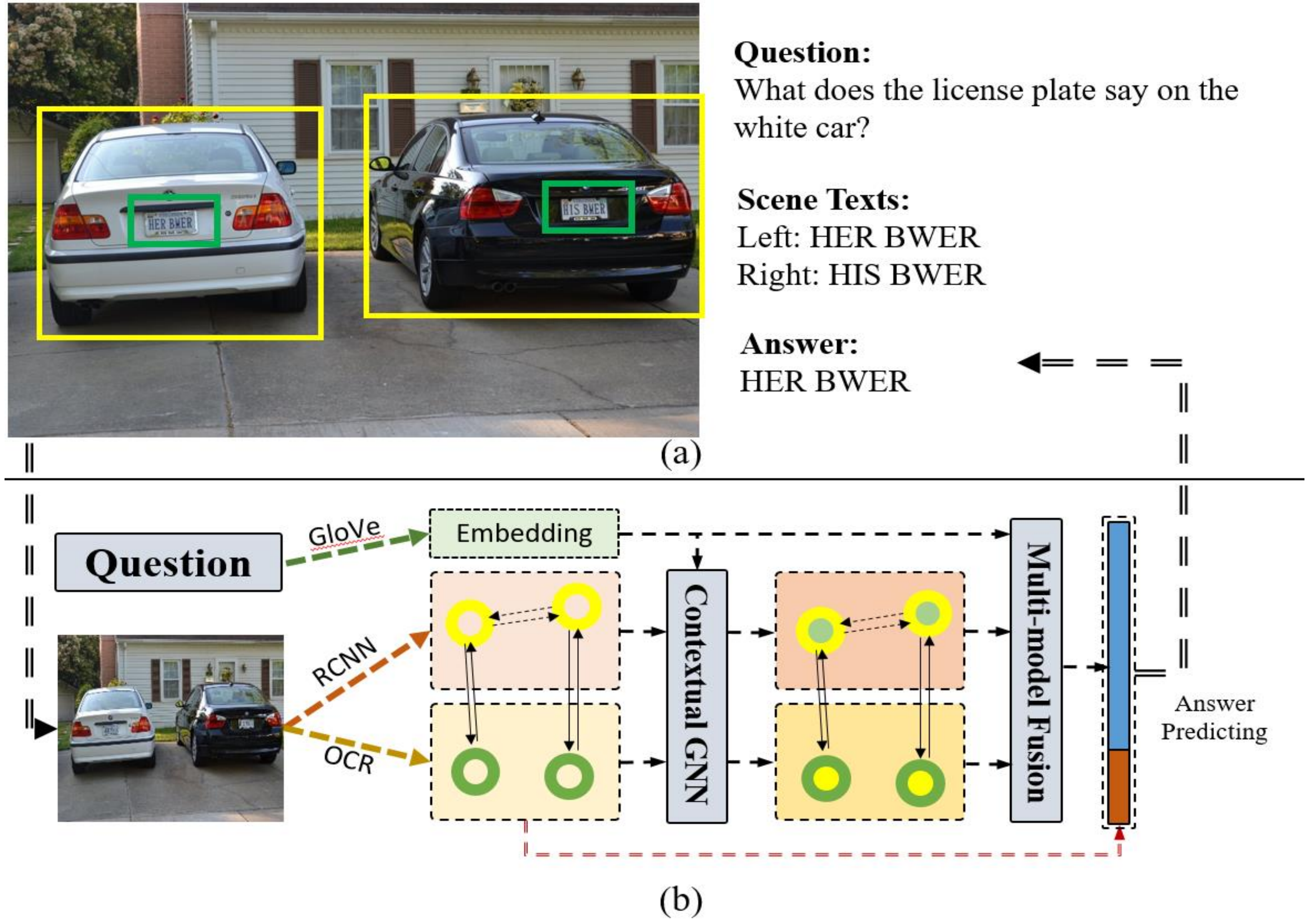
北京郵電大學
Beijing University of Posts and Telecommunications

清華大學
Tsinghua University

Multi-modal Contextual Graph Neural Network for TextVQA

Yaoyuan Liang
lyy8ztc@outlook.com

Illustration of TextVQA and MCG Model



Question:
What does the license plate say on the white car?

Scene Texts:
Left: HER BWER
Right: HIS BWER

Answer:
HER BWER

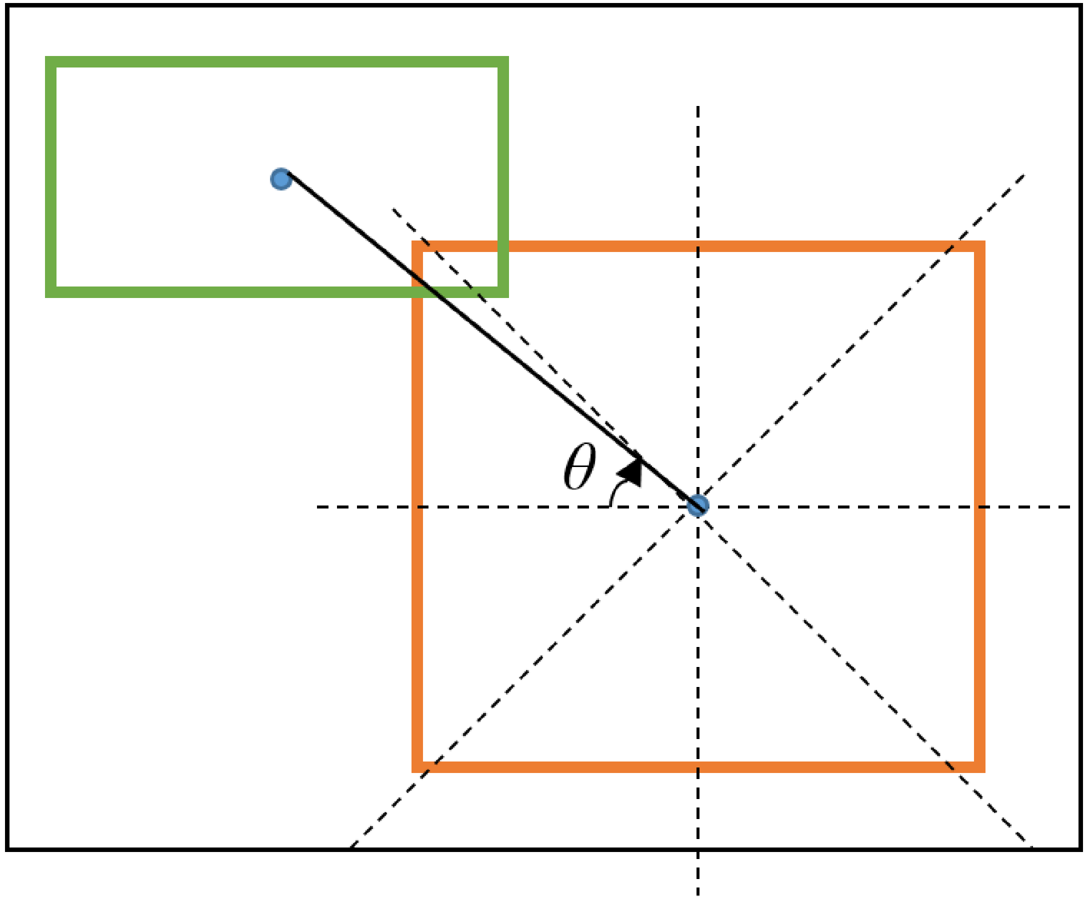
Brief description of TextVQA problem, and an illustration of our MCG model structure, which contains a GNN-based contextual information propagation mechanism.

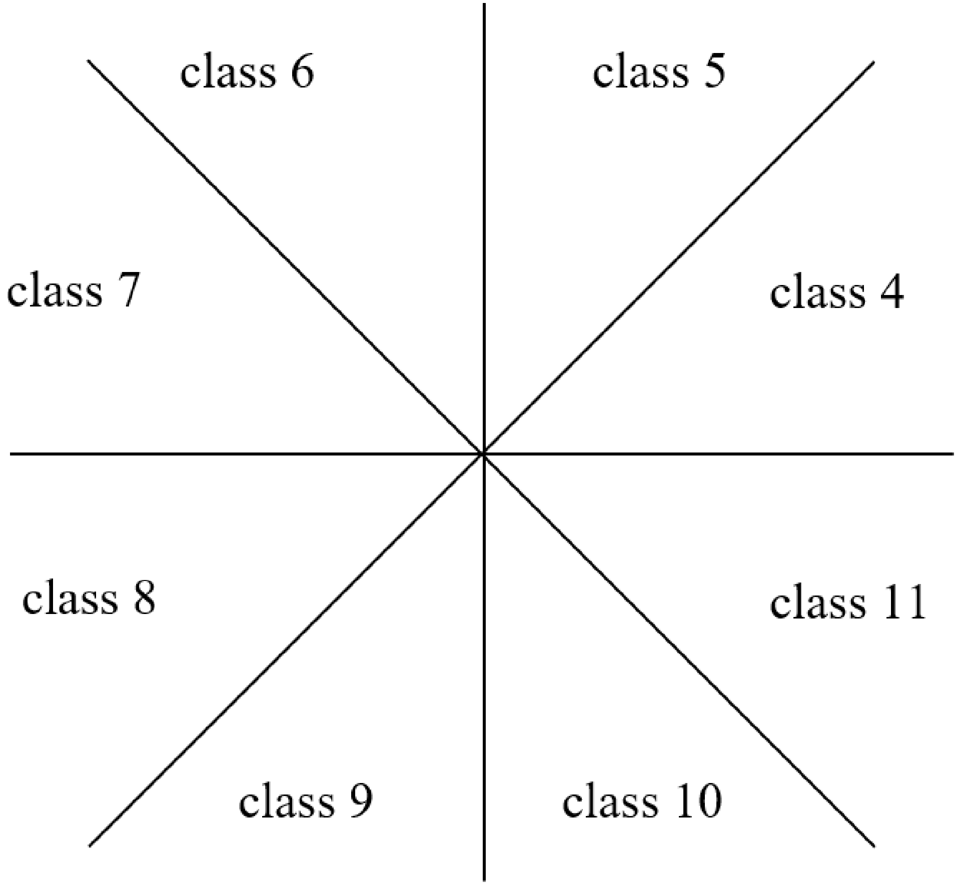
Encoding Component

- Non-textual object features are extracted with a pre-trained Faster-RCNN model.
- For the scene texts in the image, we apply scene text detector Rosetta to identify tokens in the image. We get tokens, visual bounding box, and visual feature of scene texts. The visual feature is extracted trough feeding the bounding box into the Faster-RCNN model.
- For the question, we follow the common practice as in other VQA works.

Relation Modeling Component

- Spatial Relationship modeling:





Contextual GNN Propagation Mechansim

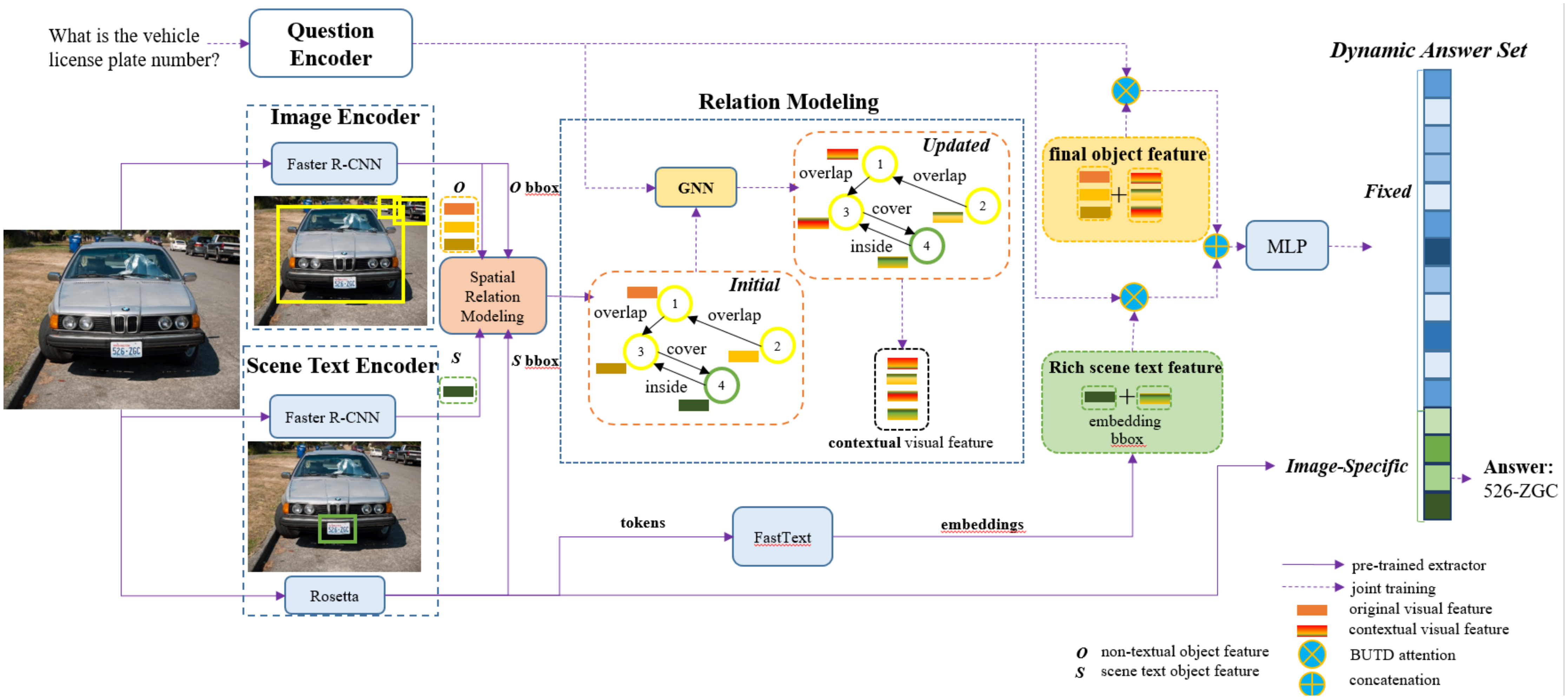
$$\boldsymbol{v}_i^{(q)} = \sigma\left(\boldsymbol{q} \cdot \boldsymbol{W}_{\boldsymbol{q}} \boldsymbol{v}_i\right),\ i = 1, 2, \cdots, K + M,$$

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

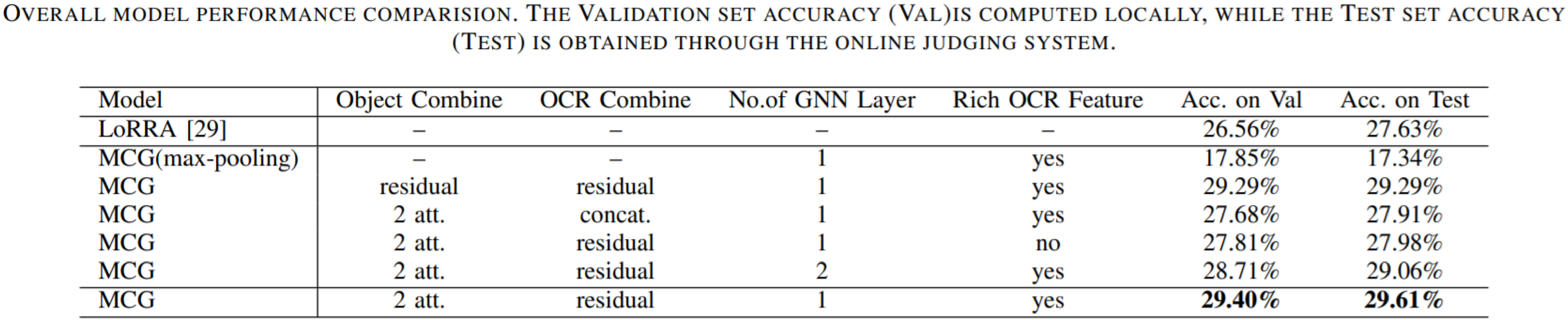
$$\alpha_{ij}^l = \frac{\exp\left(\left(\boldsymbol{U}^l \boldsymbol{v}_i^h\right)^\top \cdot \boldsymbol{V}^l \boldsymbol{v}_j^h\right)}{\sum_{j \in \mathcal{N}_i} \exp\left(\left(\boldsymbol{U}^l \boldsymbol{v}_i^h\right)^\top \cdot \boldsymbol{V}^l \boldsymbol{v}_j^h\right)},$$

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

Model Architecture



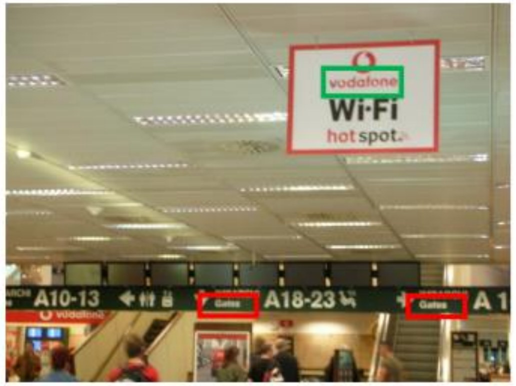
Results



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.

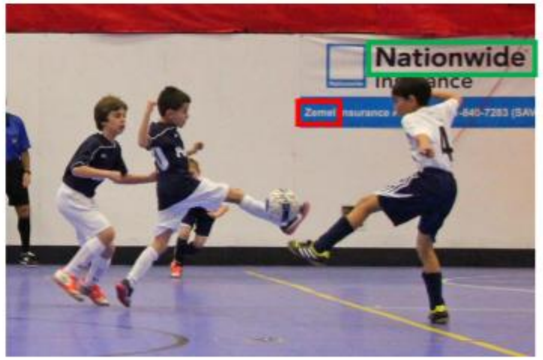
Model	Object Combine	OCR Combine	No.of GNN Layer	Rich OCR Feature	Acc. on Val	Acc. on Test
LoRRA [29]	–	–	–	–	26.56%	27.63%
MCG(max-pooling)	–	–	1	yes	17.85%	17.34%
MCG	residual	residual	1	yes	29.29%	29.29%
MCG	2 att.	concat.	1	yes	27.68%	27.91%
MCG	2 att.	residual	1	no	27.81%	27.98%
MCG	2 att.	residual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

Results-Qualitive



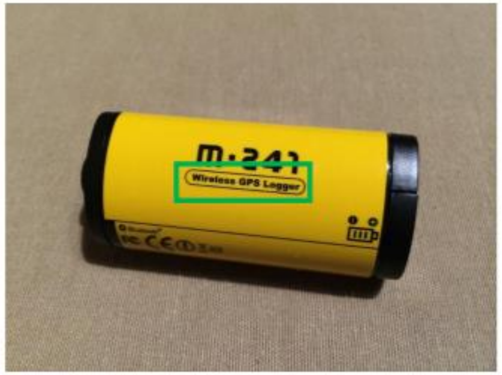
*What is the **name** of the **hotspot**?*

LoRRA: gates
MCG: vodafone



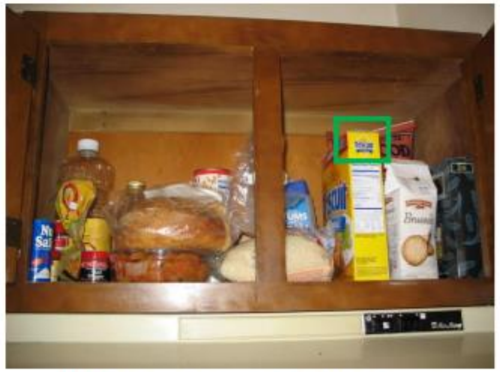
*What **company** is on the **advert**?*

LoRRA: zemel
MCG: nationwide



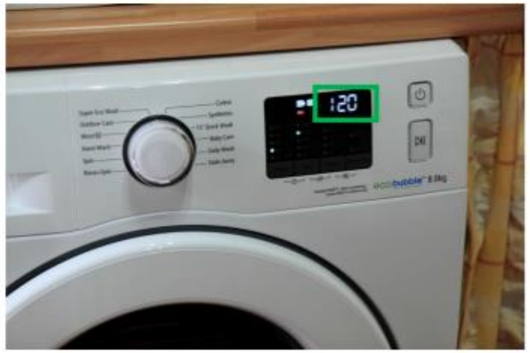
*What kind of **gps logger** is it?*

LoRRA: peceoi
MCG: wireless



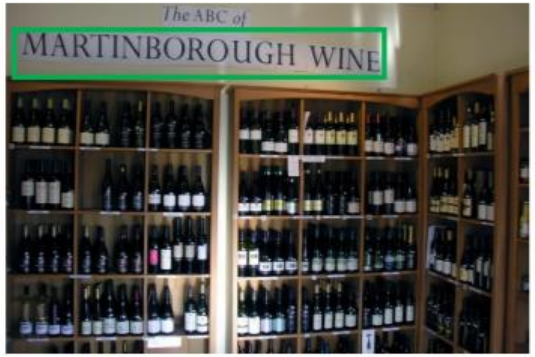
*What **brand** is the **yellow box**?*

LoRRA: eauking
MCG: triscuit



*How much **time** is left on the **washing machine**?*

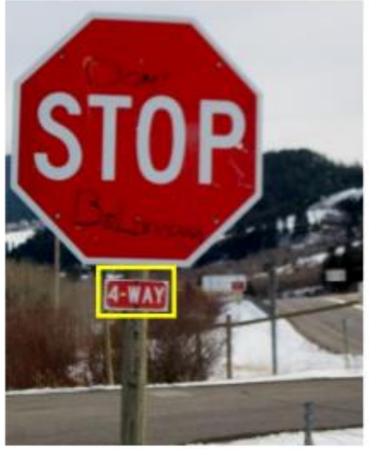
LoRRA: 0
MCG: 120



*What **city** is **named**?*

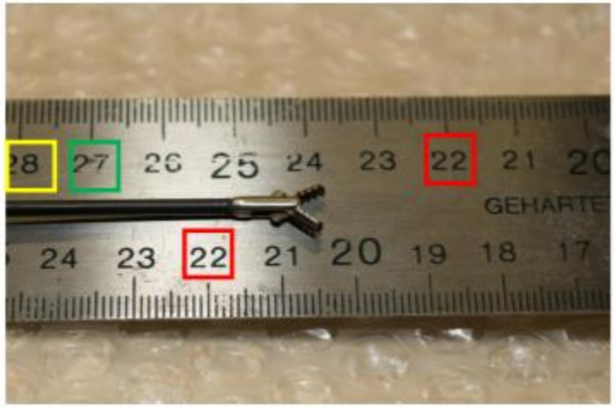
LoRRA: new york
MCG: martinborough

Results-Faulty



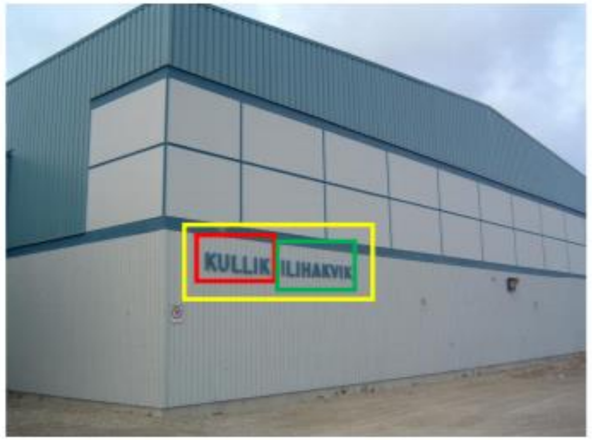
*How many **way stop** is this **sign** for?*

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