Localize, Group, and Select: Boosting Text-VQA by Scene Text Modeling

Xiaopeng Lu,* Zhen Fan,* Yansen Wang,* Jean Oh, Carolyn P. Rosé
Language Technologies Institute, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh PA, 15213
{xiaopen2, zhenfan, yansenwa, jeanoh, cp3a}@andrew.cmu.edu

Abstract

*As an important task in multimodal context understanding, Text-VQA (Visual Question Answering) aims at question answering through reading text information in images. It differentiates from the original VQA task as Text-VQA requires large amounts of scene-text relationship understanding, in addition to the cross-modal grounding capability. In this paper, we propose **Lo**calize, **Gr**oup, and **S**elect (LO-GOS), a novel model which attempts to tackle this problem from multiple aspects. LOGOS leverages two grounding tasks to better localize the key information of the image, utilizes scene text clustering to group individual OCR tokens, and learns to select the best answer from different sources of OCR (Optical Character Recognition) texts. Experiments show that LOGOS outperforms previous state-of-the-art methods on two Text-VQA benchmarks without using additional OCR annotation data. Ablation studies and analysis demonstrate the capability of LOGOS to bridge different modalities and better understand scene text.*

1. Introduction

Text-VQA, which calls for a model to answer questions based on scene text in Visual Question Answering (VQA) settings, has attracted much attention in recent years. The task is set to evaluate a model's capability to detect and understand textual information from images, and to make inference between cross-modal facts. Various high-quality Text-VQA benchmarks have been released in recent years [28, 3, 24, 31], as well as multiple proposed methods to address this task.

Previous works have followed the pipeline of first extracting the embedded scene texts, and then jointly generating the answer based on input question, extracted texts, and visual contents [28, 15, 33]. However, simply selecting tokens from detected scene text is often not enough. Text-VQA differs from other multi-modal language analy-

*Equally contributed

Figure 1: Example of Text-VQA. The question is "What is the title of the book in the middle?".

sis tasks (multi-modal sentiment analysis, image retrieval, etc.) where there is a direct connection between task and textual information, and the text itself is often complete, reliable, and can be emphasized over other modalities [32, 25]. In the Text-VQA scenario, things become much more complicated because: (1) The scene text itself may not contain enough information to fulfill the information need, especially when the given question focuses on certain visual characteristics (e.g. *name written in blue*) or certain objects (e.g. *text written on the boat*). (2) The layout of scene text is noisy in real-life images: a single sentence may be split into different rows in the image, which makes it difficult for the current OCR system to recognize them as a full sentence or paragraph. (3) OCR systems may make detection errors, especially on real-life images from very diverse domains where different OCR systems with different tendencies may excel or struggle, which is a natural bottleneck for this task.

Figure 1 shows a typical example of Text-VQA. In order to answer the question "*What is the title of the book in the middle?*" correctly, the model needs to overcome several

obstacles. It should bridge the question with corresponding textual or visual characteristics (*the book in the middle*), properly group the "lines" of words on the book covers to form book and author names, and avoid possible OCR detection errors and choose the most reasonable and reliable OCR output from multiple sources.

Considering these difficulties, in this paper, we propose a model named LOGOS (Localize, Group, and Select) for better scene text understanding. Concretely, we **localize** the question to its Region of Interest (ROI) by (1) introducing a visual grounding task to connect question text and image regions and (2) connecting regions with text modalities using object tokens as explicit alignments. Also, we **group** the individual text pieces via unsupervised, position-based clustering, and provide such positional information to the model as layout representation. Finally, we utilize OCR systems of the different capability to model words in different granularity, and train LOGOS to dynamically **select** the answer from multiple noisy OCR sources. Experiments on two benchmarks confirm that our model benefits from better modeling the text modality and can outperform the state-of-the-art models without additional OCR data annotation.

The main contributions of this paper are the following:

- A novel Text-VQA model that effectively grounds different modalities and denoise low-quality OCR inputs
- State-of-the-art results over two benchmark Text-VQA datasets
- Detailed analysis on the advantages of LOGOS in cross-modal grounding and scene-text understanding

2. Related Works

The Text-VQA task aims at reading and understanding the text captured in the images to answer visual questions. Although the importance of involving scene texts in visual question answering tasks was originally emphasized by [2], due to the lack of available large-scale datasets, early development of question answering tasks related to embedded text understanding was limited in narrow domains such as bar-charts [14] or diagrams [17, 18]. The first large-scale open-domain dataset of the Text-VQA task, TextVQA, was introduced by [28], followed by several similar works including STVQA [3], OCR-VQA [24] and EST-VQA [31].

Recent studies [28, 12, 15, 33, 7, 8, 22, 11] have proposed several models and network architectures for the Text-VQA task. Introduced together with the TextVQA dataset, the LoRRA [28] model is built upon Pythia [13] with an OCR module added to detect and recognize the scene texts. M4C [12] first models Text-VQA as a multimodal task and uses a multimodal transformer to fuse different features over a joint embedding space. Also, it attaches a pointer network that can dynamically copy words

from OCR systems. SA-M4C [15] added spatial information between objects, OCR, and question tokens to implicitly capture their relationships based on M4C to get further improvement. TAP [33] proposed to pretrain the model on several auxiliary tasks such as masked language modeling (MLM) and relative position prediction (RPP). It also leverages additional large-scale OCR datasets to enhance its ability to capture the contextualized information of scene text.

Although some previous works reported better results obtained by purely changing the OCR system [15, 33], they either don't fully realize the potential of modeling the text modality with information from the visual modality or use expensive large-scale OCR data for pretraining. In this paper, we look deeper into other solutions to ground and refine features in the textual modality within existing datasets to facilitate the understanding of scene text in the multimodal fusion process.

3. LOGOS Model

Figure 2 demonstrates the LOGOS model structure, upon which we attempt to bridge the modalities of the question, image, and OCR text with different approaches. The focus of our model is three-fold:

- Localize ROI by question-visual pretraining and question-OCR text modeling
- Group related scene text tokens by clustering OCR layout information
- Select OCR tokens dynamically from multiple noisy OCR systems

3.1. Model Architecture

Considering Text-VQA as a typical multi-modal task with inputs from several different modalities, our model utilizes the hybrid fusion technique, where the first step is to generate separate unimodal representations of different modalities. We use BERT [6] and Faster R-CNN [26] to encode the text and image respectively. Following previous works [12, 33], we also include the pyramidal histogram of characters (PHOC) representations [1] for the OCR tokens.

To let the model aware of ROI given a certain question, we leverage question-visual pretraining before Text-VQA, and question-OCR modeling during Text-VQA training, with object tokens to help align two modalities, with details shown in Section 3.2.

With the aforementioned inputs, we add grouping information for OCR tokens based on their locations. The details of the clustering algorithm are introduced in Section 3.3.

As shown in Figure 2(b), after uni-modal features are obtained, the representations of the same OCR and object tokens from different modalities are concatenated respectively. We feed the textual representation of the question and the fused representation of detected objects and OCR
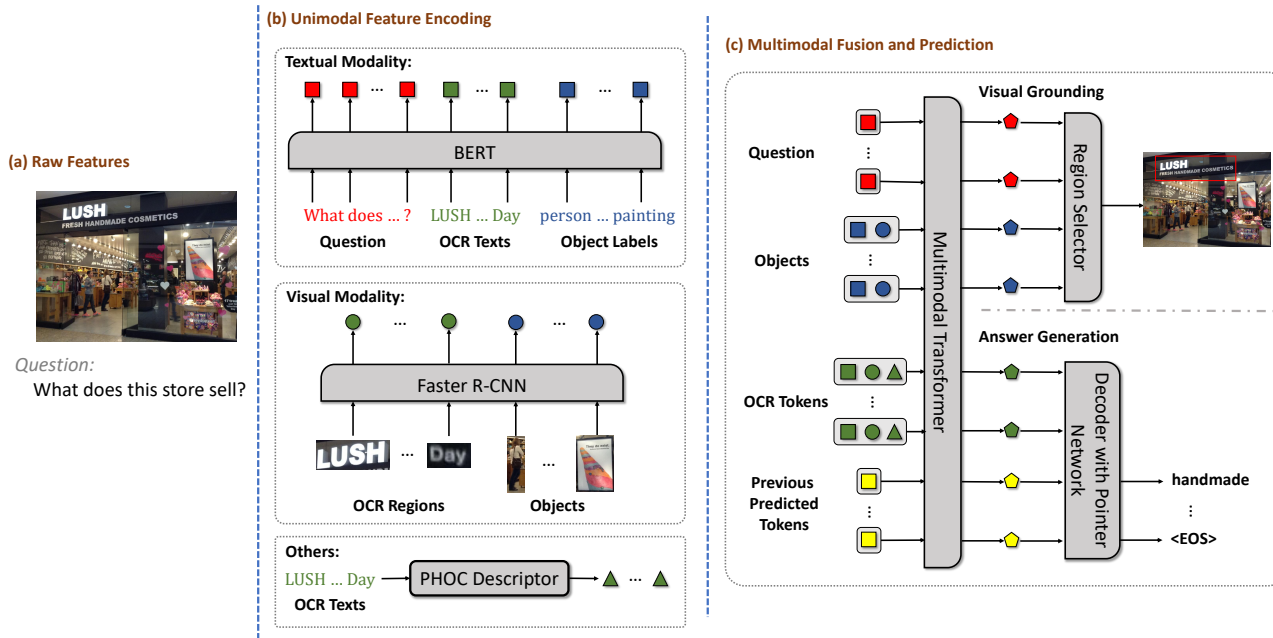
Figure 2: Overview of the LOGOS model. It illustrates how the raw features from the image and the question are processed step by step. Marks of the same **shape** represent features of the same modality, and marks of the same **color** represent features from the same text token or image region.

tokens into a multi-modal transformer. We utilize an answer decoder with a pointer network for output generation and design a selection framework for multiple OCR sources, which we discuss in Section 3.4.

3.2. ROI Localization

To successfully predict the answer given the question information, a robust model should be capable of first pointing out the specific region that is most related to the question, then relying on the OCR text of that *specific* region to generate the output.

It is the most desirable case if we have large amounts of alignment data between question and OCR regions to learn the question-OCR grounding. Since there is no publicly available large-scale question-OCR dataset, we view the question-OCR grounding problem as a two-stage learning problem. Specifically, we first learn the *general* grounding between text and image regions usin existing region-description datasets, for example, Visual Genome [20]. Second, following the intuition from Oscar [21], we use the object tokens as textual "guidance" to help the model learn to ground between question and OCR regions.

3.2.1 Question-Visual Pretraining

To better equip the model with this ability, we select referral expression selection as a pretraining task for question-visual

grounding purposes. Concretely, given an image with a description and n non-overlapping bounding boxes, the model learns to predict which bounding box is the one that is best aligned with the description. As shown in Figure 2, we reuse most parts of the network but add an extra classification layer for candidate prediction. As shown in Figure 2 (c), similar to LOGOS encoding structure, we use question encodings and object encodings as input to the multi-modal transformer module. We train the question-visual pretraining task using cross-entropy loss.

3.2.2 Question-OCR Modeling

Question-Visual Pretraining learns general grounding between question and object regions. In the second stage, we bring object tokens together with object region representations to bridge the gap between question tokens and unseen OCR regions. As shown in Figure 2 (b), we concatenate object label tokens together with question and OCR texts and feed them into a uni-modal BERT encoder. Unimodal BERT encoder models contextualized representation jointly among question, objects, and OCR texts. After passing through the uni-modal BERT layer, we follow the first stage pretraining routine with OCR concatenated representation as an additional input source to the multi-modal transformer encoder. The injection of object information in *both* uni-modal and multi-modal transformers can be seen as

a bridge to help learn the grounding information between question and OCR inputs, in the case where large amounts of question-OCR data are not available.
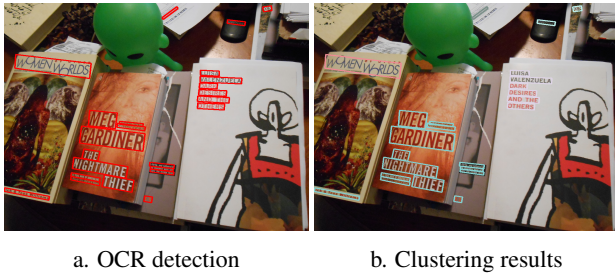

a. OCR detection   b. Clustering results

Figure 3: Example of position-based cluster. Each box represents a line detected by OCR. Same box color in Figure 3b represents same clusters.

3.3. Scene Text Clustering and Modeling

Context understanding has always been important in language processing and question answering. This problem becomes even more vital in Text-VQA, where the evidence of clear separation of text groups lies in the visual modality rather than text. The detection scope of current OCR systems also mainly remains at individual token level or "line" level (grouping tokens aligned closely forming a line).

Figure 3a shows an example of raw detection results by an OCR system. For most text groups such as names and authors of the books, its words are split into multiple lines and thus detected individually. Simply aligning text pieces by their generated sequence, and not providing extra visual evidence will often lead to misordered text and incomplete/incorrect context modeling.

A straightforward idea would be to extract potential text groups based on visual modality evidence, which may serve as additional spatial information along with raw OCR bounding box coordinates, and provide weak evidence for token realignment or token context to form complete sentences in the text domain.

In LOGOS, we perform unsupervised clustering of the OCR text bounding boxes during data preparation. Given an image containing m lines $\{L_1, \ldots, L_m\}$ detected by OCR, with $L_i = t_{i1}t_{i2} \ldots t_{il_i}$ containing l_i tokens, we define the distance between two bounding boxes as the minimum distance between any two points on the two bounding boxes, and perform clustering using DBSCAN [10] on the bounding boxes of all lines $\{B_i\}$. Through clustering, each token t_{ij} is assigned extra spatial hierarchical information in its cluster C_i, line i and token j. We generate a d dimensional sinusoidal embedding as a positional representation for each attribute to form the overall descriptor v_{ij}^{sp}, and

concatenate it as part of the final OCR representation.

$$C_i = \text{DBSCAN}(\{B_1 \ldots B_m\}, \epsilon)[i] \quad (1)$$

$$v_{ij}^{sp} = \text{Pos}(C_i, d)\|\text{Pos}(i, d)\|\text{Pos}(j, d) \quad (2)$$

Where $\text{Pos}(i, d)$ is the d-dim sinusoidal positional embedding at position i. Figure 3b shows a sample result of this clustering approach.

3.4. OCR Source Selection

As aforementioned, current OCR systems are not robust enough to perfectly extract high-quality scene text from images. Different systems also result in different kinds of errors: while a careless OCR system will not be capable of finding all text in the images correctly and make more detection error, a meticulous OCR system may detect too much text from details inside the image, which will increase the difficulty of localizing the question to the correct region. This opens up the possibility of minimizing OCR error and maximizing reasoning effectiveness by combining different OCR sources.

We modify the training and predicting process order to best utilize different OCR systems. Given OCR scene text from k systems, during the decoding stage, k independent answers are generated separately. LOGOS then calculates the confidence score for the i-th answer $S^i = t_1^i t_2^i...t_l^i$ as

$$Score(S^i) = \mathcal{P}(S^i)$$
$$= \mathcal{P}(t_1^i|x) \prod_{j=2}^{l} \mathcal{P}(t_j^i|x, t_1^i, ..., t_{j-1}^i), \quad (3)$$

where $i \in [k]$ and x is the input of other features including visual features and questions. The answer with the highest score is selected as the final answer. During training, all k OCR sources are trained equally, which also serves as extra training data for the learning process.

4. Experiments

4.1. Datasets and Evaluation Metrics

In this paper, we use TextVQA [28] and STVQA [3], two commonly used scene-text VQA datasets as our main testbeds. Visual Genome dataset [20] is used for the referral expression auxiliary training.

TextVQA is the first large-scale Text-VQA dataset with 28,408 images sampled from the Open Image Dataset [19]. A total of 45,336 questions related to the text information in the image were answered by annotators. For each question-image pair, 10 answers are provided by different annotators. The accuracy normalized by weighted voting over the 10 answers is reported on this dataset. Following previous settings [28, 12], we split the dataset into 21,953, 3,166, and 3,289 images respectively for train, validation, and test set.

STVQA is similar to the TextVQA dataset and it contains 23,038 images with 31,791 questions. We follow the setting from M4C [12] and split the dataset into train, validation, and test splits with 17,028, 1,893, and 2,971 images respectively. Compared with TextVQA dataset, the data source of STVQA is more diverse which includes data from Cocotext [29], Visual Genome [20], VizWiz [9], ICDAR [16], ImageNet [5], and IIIT-STR [23] data. We report 2 metrics, accuracy and Average Normalized Levenshtein Similarity(ANLS) on this dataset.

Visual Genome is a multi-modal grounding dataset with 108,077 images [20]. We use this dataset for the purpose of visual grounding joint training. Specifically, region descriptions and their corresponding bounding boxes are used for generating multiple-choice data. In total, we generate 1,216,806 training pairs from the Visual Genome dataset.

4.2. Experiment settings and Training Details

Our model is implemented based on the framework of M4C [12]. For the visual modality, we follow the settings of M4C and use Faster R-CNN [26] fc7 features of 100 top-scoring objects in the image detected by a Faster R-CNN detector pretrained on the Visual Genome dataset. The fc7 weights are fine-tuned during training. For the text modality, we use two OCR systems, Rosetta OCR [4] and Azure OCR [1], to recognize scene text. We use a trainable 3-layer BERT-base encoder [6] for text representation. We follow M4C and include pyramidal histogram of characters (PHOC) representations [1] of OCR text. We perform scene text clustering on normalized bounding boxes using DBSCAN with $\epsilon = 0.02$.

The candidate vocabulary for decoding consists of the top 5,000 frequent words from the answers in the training set, as well as the detected OCR tokens in the current image. The answer is generated with a pointer network [30].

LOGOS contains 116M trainable parameters. During pretraining steps, we first pretrain LOGOS using the question-visual grounding task, then continue to train on the Text-VQA task. We set the batch size to 48 and train for 24,000 iterations for both pretraining and fine-tuning stages. The initial learning rate is set to 1e-4 with a warm-up period of 1000 iterations, and the learning rate decays to 0.1x after 14,000 and 19,000 iterations respectively. We use the checkpoints that achieved the best performance on the validation set for evaluation.

4.3. Experiment Results

Table 1 lists the performance of LOGOS on the TextVQA dataset comparing to other baselines. We find

that LOGOS outperforms all current baselines which do not use extra OCR data.

We have several interesting observations from this table: (1) We see a huge gap between models using another OCR system (M4C+Azure OCR, SA-M4C, TAP, LOGOS) and models using the Rosetta system. The gap illustrates that previous performances on the Text-VQA task are severely limited by the quality of the scene texts detected. This proves the importance of better modeling and refining for the textual modality. (2) LOGOS shares the same idea of introducing auxiliary tasks for pretraining or joint training with TAP, and the results improve similarly although the training tasks are very different. However, LOGOS still outperforms TAP by 0.93% on the TextVQA dataset because of better modeling of text features. (3) It's noteworthy that the current highest score from TAP is obtained by pretraining on other large-scale OCR datasets (OCR-CC) which are designed to better utilize scene text features in multimodal tasks. This idea is similar to ours and our model is fully compatible with pretraining on this dataset. We expect a further improvement of our model when it can get access to more OCR data.

Looking at STVQA results, we can see that training using only STVQA data does not outperform the TAP model [33]. This may due to the fact that STVQA contains more short-length answers, compared with TextVQA where the questions are relatively longer and harder to answer. As the dataset size of STVQA is small and the spatial relationship of this dataset is relatively simple, we observe that LOGOS suffers from overfitting in this dataset. As introducing the joint training with TextVQA dataset, we do see a huge improvement on the STVQA dataset: our model outperforms SNE, previously best STVQA model jointly trained on TextVQA dataset by 2.9 points in ANLS. This reveals that LOGOS can also achieve great performance in STVQA dataset with the additional training data from other Text-VQA datasets.

5. Analysis

5.1. Ablation Studies

Besides the full model, we also ran several experiments on variants of LOGOS to examine the effectiveness of each component. Results are shown in Table 2. We analyze model variant performance from three perspectives.

Effect of ROI Localization Rows #1-#3 show the effectiveness of ROI localization based on question information. We see an improvement when either the question-visual pretraining or the question-OCR modeling is involved, meaning that grounding questions in both modalities are helpful. Specifically, we see better performance ($\sim$ 0.7%) for question-OCR modeling comparing to question-

OCR Data	Method	TextVQA		ST-VQA		
		Acc.(val)	Acc.(test)	Acc.(val)	ANLS(val)	ANLS(test)
External data[†]	TAP [33]	54.71	53.97	50.83	0.598	0.597
Original dataset only	LoRRA [28]	26.56	27.63	-	-	-
	SAN+STR [3]	-	-	-	-	0.135
	M4C [12]	40.55	40.46	38.05	0.472	0.462
	M4C+Azure OCR	45.22	-	42.28	0.517	0.517
	SMA [7]	40.05	40.66	-	-	0.466
	CRN [22]	40.39	40.96	-	-	0.483
	LaAP-Net [11]	40.68	40.54	39.74	0.497	0.485
	SA-M4C [15]	45.40	44.60	42.23	0.512	0.504
	TAP [33]	49.91	49.71	**45.29**	**0.551**	**0.543**
	LOGOS (Ours)	**50.79**	**50.65**	44.10	0.535	0.522
TextVQA+STVQA	LaAP-Net [11]	41.02	41.41	-	-	-
	SNE [34]	45.53	45.66	-	-	0.550
	TAP [33]	50.57	50.71	-	-	-
	LOGOS (Ours)	**51.53**	**51.08**	**48.63**	**0.581**	**0.579**

[†] Uses extra OCR annotations on the OCR-CC dataset, currently not publicly available, with TextVQA and STVQA.

Table 1: Model performance on TextVQA and STVQA.

#	Question-OCR Modeling	Question-Visual Pretraining	Scene Texts Clustering	Source(s)	TextVQA Acc.(val)
1				A	47.76
2		✓		A	48.26
3	✓			A	48.93
4	✓			R	40.84
5	✓			A+R	49.79
6	✓		✓	A+R	50.40
7	✓	✓	✓	A+R	**50.79**

Table 2: Ablation studies of LOGOS on TextVQA dataset; A refers to Azure OCR and R refers to Rosetta OCR.

visual pretraining. This is expected because the question and OCR tokens share the same embedding space in the textual modality, which leads to more efficient relationship learning.

Effect of Using Multiple OCR Systems Compared to simply selecting the highest-quality OCR system, LOGOS utilizes different capabilities of different OCR systems to contribute towards better performance. Rows #3-#5 show the results under the same modeling setting except for the source of OCR tokens. We can see that changing from a low-quality OCR system (Rosetta OCR) to a better one (Azure OCR) results in a huge gain ($\sim 8.1\%$) in terms of accuracy, but using both and selecting with a selector results in further improvement ($\sim 0.9\%$).

In order to verify that the improvement is not only achieved by more diverse training data, we conduct an ex-

Inference source	Accuracy
Rosetta Only	43.26
Azure Only	50.02
Select Better Confidence	51.53
Selection Accuracy	71.96

Table 3: Ablation study on the OCR Source Selection module. The values are obtained from the same model snapshot trained on TextVQA+STVQA.

tra experiment by using both OCR sources during training but limiting the source during inference. The results are shown in Table 3. Among the whole validation dataset of TextVQA, there are 22.4% of data where only one OCR system predicts the right answer, and in 71.96% of these
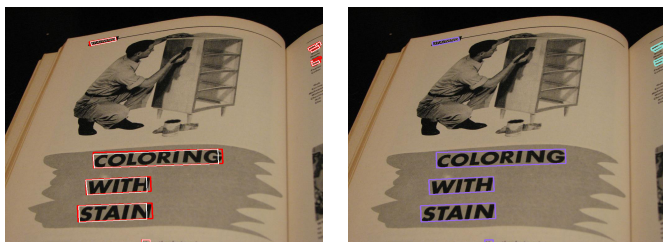
Question: What are they coloring with this instructional book?
LOGOS (w/o Scene Text Clustering): coloring
LOGOS (w. Scene Text Clustering): stain

Question: What is the number of the jersey on the left?
LOGOS (Azure): 17
LOGOS (Source Selection - Azure, Rosetta): 3

Figure 4: Qualitative comparisons between LOGOS and its ablated model predictions. Blue and red answers represent correct and incorrect predictions respectively.

cases, LOGOS correctly selects the OCR source based on confidence.

Effect of Scene Text Clustering By comparing Row #5 and #6, we find that by grouping scene text within a similar area and adding the corresponding position representation, our model achieved an improvement of around 0.6%. Note that all the methods mentioned in this paper are compatible with each other and can be used at the same time, by which we can obtain the state-of-the-art model with results reported in Row #7 compared to models using only TextVQA as OCR data for training.

5.2. Case Studies

In this section, we analyze predictions of different model variants of LOGOS on validation set questions to check module effectiveness. We show two examples in Figure 4.

The case on the left demonstrates the effect of text clustering. The two images show OCR text pieces without and with text clustering respectively. Without the clustering information and with only the bounding box coordinates, the model still struggles to find the connection between individual words to form the phrase "coloring with stain". Adding clustering-based word position information leads to much easier inference for the model to predict the correct answer.

The case on the right further emphasizes the importance of utilizing detection results from multiple OCR sources. When one OCR source (Azure) incorrectly misses the key evidence text (the "3" on the player's jersey), the model will most certainly generate an incorrect answer. After incorporating the source selection module, the model is able to dynamically choose the most reasonable source, based on its understanding of the question and relationship between text and objects. In this case, the model successfully grounds the question to the correct object (the jersey on the left) and answers the correct jersey number 3 instead of 17.

5.3. Error Analysis

In this section, we present analysis of error types to see how our model achieved better accuracy and what limits its further improvement. We randomly sampled 30 negative examples from M4C and another 30 from LOGOS and counted the number of errors caused by bad OCR quality, failure to group scene text, or failure to locate questions in the related area. The proportion of such errors reduced from 80.0% to 53.3%, which proves LOGOS's ability to better locate questions and utilize scene text clusters from multiple OCR systems.

When looking into negative examples from LOGOS, we also notice two typical types of error which we present in Figure 5. In these two cases, the model correctly recognizes and locates the question to the scene text, but still failed to generate the correct answer.

In the first case, the question asks about the size of the TV. Answering such questions requires a QA model to be equipped with proper external knowledge. Here, the model needs to understand measurements and different kinds of notations such as *8K* for resolution and *98"* for size. Other types of external knowledge in similar TextVQA error cases include: recognizing time, identifying dates, and counting and calculation.
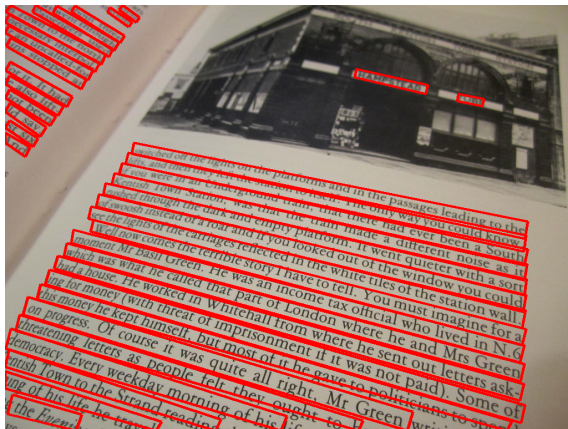
As for the second case, although all the words are detected correctly by the OCR system and grouped together, the model struggles to completely understand the long text. Answering this type of question further requires the model to fully understand the content, which is almost impossible to be achieved without a more specific design, for example, a pipeline with a reading comprehension module.

6. Conclusion and Future Work

In this paper, we propose LOGOS, a novel model that hierarchically groups evidence from both image and text modality, and outputs answers from multiple noisy OCR systems. Results reveal that LOGOS outperforms the state-

data. Detailed analysis shows that LOGOS can not only learn which region to focus on given the question but can also generate coherent answers with correct spatial order from the original image.

Based on the analysis and comparison with other models, there are also some potential areas for future work. We do not use any external OCR-related datasets to enhance the model's ability to model scene text, and we expect a higher score when more OCR data are applied to our model. Our model makes the first attempt to better utilize scene text from images in the multi-modal setting. We observe a huge improvement on Text-VQA performance, and we believe this method can also be applied to similar tasks such as Text-Caption [27] where the understanding of scene text plays an important role.

References

[1] 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.

[2] 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.

[3] 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.

[4] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 71–79, 2018.

[5] 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.

[6] 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.

[7] Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton van den Hengel, and Qi Wu. Structured multimodal attentions for textvqa. *arXiv preprint arXiv:2006.00753*, 2020.

[8] 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.

[9] 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

Question: How big is the TV?
LOGOS: 8k
Ground Truth: 98" / 98 inch

Question: What was Mr. Green's first name?
LOGOS: Hampstead
Ground Truth: Basil

Figure 5: Qualitative examples of LOGOS failed cases. Blue and red answers represent correct and incorrect predictions respectively.

of-the-art models without using additional OCR training

blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

[10] Michael Hahsler, Matthew Piekenbrock, and Derek Doran. dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, 91(1):1–30, 2019.

[11] 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.

[12] 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.

[13] 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.

[14] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.

[15] Yash Kant, Dhruv Batra, Peter Anderson, Alex Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for textvqa. *arXiv preprint arXiv:2007.12146*, 2020.

[16] 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.

[17] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251. Springer, 2016.

[18] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017.

[19] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017.

[20] R Krishna et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arxiv. *arXiv preprint arXiv:1602.07332*, 2016.

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[22] 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.

[23] 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.

[24] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[25] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pre-trained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access, 2020.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[27] 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.

[28] 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.

[29] 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.

[30] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. *arXiv preprint arXiv:1506.03134*, 2015.

[31] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135, 2020.

[32] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223, 2019.

[33] 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. *arXiv preprint arXiv:2012.04638*, 2020.

[34] Qi Zhu, Chenyu Gao, Peng Wang, and Qi Wu. Simple is not easy: A simple strong baseline for textvqa and textcaps. *arXiv preprint arXiv:2012.05153*, 2020.