TextCaps: a Dataset for Image Captioning with Reading Comprehension

Oleksii Sidorov[1] Ronghang Hu[1,2] Marcus Rohrbach[1] Amanpreet Singh[1]
[1]Facebook AI Research (FAIR) [2]University of California, Berkeley
oleksiis@fb.com, ronghang@eecs.berkeley.edu, mrf@fb.com, asg@fb.com

1. Introduction

When trying to understand man-made environments, it is not only important to recognize objects but it is frequently critical to read associated text and comprehend it in the context to the visual scene. Knowing there is "a red sign" is not sufficient to understand that one is at "Mornington Crescent" Station (see Fig. 1(a)), or knowing that an old artifact is next to a ruler is not enough to know that it is "around 40 mm wide" (Fig. 1(c)). In addition, text out of context (*e.g. '5:43p'*) may be of little help, whereas scene description (*e.g.* 'shown on a departure tableau') makes it substantially more meaningful.

In recent years, with the availability of large labelled corpora, progress in image captioning has seen steady increase in performance and quality [4, 1, 5, 7] and reading scene text (OCR) has matured [3, 8, 11]. However, while OCR only focuses on written text, state-of-the-art image captioning methods focus only on the visual objects when generating captions and fail to recognize and reason about the text in the scene. For example, the predictions in Fig. 1 clearly show an inability of current state-of-the-art image captioning methods to read and comprehend text present in images. Incorporating OCR tokens into a sentence is a challenging task, as unlike conventional vocabulary tokens which depend on the text before them and therefore can be inferred, OCR tokens often can not be predicted from the context and therefore represent independent entities. Predicting a token from vocabulary and selecting an OCR token from the scene are two rather different tasks which have to be seamlessly combined to tackle this task.

To study the novel task of image captioning with reading comprehension, we thus believe it is crucial to build a dataset containing captions which require reading and reasoning about the text in images. We find the COCO Captioning dataset [4] not to be suitable as only an estimated 2.7% of its captions mention OCR tokens present in the image.Meanwhile, in Visual Question Answering, multiple datasets [2, 9, 10] were recently introduced which focus on text-based visual question answering. However, the answers are typically short (1-2 words) and do not require switching between OCR and vocabulary to build a complete

Figure 1: **Existing captioning models cannot read!** The *image captioning with reading comprehension* task using data from our TextCaps dataset and BUTD model [1] trained on it.

sentence. These differences in task and dataset do not allow training models to generate long sentences. Furthermore and importantly, we require a dataset with human collected reference sentences to validate and test captioning models for *reading comprehension*.

Consequently, in this work, we contribute the following:

- For our novel task *image captioning with reading comprehension*, we collect a new dataset, **TextCaps**, which **contains 142,040 captions** on 28,408 images and requires models to read and reason about text in the image to generate coherent descriptions.
- We analyse our dataset, and find it has **several new technical challenges for captioning**, including the ability to switch multiple times between OCR tokens and vocabulary, zero-shot OCR tokens, as well as paraphrasing and inference about OCR tokens.
- Our evaluation shows that **standard captioning models fail on this new task**, while the state-of-the-art TextVQA [10] model, M4C [6], when trained with our dataset TextCaps, gets encouraging results.
- We conduct **human evaluations** on model predictions which show that there is a **significant gap between the best model and humans**, indicating an exciting avenue of future image captioning research.

2. TextCaps Dataset

We collect TextCaps with the goal of studying the novel task of *image captioning with reading comprehension*. Our

a

the numbers 18 and 17 on a scoreboard
the number 17 is on the scoreboard with the word rice on it
The scoreboard of a football game showes that **Rice is winning.**
The word "RICE" is displayed on the scoreboard.
A score board shows Rice with 18 points vs. ECU with 17 points.

b

the price of 17.88 that is above a lady
A Walmart sign that says Rollback $17.88 is above a shelf of weight loss products.
A display at Walmart **for a special price** on Hydroxycut.
Box of Hydroxycut **on sale** for only 17.88 at a store.
walmart has hydroxycut **for sale** for 17.88 **instead of** 19.88

c

A white Samsung smartphone shows the time is 11:19.
top part of samsung phone at 11:19 on December 30
A close up of the top half of a Samsung cell phone.
A samsung brand phone shows the current time is 11:19.
The top half of a Samsung cellphone showing the time, date and weather conditions.

Figure 2: **Illustration of TextCaps captions.** The bold font highlights instances which do not copy the text directly but require paraphrasing or some inference beyond copying. Underlined font highlights copied text tokens.

dataset allows us to test captioning models' reading comprehension ability and we hope it will also enable us to teach image captioning models how "to read", *i.e.*, allow us to design and train image captioning algorithms which are able to process and include information from the text in the image. The dataset is publicly available at textvqa.org/textcaps.

2.1. Dataset collection

With the goal of having a diverse set of images, we rely on images from Open Images v3 dataset . We use a subset of images which contain text. Specifically, we use the subset of images present in the TextVQA dataset [10].The images were annotated by human annotators in two stages:

Annotators were asked to describe an image in one sentence which would require reading the text in the image.[1]

Evaluators were asked to vote whether the caption written in the first step satisfies the requirements.The majority of 5 votes was used to filter captions of low quality.

Five independent captions were collected for each image. An additional 6th caption was collected for the test set only to estimate human performance on the dataset. In total, we collected 145,329 captions for 28,408 images. We follow the same image splits as TextVQA for training (21,953), validation (3,166), and test (3,289) sets.

2.2. Dataset analysis

Examples of our collected dataset in Fig. 2 demonstrate that our image captions combine the textual information present in the image with its natural language scene description. We asked the annotators to read and use text in the images but we did not restrict them to directly copy the text. Thus, our dataset also contains captions where OCR tokens are not present directly but were used to infer a description, *e.g.* in Fig. 2a "Rice is winning" instead of "Rice has 18 and Ecu has 17". The collected captions are not limited to trivial template "Object *X* which says *Y*". We have observed various types of relations between text and other objects in a scene which are impossible to formulate without reading comprehension. For example, in Fig. 2: "A *score board* shows Rice with 18 points vs. ECU with 17 points" (a), "*Box* of Hydroxycut on sale for only 17.88 at a *store*" (b).

Fig. 3 compares the percentage of captions with a particular number of OCR tokens between COCO and TextCaps datasets.[2]

TextCaps has a much larger number of OCR tokens in the captions as well as in the images compared to COCO (note the high percentage at 0). A small part (2.7%) of COCO

[1]Apart form direct copying, we also allowed indirect use of text, *e.g.* inferring, paraphrasing, summarizing, or reasoning about it (see Fig. 2). This approach creates a fundamental difference with existing OCR datasets where alteration of text is not acceptable. For captioning systems, however, the ability to reason about text can be beneficial.

[2]Note that OCR tokens have been extracted using Rosetta OCR system [3] which cannot guarantee exhaustive coverage of all text present in an image and presents just an estimation.

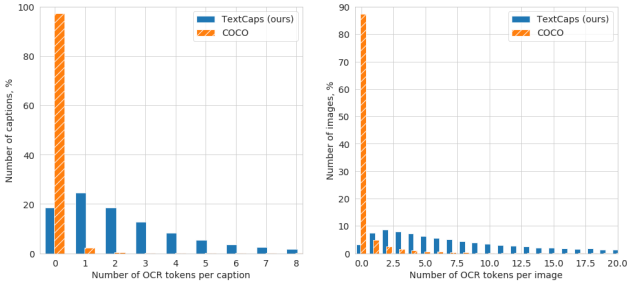

Figure 3: **Distribution of OCR tokens** in COCO [4] and TextCaps captions (left) and images (right). COCO contains **2.7%** of captions and **12.7%** of images with at least one OCR token, whereas TextCaps – **81.3%** and **96.9%**.

#	Method	Trained on	TextCaps validation set					TextCaps test set					
			B-4	M	R	S	C	B-4	M	R	S	C	H
1	BUTD [1]	COCO	12.4	13.3	33.7	8.7	24.2						
2	BUTD [1]	TextCaps	20.1	17.8	42.9	11.7	41.9	14.9	15.2	39.9	8.8	33.8	1.4
3	AoANet [7]	COCO	18.1	17.7	41.4	11.2	32.3						
4	AoANet [7]	TextCaps	20.4	18.9	42.9	13.2	42.7	15.9	16.6	40.4	10.5	34.6	1.4
5	M4C-Captioner	COCO	12.3	14.2	34.8	9.2	30.3						
6	M4C-Captioner	TextVQA	0.1	4.4	11.3	2.8	16.9						
7	M4C-Captioner w/o OCRs	TextCaps	15.9	18.0	39.6	12.1	35.1						
8	M4C-Captioner w/o copying	TextCaps	18.2	19.2	41.5	13.1	49.2						
9	M4C-Captioner	TextCaps	**23.3**	**22.0**	**46.2**	**15.6**	**89.6**	**18.9**	**19.8**	**43.2**	**12.8**	**81.0**	**3.0**
10	M4C-Captioner (w/ GT OCRs; on a subset)	TextCaps	26.0	23.8	48.7	17.0	110.6	22.1	21.7	46.7	14.1	106.0	3.4
11	Human	–						24.4	26.1	47.0	18.8	125.5	4.7

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

Table 1: **Performance of our baselines on our TextCaps dataset.**

captions which contain OCR tokens is mostly limited to one token per caption; only 0.38% of captions contain two or more tokens. Whereas in TextCaps, multi-word reading is much more common (56.8%) which is crucial for capturing real-world information (*e.g.* authors, titles, monuments, *etc.*). Moreover, while COCO Captions contain less than 350 unique OCR tokens, TextCaps contains 39.7k of them. On an average, TextCaps captions contain 2.64 OCR tokens per caption while COCO - 0.03.

In addition, we find that an impressive number of 2901 of 6329 unique OCR tokens appearing in the test set captions, have neither appeared in the training nor validation set (i.e. they are "zero-shot") which makes it necessary for models to be able to read new text in images. TextCaps dataset also creates new technical challenges for the models. Due to the common use of OCR tokens in the captions, models required to switch between OCR and vocabulary words often. The majority of the TextCaps captions require to switch twice or more, whereas most COCO and TextVQA outputs can be generated even without any switches.

3. Benchmark Evaluation

In this section, we evaluate the ability of existing approaches to read and reason about text in the image for caption generation.

3.1. Baselines

Bottom-Up Top-Down Attention model (BUTD) [1] is a widely used image captioning model which generates region proposals based visual features (Bottom-Up) in conjunction with two attention-weighted LSTM layers (Top-Down).

Attention on Attention model (AoANet) [7] is a current SoTA captioning algorithm which uses the attention-on-attention module (AoA) to create a relation between attended vectors in both encoder and decoder.

M4C-Captioner. M4C [6] is a recent model for the TextVQA task.The model fuses different modalities by embedding them into a common semantic space and processing them with a multimodal transformer. We adapt M4C to our task by removing the question input and directly use its multi-word answer decoder to generate a caption conditioned on the detected objects and OCR tokens in the image.

Human performance. In addition to our baselines, we provide an estimate of human performance on the TextCaps test set.

3.2. Results

It can be observed in results (Table 1) that the BUTD model trained on the COCO captioning dataset (line 1) achieves the lowest CIDEr score, indicating that it fails to describe text in the image. When trained on the TextCaps dataset (line 2), the BUTD model has higher scores as expected, since there is no longer a domain shift between training and evaluation. AoANet (line 3, 4), which is a stronger captioning model, outperforms BUTD but still cannot handle reading comprehension and largely underperforms M4C-Captioner. For the M4C-Captioner model, there is a large gap (especially in CIDEr scores) between training with and without OCR inputs (line 9 vs. 7). Moreover, "M4C-Captioner w/o copying" (line 8) is worse than the full model (line 9) but better than the more restricted "M4C-Captioner w/o OCRs" (line 7). The results indicate that it is crucial to both encode OCR features **and** be able to directly copy OCR tokens. However, on the test set, we still notice a large gap between the best machine performance (line 9) and the human performance (line 11) on this task. Also, using ground-truth OCRs as inputs (line 10) reduces this gap but still does not close it, suggesting that there is room for future improvement in both better reasoning and better text recognition.

Figure 4 shows qualitative examples from different methods. It can be seen that BUTD and M4C-Captioner without OCR inputs rarely mention text in the image except for common brand logos such as "pepsi" that are easy to recognize visually. On the other hand, the full M4C-Captioner approach learns to read text in the image and mention it in its generated captions. Moreover, M4C-Captioner learns

a

BUTD: a white laptop computer sitting on top of a table
M4C-Captioner (w/o OCR): the front of an lg phone that is white and black.
M4C-Captioner: the front and back of an [lg] phone that is on [october] [19].
Human: An advertisment shows the LG Optimus L9 phone.

b

BUTD: a close up of a can of soda and a can of soda
M4C-Captioner (w/o OCR): two pepsi bottles sit on a wooden shelf next to a pepsi bottle.
M4C-Captioner: two pepsi bottles are next to each other on a wooden shelf.
Human: 3 pepsi bottles next to each other on a shelf

c

BUTD: a close up of a clock on a wall
M4C-Captioner (w/o OCR): a sprint phone with the words score and score at the top.
M4C-Captioner: a digital sign says the [track] is [moved] in [kenosha] [pm] [pm] and is in the middle of the screen.
Human: A kiosk of track 13 of Metra which states that the 5:43 train has moved tracks.

Figure 4: **Illustration predictions from different models.** Square brackets indicate tokens copied from OCR.

and recognizes relations between objects and is able to combine multiple OCR tokens into one complex description. For e.g., in Fig. 4(c) the model attempts to include and combine multiple tokens into a single message ("the *track* is *moved* in *Kenosha*" instead of "the word *moved*, the word *track*, and the word *Kenosha* are on the sign"). This points to many potential directions for future development on this challenging generative task, which requires visual and textual understanding, requiring new model designs, conceptually different from previously existing captioning models.

4. Conclusion

Image captioning with reading comprehension is a novel challenging task requiring models to read text in the image, recognize the image content, and comprehend both modalities jointly to generate a succinct image caption. To enable learning and to study this task in isolation, we collected TextCaps with 142k captions. Our analysis also points out several challenges of this dataset: different from captioning datasets, nearly all our captions require integration of OCR tokens, many are unseen ("zero-shot"). In contrast to TextVQA datasets, our data requires to generate long sentences, which includes new technical challenges, including many switches between OCR and vocabulary tokens. By collecting GT OCR tokens for the validation and test set, we see that the dataset also contains several challenges for current OCR systems.

We hope our dataset with challenge server, available at textvqa.org/textcaps, will encourage the community to design novel image captioning models for this novel task and its technical challenges, especially increasing their usefulness for assisting visually disabled people.

The full paper to this extended abstract with appendix can be found at https://arxiv.org/abs/2003.12462

References

[1] 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. 1, 3

[2] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. *arXiv preprint arXiv:1905.13648*, 2019. 1

[3] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 71–79. ACM, 2018. 1, 2

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 2

[5] Priya Goyal, Dhruv Kumar Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. *International Conference on Computer Vision*, abs/1905.01235, 2019. 1

[6] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. *arXiv preprint arXiv:1911.06258*, 2019. 1, 3

[7] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *IEEE International Conference on Computer Vision*, pages 4634–4643, 2019. 1, 3

[8] Hui Li, Peng Wang, and Chunhua Shen. Towards end-to-end text spotting with convolutional recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5238–5246, 2017. 1

[9] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019. 1

[10] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 1, 2

[11] Ray Smith. An overview of the tesseract ocr engine. In *International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007. 1