Supplementary Materials of TAG: Boosting Text-VQA via Text-aware Visual Question-answer Generation

Jun Wang*[1]
junwang@umiacs.umd.edu
Mingfei Gao[2]
mingfei.gao@salesforce.com
Yuqian Hu[1]
yhu1109@umd.edu
Ramprasaath R. Selvaraju[2]
rselvaraju@salesforce.com
Chetan Ramaiah[2]
cramaiah@salesforce.com
Ran Xu[2]
ran.xu@salesforce.com
Joseph F. JaJa[1]
josephj@umd.edu
Larry S. Davis[1]
lsd@umiacs.umd.edu

[1] University of Maryland, College Park, MD. USA.
[2] Salesforce Research, Palo Alto, CA. USA.

1 Additional Details of Our Approach

Text-aware Visual Question Prediction. We leverage the powerful capability of the attention mechanism in transformers [1] to capture the interactions among the extended answer words, visual objects, and OCR tokens. Our decoding module is based on a dynamic pointer network [2], which allows both copying words via pointing, and generating words from a fixed vocabulary obtained from the training set.†

Hyper-parameters. Table. A1 overviews the hyper-parameter settings of TAG. We use the original parameter settings of downstream Text-VQA models except that we increase the maximum number of iteration in proportion to the increased size of the augmented data to accommodate the enlarged number of training samples.

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

*Part of this work was done when JW was an intern at Salesforce Research.

†Our decoding module is implemented following the implementation of the decoding module in TAP [4]. We keep their default hyper-parameters except otherwise noted.

Hyper-parameters	Value
Max length of answer words w^{ans}	20
Max length M of visual objects w^{obj}	100
Max length N of scene texts w^{ocr}	100
Max length K of extended answer words	220
max length T of decoding step	30
optimizer	Adam
batch size	128
max iterations	24K
base learning rate	1e-4
learning rate steps	14K, 19K
learning rate decay	0.1

Table A1: Hyper-parameters of TAG.

2 Additional Qualitatively Visualization of TAG

We present additional visualization results of generated QA pairs by TAG on the TextVQA training set in Figure. A1.

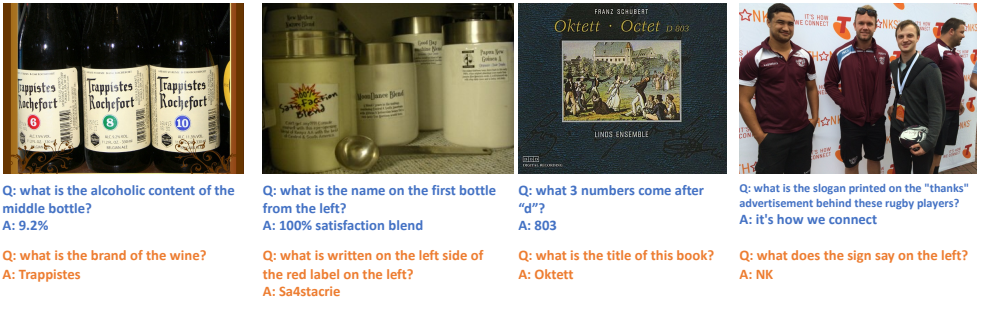

Figure A1: We visualize the examples of the generated QA pairs (bottom in orange) by the TAG module compared with the original annotated QA pairs (top in blue) on the TextVQA training set. "Q" and "A" refer to question and answer, respectively. Best viewed in color.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. *Advances in neural information processing systems*, 28, 2015.

[3] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8761, 2021.