TextCaps: a Dataset for Image Captioning with Reading Comprehension
(Supplementary Material)

We include the following material in our supplemental material:

Section A. Analysis of the influence of GT-OCR on M4C-Captioner model.
Section B. Precision and recall of Rosetta OCR tokens on TextCaps.
Section C. We show the evaluation on the COCO dataset: Our qualitative (Figure C.1) and quantitative (Table C.1) results show that COCO references captions rarely involve reading comprehension, indicating that COCO captions are not a good dataset for training or evaluating this task.
Section D. Qualitative illustration of frequent words in TextCaps images and captions.
Section E. Comparison of TextCaps Test and Validation sets.
Section F. Data collection User Interface.
Figure F.1. Additional examples of M4C-Captioner predictions.

A Analysis of the influence of GT-OCR on M4C-Captioner model

In this section, we provide additional analysis on ground-truth OCRs (GT-OCRs). So far we collected GT-OCR annotations for around 96% on the training, 96% on the validation, and 92% on the test set, we excluded OCR annotation of non-Latin/non-English characters. In Table 1 in the main paper, "M4C-Captioner (w/ GT OCRs)" (line 14 and 18) is evaluated on all TextCaps validation and test set images respectively, where an empty OCR list is used as inputs to the model on those images without GT-OCR annotations. Here, we also specifically compare the methods on the subsets of TextCaps validation and test sets, excluding those images with empty GT-OCR annotations. The results are shown in Table A.1, where ground truth OCR tokens improve the quality of generated predictions significantly.

The analysis of predictions from M4C-Captioner model with automatically extracted and ground-truth OCR tokens (trained and evaluated) shows that vocabulary size of all tokens used in predictions and OCR tokens in particular does not change significantly (Table A.2). Although, the quality of OCR tokens used increase, as indicated by the precision metric. Precision is calculated as average ratio of OCR tokens predicted by the model which match OCR tokens used by annotators from total number of OCR predicted in each sentence.

B Rosetta OCR performance analysis

In our experiments, we use Rosetta [1] to extract OCR tokens from an image. The English-only version of Rosetta is used, which is referred to as **Rosetta-en**

Table A.1: **Effect of using GT-OCR on performance of M4C-Captioner on TextCaps dataset.** Evaluated on a subsets of images with GT-OCR annotations (96% for the validation set, 92% for the test set).

#	Method	TextCaps validation set metrics				
		B-4	M	R	S	C
1	M4C-Captioner	23.0	21.9	46.1	15.4	88.7
2	M4C-Captioner (evaluated w/ GT OCRs)	24.4	22.8	47.0	16.2	99.4
3	M4C-Captioner (trained and evaluated w/ GT OCRs)	26.3	23.3	48.0	16.4	107.2
#	Method	TextCaps test set metrics				
		B-4	M	R	S	C
4	M4C-Captioner	19.0	19.7	43.2	12.7	80.7
5	M4C-Captioner (evaluated w/ GT OCRs)	20.4	20.8	44.5	13.7	95.1
6	M4C-Captioner (trained and evaluated w/ GT OCRs)	22.2	21.6	45.8	14.0	103.5
7	Human	24.4	26.1	46.9	18.8	125.1

B-4: BLEU-4; M: METEOR; R: ROUGE_L; S: SPICE; C: CIDEr

Table A.2: **Statistics of predicted sentences with automatic OCR compared to GT-OCR**

Method	vocab size			OCR
	Total	OCR	VOCAB	token precision
M4C-Captioner	3287	2957	545	0.62
M4C-Captioner (trained&evaluated w/ GT OCRs)	3391	3106	491	0.78

in [2]. To measure the performance of the Rosetta OCR system on TextCaps, we evaluated the precision and recall of OCR tokens against the human-annotated text (ground-truth OCRs) over the validation and test set images, following the ICDAR-13 evaluation protocol for end-to-end text recognition [3]. On the validation set images, the Rosetta OCR tokens have a precision of 56.50, a recall of 37.15, and an F-1 score of 44.83. On the test set images, the Rosetta OCR tokens have a precision of 53.60, a recall of 36.92, and an F-1 score of 43.72.

C Automatic evaluation on COCO captioning

Table C.1 shows the automatic evaluation metrics of the M4C-Captioner model on the COCO dataset. Here, the model trained on COCO + TextCaps[1] has lower evaluation scores than the same model trained only on COCO. We also experiment with different sampling ratios between COCO captions and TextCaps captions, and observe that higher TextCaps ratio (sampling TextCaps captions more frequently) leads to better qualitative results where more OCR tokens

[1] When training on COCO + TextCaps in this setting, we sample TextCaps captions more frequently than COCO captions to encourage learning text reading.

are described in the generated captions, but worse CIDEr scores on the COCO validation set. We inspect and find that this is mainly because the human captions in the COCO dataset rarely involve reading comprehension. For example, in Figure C.1, we see that the predicted captions from M4C-Captioner trained on COCO + TextCaps has noticeably lower CIDEr scores, although it learns to read and copy relevant text from the image.

M4C-Captioner (trained on COCO + TextCaps):
a large [china] airlines passenger jet is parked on the tarmac (CIDEr: 80.7)
M4C-Captioner (trained on COCO):
a large passenger jet is sitting on top of an airport tarmac (CIDEr: 175.1)
human captions:
china airlines plain on the ground at an airport with baggage cars nearby.
a large passenger jet sitting on top of an airport runway.
a large commercial plane with a flower on the tail
a plane parked on the runway with luggage carts parked next to it
a cargo air plane is parked on the runway

M4C-Captioner (trained on COCO + TextCaps):
a red light at the intersection of [shoreline] [blvd] (CIDEr: 16.3)
M4C-Captioner (trained on COCO):
a traffic light with a sky background (CIDEr: 86.1)
human captions:
a blue sky with puffy white clouds and the top of a stop light.
a view of the clouds beyond the stoplight.
a traffic light in front of a cloudy blue sky.
cloudy sky with a street light set to stop.
a yellow streetlight beneath a sky full of clouds.

Fig. C.1: Predicted and human captions on COCO validation set (Karpathy split), where words copied from OCR tokens are taken in square brackets. As human captions in COCO rarely describe text in the image, generated captions that mention text often have lower CIDEr scores.

Table C.1: Automatic evaluation metrics of the M4C-Captioner model on the COCO captioning validation set (Karpathy split). Here, training on TextCaps leads to lower metrics on COCO. This is mainly because the human captions in the COCO dataset do not involve reading comprehension in addition to the domain shift between the two datasets. See Sec. C and Figure C.1 for details.

#	Method	Trained on	B-4	M	R	S	C
1	M4C-Captioner	COCO	**34.3**	**27.5**	**56.2**	**20.6**	**112.2**
2	M4C-Captioner	COCO+TextCaps	27.1	24.1	51.6	17.4	87.5

B: BLEU-4; M: METEOR; R: ROUGE_L; S: SPICE; C: CIDEr

D Illustration of most frequent words in TextCaps

We visualize word clouds for the text tokens in TextCaps captions in Fig. C.2. In the left word cloud, it can be seen that OCR tokens copied from the image to the caption with high frequency mainly consist of brand names and other words which can be found on the products and their labels (*'samsung'*, *'nokia'*, *'colgate'*, *'ale'*). In the right word cloud, when all the words are taken into account, we can observe great use of words like *'sign'*, *'says'*, *'written'* which annotators used to incorporate text tokens into their captions.

E Comparison of TextCaps Test and Validation sets

We notice a difference in performance on our validation and test set in Table 1 of the main paper, specifically, the performance on the validation set is always higher. In this section, we discuss the similarities and differences of the two sets to understand the performance difference.

First, the images for TextCaps 'training' and 'validation' sets are from the OpenImages [4] 'training' set, while TextCaps 'test' images are from the Open-Images 'test' set. We observe that the image-labels of OpenImages training and test sets have slightly different distributions and categories [5]. As our training and validation set are both from the same image distribution (as we follow TextVQA's split [6]), it is likely that models trained on the training set better fit the validation set than the test set. This is partially confirmed by evaluating a model trained only on COCO captions on TextCaps validation and test set. Here, the performance difference is smaller, for example for the BUTD model the CIDEr score drops by 4% (relative) from validation to test set when trained on COCO, but 20% when trained on TextCaps.

Second, although the captions for training, validation, and test were collected jointly, the different image distributions might affect the captions. For this, we further compare their statistics. In particular, we observe from Figure E.1 that both images and captions of the validation set have a larger number of OCR

Fig. C.2: **Wordcloud visualizations** of most frequent OCR tokens (left) and all words (right) in TextCaps captions.

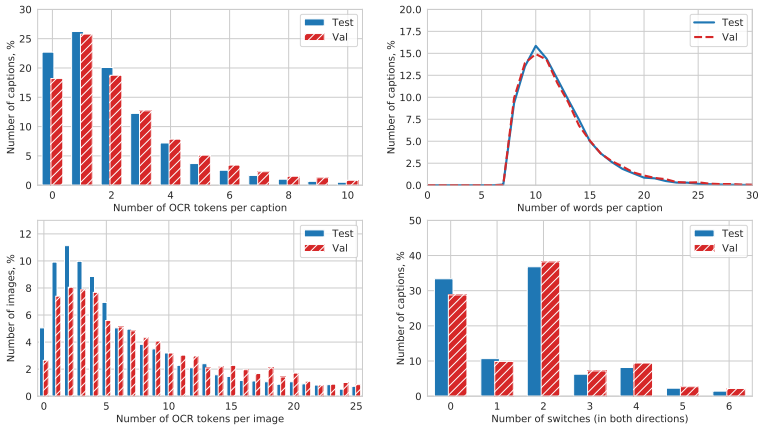


Fig. E.1: **Statistics of TextCaps Test and Validation sets.**

tokens on average (note that all these statistics are based on automatically extracted OCR tokens). This also causes a larger number of switches between OCR and vocabulary required in the validation set. On the other hand, in the test set we observe more captions without any copied OCR tokens, which could suggest more paraphrasing, reasoning, and re-formulation of the OCR tokens in this set. The distribution of captions' length is almost the same for both sets.

Third, we evaluate and compare the automatic metrics on human-written captions between the test and validation sets. Since there are only 5 human captions (instead of 6) collected on the validation set, we perform a similar leave-one-out evaluation as mentioned in Sec. 4.1 but using only 5 human captions per image (evaluating 1 human caption over the remaining 4 and averaging over the 5 runs). The results are shown in Table E.1, where the BLEU-4, METEOR, ROUGE_L, and CIDEr metrics are higher on the test set than on the validation set. This is a bit surprising, but also indicates, that there is slight domain shift between validation and test set, which humans are not affected by, rather than that the test set is more difficult in itself as the models' performance drop might suggest.

F Data collection User Interface

TextCaps was collected using the interfaces presented in Figures F.2-F.5. Before starting, users were presented with a list of detailed instructions (Fig. F.2 and Fig. F.4 for annotation and evaluation, respectively). The main interface window includes a panel with the same list of instructions on the left, a panel with an image in the center, and short instruction followed by the answer field on the right

Table E.1: **Comparison of automatic metrics on human captions between TextCaps test and validation set**, using 5 human captions per image (evaluating 1 human caption over the remaining 4 and averaging over the 5 runs).

#	Method	B-4	M	R	S	C
1	Human captions on the TextCaps validation set	22.1	24.8	44.6	20.3	118.0
2	Human captions on the TextCaps test set	22.6	25.4	45.5	20.3	127.9

B-4: BLEU-4; M: METEOR; R: ROUGE_L; S: SPICE; C: CIDEr

(Fig. F.3, F.5). It is worth noting that in the case of small or hardly-readable text users had options either to open the image in a full size in a new window or to use interactive magnifier lens with 3x zoom. Users annotated images in mini-batches of 5, and evaluated captions in mini-batches of 10. There was no restriction by time (except of very extreme limit of 15 minutes per mini-batch).

References

1. Borisyuk, F., Gordo, A., Sivakumar, V.: Rosetta: Large scale system for text detection and recognition in images. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 71–79. ACM (2018) 1
2. Hu, R., Singh, A., Darrell, T., Rohrbach, M.: Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. arXiv preprint arXiv:1911.06258 (2019) 2
3. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: Icdar 2013 robust reading competition. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1484–1493. IEEE (2013) 2
4. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages (2017) 4
5. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. (2017), https://storage.googleapis.com/openimages/web/factsfigures_v4.html 4
6. Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019) 4

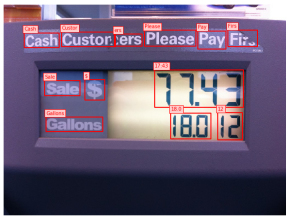
(a) a framed picture with the year [2012] on it

(b) a pair of [merrell] brand products are on a table

(c) a bottle of [deluse] sits on a table next to a small small small plastic bottle

(d) a woman is looking at a screen that says [£20] on it

(e) a trash can with a sticker on it that says [seniors] [only]



(f) a sign that says ' [cash] [custor] [pay] [firs] ' on it

(g) a plane with the number [202] on the side of it

(h) a red telephone booth with a red telephone booth

(i) a restaurant with a red sign that says [bar] [bar]

(j) a man is holding a box that says ' i ' m ' on it

(k) a black box with the word [bizhub] on it

(l) a bottle of wine with the word [chenet] on the label

Fig. F.1: **Additional examples of M4C-Captioner predictions on the test set.** Square brackets denote tokens which model selected from OCR tokens while others are from vocabulary.

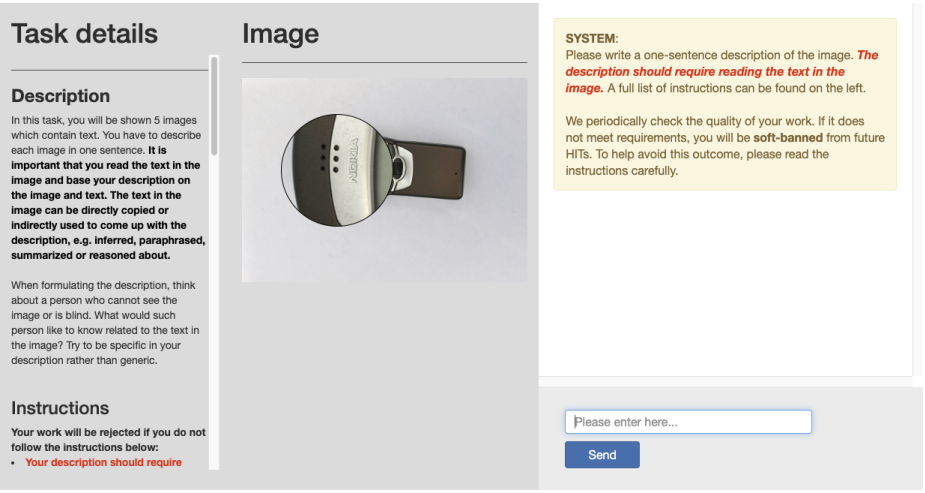


Fig. F.2: **The main interface window for the annotation stage of the data collection.** Detailed instructions are shown in the next Figure. The circle is an interactive zoom tool which users can move with the mouse cursor like a magnifier lens.

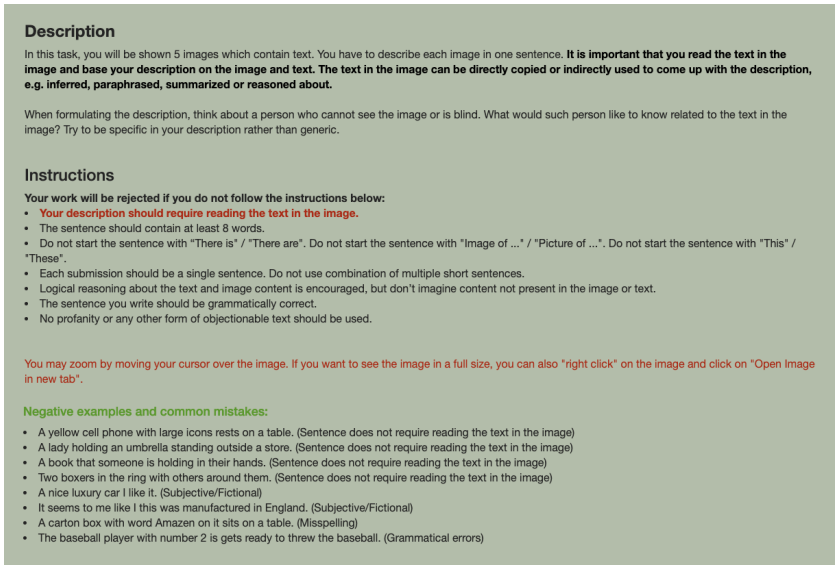


Fig. F.3: **Instructions for the annotation stage of data collection.** First time users saw the instructions before starting the task, after which they could find it on the left panel of our main task interface.

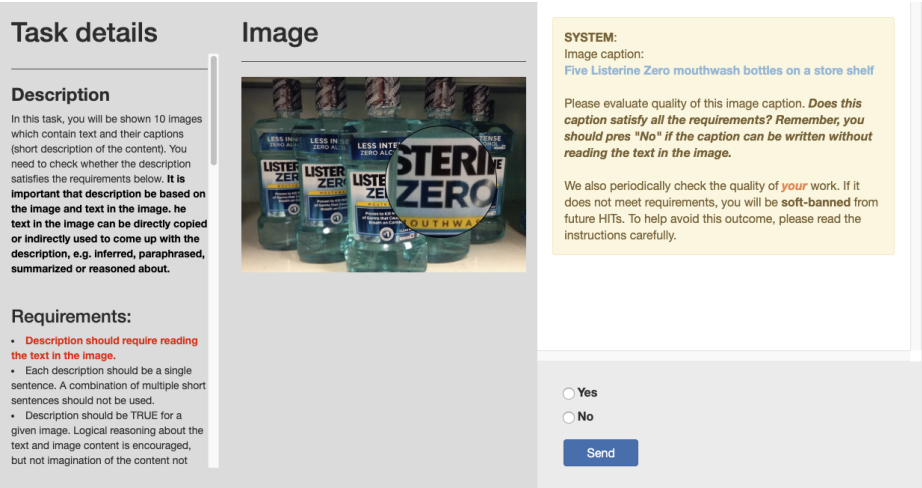


Fig. F.4: **The main interface window for the evaluation stage of the data collection.** Detailed instructions are shown in the next Figure. The circle is an interactive zoom tool which users can move with the mouse cursor like a magnifier lens.

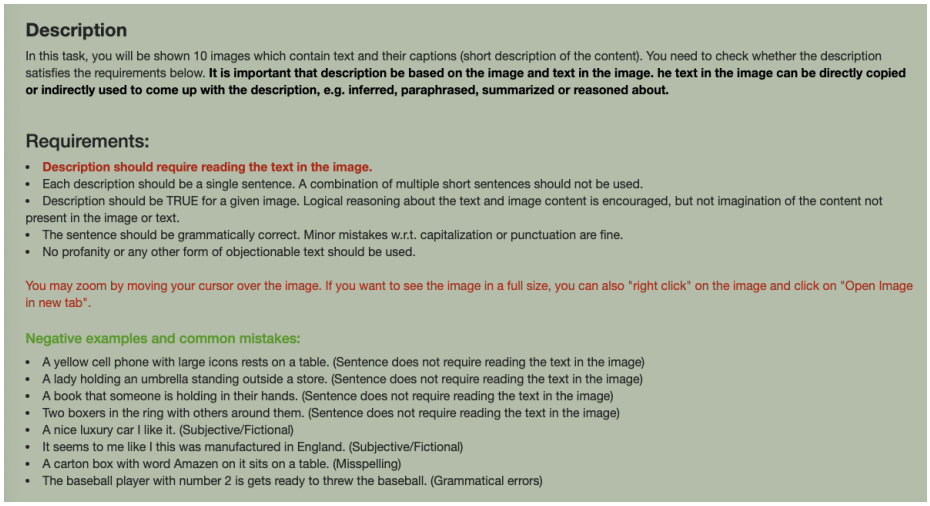


Fig. F.5: **Instructions for the evaluation stage of data collection.** First time users saw the instructions before starting the task, after which they could find it on the left panel of our main task interface.