A Knowledge-based Deep Heterogeneous Graph Matching Model for Multimodal RecipeQA

Yunjie Wu[1], Sai Zhang[1], Xiaowang Zhang[1], Zhiyong Feng[1,2]⋆, and Liang Wan[1]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] College of Intelligence and Computing, Shenzhen Research Institute of Tianjin University, Tianjin University, Tianjin, China
{yunjie_wu, zhang_sai, xiaowangzhang, zyfeng, lwan}@tju.edu.cn
⋆ Corresponding Author

Abstract. RecipeQA is a multimodal task that requires understanding the multimodal context. Since the recipe instructions are procedural, temporal relations are essential to support procedural understanding. Due to the high divergence of representation, it is challenging to model the temporal relations of multimodal and dynamic recipes. In this paper, we propose a Knowledge-based Deep Heterogeneous Graph Matching Model (DHGM) to model temporal structures of recipes. Firstly, we present a knowledge-based recipe encoder to reduce the divergence between recipe entities. Secondly, we design a two-stage heterogeneous graph matching method to guarantee neighborhoods consensus. Experimental results show that our proposed approach for RecipeQA obtains the best performance on the RecipeQA dataset.

Keywords: Multimodal Reading Comprehension · Heterogeneous Graph Matching · Knowledge Graph

1 Introduction

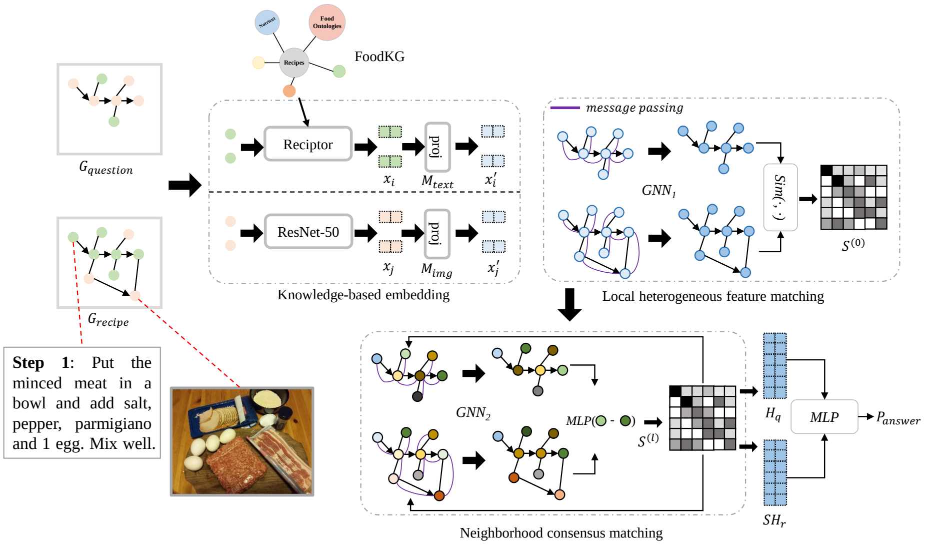
The RecipeQA [1] is a newly proposed Multimodal Machine Comprehension (M^{3}C) task, which comprises instructional recipes with multimodal context. The RecipeQA provides different multi-choice tasks that require a joint understanding of both visual and textual procedural knowledge.

Since cooking instructions are procedural, it is essential to understand temporal relations for RecipeQA. However, due to the high divergence of representations between different instructions, it is challenging to model temporal relations. In addition to the heterogeneity of multimodal data, even the same textual entity could be different. For example, the *"ground beef"* in Step 1 would change to *"patty"* in Step 3 after cooking, but they still correspond to the same entity. The existing works either learn the dynamical states of entities for procedural reasoning [2] or attempt to answer questions with attention-based alignment [3, 4]. It is not easy for them to capture the semantics of temporal structure since the structural semantics is lost during encoding the recipes. Inspired from DGMC [6], we model the temporal structure of recipes in deep graph matching.

In this paper, we propose DHGM for RecipeQA to explicitly model temporal structures in the graph matching processing. Firstly, we present a knowledge-based recipe embedding to reduce the divergence between recipe entities. Secondly, we design a two-stage heterogeneous graph matching method to guarantee neighborhoods consensus for injecting temporal structure. Finally, we conduct experiments on the RecipeQA dataset and achieve the best performance.

2 Deep Heterogeneous Graph Matching Model

The framework of DHGM is shown in Figuer 1. We build heterogeneous graphs with temporal relations for recipe and question, which are embedded based on the knowledge graph. The correct answer is chosen with a two-stage method: local heterogeneous feature matching and neighborhood consensus matching.



Step 1: Put the minced meat in a bowl and add salt, pepper, parmigiano and 1 egg. Mix well.

Fig. 1. The overall framework of DHGM for RecipeQA

2.1 Knowledge-based recipe embedding

To reduce the divergence between entity representations in recipes, we employed a pretrained model *Reciptor* [5] for recipe embedding with an external knowledge graph *FoodKG*. With an anchor recipe r_a, we extract the related triples from

FoodKG through a similarity model. Given an entity e_a, a positive partner e_p connected to e_a and a negative partner e_n not connected to e_a, the triplet loss is used to optimize the learned embeddings in semantic space:

$$\mathcal{L}_{emb}\left(e_a, e_p, e_n\right) = \max\left(0, d\left(e_a, e_p\right) + \alpha - d\left(e_a, e_n\right)\right) \tag{1}$$

where $d\left(e_i, e_j\right)$ is used to measure the distance, α is a margin parameter. Furthermore, we employ a pretrained ResNet-50 model to represent each image.

2.2 Local heterogeneous feature matching

We construct graphs for each recipe and question with the temporal relations. A graph can be defined as $G = (\mathcal{V}, A)$, where $\mathcal{V}$ is a vertex set, and A is the adjacency matrix of edges. For a heterogeneous node with type t, we employ a type-specific transformation matrix M_t to project the node into same semantic space:

$$\boldsymbol{x}_i' = \boldsymbol{M}_t \boldsymbol{x}_i \tag{2}$$

where $\boldsymbol{x}_i$ is the original feature, and $\boldsymbol{x}_i'$ is the projected feature. The node feature $\boldsymbol{h}_j^{(t-1)}$ in layer t can be updated with localized information as follows:

$$\boldsymbol{h}_i^{(0)} = \boldsymbol{x}_i' \tag{3}$$

$$\boldsymbol{a}_{ij}^{(t)} = softmax(att(\boldsymbol{h}_i^{(t-1)}, \boldsymbol{h}_j^{(t-1)})) \tag{4}$$

$$\boldsymbol{h}_i^{(t)} = \sigma(\sum_{j \in \mathcal{N}_i} a_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}) \tag{5}$$

where $\mathcal{N}_i$ is the neighbors set of node i, and $att(\cdot)$ is the self-attention model for learning the weight of different neighbors. Given the node embeddings of recipe $\boldsymbol{H_r}$ and question $\boldsymbol{H_q}$, we obtain the initial correspondences matrix as:

$$S^{(0)} = mask\_softmax(\boldsymbol{H_q} \boldsymbol{H_r}^{\top}) \tag{6}$$

2.3 Neighborhood consensus matching

We iteratively refine the correspondences matrix $S^{(l)}$ to guarantee the neighborhood consensus. The node features $\boldsymbol{H_q}$, $\boldsymbol{H_r}$ can be passed along the soft correspondence S to obtain node features $\boldsymbol{H_r}'$, $\boldsymbol{H_q}'$ in other domain:

$$\boldsymbol{H_r}' = S^{\top} \boldsymbol{H_q} \quad \text{and} \quad \boldsymbol{H_q}' = S \boldsymbol{H_r} \tag{7}$$

And then, we employ node indicator function $\boldsymbol{I}$ to map corresponding neighborhoods into sub-domain and propagate message by *GNN* model:

$$O_q = GNN(\boldsymbol{I}, X_q, A_q) \quad \text{and} \quad O_r = GNN(S_{(l)}^{\top} I, X_r, A_r) \tag{8}$$

where X is the feature matrix of nodes. We measure the consensus of nodes pair (v_i, v_j) by $\boldsymbol{d}_{ij} = \boldsymbol{o}^q_i - \boldsymbol{o}^r_j$. And the correspondence matrix can be updated as follow:

$$S^{(l+1)}_{i,j} = softmax(S^{(l)}_{i,j} + MLP(\boldsymbol{d}_{j,i}))_{i,j} \tag{9}$$

we utilized a similarity model followed by an MLP layer to obtain the final answer:

$$P(a_k) = softmax(similarity(\boldsymbol{H_q}, S\boldsymbol{H_r})) \tag{10}$$

The final objective function has been defined as follows:

$$\mathcal{L} = -\sum_{i \in \mathcal{V}_\sigma} \log(S^{(L)}_{i,\pi_{gt}(i)}) + \sum_{(a^+,a^-) \in A_i} [-\log(p(a^+)) - \log(1 - p(a^-))] \tag{11}$$

where $\pi_{gt}(i)$ is the ground truth correspondences.

3 Experiments and Evaluation

We conduct the experiments on three tasks of the RecipeQA dataset (i.e., visual cloze, visual coherence, visual ordering). The models are trained on each task separately. We employ accuracy as the metric.

Table 1 presents the quantitative results on the test set. The proposed DHGM outperforms other benchmark models on all three tasks, demonstrating that temporal structural information plays an essential role in procedural understanding. We also employ a DHGM model without knowledge-based recipe embedding to verify the impact of *FoodKG*, which still outperforms the benchmark models. The DHGM explicitly models the temporal structure in graph matching, which is beneficial to understanding the multimodal temporal context. We observe that the performance drops without the knowledge-based embedding, which shows that knowledge is helpful to enhance the recipe representations.

Table 1. Accuracy on the test sets of RecipeQA.

	Visual Cloze	Visual Coherence	Visual Ordering
HUMAN	77.60	81.60	64.00
Hasty Student	27.35	65.80	40.88
Impatient Reader	27.36	28.08	26.74
PRN (Single Task)	56.31	53.64	62.77
PRN (Multi Task)	46.45	40.58	62.67
DHGM -w/o KG	48.16	45.37	61.54
DHGM	**51.57**	**50.28**	**63.11**

4 Conclusion

In this paper, we propose DHGM for RecipeQA to explicitly model the temporal structure in graph matching. We believe that the graph-matching-based idea for modeling the temporal structure is meaningful for understanding the multimodal procedural data similar to RecipeQA. Experimental results on the RecipeQA dataset demonstrate the excellent performance of the proposed DHGM. In future work, we would extend our approach to other tasks with heterogeneous procedural data to verify the generalization.

5 Acknowledgments

The work described in this paper is partially supported by Shenzhen Science and Technology Foundation (JCYJ20170816093943197) and Key Research and Development Program of Hubei Province (No. 2020BAB026).

References

1. Yagcioglu, S., Erdem, A., Erdem, E., Ikizler-Cinbis, N.: RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In: *Proc. of EMNLP 2018*, pp.1358-1368.
2. Amac, M. S., Yagcioglu, S., Erdem, A., Erdem, E.: Procedural Reasoning Networks for Understanding Multimodal Procedures. In: *Proc. of CoNLL 2019*, pp.441-451.
3. Faghihi, H. R., Mirzaee, R., Paliwal, S., Kordjamshidi, P.: Latent Alignment of Procedural Concepts in Multimodal Recipes. In: *Proc. of ALVR 2020*, pp.26-31.
4. Liu, A., Yuan, S., Zhang, C., Luo, C., Liao, Y., Bai, K., Xu, Z.: Multi-Level Multimodal Transformer Network for Multimodal Recipe Comprehension. In: *Proc. of SIGIR 2020*, pp.1781-1784.
5. Li, D., Zaki, M. J.: Reciptor: An effective pretrained model for recipe representation learning. In: *Proc. of KDD 2020*, pp.1719-1727.
6. Fey, M., Lenssen, J. E., Morris, C., Masci, J., Kriege, N. M.: Deep graph matching consensus. In: *Proc. of ICLR 2020*, poster.