TextCaps: A Dataset for Image Captioning with Reading Comprehension

Oleksii Sidorov[1](✉), Ronghang Hu[1,2], Marcus Rohrbach[1], and Amanpreet Singh[1]

[1] Facebook AI Research, Menlo Park, USA
acecreamu@gmail.com, {mrf,asg}@fb.com
[2] University of California, Berkeley, USA
ronghang@eecs.berkeley.edu

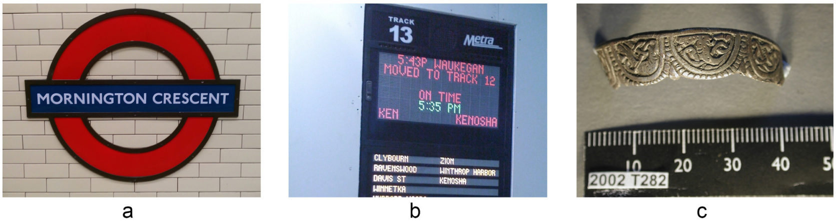
Abstract. Image descriptions can help visually impaired people to quickly understand the image content. While we made significant progress in automatically describing images and optical character recognition, current approaches are unable to include written text in their descriptions, although text is omnipresent in human environments and frequently critical to understand our surroundings. To study how to comprehend text in the context of an image we collect a novel dataset, TextCaps, with 145k captions for 28k images. Our dataset challenges a model to recognize text, relate it to its visual context, and decide what part of the text to copy or paraphrase, requiring spatial, semantic, and visual reasoning between multiple text tokens and visual entities, such as objects. We study baselines and adapt existing approaches to this new task, which we refer to as *image captioning with reading comprehension*. Our analysis with automatic and human studies shows that our new TextCaps dataset provides many new technical challenges over previous datasets.

1 Introduction

When trying to understand man-made environments, it is not only important to recognize objects but also 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 "40 mm wide" (Fig. 1(c)). Reading comprehension in images is crucial for blind people. As the VizWiz datasets [5] suggest, 21% of questions visually-impaired people asked about an image were related to the text in it. Image captioning plays an important role in starting a visual dialog with a blind user allowing them to ask for further information as required. 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.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58536-5_44) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12347, pp. 742–758, 2020.
https://doi.org/10.1007/978-3-030-58536-5_44



Model: a macdonald ' s sign that is on a brick wall

Human: A tile wall with a red circle on it reading Mornington Crescent

Model: a sign that has the time of 12 : 37 on it

Human: A kiosk of track 13 of Metra which states that the 5:43 train has moved tracks

Model: a ruler that has the number 2003 on it

Human: An old artifact being measured by a ruler that shows it is around 40 millimeters wide

Fig. 1. Existing captioning models cannot read! The *image captioning with reading comprehension* task using data from our TextCaps dataset and BUTD model [4] trained on it. (Color figure online)

In recent years, with the availability of large labelled corpora, progress in image captioning has seen steady increase in performance and quality [4,10,12,13,34] and reading scene text (OCR) has matured [8,16,19,21,31]. 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, Fig. 1 shows predictions of a state-of-the-art model [4] on a few images that require reading comprehension. The predictions 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.

Considering the images and reference captions in Fig. 1, we can breakdown what is needed to successfully describe these images: First, detect and extract text/OCR tokens[1] (*'Mornington Crescent'*, *'moved track'*) as well the visual context such as objects in the image (*'red circle'*, *'kiosk'*). Second, generate a grammatically correct sentence which combines words from the vocabulary and OCR tokens. In addition to the challenges in normal captioning, *image captioning with reading comprehension* can include the following technical challenges:

1. Determine the relationships **between different OCR tokens** and between **OCR tokens and the visual context**, to decide if an OCR token should be mentioned in the sentence and which OCR tokens should be joined together (*e.g.* in Fig. 1b: "5:35" denotes the current time and should not be joined with

[1] The remainder of the manuscript we refer to the text in an image as "OCR tokens", where one token is typically a word, i.e. a group of characters.

"ON TIME"), based on their (a) *semantics* (Fig. 2b), (b) *spatial* relationship (Fig. 1c), and (c) *visual* appearance and context (Fig. 2d).
2. **Switching multiple times** during caption generation between the words from the model's vocabulary and OCR tokens (Fig. 1b).
3. **Paraphrasing and inference** about the OCR tokens (Fig. 2 bold).
4. Handling of OCR tokens, including ones never seen before (**zero-shot**).

While this list should not suggest a temporal processing order, it explains why today's models lack capabilities to comprehend text in images to generate meaningful descriptions. It is unlikely that the above skills will naturally emerge through supervised deep learning on existing image captioning datasets as they are not focusing on this problem. In contrast, captions in these datasets are collected in a way that implicitly or explicitly avoids mentioning specific instances appearing in the OCR text. To study the novel task of image captioning with reading comprehension, we thus believe it is important to build a dataset containing captions which require reading and reasoning about text in images. We find the COCO Captioning dataset [9] not suitable as only an estimated 2.7% of its captions mention OCR tokens present in the image, and in total there are less than 350 different OCRs (i.e. the OCR vocabulary size), moreover most OCR tokens are common words, such as "stop", "man", which are already present in a standard captioning vocabulary. Meanwhile, in Visual Question Answering, multiple datasets [6,23,30] were recently introduced which focus on text-based visual question answering. This task is harder than OCR recognition and extraction as it requires understanding the OCR extracted text in the context of the question and the image to deduce the correct answer. However, although these datasets focus on text reading, the answers are typically shorter than 5 words (mainly 1 or 2), and, typically, all the words which have to be generated are either entirely from the training vocabulary *or* OCR text, rather than requiring switching between them to build a complete 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 [30] model, M4C [17], when trained with our dataset TextCaps, gets encouraging results. Our ablation study shows that it is important to take into account all semantic, visual, and spatial information of OCR tokens to generate high-quality captions.

- 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 Related Work

Image Captioning. The Flickr30k [35] and COCO Captions [9] dataset have both been collected similarly via crowd-sourcing. The COCO Captions dataset is significantly larger than Flickr30k and acts as a base for training the majority of current state-of-the-art image captioning algorithms. It includes 995,684 captions for 164,062 images. The annotators of COCO were asked “Describe all the important parts of the scene” and “Do not describe unimportant details”, which resulted in COCO being focused on objects which are more prominent rather than text. SBU Captions [24] is an image captioning dataset which was collected automatically by retrieving one million images and associated user descriptions from Flickr, filtering them based on key words and sentence length. Similarly, Conceptual Captions (CC) dataset [27] is also automatically constructed by crawling images from web pages together with their ALT-text. The collected annotations were extensively filtered and processed, e.g. replacing proper names and titles with object classes (*e.g.* man, city), resulting in 3.3 million image-caption pairs. This simplifies caption generation but at the same time removes fine details such as unique OCR tokens. Apart from conventional paired datasets there are also datasets like NoCaps [1], oriented to a more advanced task of captioning with zero-shot generalization to novel object classes.

While our TextCaps dataset also consists of image-sentence pairs, it focuses on the text in the image, posing additional challenges. Specifically, text can be seen as an additional modality, which models have to read (typically using OCR), comprehend, and include when generating a sentence. Additionally, many OCR tokens do not appear in the training set, but only in the test (zero-shot). In concurrent work, [15] collect captions on VizWiz [5] images but unlike TextCaps there isn’t a specific focus on reading comprehension.

Optical Character Recognition (OCR). OCR involves in general two steps, namely (i) detection: finding the location of text, and (ii) extraction: based on the detected text boundaries, extracting the text as characters. OCR can be seen as a subtask for our *image captioning with reading comprehension* task as one needs to know the text present in the image to generate a meaningful description of an image containing text. This makes OCR research an important and relevant topic to our task, which additionally requires to understand the importance of OCR token, their semantic meaning, as well as relationship to visual context and other OCR tokens. Recent OCR models have shown reliability and performance improvements [8,16,19,21,31]. However, in our experiments we observe that OCR is far from a solved problem in real-world scenarios present in our dataset.

Visual Question Answering with Text Reading Ability. Recently, three different text-oriented datasets were presented for the task of Visual Question Answering. TextVQA [30] consists of 28,408 images from selected categories of Open Images v3 dataset, corresponding 45,336 questions, and 10 answers for each question. Scene Text VQA (ST-VQA) dataset [6] has a similar size of 23,038 images and 31,791 questions but only one answer for each question. Both these datasets were annotated via crowd-sourcing. OCR-VQA [23] is a larger dataset (207,572 images) collected semi-automatically using photos of book covers and corresponding metadata. The rule generated questions were paraphrased by human annotators. These three datasets require reading and reasoning about the text in the image while considering the context for answering a question, which is similar in spirit to TextCaps. However, the image, question and answer triplet is not directly suitable for generation of descriptive sentences. We provide additional quantitative comparisons and discussion between our and existing captioning and VQA datasets in Sect. 3.2.

3 TextCaps Dataset

We collect TextCaps with the goal of studying the novel task of *image captioning with reading comprehension*. Our 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. In this section, we describe the dataset collection and analyze its statistics. The dataset is publicly available at textvqa.org/textcaps.

3.1 Dataset Collection

With the goal of having a diverse set of images, we rely on images from Open Images v3 dataset (CC 2.0 license). Specifically, we use the same subset of images as in the TextVQA dataset [30]; these images have been verified to contain text through an OCR system [8] and human annotators [30]. Using the same images as TextVQA additionally allows multi-task and transfer learning scenarios between OCR-based VQA and image captioning tasks. The images were annotated by human annotators in two stages.[2]

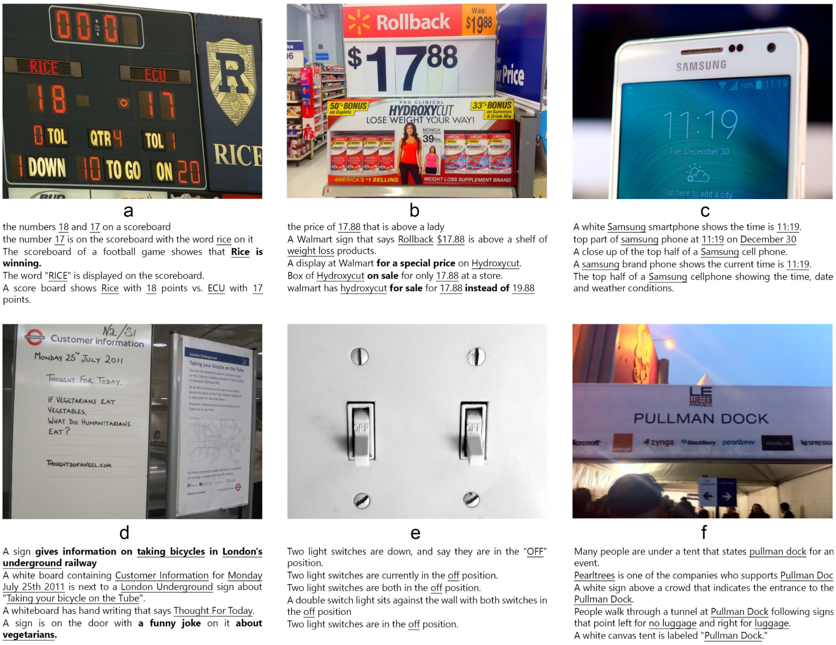
Annotators were asked to describe an image in one sentence which would require reading the text in the image.[3]

[2] The full text of the instructions as well as screenshots of the user interface are presented in the Supplemental (Sec. F).

[3] Apart from 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 from OCR datasets where alteration of text is not acceptable. For captioning, however, the ability to reason about text can be beneficial.

Evaluators were asked to vote yes/no on whether the caption written in the first step satisfies the following requirements: requires reading the text in the image; is true for the given image; consists of one sentence; is grammatically correct; and does not contain subjective language. The majority of 5 votes was used to filter captions of low quality. The quality of the work of evaluators was controlled using gold captions of known good/bad 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. The annotators did not see previously collected captions for a particular image and did not see the same image twice. 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. An estimation performed using ground-truth OCR shows that on average, 39.5% out of all OCR tokens present in the image are covered by the collected human annotations.



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.

d

A sign **gives information on taking bicycles in London's underground railway**
A white board containing Customer Information for Monday July 25th 2011 is next to a London Underground sign about "Taking your bicycle on the Tube".
A whiteboard has hand writing that says Thought For Today.
A sign is on the door with **a funny joke** on it **about vegetarians.**

e

Two light switches are down, and say they are in the "OFF" position.
Two light switches are currently in the off position.
Two light switches are both in the off position.
A double switch light sits against the wall with both switches in the off position
Two light switches are in the off position.

f

Many people are under a tent that states pullman dock for an event.
Pearltrees is one of the companies who supports Pullman Doc
A white sign above a crowd that indicates the entrance to the Pullman Dock.
People walk through a tunnel at Pullman Dock following signs that point left for no luggage and right for luggage.
A white canvas tent is labeled "Pullman Dock."

Fig. 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.

3.2 Dataset Analysis

We first discuss several properties of the TextCaps qualitatively and then analyse and compare its statistics to other captioning and OCR-based VQA datasets.

Qualitative Observations. 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". In a human evaluation of 640 captions we found that about 20% of images have at least one caption (8% of captions) which require more challenging reasoning or paraphrasing rather than just direct copying of visible text. Nevertheless, even the captions which require copying text directly can be complex and may require advanced reasoning as illustrated in multiple examples in Fig. 2. 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), "Two *light switches* are both in off position" (e).

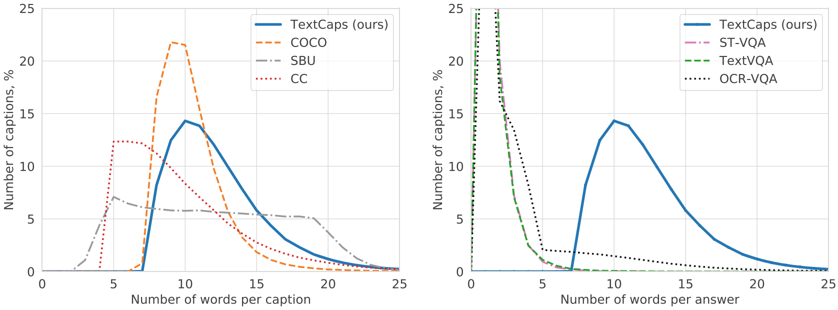


Fig. 3. Distribution of caption/answer lengths in Image Captioning (left) and VQA (right) datasets. VQA answers are significantly shorter than image captions and mostly concentrated within 5 words limit.

Dataset Statistics. To situate TextCaps properly w.r.t. other image captioning datasets, we compare TextCaps with other prominent image captioning datasets, namely COCO [9], SBU [24], and Conceptual Captions [27], as well as reading-oriented VQA datasets TextVQA [30], ST-VQA [6], and OCR-VQA [23]. The average caption length is 12.0 words for SBU, 9.7 words for Conceptual Captions, and 10.5 words for COCO, respectively. The average length for TextCaps is 12.4, slightly larger than the others (see Fig. 3). This can be explained by the fact that captions in TextCaps typically include both scene description as well as the text from it in one sentence, while conventional captioning datasets only cover the scene description. Meanwhile, the average answer length is 1.53 for TextVQA, 1.51 for ST-VQA and 3.31 for OCR-VQA – much smaller than the captions in

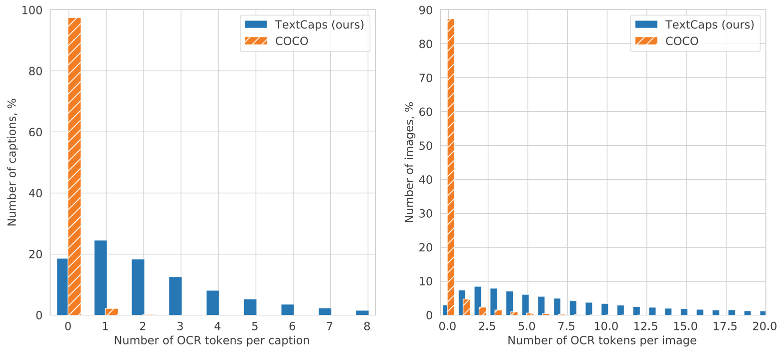
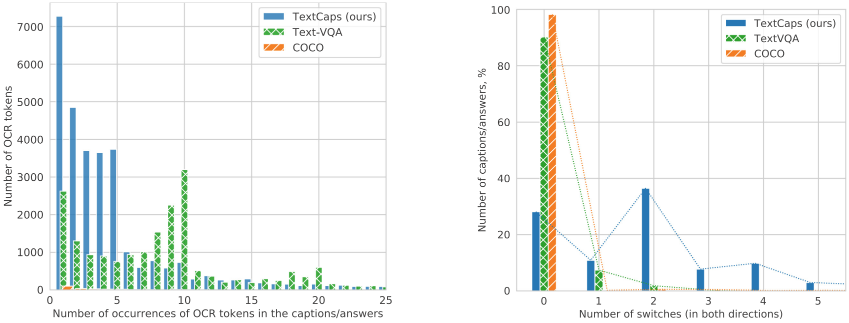


Fig. 4. Distribution of OCR tokens in COCO and TextCaps captions (left) and images (right). In total, COCO contains **2.7%** of captions and **12.7%** of images with at least one OCR token, whereas TextCaps – **81.3%** and **96.9%**.



(a) **OCR frequency distribution** shows how many OCR tokens occur once, twice, *etc.* TextCaps has the largest amount of unique and rare (< 5) OCR tokens. Note that TextVQA has 10 answers for each question which are often identical.

(b) **Number of switches between OCR $\rightleftharpoons$ Vocab** illustrates the technical complexity of the datasets. An approach which cannot make switches will be sufficient for most of COCO captions and TextVQA but not for TextCaps.

Fig. 5. Analysis of OCR in our dataset vs. others

our dataset. Typical answers like *'yes'*, *'two'*, *'coca cola'* may be sufficient to answer a question but insufficient to describe the image comprehensively.

Figure 4 compares the percentage of captions with a particular number of OCR tokens between COCO and TextCaps datasets.[4] 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 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-

[4] Note that OCR tokens are extracted using Rosetta OCR system [8] which cannot guarantee exhaustive coverage of all text in an image and presents just an estimation.

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.

We also measured the frequency of OCR tokens in the captions. Figure 5a illustrates the number of times a particular OCR token appears in the captions. More than 9000 tokens appear only once in the whole dataset. The curve drops rapidly after 5 occurrences and only a small part of tokens occur more than 10 times. Quantitatively, 75.7% of tokens are presented less then 5 times, and only 12.9% are presented more than 10 times. The distribution specifically demonstrates the large variance in text occurring in natural images which is challenging to model using a fixed word vocabulary. In addition to this long-tailed distribution, 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. Figure 5b illustrates that 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.

4 Benchmark Evaluation

4.1 Baselines

Our baselines aim to illustrate the gap between performance of conventional state-of-the-art image captioning models (BUTD [4], AoANet[18]) in comparison to recent architectures which incorporate reading (M4C [17]).

Bottom-Up Top-Down Attention Model (BUTD). [4] is a widely used image captioning model based on Faster R-CNN [26] object detection features (Bottom-Up) in conjunction with attention-weighted LSTM layers (Top-Down).

Attention on Attention Model (AoANet). [18] 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 [17] is a recent model with state-of-the-art performance on the TextVQA task. The model fuses different modalities by embedding them into a common semantic space and processing them with a multimodal transformer. Apart from that, unlike conventional VQA models where a prediction is made via classification, it enables iterative answer decoding with a dynamic pointer network [22,33], allowing the model to generate a multi-word answer, which is not limited to a fixed vocabulary. This feature makes it also suitable for reading-based caption generation. 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 (we refer to this model as **M4C-Captioner** and illustrate it in Fig. 6).

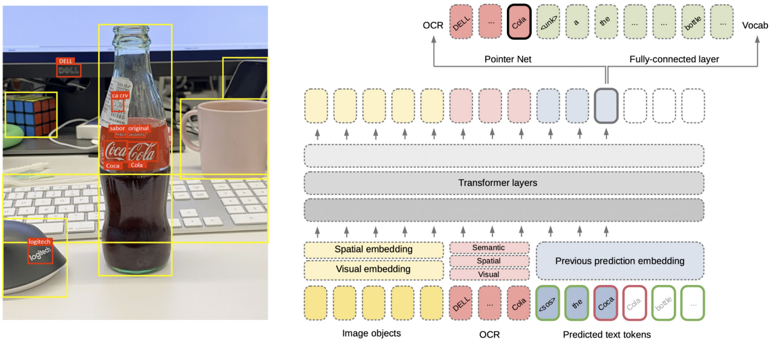


Fig. 6. M4C-Captioner architecture for the image captioning with reading comprehension task.

M4C-Captioner Ablations. In comparison to its full version, we also evaluate a restricted version of this model without access to OCR results (referred to as **M4C-Captioner w/o OCRs**), where we use an empty OCR token list as input to the model. Additionally, we experiment with removing the pointer network (described in details in [17]) from M4C-Captioner, so that the model still has access to OCR features but cannot directly copy OCR tokens, and must use its fixed vocabulary for caption generation (referred to as **M4C-Captioner w/o copying**). As multiple types of features are used for OCR tokens in M4C-Captioner by default (same as in [17]), we further study the impact of each OCR feature type and use only **spatial** information (4-dimensional relative bounding box coordinates $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ of OCR tokens), **semantic** information (FastText [7] and PHOC [2]), and **visual** (Faster R-CNN [26]) features in different experiments. Additionally, we use ground truth OCR tokens annotated by humans (referred to as **M4C-Captioner w/ GT OCRs**) for training and prediction[5] to study the influence of mistakes of automatic OCR methods.

Human Performance. In addition to our baselines, we provide an estimate of human performance by using the same metrics on the TextCaps test set to benchmark the progress that models still need to make. As discussed in Sect. 4.3, we collected one more caption for each image in the test set. The metrics are then calculated by averaging the results over 6 runs, each time leaving out one caption as a prediction, similar to [14]. On the test set, we use the same approach to evaluate machine-generated captions, so numbers are comparable.

[5] This includes a small number of images without GT-OCRs (Supplemental Sec. A).

Table 1. Performance of our baselines on our TextCaps dataset. M4C-Captioner significantly benefits from OCR inputs and achieves the highest CIDEr score, suggesting that it is important to copy text from image on this task. However, there is still a large gap between the current machine performance and human performance, which we hope can be closed by future work.

#	Method	Trained on	TextCaps validation set metrics					
			B-4	M	R	S	C	
1	BUTD [4]	COCO	12.4	13.3	33.7	8.7	24.2	
2	BUTD [4]	TextCaps	20.1	17.8	42.9	11.7	41.9	
3	AoANet [18]	COCO	18.1	17.7	41.4	11.2	32.3	
4	AoANet [18]	TextCaps	20.4	18.9	42.9	13.2	42.7	
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 (OCR semantic)	TextCaps	21.4	20.4	44.0	14.1	69.0	
10	M4C-Captioner (OCR spatial)	TextCaps	21.7	20.6	44.6	13.7	72.0	
11	M4C-Captioner (OCR visual)	TextCaps	22.5	21.3	45.3	14.4	84.0	
12	M4C-Captioner (OCR semantic & visual)	TextCaps	23.4	21.5	45.8	14.9	86.0	
13	M4C-Captioner	TextCaps	**23.3**	**22.0**	**46.2**	**15.6**	**89.6**	
14	M4C-Captioner (w/ GT OCRs)	TextCaps	26.0	23.2	47.8	16.2	104.3	
#	Method	Trained on	TextCaps test set metrics					
			B-4	M	R	S	C	H
15	BUTD [4]	TextCaps	14.9	15.2	39.9	8.8	33.8	1.4
16	AoANet [18]	TextCaps	15.9	16.6	40.4	10.5	34.6	1.4
17	M4C-Captioner	TextCaps	**18.9**	**19.8**	**43.2**	**12.8**	**81.0**	**3.0**
18	M4C-Captioner (w/ GT OCRs)	TextCaps	21.3	21.1	45.0	13.5	97.2	3.4
19	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

4.2 Experimental Setup

[6]We follow the default configurations and hyper-parameters for training and evaluation of each baseline. For AoANet we use original implementation and feature extraction technique. For BUTD [4], we use the implementation and hyper-parameters from MMF [28,29]. For M4C-Captioner [17], we follow the same implementation details as used for TextVQA task [17]. We train both models for the same number of iterations on the TextCaps training set. During caption generation, we remove the `<unk>` token (for unknown words).

Datasets. We first evaluate the models trained using COCO dataset on TextCaps to demonstrate how existing datasets and models lack reading comprehension. Then we train and evaluate each baseline using TextCaps.

[6] Code for experiments is available at https://git.io/JJGuG.

Metrics. Apart from automatic captioning metrics including BLEU [25], METEOR [11], ROUGE_L [20], SPICE [3], and CIDEr [32], we also perform human evaluation. We collect 5000 human scores on a Likert scale from 1 to 5 for a random sample of 200 images and compute median score for each caption. Figure 7 shows that ranking of the sentence quality is the same as for automatic metrics. Moreover, all the metrics show very high correlation with human scores but CIDEr and METEOR have the highest. For comparison between different methods, we focus on the CIDEr, which puts more weight on informative n-grams in the captions (such as OCR tokens) and less weight on commonly occurring words with TF-IDF weighting.

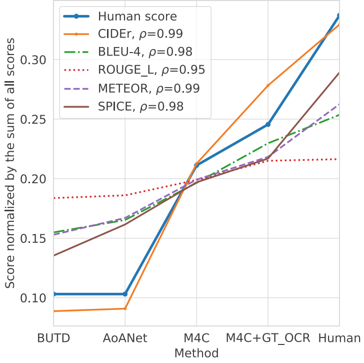
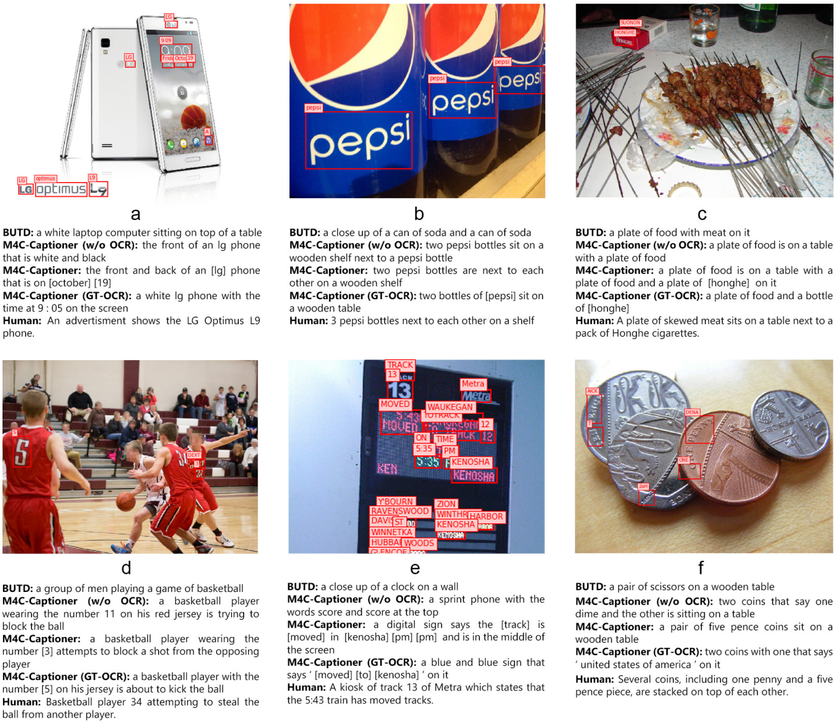


Fig. 7. Human evaluation in comparison to automatic metrics.



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]
M4C-Captioner (GT-OCR): a white lg phone with the time at 9 : 05 on the screen
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
M4C-Captioner (GT-OCR): two bottles of [pepsi] sit on a wooden table
Human: 3 pepsi bottles next to each other on a shelf

c

BUTD: a plate of food with meat on it
M4C-Captioner (w/o OCR): a plate of food is on a table with a plate of food
M4C-Captioner: a plate of food is on a table with a plate of food and a plate of [honghe] on it
M4C-Captioner (GT-OCR): a plate of food and a bottle of [honghe]
Human: A plate of skewed meat sits on a table next to a pack of Honghe cigarettes.

d

BUTD: a group of men playing a game of basketball
M4C-Captioner (w/o OCR): a basketball player wearing the number 11 on his red jersey is trying to block the ball
M4C-Captioner: a basketball player wearing the number [3] attempts to block a shot from the opposing player
M4C-Captioner (GT-OCR): a basketball player with the number [5] on his jersey is about to kick the ball
Human: Basketball player 34 attempting to steal the ball from another player.

e

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
M4C-Captioner (GT-OCR): a blue and blue sign that says ' [moved] [to] [kenosha] ' on it
Human: A kiosk of track 13 of Metra which states that the 5:43 train has moved tracks.

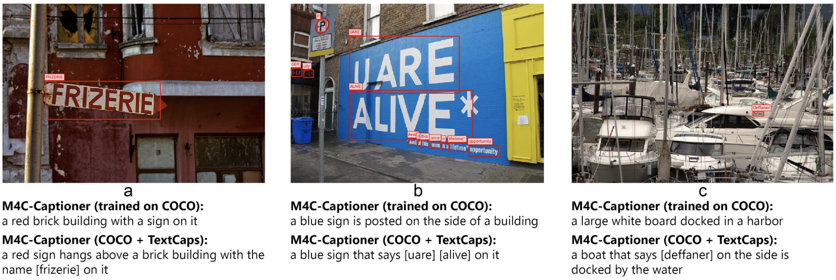
f

BUTD: a pair of scissors on a wooden table
M4C-Captioner (w/o OCR): two coins that say one dime and the other is sitting on a table
M4C-Captioner: a pair of five pence coins sit on a wooden table
M4C-Captioner (GT-OCR): two coins with one that says ' united states of america ' on it
Human: Several coins, including one penny and a five pence piece, are stacked on top of each other.

Fig. 8. Illustration of positive and negative predictions from different models on TextCaps validation set. For M4C-Captioner, square brackets indicate tokens copied from OCR. While most of the time OCR tokens are very important for correct copying of the text from the images, for common terms such as "pepsi" or "pence", the model sometimes prefer to select them from the vocabulary.

4.3 Results

TextCaps Dataset. 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 13 vs. 7). Moreover, "M4C-Captioner w/o copying" (line 8) is worse than the full model (line 13) but better than the more restricted "M4C-Captioner w/o OCRs" (line 7). The results indicate that it is important to both encode OCR features **and** be able to directly copy OCR tokens. We also observe (in line 13 vs. 9–12) that it is important for a model to use spatial, visual, and semantic features of OCR tokens together, especially in the complex combinations of OCR tokens where both spatial relation and semantics play an important role in finding a connection between words. However, on the test set, we still notice a large gap between the best machine performance (line 17) and the human performance (line 19) on this task. Also, using ground-truth OCRs (line 18) 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.



a

M4C-Captioner (trained on COCO):
a red brick building with a sign on it

M4C-Captioner (COCO + TextCaps):
a red sign hangs above a brick building with the name [frizerie] on it

b

M4C-Captioner (trained on COCO):
a blue sign is posted on the side of a building

M4C-Captioner (COCO + TextCaps):
a blue sign that says [uare] [alive] on it

c

M4C-Captioner (trained on COCO):
a large white board docked in a harbor

M4C-Captioner (COCO + TextCaps):
a boat that says [deffaner] on the side is docked by the water

Fig. 9. Examples of M4C-Captioner's predictions on COCO data when trained on COCO and TextCaps. It can be observed that *despite of availability of OCR module in both cases*, using TextCaps pushes model to read the text. Square brackets indicate tokens copied from OCR.

Figure 8 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.[7] Moreover,

[7] More predictions from M4C-Captioner are presented in Supplemental (Fig. F.1).

M4C-Captioner learns and recognizes relations between objects and is able to combine multiple OCR tokens into one complex description. For e.g., in Fig. 8(d) the model uses a OCR token to correctly name a player who is blocking another player; in Fig. 8(e) 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"). In Fig. 8(b) prediction is constructed fully from vocabulary, and even then the model counts similar objects and returns "two pepsi bottles" instead of "pepsi bottle and pepsi bottle". We also observe a large amount of mistakes in model predictions. Many mistakes are due to wrong scene understanding and object identification, which is a common problem in captioning algorithms. We also observe placing OCR tokens in the wrong object or semantic context in the caption (Fig. 8(c, e)), incorrect repetition of an OCR token in a caption (Fig. 8(a, e)), or insufficient use of them (Fig. 8(f)) by the model. Some mistakes (as "number *3*" in Fig. 8(d) are due to the errors of OCR detection algorith m and not the captioning model. 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.

Transferring to COCO. We further qualitatively show that when integrated with other datasets such as COCO [9], our dataset also enables text-based captioning on other datasets. In this setting, we experiment training M4C-Captioner (Table 1's best) on both TextCaps dataset and COCO dataset together. We balance the number of samples seen by the model from both COCO and TextCaps during training, and apply the trained model on the COCO validation set. COCO Captions mostly focus on visual objects but we show several examples where reading is necessary to describe the scene in Fig. 9. When trained on the union of our dataset and COCO, the M4C-Captioner learns to generate captions containing text present in the images. On the other hand, the same model only describes visual objects without mentioning any text when trained on COCO alone. Quantitative results can be found in Supplemental (Sec. C).

5 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 models to learn this ability and study this task in isolation, we collected TextCaps with 142k captions. The captions include a mix of objects and/or visual scene descriptions in relation to OCR tokens copied or rephrased from the images. In most cases, OCR tokens have to be copied and related to the visual scene, but sometimes the OCR tokens have to be understood, and sometimes spatial or visual reasoning between text and objects in the image is required, as shown in our ablation study. Our analysis also points out several challenges of this dataset: Different from other captioning datasets, nearly all our captions require integration of OCR tokens, many are unseen ("zero-shot"). In contrast to TextVQA datasets, TextCaps requires generating long sentences and involves new technical challenges, including many switches between OCR and vocabulary tokens.

We find that current state-of-the-art image captioning models cannot read when trained on existing captioning dataset. However, when adapting the recent M4C VQA model to our task and training it on our TextCaps dataset, we are able to generate impressive captions on both TextCaps and COCO, which involve copying multiple OCR tokens and correctly integrating them in the captions. Our human evaluation confirms the result of the automatic metrics with very high correlation, and also shows that human captions are still significantly better than automatically generated ones, leaving room for many advances in future work, including better semantic understanding between image and text content, missing reasoning capabilities, and reading long text or single characters.

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

Acknowledgments. We would like to thank Guan Pang and Mandy Toh for helping us with OCR ground-truth collection. We would also like to thank Devi Parikh for helpful discussions and insights.

References

1. Agrawal, H., et al.: nocaps: novel object captioning at scale. In: International Conference on Computer Vision (ICCV) (2019)
2. Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Word spotting and recognition with embedded attributes. IEEE Trans. Pattern Anal. Mach. Intell. **36**(12), 2552–2566 (2014)
3. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: semantic propositional image caption evaluation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 382–398. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_24
4. Anderson, P., et al.: 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, pp. 6077–6086 (2018)
5. Bigham, J.P., et al.: Vizwiz: nearly real-time answers to visual questions. In: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, pp. 333–342. ACM (2010)
6. Biten, A.F., et al.: Scene text visual question answering. arXiv preprint arXiv:1905.13648 (2019)
7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
8. 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)
9. Chen, X., et al.: Microsoft coco captions: data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)

10. Chen, Y.C., et al.: Uniter: learning universal image-text representations. arXiv preprint arXiv:1909.11740 (2019)
11. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380 (2014)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019)
13. Goyal, P., Mahajan, D.K., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: International Conference on Computer Vision, abs/1905.01235 (2019)
14. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: VQA 2.0 evaluation. https://visualqa.org/evaluation.html
15. Gurari, D., Zhao, Y., Zhang, M., Bhattacharya, N.: Captioning images taken by people who are blind. arXiv preprint arXiv:2002.08565 (2020)
16. He, T., Tian, Z., Huang, W., Shen, C., Qiao, Y., Sun, C.: An end-to-end textspotter with explicit alignment and attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5020–5029 (2018)
17. Hu, R., Singh, A., Darrell, T., Rohrbach, M.: Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA. arXiv preprint arXiv:1911.06258 (2019)
18. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: IEEE International Conference on Computer Vision, pp. 4634–4643 (2019)
19. Li, H., Wang, P., Shen, C.: Towards end-to-end text spotting with convolutional recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5238–5246 (2017)
20. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text summarization Branches Out, pp. 74–81 (2004)
21. Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., Yan, J.: Fots: fast oriented text spotting with a unified network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5676–5685 (2018)
22. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7219–7228 (2018)
23. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: OCR-VQA: visual question answering by reading text in images. In: ICDAR (2019)
24. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2Text: describing images using 1 million captioned photographs. In: Neural Information Processing Systems (NIPS) (2011)
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
27. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), pp. 2556–2565 (2018)
28. Singh, A., et al.: MMF: a multimodal framework for vision and language research (2020). https://github.com/facebookresearch/mmf

29. Singh, A., et al.: Pythia-a platform for vision & language research. In: SysML Workshop, NeurIPS, vol. 2018 (2018)
30. Singh, A., et al.: Towards VQA models that can read. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8317–8326 (2019)
31. Smith, R.: An overview of the tesseract OCR engine. In: International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 629–633. IEEE (2007)
32. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015)
33. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: Advances in Neural Information Processing Systems, pp. 2692–2700 (2015)
34. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of International Conference on Learning Representations (2019)
35. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguist. **2**, 67–78 (2014)