Supplementing Omitted Named Entities in Cooking Procedural Text with Attached Images

Yixin Zhang
Kyoto University
Kyoto, Japan
zhangyx@dl.soc.i.kyoto-u.ac.jp

Yoko Yamakata
The University of Tokyo
Tokyo, Japan
yamakata@mi.u-tokyo.ac.jp

Keishi Tajima
Kyoto University
Kyoto, Japan
tajima@i.kyoto-u.ac.jp

***Abstract*—In this research, we aim at complementing named entities, such as food, omitted in the procedural text of recipe data. It helps users understand the recipe and is also necessary for the machine to automatically understanding the recipe data. The main work of this research is as follows. (1) We construct a dataset of Chinese recipes consisting of 12,548 recipes. To detect sentences in which food entities are omitted, we label named entities such as food, tool, and cooking actions in the procedural text by using the automatic recipe named entity recognition method. (2) We propose a method of recognizing food from the attached images. A procedural text of recipe data is often associated with an image, and the attached image often contains the food even when it is omitted in the procedural text. Tool entities in images in recipe data can be identified with high accuracy by conventional general object recognition techniques. On the other hand, the general object recognition methods in the literature, which assume that the properties of an object are constant, perform not well for food in recipe image data because food states change during cooking procedures. To solve this problem, we propose a method of obtaining food entity candidates from other steps that are similar to the target step, both in sentence similarity and image feature similarity. Among all the 246,195 procedural steps in our dataset, there are 16,593 steps in which the food entity is omitted in the procedural text. Our method is applied to complement the food entities in these steps and achieves the accuracy of 70.19%.**

***Index Terms*—Recipe Data, Recipe Named Entity, Word Complement**

I. Introduction

In recent years, many user-submitted recipe sites, such as Allrecipes[1] in North America and the UK, Cookpad[2] in Japan, and Haodou[3] in China, have become popular. Nowadays, several millions of recipe data posted by users are shared on such recipe sites. Each recipe data entry consists of titles of recipes, outlines, material and ingredient lists, cooking procedures, and tips. Many recipe data are multi-modal and include both text and images. In some cases, images are attached not only with the captions but also with specific procedural steps in the recipe. Figure 1 shows an example of such recipe data on Haodou Recipe Site. On this site, each procedural step is attached with a corresponding image in every recipe.

[1]https://www.allrecipes.com/
[2]http://cookpad.com/
[3]http://www.haodou.com/recipe/

Fig. 1. Example of recipe data posted on Haodou.

The contents of recipes are uniformly structured, and it helps us to extract information from it. In addition, this kind of recipe with a one-to-one correspondence between images and procedural steps contains rich and valuable multimedia information regarding food. As a result, research on extracting and analyzing useful information in recipe data has become an important research issue in natural language processing and computer vision communities [2], [13].

Problem Statement: In this work, we study the problem of complementing food entities omitted in cooking procedural text associated with images.

Some literature has investigated the automatic understanding of recipe data recently [2], [6], [9]. With the rise of smart

devices (e.g., smart speakers and smartphone voice assistants) and image retrieval methods, more and more users are using these devices and methods to retrieve recipe information such as specific cooking procedures or ingredients in a meal. The method of analyzing these multimedia data is useful in numerous recipe retrieval applications or scenarios and provides important implications in natural language processing and image recognition in general.

However, the proposed problem has a few challenges and also provides more research opportunities:

- When users upload recipes onto these websites, partly due to users' word habits, text descriptions often include typos and infrequent expressions, and sometimes even omit some important information such as tool or food entities. When we use smart devices (e.g., smart speakers) for automatic reading of recipes and cooking support, it is difficult for a machine to explain the omitted entities in the sentences. For example, if we ask for a specific step, and the smart speaker responds with "cut into slices", we may need to ask "cut what into slices?" and the machine may not be able to answer. However, those omitted food entities in text descriptions are sometimes shown in the attached images.
- Object recognition is a useful method to recognize named entities in images for automatically complementing text descriptions. It is useful for tool entities. However, food state and shapes are always changing during cooking procedures. The existing image recognition method barely takes this problem into account and could not be able to achieve high accuracy in some situations. In this paper, we develop a method of complementing procedural text with omitted food entities by obtaining food entity candidates from other steps that are similar to the target step both in sentence similarity and image feature similarity.

The main contributions of this paper can be summarized as follows: (i) We construct our own text-image procedural recipe dataset in Chinese, which provides a wealth of data and language options for future research. (ii) We develop a recipe Named Entity (r-NE) recognizer [14] in Chinese after in Japanese and English, which improves language versatility in the recipe research field. (iii) To solve the difficulty of recognition of food in different cooking stages, we propose a method of obtaining food entity candidates from other steps that are similar to the target step, both in sentence similarity and image feature similarity.

When a food entity is omitted in the target step, the proposed method aims to find steps which have similar sentence and image with the target step in order to obtain their food entities as candidates. The likelihood of the candidates is brought from the similarities of sentences and images. Then the candidates could be narrowed down using the ingredient list of the target recipe, and finally, the most likely food entity is selected.

The rest of this paper is organized as follows. Related work is reviewed in Section II. Section III explains the details of our proposed methods. After that, experiments and evaluation results are shown in Section IV. Section V summarizes this paper and discuss the future work.

II. Related Work

Generally speaking, information complementing in multi-modal recipe descriptions using features of procedural images is a practical research topic. It is related to the following research lines.

A. Recipe Text Processing

There has been researched on recipe text processing. Recipe text processing has some differences from general text processing, which makes it difficult to apply the existing text processing method easily to recipe data [8]. A specific method for the recipe domain, which could even be applied for the multilingual environment, is desired.

The analysis of a set of words attached to images is also a research issue nowadays, such as tag identification from a tag set attached with an image [7] and inferring the semantic relationship between them [5]. They focus on the data from image posting sites like Flickr, where the images are attached with tags already. In this paper, we use the text description along with the attached images and try to infer the relationship between images and text and complement the word omitted in the text.

B. Recipe Image Recognition

Currently, research on the recognition of images in cooking recipes mainly focuses on the whole dish's appearance without explicit analysis of ingredient composition [2]. Ingredient and material estimation only from a completed food image, is a task far harder than food categorization. Our method intends to recognize the food omitted in the intermediate procedural steps, i.e., not a final completed dish, in order to complement the needed information in text data.

C. Multi-modal Correlation Learning

Multi-modal Correlation Learning between images and texts is a hot research issue in computer vision and natural language processing in recent years, such as CCA [4], which study the general cross-modal correlation between images and text information and Bi-linear model [12]. However, multi-modal learning about procedural text-image data in recipes has more unique characteristics compared with general problems. For example, food is changing gradually with procedures. This means that the study of recipes requires a more specific analysis and thus could be applied in practice better.

In summary, recipe text processing and image feature analysis are important issues, and in this paper, we specifically focus on the automatic complementing of food entities in procedural steps attached to images, which needs specific consideration and has not been studied in the existing research.

III. Our Methods

In this section, we explain some details of our proposed methods of complementing omitted food entities in recipe data.

TABLE I
RECIPE NAMED ENTITY (R-NE) TAGS

Tag	Meaning	Remarks
Ac	Action by chef	Verb representing a chef's action
Ac2	Discontinuous Ac	Second, non-contiguous part of a single action by chef
F	Food	Eatable, also intermediate products
T	Tool	Knife, container, etc.
Sf	Food state	Food's initial or intermediate state
St	Tool state	Tool's initial or intermediate state
D	Duration	Duration of cooking
Q	Quantity	Quantity of food
At	Action by tool	Verb representing a tool's action
Af	Action by chef	Verb representing action of a food

A. Dataset

We collected 12,548 recipe data posted on Haodou Recipe[4], a user-submitted recipe site in China. Each data item consists of the following components: a recipe ID, a general description, ingredients, tips, and a sequence of cooking procedural steps, each of which is a pair of a text description and an optional image. Figure 1 shows an example.

We extract text data of the procedural steps, segment Chinese sentences into words, and add the Part-of-Speech tagging (POS tagging) by using Chinese language segmentation and POS tagging tool named jieba[5].

B. Target Sentence Detection

To complement omitted entities, we first need to detect sentences in which food entities are omitted. We use the Recipe Named Entity Recognizer method to annotate the texts. Recipe Named Entity, which is an example of a domain-specific Named Entity (NE) definition for the recipe, is used to recognize tokens of food, tools, actions performed by a chef, and so on in a procedural text. Mori et al. constructed a Japanese recipe corpus consisting of 208 recipes randomly sampled from the Cookpad website [8] and Yamakata et al. adopted it into English and extended it by adding two more tags, Ac2 and At, in order to account for additional phenomena in English text [14]. In this paper, we adopt this method to recipes procedural text in the Chinese language.

We first ask Chinese native speakers to annotate 50 recipes according to the guidelines for English recipes [14]. Table I shows ten types of Recipe Named Entity (r-NE) and their meanings. We then train a named entity recognizer model named BERT-NER[6] (a state-of-the-art named entity recognizer which is constructed based on the BERT neural network architecture [3]) on Chinese recipe corpus in our dataset.

For example, we could annotate the sentence "豆芽摘洗干净 " as "豆芽**/F** 摘洗**/Ac** 干净**/Sf** ", which means "Pick**/Ac** and wash**/Ac** the sprouts**/F** clean**/Sf**". In this way, we can clearly obtain words and their tags in order to easily detect sentences where food entities are omitted.

[4]http://www.haodou.com/recipe/

[5]https://github.com/fxsjy/jieba

[6]https://github.com/kyzhouhzau/BERT-NER

C. Vectorization of Text and Associated Images

As shown in Figure 2, the main processing flow of our proposed method is as follows:

- For each procedural step i in which food entity is omitted (target sentence), we vectorize pairs of text s_i and image g_i as v_{sent_i} and v_{img_i} and obtain the corresponding ingredient list Ing.
- For each $sent_i$, we compute the similarities between $sent_i$ and sentences in recipe dataset and obtain the sentence similarity ranking list sim_{sent_i} along with food entities.
- For each img_i, we compute the similarities between img_i and sentences in recipe dataset and obtain the image similarity ranking list sim_{img_i} along with food entities.
- Then we merge the above two similarity lists in order to obtain the final candidate list using $\alpha sim_{sent_i} + \beta sim_{img_i}$, where $\alpha = 0.7$ and $\beta = 0.3$
- For the food entity obtained in step 3, we narrow down the candidates by using the ingredient list Ing. The name of the ingredient with the highest similarity is assumed to be the omitted food entity.

We use sentence embedding to process the text of procedural steps. We use one of the standard sentence-embedding method, Sentence2vec [1]. We learn sentence-embedding by using the corpus of all procedural text data in our recipe data and transform each sentence into a 100-dimensional vector.

sfccGiven word embedding vectors of procedural step sentences, the similarity between two sentences is computed by the cosine similarity of two vectors. In the procedural step sentence dataset, we form a sentence set S which contains 300 sentences s_i, and select the top 10 sentences $s_{i,j}$ with high similarity with them, then vectorize them as $v_{s_{i,j}}$ in set V_{s_i} (for $i = 0, \ldots, 299$ and $j = 0, \ldots, 9$).

For the image data, we vectorize the associated images by using a convolutional neural network VGG16 [10], which is trained on ImageNet data and widely used for image recognition. We use the output of two fully-connected layers, which is a 4,096-dimensional vector. For each sentence $s_{i,j}$, we vectorize the associated image $g_{i,j}$ as $v_{g_{i,j}}$ in set V_{g_i} (for $i = 0, \ldots, 299$ and $j = 0, \ldots, 9$).

D. Food Entity Complement

We first traverse all the text data that has been tagged by BERT-NER. For a procedural step where food entity is omitted, we compute the similarity between the feature vector of image v_g associated with this step and the feature vectors in each set V_{g_i} from V_{g_0} to $V_{g_{299}}$. We calculate the average value of similarities between the vector v_g and each vector in V_{g_i}. We select set G_i, which has the highest average similarity.

The likelihood of the candidates is brought from the similarities of images in set G_i. We extract the food entities from corresponding sentences. Then the candidates are narrow down with the ingredient list Ing_n in which this step is located. The name of the ingredient with the highest similarity is assumed to be the omitted food entity.

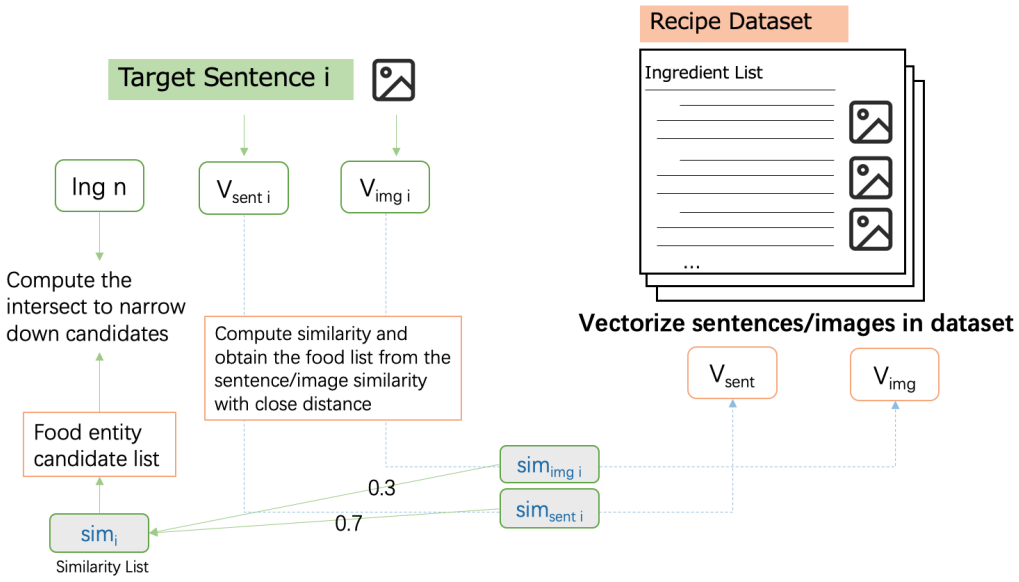

Fig. 2. Processing flow of our proposed method

IV. Experiment

In this section, we explain the contents of the experiment and then discuss the results of the experiment in order to evaluate the proposed method.

A. Sentence Detection and Recognition Accuracy of r-NE

There are 12,548 recipes in our dataset. Among all the 246,195 steps, there are 16,593 steps in which food entities are omitted in the text.

In sentences in which the food entity is not missing, there are 5,685 distinct food names, where 24 food names appear more than 1,000 times, 298 food names more than 100 times but less than 1000 times, and 5,355 food names only 50 times or less, which accounts for 94.20% in total.

To evaluate the accuracy of the r-NE recognizer in Chinese, we process 2142 tokens with 1472 phrases, and the number of correct ones is 1283. The results show that the overall accuracy is 91.27%, the precision is 87.28%, recall is 87.16%, and F1 score is 87.22. For each entity, the precision, recall, and F1 score are as shown in table II.

Figure 3 shows the proportions of each r-NE tag in the Chinese corpus. We could find that, besides the Ac (action verbs) tag, food tags have high proportions. Therefore, food entities play an important role in the recipe data analysis, and if this kind of information is omitted, it will lead to great challenges in understanding the semantics for machines.

B. Food Entity Complement

Among the candidates of r-NE tags, Ac, F and T account for a large proportion since these are the most basic but important information in recipe text. Therefore the analysis of information containing these attributes is very valuable.

TABLE II
Recipe Named Entity (r-NE) Recognition Accuracy

Tag	Precision	Recall	F1 score
Total	87.28%	87.16%	87.22
Ac	93.70%	92.94%	93.32
Ac2	68.97%	76.92%	72.73
F	83.15%	93.77%	90.34
T	87.10%	76.06%	81.20
Sf	73.23%	76.23%	74.70
St	58.82%	47.62%	52.63
D	92.31%	94.74%	93.51
Q	88.89%	100.00%	94.12
Af	0%	0%	0
At	0%	0%	0

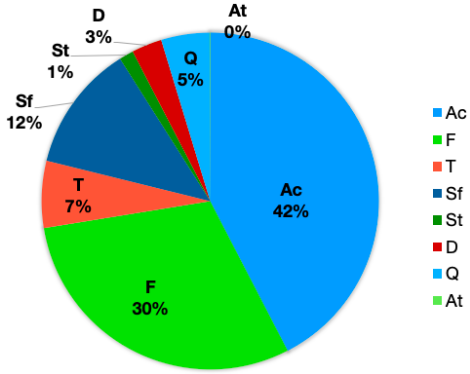

Fig. 3. Proportions of the 10 r-NE tag types

As mentioned in Section III-C, we propose a method of obtaining food entity candidates from other steps that are similar to the target step both in sentence similarity and image feature similarity.

From the experiment result, we can know that, among the

TABLE III
EXAMPLE: SIMILAR PROCEDURAL SENTENCES BY SENTENCE2VEC

Add 10g of water 加入10 克清水

	text		sim
0	Add 10g water	加入10克清水	1.0000
1	Add 30g water	加入30克清水	0.8989
2	Add 20g water	加入20克清水	0.8895
3	Add 800g water	加入800克清水	0.8819
4	Add condensed milk 10g	加入炼乳10克	0.8781
5	Add 30 g water	加入30克清水	0.8683
6	Brown sugar with water 60g	黑糖 加入 清水60克	0.8606
7	Add 70g water	加入70克清水	0.8484
8	Water 10g	水 10 克	0.8424
9	Add 18g water	加入18克清水	0.8415

Quickly add egg pieces 快速加入鸡蛋块

	text		sim
0	Quick Add Egg Pieces	快速加入 鸡蛋块	1.0000
1	Add egg pieces	加入鸡蛋块	0.7512
2	Pour egg pieces	倒入鸡蛋块	0.7086
3	Pour in egg pieces and stir fry	倒入鸡蛋 块翻炒	0.6830
4	Quick Pour Meat Slices	快速倒入肉片	0.6682
5	Put in egg pieces	放入鸡蛋块	0.6613
6	Add egg pieces	加鸡蛋块	0.6355
7	Add 70g water	鸡蛋洗干净	0.6303
8	Wash the eggs	加入鸡蛋 搅拌均匀	0.6269
9	Add egg pieces and fry for a while	加入鸡蛋块 稍微炒一会	0.6237

246,195 procedural steps in our dataset, there are 16,593 steps in which the food entity is omitted in the text. Then we calculate the similarity of both the vectors of text and the vectors of images between these steps and the whole steps in the dataset and obtain the food entity candidates. For the food entity obtained, we narrow down the candidates in the corresponding ingredient list Ing_n of procedural steps in which the food entity is omitted. The name of the ingredient with the highest similarity is assumed to be the omitted food entity.

The example results of the sentence embedding method for calculating text similarity are shown in Table III. We could find that each similar sentence set may contain similar food such as water, egg, and others.

Due to the contextual implication, the author may omit food in some steps, but we can use the relationships between sentences and sentences as well as images and images to recognize the omitted food entities. For instance, as shown in Figure 4, the images on the left corresponds to "stir fry well and stand by (炒散备用)" where the food entity is omitted. Based on the calculating of similarity between the vector of this image and the 300 sets of image vectors, the image set on the right side which is associated with sentences in Table III has the highest similarity with the target image. Therefore the word "egg" is complemented as the food entity into the text.

In order to compute the accuracy, the complemented result

Fig. 4. Food entity complement

TABLE IV
EXAMPLE: COMPARISON OF OUR METHOD AND MANUAL LABELED METHOD

Sentence	Our Method	Manual Labeled
腌制入味 Marinate for flavor	鱼 fish	鱼 fish
切丝备用 Shredding for later use	土豆 potato	土豆 potato
打散 beat	鸡蛋 egg	蛋 egg
放入烤箱 Put in the oven	面团 dough	披萨饼坯 pizza batter
蒸熟后碾碎 Steam and then crush	土豆 potato	南瓜 pumpkin

is then compared with the result of the manual annotation of the food entity. We randomly select 3000 pieces of the procedural step text and annotate them manually according to the corresponding images. Some comparison results are shown in Table IV. According to the comparison, among the steps selected in which the food entity is omitted, 70.19% of the results of our method are consistent with the manual labeled result. Therefore, the accuracy is 70.19 %.

We choose multi-label image classification method by using Inception Net v3 [11] as the ground truth. It is a deep convolutional neural network trained for single-label image classification and trained on ImageNet data[7]. First of all, we prepare the network with correct labels over 20 classes for each image in the training set. Then we retrain the last layer of the model and modify the method of evaluating generated predictions to be actually able to train it with regard to multiple possible correct classes for each image.

The accuracy of Inception Net v3 model is over 86%. We also use the selected 3000 pieces of procedural steps to compute the intersect of the result of the model and the manual result in order to test the accuracy. Inception Net v3 model achieves 43.57% for food. We also compute the intersect of the result of only the sentence-similarity result and the manual result. Sentence-similarity achieves 56.21%. The above two methods are much lower than the proposed method since the limitation of food classes and the shape and states of food are diverse. We take both text and visual similarity to improve recognition accuracy. Our method solves

[7]https://github.com/IntelAI/models/tree/master/benchmarks/imagerecognition/tensorflow/inceptionv3

TABLE V
COMPARE THE RESULT OF METHODS WITH THE MANUAL COMPLEMENT

	Accuracy
Proposed Method	70.19%
Sentence Similarity	56.21%
Inception Net v3 Model	43.57%

the problems mentioned above to a certain extent and obtains higher accuracy.

We could find from the result in Table IV that although the accuracy is higher than the multi-label classification result, in some cases, there may have some misidentification too. For example, when the action contents in several sentences are similar or even the same, these sentences are also classified as a set of high similarity, but the objects, that is, the food, maybe totally different. Also, there may exist different expressions of the same food. This method still has some limitations that need to be solved and improved in the future. The features of changing state of food during being cook could be taken into account in order to improve food identification accuracy.

V. CONCLUSION

In this paper, we propose a method of complementing omitted food entities in procedural text descriptions. We first detect target sentences where food entities are omitted by adopting the method of Recipe Named Entity into Chinese recipes. Then the vectorization of sentences and images is adopted in order to complement the food entity omitted into the procedural text description.

In general, the method proposed in this paper complements the omitted food entity to a certain extent. This method could solve the problem of low identification of food in changing states and shapes to some extent. Because not only do we use image similarity, we also take contextual similarity into account; in other words, we use the contextual information to avoid the problem of limitation of classes in multi-label classification and improve the accuracy. However, there are some factors that can cause misidentification, such as the structure of actions are too similar, but the foods are totally different.

In future work, since food states change over time during cooking, we could use this special feature to greater effect in order to improve the identification of food. Also, a single procedural step associated with a single image often includes more than two action verbs but omitted key information such as food and tool. We would like to use the relationship between text and images to enrich the information content and structure of recipes so as to be more conducive to the application of recipe retrieval or automatic translation.

ACKNOWLEDGMENT

This work was supported by JST CREST Grant Number JPMJCR16E3 and JSPS KAKENHI Grant Number 18K11425, Japan.

REFERENCES

[1] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," *ICLR*, 2017.

[2] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 32–41.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

[5] M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1059–1074, 2014.

[6] Y. Kawano and K. Yanai, "Foodcam: A real-time food recognition system on a smartphone," *Multimedia Tools and Applications*, vol. 74, no. 14, pp. 5263–5287, 2015.

[7] S. Li, S. Purushotham, C. Chen, Y. Ren, and C.-C. J. Kuo, "Measuring and predicting tag importance for image retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2423–2436, 2017.

[8] S. Mori, H. Maeta, Y. Yamakata, and T. Sasada, "Flow graph corpus from recipe texts." in *LREC*, 2014, pp. 2370–2377.

[9] T. Sasada, S. Mori, T. Kawahara, and Y. Yamakata, "Named entity recognizer trainable from partially annotated data," in *Conference of the Pacific Association for Computational Linguistics*. Springer, 2015, pp. 148–160.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[12] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[13] X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso, "Recipe recognition with large multimodal food dataset," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6.

[14] Y. Yamakata, J. Carroll, and S. Mori, "A comparison of cooking recipe named entities between japanese and english," in *Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in conjunction with The 2017 International Joint Conference on Artificial Intelligence*, 2017, pp. 7–12.