A DEEP MULTIMODAL APPROACH FOR MAP IMAGE CLASSIFICATION

Tomoya Sawada and Marie Katsurai

A DEEP MULTIMODAL APPROACH FOR MAP IMAGE CLASSIFICATION

Tomoya Sawada and Marie Katsurai

Department of Intelligent Information Engineering and Sciences, Doshisha University
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto, 610-0394 Japan
E-mail: {sawada, katsurai}@mm.doshisha.ac.jp

ABSTRACT

Map images (e.g., illustrated maps, historical maps, and geographic maps) have been published around the world, not only for giving location but also to attract tourists or hand down the histories of locations. The management of map data, however, has been an open issue for several research fields, including digital library, humanities, and tourism studies. This paper explores an approach for classifying diverse map images by their themes using map content features. Specifically, we present a novel strategy for preprocessing text data that are positioned inside the map images, which are extracted using OCR. The activation of the textual feature-based model is joint with the visual features in an early fusion manner. Finally, we train a classifier model comprising a convolutional layer and a fully connected layer, which predicts the belonging class of the input map. In experiments conducted on a new labeled dataset of map images, we demonstrate that our approach that uses the fused features achieved the best classification performance over single modality. We have made our dataset available on the Internet to facilitate this new task.

***Index Terms*—** map images, multimodal classification, OCR features

1. INTRODUCTION

Map images are published around the world, and huge number of maps have been digitized and archived as open-source data. For example, the United States (U.S.) Geological Survey scanned and archived over 178,000 topographical images of the U.S.,[1] and they are still increasing its number. The contents of maps varies in their theme: maps for strolling can recommend people restaurants and tourist spots; hazard maps, obtained from municipal websites, can provide immediate reference of safe places and hazardous areas when a natural disaster occurs. In addition, because attractive maps of regions can call in tourists, leading to regional development, local governments and companies often attempt to publish illustrated maps on the Internet. These diverse maps have attracted much attention in several research fields such as digital library, humanities, and tourism studies. However, these maps do not usually contain metadata to describe their themes, and the means to make them machine-readable have not been studied yet. This technological gap prevents the maps from being indexed and searched and limits the opportunity for both researchers and users to access valuable location information. By automatically recognizing the theme of map images, they will become easy to be retrieved, enabling several applications [1, 2]. Therefore, as the first step of computational map database management, we tackle a novel task: *automatic map image classification*.

The appearance of the image differs by theme. Intuitively, maps for tourism use bright colors to attract attention or historical maps will have dark colors because of aging degradation. Still, some maps share similarities: most tourism maps, for example, have in common that they visualize the illustrates or photos of tourist spots, such as temples in maps published to present historical buildings. On the other hand, the texts positioned inside the image can be valid suggestions for detecting the difference between visually similar images. Maps with different themes will use words characteristic to their content. Words like "walking course" occur more often inside walking maps than guide maps, for instance. Hazard maps will have words to call attention to dangerous places, and maps presenting restaurants frequently use the names of food. Compared to using only a single feature, effectively combining the two modalities (i.e., visual and textual features) have high expectation on improving the accuracy of classification.

This paper proposes a deep multimodal approach to map image classification using the features extracted from within the image. The text data are first extracted from the image using an optical character recognition (OCR) system, and preprocessed to reduce misrecognized letters by the system. Then they are embedded to feature vectors, using a text encoder that is capable of handling multilingual input. The features of two modalities in the training data were then joined by a multimodal fusion method and used to train a classification model. To validate the effectiveness of the proposed approach, we constructed a labeled dataset of map images. Our experiments demonstrated that fusing visual and textual features achieved the best classification accuracy over single modality-based methods.

The main contributions of this paper are three-fold:

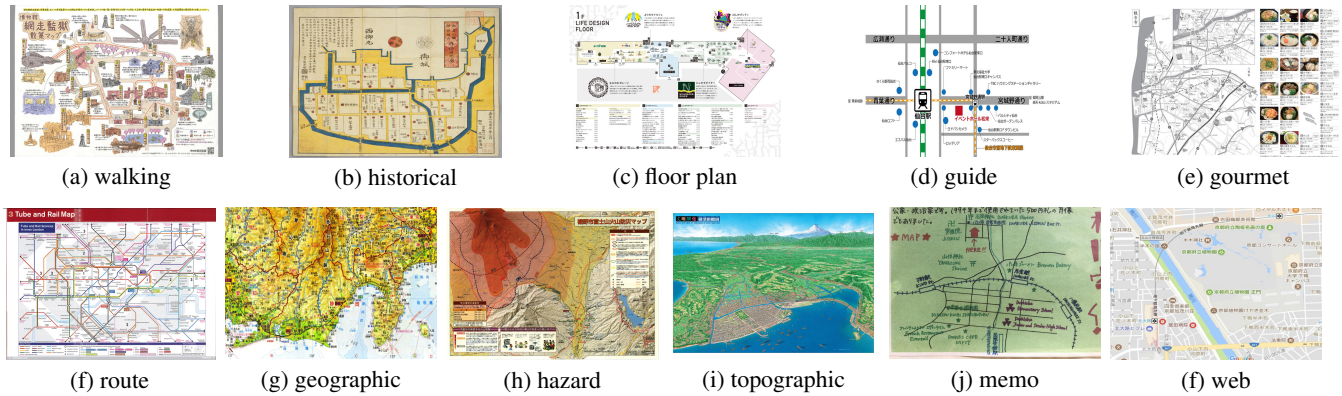
1. To the best of our knowledge, our work is the first to tackle the classification of diverse map images, which can be a valuable tool for digital library management.
2. We present useful tips to complementarily use text data produced by an OCR system with visual features as a multimodal input to a classification model. This is applicable to other domains with similar characteristics, such as advertisement images.
3. We have made our dataset available to encourage the study of map image classification.[2]

2. RELATED WORK

Although the importance of map database management and its potential for promoting interdisciplinary research has been acknowl-

[1] https://ngmdb.usgs.gov/topoview/

[2] https://mm.doshisha.ac.jp/map/MultimodalClassification.html



(a) walking (b) historical (c) floor plan (d) guide (e) gourmet

(f) route (g) geographic (h) hazard (i) topographic (j) memo (f) web

Fig. 1. Examples of images belonging to each class. Note that the images are reshaped to 700×500 pixels for construction.

Table 1. Map image categories and corresponding map counts.

Category	#images	Category	#images
walking	397	historical	405
floor-plan	133	guide	538
gourmet	149	route	67
geographic	149	hazard	55
topographic	26	memo	83
web	35		

edged recently [1, 3], few works had studied automatic map image classification. The work most related to ours was presented by Mandal et al. [3], which classified 446 land map images into the following four specific subcategories: political, physical, resource, and topographic maps. It used contour lines, identical icons, hue, and texture features to train a support vector machine (SVM). Our focus differs from [3] in the following two viewpoints: (i) the classification in [3] is focused on geological significance, while we challenged to classify map images into more wider scoped categories; and (ii) the features used in [3] are low-level features, whereas we learned deep features, which are known to exhibit stronger representation power [4].

When multiple modalities are available for given input images (e.g., textual comments or audio data), the method to fuse different types of features is a common argument roughly categorized into two approaches: late fusion and early fusion. Late fusion performs the actual integration in the decision-level using separate classification results for each modality, while early fusion focuses on creating a single representation for multiple features. Recent works have shown the superiority of the early fusion in deep learning frameworks. For example, Hu et al. [5] trained a fully connected layer after concatenating the text and visual features and achieved the state-of-the art performance in emotion analysis. Following to these works, we also exploit an early fusion strategy to complementarily use visual and textual features, which are extracted from map images.

If the input image is not surrounded by text, external text generators can be used to produce auxiliary semantic information [6, 7]. Ye et al. [7] used pseudo captions of the advertisement images as the textual data, which were automatically generated by object detector named DenseCap [8], to understand the content of advertisements. This paper presents tips for preprocessing the OCR results and embedding those to a deep learning framework.

3. DATASET

Because no public dataset exists for map image classification, we create our own dataset as follows: we first collected 1,977 map images from Stroly.com,[3] which is an Internet service targeted sightseeing using maps uploaded by users. Each image was examined by a human expert and was manually classified into 11 categories we defined. In our work we used manual procedure to label the map images, due to the novelty of our task. Automating this procedure will be our future work. Figure 1 presents examples of images in each category. As shown, the purposes of maps cover a wide range of topics, not limited to land maps studied in [3]. Table 1 shows the 11 categories and their corresponding numbers of images.

As maps of different countries and maps for tourists from foreign countries are included in our dataset, the texts positioned inside the map images contain several languages such as Japanese, Chinese, Korean, and English. In this study, the text corresponding to the map images were extracted using Google Vision API,[4] which is one of the most widely-used OCR systems, capable of extracting text regardless of the language. We obtained a total of 438,037 words from 1,977 images, and the total number of unique words was 67,397.

The language of each text data were further identified by a language identification tool, introduced by Lui et al. [9, 10]. This revealed that there are 16 languages in the dataset. In our preliminary analysis, Japanese, Korean, and English text were labeled mostly correct, but difficulties remained on labeling Chinese as many Chinese characters were confused with similar shaped Japanese character. Accordingly, Chinese texts were provisionally labeled as Japanese.

4. PROPOSED APPROACH

4.1. Textual Feature Extraction

Because the extracted text data might contain misrecognitions, including unnecessary or omitted letters, cleaning them tends to improve the accuracy of the classification. Figure 2 presents the overall process of the preprocessing and feature extraction of the text data. This subsection describes the preprocessing and embedding sequence we used to vectorize the text data.

[3]https://stroly.com/ja/

[4]https://cloud.google.com/vision/

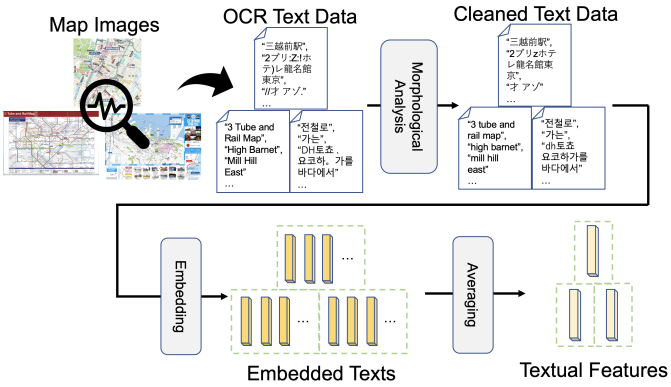


Fig. 2. Overall process of textual feature extraction.

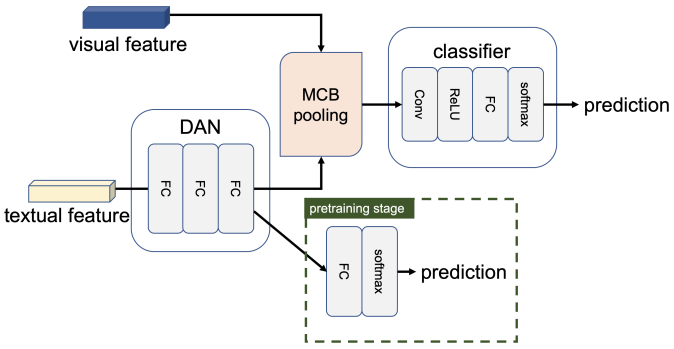


Fig. 3. Outline of the proposed architecture. Conv implies convolutional layer, and FC implies fully connected layer.

Text preprocessing: We perform a morphological analysis to the text data, each using suitable tools corresponding to the language. Specifically, we used MeCab[5] for Japanese and Chinese texts, MeCab-ko[6] for Korean texts, and NLTK[7] for the remaining languages. The part of speech of each word was detected and used to eliminate symbols from the data to reduce noise, which impede the evaluation of classification models. Unnecessary symbols are results of misrecognizing a section of the map contained in an image. For example, "+" were recognized from junctions in the map.

Text embedding: The next step is to aggregate the preprocessed text data of a map image to a single feature vector. Since the map image dataset contains multiple languages, we could not apply popular pretrained models such as word2vec [11] or GloVe [12]. Recently, Yang et al. [13] presented Multilingual Universal Sentence Encoder, which is an embedding model that can embed sentences of several languages into a common space. Using this encoder, we obtain a 512-dimensional vector for each word or sentence in the OCR results.

The positions of the text inside map images had no particular order. In using unordered text data, we took the average of the embedded texts of each image and the resulting vector is used as the textual feature.

4.2. Visual Feature Extraction

The visual features are extracted by a 18-layer Residual Network [14], pretrained on the ImageNet dataset [15]. The input images are resized to 448×448 pixels. We use the output of the layer, right before the 1000-way classification layer. We perform an L2 normalization to the 512-dimensional vector extracted by the model.

4.3. Fusion of Textual and Visual Features

To fuse the two types of features, we use Multimodal Compact Bilinear (MCB) pooling [16]. MCB pooling is an extension of Compact Bilinear Pooling [17] from single modality to concatenation of multiple modalities, and it has shown strong performance on tasks using multimodal features. Bilinear models [18], the origin of MCB pooling, calculates the outer product of the two feature vectors, enabling all the elements in the two vectors to participate in the training. However, high dimensionality of the created representation is making the method unable to be widely used. The key point of MCB pooling is that it reduces enormous numbers of dimensions, compared to the Bilinear models without losing its expressive power. In experiments, we investigate the effectiveness of MCB pooling compared to simple concatenation of the textual and visual features.

4.4. Architecture for Classification

Figure 3 shows an outline of the proposed architecture for map image classification. Below, we present the details of each of tree networks enclosed by rectangles. To improve the averaged text feature representation, we composed a variant model of Deep Averaging Network (DAN) [19]. In this model, the averaged word embeddings were fed through three fully connected (FC) layers of 512 units each. The activations of each layers of DAN are ReLU. As in our previous work [20], we first pretrain the DAN with the textual features as a single task to optimize the map image classification. An FC layer followed by a softmax layer was used to calculate the accuracy and loss for the pretraining, as shown in the dotted box in Fig. 3.

The input to the whole model are the textual and visual features. The textual feature is first fed through the pretrained DAN. After fusing the visual and textual features via the MCB pooling layer, we use a simple two-layer model with one convolutional (Conv) layer and one FC layer, which predicts the belonging class of the input map image. We apply ReLU and softmax to these layers, respectively. The cross-entropy loss is used to learn the classifier.

5. EXPERIMENTS

This section presents experimental results that verify the effectiveness of the proposed approach. Section 5.1 describes the experiment setup in detail. Section 5.2 then presents the results of the classification.

5.1. Experiment Settings and Baselines

Our dataset was split to 1,581 images for training and 396 images for testing. Then the training images were split again to 1,264 images for training and 317 images for validation. For the proposed deep architecture, we first pretrained the DAN composed in the proposed model, using map images and labels in the training data. After the pretrained weights of the DAN were loaded, then the whole model

[5]https://taku910.github.io/mecab/
[6]https://bitbucket.org/eunjeon/mecab-ko/src/master/
[7]https://www.nltk.org/

Table 2. The average accuracies of classification results on test data for five runs. FC implies fully connected layer, DAN implies the DAN section of the proposed method. Note that for singleton features, the vectors are fed directly to the classifier section, and for concatenated feature, the MCB pooling layer is switched to concatenation.

feature	SVM	NN Classifier
Textual	0.568	0.618
Textual + FC	-	0.627
Textual + DAN	-	0.694
Visual	0.719	0.655
Visual + FC	-	0.659
Concat	**0.732**	0.759
DAN + Concat	-	0.766
Concat + FC	-	0.723
DAN + Concat + FC	-	0.755
MCB pooling	0.727	0.762
DAN + MCB pooling	-	**0.774** (proposed)

was trained. We used Adam [21] as an optimizer for the loss function and also for pretraining the DAN.

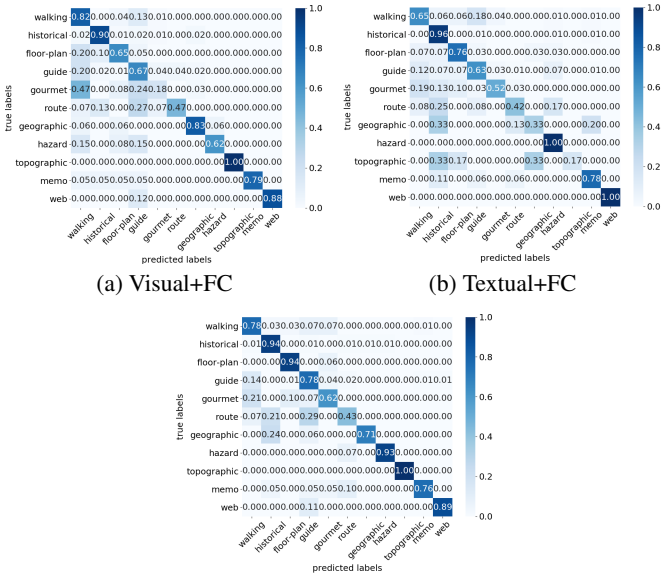
We evaluated the performance of our deep multimodal approach with several methods: first, we tested models that were trained on single modality, that is, visual or textual features. When using a single modality feature, the inputs were fed directly to the classifier section of the model. This approach is denoted as **Text** or **Visual** depending on the modality input. When textual features were fed through DAN section, they will be denoted as **DAN**. Using just a concatenation of the two features as an early fusion method, instead of the MCB pooling layer, is denoted as **Concat**. We also trained the model with the resulting vector of a FC layer, composed right after the input, which is denoted as **Text+FC**, **Visual+FC** or **Concat+FC**.

All features were also used to train linear SVMs, which were used in the related work [3]. Such SVMs are usually adopted when the dimensionality of feature vectors is high, like our textual feature vectors. In the tested model, the SVMs classified the map images instead of our layered classifier. The hyperparameters for the linear SVM were found by grid search based on cross-validation.

5.2. Results

For performance evaluation on testing data, we ran each classification model five times and calculated the averaged accuracy. Table 2 presents the average accuracy of each classification model. As shown, the proposed architecture, the layered classifier of which was trained using the feature joint by MCB pooling, exhibited the most highest accuracy in all of the conation. Simple concatenation of features exhibited the highest accuracy for SVMs. With each classification model, the use of multimodal representation over-performed those with singleton features. Figure 4 shows confusion matrices for Visual+FC, Textual+FC, and the proposed method. We can see from Fig. 4(c) that our method effectively labeled most of the images with their true categories. Visual+FC and Textual+FC each showed high accuracy in classificating topographic maps and hazard maps, and our method was possible to efficiently combined these beneficial inclination. The MCB pooling emphasized the commonalities of the two input vectors, the consideration of which can improve the accuracy of classification.

We also found from Fig. 4(c) that images belonging to "walking," "guide," and "gourmet" tend to be confused. Models trained by other features also showed the same inclination with the testing data. These results can be used to indicate how to design class labels in the map classification task.



(a) Visual+FC (b) Textual+FC

(c) Proposed method

Fig. 4. Confusion matrices of models trained by single modality features and our proposed method.

6. CONCLUSIONS AND FUTURE WORK

This paper presented a deep multimodal approach for map image classification, which is required for digital library management. By using MCB pooling as the fusion method, the joint features over-performed the singleton modality feature. Especially, the textual features by themselves had low expressiveness to show accurate performance, but they were able to draw out their strength by collaborating with the visual features.

Our work in this paper is the first step toward automatic recognition of map images. The next issue will be to refine category divisions by using the results we presented. It would be valuable to develop a method for automatically finding topics from a map collection. We have presented that text data positioned inside the images improve the accuracy of classification tasks, and we believe so with our next challenge. Still, the preprocessing of the OCR text data require improvement, as we can see from the accuracy of classification using the text only feature. In our work, Chinese were analysed as Japanese, and the elimination of symbols in Chinese texts was incomplete. The use of more-sophisticated language-specific preprocessing should be discussed in future works. We will also develop several applications featuring content-based map image retrieval based on the method proposed in this paper.

7. ACKNOWLEDGEMENTS

This work was partly supported by a Grant-in-Aid for Young Scientists (B) 17K12794. We also thank Stroly Inc. for providing the map images that were used in this study.

8. REFERENCES

[1] Y.-Y. Chiang, "Querying historical maps as a unified, structured, and linked spatiotemporal source: Vision paper," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, 2015, pp. 16:1–16:4.

[2] H. Vermeulen, T. Takahashi, M. Takahashi, K. Ohtsuka, T. Nakagawa, and H. Ueda, "Stroly: A historic and illustrated maps platform," in *Proceedings of the Second International Conference on Culture and Computing,*, 2012, pp. 195–196.

[3] S. Mandal, S. Biswas, A. Kumar Das, and B. Chanda, "Land map image dataset: Ground-truth and classification using visual and textual features," *Image Processing & Communications*, vol. 19, 2014.

[4] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014, pp. 806–813.

[5] A. Hu and S. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2018, pp. 350–358.

[6] M. Elhoseiny, J. Liu, H. Cheng, H. Sawhney, and A. Elgammal, "Zero-shot event detection by multimodal distributional semantic embedding of videos," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2016, pp. 3478–3486.

[7] K. Ye and A. Kovashka, "ADVISE: Symbolism and external knowledge for decoding advertisements," in *Computer Vision – ECCV 2018*. 2018, pp. 868–886, Springer International Publishing.

[8] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4565–4574.

[9] M. Lui and T. Baldman, "Cross-domain feature selection for language identification," in *Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP)*, 2011, pp. 553–561.

[10] M. Lui and T. Baldman, "langid.py: An off-the-shelf language identification tool," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012, pp. 25–30.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013s.

[12] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[13] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernández Ábrego, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil, "Multilingual universal sentence encoder for semantic retrieval," *arXiv preprint arXiv:1907.04307*, 2019.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual neural network for image recognition," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 770–778.

[15] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet : A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[16] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[17] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 317–326.

[18] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, pp. 1247–1283, 2000.

[19] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015, pp. 1681–1691.

[20] S. Sanjo and M. Katsurai, "Recipe popularity prediction with deep visual-semantic fusion," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*, 2017, pp. 2279–2282.

[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1424.6980*, 2015.