A Study on Intra-modal Constraint Loss toward Cross-modal Recipe Retrieval

Jiahang LU[†], Haruya KYUTOKU[††,†], Keisuke DOMAN[†††,†], Takahiro KOMAMIZU[†], Yasutomo KAWANISHI[††††,†], Takatsugu HIRAYAMA[†††††,†], and Ichiro IDE[†]

y Nagoya University Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601 Japan
yy Aichi University of Technology 50-2 Manori, Nishihasama-cho, Gamagori, Aichi, 443-0047 Japan
yyy Chukyo University 101 Tokodachi, Kaizu-cho, Toyota, Aichi, 470-0393 Japan
yyyy RIKEN 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
yyyyy University of Human Environments 6-2 Kami Sanbonmatsu, Motojuku-cho, Okazaki, Aichi 444-3505 Japan

E-mail: *y* luj@cs.is.i.nagoya-u.ac.jp, taka-coma@acm.org, ide@i.nagoya-u.ac.jp, *yy* kyutoku-haruya@aut.ac.jp, *yyy* kdoman@sist.chukyo-u.ac.jp, *yyyy* yasutomo.kawanishi@riken.jp, *yyyyy* t-hirayama@uhe.ac.jp

Abstract Cross-modal recipe retrieval has become a popular task in multimedia research due to the importance of food in daily lives. With the development of neural networks, the limit for this task is no longer in the encoders of image or text but to learn a better embedding space across image-text modalities. In this presentation, we propose the usage of an Intra-Modal Constraint (IMC) loss function for learning the joint embedding of image and text. The IMC loss penalizes when negative pairs from the same homogeneous modality when they are close in the joint space. We investigated the e ectiveness of the proposed method through experiments on a cooking recipe dataset Recipe1M.

Key words Recipe retrieval, Multimedia, Intra-modal constraint loss function

1. Introduction

Meal is one of the most essential symbol for human civilization. It has become a crucial element for healthy lives, sharing cultural differences, and sense of community. Moreover, recipes has become a popular means of sharing meals, and many people follow such recipes to reproduce the meals.

We usually have experiences in daily life about seeing some delicious-looking meal images and attempting to try to cook them but fail to find detailed recipes explaining how to reproduce them. On the other hand, when we find an interesting recipe but hesitate if we should try cooking it, its image will help us make the decision. To deal with these problems, the recipe retrieval task has gathered much attentions [1]. The goal of recipe retrieval is finding a relevant recipe from an input meal image by retrieving the right image accompanying a recipe. Retrieving a relevant image from an input recipe is also available by the same system. Since some recipes have

[1, 2] have provided an opportunity to accelerate system developments of this task. These massively labeled datasets could help us achieve the goal of learning representations from various samples.

Recent efforts addressing the problem of extracting text features by Recurrent Neural Network (RNN) architectures, such as the Long Short-Term Memory (LSTM) [3] units and Gated Recurrent Units (GRU) [4] have emerged as effective models to capture long-term dependencies in sequential data. At the same time, with the development of computer vision, widely pre-trained models can be used for extracting image features effectively such as ResNet-50 [5]. In addition, some previous studies aim to learn joint representations for textual and visual modalities in the context of food images and cooking recipes [1, 6–10]. To learn an appropriate joint embedding space, loss function is a key.

In this work, for dealing with the problem of heterogeneous modality restriction, we use Intra-Modal Constraint (IMC) loss [11] for reducing the violation between negative pairs within the same modality. The violation means some representations from the same modality might be similar and mislead the model to retrieve the right one from numerous alternatives. To be specific, the IMC loss levers the Intra-Modal Constraint item which can measure the distances between homogeneous modalities. In the process of calculation, the loss will increase with another penalty term in the case of negative pairs in the same modality. In other words, this process could increase the intra-modal non-pair distance, the distance corresponding to the similarity of two representations and the higher distance means the lower similarity.

We also performed experiments using different similarity functions, such as the L1 normalization and L2 normalization which can effectively obtain the distance between two features.

In the following, we introduce related work in Section 2, proposed method in Section 3, experiment in Section 4, and conclusion in Section 5.

2. Related Work

2.1 Cross-Modal Recipe Retrieval

This research explores cross-modal recipe retrieval, which has attracted attentions from many researchers. They try to construct a system to retrieve relevant information from a query in different modalities as shown in Fig. 1. These models are usually trained via direct correspondence between pairs of instances in different modalities. To be specific, the recipe retrieval task uses food images and corresponding text descriptions. The text description is composed of a title, ingredients, and instructions. One of the challenges is the media gap [12], which indicates the features from different modalities are inconsistent, so it can be difficult to calculate their similarity accurately.

A system for this task is usually composed of an image encoder and a text encoder which uses pre-trained image recognition models and natural language understanding models. They are followed by projection onto a joint embedding space calculated by a loss function. Most studies aim to train a better space via diverse strategies.

Compared to some other tasks which involve the descriptions of an image such as the image captioning task, the text description part in cross-modal recipe retrieval is longer and more complex. This characteristics also leads to the difficulty of encoding. For dealing with the structured nature of the recipe text, some previous methods extract features from these components independently and concatenate them in a late-fusion layer. In this way, the problem of the structured nature of recipe text is solved and these components could be merged as a fixed-length recipe embedding. For learning appropriate embeddings in text, word2vec [13] and GloVe [14] are popular strategies and the representations from these strategies can be used as the input.

Some researches focus on training a better joint embedding space via different strategies. Chen et al. [7] proposed a deep hierarchical attention network of words and sentences. The method projects them onto a common embedding space. In this way, the model can understand the instructions better and predict the consequences from the visual representations precisely. Cao et al. [15] presented a retrieval framework based on co-attention network. This network can learn the representations of texts and images. The co-attention network computes the attention weights of cooking procedures and retrieve appropriate components from them.

2.2 Image-Recipe Representation Computation and Understanding

For this research, as the base of retrieving the corresponding representation from another modality, understanding the representation in both directions is important. Food understanding has become a popular topic, because it involves many tasks related to food. With the infiltration of social media into our daily lives, people tend to share food images online. It offers us a rich data source to analyze foods and related information.

Many studies pay attention on food image classification [16–18], some interesting ones aim to predict the calories [19] or estimate the ingredients from a dish [20]. These techniques offer assistance to researches such as that by Salvador et al. [1] which learn the representation from the information of food categories and propose the Recipe1M dataset. This dataset also provides the nutrition information for many foods.

2.3 Loss Function Learning in Cross-Modal Retrieval

An important part of this research is constructing a better joint-embedding space by loss function. For learning a better common space between two different modalities, a variety of loss functions are proposed to fill the gap. One of the popular methods is called Sum of Hinges (SH) loss [21], which can reduce the retrieval distance in both directions (between text and image). Faghri et al. [21] also proposed the Max of Hinges (MH) loss based on the SH loss. The MH loss pays more attention on hard negatives for training, and achieves better performances than the SH loss. The above two loss functions pay more attention on heterogeneous modality pair, but the effect of homogeneous modality pairs is neglected.

3. Proposed Method

In this section, we propose a cross-modal recipe retrieval model based on the Intra-Modal Constraint (IMC) loss. As illustrated in Fig. 2, this model is composed of an image encoder, a text encoder, and a joint embedding space. The features extracted from these two encoders are projected onto the joint embedding space. In this work, we take a food image or a cooking recipe as an input, and then give a retrieval result of the opposite modality as an output.

3.1 Food Image Encoder

The purpose of the image encoder is to learn an appropriate function to project the input image onto the joint image-text embedding space. In our work, the pre-trained ResNet-152 [5] was used as the image encoder. It is followed by a Fully Connected (FC) layer to extract the image feature vector e_i^n (dimension $D_{\text{img}} = 1,024$).

3.2 Recipe Text Encoder

The recipe encoder aims to learn an appropriate function to project the input text onto the joint image-text embedding space. The word representation is obtained from the pre-trained GloVe [14] and we employ the Bi-direction Long Short-Term Memory (Bi-LSTM) [3] which considers both forward and backward orderings to obtain the text feature vector. In addition, due to the structured nature of a recipe, we use three separate encoders to process sentences from the title, ingredients, and instructions. They are followed by a FC layer to extract the recipe feature vector e_t^n (dimension $D_{\text{txt}} = 1,024$).

3.3 Intra-Modal Constraint Loss

The image feature e_i^n and text feature e_t^n from the n-th image and its corresponding text are extracted from two separate networks. They are projected onto a common embedding space called Intra-Modal Constraint Loss Joint Embedding Space. The non-pair features are represented as e_i^n and e_i^m which indicates the negative pair from the n-th image and the m-th image, respectively.

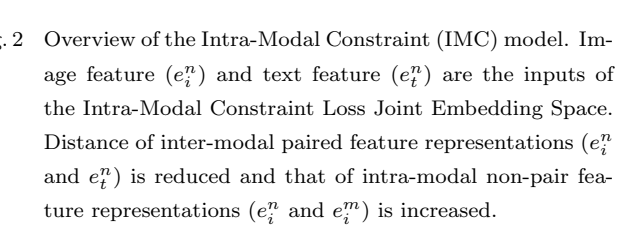

Fig. 2 Overview of the Intra-Modal Constraint (IMC) model. Image feature (e_i^n) and text feature (e_t^n) are the inputs of the Intra-Modal Constraint Loss Joint Embedding Space. Distance of inter-modal paired feature representations (e_i^n and e_t^n) is reduced and that of intra-modal non-pair feature representations (e_i^n and e_i^m) is increased.

In this work, we use the IMC loss [11] to train the common joint space and the encoders. The IMC loss is inspired by MH loss [21], unlike the MH loss, it focuses not only on heterogeneous modality pairs but also on the non-pair features from homogeneous modality. IMC loss is composed by MH loss and two Intra-Modal Constraint terms as:

$$\begin{aligned} \text{L}_{\text{IMC}}(m,n) = & \max_{m,n \in N}[\alpha + \theta(e_i^n + e_t^m) - \theta(e_i^n + e_t^n)] \\ & + \max_{m,n \in N}[\alpha + \theta(e_t^n + e_i^m) - \theta(e_i^n + e_t^n)] \\ & + \text{IMC}(e_i^n, e_i^m) + \text{IMC}(e_t^n, e_t^m) \end{aligned} \quad (1)$$

The first item is utilized for dealing with all negative texts and queried image and the second item is utilized for dealing with negative images and queried text. Each term is proportional to the expected loss over sets of negative samples. In this function, the winner takes all the gradients, so. In this way, the effectiveness of optimization can be enhanced and also benefit to constructing a better joint embedding space. The coefficient α serves as a margin parameter and the function $\theta(x, y)$ is a similarity function.

The IMC items can constrain the features from a homogeneous modality which is defined as follows:

$$\text{IMC}(e^n, e^m) = \tau \sum_{n,m \in N} \begin{cases} 0, & \beta(e^n, e^m) < \mu_{\text{down}} \\ \beta(e^n, e^m), & \mu_{\text{down}} \leqq \beta(e^n, e^m) \leqq \mu_{\text{up}} \\ 0, & \mu_{\text{up}} < \beta(e^n, e^m) \end{cases} \quad (2)$$

The coefficient τ is a weight parameter for balancing the loss and we defined it as 1.0 here. The function $\beta(x, y)$ is a similarity function. The constants μ_{up} and μ_{down} are boundary thresholds.

In this way, we can consider both paired/non-paired features in the heterogeneous/homogeneous modalities. There is a variety of normalization functions that is effective as a similarity function.

In contrast to some previous work that only considers the relation among the heterogeneous modalities, we also pay attention to the relation of a homogeneous modality. To be specific, the process of this part is to reduce the distance from the paired data and increase the distance from the non-paired data. In this way, the model can distinguish the corresponding result from an input. The first two items in the loss function also inherit this principle.

However, some representations from a homogeneous modality tend to be similar and mislead the model. The latter two items in the loss function will increase with another penalty term within the boundaries in the situation of negative pairs. In this way, the distance of similar representations from the same homogeneous modality will be increased and the model can distinguish them well.

4. Experiment

In this section, we report the results of an evaluation experiment for validating the effectiveness of the proposed approach.

4.1 Dataset

As same as some previous studies, for training the joint embedding model and capturing the joint cross-modal information, experiments were conducted using Recipe1M [1], which contains one million structured cooking recipes and corresponding images.

In the task of recipe retrieval, each image should be associated to corresponding texts composed of title, ingredients, and instructions. This dataset also consists of the above essential information which is extracted from cooking Web sites.

Due to the structure of the available recipes on the Web, Recipe1M largely consists of text-only samples and multiple recipes corresponding to one image. This actually influences the performance of this model. Additionally, 25 percents of images are associated with 1 percent of recipes while half of all images belong to 10 percents of recipes.

4.2 Experimental Conditions

Following some previous studies, the retrieval performance of the experimental results was evaluated with the R@K (refer to as R@1, R@5, R@10), medR, and meanR. R@K indicates Recall at K, the proportion of correct matches in the top $K = \{1, 5, 10\}$ retrieved results. These metrics are used to quantify the performance of the model. There are also some options for similarity functions applied in IMC items. We also performed an experiment based on these normalization functions, including the L1 and L2 distance:

Table 1 Comparison of the results of Image-to-Recipe retrieval sub-task. Similarity distances of the proposed method is measured in Manhattan distance (L1) and Euclidean distance (L2).

	R1↑	R5↑	R10↑	medR↓	meanR↓
Salvador et al. [1]	24.0	51.0	65.0	5.2	—
Chen et al. [7]	**25.6**	53.7	66.9	4.6	—
Proposed (β_{L_1})	17.2	**62.9**	88.4	**4.0**	5.6
Proposed (β_L)	18.1	62.8	**89.4**	**4.0**	**5.5**

Table 2 Comparison of the results of Recipe-to-Image retrieval sub-task. Similarity distances of the proposed method is measured in Manhattan distance (L1) and Euclidean distance (L2).

	R1↑	R5↑	R10↑	medR↓	meanR↓
Salvador et al. [1]	25.0	52.0	65.0	5.1	—
Chen et al. [7]	**25.7**	**53.9**	67.1	**4.6**	—
Proposed (β_{L1})	7.5	37.5	**74.8**	7.0	**7.0**
Proposed (β_{L2})	7.4	37.3	74.5	7.0	**7.0**

Manhattan distance (L1):

$$\mathrm{L1}(e^n, e^m) = \sum_{n,m \in N} |e^n - e^m| \tag{3}$$

Euclidean distance (L2):

$$\mathrm{L2}(e^n, e^m) = \sqrt{\sum_{n,m \in N} (e^n - e^m)^2} \tag{4}$$

where (e^n, e^m) are negative pairs in the same modality.

4.3 Training Details

The model is implemented in PyTorch running on an NVIDIA GeForce RTX 3090 GPU. We trained models with a batch size of 128 with a base learning rate of 2×10^{-4} and updated every 8 epochs. The Bi-LSTM is initialized with Xavier init [22] and applies dropout with a probability of 0.5 to avoid over-fitting. The model was trained for 20 epochs with the Adam [23] optimizer. The thresholds μ_{down} and μ_{up} were empirically set to 0.05 and 0.5, respectively. Different distance functions were used in the experiment, which will be explained in the next section.

4.4 Experimental Results

Tables 1 and 2 show the results of the proposed method on the Recipe1M [1] dataset.

The used similarity functions here were the L1 distance and L2 distance as defined in Eqs. (3) and (4). We found that the Euclidean distance (L2) had better accuracy compared with the Manhattan distance (L1) in general. Unlike some previous methods, such as those proposed by Chen et al. [7] and Salvador et al. [1], only project the heterogeneous modality representations onto the joint embedding

space, our method takes homogeneous modality representations into account.The result also shows the effectiveness of the intra-modal constraint loss, notably in the metric of R@10. The precision of Image-to-Recipe sub-task was also higher than the Recipe-to-Image sub-task. This might be because of the complexity of the text. Especially, concatenating three components by a linear layer might have influenced the performance of the model.

5. Conclusion

In this presentation, we studied the cross-modal retrieval task in the food domain by the IMC loss. We tried to address the problem of traditional methods that only focus on the heterogeneous modality but ignore the features in the homogeneous modality. This loss allows us to train using both paired and unpaired recipe data. The result shows this method could achieve good results in R@10, especially in the sub-task of Image-to-Recipe. In the future, we will try to enhance the performance of this model, especially for R@1, medR, and meanR.

Acknowledgement

Parts of this work were supported by JSPS Grants-in-Aid for Scientific Research (20K12038, 22H00548).

References

[1] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp.3020–3028, 2017.

[2] F. Ofli, Y. Aytar, I. Weber, R. Al Hammouri, and A. Torralba, "Is saki# delicious? The food perception gap on instagram and its relation to health," Proceedings of the 26th International Conference on World Wide Web, pp.509–518, 2017.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol.9, no.8, pp.1735–1780, 1997.

[4] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," Computing Research Repository arXiv preprint, arXiv:1406.1078, pp.1–15, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp.770–778, 2016.

[6] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord, "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings," Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp.35–44, 2018.

[7] J.-J. Chen, C.-W. Ngo, F.-L. Feng, and T.-S. Chua, "Deep understanding of cooking procedure for cross-modal recipe retrieval," Proceedings of the 26th ACM International Conference on Multimedia, pp.1020–1028, 2018.

[8] M. Fain, N. Twomey, A. Ponikar, R. Fox, and D. Bollegala, "Dividing and conquering cross-modal recipe retrieval: From nearest neighbours baselines to SOTA," Computing Research Repository arXiv preprint, arXiv:1911.12763, pp.1–11, 2019.

[9] A. Salvador, E. Gundogdu, L. Bazzani, and M. Donoser, "Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning," Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.15475–15484, 2021.

[10] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, and S.C. Hoi, "Learning cross-modal embeddings with adversarial networks for cooking recipes and food images," Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.11572–11581, 2019.

[11] J. Chen, L. Zhang, Q. Wang, C. Bai, and K. Kpalma, "Intra-modal constraint loss for image-text retrieval," Proceedings of the 29th IEEE International Conference on Image Processing, pp.4023–4027, 2022.

[12] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," IEEE Transactions on Circuits and Systems for Video Technology, vol.28, no.9, pp.2372–2385, 2017.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Computing Research Repository arXiv preprint, arXiv:1301.3781, pp.1–12, 2013.

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

[15] D. Cao, Z. Yu, H. Zhang, J. Fang, L. Nie, and Q. Tian, "Video-based cross-modal recipe retrieval," Proceedings of the 27th ACM International Conference on Multimedia, pp.1685–1693, 2019.

[16] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, "Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment," Proceedings of the 14th International Conference on Smart Homes and Health Telematics, pp.37–48, 2016.

[17] S. Mezgec and B. Koroušić Seljak, "Nutrinet: A deep learning food and drink image recognition system for dietary assessment," Nutrients, vol.9, no.7, p.657, 2017.

[18] K.-H. Lee, X. He, L. Zhang, and L. Yang, "Cleannet: Transfer learning for scalable image classifier training with label noise," Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, pp.5447–5456, 2018.

[19] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K.P. Murphy, "Im2calories: Towards an automated mobile vision food diary," Proceedings of the 16th IEEE International Conference on Computer Vision, pp.1233–1241, 2015.

[20] J. Li, R. Guerrero, and V. Pavlovic, "Deep cooking: Predicting relative food ingredient amounts from images," Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, pp.2–6, 2019.

[21] F. Faghri, D.J. Fleet, J.R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," Computing Research Repository arXiv preprint, arXiv:1707.05612, pp.1–14, 2017.

[22] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp.249–256, 2010.

[23] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Computing Research Repository arXiv preprint, arXiv:1412.6980, pp.1–15, 2014.