Guiding Scene Text Question Answering with Region Priors Appendix

Anonymous submission
Paper ID: 321

1 Feature Encoder

Question Encoder. We use a pretrained Faster R-CNN detector (Ren et al. 2015) to obtain M visual object labels and use the corresponding OCR system to obtain N OCR tokens. To better model the interaction between the question and texts in the image, we first extend question words with scene text and object labels, and then we adopt a four-layer BERT (Devlin et al. 2019) model to encode the enriched texts into a sequence of d-dimensional feature vectors $\{x_k^{ques}\}_{k=1}^K$,$\{x_m^{obj,text}\}_{m=1}^M$ and $\{x_n^{ocr,text}\}_{n=1}^N$. We take $\{x_k^{ques}\}_{k=1}^K$ as the final question features.

Object Encoder. We extract appearance features $\{x_m^{fr}\}_{m=1}^M$ and 4-dimensional location features $\{x_m^{\text{b}}\}_{m=1}^M$ for the M detected objects. For the m-th object, we define its location feature x_m^{b} as its normalized bounding box coordinates as follows,

$$[\frac{x_m^{\text{ul}}}{H}, \frac{x_m^{\text{ul}}}{W}, \frac{x_m^{\text{lr}}}{H}, \frac{x_m^{\text{lr}}}{W}], \tag{1}$$

where $(x_m^{\text{ul}}, y_m^{\text{ul}})$, $(x_m^{\text{lr}}, y_m^{\text{lr}})$, $(x_m^{\text{c}}, y_m^{\text{c}})$ are the bounding box coordinates of the upper left and lower right, respectively. And the (h_m, w_m) and (H, W) are the height and width of the bounding box and the image, respectively. We enhance the appearance features by concatenating x_m^{fr} and $x_m^{obj,text}$. Following prior work (Hu et al. 2020), we fuse them with location features to obtain the final object feature representation $\{x_m^{\text{obj}}\}_{m=1}^M$ via the stack of learned linear transforms, element-wise addition and the layer normalization (Ba, Kiros, and Hinton 2016).

OCR Encoder. Given N OCR tokens in the image, similar to prior works (Hu et al. 2020; Yang et al. 2021), we extract their OCR feature representation $\{x_n^{\text{ocr}}\}_{n=1}^N$ and location feature representation $\{x_n^{\text{b}}\}_{n=1}^N$. Specifically, the location feature representation of OCR tokens is enriched with the bounding box coordinate of center, the height and width of the bounding box. The OCR feature representation is fused from the BERT vector, the FastText vector (Bojanowski et al. 2017), the Faster R-CNN feature (Ren et al. 2015), the PHOC vector (Almazán et al. 2014) and location feature via the same fusion method used in object encoder.

Hyper-parameters	Value of TAP	Value of ROQ
max length of question word, K	20	20
max number of visual object, N	100	100
max number of OCR token, M	100	**50**
optimizer	Adam	Adam
batch size	128	128
base learning rate	1e-4	1e-4
learning rate decay	0.1	0.1
learning rate steps	14k, 19k	**14k, 15k**
max iterations of pre-training	24k	**0**
max iterations of fine-tuning	24k	24k

Table 1: Hyper-parameters of the ROQ and TAP. We compare our method ROQ with TAP and highlight the changed hyper-parameters in bold.

2 Answer Generation

We perform a query selection to make the query focus on the information to predict the answers. Given the set of queries $\boldsymbol{x}^{\text{ques}} \in R^{d\times K}$, the computation for the selected query vector q^{ques} is formulated as follows,

$$\begin{aligned} \alpha_k &= \text{Softmax}(\text{MLP}(\boldsymbol{x}_k^{ques})), \\ q^{\text{ques}} &= \sum_{k=1}^{K} \alpha_k x_k^{\text{ques}}, \end{aligned} \tag{2}$$

where $\text{MLP}(\cdot)$ denotes the multilayer perceptron. Also, we compute the selected query vectors q^{reg} and q^{ocr} from the query results $\boldsymbol{f}^{\text{reg}}$ and $\boldsymbol{f}^{\text{ocr}}$ from Region Query module and OCR Query module, respectively.

Following previous works (Singh et al. 2019; Hu et al. 2020), we extend the fixed answer vocabulary with M OCR tokens, which forms our answer space. During decoding process, we pick the item that gets highest score among the answer space as the answer token at each step. The maximum decoding step is 12.

3 Implementation Details

The main hyper-parameters are listed in Table 1. Our most hyper-parameters are in the same setting as TAP (Yang et al. 2021) and the changed parameters are highlighted in bold.

4 Datasets

TextVQA dataset (Singh et al. 2019) contains 45,336 human-written questions on 28,408 images from Open Image dataset (Kuznetsova et al. 2020). Following the previous work (Hu et al. 2020), we adopt the same train/validation/test split setting. Specifically, 21,953, 3,166 and 3,289 images for train, validation, and test set. Follwing VQAv2 (Antol et al. 2015), the accuracy is evaluated by soft voting over 10 answers provided from different annotators.

ST-VQA dataset (Biten et al. 2019) contains 23,038 images with 31,791 questions from six different datasets to reduce dataset bias, including ICDAR 2013 (Karatzas et al. 2013) and ICDAR2015 (Karatzas et al. 2015), ImageNet (Kuznetsova et al. 2020), VizWiz (Gurari et al. 2018), IIIT STR (Mishra, Alahari, and Jawahar 2013), Visual Genome (Krishna et al. 2017) and COCO-Text (Veit et al. 2016). We follow the same setting from M4C (Hu et al. 2020) that splits the dataset into train, validation, and test sets with 17,208, 1,893 and 2,971 images respectively. Two metrics including accuracy and Average Normalized Levenshtein Similarity(ANLS) are reported on this dataset.

5 Qualitative Examples

In this section, we present additional qualitative results of our ROQ model.

In the Figure 1, we visualize the explainable visual evidence of the stepwise query processes. Qualitatively, the proposed ROQ can firstly ground the salient object region and then identify the target OCR tokens that contribute to the answer.

In the Figure 2, we display several challenging examples along with the predicted results of our ROQ, M4C (Hu et al. 2020), and TAP (Yang et al. 2021). Compared to M4C and TAP, our model could correctly ground the corresponding object region when the scenario is complex (*e.g.*, the third image in the top row).

Moreover, we show some typical failed cases of our model in the Figure 3. The wrong OCR recognition of the OCR system and lack of external knowledge are the main reasons why our model fails to predict correctly. For the third image in the top row, the OCR system misidentifies "2013" to "20d", which causes the failure. For the first image in the bottom row, our method fails to determine the type of the game without external knowledge.

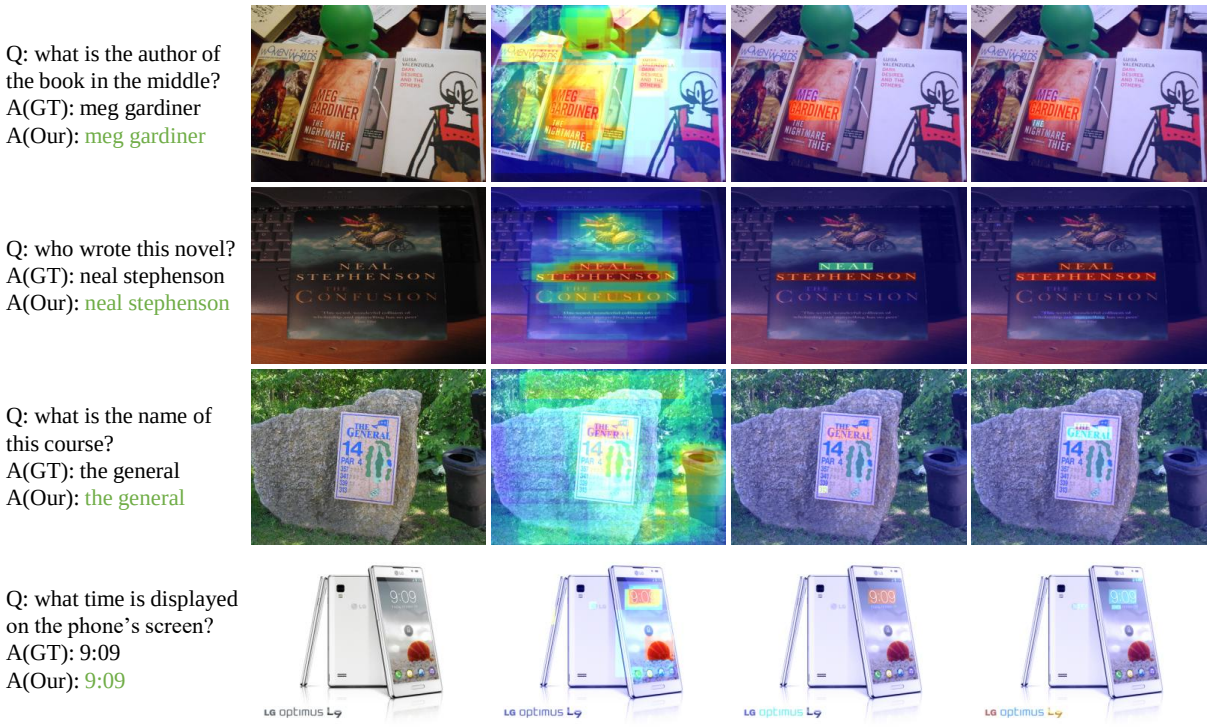


Figure 1: Additional qualitative results show explainable visual evidence of the stepwise query processes.

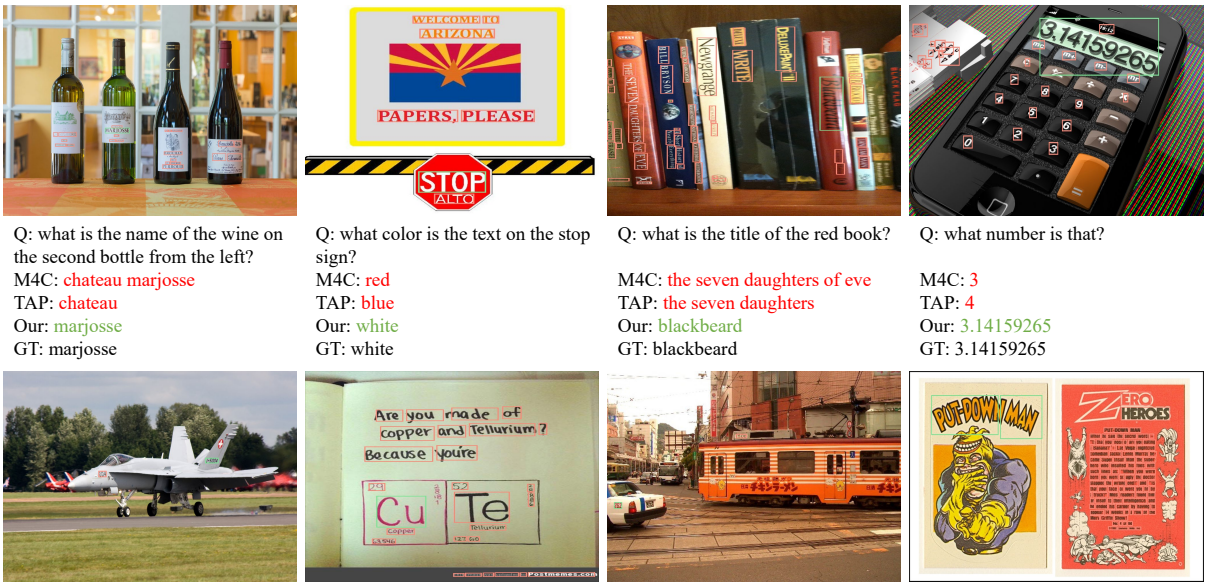


Figure 2: Additional examples along with the predicted results of our ROQ, M4C, and TAP on the TextVQA validation set.

Q: what is on his shirt?

Our: f
GT: w

Q: what letter is on the red shorts?
Our: a
GT: m

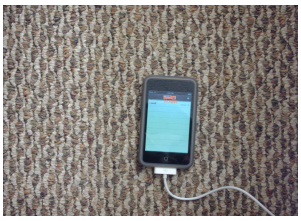
Q: when was this drawn?

Our: 20d
GT: 2013

Q: what team name is the little guy wearing?
Our: cyclo
GT: cyclones

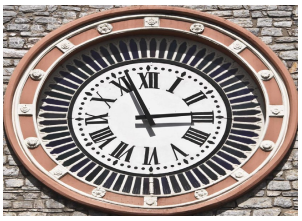
Q: what game is being plauyed?

Our: mets
GT: baseball



Q: this phones brand is?

Our: no
GT: iphone



Q: what is the time right now?

Our: 8:15
GT: 2:57

Q: what state does this mug have on it?
Our: daytona
GT: florida

Figure 3: Examples that our model fails on the TextVQA validation set.

References

Almazán, J.; Gordo, A.; Fornés, A.; and Valveny, E. 2014. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12): 2552–2566.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Biten, A. F.; Tito, R.; Mafla, A.; Gomez, L.; Rusinol, M.; Valveny, E.; Jawahar, C.; and Karatzas, D. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4291–4301.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5: 135–146.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3608–3617.

Hu, R.; Singh, A.; Darrell, T.; and Rohrbach, M. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9992–10002.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 1156–1160. IEEE.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, 1484–1493. IEEE.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1): 32–73.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4. *International Journal of Computer Vision*, 128(7): 1956–1981.

Mishra, A.; Alahari, K.; and Jawahar, C. 2013. Image retrieval using textual cues. In *Proceedings of the IEEE International Conference on Computer Vision*, 3040–3047.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Veit, A.; Matera, T.; Neumann, L.; Matas, J.; and Belongie, S. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*.

Yang, Z.; Lu, Y.; Wang, J.; Yin, X.; Florencio, D.; Wang, L.; Zhang, C.; Zhang, L.; and Luo, J. 2021. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8751–8761.