Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text – Supplementary Material

Difei Gao[1,2*], Ke Li[1,2*], Ruiping Wang[1,2], Shiguang Shan[1,2], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
{difei.gao, ke.li}@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

Overview

In the supplementary material, we display more implementation details of number encoder in graph construction (Sec.3.1 of the main body) and more experimental results about our method Multi-Modal Graph Neural Network (MM-GNN), including quantitative (on Sec. 2) and qualitative analysis (on Sec. 3).

1. Number encoder

In MM-GNN, digital numbers and days of the week which are similar to time-related strings containing periodic information, e.g. "Sunday", are considered as numeric type strings. Besides, for periodic numbers, e.g. 10:00, we first normalize it into $\frac{10}{24}$, then we apply the cosine embedding function $x \rightarrow cos\, 2\pi x$. Note that, how to encode the numeric strings is still an open problem, different encoders can capture different relations between numbers, e.g. an alternative representation that uses the polar coordinate system which uses two functions cosine and sine to encode a number could be better in representing the periodic numbers.

2. Quantitative Analysis

The impact of the order of aggregators. In Multi-Modal Graph Neural Network (MM-GNN), the three aggregators which update the representations of nodes in different sub-graphs are performed in a particular order, that is, first perform Visual-Semantic aggregator (VS), then perform Semantic-Semantic aggregator (SS), finally Semantic-Numeric aggregator (SN). In this part, we evaluate the influences of all different orders of aggregators. The results are shown in Table 1.

From the results, we can see that the performances of different variants are similar to each other, which indicates that our proposed MM-GNN is robust to changes in the order. This probably thanks to the functions of three aggregators have relatively low dependencies on each other.

Method	Answerable		Overall
	Vocab	OCR	
SS-VS-SN	26.71	42.99	30.54
SS-SN-VS	26.88	43.11	30.65
VS-SN-SS	25.80	43.08	30.27
SN-SS-VS	26.58	42.97	30.46
SN-VS-SS	27.66	41.63	30.33
VS-SS-SN (ours)	**27.85**	**43.36**	**31.21**

Table 1. VQA accuracy (%) of variants of MM-GNN with different aggregators order on validation set of TextVQA dataset.

Results on different question types. Similar to VQA dataset [1], we categorize the questions in TextVQA into three groups based on their question-type, i.e., yes/no, number and others. The performances of MM-GNN and baseline No-GNN on different question types are shown in Table 2. We can see that our method mainly outperforms the baseline on others-type questions as expected, because these questions are mostly related to understanding diverse scene texts.

Model	yes/no	number	others	Final
No-GNN	88.79	35.14	22.65	27.55
MM-GNN	**88.93**	**36.13**	**27.36**	**31.21**

Table 2. VQA accuracy (%) of MM-GNN with on different types questions of TextVQA dataset compared to No-GNN.

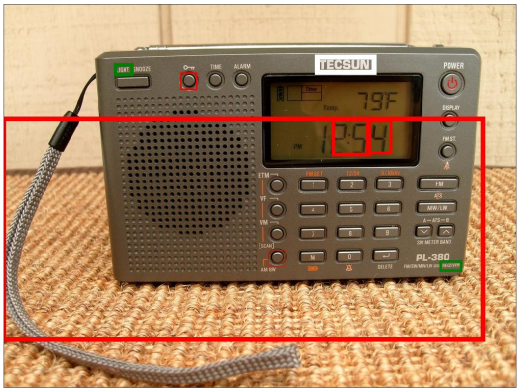
3. Qualitative Analysis

In Fig. 1, we show more successful cases for our MM-GNN model on TextVQA dataset. We show that MM-GNN obtains the correct answer, along with reasonable attention results when utilizing the multi-modal contexts. In Fig. 2, we show some failure cases of MM-GNN and analyze the possible reason for each example in the image.

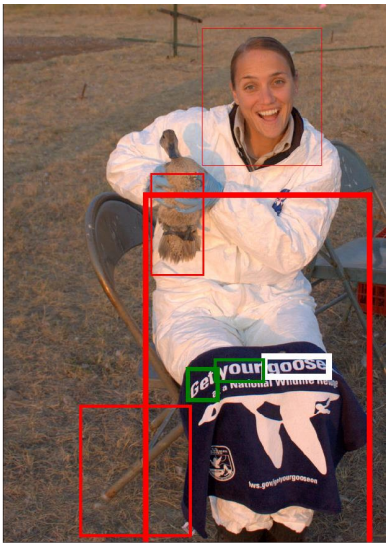
* indicates equal contribution.

Question: What kind of cidar is it?
OCR: sheppy 's, sheppy 's, dm, sheppy 's, molel, farmhgest, don, kaing, afmhousecide, witre
MM-GNN: sheppy's
Baseline: sheppy's

Question: What kind of fast food?
OCR: team, greenbaud, xoyxx, blackbaud' 60, 50, 45, 20, 20, 35, min, 25, 30, 15, 15minute, meals, delicious, food, fast, www.sonesoprictacooerscoo.om
MM-GNN: delicious
Baseline: 50

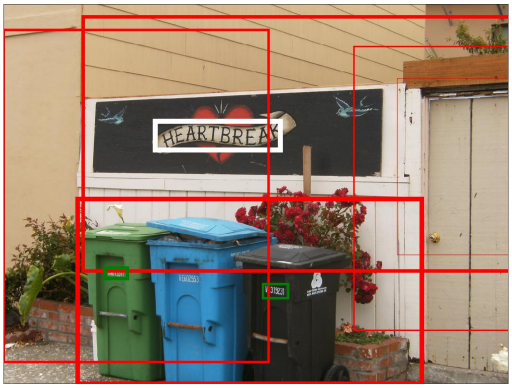


Question: What brand of radio is this?
OCR: light/, snooze, alarm, tecsun, power, 79f, display, mw/lw, swmeterband, pl-380, delete, fm/sw/mw/lw, receiver
MM-GNN: tecsun
Baseline: snooze



Question: Get you what?
OCR: get, ata, national, wildlife, reiwe, your, goose
MM-GNN: goose
Baseline: wildlife

Question: What kind of ale is this?
OCR: a6g'a, wedding, ale, encalireer, l.m.
MM-GNN: wedding
Baseline: wedding

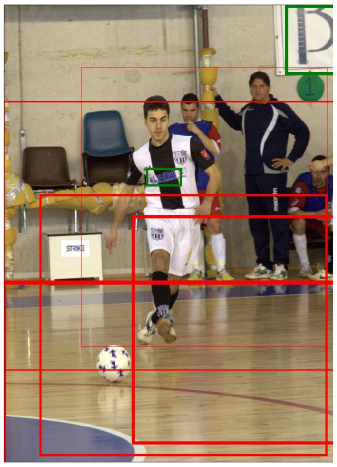


Question: What kind of break is mentioned above the trashcan?
OCR: heartbreak, bn613317, veb02953, vk319231
MM-GNN: heartbreak
Baseline: perfect

Question: What is the right number?
OCR: 8, 9, 10, 11
GT: 11
Baseline: 11

Figure 1. **Successful case analysis**. The predicted OCR is bounded in a white box. We show the attention from the predicted OCR tokens to the most attended five visual objects in Visual-Semantic aggregator (in red bounding boxes) and the attention between OCR tokens to the most attended two OCR tokens in Semantic-Semantic aggregator (in green bounding boxes), where bolder bounding box indicates higher attention value.

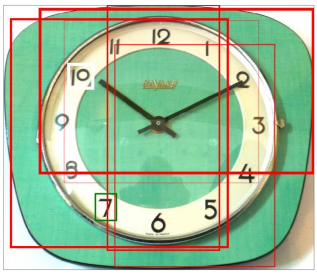
Question: Which brewing company makes this beer?
OCR: phillip, ldie, speeid, grazy8s, iua
GT: phillips
Our Answer: phillip
Possible Reason: MM-GNN attends to the right place, but the OCR system result is different to all annotators' answers. OCR failure makes up a large part of the error cases.



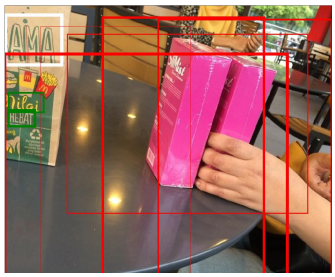
Question: What does the box say in the back?
OCR: eleom, srik
GT: strike
Our Answer: srik
Possible Reason: Similar to left, this time the attended text is printed in a confusing typeset. OCR can make various mistakes, due to vertically arranged words, too many words in image, widely spaced characters, bad illumination, etc.

Question: Where will I be if I turn right?
OCR: kumamoto, shin-yatsushiro, shin-tamana, uhtou3
GT: shin-tamana **Our Answer**: kumanoto
Possible Reason: MM-GNN probably has not seen white triangles indicating directions. Even for humans, unseen road signs can be bewildering.

Question: What drink are these?
OCR: oneli, onel, mbrusco, mbrusco, ell'emilia, azione, bgbograpica, tipica:, emilia, ddotto, in, italia, osato, ozzante
GT: ell emilia **Our Answer**: mbrusco
Possible Reason: Too many out-of-vocabulary words for MM-GNN in a single image. Image reading books written in a foreign language, when you cannot infer the meaning of one unknown word by another.



Question: What time does the clock read?
OCR: bavar!, 9, 3, 6, 4, 7, 6, 5
GT: 10:10
Our Answer: 10
Possible Reason: Reading wall clock requires special training even for humans. We do not equip MM-GNN with such one because we hope to focus on modality fusing. Imagine someone designs a weird new clock and not tell you how to read it, that's how MM-GNN feels.



Question: Where is this person eating at?
OCR: ama, m, ilai, hebat
GT: mcdonald's
Our Answer: ama
Possible Reason: Answering this question requires knowledge about the "M" logo for McDonald's (printed on the paper bag), together with surrounding visual information, to infer what the scene is. The provided information is not sufficient alone to give the right answer.

Question: What does the board say?
OCR: welcome, to, the, deep.spacediner
GT: welcome to the deep space diner
Our Answer: welcome
Possible Reason: MM-GNN does not have a sequential decoder, with accuracy severely damaged.

Figure 2. **Failure case analysis**. The predicted OCR is bounded in a white box. We show the attention from OCR tokens to the most attended five visual objects in Visual-Semantic aggregator (in red bounding boxes) and the attention between OCR tokens to the most attended two OCR tokens in Semantic-Semantic aggregator (in green bounding boxes), where bolder bounding box indicates higher attention value.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 1