$ and $\mathbf{v}_v$ are trainable matrices and vector. $\mathbf{q}_v$ is the attention query vector. Here, it is a trainable vector initialized from scratch. For sake of writing convenience, we call such attention layer as *Attention Pooling* and the input annotations ($\mathbf{H}^v$) as *Attention Context*.

3.3. Recipe Encoder

In the recipe branch, ingredients and instructions are encoded separately with similar networks. Since the ingredients or instructions of a recipe usually comprise multiple sentences, we use a hierarchical attention-based model to extract textual features. Each instruction/ingredient is first fed to a word-level bi-directional recurrent neural network (bi-RNN) with gated recurrent unit (GRU) [10] and the final word-level representations are calculated with attention pooling mechanism (Equation 1-2) where the RNN hidden states are used as the attention contexts. Denote $\mathbf{H}^{ins} = (\mathbf{h}_1^{ins} \cdots \mathbf{h}_m^{ins})$ and $\mathbf{H}^{ing} = (\mathbf{h}_1^{ing} \cdots \mathbf{h}_n^{ing})$ as the feature sequences of instructions and ingredients respectively, where m and n are the numbers of instructions and ingredients of a recipe, and each element $\mathbf{h}_t^{ins}/\mathbf{h}^{ing}$ is the abstract representation of an instruction/ingredient. To model the correlations between instructions and ingredients, we employ the attention-based RNN decoder [2], which takes $\mathbf{H}^{ins}$ as the sequential input and $\mathbf{H}^{ing}$ as the contexts respectively. The output of the RNN decoder is denoted as $\mathbf{H}^c = (\mathbf{h}_1^c, \cdots, \mathbf{h}_m^c)$ which contains the joint information of both instructions and ingredients. Then, $\mathbf{H}^c$, $\mathbf{H}^{ins}$ and $\mathbf{H}^{ing}$ are fed to independent sentence-level bi-RNNs and attention pooling layers to obtain the sentence-level representations, denoted as $\mathbf{s}^c$, $\mathbf{s}^{ins}$, and $\mathbf{s}^{ing}$ respectively. The final feature representation of the recipe is obtained by concatenating the three sentence representations as:

$$\mathbf{s}^r = [\mathbf{s}^{c\top}, \mathbf{s}^{ins\top}, \mathbf{s}^{ing\top}]^\top. \quad (3)$$

3.4. Modality-Consistent Embedding

It is challenging to align feature representations of multiple modalities when the features are extracted with independent networks. To alleviate this issue, we incorporate latent variables to capture the interactions between modalities. This method converts the embedding computation into a generative process. Taking the image side for instance, the probability to generate a specific embedding $\mathbf{e}^v$ for a given image $\mathbf{v}$ is modeled as:

$$p(\mathbf{e}^v|\mathbf{v}) = p(\mathbf{e}^v|\mathbf{z}^v, \mathbf{v})p(\mathbf{z}^v|\mathbf{v}), \tag{4}$$

where the latent vector $\mathbf{z}^v$ is assumed to capture the correlations between $\mathbf{v}$ and the corresponding recipe $\mathbf{r}$. The posterior of $\mathbf{z}^v$ should hence be conditioned on both the recipe $\mathbf{r}$ and image $\mathbf{v}$, denoted as $p(\mathbf{z}^v|\mathbf{v}, \mathbf{r})$. The prior of latent variables is usually formulated as a standard Gaussian distribution, which may reduce the effectiveness in generation [11]. Here we propose to estimate the prior distribution with a neural network model that jointly learns the prior knowledge and excavates cross-modal alignments based on single modality, denoted as $p(\mathbf{z}^v|\mathbf{v})$. To simplify the generative process, both prior and posterior distributions for latent variables are assumed to be Gaussian distributions. Concretely, the generative story is as follows. We sample a latent variable $\mathbf{z}^v$ from the prior Gaussian distribution as:

$$\mathbf{z}^v|\mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}_v, \mathrm{diag}(\boldsymbol{\sigma}_v^2)) \tag{5}$$

$$\boldsymbol{\mu}_v = \mathbf{W}_\mu^v \mathbf{s}^v + \mathbf{b}_\mu^v \tag{6}$$

$$\boldsymbol{\sigma}_v = \mathrm{softplus}(\mathbf{W}_\sigma^v \mathbf{s}^v + \mathbf{b}_\sigma^v), \tag{7}$$

where $\mathbf{W}_\mu^v$, $\mathbf{W}_\sigma^v$ and $\mathbf{b}_\mu^v$, $\mathbf{b}_\sigma^v$ are weight matrices and bias. Conditioned on the latent variable $\mathbf{z}^v$, we generate the final image embedding as:

$$\mathbf{e}^v = f_v(\mathbf{z}^v), \tag{8}$$

where f_v is a mapping function implemented as a one-layer neural network with tanh activation.

Estimation of Equation 4 can be challenging since the distributions are intractable. We leverage neural variational inference [26, 47] to optimize the evidence lowerbound (ELBO) as:

$$\mathbb{E}_{q(\mathbf{z}^v|\mathbf{v},\mathbf{r})}(\log p(\mathbf{e}^v|\mathbf{z}^v, \mathbf{v})) - D_{KL}(q(\mathbf{z}^v|\mathbf{v}, \mathbf{r})\|p(\mathbf{z}^v|\mathbf{v})), \tag{9}$$

where $D_{KL}(\cdot)$ is the Kullback-Leibler divergence and $q(\mathbf{z}^v|\mathbf{v}, \mathbf{r})$ is the approximate posterior, estimated as:

$$\mathbf{z}^v|\mathbf{v}, \mathbf{r} \sim \mathcal{N}(\boldsymbol{\mu}_v^*, \mathrm{diag}(\boldsymbol{\sigma}_v^{*2})) \tag{10}$$

$$\boldsymbol{\mu}_v^* = \mathbf{W}_\mu^{v*} \mathbf{s}^{v*} + \mathbf{b}_\mu^{v*} \tag{11}$$

$$\boldsymbol{\sigma}_v^* = \mathrm{softplus}(\mathbf{W}_\sigma^{v*} \mathbf{s}^{v*} + \mathbf{b}_\sigma^{v*}), \tag{12}$$

where $\mathbf{W}_\mu^{v*}$, $\mathbf{W}_\sigma^{v*}$ and $\mathbf{b}_\mu^{v*}$, $\mathbf{b}_\sigma^{v*}$ are trainable matrices and bias, which are independent from the prior model. The cross-modal representation $\mathbf{s}^{v*}$ is obtained with an attention pooling layer which takes the recipe representation $\mathbf{s}^r$ as the query vector and image region features $\mathbf{H}^v$ as the attention contexts. The lowerbound of the likelihood can be optimized by minimizing the triplet loss, formalized as:

$$\mathcal{L}_{ret}^v = [s(\mathbf{e}_a^v, \mathbf{e}_n^i) - s(\mathbf{e}_a^v, \mathbf{e}_p^i) + m]_+ \tag{13}$$

where $s(\cdot)$ expresses the cosine similarity between two vectors, and m is the margin of error. Subscripts p, n and a refer to positive, negative and anchor of a triplet respectively.

Cases are similar on the recipe side and the distinction lies in the calculation of cross-modal representation $\mathbf{s}^{r*}$ for posterior approximation $q(\mathbf{z}^r|\mathbf{v}, \mathbf{r})$. Here, we obtain $\mathbf{s}^{r*}$ with the similar manner to $\mathbf{s}^r$ (Equation 3) but replace the original trainable query vector with the image feature $\mathbf{s}^i$. Formally, the final retrieval learning object is defined as:

$$\mathcal{L}_{ret} + \alpha\mathcal{L}_{KL}, \tag{14}$$

where $\mathcal{L}_{ret}$ is the summation of the triplet losses for image-to-recipe and recipe-to-image retrieval, and α is a trade-off hyper-parameter. $\mathcal{L}_{KL}$ is the summation of the KL divergences on both sides:

$$\begin{aligned}\mathcal{L}_{KL} = D_{KL}(q(\mathbf{z}^v|\mathbf{v}, \mathbf{r})\|p(\mathbf{z}^v|\mathbf{v}))+\\ D_{KL}(q(\mathbf{z}^r|\mathbf{v}, \mathbf{r})\|p(\mathbf{z}^r|\mathbf{r})).\end{aligned} \tag{15}$$

Moreover, as discussed, we aim to align the distributions of both modalities. For this end, we simply push the prior embedding distributions of both modalities together by minimizing the following KL-divergence:

$$\mathcal{L}_{cos} = D_{KL}(p(\mathbf{z}^v|\mathbf{v})\|p(\mathbf{z}^r|\mathbf{r})). \tag{16}$$

3.5. Cross-Modal Reconstruction

Recent work [62, 54] has proved the effectiveness of reconstruction loss on cross-modal recipe retrieval, since it encourages the embedding of one modality covers the corresponding information of the other modality. However, such an approach introduces additional network parameters to reconstruct the original images and recipes, which are too cumbersome for training a retrieval system. In this work we propose a much conciser method for cross-modal reconstruction. Instead of recovering the entire information of the original inputs, we only reconstruct the latent representations with the learned embeddings as:

$$\mathbf{s}^{r'} = f_r^v(\mathbf{e}^v), \tag{17}$$

$$\mathbf{s}^{v'} = f_v^v(\mathbf{e}^r), \tag{18}$$

where f_r^v and f_v^v are mapping functions, implemented as two-layer neural networks. The formal reconstruction loss is formulated as:

$$\mathcal{L}_{rec} = P(\mathbf{s}^{r'}, \mathbf{s}^r) + P(\mathbf{s}^{v'}, \mathbf{s}^v), \tag{19}$$

where $P(\cdot)$ computes Pearson's correlation coefficient.

3.6. Training and Inference

The overall training object of MCEN is formulated as:

$$\mathcal{L} = \mathcal{L}_{ret} + \alpha\mathcal{L}_{KL} + \beta\mathcal{L}_{cos} + \gamma\mathcal{L}_{rec}, \tag{20}$$

where α, β and γ are hyper-parameters which balance the preference of different components. The entire model can be trained end-to-end with the reparameterization trick [26, 47]. During inference, the latent variables are fixed to the expectation of prior distribution to stabilize the retrieval performance.

4. Experiments

4.1. Settings

Dataset. The experiments are conducted on Recipe1M benchmark [48], a large-scale collection for recipe retrieval, including cooking instructions along with food images. The dataset consists of over 1M textual recipes and around 900K images. We use the same preprocessed samples provided by [48] and we finally obtain 238,399 matching pairs of recipes and images for training, 51,119 pairs for validation and 51,303 pairs for test respectively. Moreover, it should be noted that we do not incorporate the additional semantic labels used by prior work [48, 6, 62], such as food-classes and labels of commonly used ingredients.

Metrics. We utilize the same metrics as the prior work [48, 6, 62]. Concretely, we compute median rank (MedR) and recall rate at top K (R@K) on sampled subsets in the test partition to evaluate the retrieval performance. The sampling process is repeated for 10 times and the mean scores are reported. MedR measures the median retrieval rank position of true positives over all test samples, and the ranking position starts from 1. R@K refers to the percentage of queries for which matching instances are ranked among the top K results.

Implementation. For the image encoder, ResNet-50 [20] pretrained on ImageNet [12] is used as the initialization weight. On the recipes side, the dimension of all hidden states is set to 300. Different from prior work, we do not use pretrained word embeddings. The entire recipe encoder is trained from scratch and the trainable parameters are initialized uniformly between $[-0.02, 0.02]$.

The dimension of final embeddings and all hidden states for neural inference is 1024. The margin of error m is 0.3 and the hyper-parameters α, β, γ are set to 0.1, 0.002 and 0.008 respectively. The norm of gradient is clipped to be between $[-5, 5]$. We employ Adam solver [25] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ as the optimizer and the corresponding initial learning rate is set to 10^{-4}. The model is trained end-to-end with batch-size 32.

To train the model efficiently, we utilize two training strategies. First, as it is observed by other work [5], the loss for sequence modeling suffers from KL-divergence vanishing. To address this issue, we initialize α as 10^{-4} and gradually increase it to 0.1 as the training progress runs. Moreover, incorporating two independent stochastic variables can reduce the convergence speed. We therefore leverage a stage-wise strategy. Specifically, we fix the latent representation on the image side $\mathbf{z}^r$ as the mean of prior $\boldsymbol{\mu}^r$ and focus on training the recipe part. Then we alternatively train the posterior parameters on the image side after several epochs. Finally, early stopping strategy is applied and the model with best R@1 score on validation set is selected for testing.

Comparison. The proposed MCEN is compared against several SOTA approaches:

- CCA [21], the Canonical Correlation Analysis method. The results are from [48].
- JE [48], a method to learn the joint embedding space of images and texts with pairwise cosine loss. This method also incorporates the classification task as a regularization.
- ATTEN [9], a hierarchical attention model for cross-modal recipe retrieval. This approach also incorporates title information to extract recipe features.
- AdaMine [6], a two-level retrieval approach which injects the semantic information into the triplet object.
- R^2GAN [62], a GAN-based method which learns cross-modal retrieval and multi-modal generation simultaneously.
- ACME [54], the state-of-the-art method on cross-modal recipe retrieval task, which improves modality alignment using multiple GAN components. In our experiments, we use the released pre-trained model and report the results on our sampled test set.

4.2. Main Results

The main results on cross-modal retrieval task are listed in Table 1. Generally, the proposed MCEN consistently outperforms all baselines with obvious margin across all evaluation metrics and test sets. On the 1K set, MCEN achieves 2.0 median rank, which matches the SOTA results. In terms of R@K, MCEN achieves promising performance, beating all baselines including the to-date best approach ACME across all metrics on both image-to-recipe and recipe-to-image tasks.

On the 10K setting, the performances of all models decrease significantly since the retrieval task becomes much harder. As the size of subset increases, the gap between MCEN and previous methods becomes larger. Compared with the SOTA ACME method, our model achieves almost 30% improvements on MedR metric over both im2recipe and recipe2im tasks, indicating the robustness of MCEN.

4.3. Ablation Studies

To evaluate the contributions of different components, we conduct ablation study on several variants of architectures detailedly. We depict the variants of MCEN in Figure 3. MCEN-vanilla (Figure 3 (b)) is the simplest architecture

Size	Methods	Image-to-Recipe				Recipe-to-Image			
		MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10
1K	Random	500	0.1	0.5	1.0	500	0.1	0.5	1.0
	CCA [48]	15.7	14.0	32.0	43.0	24.8	9.0	24.0	35.0
	JE [48]	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0
	ATTEN [9]	4.6	25.6	53.7	66.9	4.6	25.7	53.9	67.1
	AdaMine [6]	2.0	39.8	69.0	77.4	2.0	40.2	68.1	78.7
	R^2GAN [62]	2.0	39.1	71.0	81.7	2.0	40.6	72.6	83.3
	ACME [54]	2.0	44.3	72.9	81.7	2.0	45.4	73.4	82.0
	MCEN (ours)	**2.0**$\pm$0.0	**48.2**$\pm$0.9	**75.8**$\pm$1.1	**83.6**$\pm$0.9	**1.9**$\pm$0.3	**48.4**$\pm$1.0	**76.1**$\pm$0.9	**83.7**$\pm$1.1
10K	JE [48]	41.9	-	-	-	39.2	-	-	-
	ATTEN [9]	39.8	7.2	19.2	27.6	38.1	7.0	19.4	27.8
	AdaMine [6]	13.2	14.9	35.3	45.2	12.2	14.8	34.6	46.1
	R^2GAN [62]	13.9	13.5	33.5	44.9	12.6	14.2	35.0	46.8
	ACME [54]	10.0	18.1	39.9	50.8	9.2	20.1	41.5	51.9
	MCEN (ours)	**7.2**$\pm$0.4	**20.3**$\pm$0.3	**43.3**$\pm$0.3	**54.4** $\pm$0.2	**6.6**$\pm$0.5	**21.4**$\pm$0.3	**44.3**$\pm$0.3	**55.2**$\pm$0.3

Table 1. **Retrieval Results of baselines.** The cross-modal retrieval performance is evaluated with MedR (lower is better) and R@K (higher is better). It should be noted that we do not incorporate pretraining embeddings and additional food-class labels which are utilized by prior approaches.

which does not incorporate any latent variables. The final embeddings $\mathbf{e}^r$ and $\mathbf{e}^v$ are obtained by:

$$\mathbf{e}^r = g_r(\mathbf{s}^r), \tag{21}$$

$$\mathbf{e}^v = g_v(\mathbf{s}^v), \tag{22}$$

where $\mathbf{s}^r$ and $\mathbf{s}^r$ are the output of recipe encoder (Equation 3) and image encoder (Equation 1) respectively. The mappings g_r and g_v are implemented as two-layer neural networks with tanh activations. We also propose two variant models which leverage latent variables on either image (Figure 3 (c)) or recipe side (Figure 3 (d)). Besides, the performance of MCEN without reconstruction component (Equation 17-18) is also reported. For all variants derived from MCEN, the modality-consistency loss (Equation 15) is removed.

The retrieval results of different variant models on 1K subset are listed in Table 2. Not surprisingly, MCEN outperforms all variants with all evaluation metrics. It can be observed that the performance of MCEN-vanilla is similar to ACME (Table 1), indicating the effectiveness of the proposed architecture of the recipe encoder. Moreover, an interesting finding is that MCEN-image outperforms MCEN-recipe. A possible reason could be that, compared with rigmarole instructions, the relative semantic weights of different regions in an image are easier to be exploited.

4.4. Analysis

Parameters and Speed. We list the numbers of parameters and speeds of different systems in Table 3. We can observe that although the inference network on either side

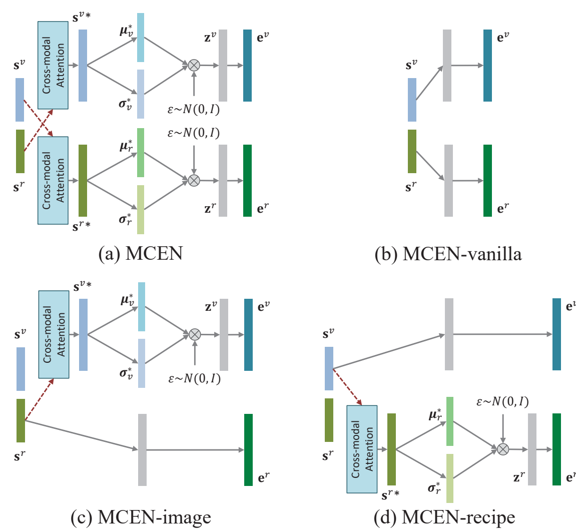


Figure 3. Variants of architectures derived from MCEN.

introduces about 10.4M parameters, the additional parameters do not significantly decrease the training and test speed. Compared with the current SOTA ACME [54], MCEN contains about 30% less parameters and generates cross-modal embeddings with almost double speed, proving the high efficiency of the proposed architecture. The major reason for the gap between MCEN and ACME is that ACME requires additional overhead for adversarial learning.

Effectiveness of Cross-modal Attention. To better understand what has been learned by the cross-modal attention

Methods	Image-to-Recipe			Recipe-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
MCEN-vanilla	44.5	72.3	80.7	44.9	72.8	80.9
MCEN-recipe	45.8	73.1	81.3	46.1	73.3	81.5
MCEN-image	47.6	75.1	83.0	47.8	75.4	83.3
MCEN w/o reconstruction	46.4	75.4	83.1	47.8	75.7	83.3
MCEN	**48.2**	**75.8**	**83.6**	**48.4**	**76.1**	**83.7**

Table 2. **Ablation Study**. The models are evaluated in terms of R@K with 1K subset.

Methods	#Para	Speed	
		Train	Test
AdaMine [6]	46.3M	117.8	197.9
R^2GAN [62]	89.9M	30.3	195.4
ACME [54]	98.6M	30.7	111.7
MCEN-vanilla	48.9M	57.6	194.9
MCEN-recipe	59.3M	45.0	189.1
MCEN-image	59.3M	45.2	188.7
MCEN	69.6M	42.7	185.8

Table 3. **Statistics of parameters, training and testing speed (pairs/second).** All models are evaluated with the same settings on a single Titan XP GPU with batch-size 32. This comparison could be unfair since all the baselines require additional computational overhead for pre-training skip-thought vectors.

components, we visualize the intermediate results with attention. As shown in Figure 4, the attention model learns to focus more on the valid regions containing food and ignore the background. Consequently, the final image embeddings are more constrained and not likely to be affected by noises (i.e. fork and tablecloth) or polysemous instances.

On the recipe side, as shown in Figure 5, the attention model learns to focus on ingredients which can be interpreted based on visual connections with the food images. Taking the first sub-picture in Figure 5 for instance, the attention model attaches highest weights to the three ingredients: *steak*, *ketchup* and *baguettes*, which make up nearly the entire dish. These observations demonstrate that the proposed MCEN learns to capture the semantic alignment relationships between images and recipes.

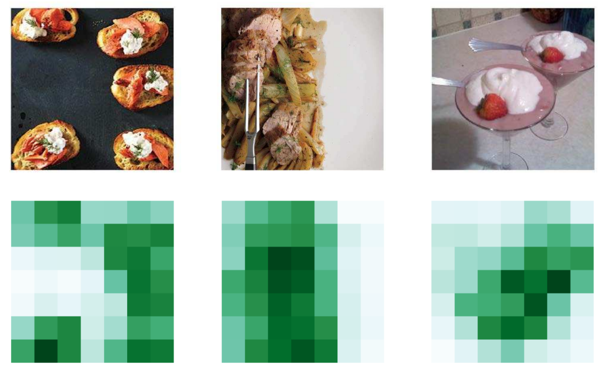


Figure 4. **Attention map of sampled images.** The darker color, the higher attention score.

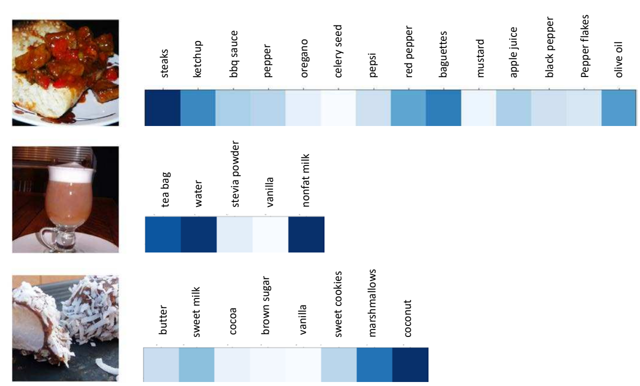


Figure 5. **Visualization of ingredient attention.** The model focuses on important ingredients with high attention scores.

5. Conclusion and Future Work

In this paper, we propose a Modality-Consistent Embedding Network, namely MCEN, for cross-modal recipe retrieval. The proposed model focuses on modeling the interactions between food images and textual recipes during training with latent variables. Concretely, the latent variables are modeled based on cross-modal attention mechanisms during training while the embeddings of different modalities are still calculated independently during inference. We conduct experiments on the challenging Recipe1M dataset and the evaluation results with different metrics demonstrate the efficiency and effectiveness of MCEN. In the future, we are interested in incorporating pre-trained language models into cross-modals analysis tasks.

Acknowledge

We would like to thank the reviewers for their detailed comments and constructive suggestions.

References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255, 2013.

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

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

[4] Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. Simulating action dynamics with neural process networks. In *International Conference on Learning Representation (ICLR)*, 2018.

[5] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

[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 *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 35–44. ACM, 2018.

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

[8] Jing-jing Chen, Chong-Wah Ngo, and Tat-Seng Chua. Cross-modal recipe retrieval with rich food attributes. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1771–1779. ACM, 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] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[11] Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] David Elsweiler, Christoph Trattner, and Morgan Harvey. Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pages 575–584. ACM, 2017.

[14] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2018.

[15] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

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

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

[18] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014.

[19] Marvin Harris. *Good to eat: Riddles of food and culture*. Waveland Press, 1998.

[20] 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.

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

[22] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318, 2017.

[23] Hokuto Kagaya, Kiyoharu Aizawa, and Makoto Ogawa. Food detection and recognition using convolutional neural network. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1085–1088. ACM, 2014.

[24] Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, 2016.

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

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[27] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *International Conference on Machine Learning*, pages 595–603, 2014.

[28] 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.

[29] Tomasz Kusmierczyk and Kjetil Nørvåg. Online food recipe title semantics: Combining nutrient facts and topics. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2013–2016. ACM, 2016.

[30] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[31] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2018.

[32] Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):1–19, 2006.

[33] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019.

[34] 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.

[35] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. *arXiv preprint arXiv:1503.01558*, 2015.

[36] Jonathan Malmaud, Earl Wagner, Nancy Chang, and Kevin Murphy. Cooking with semantics. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 33–38, 2014.

[37] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m: A dataset for learning cross-modal embeddings for cooking recipes and food images. *arXiv preprint arXiv:1810.06553*, 2018.

[38] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. Recognition of multiple-food images by detecting candidate regions. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 25–30. IEEE, 2012.

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

[40] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys (CSUR)*, 52(5):92, 2019.

[41] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia*, 19(5):1100–1113, 2017.

[42] Weiqing Min, Shuqiang Jiang, Shuhui Wang, Jitao Sang, and Shuhuan Mei. A delicious recipe analysis framework for exploring multi-modal recipes with various attributes. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 402–410. ACM, 2017.

[43] Andrea Moed, Daniela Rosner, and Nancy Van House. Is food scenery? generative situations in urban networked photography. In *CHI 2007 Workshop: Image the city: Exploring the practices and technologies of representing the urban environment in HCL. San Jose, CA, USA*, 2007.

[44] Shinsuke Mori, Hirokuni Maeta, Tetsuro Sasada, Koichiro Yoshino, Atsushi Hashimoto, Takuya Funatomi, and Yoko Yamakata. Flowgraph2text: Automatic sentence skeleton compilation for procedural text generation. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 118–122, 2014.

[45] Yuxin Peng, Xin Huang, and Jinwei Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. In *IJCAI*, pages 3846–3853, 2016.

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

[47] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

[48] 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. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3068–3076. IEEE, 2017.

[49] Satoshi Sanjo and Marie Katsurai. Recipe popularity prediction with deep visual-semantic fusion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2279–2282. ACM, 2017.

[50] 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.

[51] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.

[52] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[53] Christoph Trattner and David Elsweiler. 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*, pages 489–498. International World Wide Web Conferences Steering Committee, 2017.

[54] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. 2019.

[55] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[56] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5764–5773, 2019.

[57] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, 2018.

[58] Yoko Yamakata, Shinji Imahori, Hirokuni Maeta, and Shinsuke Mori. A method for extracting major workflow composed of ingredients, tools, and actions from cooking procedural text. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2016.

[59] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3441–3450, 2015.

[60] Longqi Yang, Cheng-Kang Hsieh, Hongjian Yang, John P Pollak, Nicola Dell, Serge Belongie, Curtis Cole, and Deborah Estrin. Yum-me: a personalized nutrient-based meal recommender system. *ACM Transactions on Information Systems (TOIS)*, 36(1):7, 2017.

[61] Feng Zhou and Yuanqing Lin. Fine-grained image classification by exploring bipartite-graph labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1124–1133, 2016.

[62] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11477–11486, 2019.