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

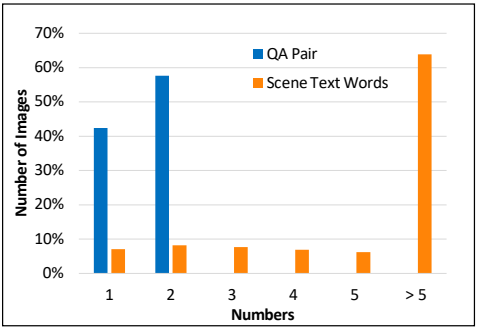
Jun Wang*[1]
junwang@umiacs.umd.edu
Mingfei Gao[2]
mingfei.gao@salesforce.com
Yuqian Hu[1]
yhu1109@umd.edu
Ramprasaath R. Selvaraju[2]
rselvaraju@salesforce.com
Chetan Ramaiah[2]
cramaiah@salesforce.com
Ran Xu[2]
ran.xu@salesforce.com
Joseph F. JaJa[1]
josephj@umd.edu
Larry S. Davis[1]
lsd@umiacs.umd.edu

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

Abstract

Text-VQA aims at answering questions that require understanding the textual cues in an image. Despite the great progress of existing Text-VQA methods, their performance suffers from insufficient human-labeled question-answer (QA) pairs. However, we observe that, in general, the scene text is not fully exploited in the existing datasets– only a small portion of the text in each image participates in the annotated QA activities. This results in a huge waste of useful information. To address this deficiency, we develop a new method to generate high-quality and diverse QA pairs by explicitly utilizing the existing rich text available in the scene context of each image. Specifically, we propose, TAG, a text-aware visual question-answer generation architecture that learns to produce meaningful, and accurate QA samples using a multimodal transformer. The architecture exploits underexplored scene text information and enhances scene understanding of Text-VQA models by combining the generated QA pairs with the initial training data. Extensive experimental results on two well-known Text-VQA benchmarks (TextVQA and ST-VQA) demonstrate that our proposed TAG effectively enlarges the training data that helps improve the Text-VQA performance without extra labeling effort. Moreover, our model outperforms state-of-the-art approaches that are pre-trained with extra large-scale data. Code is available here.

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



Statistics in TextVQA Training Set

Training Example

Figure 1: **Left:** statistics of the numbers of annotated QA pairs (blue) and scene text words (orange) for each image in the TextVQA training set [41]. Clearly, the scene text words are not fully leveraged in the annotations. **Right:** an example of a training image with more than 5 scene text words, which is typical. Best viewed in color.

1 Introduction

Visual question answering (VQA) task [5] aims at inferring the answer to a question based on a holistic understanding of an image. It facilitates many AI applications such as robot interactions [4], document analysis [36] and assistance for visually impaired people [2]. Text-VQA specifically addresses question answering requests where reasoning text in an image is essential to answer a question. It is a more challenging task in a sense that it requires not only understanding the question and the visual context, but also the embedded text in an image [41]. To achieve this goal, Text-VQA methods [22, 25, 49] aim at studying the interactions among question words, visual objects, and scene text in an image. Recent approaches have focused on either improving transformer-based architectures [43] in a multi-modal manner [16, 21, 22, 25, 32, 34, 54], or adopting pre-training using additional large-scale data [49] to further boost their model performance. All of these methods heavily rely on annotations of question-answer (QA) pairs for model training. Intuitively, the more annotated pairs are leveraged, the better performance a model can achieve. Thanks to the development of text-related VQA datasets [8, 20, 36, 47], Text-VQA has achieved rapid progress.

However, the amount of Text-VQA annotations available is still limited due to the sparse labeling of QA pairs in recent datasets. Consider for example the TextVQA dataset [41] whose statistics are illustrated in Figure 1. It shows that only one or at most two QA pairs are annotated in the training images. Meanwhile, we also compute the number of text words presented in each image* and observe that most of the images contain at least 5 text words. This observation indicates that scene text is not fully utilized in the annotations, and hence not fully leveraged by recent methods. A natural question would be – can we fully take advantage of text words in images without incurring extra annotation costs?

As illustrated in Figure 2, we propose to tackle the problem by learning to generate large-scale and diverse text-related QA pairs from existing Text-VQA datasets, using the generated QA pairs to expand the training set and ultimately improving Text-VQA models.

*We compute the average of different OCR tokens acquired by the Microsoft-OCR system [49] for the entire training set.

Towards this end, we introduce TAG, a text-aware QA generation model, that generates novel text-related QA pairs at scale. It takes text words (the answer) as one of the inputs and aims at generating a question corresponding to this answer by leveraging the rich visual and scene textual cues. TAG is trained using the originally annotated QA pairs and adapts to generate new QA pairs containing scene text words in images that are not utilized in original annotations. No extra human annotation is required in our framework, so the size and diversity of the training data could be easily and largely increased. Since our generation process is disentangled with the training of Text-VQA models, our generated QA pairs can be used by most of the recent methods.

In summary, we introduce a simple yet efficient text-aware generation approach, which automatically and efficiently generates new QA pairs to improve the performance of the current Text-VQA methods. The main contributions of our work are three-fold:

- We identify and analyze possible deficiencies of current Text-VQA datasets– sparse annotations of QA pairs - and propose to better utilize unused scene text information within each image to improve the model performance.
- To the best of our knowledge, TAG is the first method that explores scene text-related QA pairs generation for improving Text-VQA tasks without additional labeled data.
- We consistently demonstrate the effectiveness of our method with two recent Text-VQA models on two Text-VQA datasets. The experimental results suggest that the existing Text-VQA algorithms can benefit from training with the high-quality and diverse QA pairs generated by our method.

2 Related Work

2.1 Text-related VQA

To study and evaluate the Text-VQA task, several scene text-based datasets are introduced, including VizWiz [20], OCR-VQA[36], TextVQA [41], and ST-VQA [8]. With the help of these datasets, numerous approaches have been proposed in recent years which increasingly improve Text-VQA performance [3, 9, 16, 17, 21, 22, 23, 25, 32, 34, 41, 49, 50, 51, 54]. LoRRA [41] is an early work that extends the original VQA models [3, 23] with an extra OCR attention branch to select the answer from either a fixed word vocabulary or OCR tokens. Recent studies [6, 11, 12, 14, 15, 19, 33, 45, 46, 52, 53] show the benefits of transformer for different vision, language and speech tasks. M4C [22] develops a transformer-based architecture to fuse different input modalities and iteratively predicts answers through a multi-step answer decoder. Inspired by M4C, more transformer-based models have been proposed with varied structure modifications. Among them, CRN [32] constructs a graph network to model the interactions between text and visual objects. LaAP-Net [21] predicts a bounding box to explain the generated answer. SSBaseline [54] proposes to split the OCR token features into separate visual and linguistic attention branches. SMA [16] reasons over structural text-object graphs and produces answers in a generative way. LOGOS [34] introduces a question-visual grounding pre-training task to connect question text and image regions. SA-M4C [25] builds a spatial graph to explicitly model relative spatial relations between visual objects and OCR tokens. TAP [49] presents three text-aware pre-training tasks to align representations among scene text, text words, and visual objects. However, most of

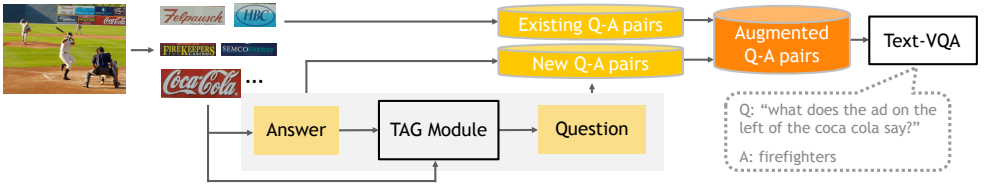


Figure 2: **The proposed Text-VQA framework.** It consists of two parts: a text-aware visual question-answer generation module (TAG), followed by a downstream Text-VQA model. TAG is based on a multi-modal transformer architecture, which takes a single image and text words (the answer) as input, and outputs a newly-generated question corresponding to the input answer. The generated QA pairs from TAG together with the originally labeled data are subsequently used to train Text-VQA models, leading to better Text-VQA performance.

the existing works focus on designing sophisticated architectures that leverage the annotated text in an image and overlook the rich text information that is underused by the annotated QA activities. We fully explore the embedded scene text in images and explicitly generate novel QA pairs that can be used to boost the performance of downstream Text-VQA models.

2.2 Data Augmentation for VQA

Data augmentation has been demonstrated to be an effective approach to improve the performance of the VQA task [1, 24, 26, 37, 39, 42, 48]. Kafle et al. [24] propose to generate new questions using the existing semantic segmentation annotations and templates. Shah et al. [39] introduce a cycle-consistent scheme generating question rephrasings to make VQA models more robust to linguistic variations. Ray et al. [37] propose a consistency-improving data augmentation module to make VQA models answer consistently. Agarwal et al. [1] explore data augmentation to improve the VQA model's robustness to semantic visual variations. Tang et al. [42] use data augmentation to inject proper inductive biases into the VQA model. Wang et al. [48] introduce a generative model for cross-modal data augmentation on VQA. Kant et al. [26] adopt the contrastive loss to make the VQA model robust to linguistic variations in generated questions. However, these approaches are designed for the traditional VQA systems that do not emphasize the importance of scene text in their QA tasks. Our method is tailored for the problem of Text-VQA. It takes advantage of the underexploited scene text in images and enlarges the training samples by generating novel text-related QA pairs without the extra labeling cost.

3 Our Approach

The proposed framework is illustrated in Figure 2, which consists of a transformer-based text-aware visual QA generation module named TAG, followed by a downstream Text-VQA model. Our core module, TAG, carries out text-aware data augmentation tailored for the Text-VQA task and generates novel QA pairs by leveraging underused scene text in an image. After the TAG module generates a large amount of new QA pairs, we directly augment the training data by combining the generated set and the originally labeled set. The augmented set is used by the downstream Text-VQA models to boost the model performance.

The workflow of our method is as follows. Given an image, an OCR system and an

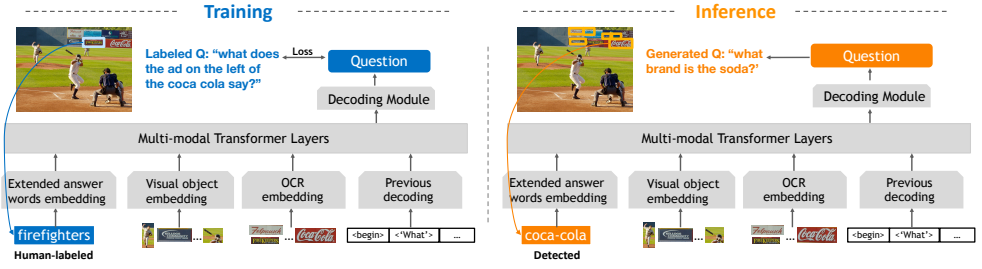


Figure 3: **The illustration of our proposed TAG.** High-dimensional feature representations are first extracted for three modalities, including extended answer words, visual objects, and scene text. Then, a multi-modal transformer is used to model the interactions of different modalities. Finally, a decoding module is used to predict the question corresponding to the answer through iterative decoding with an auto-regressive mechanism. **Left:** Originally labeled QA pairs are used for training. **Right:** During inference, detected OCR words are used as a novel answer to generate a question. Best viewed in color.

object detector are used to detect scene text and visual objects, respectively. As illustrated in Figure 3, our TAG takes the scene text words of interest (the answer words), the visual objects and all the detected OCR tokens in the image as inputs and generates a question explicitly corresponding to the answer. Specifically, the answer words, visual objects, and all the OCR tokens are first represented by high-dimensional features (Section 3.1). Then, the multi-modality information is fully aggregated through a transformer architecture with the attention mechanism (Section 3.2). Finally, the enriched features are used to predict a question to the answer through iterative decoding in an auto-regressive manner (Section 3.3). More details can be found in the supplementary.

3.1 Multi-modality Feature Embeddings

We describe the feature embedding strategy of our work. The answer words, detected visual objects, and all the detected OCR tokens are embedded as high-dimensional features and then projected into a common d-dimensional embedding space.

Embedding of extended answer words. We follow [49] to use an extended representation to embed answer words. Given an answer input w^{ans}, we extend the words with labels of objects w^{obj} (detected from the object detector) and scene text OCR words w^{ocr} (generated from the OCR system) as a set of K text words. A trainable BERT-style model [14] is adopted to extract the embedding of those text words, $\boldsymbol{F}^{ans} = \{\boldsymbol{f}_1^{ans}, \boldsymbol{f}_2^{ans}, ..., \boldsymbol{f}_K^{ans}\}$, where $k = \{1, 2, ..., K\}$, and $\boldsymbol{f}_k^{ans}$ is the d-dimensional feature vectors for k_{th} text word. The embeddings of the set of words are used jointly as the feature of the answer.

Embedding of detected objects. Following M4C [22], we run a pre-trained 2D object detector, Faster R-CNN [38] to localize M visual objects for each image. Two visual object features, including appearance and location features are extracted and then combined together to encode each detected object, $\boldsymbol{F}^{obj} = \{\boldsymbol{f}_1^{obj}, \boldsymbol{f}_2^{obj}, ..., \boldsymbol{f}_m^{obj}\}$, where $m = \{1, 2, ..., M\}$ and $\boldsymbol{f}_m^{obj}$ is the projected d-dimensional feature vectors for m_{th} object. Specifically, the feature vector output of the object detector (from the fc7 layer) is used to encode the appearance feature and the relative bounding box coordinates are employed as the location feature.

Embedding of OCR tokens. For the N OCR tokens extracted by an OCR system, we

construct the embedding for each token containing both its visual and text feature. The visual feature extraction follows the strategy of the above visual object embedding. Additionally, FastText [10] and PHOC features [2] are extracted for each OCR token to represent its textual cues. A rich OCR representation is thus obtained, $\boldsymbol{F}^{ocr} = \{\boldsymbol{f}_1^{ocr}, \boldsymbol{f}_2^{ocr}, ..., \boldsymbol{f}_n^{ocr}\}$, where $n = \{1,2,...,N\}$ and $\boldsymbol{f}_n^{ocr}$ is the projected d-dimensional feature vectors for n_{th} OCR token.

3.2 Multi-modality Fusion

Once the feature embedding representation from individual modality, $\boldsymbol{F}^{ans}$, $\boldsymbol{F}^{obj}$ and $\boldsymbol{F}^{ocr}$ are generated, they are able to dynamically attend to each other from a stack of L transformer layers [43] as shown in Figure 3. The input sequence to the multi-modal transformer is $\boldsymbol{F} = \{\boldsymbol{F}^{ans}, \boldsymbol{F}^{obj}, \boldsymbol{F}^{ocr}\}$. The multi-modal transformer leverages feature embeddings from different modalities and accordingly models interaction among them through the multi-head attention mechanism. From the output of the multi-modal transformer, we extract a sequence of d-dimensional feature vectors for each modality, which is an enriched feature from a joint semantic embedding space.

3.3 Text-aware Visual Question Prediction

With the enriched embedding from the multi-modal transformer, the multi-step decoding module predicts a question to the input answer and iteratively generates the question word by word. At each iterative decoding step, we feed in an embedding of previously predicted words, and then the next output word could be either selected from the fixed frequent word vocabulary or from the extracted OCR tokens. Similar to [22, 49], two special tokens $< begin >$ and $< end >$ are appended to the word vocabulary, where $< begin >$ is used as the input to the first decoding step and $< end >$ indicates the end of the decoding process. Alternatively, the decoding process ends when the maximum number of steps T is reached.

During training, our TAG is supervised with the binary cross-entropy loss applied using the originally annotated QA pairs and adapts to generate novel QA pairs during generation. During the QA pairs generation process, we pass an input answer, each of which is selected from the extracted OCR tokens, into the TAG module and generate the corresponding question accordingly. In this way, the generated QA pairs cover a diverse set of scene text which was not directly exploited in the original annotation set. For answer selection, we perform a simple yet efficient strategy that is feeding the OCR token with the largest bounding box as the answer candidate to the proposed TAG. The intuition behind this design is that the scene text with the largest bounding box region is likely to encode semantically meaningful information for scene text-based understanding and reasoning. Also, scene text with a larger font size has a higher chance to be detected correctly without recognition error in general. As we illustrate in our experiments, our simple design facilitates a better understanding of the visual content and provides promising Text-VQA performance. Note that, more high-quality QA pairs could be continuously augmented with a more sophisticated answer-candidate selection strategy. We leave this direction as future work.

4 Experiments

We evaluate TAG both qualitatively and quantitatively on the TextVQA [41] and the ST-VQA [8] datasets. We first present a brief overview of the datasets and implementation details.

Method	OCR system	Extra Data	Val Acc.	Test Acc.
CRN [32]	Rosetta-en	×	40.39	40.96
LaAP-Net [21]	Rosetta-en	×	40.68	40.54
SMA [16]	SBD-Trans OCR	×	43.74	44.29
SSBaseline [54]	SBD-Trans OCR	×	43.95	44.72
LOGOS [34]	Microsoft-OCR	×	50.79	50.65
M4C† [22]	Microsoft-OCR	×	44.50	44.75
M4C† + TAG	Microsoft-OCR	×	**45.68**	**45.96**
TAP [49]	Microsoft-OCR	×	49.91	49.71
TAP + TAG	Microsoft-OCR	×	**52.54**	**52.57**
LaAP-Net [21]	Rosetta-en	ST-VQA	41.02	41.41
SA-M4C [25]	Google-OCR	ST-VQA	45.40	44.60
SMA [16]	SBD-Trans OCR	ST-VQA	44.58	45.51
SSBaseline [54]	SBD-Trans OCR	ST-VQA	45.53	45.66
LOGOS [34]	Microsoft-OCR	ST-VQA	51.53	51.08
M4C† [22]	Microsoft-OCR	ST-VQA	45.22	-
M4C† + TAG	Microsoft-OCR	ST-VQA	**46.33**	**46.38**
TAP [49]	Microsoft-OCR	ST-VQA	50.57	50.71
TAP + TAG	Microsoft-OCR	ST-VQA	**53.63**	**53.69**

Table 1: **TAG's outperformance on the TextVQA dataset when trained on original and augmented dataset under two settings.** Note that M4C† is the improved version from [49].

Then, we empirically validate the effectiveness of our proposed method by comparing it with the existing Text-VQA approaches. Our framework clearly outperforms previous work by a significant margin on both datasets.

4.1 Datasets and Evaluation Metrics

TextVQA dataset [41] is a widely used benchmark for the Text-VQA task. It consists of 28,408 images sourced from the Open Images dataset [31], with human-annotated questions that require reasoning over text in the images. We follow the standard split on the training, validation and test sets [22, 49]. For each question, the answer prediction is evaluated based on the soft-voting accuracy of 10 human-annotated answers [18, 22, 49].

ST-VQA dataset [8] is another popular dataset for the Text-VQA task. It contains 23,038 images from multiple sources including ICDAR 2013 [27], ICDAR 2015 [28], ImageNet [13], VizWiz [20], IIIT STR [35], Visual Genome [30], and COCO-Text [44]. The standard evaluation protocol on the ST-VQA dataset consists of accuracy and Average Normalized Levenshtein Similarity (ANLS) [8].

4.2 Implementation Details

We use PyTorch to implement our TAG† that is used to augment the initially labeled data. The augmented dataset is used to improve two recent Text-VQA models, M4C [22] and TAP [49]. M4C† is a variant version [49] of M4C, where the detected object labels and scene text tokens are also included in the text encoder, which further improves the performance.

†Our implementation is built upon the codebase: https://github.com/microsoft/TAP.

Figure 4: We visualize the examples of the generated QA pairs (bottom in orange) by 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.

TAG. We project the multi-modality feature embedding to be $d = 768$ channels. We extract the embedding of extended answer words using the same trainable structure as $BERT_{BASE}$ [14]. Specifically, we initialize the weights of the model from the first three layers of $BERT_{BASE}$ and eliminate the separate text transformer. In terms of the object embedding, a Faster R-CNN object detector [38] pre-trained on the Visual Genome dataset [30] is adopted to extract $M = 100$ top-scoring objects on each image and represents each object with its appearance and location features. The Microsoft-OCR system [49] is used to extract OCR tokens per image with each token represented with its appearance, location, FastText [10] and PHOC features [2]. The multi-modality fusion module is a four-layer transformer with 12 attention heads, which has the same hyper-parameters as $BERT_{BASE}$. We use $T = 30$ decoding steps to predict the output question word by word in an auto-regressive manner.
Training parameters. Experiments are conducted on 4 Nvidia P6000 GPUs. We train TAG for 24K iterations with a batch size of 128. We adopt the Adam optimizer [29] with a learning rate of 1e-4 and a staircase learning rate schedule, where we multiply the learning rate by 0.1 at 14K and at 19K iterations. We keep the original parameter settings of downstream Text-VQA models except that we increase their maximum iteration in proportion to the increased size of the augmented data to accommodate the enlarged number of training samples.

4.3 Main Results

TextVQA dataset. To perform a fair comparison with prior work, we conduct experiments in both the constrained setting (top part of Table 1) and the unconstrained setting (bottom part of Table 1) on the TextVQA dataset [41].[‡] The number of our augmented training QA examples for Text-VQA is 69.2K compared with 34.6K for the original one. In the constrained setting (top), our TAG improves the corresponding M4C and TAP baselines by 1.18% and 2.63% on the validation set, respectively. We note that although LOGOS [34] in Table 1 uses an extra grounding dataset with 1.1 million images for pre-training and yet our method performs better. In the unconstrained setting (bottom), TAG further boosts M4C and TAP baselines by 1.11% and 3.06% on the validation set, respectively. On the TextVQA test set, TAG also obtains significant performance gains over existing methods. This validates the effectiveness of TAG. We also visualize the generated QA pairs of our TAG in Figure 4. It shows that our TAG generates meaningful QA pairs that are novel compared to the originally annotated ones.

[‡]The constrained setting means training without extra data and the unconstrained one indicates otherwise.

Method	Extra Data	Val Acc.	Val ANLS	Test ANLS
CRN [32]	×	-	-	0.483
LaAP-Net [21]	×	39.74	0.497	0.485
SMA [16]	×	-	-	0.486
SA-M4C [25]	×	42.23	0.512	0.504
SSBaseline [54]	×	-	-	0.509
LOGOS [34]	×	44.10	0.535	0.522
M4C† [22]	×	42.28	0.517	0.517
M4C† + TAG	×	**44.52**	**0.540**	**0.529**
TAP [49]	×	45.29	0.551	0.543
TAP + TAG	×	**50.18**	**0.595**	**0.586**
SSBaseline [54]	TextVQA	-	-	0.550
LOGOS [34]	TextVQA	48.63	0.581	0.579
M4C† [22]	TextVQA	46.60	0.560	0.552
M4C† + TAG	TextVQA	**48.69**	**0.579**	**0.571**
TAP†† [49]	TextVQA, TextCaps, OCR-CC	50.83	0.598	0.597
TAP + TAG	TextVQA	**53.53**	**0.620**	**0.602**

Table 2: **Our framework outperforms prior work on the ST-VQA dataset.** Note that M4C† is the improved version from [49]. Specifically, our model with TextVQA outperforms the SOTA approach TAP†† [49] that is pre-trained with extra large-scale data from external TextCaps [40] and OCR-CC [49] datasets.

ST-VQA dataset. We also compare our approach with the state-of-the-art (SOTA) methods under both the constrained setting and the unconstrained setting on the ST-VQA dataset [8]. We compute the accuracy and ANLS score as the evaluation metrics. The number of the newly built training QA examples for the ST-VQA task after augmentation is 46.8K compared with 23.4K for the original one. Table 2 suggests that TAG achieves SOTA performance and significantly outperforms the baselines. In particular, TAP [49] achieves 50.83%, and 0.598 in terms of the accuracy and ANLS score on the validation set with additional TextVQA and 1.4 million large-scale pre-training data, while TAG improves these results by a significant 2.70% and 0.022 with only additional TextVQA data. In addition, we submit the prediction results of test set on the ST-VQA test server. The results show that TAG with TAP achieves the SOTA performance with ANLS score of 0.602 on the test set. Without bells and whistles, our approach greatly outperforms the baselines, M4C [22] and TAP [49].

4.4 Ablation Studies

We conduct extensive ablation studies to demonstrate the effectiveness of TAG using TAP [49] under the constrained setting on the TextVQA validation set.

Contribution of each modality in TAG. To understand the contribution of different input modalities to the success of TAG, Table. 3 summarizes the performance of our framework when a certain modality is removed. It suggests that when both the visual objects and OCR tokens modalities are removed, the performance of our TAG decreases by 3.78%. On the other side, when removing the visual objects modality and OCR tokens modality separately, the performance drops by 3.59% and 3.41%, respectively.

Impact of the answer selection strategy. To better explore the performance of our TAG, and

Ans.	Obj.	OCR.	Val Acc.
✓			48.76
✓		✓	48.95
✓	✓		49.13
✓	✓	✓	**52.54**

Table 3: Ablation study of TAG with TAP [49] under constrained setting on TextVQA validation set. "Ans.", "Obj." and "OCR." refer to embedding of answer words, detected objects and OCR tokens, respectively.

Answer Selection	Val Acc.
random	49.26
largest	52.54
top three	52.73
top five	52.19

Table 4: Ablation study of TAG with TAP [49] under constrained setting on TextVQA validation set. *Random* means a random OCR token is selected as the answer input to TAG, while *top three* means the top three largest OCR tokens are selected.

understand how different answer selection strategies would affect the model performance, we design several experiments over the choice of input answer selection strategy. Our method adopts the *largest* OCR word as the answer candidate for TAG. We compare this strategy with other possibilities in Table. 4. The table shows that, if we use a *random* OCR token as the input answer, the performance drops by 3.28%. On the other hand, if we increase the number of answer candidates by including the top-3 largest OCR tokens to augment the labeled data by 3×, the performance boosts additional 0.19% as compared to the *largest* strategy while it introduces 3× training time. To achieve a better balance between training efficiency and accuracy, we consider the OCR token with the *largest* bounding box as our final setting for the input answer to TAG. As we have mentioned previously, more high-quality QA pairs could be continuously augmented with a more sophisticated answer-candidate selection strategy. We leave this direction for future work.

5 Conclusion

We propose a novel architecture TAG, a text-aware visual question-answer (QA) generation method to deal with the sparse annotation of existing Text-VQA datasets. Our approach leverages the rich yet underexplored visual and scene text information and directly enlarges the existing training set by generating high-quality and rich QA pairs without extra labeling cost. Without bells and whistles, experimental results show that our generated QA pairs boost the performance of recent Text-VQA models by a large margin on both TextVQA and ST-VQA datasets.

References

[1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020.

[2] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566, 2014.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[5] 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*, pages 2425–2433, 2015.

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[7] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

[8] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019.

[9] Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16548–16558, June 2022.

[10] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton Van den Hengel, and Qi Wu. Structured multimodal attentions for textvqa. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[17] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12746–12756, 2020.

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[19] Tianrui Guan, Jun Wang, Shiyi Lan, Rohan Chandra, Zuxuan Wu, Larry Davis, and Dinesh Manocha. M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 772–782, 2022.

[20] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

[21] Wei Han, Hantao Huang, and Tao Han. Finding the evidence: Localization-aware answer prediction for text visual question answering. *arXiv preprint arXiv:2010.02582*, 2020.

[22] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002, 2020.

[23] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[24] Kushal Kafle, Mohammed A Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In *INLG*, pages 198–202, 2017.

[25] Yash Kant, Dhruv Batra, Peter Anderson, Alexander Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for textvqa. In *European Conference on Computer Vision*, pages 715–732. Springer, 2020.

[26] Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. Contrast and classify: Training robust vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1604–1613, 2021.

[27] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE, 2013.

[28] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[31] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[32] Fen Liu, Guanghui Xu, Qi Wu, Qing Du, Wei Jia, and Mingkui Tan. Cascade reasoning network for text-based visual question answering. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4060–4069, 2020.

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[34] Xiaopeng Lu, Zhen Fan, Yansen Wang, Jean Oh, and Carolyn P Rosé. Localize, group, and select: Boosting text-vqa by scene text modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2631–2639, 2021.

[35] Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Proceedings of the IEEE international conference on computer vision*, pages 3040–3047, 2013.

[36] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

[37] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. *arXiv preprint arXiv:1909.04696*, 2019.

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[39] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019.

[40] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer, 2020.

[41] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

[42] Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. Semantic equivalent adversarial data augmentation for visual question answering. In *European Conference on Computer Vision*, pages 437–453. Springer, 2020.

[43] 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.

[44] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

[45] Jun Wang. ESSumm: Extractive Speech Summarization from Untranscribed Meeting. In *Proc. Interspeech 2022*, pages 3243–3247, 2022. doi: 10.21437/Interspeech.2022-945.

[46] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *arXiv preprint arXiv:2209.07526*, 2022.

[47] Qingqing Wang, Liqiang Xiao, Yue Lu, Yaohui Jin, and Hao He. Towards reasoning ability in scene text visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2281–2289, 2021.

[48] Zixu Wang, Yishu Miao, and Lucia Specia. Cross-modal generative augmentation for visual question answering. *arXiv preprint arXiv:2105.04780*, 2021.

[49] 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.

[50] Gangyan Zeng, Yuan Zhang, Yu Zhou, and Xiaomeng Yang. Beyond ocr+ vqa: involving ocr into the flow for robust and accurate textvqa. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 376–385, 2021.

[51] Xuanyu Zhang and Qing Yang. Position-augmented transformers with entity-aligned mesh for textvqa. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2519–2528, 2021.

[52] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021.

[53] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.

[54] Qi Zhu, Chenyu Gao, Peng Wang, and Qi Wu. Simple is not easy: A simple strong baseline for textvqa and textcaps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3608–3615, 2021.