Paper ID:1025

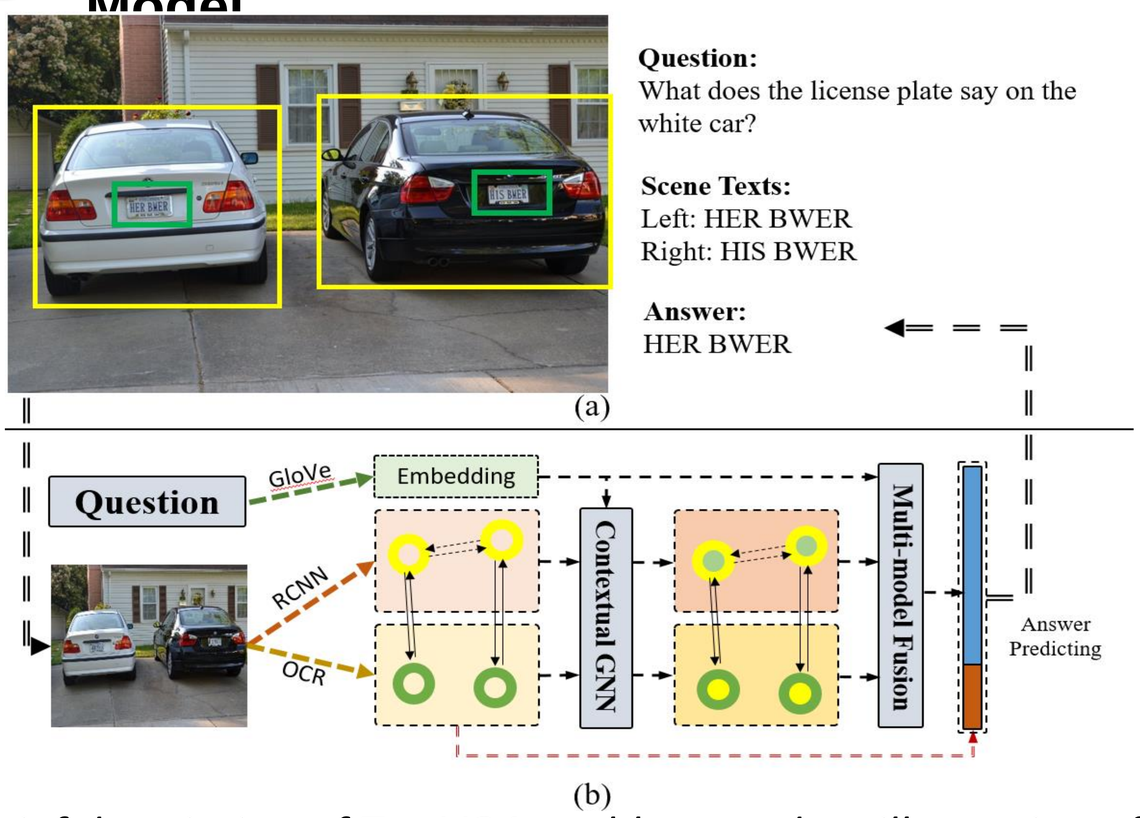
Multi-modal Contextual Graph Neural Network for Text Visual Question Answering

Yaoyuan Liang, Xin Wang, Xuguang Duan*, Wenwu Zhu*

Beijing University of Posts and Telecommunications

Tsinghua University

1 Illustration of TextVQA and MCG Model



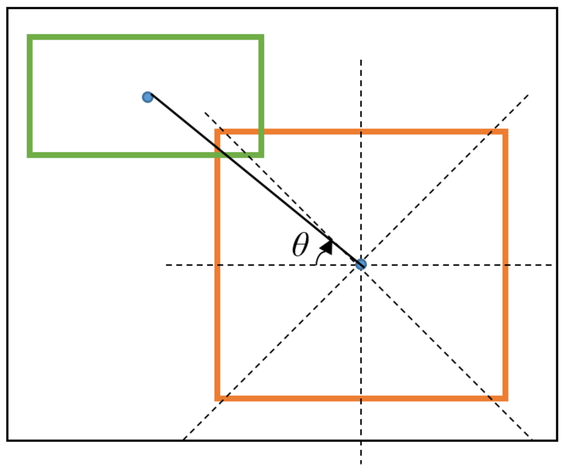
Brief description of TextVQA problem, and an illustration of our MCG model structure, which contains a GNN-based contextual information propagation mechanism.

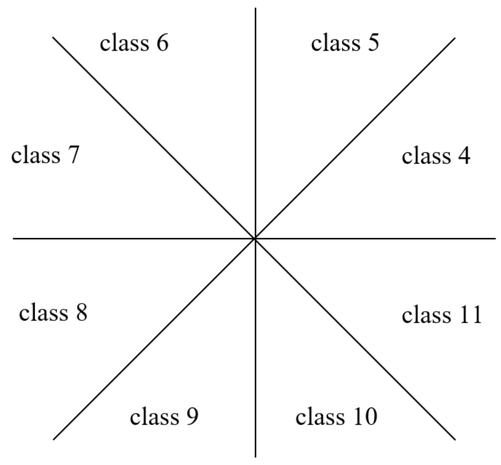
2 Visual and Textual Feature Extractor

Textual Object: Rosetta and Faster-RCNN ROI pooling.

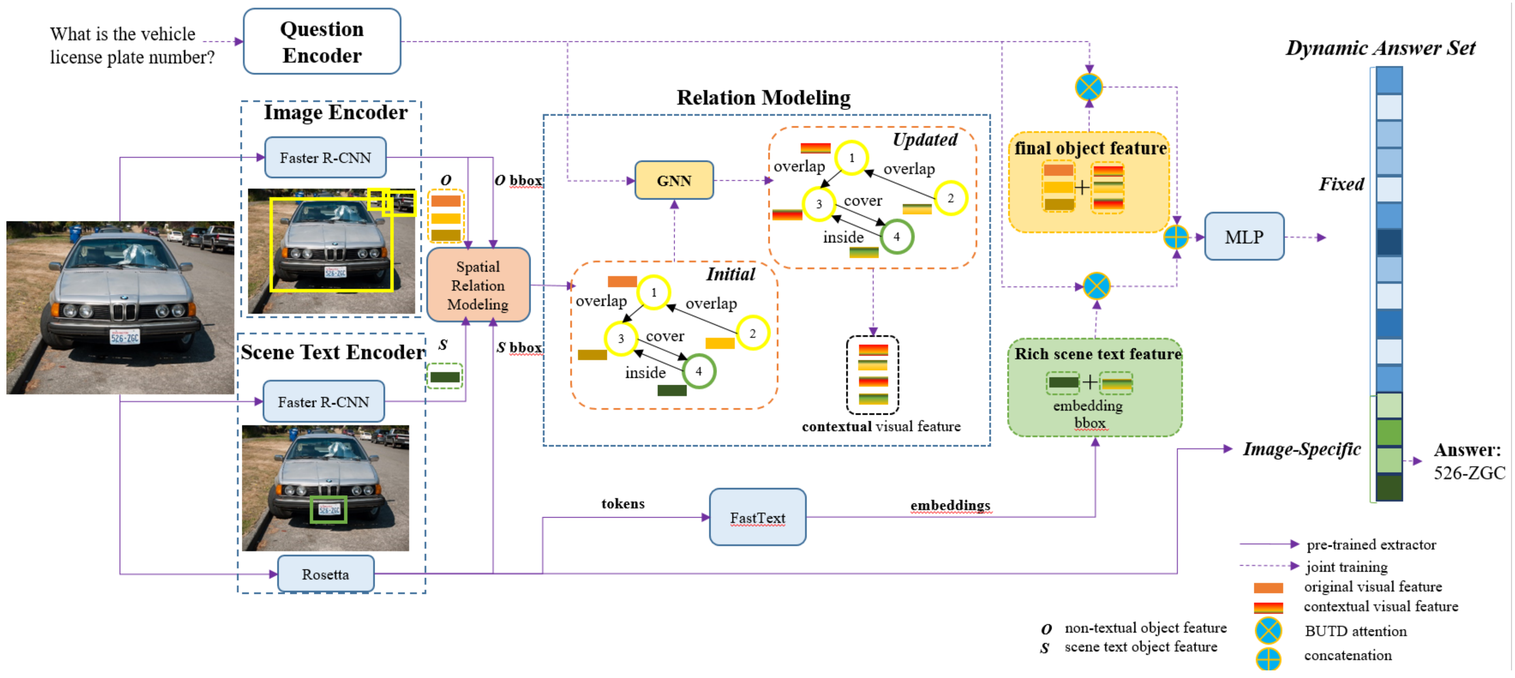
Non-textual Object: Faster-RCNN

3 Spatial Relationship Modelling





4 Overall Model Architecture and GNN Propagation Mechanism



$$\boldsymbol{v}_i^{(q)} = \sigma\left(\boldsymbol{q} \cdot \boldsymbol{W_q} \boldsymbol{v}_i\right), \ i = 1, 2, \cdots, K+M,$$

$$\boldsymbol{v}_i^{h+1} = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \cdot \boldsymbol{W_h} \boldsymbol{v}_j^h\right),$$

$$\alpha_{ij}^l = \frac{\exp\left(\left(\boldsymbol{U}^l \boldsymbol{v}_i^h\right)^\top \cdot \boldsymbol{V}^l \boldsymbol{v}_j^h\right)}{\sum_{j \in \mathcal{N}_i} \exp\left(\left(\boldsymbol{U}^l \boldsymbol{v}_i^h\right)^\top \cdot \boldsymbol{V}^l \boldsymbol{v}_j^h\right)},$$

$$\boldsymbol{v}_i^{h+1} = \Big\Vert_{l=1}^{L} \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^l \cdot \boldsymbol{W_h}^l \boldsymbol{v}_j^h\right)$$

5 Results & Illustration

OVERALL MODEL PERFORMANCE COMPARISION. THE VALIDATION SET ACCURACY (VAL)IS COMPUTED LOCALLY, WHILE THE TEST SET ACCURACY (TEST) IS OBTAINED THROUGH THE ONLINE JUDGING SYSTEM.

Model	Object Combine	OCR Combine	No.of GNN Layer	Rich OCR Feature	Acc. on Val	Acc. on Test
LoRRA [29]	–	–	–	–	26.56%	27.63%
MCG(max-pooling)	–	–	1	yes	17.85%	17.34%
MCG	residual	residual	1	yes	29.29%	29.29%
MCG	2 att.	concat.	1	yes	27.68%	27.91%
MCG	2 att.	residual	1	no	27.81%	27.98%
MCG	2 att.	residual	2	yes	28.71%	29.06%
MCG	2 att.	residual	1	yes	**29.40%**	**29.61%**

*What is the **name** of the **hotspot**?*
LoRRA: gates
MCG: vodafone

*What **company** is on the **advert**?*
LoRRA: zemel
MCG: nationwide

*What kind of **gps logger** is it?*
LoRRA: peceoi
MCG: wireless

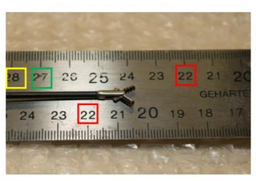
*What **brand** is the **yellow box**?*
LoRRA: eauking
MCG: triscuit

*How much **time** is left on the **washing machine**?*
LoRRA: 0
MCG: 120

*What **city** is **named**?*
LoRRA: new york
MCG: martinborough

Qualitative examples from our MCG model on TextVQA test set

*How many **way stop** is this **sign** for?*
LoRRA: 3
MCG: all
Human: 4



*What is the **largest number** on the **top row** of this **ruler**?*
LoRRA: 22
MCG: 27
Human: 28

*What does it say in **blue**?*
LoRRA: kullik
MCG: ilihakvik
Human: kullik ilihakvik

Faulty examples from our MCG model on TextVQA test set. We can infer that previous work LoRRA and our MCG model are weak in 2 aspects:

1) do not have the ability in predicting answers that require more than 1 token.
2) unable to split extracted OCR tokens according to the semantic clue given in question.

6 Reference

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8317–8326, 2019.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, pages 10313– 10322, 2019.

7 Contacts

lyy8ztc@outlook.com