Being a Supercook: Joint Food Attributes and Multimodal Content Modeling for Recipe Retrieval and Exploration

Weiqing Min, Shuqiang Jiang, *Senior Member, IEEE*, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz

***Abstract*—This paper considers the problem of recipe-oriented image-ingredient correlation learning with multi-attributes for recipe retrieval and exploration. Existing methods mainly focus on food visual information for recognition while we model visual information, textual content (e.g., ingredients), and attributes (e.g., cuisine and course) together to solve extended recipe-oriented problems, such as multimodal cuisine classification and attribute-enhanced food image retrieval. As a solution, we propose a multimodal multitask deep belief network (M^3TDBN) to learn joint image-ingredient representation regularized by different attributes. By grouping ingredients into visible ingredients (which are visible in the food image, e.g., "chicken" and "mushroom") and nonvisible ingredients (e.g., "salt" and "oil"), M^3TDBN is capable of learning both midlevel visual representation between images and visible ingredients and nonvisual representation. Furthermore, in order to utilize different attributes to improve the intermodality correlation, M^3TDBN incorporates multitask learning to make different attributes collaborate each other. Based on the proposed M^3TDBN, we exploit the derived deep features and the discovered correlations for three extended novel applications: 1) multimodal cuisine classification; 2) attribute-augmented cross-modal recipe image retrieval; and 3) ingredient and attribute inference from food images. The proposed approach is evaluated on the constructed Yummly dataset and the evaluation results have validated the effectiveness of the proposed approach.**

Manuscript received June 26, 2016; revised September 26, 2016 and November 6, 2016; accepted December 5, 2016. Date of publication December 14, 2016; date of current version April 15, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61322212, Grant 61532018, Grant 61550110505, Grant 61602437, and Grant 61373122, in part by the National High Technology Research and Development 863 Program of China under Grant 2014AA015202, in part by the Beijing Municipal Commission of Science and Technology under Grant D161100001816001, in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-Notch Young Professionals, and in part by China Postdoctoral Science Foundation under Grant 2016M590135. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Benoit Huet.

W. Min, S. Jiang, H. Wang, and L. Herranz are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: weiqing.min@vipl.ict.ac.cn; shuqiang.jiang@vipl.ict.ac.cn; luis.herranz@vipl.ict.ac.cn; huayang.wang@vipl.ict.ac.cn).

J. Sang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jtsang@nlpr.ia.ac.cn).

X. Liu is with State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: xinda.azz@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2639382

***Index Terms*—Cuisine classification, recipe image retrieval, ingredient inference, multitask deep belief network.**

I. Introduction

FOOD differs in many aspects, such as ingredients, cuisine, course, nutrition and taste. The diversity of food leads to different food preference, which has a strong effect on our personal health and social lives [1], [2]. Effectively modeling these various food information plays important roles in applications like food preference learning, food image calorie estimation and personalized recipe recommendation. The proliferation of online recipe-sharing websites has provided rich data for recipe-oriented research. For example, one of the most popular recipe-sharing websites, Yummly[1] hosts over one million recipes with rich metadata information. Fig. 1 shows some example recipes from Yummly. Each food item consists of the visual food photo, textual content (e.g., name and ingredients) and attributes (e.g., cuisine and course). The huge food images from these websites often have multimodalities and multi-attributes. Such kind of food data opens up many opportunities to recipe-related research communities. Specifically, three major problems have been investigated as cuisine classification [3], recipe retrieval [4] and food image recognition [5]. The top of Fig. 2 illustrated these problems.

Most of the existing studies addressing the above problems focus on exploiting the binary correlations between visual content, textual content, and attributes. For example, Han *et al.* [3] utilized associative classification techniques to discover the underlying correlations between textual ingredient content and the cuisine attribute to address the cuisine classification problem. Freyne *et al.* [4] developed an intelligent food planning system to exploit the correlations between ingredients and recipe names towards recipe recommendation. Hessel *et al.* [5] treated food image recognition as an image captioning problem and investigated different variants of the Convolutional Neural Network (CNN) to exploit the correlations between visual photo content and textual recipe name. However, as shown in Fig. 1, each recipe consists of textual content, visual content and multiple attributes. While existing studies modeled only partial correlations among these included information, we believe both multimodal content and attributes are critical to solve the above recipe-oriented problems. 1) For cuisine classification,

[1][Online]. Available: http://www.yummly.com/

Photos			
Name	Chicken Enchiladas	Ancho Chicken Soup with Masa Dumplings	Hearty Italian Basil Sausage Soup
Ingredients	"1 medium onion, chopped fine", "1 teaspoon canola oil", "3 teaspoons sugar", "1 15-oz can tomato sauce", "1 pound boneless, skinless chicken breasts	"6 bone-in chicken thighs", "1/2 cup maseca masa flour plus 1/2 cup water", "2 dried ancho chiles, stems removed", "2 medium tomatoes, cut in half", "3 cloves garlic, peeled", "2 chayote squash, or 2 large potatoes", ...	"1 pound mild Italian chicken sausage", "3 carrots", "1 stalk celery", "1 tablespoon extra virgin olive oil", "8 cups chicken broth "1 tablespoon dried Italian seasoning",
Cuisine	Mexican	Mexican	Italian
Course	Main Dishes	Soups	Soups

Fig. 1. Food items from Yummly

similar ingredients will lead to very different recipes in different cuisines. For example, we can use the ingredients "chicken", "mushroom", "butter" and "flour" to make a variety of recipes, such as "Braised Chicken with Mushrooms" from Italy and "Mushroom Chicken Rice Soup" from China. Ingredient-based methods probably fail in these cases. The visual content will significantly improve the accuracy for cuisine classification. 2) For recipe retrieval, there are a huge number of internet food images without ingredients, and therefore simple ingredient-matching methods do not work on these unstructured food data. 3) For food image annotation, visual content alone cannot capture all details of food, and it is very difficult to directly retrieve high-level recipe names from the low-level visual content. Taking the recipe attributes and textual ingredients into consideration provides important mid-level features and contributes to inferring richer ingredients and attributes from food images.

Therefore, in this work, we are motivated to study a unified recipe modeling framework, to jointly model both the recipe attributes and the multimodal content information. The proposed framework is expected to capture the underlying rich correlations between various recipe information, and extend the scenarios of traditional recipe-oriented problems to three novel application problems, i.e., multimodal cuisine classification, attribute-augmented cross-modal recipe image retrieval, and ingredient and attribute inference from food images. As shown in Fig. 2, we illustrate and compare between the traditional and extended recipe-oriented problems.

It is non-trivial to model the multimodal and multi-attribute information in a unified framework. The challenge is two-fold: 1) The correlations between the visual recipe photo and the ingredients are weakly-labeled. Some ingredients correspond to explicit regions from images while the others are completely non-visible in the image. As shown in Fig. 1, for the "Chicken Enchiladas" recipe, the ingredients "tomato sauce" and "chicken breasts" are visible while the ingredients "onion", "oil" and "sugar" are not visible. Existing methods either neglect the non-visible ingredients or directly build correlations between the visual content and all ingredients at image-level. 2) Multiple attributes are involved and should be modeled integrated. For example, the recipes from Yummly include different attributes such as the cuisines and courses. Different attributes reflect respective aspects of the recipes and jointly contribute to accurate and complete recipe descriptions. If these attributes are incorporated into the model in an ad hoc manner, this results in models with more sophisticated structures and complicated inference procedure. How to find a general solution to encourage different attributes to collaborate each other is the other challenging problem.

In order to address these challenges, we propose a solution framework called MultiModal MultiTask Deep Belief Network (M^3TDBN): 1) the weak correlation between visual content and ingredients is improved by considering multiple attributes; 2) multiple attributes are simultaneously considered in a multi-task formulation. Specifically, we define two different types of ingredients, i.e., visible ingredients and non-visible ingredients. The visible ingredients are generally visible in the food image (e.g., "chicken" and "mushroom") and the non-visible ingredients are non-visible in the food image (e.g.,"salt", "sugar" and "oil"). The goal of M^3TDBN is to learn the mid-level representation that can capture both the visual representation and non-visual ingredient representation regularized by different attributes. Firstly, by annotating the ingredients with visible ingredients and non-visible ingredients, M^3TDBN can learn joint visual representation between images and visible ingredients, and non-visual ingredient representation, respectively. Secondly, M^3TDBN incorporates multi-task learning [6] to make different attributes reinforce each other. Different attributes and different modality information are correlated through the mid-level representation. Based on the proposed M^3TDBN, we exploit the derived deep features and the discovered correlations in three extended recipe-oriented problems.

The contributions of the proposed approach can be summarized as follows:

1) To our knowledge, this is the first time to simultaneously model visual content, textual ingredients and multiple recipe attributes together into a unified framework to enable various recipe-oriented research problems and applications.
2) We propose a novel MultiModal MultiTask Deep Belief Network (M^3TDBN) model to address the problem of weak dependence between visual content and textual ingredients, as well as the collaboration among different attributes. Two pathways are designed to learn mid-level visual joint representation and ingredient representation respectively by incorporating multi-attributes into a multi-task mechanism.
3) We present a wide variety of recipe-oriented applications based on the proposed M^3TDBN, including 1) multimodal cuisine classification, 2) attribute-augmented cross-modal recipe image retrieval, and 3) ingredient and attribute inference from food images.
4) We collected a real-world food dataset Yummly-28k, where we have validated the effectiveness of our proposed approach.

The rest of the paper is organized as follows. Section II reviews the related work. Section III presents the network architecture of the proposed MultiModal MultiTask Deep Belief Network (M^3TDBN). The model learning and parameter estimation is also introduced in this section. Section IV introduces three recipe-oriented applications derived from our proposed model, including multimodal cuisine classification, attribute-augmented cross-modal recipe image retrieval, and

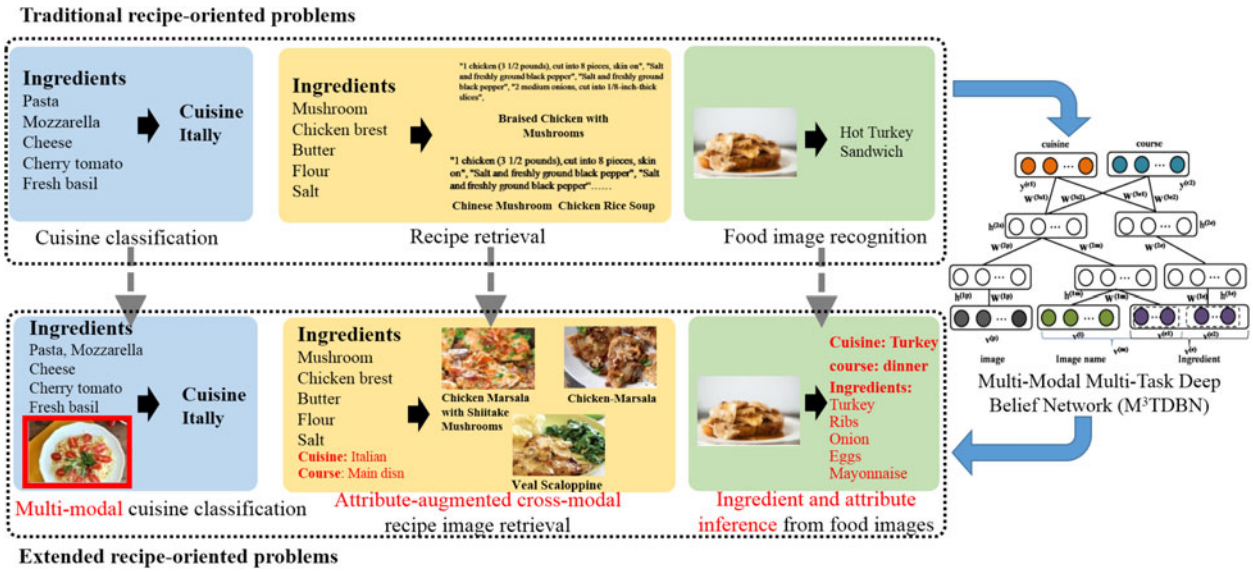


Fig. 2. Traditional versus extended recipe-oriented problems. The extended inputs and outputs are highlighted with red color.

ingredient and attribute inference from food images. Experimental results are reported in Section V. Finally, we conclude the paper and give the future work in Section VI.

II. Related Work

Our work is closely related to the following four research areas: 1) cuisine classification, 2) recipe retrieval and recommendation, 3) food image recognition and annotation and 4) multimodal restricted boltzmann machine [7], [8].

A. Cuisine Classification

Cuisine is a style of cooking characterized by distinctive ingredients, techniques and dishes, and usually associated with a specific geographic region[2]. Recipe cuisines are believed to be one of the main considerations when users choose to eat. Automatically presenting the cuisine of a recipe of interest can boost the quality of recipe recommendation. For cuisine classification, Ahn *et al.* [9] used the graph model to explore the ingredient-based relationship between various regional cuisines to illustrate the ingredients shared by various cuisines and those that are unique to a particular region. Han *et al.* [3] used the ingredients as the feature, namely 0-1 boolean representation to investigate the underlying cuisine ingredient connections by exploiting the classification technologies, including the associative classification and support vector machine. In addition, a famous platform for predictive modelling and analytics competitions, Kaggle organized a competition "What's Cooking?",[3] where the task is to predict the category of a dish's cuisine given a list of its ingredients. These methods mainly use food ingredients for learning tasks. However, similar ingredients will lead to very different recipes in different cuisines and the ingredient information is not enough to classify cuisines. In order to overcome the limitation, our work investigates the multimodal cuisine classification in the multimedia context. Besides the textual ingredients, we provide a multimodal framework, which is capable of simultaneously modeling the visual content and textual ingredients.

[2][Online]. Available: https://en.wikipedia.org/wiki/Cuisine
[3][Online]. Available: https://www.kaggle.com/c/whats-cooking

B. Recipe Image Retrieval and Recommendation

Recipe retrieval and recommendation has many applications. For example, we often need to find recipes based on the ingredients on hand. Wang *et al.* [10] represented the recipes as cooking graphs consisting of ingredients and cooking directions and used the graph representation to characterize Chinese dishes. However, their work fails to take into account the relationships between ingredients. Teng *et al.* [11] constructed two types of networks to capture the relationships between ingredients to represent the recipe features for recipe recommendation. Freyne *et al.* [4] proposed an Intelligent Food Planning (IFP) system to consider the ingredients of a recipe and gave each ingredient a weight. Then, IFP used the weights of the ingredients to predict a new recipe. In addition, some work [12], [13] proposed a matrix factorization method to model hidden factors between users and ingredients for recommendation. Our work is different from them in that besides the images and the textual ingredients, we incorporated the multi-attributes (e.g., cuisine and course) information into our proposed framework for attribute-enhanced cross-modal recipe image retrieval.

C. Food Image Recognition

Many works focus on food image recognition [14]–[17]. Yang *et al.* [15] proposed a visual representation for food items that calculates pairwise statistics between local features computed over a soft pixellevel segmentation of the image into eight ingredient types for food recognition. This approach is bound to work only for standardized meals. Lukas *et al.* [18] mined discriminative parts of food images using random forests for dish

recognition. Compared with these shallow models, Kagaya *et al.* [16] mainly applied the CNN for food detection and recognition. Some works [19]–[22] developed restaurant-specific dish recognition. For example, Xu *et al.* [21] proposed a framework incorporating discriminative classification in geolocalized settings and introduced the concept of geolocalized models for food recognition. There are also some works such as [23]–[26], which focused on mobile food recognition. Based on the food image recognition, Myers *et al.* [27] further proposed a system which can recognize the contents of the meal from one image, and then predicted its nutritional contents. Recent work [28] utilized additional text information for multimodal food recognition. Hessel *et al.* [5] further discussed different methods of CNN for food image caption. In addition, Amano *et al.* [29] proposed a method to generate classification categories automatically from the raw meal names. Some food dataset are also available, such as FoodCam-256 [24], ETHZ Food-101 [18], UPMC Food-101 [28], and Food201-MultiLabel dataset [27]. These methods focus on food recognition based on the image name or a few ingredients. Rich attribute information is not explored, such as the cuisine and course information. Recent work [30] proposed deep architectures for simultaneous learning of ingredient recognition and food categorization by exploiting the visual features, ingredients and image categories. Different from their work, we also take the recipe attributes into account for richer food image annotations by inferring ingredients and attributes from the food images.

D. Multimodal Restricted Boltzmann Machine

Due to the power of representation learning [31], Restricted Boltzmann Machine (RBM) has been successfully applied to various tasks, such as image classification [7], information recommendation [32] and retrieval [33], especially multimodal learning [34]–[37]. RBMs were originally developed for modeling binary vectors. In multimodal learning, different kinds of features have different distributions. In order to model different distributions, Gaussian-Bernoulli RBMs were developed [7] for continuous values, such as visual features and audio features. Salakhutdinov *et al.* [38] further developed another variant of RBM, namely Replicated Softmax Model (RSM) for modeling the textual featues with the count data. Based on these basic units, Srivastava *et al.* [34] proposed RBM based deep network for multimodal data modeling, where the data from each modality is firstly learned from the corresponding single-modality pathway and the learned high-level features are fused on the top layer. In contrast, Pang *et al.* [36] introduced the third pathway to learn audio-specific features and then learned joint representation among visual features, textual features and audio features for affective analysis based video retrieval. Huang *et al.* [36] proposed a multi-label conditional RBM for multimodal multi-label classification. Yuan *et al.* [37] designed a novel relational generative deep learning model to solve the social media link analysis problem. Our work is also inspired by the work [34], in that we are devoted to using RBM based framework for exploring the joint multimodal representation. However, we have differences in two-fold. (1) Motivation. We aim to apply RBM-based framework to model different modalities and attributes to enable recipe-related applications. (2) Methodology. At the bottom layer, we group the ingredients into visual ingredients and non-visual ingredients and introduce another pathway to model the non-visual ingredients; At the top layer, we incorporate multi-task learning in our model to utilize different kinds of attributes.

III. Multimodal Multitask Deep Belief Network (M^3TDBN)

This section introduces the MultiModal MultiTask Deep Belief Network (M^3TDBN), which is to model both multi-modality and multi-attributes from the food items. We firstly describe the basic ideas and the network architecture, and then introduce the inference and learning process of the proposed model.

A. M^3TDBN Design

Two key ideas are exploited in our deep network. Firstly, in order to learn mid-level visual joint representation and non-visual representation, as shown in Fig. 3(a), there are two pathways namely Pathway-A and Pathway-B. Pathway-A is to learn the joint representation of image features and visible ingredients. Pathway-B is to learn the representation of ingredients including non-visible ingredients. Secondly, in order to incorporate the cuisine and course attribute into our deep network, we connect these two kinds of information to the top level layers, which enables the fine-tuning of the whole architecture in a multitask fashion [see Fig. 3(b)].

B. M^3TDBN Representation

Each food item is denoted as $\{\mathbf{v}^{(p)}, \mathbf{v}^{(o)}, \mathbf{y}, \mathbf{z}\}$, where $\mathbf{v}^{(p)}$, $\mathbf{v}^{(o)}$, $\mathbf{y}$, $\mathbf{z}$ are the input visual units of the image, ingredients, the cuisine and course, respectively. The visual units $\mathbf{v}^{(o)}$ are divided into two parts, one of which $\mathbf{v}^{(m)}$ denotes the units of visible ingredients and $\mathbf{v}^{(e)}$ the units of non-visible ingredients. Next, we will show the network representation in details.

1) Pathway-A: As shown in Fig. 3(a), for Pathway-A, because the input units $\mathbf{v}^{(p)}$ are visual features, the connections between $\mathbf{v}^{(p)}$ and $\mathbf{h}^{(1p)}$ are modeled with Gaussian Restricted Boltzmann Machine [7]. The joint distribution of binary visible units and binary hidden units is written as follows:

$$P(\mathbf{v}^{(p)}, \mathbf{h}^{(1p)}) = \frac{1}{\mathcal{Z}} \exp(-E(\mathbf{v}^{(p)}, \mathbf{h}^{(1p)}; \boldsymbol{\theta})). \tag{1}$$

The energy function $E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})$ is

$$E(\mathbf{v}^{(p)}, \mathbf{h}^{(1p)}; \boldsymbol{\theta}) = \sum_i \frac{(v_i^{(p)} - b_i^{(p)})}{2\sigma_i^2} - \sum_i \sum_j \frac{v_i^{(p)}}{\sigma_i} W_{ij}^{(1p)} h_j^{(1p)} - \sum_j c_j^{(p)} h_j^{(1p)} \tag{2}$$

where $\mathcal{Z}$ is the normalizing constant. The weights $\mathbf{W}^{(1p)} = \{w_{i,j}^{(1p)}\}$ are associated with the connection between the visible units $\mathbf{v}^{(p)} = \{v_i^{(p)}\}$ and the hidden units $\mathbf{h}^{(1p)} = \{h_j^{(1p)}\}$, as well as bias weights $\mathbf{b}^{(p)} = \{b_i^{(p)}\}$ for the visible units and

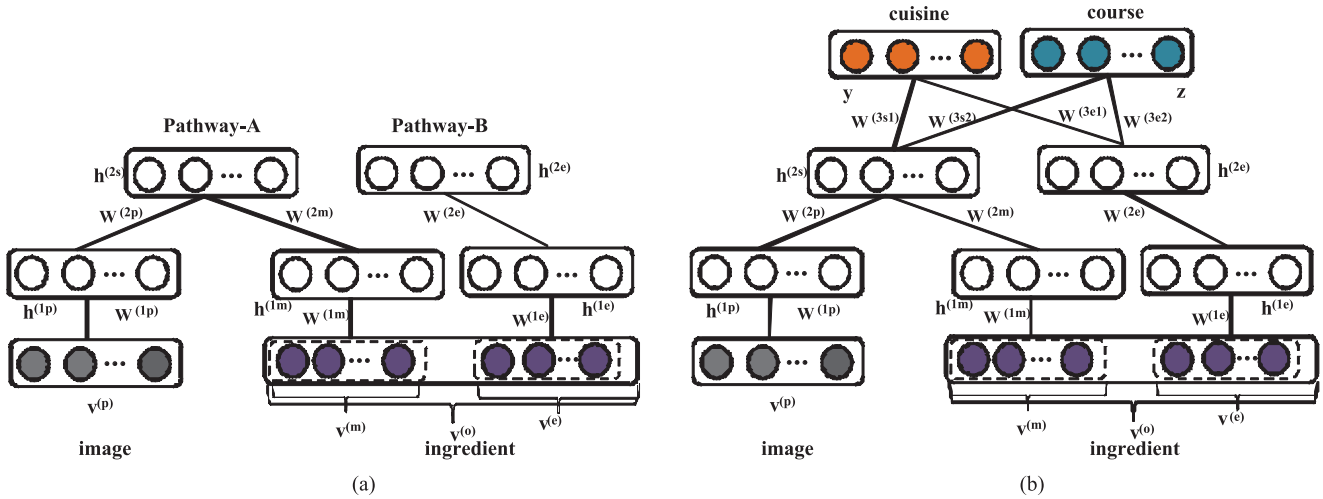


Fig. 3. Proposed deep network: (a) joint visual and non-visual learning learning network and (b) M^3TDBN

$\mathbf{c}^{(p)} = \{c_j^{(p)}\}$ for the hidden units. $\boldsymbol{\theta} = \{\mathbf{c}^{(p)}, \mathbf{b}^{(p)}, \mathbf{W}^{(1p)}, \sigma\}$ are the model parameters. In order to simplify the representation, we will neglect the parameters $\boldsymbol{\theta}$ in the following and represent $E(\mathbf{v}^{(p)}, \mathbf{h}^{(1p)}; \boldsymbol{\theta})$ as $E(\mathbf{v}^{(p)}, \mathbf{h}^{(1p)})$. The corresponding probabilistic forms of the energy functions are ignored. The conditional distributions can be written as follows:

$$v_i^{(p)} = 1|\mathbf{h}^{(1p)} \sim \mathcal{N}\Big(b_i^{(p)} + \sigma_i \sum_j W_{ij}^{(1p)} h_j^{(1p)}, \sigma_i^2\Big) \tag{3}$$

$$P(h_j^{(1p)} = 1|\mathbf{v}^{(p)}) = g\Big(c_j^{(1p)} + \sum_i \frac{v_i^{(p)}}{\sigma_i} W_{ij}^{(1p)}\Big) \tag{4}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2. $g(x) = \frac{1}{1+\exp(-x)}$ is the logistic function. Note that variance is empirically set to unit variance $\sigma_i = 1$. This is because learning the variance of this network made the training unstable [34].

Different from the visual modalities, the input units $\mathbf{v}^{(m)}$ are the count data, correspondingly, the connections between $\mathbf{v}^{(m)}$ and $\mathbf{h}^{(1m)}$ are modeled with Replicated Softmax Model (RSM) [39]. The energy function $E(\mathbf{V}^{(m)}, \mathbf{h}^{(1m)})$ is

$$E(\mathbf{V}^{(m)}, \mathbf{h}^{(1m)}) = -\sum_k \sum_j \hat{v}_k^{(m)} W_{kj}^{(1m)} h_j^{(1m)} - \sum_i b_k^{(m)} \hat{v}_k^{(m)} - M^{(m)} \sum_j c_j^{(1m)} h_j^{(1m)} \tag{5}$$

where $\hat{\mathbf{v}}^{(m)} = \{\hat{v}_{(1)}^{(m)}, \ldots, \hat{v}_k^{(m)}, \ldots \hat{v}_{K^{(m)}}^{(m)}\}$ and $\hat{v}_k^{(m)} = \sum_{i=1}^{M^{(m)}} v_{ik}^{(m)}$ denotes the count for the k^{th} word. $M^{(m)}$ is the length of the document. $K^{(m)}$ is the size of the ingredient dictionary in our work. The corresponding conditional distributions are given by

$$P(\hat{v}_k^{(m)} = 1|\mathbf{h}^{(1m)}) = \frac{\exp(b_k^{(m)} + \sum_j W_{kj}^{(1m)} h_j^{(1m)})}{\sum_{q=1} \exp(b_q^{(m)} + \sum_j W_{qj}^{(1m)} h_j^{(1m)})} \tag{6}$$

$$P(h_j^{(1m)} = 1|\mathbf{V}^{(m)}) = g\left(M^{(m)} c_j^{(1m)} + \sum_k \hat{v}_k^{(m)} W_{kj}^{(1m)}\right). \tag{7}$$

$\mathbf{V}^{(m)}$ is $M^{(m)} \times K^{(m)}$ observed binary matrix with $v_{ik} = 1$ iff the multinomial visual unit i takes on the k^{th} value.

The second layer of Pathway-A learns joint visual representation, which consists of two binary-value RBM.

The energy function $E(\mathbf{h}^{(1p)}, \mathbf{h}^{(1m)}, \mathbf{h}^{(2s)})$ is

$$E(\mathbf{h}^{(1p)}, \mathbf{h}^{(1m)}, \mathbf{h}^{(2s)}) = -\sum_i \sum_j h_i^{(1p)} W_{ij}^{(2p)} h_j^{(2s)} - \sum_k \sum_j h_k^{(1m)} W_{kj}^{(2m)} h_j^{(2s)} - \sum_i c_i^{(1p)} h_i^{(1p)} - \sum_k c_k^{(1m)} h_k^{(1m)} - \sum_j c_j^{(2s)} h_j^{(2s)}. \tag{8}$$

The corresponding conditional distributions are

$$P(h_j^{(2s)} = 1|\mathbf{h}^{(1p)}, \mathbf{h}^{(1m)}) = g\Big(c_j^{(2s)} + \sum_i h_i^{(1p)} W_{ij}^{2p} + \sum_k h_k^{(1m)} W_{kj}^{(2m)}\Big) \tag{9}$$

$$P(h_i^{(1p)} = 1|\mathbf{h}^{(2s)}) = g\Big(c_i^{(1p)} + \sum_j W_{ij}^{(2p)} h_j^{(2s)}\Big) \tag{10}$$

$$P(h_k^{(1m)} = 1|\mathbf{h}^{(2s)}) = g\Big(c_k^{(1m)} + \sum_j W_{kj}^{(2m)} h_j^{(2s)}\Big). \tag{11}$$

2) Pathway-B: For Pathway-B, the input units of the first layer is count data, therefore, the energy function and corresponding conditional distributions are similar to (5)–(7). The energy function $E(\mathbf{V}^{(e)}, \mathbf{h}^{(1e)})$ is

$$E(\mathbf{V}^{(e)}, \mathbf{h}^{(1e)}) = -\sum_k \sum_j \hat{v}_k^{(e)} W_{kj}^{(1e)} h_j^{(1e)} - \sum_i b_k^{(e)} \hat{v}_k^{(e)} - M^{(e)} \sum_j c_j^{(1e)} h_j^{(1e)} \quad (12)$$

where $\hat{\mathbf{v}}^{(e)} = \{\hat{v}_{(1)}^{(e)}, \ldots, \hat{v}_k^{(e)}, \ldots \hat{v}_{K^{(e)}}^{(e)}\}$ and $\hat{v}_k^{(e)} = \sum_{i=1}^{M^{(e)}} v_{ik}^{(e)}$ denotes the count for the k^{th} word. $M^{(e)}$ is the length of the document. $K^{(e)}$ is the size of the ingredient dictionary in our work. The corresponding conditional distributions are given by

$$P(\hat{v}_k^{(e)} = 1 | \mathbf{h}^{(1e)}) = \frac{\exp(b_k^{(e)} + \sum_j W_{kj}^{(1e)} h_j^{(1m)})}{\sum_{q=1} \exp(b_q^{(e)} + \sum_j W_{qj}^{(1e)} h_j^{(1m)})} \quad (13)$$

$$P(h_j^{(1e)} = 1 | \mathbf{V}^{(e)}) = g\Big(M^{(e)} c_j^{(1e)} + \sum_k \hat{v}_k^{(e)} W_{kj}^{(1e)}\Big). \quad (14)$$

$\mathbf{V}^{(e)}$ is $M^{(e)} \times K^{(e)}$ observed binary matrix with $v_{ik} = 1$ iff the multinomial visual unit i takes on the k^{th} value.

The second layer is binary RBM, where the energy function $E(\mathbf{h}^{(1e)}, \mathbf{h}^{(2e)})$ is

$$E(\mathbf{h}^{(1e)}, \mathbf{h}^{(2e)}) = -\sum_i \sum_j h_i^{(1e)} W_{ij}^{(1e)} h_j^{(2e)} - \sum_i c_i^{(1e)} h_i^{(1e)} - \sum_j c_j^{(2e)} h_j^{(2e)}. \quad (15)$$

The corresponding conditional distributions are

$$P(h_i^{(1e)} = 1 | \mathbf{h}^{(2e)}) = g\Big(c_i^{(1e)} + \sum_j W_{ij}^{(2e)} h_j^{(2e)}\Big) \quad (16)$$

$$P(h_j^{(2e)} = 1 | \mathbf{h}^{(1e)}) = g\Big(c_j^{(2e)} + \sum_i h_i^{(1e)} W_{ij}^{(2e)}\Big). \quad (17)$$

3) Joint Layer With Attributes: When introducing the cuisine and course information, the top layer of the two pathways connects two kinds of attributes units. These two kinds of attributes can be considered as the part of the input [see Fig. 3(b)] and used to fine-tune the whole architecture jointly. The energy function of the top layer is

$$E(\mathbf{h}^{(2s)}, \mathbf{h}^{(1p)}, \mathbf{h}^{(1m)}, \mathbf{h}^{(2e)}, \mathbf{h}^{(1e)}, \mathbf{y}, \mathbf{z}) = E(\mathbf{h}^{(1e)}, \mathbf{h}^{(2e)}) + E(\mathbf{h}^{(1p)}, \mathbf{h}^{(1m)}, \mathbf{h}^{(2s)}) + E(\mathbf{h}^{(2s)}, \mathbf{h}^{(2e)}, \mathbf{y}, \mathbf{z}) \quad (18)$$

where

$$\begin{aligned} E(\mathbf{h}^{(2s)}, \mathbf{h}^{(2e)}, \mathbf{y}, \mathbf{z}) = &-\sum_a \sum_l y_a W_{al}^{(3s1)} h_l^{(2s)} \\ &- \sum_p \sum_l z_p W_{pl}^{(3s2)} h_l^{(2s)} \\ &- \sum_a \sum_f y_a W_{af}^{(3e1)} h_f^{(2e)} \\ &- \sum_p \sum_f z_p W_{pf}^{(3e2)} h_f^{(2e)} \\ &- \sum_a d_a^{(1)} y_l - \sum_p d_p^{(2)} z_p. \end{aligned} \quad (19)$$

The conditional probabilities can be written as follows:

$$P(h_i^{(1p)} = 1 | \mathbf{h}^{(2s)}) = g\Big(c_i^{(1p)} + \sum_j W_{ij}^{(2p)} h_j^{(2s)}\Big) \quad (20)$$

$$P(h_i^{(1m)} = 1 | \mathbf{h}^{(2s)}) = g\Big(c_i^{(1m)} + \sum_j W_{ij}^{(2m)} h_j^{(2s)}\Big) \quad (21)$$

$$P(h_i^{(1e)} = 1 | \mathbf{h}^{(2e)}) = g\Big(c_i^{(1e)} + \sum_j W_{ij}^{(2e)} h_j^{(2e)}\Big) \quad (22)$$

$$P(y_a = 1 | \mathbf{h}^{(2s)}, \mathbf{h}^{(2e)}) = \frac{\exp(d_a^{(1)} + \sum_l W_{al}^{(3s1)} \mathbf{h}_l^{(2s)} + \sum_f \mathbf{W}_{af}^{(3e1)} \mathbf{h}^{(2e)})}{\sum_q \exp(d_q^{(1)} + \sum_l \mathbf{W}_{ql}^{(3s1)} \mathbf{h}_l^{(2s)} + \sum_f W_{qf}^{(3e1)} \mathbf{h}_f^{(2e)})} \quad (23)$$

$$P(z_p = 1 | \mathbf{h}^{(2s)}, \mathbf{h}^{(2e)}) = \frac{\exp(d_p^{(2)} + \sum_l W_{pl}^{(3s2)} \mathbf{h}_l^{(2s)} + \sum_f \mathbf{W}_{pf}^{(3e2)} \mathbf{h}^{(2e)})}{\sum_q \exp(d_q^{(2)} + \sum_l \mathbf{W}_{ql}^{(3s2)} \mathbf{h}_l^{(2s)} + \sum_f W_{qf}^{(3e2)} \mathbf{h}_f^{(2e)})} \quad (24)$$

$$P(h_j^{(2s)} = 1 | \mathbf{h}^{(1p)}, \mathbf{h}^{(1m)}, \mathbf{y}, \mathbf{z}) = g\Big(c_i^{(2s)} + \sum_i h_i^{(1p)} W_{ij}^{(2p)} + \sum_i h_i^{(1m)} W_{ij}^{(2m)} + \sum_a y_a W_{aj}^{(3s1)} + \sum_p z_p W_{pj}^{(3s2)}\Big) \quad (25)$$

$$P(h_j^{(2e)} = 1 | \mathbf{h}^{(1e)}, \mathbf{y}, \mathbf{z}) = g\Big(c_j^{(2e)} + \sum_i h_i^{(1e)} W_{ij}^{(2e)} + \sum_i y_a W_{aj}^{(3e1)} + \sum_i z_p W_{pj}^{(3e2)}\Big) \quad (26)$$

where the attribute labels are represented as the softmax function. $\mathbf{d}^{(1)} = \{d_a^{(1)}\}$ and $\mathbf{d}^{(2)} = \{d_p^{(2)}\}$ are the bias terms of $\mathbf{y}$ and $\mathbf{z}$, respectively.

C. Inference and Learning

We train the network with stochastic gradient descent using Contrastive Divergence (CD) [40]. Since the exact inference is intractable, we use mean-field or alternating Gibbs sampling for approximate inference. Particularly, each RBM component of the proposed M^3TDBN is pretrained using the greedy layer-wise pretraining strategy. In this stage, we adopt k-step

Algorithm 1: Attribute-Enhanced Cross-Modal Recipe Image Retrieval

1: Clamp observed visual feature $\mathbf{v}_q^{(o)} = \{\mathbf{v}_q^{(m)}, \mathbf{v}_q^{(e)}\}$, cuisine features $\mathbf{y}_q$ and course features $\mathbf{z}_q$ at the input.
2: Infer the values of the hidden variables $\mathbf{h}_q^{(1m)}$ in Pathway-A by forward propagating $\mathbf{v}_q^{(m)}$ through to the first hidden layer using (7).
3: Infer the values of the hidden variables $\mathbf{h}_q^{(1e)}$ in Pathway-B by forward propagating $\mathbf{v}_q^{(1e)}$ through to the first hidden layer using (14).
4: Perform alternating Gibbs sampling to infer $\mathbf{h}_q^{(1p)}$ using (20-26) conditioned on $\mathbf{y}_q$, $\mathbf{z}_q$, $\mathbf{h}_q^{(1m)}$ and $\mathbf{h}_q^{(1e)}$
5: Infer $\mathbf{h}^{(1p)}$ from all food images of the query dataset by forward propagating $\mathbf{v}^{(p)}$ through to the first hidden layer using (4).

contrastive divergence (CD_k) to learn the parameters. For CD_k, a k-step Markov chain is initialized with the training sample. The stochastic reconstruction of the training sample from Markov chain by Gibbs sampling has a decreased free energy. Hence, this reconstruction can be approximately treated as the distribution generated by the RBM model. For example, in the first layer of image pathway in Pathway-A, we use Gibbs sampling through (3) and (2) for the reconstructed input. In practice, we apply CD_1 in RBM training, this is because the good approximation of the changing direction is already obtained when $k = 1$.

IV. Recipe-Oriented Applications

M^{3}TDBN can be potentially applied to many recipe-related problems based on its power of representation and inference. In this section, we apply the proposed M^{3}TDBN into three recipe-oriented problems, namely multimodal cuisine classification, attribute-augmented cross-modal recipe image retrieval, and ingredient and attribute inference from food images.

A. Multimodal Cuisine Classification

For each food item with multimodal inputs including the recipe food and ingredients, we should infer the visual representation $\mathbf{h}^{(2s)}$ and non-visual representation $\mathbf{h}^{(2e)}$ based on the learned model. In order to generate $\mathbf{h}^{(2s)}$, we first utilize (4) and (7) to infer $\mathbf{h}^{(1p)}$ and $\mathbf{h}^{(1m)}$, respectively, and then infer $\mathbf{h}^{(2s)}$ using (9). In order to generate $\mathbf{h}^{(2e)}$, we infer $\mathbf{h}^{(1e)}$ using (14) and then infer $\mathbf{h}^{(2e)}$ using (17) based on $\mathbf{h}^{(1e)}$. The concatenated feature representations $< \mathbf{h}^{(2s)}, \mathbf{h}^{(2e)} >$ can be used for multimodal cuisine classification.

B. Attribute-Augmented Cross-Modal Recipe Image Retrieval

This task is described as follows: given the query ingredients $\mathbf{v}_q^{(o)} = \{\mathbf{v}_q^{(m)}, \mathbf{v}_q^{(e)}\}$ and food attributes, including cuisine $\mathbf{y}_q$ and course $\mathbf{z}_q$, the goal is to return a ranked food image list. Such task can be used in many scenarios. For example, one user wants to cook *American breakfast* given some ingredients. In this case, given the query of both the ingredients and two attributes: the cuisine *American* and the course *breakfast*, the user can obtain more personalized retrieval results from this task. The details of this inference process can be shown in Algorithm I.

Note that there are two kinds of query ingredients from users' input. The ingredients contain or do not contain non-visible ones. For query ingredients without non-visible ones, we just retrieve recipe images through Pathway-A. For query ingredients with non-visible ones, we should jointly use the Pathway-A and Pathway-B to retrieve recipe images. Through the algorithm, we can obtain the inferred query features $\mathbf{h}_q^{(1p)}$ and all the hidden features $\mathbf{h}^{(1p)}$. The cosine similarity is used to match queries to data points.

$$\mathrm{sim}(\mathbf{h}_q^{(1p)}, \mathbf{h}^{(1p)}) = \frac{\mathbf{h}_q^{(1p)}\mathbf{h}^{(1p)}}{\|\mathbf{h}_q^{(1p)}\|\|\mathbf{h}^{(1p)}\|} \tag{27}$$

$$\max_{\hat{y}\in Y(x_i)} s(CVG(x_i, \hat{y})) + \Delta(y_i, \hat{y}) - s(CVG(x_i, y_i)) \tag{28}$$

C. Ingredient and Attribute Inference From Food Images

For this application, given one recipe image, our goal is to infer rich textual and attribute information including ingredients, cuisine and course information with high probability. Since our model can generate the unknown modalities conditioned on a given modality, we can alternatively perform Gibbs sampling to draw samples from $p(\mathbf{v}^{(m)}, \mathbf{y}, \mathbf{z}|\mathbf{v}^{(p)})$ based on the conditional distributions as follows:

$$P(h_i^{(1p)} = 1|\mathbf{h}^{(2s)}) = g\Big(c_i^{(1p)} + \sum_j W_{ij}^{(2p)} h_j^{(2s)}\Big) \tag{29}$$

$$P(h_i^{(1m)} = 1|\mathbf{h}^{(2s)}) = g\Big(c_i^{(1m)} + \sum_j W_{ij}^{(2m)} h_j^{(2s)}\Big) \tag{30}$$

$$P(y_a = 1|\mathbf{h}^{(2s)}) = \frac{\exp(d_a^{(1)} + \sum_l W_{al}^{(3s1)}\mathbf{h}_l^{(2s)})}{\sum_q \exp(d_q^{(1)} + \sum_l \mathbf{W}_{ql}^{(3s1)}\mathbf{h}_l^{(2s)})} \tag{31}$$

$$P(z_p = 1|\mathbf{h}^{(2s)}) = \frac{\exp(d_p^{(2)} + \sum_l W_{pl}^{(3s2)}\mathbf{h}_l^{(2s)})}{\sum_q \exp(d_q^{(2)} + \sum_l \mathbf{W}_{ql}^{(3s2)}\mathbf{h}_l^{(2s)})} \tag{32}$$

$$P(h_j^{(2s)} = 1|\mathbf{h}^{(1p)}, \mathbf{h}^{(1m)}, \mathbf{y}, \mathbf{z}) \tag{33}$$

$$= g\Big(c_i^{(2s)} + \sum_i h_i^{(1p)} W_{ij}^{(2p)} + \sum_i h_i^{(1m)} W_{ij}^{(2m)} + \sum_a y_a W_{aj}^{(3s1)} + \sum_p z_p W_{pj}^{(3s2)}\Big) \tag{34}$$

where $\mathbf{h}^{(1p)}$ in Pathway-A is inferred by forward propagating $\mathbf{v}^{(p)}$ through to the first hidden layer using (7). After the iteration, we can obtain the sample values $\mathbf{y}$ and $\mathbf{z}$. We then sample $\mathbf{v}^{(m)}$ using the inferred hidden variables through (6).

V. Experiment

In this section, we firstly describe the experimental setting including the dataset and implementation details. We then evaluate the performance of the proposed three applications including multimodal cuisine classification, attribute-augmented cross-modal recipe image retrieval, ingredient and attribute inference from food images, respectively.

A. Yummly-28K Dataset

To build the food dataset, we crawled the data through Yummly API[4] and obtained 63,492 recipe items in total. Since our work needs to utilize cuisine and course information, we removed food items without cuisine or course labels. As shown in Fig. 1, each recipe item includes the food image, the recipe name, the ingredients, cuisine and course information.

In order to build the ingredient vocabulary, the ingredients are preprocessed as follows [11]. Since each ingredient is usually listed on a separate line, we first found the maximal match between a pre-curated list of ingredients and the text of the line. Here the pre-curated list is from the training set in Kaggle competition.[5] We then used the regular expression matching to remove non-ingredient terms from the line and identified the remainder as the ingredient. We also removed quantifiers, such as "1 pound" and "4 cups", the words representing consistency or temperature, such as "diced" and "hot". We also removed content in parentheses. In addition, highly similar ingredients, e.g. "cheddar cheese", used in 1338 recipes, were considered different from "shredded cheddar cheese", occurring in 543 recipes. Finally, because one of our application is cuisine classification, we further removed food items with less than 100 samples for each cuisine.

After preprocessing, we constructed a vocabulary of 3000 ingredients. The ingredients are further divided into 2208 visible ingredients and 792 non-visible ingredients. Note that there are some ingredients which are visible in one food image and non-visible in another. In our experiment, we simply calculated the proportion of the visible-annotated recipe items for each ingredient. If the ratio is larger than 0.5,[6] then we annotated ingredients as visible ingredients and non-visible ones, otherwise. The final dataset has 27,638 items, which we denote as the Yummly-28K dataset. The statistic of Yummly-28K dataset is shown in Table I. Table II lists the values of different attributes: there are 16 kinds of cuisines (e.g, "American", "Italian" and "Mexican") and 13 kinds of recipe courses (e.g, "Main Dishes", "Desserts" and "Lunch and Snacks"). Fig. 4 shows some examples. Note that each item can have multiple labels for each attribute. For example, the dish "Black-Eyed Pea Salad" is labeled with two cuisine attributes "Southern and Soul Food" and "American". It is labeled with three course attributes "Main Dishes", "Breakfast and Brunch" and "Lunch and Snacks".

[4][Online]. Available: https://developer.yummly.com/documentation

[5][Online]. Available: https://www.kaggle.com/c/whats-cooking/data

[6]We conducted the experiment using the validation set for cuisine classification in our dataset. For the ratio $r = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.8, 0.9, 1.0\}$, we found that with the increase of r, the performance of MAP is rising; when $r \geq 0.5$, the performance is stable. Therefore, we chose 0.5 as the threshold.

TABLE I
TATISTICS OF YUMMLY-28K

♯ Items	♯ Cuisine	♯ Course	♯ Vocabulary
27,638	16	13	3000

TABLE II
VALUES OF DIFFERENT ATTRIBUTES

♯ Type	♯ Value
Cuisine	American, Italian, Mexican, Asian, Indian, Mediterranean, Southwestern, Kid-Friendly, Chinese, Barbecue, Spanish Southern& Soul Food, Cajun& Creole, French, Thai, Greek
Course	Main Dishes, Desserts, Side Dishes, Salads, Afternoon Tea, Soups, Lunch and Snacks, Condiments and Sauces, Breads, Breakfast and Brunch, Beverages, Cocktails, Appetizers

Fig. 4. Some examples with different cuisines and courses

B. Implementation Details

For each item, the ingredients are represented by 3000-D Bag Of Ingredients (BOI), similar to Bag Of Words. For visual features, we extract CNN deep features. Yummly-28K dataset is not large enough to train a CNN model. In order to represent food images properly, following [5], we fine-tuned the AlexNet [41] by the Food-101 dataset which contains 101 classes of food with 101K food images containing 75,250 images for training and 25,250 images for testing [18]. The last 1000-way softmax is replaced by 101 and the base learning rate is set to 0.001. The resulting model has a classification accuracy on the Food-101 test set of 68.3%. Once the network is tuned, we computed 4096 dimensional vector representations for each image in Yummly-28K by extracting the network activations in the final fully-connected layer. We call such visual features as CNN Visual Features (CNN-VF).

For the paramers of M^3TDBN, in Pathway-A, the image pathway consists of a Gaussian RBM with 4096 linear visible units and 1000 hidden units. The visible ingredient pathway consists of a RSM with 2208 visible units and 1000 hidden units. The joint layer contains 2000 hidden units, which connect 16 visible cuisine units and 13 course units. In Pathway-B, the non-visible ingredient pathway consists of a RSM with 792 visible units and 1000 hidden units, followed by another layer of 1000 hidden units. The hidden units from the top layer connect 16 visible cuisine units and 13 course units. Table III lists the number of

TABLE III
NUMBER OF NEURONS IN EACH LAYER OF OUR M^3TDBN

Features	$\mathbf{v}$	$\mathbf{h}^{(1)}$	$\mathbf{h}^{(2)}$	$\mathbf{y}+\mathbf{z}$
CNN	4096	1000	2000	16 + 13
Visual-BOI	2208	1000		
Non-Visual-BOI	792	1000	1000	16 + 13

neurons in each layer. The network contains 12,183,000 weight parameters and 13,125 bias parameters. Like RBM based deep network [36], [42], the number of parameters in our network is also high. In order to avoid over-fitting and speed up the learning, we firstly used the greedy layer-wise pre-training strategy to train our network, which is developed by Hinton *et al.* [43] and has been widely used in RBM-based deep network [36], [42]. That means, we did not train the whole network at once, but only trained one layer of RBM at each time. Secondly, we introduced some hyper-parameters, such as the weight decay and sparsity terms [44] to effectively reduce the complexity of the model. Thirdly, the deep network is suitable for parallel training of parameters on each pathway. Furthermore, Hinton *et al.* [44] found that if each image contains 1,000 pixels, using 10,000 training examples to learn weights of a million parameters in one RBM is quite reasonable. In the proposed network, although the largest RBM has 4,096 $\times$ 1,000 or around 4 million parameters, in the following experiment with 20,000 samples as the training set, we did not observe the tendency of over-fitting when training the network in the layer-wise pre-training strategy.

For the hyper-parameters, the learning rate of RSM layer is 0.10 and other layers is 0.01. They are all selected by grid search. For the Gaussian-RBM, each Gaussian visible unit is empirically set to have unit variance $\sigma_i = 1$ [34]. For the inference of all three tasks, empirically, about 10 times of Gibbs sampling is enough to obtain a good result.

Considering that each recipe item in Yummly may have multiple labels for each attribute, similar to [34], [42], we performed 1-vs-all classification using logistic regression classifiers to classify all learned features for all the tasks involving classification and prediction. Since the logistic regression can be considered as one layer forward neural network, the parameters include the learning rate and the weight decay, and they are selected by grid search on the validation set.

Yummly-28K is split into three subsets: 20,000 items as the training set, 3000 items as validation set and 4638 items as the testing set.

C. MultiModal Cuisine Classification

We evaluate the effectiveness of learned joint representation by our model M^3TDBN in cuisine classification. Considering that recipe items in Yummly may have multiple cuisine labels, we evaluate our models using Mean Average Precision (MAP) and Precision at top-50 predictions (Prec@50), which are standard metrics used for multi-label classification [45].

We consider the following baselines for comparison:

TABLE IV
PERFORMANCE OF ALL DIFFERENT ALGORITHMS FOR MULTIMODAL CUISINE CLASSIFICATION

Method	MAP	Pre@50
CNN-VF-O	0.213	0.373
CNN-VF [5], [27]	0.376	0.534
BOI [3]	0.548	0.645
CNN-VF+BOI [46]	0.588	0.672
DIF [39]	0.656	0.725
Bimodal DBN	0.682	0.731
Joint-VF-NVF	0.721	0.814
Cu-Joint-VF-NVF	0.750	0.837
M^3TDBN	0.789	0.853

1) CNN-VF-O: CNN Visual Features. The features are generated by fine-tuning the AlexNet network using our 20,000 training set.
2) CNN-VF: CNN Visual Features [5], [27]. Similar to CNN-VF-O, the features are generated by fine-tuning the AlexNet network, but use the Food 101 dataset.
3) BOI: Bag Of Ingredients [3]. Each item is represented by 3000-dimensional Bag Of Ingredients.
4) CNN-VF+BOI [46]. This baseline concatenates the CNN-VF and BOI as the features.
5) DIF: Deep Ingredient Features [39]. Considering the BOI ingredients as the input, this baseline uses the learned hidden representation of one-layer RSM and another one-layer RBM from our model trained on only the ingredients.
6) Multimodal DBN: [34], [47]. Considering the visual features CNN-VF and textual features BOI as the input, This baseline is trained on the image and ingredient modalities without annotating visible ingredients and non-visible ingredients.
7) Joint-VF-NVF: Joint Visual Representation and Non-Visual Representation. Considering the visual features CNN-VF and textual features BOI as the input, this method is trained on both the visual and ingredient information with differentiating between visible ingredients and non-visible ingredients. The visual representation is obtained from learning the joint representation between visible ingredients and the non-visual representation from learning the ingredient.
8) Cu-Joint-VF-NVF. This baseline is similar to Joint-VF-NVF, but fine-tunes the network by the cuisine attributes.

The results are shown in Table IV. From these comparison results, some observations and analysis are included as follows: 1) The performance of CNN-VF-O is lower than CNN-VF. The probable reason is that the training set for fine-tuning the AlexNet is not enough. 2) The performance of BOI outperforms CNN-VF. This shows that ingredient based features are more discriminative than visual features. 3) CNN-VF+BOI outperforms both CNN-VF and BOI. This shows that the performance of combination of multimodality features is better than one-single features. When the ingredients from different cuisines are probably similar, as a complement, the

visual content can improve the accuracy for cuisine classification. 4) The performance of Joint-VF-NVF is better than Multimodal DBN. This means that differentiating between visible ingredients and non-visible ingredients is useful. Multimodal DBN maps the non-visual ingredient features and visual features into the same space, resulting in less efficient joint representation. 5) The proposed M^3TDBN outperforms all other baselines. There is a 5% improvement in MAP and 3 % improvement in Pre@50 compared with the best baseline. This result coincides with the motivation we introduce at the beginning of the paper: Firstly, M^3TDBN differentiates between two kinds of ingredients and thus can learn robust visual and non-visual joint representation. Secondly, M^3TDBN incorporates multitask learning into our model, where the course and the cuisine information collaborate each to enhance the correlation between the images and ingredients, leading to improved cuisine classification. In addition, we also tested the performance using the deep network GoogLeNet [48] fine-tuned by the food101 in our experiment. Because of better power of GoogLeNet than AlexNet, the MAP and Pre@50 of our model achieved 0.803 and 0.871, respectively.

D. Attribute-Augmented Cross-Modal Recipe Image Retrieval

We divided the test set into two parts: one part contains the ingredients, course and cuisine information, and the other part contains the food images. We then use the ingredients, course and cuisine information as the query to retrieve food images.

As for the evaluation metric, since there is only one groundtruth match between the query $<$ ingredients, course, cuisine $>$ and the recipe image, we rely on the position of the groundtruth recipe image in the ranked list to evaluate the performance. Similar to [49], we use Top K % for evaluation, which is the relative number of images $<$ ingredients, course, cuisine $>$ correctly retrieved in the first K% of the ranked list. Some work such as [50] has used this metric in cross-modal retrieval tasks. Specifically, we set $K \in \{1,\ 10,\ 20,\ 40,\ 60,\ 80\}$ to give us the accurate rate. Based on the learned representation, the cosine similarity is calculated to obtain the ranked results.

We consider the following baselines for comparison.

1) Multimodal DBN [34], [47]. A Multimodal DBN is obtained by connecting image and ingredient features with a joint layer. The model neither differentiates between visible ingredients and non-visible ingredients nor incorporates attributes information in the deep architecture. Therefore, in this baseline, we can generate the food images only on a given ingredient list.
2) Corr-AE [50]. Different from Multimodal DBN, this method adopts the architecture of autoencoder [51]. This baseline differentiates common information and modality-specific information through the introduced constraints, but does not consider the attribute information.
3) Joint-VF-NVF. This baseline uses Pathway-A in Fig. 3(a) and differentiates between visible ingredients and non-visible ingredients, but not incorporates attribute information in the deep architecture.

TABLE V
PERFORMANCES OF ALL DIFFERENT ALGORITHMS FOR ATTRIBUTE-AUGMENTED CROSS-MODAL RECIPE IMAGE RETRIEVAL WITH TOP K%

Method	K = 1	K = 10	K = 20	K = 40	K = 60	K = 80
Multimodal DBN	0.014	0.126	0.230	0.415	0.600	0.800
Corr-AE	0.015	0.130	0.241	0.440	0.620	0.825
Joint-VF-NVF	0.017	0.133	0.249	0.453	0.633	0.834
Cu-Joint-VF	0.017	0.140	0.253	0.466	0.642	0.836
Cu-Joint-VF-NVF	0.017	0.140	0.255	0.475	0.664	0.840
Co-Joint-VF	0.019	0.140	0.270	0.486	0.680	0.855
Co-Joint-VF-NVF	0.019	0.140	0.274	0.499	0.686	0.864
M^3TDBN	0.020	0.168	0.304	0.523	0.706	0.866

4) Cu-Joint-VF. This baseline uses Pathway-A in Fig. 3(a) and introduces the cuisine attribute at the top layer, but does not use the non-visible ingredients.
5) Cu-Joint-VF-NVF. This baseline uses both Pathway-A and Pathway-B in Fig. 3(a), and introduces the cuisine attribute at the top layer.
6) Co-Joint-VF. This baseline uses Pathway-A in Fig. 3(a) and introduces the course attribute at the top layer, but does not use the non-visible ingredients.
7) Co-Joint-VF-NVF. This baseline uses both Pathway-A and Pathway-B in Fig. 3(a), and introduces the course attribute at the top layer.

Not that for all the baselines, the input of visual features and textual features is the visual features CNN-VF and textual features BOI, respectively. Table V summarizes the Top K% results of attribute-augmented cross-modal recipe image retrieval task. We can see that the proposed M^3TDBN consistently outperforms other baselines with different k values. Joint-VF-NVF is better than Corr-AE. Because Corr-AE introduces the correlation cost to differentiate between common information and modality-specific information (such as non-visible ingredients) while we group ingredients into two parts in advance for more accurate correlation learning. When we further consider the attributes, the performance is improved. We can also see that from K = 20, the performance of Cu-joint-VF-NVF is better than Cu-Joint-VF. The reason is that the network builds the correlation between the visual layer and the non-visual layer through the attribute layer. For the query including non-visual ingredients, the non-visual ingredients and visual ingredients work together to retrieve food images. Similarly, the performance of Co-joint-VF-NVF is better than Co-Joint-VF. Furthermore, we found that the performance of Co-Joint-VF-NVF is better than Cu-Joint-VF-NVF. The reason may be that the course information is more specific than cuisine, and thus more discriminative. After incorporating two attributes in a multitask framework, our method M^3TDBN achieves the best performance, and outperforms the best baseline by 10% when K = 20.

Fig. 5 shows three examples of our M^3TDBN and the best baseline method. We observed that in these examples, our method can retrieval relevant images in top-3 ranked results. In contrast, there are no groundtruth results in some cases for the best baseline method. This again verified the effectiveness of our method.

Fig. 5. Three recipe image retrieval examples using our M^3TDBN and best baseline method. In each example, the query ingredients and its corresponding cuisine and course information are shown on the left; retrieved food images of our M^3TDBN are presented on the top; retrieved images of the best baseline model are presented at the bottom. Ranked score is shown below each image. The groundtruth food images are highlighted with red color.

E. Ingredient and Attribute Inference From Food Images

The learned model can be applied to infer the ingredients, courses and cuisines from the recipe image. We evaluate the effectiveness of M^3TDBN in ingredient and attribute inference from food images.

We can perform Gibbs sampling use (28)–(33) to draw samples of $\mathbf{v}^{(m)}$, $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$. The returned results are ranked with probability. Therefore, we resort to MAP@K for evaluation. Here MAP@K is defined as follows:

$$MAP@K == \frac{1}{Q}\sum_{q=1}^{Q}\frac{\sum_{k=1}^{K} Precision@qk * r(qk)}{\sum_{k=1}^{K} r(qk)} \quad (35)$$

where Q is the number of queries. r_k is the relevance level at position k, which is 1 if the ingredient (or attributes) is in the list of groundtruth ingredients (or attributes) from the query image, and 0 otherwise. r_{qk} is the relevance level at position k for query q. $Precision@qk$ is the precision at position k for the query q, and K is the truncation level. In our experiment, we use MAP@10 for ingredients and MAP@3 for cuisine and course, respectively. We select five methods CNN-VF, Multimodal DBN, Joint-VF-NVF, Cu-Joint-VF-NVF and Co-Joint-VF-NVF as baselines for comparison. The baseline CNN-VF considers the ingredients, cuisine and course information as the supervised label information and uses the logistic regression to obtain the score of each label for each kind of supervised information, respectively.

We firstly evaluate the performance of ingredient inference.

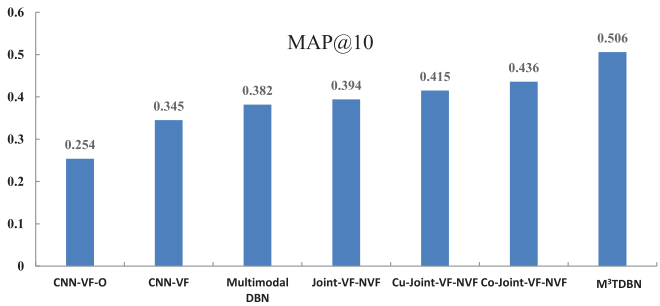


Fig. 6. Performance of ingredient inference from food images using MAP@10.

TABLE VI
PERFORMANCE OF ATTRIBUTES INFERENCE USING MAP@3

Method	Cuisine	Course
CNN-VF	0.403	0.432
Co-Joint-VF-NVF	–	0.515
Cu-Joint-VF-NVF	0.425	–
M^3TDBN	0.472	0.532

Fig. 6 shows the experimental results. We can see that 1) introducing the cuisine or course information leads to the improvement of performance. The baselines Cu-Joint-VF-NVF and Co-Joint-VF-NVF outperforms Joint-VF-NVF by 5% and 10%, respectively. This shows the regularization effect by the attributes. 2) our model achieves the best performance than

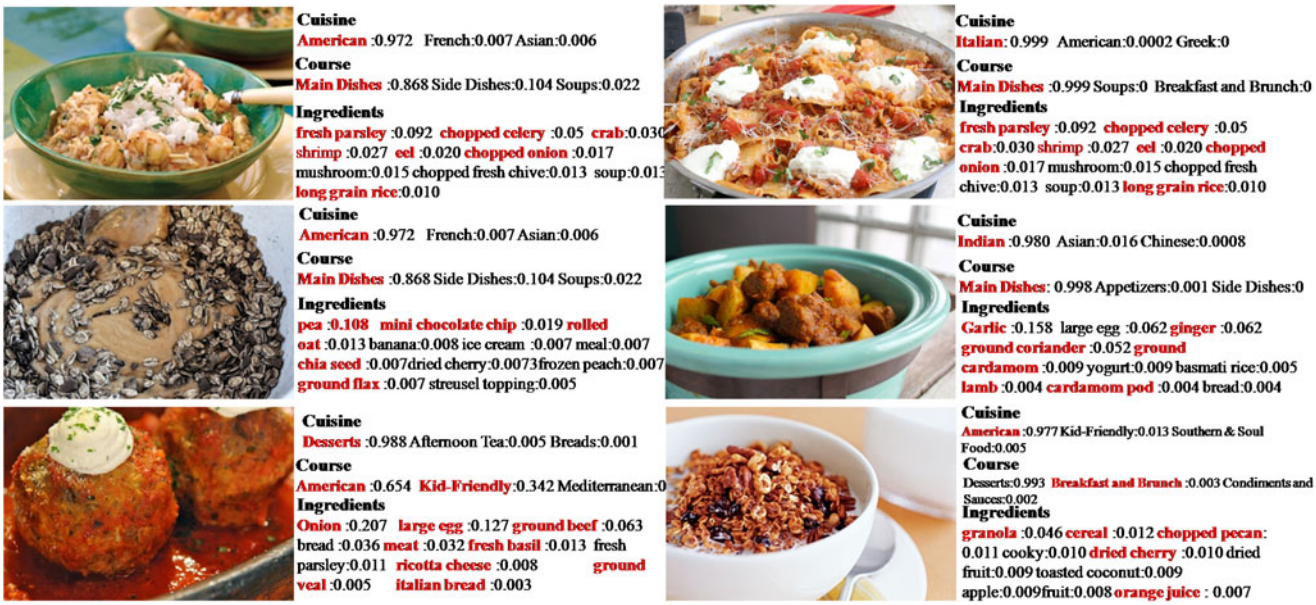


Fig. 7. Some examples of ingredient and attribute inference from food images. The probability value derived from our model is provided behind the inference result. The groundtruth labels are highlighted with red color.

the best baseline, and outperforms it by 16%. This shows that the course and cuisine attributes enforce each other in a multi-task fashion. This again validates the advantage of the proposed method in incorporating two kinds of attributes.

We then show the experimental results on attribute inference. From Table VI, we can see that M^3TDBN introduces the multi-task learning to make two attributes enforce each other and thus has a better performance.

Fig. 7 shows some examples of inferred results by our M^3TDBN. In these cases, In top-10 ranked ingredients, at least 6 ingredients are the ground truth. The first returned results in attribute inference by the query are almost ground truth. For each attribute, even there are more than one values, our method can also infer accurate results, for example, our model inferred two values "American" and "Kid-Friendly" for the course attribute of the food image in the left bottom.

VI. Conclusion

In this paper, we proposed a MultiModal MultiTask Deep Belief Network (M^3TDBN) to explore multimodality content and multi-attribute information in the food domain. This deep network consists of two pathways, where one pathway is to learn joint mid-level representation between visual and visible ingredients, and the other is mainly to learn non-visual mid-level representation. In addition, M^3TDBN incorporates multitask learning to make different attributes collaborate each other. We applied M^3TDBN into three extended novel problems, including multimodal cuisine classification, attribute-augmented cross-modal recipe image retrieval, ingredient and attribute inference from food images. The experimental results demonstrate the effectiveness of our model in the learned mid-level representation and discovered correlation.

This work can be extended in the following four directions: 1) Exploring more information from Yummly for supporting more applications. For example, we can utilize the accurate calorie information for improving the performance of existing food image calorie estimation [27]; 2) Adding user information into our model for personalized food or recipe recommendation [12], [13] and food balance estimation [52] based on the individual's taste, ingredient, diet, allergy, nutrition, calories and so on. 3) The proposed recipe retrieval and exploration framework provides one attempt to address some challenging problems in multimodality and multi-attributes analysis. With the fast development of Artificial Intelligence (AI), it is important to find the correlations with emerging methods from AI-related areas in the future. 4) Since our proposed framework on multimodality and multi-attributes modeling can be easily generalized, we hope to apply our framework into other fields such as Flickr images with multimodality and multi-context information (each kind of context information can be considered as one attribute) to enable interesting and valuable analysis.

References

[1] P. Rozin, C. Fischler, S. Imada, A. Sarubin, and A. Wrzesniewski, "Attitudes to food and the role of food in life in the USA, Japan,Flemish Belgium and France: Possible implications for the diet–health debate," *Appetite*, vol. 33, no. 2, pp. 163–180, 1999.

[2] K. Aizawa and M. Ogawa, "Foodlog: Multimedia tool for healthcare applications," *IEEE Multimedia Mag.*, vol. 22, no. 2, pp. 4–8, Apr.–Jun. 2015.

[3] H. Su, T.-W. Lin, C.-T. Li, M.-K. Shan, and J. Chang, "Automatic recipe cuisine classification by ingredients," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Adjunct Pub.*, 2014, pp. 565–570.

[4] F.-F. Kuo, C.-T. Li, M.-K. Shan, and S.-Y. Lee, "Intelligent menu planning: Recommending set of recipes by ingredients," in *Proc. ACM Multimedia Workshop Multimedia Cooking Eating Activities*, 2012, pp. 1–6.

[5] J. Hessel, N. Savva, and M. J. Wilber, "Image representations and new domains in neural image captioning," *CoRR*, 2015. [Online]. Available: http://arxiv.org/abs/1508.02091

[6] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[7] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[8] A. Fischer and C. Igel, "Training restricted Boltzmann machines: An introduction," *Pattern Recog.*, vol. 47, no. 1, pp. 25–39, 2014.

[9] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, "Flavor network and the principles of food pairing," *Scientific Rep.*, vol. 1, 2011, Art. no. 196.

[10] L. Wang, Q. Li, N. Li, G. Dong, and Y. Yang, "Substructure similarity measurement in chinese recipes," in *Proc. 17th ACM Int. Conf. World Wide Web*, 2008, pp. 979–988.

[11] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Proc. 4th Annu. ACM Web Sci. Conf.*, 2012, pp. 298–307.

[12] P. Forbes and M. Zhu, "Content-boosted matrix factorization for recommender systems: Experiments with recipe recommendation," in *Proc. 5th ACM Conf. Recommender Syst.*, 2011, pp. 261–264.

[13] C.-J. Lin, T.-T. Kuo, and S.-D. Lin, "A content-based matrix factorization model for recipe recommendation," in *Advances in Knowledge Discovery and Data Mining*. Cham, Switzerland: Springer, 2014, pp. 560–571.

[14] G. M. Farinella, M. Moltisanti, and S. Battiato, "Classifying food images represented as bag of textons," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 5212–5216.

[15] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 2249–2256.

[16] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 1085–1088.

[17] S. Christodoulidis, M. Anthimopoulos, and S. Mougiakakou, "Food recognition for dietary assessment using deep convolutional neural networks," in *Proc. Int. Conf. Image Anal. Process.*, 2015, pp. 458–465.

[18] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 446–461.

[19] V. Bettadapura, E. Thomaz, A. Parnami, G. D. Abowd, and I. Essa, "Leveraging context to support automated food recognition in restaurants," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 580–587.

[20] O. Beijbom, N. Joshi, D. Morris, S. Saponas, and S. Khullar, "Menu-Match: Restaurant-specific food logging from images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 844–851.

[21] R. Xu *et al.*, "Geolocalized modeling for dish recognition," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1187–1199, Aug. 2015.

[22] L. Herranz, R. Xu, and S. Jiang, "A probabilistic model for food image recognition in restaurants," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun.-Jul. 2015, pp. 1–6.

[23] L. Oliveira, "A mobile, lightweight, poll-based food identification system," *Pattern Recog.*, vol. 47, no. 5, pp. 1941–1952, 2014.

[24] Y. Kawano and K. Yanai, "Foodcam-256: A large-scale real-time mobile food recognition system employing high-dimensional features and compression of classifier weights," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 761–762.

[25] T. Maruyama, Y. Kawano, and K. Yanai, "Real-time mobile recipe recommendation system using food ingredient recognition," in *Proc. 2nd ACM Int. Workshop Interactive Multimedia Mobile Portable Devices*, 2012, pp. 27–34.

[26] J. Dehais, M. Anthimopoulos, and S. Mougiakakou, "Dish detection and segmentation for dietary assessment on smartphones," in *Proc. Int. Conf. Image Anal. Process.*, 2015, pp. 433–440.

[27] A. Meyers *et al.*, "Im2calories: Towards an automated mobile vision food diary," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1233–1241.

[28] X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso, "Recipe recognition with large multimodal food dataset," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jun.-Jul. 2015, pp. 1–6.

[29] S. Amano, K. Aizawa, and M. Ogawa, "Food category representatives: Extracting categories from meal names in food recordings and recipe data," in *Proc. IEEE Int. Conf. Multimedia Big Data*, Apr. 2015, pp. 48–55.

[30] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proc. ACM Multimedia Conf.*, 2016, pp. 32–41.

[31] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[32] Y. Liu, Q. Tong, Z. Du, and L. Hu, "Content-boosted restricted Boltzmann machine for recommendation," in *Proc. Int. Conf. Artif. Neural Netw.*, 2014, pp. 773–780.

[33] X. Liu *et al.*, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," in *Proc. Annu. Conf. North Amer. Chapter ACL*, 2015, pp. 912–921.

[34] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2949–2980, 2014.

[35] Y. Huang, W. Wang, and L. Wang, "Unconstrained multimodal multi-label learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1923–1935, Nov. 2015.

[36] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.

[37] Z. Yuan, J. Sang, C. Xu, and Y. Liu, "A unified framework of latent feature learning in social media," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1624–1635, Oct. 2014.

[38] G. E. Hinton and R. R. Salakhutdinov, "Replicated softmax: An undirected topic model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1607–1614.

[39] G. E. Hinton and R. Salakhutdinov, "Replicated softmax: An undirected topic model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1607–1614.

[40] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[42] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[43] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.

[44] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, 2010, Art. no. 926.

[45] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: The MIR Flickr retrieval evaluation initiative," in *Proc. Int. Conf. Multimedia Inf. Retrieval*, 2010, pp. 527–536.

[46] L. Yang *et al.*, "Plateclick: Bootstrapping food preferences through an adaptive visual interface," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 183–192.

[47] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Proc. ICML Representation Learn. Workshop*, 2012.

[48] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1–9.

[49] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2407–2414.

[50] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 7–16.

[51] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[52] K. Aizawa, Y. Maruyama, H. Li, and C. Morikawa, "Food balance estimation by using personal dietary tendencies in a multimedia food log," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2176–2185, Dec. 2013.

Weiqing Min received the B.E. degree from Shandong Normal University, Jinan, China, in 2008, and the M.E. degree from Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015.

He is currently a Postdoc with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences. His current research interests include multimedia search and mining, and multimodal learning.

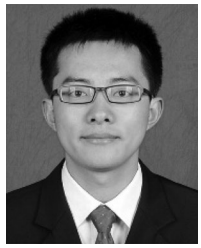
Shuqiang Jiang (S'04–M'06–SM'08) is a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, and a Professor with the University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. He was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008, NSFC Excellent Young Scientists Fund in 2013, and Young top-notch talent of Ten Thousand Talent Program in 2014. He has authored or coauthored more than 100 papers on the related research topics. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision.

Prof. Jiang is a Senior Member of CCF, a Member of ACM, and an Associate Editor of the IEEE MULTIMEDIA, MULTIMEDIA TOOLS AND APPLICATIONS. He is the General Secretary of the IEEE CASS Beijing Chapter and Vice Chair of the ACM SIGMM China chapter. He is the General Chair of ICIMCS 2015, Program Chair of ICIMCS2010, Special Session Chair of PCM2008 and ICIMCS2012, Area Chair of PCIVT2011, Publicity Chair of PCM2011, Web Chair of IS-CAS2013, and Proceedings Chair of MMSP2011. He has also served as a TPC Member for more than 20 well known conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICIP, and PCM. He was the recipient of the Lu Jiaxi Young Talent Award from the Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012.

Jitao Sang received the B.E. degree from the South-East University, Dhaka, Bangladesh, in 2007, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2012 with the highest honor, the Special Prize of President Scholarship.

He is an Associate Professor with the National Laboratory of Pattern Recognition, CASIA. He has authored one book, filed three patents, and coauthored more than 60 peer-referenced papers in multimedia related journals and conferences. His research interests include social multimedia computing, user modeling, and web data mining.

Prof. Sang is an Associate Editor of *Neurocomputing* and *KSII Transactions on Internet and Information Systems*. He served as a Program Co-Chair for PCM 2015 and ICIMCS 2015. He was the recipient of the Best Paper Finalist in MM 2012 and MM 2013, Best Student Paper in MMM 2013, Best Student Paper in ICMR 2015, and Best Paper in PCM 2016.



Huayang Wang received the B.S degree in computer science from the Shandong University of Science and Technology, Qingdao, China, in 2013, and is currently working toward M.S. degree at the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

His current research interests include multimedia content analysis, computer vision, and multimodal knowledge graph.

Xinda Liu received the B.E. degree from the China University of Mining and Technology, Beijing, China, in 2013, the M.E. degree from Ningxia University, Yinchuan, China, in 2016, and is currently working toward the Ph.D. degree at the China State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science & Engineering, Beijing University of Aeronautics & Astronautics, Beijing, China.

His research interests include machine learning and image processing.

Luis Herranz received the Telecommunication Engineer degree from the Universidad Politcnica de Madrid, Madrid, Spain, in 2003, and the Ph.D. degree in computer science and telecommunication from the Universidad Autnoma de Madrid, Madrid, Spain, in 2010.

From 2003 to 2010, he was a Researcher and a Teaching Assistant with the Escuela Politcnica Superior, Universidad Autnoma de Madrid. From 2010 to 2011, he was a Research Engineer with Mitsubishi Electric R&D Centre Europe, Livingston, U.K. From 2012 to 2016, he is a Postdoctoral Research Fellow with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His current research interests include image understanding, video abstraction, and multimedia indexing and retrieval.