EasyChair Preprint

№ 10256

Re-thinking Text Clustering for Images with Text

Shwet Kamal Mishra, Soham Joshi and Viswanath Gopalakrishnan

May 24, 2023

Re-thinking Text Clustering for Images with Text

Shwet Kamal Mishra[1[0009−0008−5656−3490]], Soham Joshi[1[0000−0002−8201−2253]], and Viswanath Gopalakrishnan[1[0000−0001−7813−877X]]

International Institute of Information Technology Bangalore, Bengaluru, India

Abstract. Text-VQA refers to the set of problems that reason about the text present in an image to answer specific questions regarding the image content. Previous works in text-VQA have largely followed the common strategy of feeding various input modalities (OCR, Objects, Question) to an attention-based learning framework. Such approaches treat the OCR tokens as independent entities and ignore the fact that these tokens often come correlated in an image representing a larger 'meaningful' entity. The 'meaningful' entity potentially represented by a group of OCR tokens could be primarily discerned by the layout of the text in the image along with the broader context it appears. In the proposed work, we aim to cluster the OCR tokens using a novel spatially-aware and knowledge-enabled clustering technique that uses an external knowledge graph to improve the answer prediction accuracy of the text-VQA problem. Our proposed algorithm is generic enough to be applied to any multi-modal transformer architecture used for text-VQA training. We showcase the objective and subjective effectiveness of the proposed approach by improving the performance of the M4C model on the Text-VQA datasets.

Keywords: Text VQA · Scene Text Clustering · Knowledge Graph.

1 Introduction

Text-VQA plays an integral role in the automatic understanding of images that come along with rich contextual text data. Specific questions regarding the content of text in an image can only be answered with the contextual understanding of the various objects in the image along with the detected text. The success of text-VQA approaches not only relies on proper reasoning regarding the inter-dependency between visual and textual content but also on the correlation between different words present in the textual content.

Previous works in text-VQA have focused on establishing inter-relationships between multiple modalities [1],[12],[22],[18] involving objects in the image, OCR-detected text [10], and questions asked about the textual content. The different modalities are fed as inputs to a multi-modal attention framework involving transformers and learned in an end-to-end fashion with the answer as the ground truth data. While it makes sense to learn the cross-correlation between the question, image content(objects) and the detected OCR tokens using the guidance

from answer ground truths, the correlation between various OCR tokens in an image cannot be learned in a similar way. Though the OCR tokens are detected separately, in many cases they form a group or cluster with a larger context involved. The understanding of this broader group of OCR tokens is imperative to rightly answer many questions involved in a text-VQA task. In this work, we focus on understanding the broader context in which the OCR tokens can be grouped and subsequently feed the grouping information to a transformer-based attentional framework with the aim to improve the accuracy of the text-VQA task.

Fig. 1: Our approach clusters the OCR tokens based on their spatial layout and an external knowledge graph. The clustering information functions as an additional input to a multi-modal transformer framework to improve the accuracy of the text-VQA problem.

The grouping of OCR tokens under the broader context can be better understood by considering the example shown in Figure 1. In Figure 1, the OCR tokens detected are 'Making', 'Visible', 'Mira', 'Schendel', and so on. Though individual OCR tokens 'Mira' and 'Schendel' can occur with independent meanings, in Figure 1, they represent the name of the person 'Mira Schendel'. To correctly answer the question 'Whose name is written on the white booklet?', the training network will need knowledge of the aforementioned grouping. Our work proposes a novel method for OCR token grouping using a joint approach that exploits the spatial layout of the tokens as well as the information available from external knowledge bases. We can easily verify from Figure 1 that the spatial layout of the OCR token indeed holds a strong clue regarding their grouping while the presence of such a meaningful entity could be further established with the help of an external knowledge graph.

The proposed work can be summarized in the following key points:

i) We improve the scene text clustering technique proposed in [26] by considering

the impact of spatial layout in grouping the tokens
ii) We modify the clustering algorithm parameters based on the inputs from external knowledge bases and propose a novel method to ingrain the external knowledge into the clustering technique.
iii) We devise a novel method to incorporate the OCR clustering information into the multi-modal transformer architecture and train it in an end-to-end fashion to showcase improved objective and subjective results in text-VQA problems.

2 Background

2.1 Text Visual Question Answering (Text-VQA)

Text-VQA involves answering questions that require models to explicitly reason about the text present in the image. Multiple datasets [18],[1], [10], [7], [9] have been proposed for the text-VQA task. Subsequently, several methods have also been proposed for this task. Broadly, the popular methods can be categorized into vanilla attention-based models [1], graph-based models [21], and transformer-based models [12], [22]. An example of attention-based models is LoRRA [1] which extends the Pythia framework [19] by including an additional attention mode for OCR tokens to reason over a combined list of vocabulary and detected OCR tokens. Multiple approaches have also incorporated the OCR tokens as a modality in their models [10], [20], and [18]. [21] builds a three-layer multi-modal graph comprising numeric, semantic and visual features and then trains a graph neural network for the Text-VQA task. Recently, transformer-based methods [12], [22] have been widely re-used for text VQA tasks by adapting the framework to accept multiple modalities for OCR tokens, Visual features and Question tokens. We elaborate on [12] (which is our baseline) in the following subsection.

There are other recent works that leverage large pre-trained encoder-decoder language models and Vision Transformers (ViTs) which are topping the leaderboard of the text-VQA tasks [23], [24], [25]. Also, these models use OCR systems (Google-OCR, Azure-OCR, etc.) that are more accurate than the Rosetta-OCR results provided with the Text-VQA dataset [1].We chose M4C as a baseline model and used Rosetta-OCR results [12] to best demonstrate the advantage of the proposed clustering strategy. However, our proposed method is modular and scalable and thus can be plug-and-played with any transformer-based architecture.

2.2 Multimodal Multi-Copy Mesh (M4C)

M4C [12] builds on top of the [13] to create a multimodal transformer module. This multimodal transformer has input modalities namely, question tokens, object embeddings, and OCR tokens. The feature extraction procedure for these three modalities is as follows:

1. Question Words: The question words are encoded using a pre-trained BERT model [14]. The question embedding is fed into the transformer through the question modality.
2. Detected Objects: The objects in the image are detected by passing the image through a Faster R-CNN network [15] to detect the proposals. The object embedding is generated by adding positional information (about the bounding box) about the normalized coordinates of the object, thus making the embedding richer.
3. OCR tokens: The OCR features are extracted through FastText [16], Faster R-CNN detector [15] and Pyramidal Histogram of Characters (PHOC) [17]. These features form parts of the OCR embedding. Additionally, positional information (about the bounding box) is also added to enrich the OCR embedding.

The M4C model projects the feature representations from these three modalities as vectors in a learned common embedding space. The model learns to predict the answer through iterative decoding accompanied by a dynamic pointer network. The work M4C was the first breakthrough in text-VQA which demonstrated the use of multimodal transformer architecture. M4C was benchmarked on text-VQA datasets like Text-VQA [1] and ST-VQA [18] and achieved SOTA results on the same.

2.3 Scene Text Clustering

Scene Text clustering is relatively a new idea in the text-VQA domain. There has been only one such attempt in the past where Lu et. al [26] clustered the tokens based on the bounding box coordinates and passed on that information to a multimodal transformer through positional embeddings. However, this approach does not cluster the tokens at a local level, due to which a larger set of tokens are grouped together. It is also prone to grouping unrelated tokens with significant differences in font size together just because they might be in close vicinity. Apart from this, it uses positional embeddings for token numbers and line numbers as well. In contrast to the clustering approach in [26], we leverage the spatial alignment of the OCR tokens and the related external knowledge to make the clusters more localized and meaningful. Furthermore, we explored the idea of passing the clustering information through a simple mechanism instead of positional encodings.

3 Spatially-Aware and Knowledge-Enabled Clustering

In this paper, we propose an approach that uses features of OCR bounding boxes to cluster tokens together. Our contribution is novel in the following ways:

1. Implemented a localized clustering that works at an entity level unlike the approach proposed in [26].

2. Utilized the spatial layout of OCR tokens by introducing a height penalty parameter in the clustering method.
3. Clustered the OCR tokens to meaningful entities by combining the spatial layout information of tokens with the external knowledge of WikiData.

Thus, our approach is spatially-aware by clustering the tokens based on spatial features and knowledge-enabled by identifying the group of tokens based on their actual meaning and presence in the knowledge graph.
Clustering is done in the pre-processing stage and the outputs of clustering are then fed to a multimodal transformer.

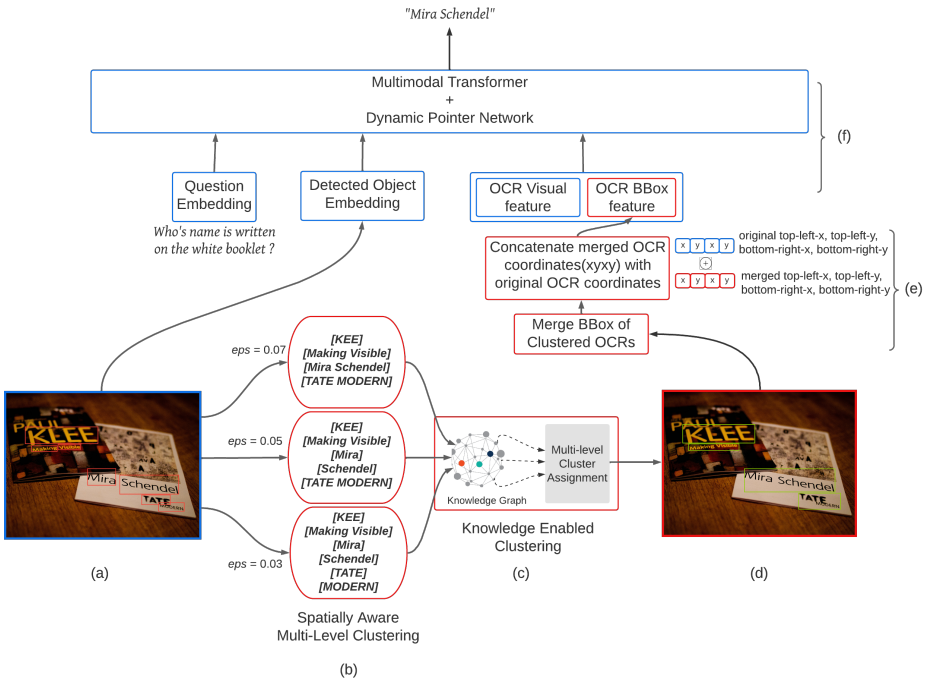


Fig. 2: An overview of our spatially-aware and knowledge-enabled approach. (a) Image with OCR Bounding boxes. (b) The OCR tokens are clustered at different levels based on the *eps* parameter that produces clusters with larger group sizes to smaller group sizes. (c) WikiData Knowledge graph is used to identify real-world entities from the clustered tokens. (d) The clusters are identified based on KG modifications and, (e) bounding boxes of grouped OCR tokens are concatenated with the original bounding boxes of tokens. (f) The concatenated bounding box vector is passed on to a multimodal transformer that eventually uses a Dynamic Pointer Network to produce the output.

3.1 Spatially Aware Clustering

We use the DBSCAN[29] algorithm to cluster bounding boxes in each image during the preprocessing stage. Each bounding box is represented by 17 features, 16 features coming from x, and y coordinates of the top left, top right, bottom left, bottom right, top midpoint, bottom midpoint, left midpoint, and right midpoint points of the bounding box and 17th feature is the height of the box. These features are passed to the DBSCAN algorithm for all the images and the clustering is tuned by the epsilon(*eps*) parameter that specifies how close boxes should be to each other to be considered a part of a cluster. The difference from traditional DBSCAN is that we use custom distance computation for clustering.

To compute the distance we first choose the two nearest points between the boxes using euclidean distance, and then the distance is penalized by the height difference between the two boxes. The idea here is to increase the distance between two boxes for clustering that has significant differences in their height even after being in close vicinity.

$$distance = d + \lambda \times \Delta H \tag{1}$$

Here, d is the Euclidean distance between the two nearest points of bounding boxes, λ is a penalty parameter and ΔH is the height difference between the boxes.
The clustering is tuned in such a way that it focuses on grouping together tokens at a localized level.

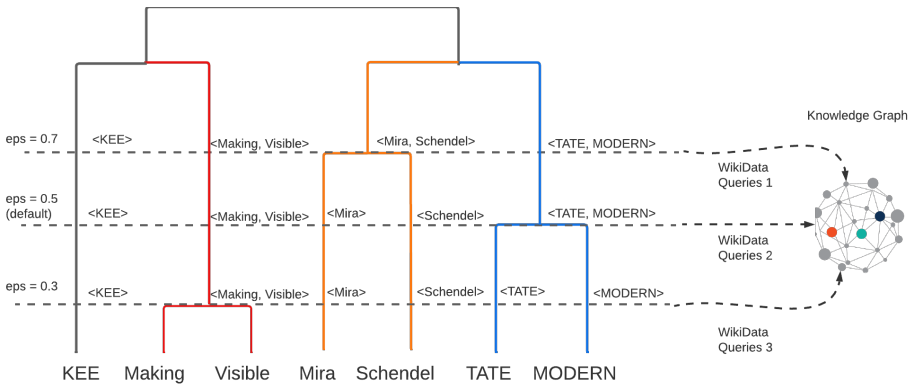


Fig. 3: Multi-level clustering happens at three *eps* values 0.07, 0.05, and, 0.03. Starting with *eps*=0.07, the clustered tokens are concatenated and queried on WikiData. The same is done for other *eps* values and the largest string that finds a match in the WikiData query is retained in the final clustering. Refer to Section 3.2 for more details.

3.2 Knowledge Enabled Clustering

As demonstrated in Figure 2, oftentimes, the text present in the image contains a real-world entity like the author's name, organization name, brand name, etc. Such real-world entities are stored in publicly available knowledge graphs. Our idea is to leverage a knowledge graph - WikiData to cluster real-world entities that are present in the image.

To cluster larger entities and to reduce the number of queries hit on WikiData APIs we take a sequential clustering approach where multi-level spatially aware clustering is done for each image. Initial clustering is done with a larger *eps* value, creating clusters with more tokens. In subsequent clustering, the *eps* value is decreased and clusters get smaller. This is done for *eps* values 0.07, 0.05, and 0.03. The sequence of steps in our proposed clustering is described below:

1. A global clustering variable(*global_cluster*) is initialized with -1 for all the tokens.
2. The clustering result with the same order is picked, and tokens from each cluster are joined into a single string S.
3. S is then queried on WikiData and if S exists as an entity then the tokens present in S are assigned the same cluster in the *global_cluster* and this cluster assignment cannot be changed by any further smaller cluster.
4. The above(3) process continues until the last cluster.
5. Finally, the tokens that still have no cluster assigned are passed through a cluster reassignment method, which simply assigns new cluster ids based on the *eps* 0.05 results.

Refer to Algorithm 1 for the pseudo-code of the approach.

3.3 Feature Engineering of Clustering Annotations

Scene Text Clustering is done during the preprocessing stage where every token is assigned an additional key *cluster_bbox* that stores the bounding box coordinate of the cluster that the token belongs to. During the training, this additional information is concatenated with every token's original bounding box coordinates vector. Thus, making the final OCR bounding vector 8-dimensional

$$\begin{aligned} x_n^{b'} = [&x_{top\_left}, y_{top\_left}, x_{bottom\_right}, y_{bottom\_right}, \\ &x'_{top\_left}, y'_{top\_left}, x'_{bottom\_right}, y'_{bottom\_right}] \end{aligned} \tag{2}$$

This information is further linearly projected (into the same hidden-size space as described in [12]) before adding it to the embedding formed from the FastText, Faster-RCNN and PHOC features. The OCR feature embedding is created as follows (similar to [12]):

$$x_n^{ocr} = LN(W_3 x_n^{ft} + W_4 x_n^{fr} + W_5 x_n^{p}) + LN(W_6 x_n^{b'}) \tag{3}$$

Algorithm 1 Algorithm to Assign Cluster Ids to each token in a single Image

$n \leftarrow$ number of OCR tokens in the image
$token\_list \leftarrow$ list of OCR Tokens
$global\_cluster \leftarrow [-1, -1, ..., -1]_{1 \times n}$
for eps_value `in [0.07, 0.05`$_{default}$`, 0.03]` **do** ▷ Decreasing order of *eps* value to ensure multi-level clustering
  $cluster\_ids$ `= DBSCAN(eps=`eps_value`)`
  for `S in stringify(token_list, cluster_ids)` **do** ▷ Stringify joins tokens in the same cluster and creates a list of Strings
    if S in WikiData **then**
      for $token$ `in S.tokens()` **do**
        $cluster\_id = cluster\_ids[token]$
        if $global\_cluster[token] == -1$ **then**
          $global\_cluster[token] \leftarrow cluster\_id$
        end if
      end for
    end if
  end for
  for `token in token_list` **do**
    if $global\_cluster_{token}$ == -1 **then**
      $global\_cluster[token] \leftarrow token\_cluster\_id_{default}$ ▷ Cluster ids assigned here does not overlap with any existing id in the global cluster
    end if
  end for

where W_3, W_4, W_5 and W_6 are learned projection matrices and $LN()$ is a layer normalization; x_n^{ft}, x_n^{fr}, x_n^p and $x_n^{b'}$ are FastText vector, Faster-RCNN (appearance) feature, Pyramidal Histogram of Character (PHOC) feature and concatenation of location feature and cluster aggregation result respectively.

This embedding is now richer in terms of the positioning of the related "meaningful" entity to which the OCR token belongs.

Another variation experimented with the OCR cluster embedding was feeding it through a linear layer instead of concatenating it with the OCR token embedding. In this approach, we project the OCR cluster Bounding box ($x_n^{b''}$) and the OCR token Bounding Box (x_n^b) differently before adding them as described in the equation below.

$$x_n^{ocr} = LN(W_3 x_n^{ft} + W_4 x_n^{fr} + W_5 x_n^p) + LN(W_6 x_n^b) + LN(W_7 x_n^{b''}) \quad (4)$$

where W_7 is a learned projection matrices and $LN()$ is a layer normalization; $x_n^{b''}$ is cluster aggregation result.

4 Experiments

4.1 Datasets

Text-VQA One of the datasets extensively used for Text Visual Question Answering experiments is Text-VQA [1]. Text-VQA dataset contains images from the Open Images dataset [2] from categories containing text like "billboard", "traffic sign" and "whiteboard". The dataset contains 28,408 images and 45,336 questions asked by (sighted) humans over them. Each question-image pair has 10 ground truth annotations (given by humans). The training set contains 34,602 questions based on 21,953 images whereas the validation set contains 5,000 questions based on 3,166 images.

ST-VQA The ST-VQA dataset [18] contains images from a combination of public datasets used for scene text understanding and general computer vision tasks. The ST-VQA comprises images from six datasets namely: ICDAR 2013 [3] and ICDAR 2015 [4], ImageNet [5], VizWiz [6], IIIT Scene Text Retrieval [7], Visual Genome [8], and COCO-Text [9]. ST-VQA dataset contains a total of 31,791 questions over 23,038 images. The training set contains 26,308 questions based on 19,027 images. We only use the training set for our experiments.

OCR-VQA-200K The dataset OCR-VQA [10] is derived from [11]. This dataset contains cover images of the books including meta-data containing author names, titles and genres. The OCR-VQA dataset comprises 207,572 images and 1,002,146 question-answer pairs. The training set contains approximately 800,000 question-answer pairs whereas the validation set contains 100,000 pairs.

4.2 Implementation Details

Concatenation of OCR Bounding Boxes (Concat-Boxes) The image consists of multiple OCR tokens which are clustered by the Multi-level clustering module (refer to Figure 2) according to the information from the Knowledge Graph. This process is described in greater detail in Section 3. The clustering algorithm identifies the labels of the OCR tokens and their corresponding cluster. Thus, every OCR token forms a part of a larger group of OCR tokens. In this experiment, the OCR bounding box feature is concatenated with the information of the cluster (minimum and maximum boundaries of the box in both dimensions). The aim is to find the tightest bounding box which can cover all the OCR tokens in the cluster. We utilise the annotations of the Rosetta OCR system [28] for our experiments. The resultant bounding box OCR embedding ($x_n^{b'}$) is thus 8-dimensional (4 (token) + 4 (concat box)). This information is further linearly projected (into the same hidden-size space as described in [12]) before adding it to the embedding formed from the FastText, Faster-RCNN and PHOC features. The OCR feature embedding is created as discussed in Subsection 3.3 (similar to [12]).

Following this, we apply a similar training strategy as described by [12] for the transformer module. We conduct two runs with this model, (i)Only Text-VQA training data, and (ii)Text-VQA + ST-VQA training data.

The experiment results are presented in Table 1. The validation set questions were divided into three subsets to evaluate the impact of the clustering strategy: (i) Single-word answers – QA pairs with the answer as a single word, (ii) Multi-word answers – QA pairs with the answer as two or more words, and (iii) Limited OCR tokens – QA pairs where the number of OCR tokens is within the 75th percentile of the overall number of OCR tokens distribution. The third subset was selected to further evaluate the effectiveness of clustering, as clustering is most effective when the number of tokens in the scene text is limited.

This strategy (Concat-Boxes) helps the model increase the single-word accuracy (in the first run with only Text-VQA training data) as compared to the baseline by **0.03%**. Additionally, it also boosted the **overall model accuracy** by nearly **0.4%** in the second run. Moreover, it also pushed the **multi-word** accuracy by **0.72%**, and **0.61%** in **limited OCR tokens** set. The results of the experiments on the dataset OCR-VQA-200K [10] are presented in Table 2.

Linear Projection of the Concatenation of OCR Bounding Boxes (Concat-Boxes Linear Projection) The first experiment involved directly concatenating the OCR cluster Bounding Box with the OCR token Bounding Box. In this approach, we project the OCR cluster Bounding box ($x_n^{b''}$) and the OCR token Bounding Box (x_n^b) differently before adding them as described in Subsection 3.3.

This experiment was designed to investigate whether different projection matrices for OCR tokens and cluster Bounding Box would help the transformer get better information. Similar to the previous experiment, we conduct two runs with this model, (i)Only Text-VQA training data and (ii)Text-VQA + ST-VQA training data.

The results of the experiment are tabulated in Table 1. The results demonstrate that we only improve the single-word accuracy at a marginal cost of overall accuracy, in the first run. In the second run, there is a marginal increment in the single-word and multi-word data subsets.

Thus, we can conclude that the addition of the linear projection of the cluster bounding box coordinates makes the model less attentive to the OCR clusters.

Model	Dataset for training	Validation Accuracy			
		TextVQA (entire)	Single-word	Multi-word	Limited OCR
Baseline	TextVQA	39.65	47.13	34.28	41.52
	TextVQA + STVQA	40.24	**48.36**	34.42	41.82
LOGOS[1] [26]	TextVQA	38.55	46.57	32.81	40.11
	TextVQA + STVQA	39.51	47.6	33.7	41.2
(1) Concat-Boxes	TextVQA	39.32	47.16	33.7	40.92
	TextVQA + STVQA	**40.64**	48.31	**35.14**	**42.43**
(2) Concat-Boxes Linear Projection	TextVQA	39.28	47.94	33.07	40.72
	TextVQA + STVQA	40.19	47.88	34.68	41.32

Table 1: Accuracy scores for experiments 1 and 2. Results in the second row represent the validation accuracy of LOGOS [26]. The improvements over the baseline are shown in bold. The best configuration is Concat-Boxes with training data from TextVQA and ST-VQA. There is a considerable improvement of nearly 0.4% in the overall accuracy, 0.72% in the Multi-word subset, and 0.61% in the limited OCR (as compared to the baseline). As the code for [26] is not public, we created our own implementation of the algorithm.

Model	Dataset for training	Validation Accuracy			
		OCR-VQA-200K (entire)	Single-word	Multi-word	Limited OCR
Baseline	OCR-VQA-200K	63.48	86.80	44.80	63.47
(1) Concat-Boxes	OCR-VQA-200K	**63.61**	**86.87**	**44.98**	**63.53**

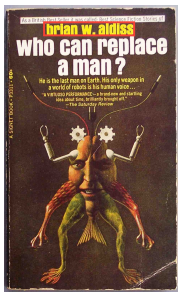
Table 2: Accuracy numbers for Concat-Boxes model on OCR-VQA-200K citeocrvqa dataset. There is an improvement of 0.13% in the overall accuracy. Additionally, the multi-word analysis also shows an improvement of 0.18%.

[1] our implementation of LOGOS[26]

Q: What band is featured on all three items?
Baseline: beatles beates
Ours: the beatles

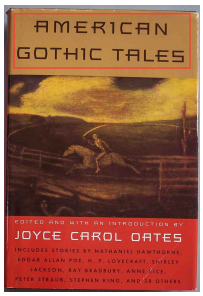
Q: This establishment is called coach and what?
Baseline: coach youngers
Ours: coach & horses

Q: What does the box say at the bottom?
Baseline: christmes
Ours: merry christmas



Q: Who is the author of this book?
Baseline: brian aldiss
Ours: brian w. aldiss

Q: Where is this mug featuring?
Baseline: daytona beach fla
Ours: daytona beach



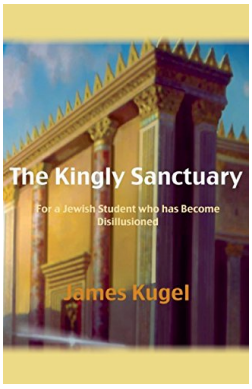
Q: What's the book title?
Baseline: american tales
Ours: american gothic tales

Q: What is this publication called?
Baseline: global book
Ours: the open book

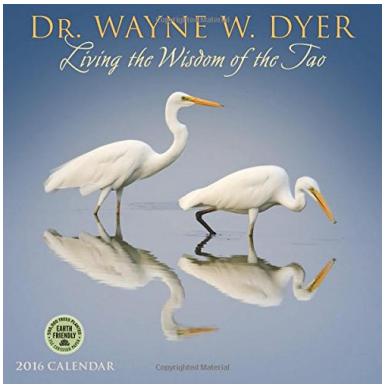
Q: What does the light sign read on the farthest right window?
Baseline: all light
Ours: bud light

Q: What is on the bottom left?
Baseline: ouble bonus
Ours: double bonus

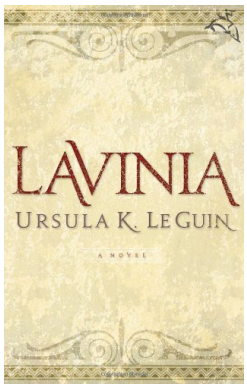
Fig. 4: Qualitative Examples: These examples show the QA pairs (from TextVQA dataset [1]) along with the bounding box highlighting the captured meaningful entity based on the question. The baseline considered was M4C[12] and our model is as shown in Table 1(1).



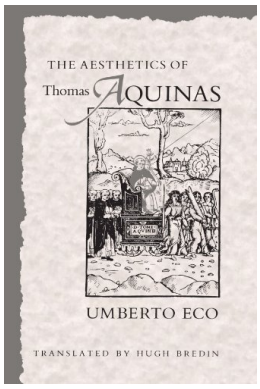
Q: What is the title of this book?
Baseline: the kingly sanctuary for a jewish student has become disillusioned
Ours: the kingly sanctuary



Q: Who is the author of this book?
Baseline: wayne w. dyer
Ours: dr. wayne w. dyer



Q: Who wrote this book?
Baseline: ursula k. guin
Ours: ursula k. le guin



Q: What is the title of this book?
Baseline: the aesthetics of aquinas
Ours: the aesthetics of thomas aquinas

Fig. 5: More Qualitative Examples: These examples demonstrate the subjective improvement on OCR-VQA-200K [10] dataset. The baseline considered was M4C[12] and our model is as shown in Table 2(1).

5 Discussion

1. Text-KVQA dataset [27] contains questions that need to be answered using external knowledge whereas Text-VQA [1] doesn't need the same. The QA pairs in Text-VQA dataset mostly refer to a "meaningful entity" present in the scene text data. As opposed to the knowledge-graph-based works on the Text-KVQA datasets, this work focuses on finding a meaningful entity in the scene text data using external knowledge. To the best of our understanding, our knowledge-graph-based clustering approach is the first such attempt to

answer questions on Text-VQA dataset. The clustering algorithm is transferable to a different domain which can expect the integration of multiple and new knowledge graphs. This emphasizes the scalability and modularity of the algorithm.

2. Although, clustering the meaningful entities in the scene text improves the accuracy for many QA pairs, for some examples as shown in Figure 6, the answers need to be extracted partially from that entity. In those scenarios, clustering information can force the multimodal transformer to output the entire clustered entity as the answer.
3. Aside from the height discrepancy penalty, which aims to address variations in size among OCR tokens, features based on font style can also enhance clustering and open up the potential for further investigation. Another potential avenue for model improvement is the integration of a dynamic vocabulary of OCR clusters, similar to the approach used with OCR tokens in the M4C decoder [12]. This can lead to improved predictions and provide an alternative option to simply choosing a cluster as the answer.

Q: How many sides of this intersection have a stop sign?
Baseline: 4
Ours: 4-way

Q: What city is listed on the bottle?
Baseline: milan
Ours: milan mich.

Q: what does the sign say is "back"?
Baseline: driving
Ours: driving is back.

Fig. 6: Failure cases of the clustering approach.

6 Conclusion

The paper presents a new approach to scene text clustering, based on external knowledge, to address the text-VQA problem. By clustering OCR tokens and prioritizing spatial alignment of scene text, our approach generates more informative queries for the knowledge graph. We also show that the information about grouped tokens can be efficiently transmitted to a multimodal transformer-based framework through box concatenation embeddings. Our objective and subjec-

tive evaluations on the Text-VQA dataset demonstrate the significance of the proposed method, particularly for multi-word answers.

References

1. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D. & Rohrbach, M. Towards VQA Models That Can Read. (arXiv,2019), https://arxiv.org/abs/1904.08920
2. Krasin, I., Duerig, T., Alldrin, N., Veit, A., Abu-El-Haija, S., Belongie, S., Cai, D., Feng, Z., Ferrari, V. & Gomes, V. OpenImages: A public dataset for large-scale multi-label and multi-class image classification.. (2016,1)
3. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Bigorda, L., Mestre, S., Mas, J., Mota, D., Almazàn, J. & Heras, L. ICDAR 2013 Robust Reading Competition. *2013 12th International Conference On Document Analysis And Recognition*. pp. 1484-1493 (2013)
4. Karatzas, D., Gomez, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V., Lu, S., Shafait, F., Uchida, S. & Valveny, E. ICDAR 2015 competition on Robust Reading. (2015,8)
5. Deng, J., Dong, W., Socher, R., Li, L., Li, K. & Fei-Fei, L. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference On Computer Vision And Pattern Recognition*. pp. 248-255 (2009)
6. Gurari, D., Li, Q., Stangl, A., Guo, A., Lin, C., Grauman, K., Luo, J. & Bigham, J. VizWiz Grand Challenge: Answering Visual Questions from Blind People. (arXiv,2018), https://arxiv.org/abs/1802.08218
7. Mishra, A., Alahari, K. & Jawahar, C. Image Retrieval Using Textual Cues. *2013 IEEE International Conference On Computer Vision*. pp. 3040-3047 (2013)
8. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D., Bernstein, M. & Li, F. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. (arXiv,2016), https://arxiv.org/abs/1602.07332
9. Veit, A., Matera, T., Neumann, L., Matas, J. & Belongie, S. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. (arXiv,2016), https://arxiv.org/abs/1601.07140
10. Mishra, A., Shekhar, S., Singh, A. & Chakraborty, A. OCR-VQA: Visual Question Answering by Reading Text in Images. *ICDAR*. (2019)
11. Iwana, B., Rizvi, S., Ahmed, S., Dengel, A. & Uchida, S. Judging a Book By its Cover. (arXiv,2016), https://arxiv.org/abs/1610.09204
12. Hu, R., Singh, A., Darrell, T. & Rohrbach, M. Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA. (arXiv,2019), https://arxiv.org/abs/1911.06258
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. & Polosukhin, I. Attention Is All You Need. (arXiv,2017), https://arxiv.org/abs/1706.03762
14. Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (arXiv,2018), https://arxiv.org/abs/1810.04805
15. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. (arXiv,2015), https://arxiv.org/abs/1506.01497

16. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. (arXiv,2016), https://arxiv.org/abs/1607.04606
17. Almazán, J., Gordo, A., Fornés, A. & Valveny, E. Word Spotting and Recognition with Embedded Attributes. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **36**, 2552-2566 (2014)
18. Biten, A., Tito, R., Mafla, A., Gomez, L., Rusiñol, M., Valveny, E., Jawahar, C. & Karatzas, D. Scene Text Visual Question Answering. (arXiv,2019), https://arxiv.org/abs/1905.13648
19. Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D. & Parikh, D. Pythia v0.1: the Winning Entry to the VQA Challenge 2018. (arXiv,2018), https://arxiv.org/abs/1807.09956
20. Biten, A., Pérez Tito, R., Mafla, A., Gomez, L., Rusinol, M., Mathew, M., Jawahar, C., Valveny, E. & Karatzas, D. ICDAR 2019 Competition on Scene Text Visual Question Answering. (2019,9)
21. Gao, D., Li, K., Wang, R., Shan, S. & Chen, X. Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text. (arXiv,2020), https://arxiv.org/abs/2003.13962
22. Kant, Y., Batra, D., Anderson, P., Schwing, A., Parikh, D., Lu, J. & Agrawal, H. Spatially Aware Multimodal Transformers for TextVQA. (arXiv,2020), https://arxiv.org/abs/2007.12146
23. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B., Riquelme, C., Steiner, A., Angelova, A., Zhai, X., Houlsby, N. & Soricut, R. PaLI: A Jointly-Scaled Multilingual Language-Image Model. (arXiv,2022), https://arxiv.org/abs/2209.06794
24. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C. & Wang, L. GIT: A Generative Image-to-text Transformer for Vision and Language. (arXiv,2022), https://arxiv.org/abs/2205.14100
25. Kil, J., Changpinyo, S., Chen, X., Hu, H., Goodman, S., Chao, W. & Soricut, R. PreSTU: Pre-Training for Scene-Text Understanding. (arXiv,2022), https://arxiv.org/abs/2209.05534
26. Lu, X., Fan, Z., Wang, Y., Oh, J. & Rose, C. Localize, group, and select: Boosting text-VQA by scene text modeling. (2021,8), https://arxiv.org/abs/2108.08965
27. Singh, A., Mishra, A., Shekhar, S. & Chakraborty, A. From Strings to Things: Knowledge-enabled VQA Model that can Read and Reason. *ICCV*. (2019)
28. Borisyuk, F., Gordo, A. & Sivakumar, V. Rosetta: Large scale system for text detection and recognition in images. *CoRR*. **abs/1910.05085** (2019), http://arxiv.org/abs/1910.05085
29. Ester, M., Kriegel, H., Sander, J. & Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Knowledge Discovery And Data Mining*. (1996)