Captioning Images Taken by People Who Are Blind - Supplementary Materials

Danna Gurari, Yinan Zhao, Meng Zhang, Nilavra Bhattacharya
University of Texas at Austin

This document supplements Sections 3 and 4 of the main paper. In particular, it includes the following:

– Implementation description of the crowdsourcing system (supplements Section 3.1)
– Analysis of the consistency of captions collected from different crowd workers for each image (supplements Section 3.1)
– Examples of images that are deemed insufficient or low quality for captioning (supplements Section 3.2)
– Visualizations and quantitative analysis demonstrating the diversity of content in VizWiz-Captions and how it compares to that in MSCOCO-Captions and the image classification datasets (supplements Section 3.2)
– Algorithm performance when using data augmentation during training by blurring images (supplements Section 4)

1 Dataset Creation (supplements Section 3.1)

1.1 Crowdsourcing Task Design

A screen shot of our crowdsourcing interface is shown in Figure 1. The interface prevented the crowdworker from proceeding to the next image (for the sequential set of five images) or submitting the work until the following criteria was met for each image description:

– Contains at least eight words (to encourage rich content)
– Contains only one period followed by a space (to restrict crowd worker to one-sentence descriptions)
– Sentences do not begin with the following prefixes (to discourage uninformative content): “There is”, “There are”, “This is”, “These are”, “The image”, “The picture”, “This image”, “This picture”, “It is”, and “It’s”

We collected all the annotations over five batches of Human Intelligence Tasks (HITs) in order to minimize the impact of inadequate workers. After each batch, we identified workers who we considered to be problematic and blocked them from participating in subsequent batches. To identify problematic workers, the authors reviewed a subset of the crowdworkers’ results. The following mechanisms were used to determine which workers’ captions to review:
Workers who were a statistical outlier in time-to-submit (taking either too little or too much time) by 1.95 times the standard deviation for all the results
- Workers who used CAPS LOCK for more than 50% of the caption text
- Workers who used the canned text ("Quality issues are too severe to recognize visual content") for more than 50% of the images that they captioned
– Workers who were the only one to either use or not use the canned text ("Quality issues are too severe to recognize visual content") for an image
– Workers who used words like "quality", "blur" and "blurry" (but not the canned text), and so were not focusing on content in the image
– Random sample from all results

We also included numerous additional quality control mechanisms. First, crowdworkers could not submit their results until their work passed an automated check that verified they followed a number of the task instructions, including writing at least 8 words, providing only one sentence, and not starting the description with “There is...” or other unsubstantial starting phrases. We also only accepted crowdworkers who previously had completed over 500 HITs with at least a 95% acceptance rate. We will publicly-share the crowdsourcing code to support reproducibility of this interface.

1.2 Caption Post-processing

We employed Microsoft Azure’s spell-checking API\(^1\) to find and correct misspelled words in the submitted captions. We chose this approach because we found from initial testing that it outperforms other tested methods, including because it can recognize brand names (which are common in our dataset). It also does a good job of correcting grammar and capitalizing words when appropriate (e.g. changing "dell" to "Dell"). When spell-checking all captions which are neither canned text nor from blocked workers (i.e., 169,073 captions), 14% (i.e., 23,424) were flagged as containing unknown “tokens” (aka - words). We replace each unknown “token” with the most confidently recommended word suggested by the Azure API.

\(^1\) https://azure.microsoft.com/en-us/services/cognitive-services/spell-check/

![Histogram of Specificity Scores of VizWiz Captions](image1.png)

![Histogram of Specificity Scores of MSCOCO Captions](image2.png)

Fig. 2: Distribution of image specificity scores for images in (a) VizWiz-Captions and (b) MSCOCO-Captions.
Caption Consistency (supplements Section 3.2)

We report the distributions of specificity scores that indicate the similarity between the five captions per image generated by different humans for all images in our VizWiz-Captions dataset and the MSCOCO-Captions validation set independently. Scores range between 0 and 1, with numbers closer to 1 indicating greater consistency between the five captions per image. Results are shown in Figure 2.

While the distributions are similar overall, we observe scores are skewed more towards 0 for VizWiz-Captions. We attribute these slightly greater annotation differences to annotators providing different level of detail and different types

Fig. 3: Examples of images that lead to a range of specificity scores when analyzing the captions collected from different crowdworkers.
of detail, as exemplified in Figure 3 in the main paper. We show examples of the diversity of captions that arise from different crowdworkers for a range of specificity scores in Figure 3.

3 Dataset Analysis (supplements Section 3.2)

3.1 Insufficient Quality Images

Figure 4 exemplifies images that were deemed insufficient or lower quality for captioning based on the agreement of five crowdworkers.

![Fig. 4: Examples of images that are unanimously labeled as insufficient quality to be captioned.](image)

We show examples of images that we deem medium or high difficulty based on the agreement of five crowdworkers who indicated the image is insufficient quality to generate a meaningful caption (i.e., 1-2 for medium and 3-4 for high difficulty) in Figure 5.

![Fig. 5: Examples of images that are labeled as insufficient quality by (a) 1-2 crowdworkers and (b) 3-4 crowdworkers.](image)
3.2 Caption Characterization

We visualize the most popular words included in the captions for each of the following word types analyzed in Table 1 of the main paper: nouns, verbs, and adjectives. We do so both for VizWiz-Captions and MSCOCO-Captions to support comparison to today’s mainstream captioning dataset. For each word type we show the most common 100 words in VizWiz-Captions and MSCOCO-Captions separately as well as the most common 100 words that are in VizWiz-Captions but not found in MSCOCO-Captions. Results for nouns, verbs, and adjectives are shown in Figures 6, 7, and 8 respectively. We observe in Figure 6 that many popular words focus on items from daily living such as ‘table’, ‘person’, ‘box’,
‘food’, and ‘monitor’. Nouns that are absent from MSCOCO-Captions and popular in VizWiz-Captions largely focus on text and numbers (Figure 6c), including ‘expiration’, ‘captcha’, ‘thermostat’, ‘password’, and ‘currency’.

We also quantify the extent to which people are present in the images, given the high prevalence in many mainstream vision datasets. When tallying how many images are captioned using words related to people (i.e., people, person, man, woman, child, hand, foot, torso), the result is 27.6% (i.e., 10,805/39,181). When applying a person detector [2], people are detected for 29.4% (i.e., 11,499/39,181) of images. We suspect this latter result is slightly larger than observed for captions because of person detections on background objects, such as newspaper or TV screens. We suspect crowdworkers found such person detections to be insufficiently salient to be described as part of the captions. Altogether, the relatively low prevalence of humans may in part be attributed to the fact that any images showing people’s faces were filtered from the publicly-shared dataset to preserve privacy. We suspect that the presence of people in our VizWiz-Captions compared to popular vision datasets will differ in that either (1) only parts of people appear in our images, such as only hands, legs, and torsos or (2) when the full body is shown, often it is because the person is on the cover of media (book, magazine, cd, dvd) or a product box.

We next report the percentage of overlap between the most common 3,000 words in VizWiz-Captions and the most common 3,000 words in MSCOCO-Captions for all words as well as with respect to each of the following word types: nouns, adjectives, and verbs. Results are shown in Table 1. The higher percentage across all words than for the different word types is likely because there are many common stopwords that are shared across both datasets that do not belong to each word type.

| words | nouns | adj | verbs |
|-------|-------|-----|-------|
| 54.4% | 45.1% | 31.6% | 42.8% |

Table 1: Percentage of overlap between most common 3,000 words in VizWiz-Captions and the most common 3,000 words in MSCOCO-Captions.

---

We employed a faster-rcnn model pretrained on COCO. It has a ResNeXt-101 as the backbone and Feature Pyramid Network (FPN) to deal with objects of different scales. We only used the “person” category out of the 80 categories. We filtered the detections with a threshold of 0.3, meaning we only counted a detection as valid if the confidence score is above 0.3.
Finally, we provide histograms showing the relative prevalence of categories found in the mainstream computer vision datasets versus our dataset for three image classification tasks: recognizing objects, scenes, and attributes. Results are shown in Figure 9, complementing those shown in the main paper in Figure 2. Exemplar images in our dataset that show concepts overlapping with those in the mainstream computer vision datasets are shown in Figures 10, 11, and 12.

Fig. 9: Parallel results to those in Figure 2 of the main paper, showing the fraction of all images in our VizWiz-Captions and popular vision datasets that contain each category for the following vision problems: (a) object recognition, (b) scene classification, and (c) attribute recognition.
Fig. 10: Examples of images showing visual concepts that are common in existing computer vision datasets for the object recognition task.
Fig. 11: Examples of images showing visual concepts that are common in existing computer vision datasets for the scene recognition task.
Fig. 12: Examples of images showing visual concepts that are common in existing computer vision datasets for the attribute classification task.
4 Algorithm Performance with Data Augmentation (supplements Section 4)

Given the need for improved algorithm performance for low quality images, we examined the potential for using data augmentation during training to help the models better cope with low quality images at test time. For this analysis, we chose the AoANet algorithm since it has the best performance when training from scratch. We re-trained this algorithm from scratch again using VizWiz-Captions, but this time augmented a copy of all the training images after blurring them using a 15x15 averaging filter kernel on all the training images.

Results are shown in Table 3. We observe that the performance is worse with data augmentation; e.g., with respect to the CIDEr score, overall performance falls from 60.5 to 56.2. We suspect this performance drop is because artificial image distortions are unsuitable for mimicking real-world quality issues [1], and thus distract from model training.

| B@1  | B@2  | B@3  | B@4  | METEOR | ROUGE | CIDEr | SPICE |
|------|------|------|------|--------|-------|-------|-------|
| AOANet |
| All  | 66.4 | 47.0 | 32.3 | 22.1   | 20.0  | 46.5  | 56.2  | 13.8  |
| Easy | 69.7 | 50.6 | 35.5 | 24.6   | 21.2  | 49.1  | 60.3  | 14.2  |
| Medium | 63.3 | 42.7 | 27.7 | 18.2   | 18.5  | 43.2  | 50.8  | 13.6  |
| Difficult | 36.3 | 19.8 | 11.0 | 6.4    | 12.3  | 29.7  | 40.3  | 11.3  |

Table 2: Analysis of the top-performing image captioning algorithm when trained from scratch using data augmentation of blurred images. Results are shown on all test images as well as only the subsets deemed easy, medium and difficult. (B@ = BLEU-)

We also report human performance based on the same set of evaluation metrics. To do so, we only consider images with five valid captions (i.e., “easy” images). We randomly choose one caption per image as the prediction, and use the remaining four captions for evaluation. Results are shown in Table 3.

| B@1  | B@2  | B@3  | B@4  | METEOR | ROUGE | CIDEr | SPICE |
|------|------|------|------|--------|-------|-------|-------|
| Easy | 60.3 | 40.7 | 27.7 | 18.8   | 22.0  | 43.4  | 83.5  | 17.5  |

Table 3: Human performance on all test images deemed easy. (B@ = BLEU-)

References

1. T.-Y. Chiu, Y. Zhao, and D. Gurari. Assessing image quality issues for real-world problems. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3646–3656, 2020.
2. S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015.