Exploring Variances in Image Captions by Crowdsourcing Work

Yiyan Huang
The University of Texas at Austin
huang.yiyan@utexas.edu

Te-Yuan Chen
The University of Texas at Austin
yuan0061@utexas.edu

Abstract

Image caption is the primary instrument in assisting blind and visually impaired people to understand images and their surrounding environments. Image captioning in computer vision generally depended on the captions annotated by workers, which, however, could vary from each other. This paper aims to address caption variances and use crowdsourcing platforms to prioritize key elements in these captions, in a bid to explore the information that needs to be included for generating satisfying and precise image captions. We screened out 100 images with differently annotated captions and invited Amazon Mechanical Turk workers to highlight and rank 5 elements they regarded important after reviewing images. We analyzed the data collected by their word classes and content denotation, with special attention paid to whether subjective description was highlighted and whether there is any difference in the annotations from workers who have been informed of the special application and purpose of this study. This paper hopes to provide a sample dataset of key elements in image captions, shed light into what information needs to be paid attention to in object recognition and detection, and share thoughts on whether the usage of image captions will modify the contents to be included in image captions.

1. Introduction

Motivation. Eyes not only enable us to recognize and detect objects, but also allow us to read and understand scenes. In a shared path, the development of computer vision enables communication grounded on images between AI and human beings, combining image understanding and natural language processing. In this process, image captioning, or automatically using words to describe an image, remains an important and heatedly discussed topic. By translating images into words, it could assist people to capture the image contents, speed up writing process or facilitate their interaction with voice-controlled assistants. Beyond that, considered as an attempt to answer the question of "What is this picture about?", image captioning can help blind and visually impaired people (BVIP) to gain insights about the surrounding situation and understand social media contents. As studies showed that BVIP placed great trust in AI-generated captions including those incorrect ones [11, 3], improving the precision and accuracy of image captioning could optimize BVIP's experience in understanding the world with automatically generated descriptions. VQA development could also benefit from good descriptions of images as a number of questions could be answered by the captions.

Previous work. Meanwhile, there are a myriad of challenges to improve the accuracy and precision of image captions automatically generated. Related previous work covered a wide scope spanning from forming image caption datasets (MSCOCO, VizWiz etc.) with the help of crowdsourcing platforms [4, 3], and generating image captions automatically by neural image caption, i.e. using RNN to generate sentences based on the classification of objects detected by computer vision [14]. Whilst the evaluation of image caption quality strongly relies on the annotations by crowdsourcing workers, not much work pays enough attention to the variance of captions annotated by human workers in image datasets, which can be prominent in terms of granularity, objects included and subjective speculation, etc., indicating variant verbal saliency (caption focus).

Reframing the variance issue. As image captioning is an endeavor to grasp the gist of an image and translate graphic information to verbal communication, a good caption should use verbal saliency to reflect visual saliency. Current algorithms used to generate captions, however, paid more attention to generate natural-toned captions in the verbal sense than aligning them with the visual focus conveyed by the images. In addition, object recognition or detection might not be sufficient for capturing the image saliency since vibe, atmosphere, weather, and background knowledge among other invisible components can hardly be recognized by computer vision. The variance in image captions and selecting out the most appropriate captions could be reframed into how to create verbal saliency in line with visual saliency.

Addressing the variance issue. We propose to identify the important elements in image captions by human annotators in line with the visual saliency in images, based on which we can realign image saliency and the verbal saliency in captions and help to generate descriptions closest to appropriateness and accuracy, which can be especially helpful for BVIP who heavily depend on image descriptions for receiving information[11]. Here, by using the image captions provided by VizWiz [3], a dataset of images shot by BVIP with questions asked by them, we propose to 1) screen out the images with various captions, 2) collect the visual elements deemed significant by people in the form of words, through asking crowdsourcing workers to prioritize the elements tokenized from the existing captions after looking at given images and captions, and 3) analyze data collected to provide a reference to the information needed for image captioning, especially those needed in addition to detected objects for understanding images. Specifically, we aim to collect the data by using Amazon Mechanical Turk to invite workers to highlight and prioritize the elements in these captions they deemed important.

2. Related Work

The collection of image caption. Microsoft COCO dataset [4] and Stair captions [15] collected image captions by using crowdsourcing platforms. When setting up AMT HIT tasks, the authors established a series of requirements to describe the image. Automated evaluation methods were used to assess the performance. In 2019, Shuster [13] added the element of personality to make the caption more human-like. Literature on caption collection provided the dataset of image captions and laid a foundation for future work. Nevertheless, while noticing there are variances in the captions collected, these work failed to discuss the reasons for the variance or set up criteria of what might be the most appropriate caption. Our work extended the previous work by exploring further reasons behind the variances and finding out what elements in a caption are deemed important by people.

Automatic caption generation and evaluation. Based on the captions collected in image datasets, work has been done to generate image captions automatically by using CNN to recognize objects and identify their locations in the picture, and then using RNN or LSTM to automatically generate the sentence[14, 5]. Furthermore, work was carried out in the fusion of all results to generate a high-quality image caption. In terms of evaluating the captions generated, metrics such as BLEU and ROUGE as well as crowdsourcing platforms were adopted. Beyond the n-gram featured metrics mentioned above, Andersen et al. proposed SPICE or semantic map in caption evaluation, which can really help improve performance on complex queries and image and video retrieval systems [1, 10, 12]. Nevertheless, in all these scenarios, the metrics strongly depended on the human-annotated captions for verification. Besides, these work focuses on extracting elements from the images to generate sentences, while neglecting the importance of producing captions that capture major components in the images which could be a future study direction. Studying on the caption variance and deciding on the verbal saliency that matches the images in this sense will help to improve the algorithm.

Image ambiguity. Previous work paid attention to the caption variance generated by people, based on which images are categorized into specific and ambiguous ones [7]. Further, research studied over the consistency between the way people view images and the way they describe them, i.e. visual saliency and verbal saliency to help distinguish specific and ambiguous images [8, 6]. Studies in this area found out that relatively clean images with few objects and usual activities are generally “specific” and usually had unanimous captions. For images with multiple objects the variances in describing the images grew. In cases where pictures contain humans and animals, they are more verbally salient, even though the inanimate objects might be the focus of pictures. In these cases, images engendered more diversity in captions annotated by human beings. Most pictures in VizWiz were taken by BVIP using their phones, indicating that these pictures can be “ambiguous” in containing several objects with less prominent foreground objects. In this context, the correlation between image saliency and verbal saliency might be low and it is worthwhile to delve into the captions generated and explore restructuring the verbal saliency to better understand the images.

3. Methods

As a caption aims to equivalently describe the information an image conveyed, a good caption should be constructed based on the image context. It requires an overall understanding of the image, not only describing the existing objects but also referring to those not in the image, such as a train that has not arrived, or introducing background information. Moreover, in a cluttered scene, the image is imperative to capture the key elements instead of everything [9, 2]. It should be comprehensive, concise and helpful in understanding the image.

With that in mind, this study will use Amazon Mechanical Turk to request help from crowdsourcing workers to highlight and rank the components they deemed most important among the 5 captions given to an image. Specifically, the crowdsourcing task is divided into the following steps:

1) Pre-selection of the images with different caption annotations. To guarantee the collection of good quality data and pay attention to the variance in image captions, we will pre-screen out image captions with prominent differences in human annotation. Image captions that met such criteria would be the ambiguous images with subjective or speculative descriptions, distinctive object description or various granularity in Figure 1. In line with the differences, the images are categorized into 5 classes, respectively, ”subjective description”, ”status”, ”background knowledge”, ”focus” and ”granularity”. The number of images in different categories is shown in Figure 2. Due to a time limit, this study will select 100 images with captions for performing this crowdsourcing task.

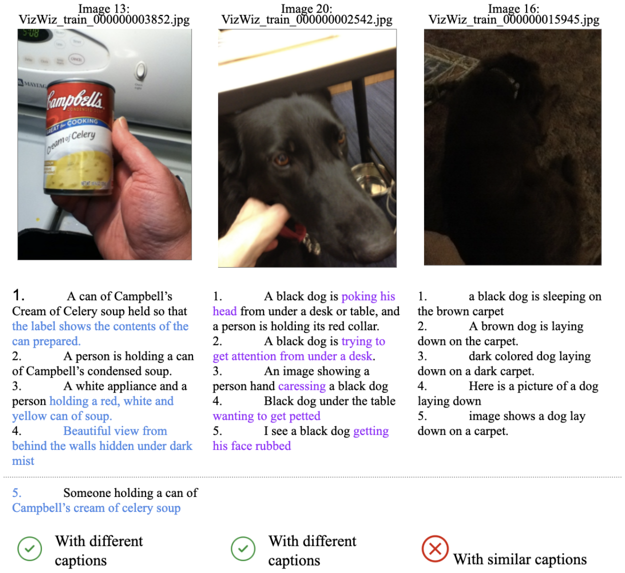


Image 13:
VizWiz_train_000000003852.jpg
Image 20:
VizWiz_train_000000002542.jpg
Image 16:
VizWiz_train_000000015945.jpg

1. A can of Campbell’s Cream of Celery soup held so that the label shows the contents of the can prepared.
2. A person is holding a can of Campbell’s condensed soup.
3. A white appliance and a person holding a red, white and yellow can of soup.
4. Beautiful view from behind the walls hidden under dark mist
5. Someone holding a can of Campbell’s cream of celery soup

✓ With different captions

1. A black dog is poking his head from under a desk or table, and a person is holding its red collar.
2. A black dog is trying to get attention from under a desk.
3. An image showing a person hand caressing a black dog
4. Black dog under the table wanting to get petted
5. I see a black dog getting his face rubbed

✓ With different captions

1. a black dog is sleeping on the brown carpet
2. A brown dog is laying down on the carpet.
3. dark colored dog laying down on a dark carpet.
4. Here is a picture of a dog laying down
5. image shows a dog lay down on a carpet.

☒ With similar captions

Figure 1. Examples of images with different (left 2) and similar (right 1) captions

2) Crowdsourcing task. This study intends to use AMT to collect elements that workers deemed most important after looking at the image. The basic procedure is to provide instructions on operation and good caption standard, and ask workers to highlight elements in the 5 captions provided with substantial meanings which they think are the most important ones for describing the image, choose the one they think most appropriately describe the image and provide a reason for their previous actions in Figure 3. Given that usually a complete sentence contains elements of subject, predicate, object, adjectives and attributes, here we require workers to highlight and rank 5 words or phrases based on their perceived importance to the image provided. For each set of captions or image we will have 3 different workers to make annotations.

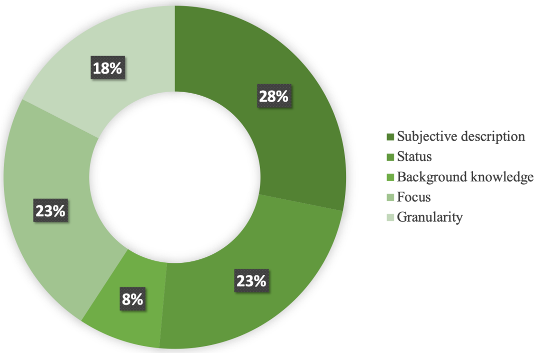


Figure 2. Categories of images with different captions

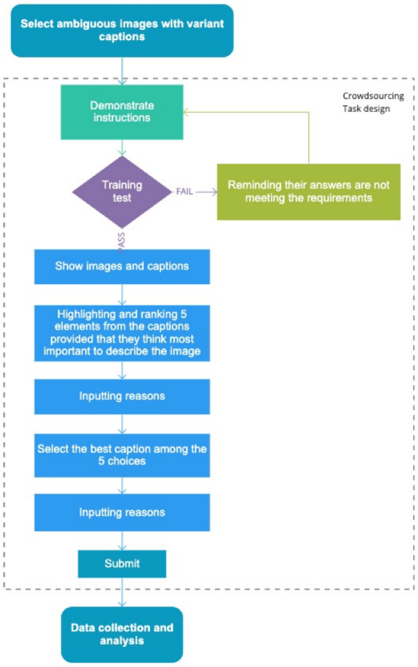

Select ambiguous images with variant captions
Crowdsourcing Task design
Demonstrate instructions
Training test
FAIL
Reminding their answers are not meeting the requirements
PASS
Show images and captions
Highlighting and ranking 5 elements from the captions provided that they think most important to describe the image
Inputting reasons
Select the best caption among the 5 choices
Inputting reasons
Submit
Data collection and analysis

Figure 3. Flow chart of the research method and crowdsourcing task

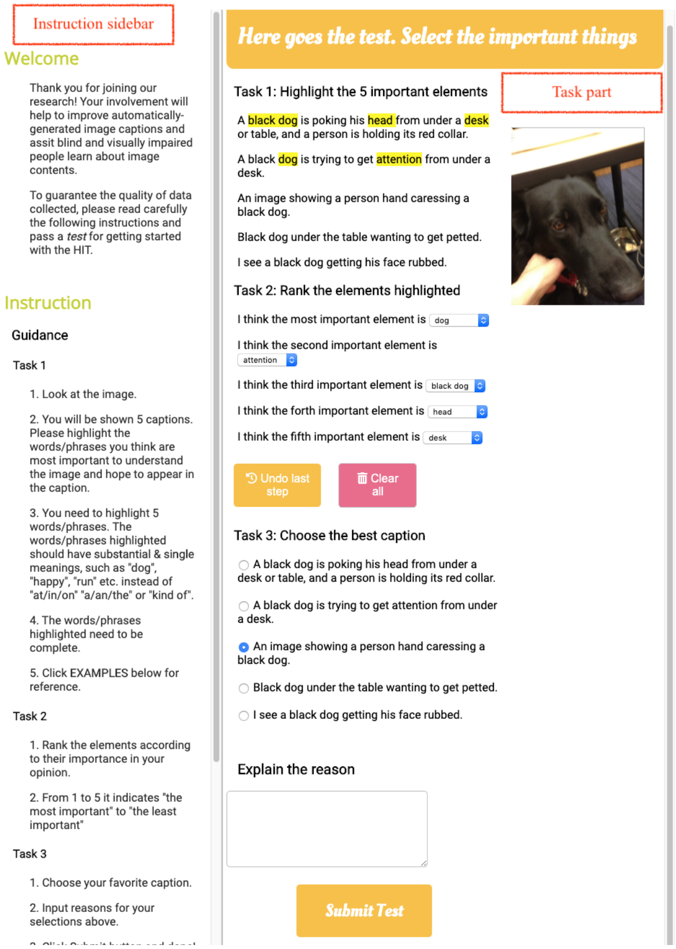
3) Improving efficiency and quality control measures. Making sure we collected complete words and avoiding arbitrary annotation are the two challenges we aim to address when designing the crowdsourcing interface. For the previous one, we deploy a system that allows crowdsourcing workers to highlight the words they deemed most important, to avoid typos from typing in the words and improve working efficiency. For the latter, workers need to pass a test before starting to work. The test presents the same test structure with the official one, only with one word which corresponds to the main object in the picture. If the answers submitted by the workers include this key word, they could start to work; otherwise, they will be asked to read the instructions and do the test again. In addition, based on our previous experience as a worker on AMT, we presented appreciation to all workers for their contribution to this work at the beginning of instruction part.

4. Experimental Design

Dataset. We use VizWiz[3], the dataset specifically created aiming to develop assistive technologies for BVIP to address their daily challenges. The dataset is composed of images and questions, respectively taken and verbally asked by BVIP. In addition, the dataset includes captions collected by UT Austin Image and Video Computing (IVC) Group, 5 captions on average for each image. This research mainly uses 100 sets of captions of VizWiz for data collection and analysis, with each image or set of captions annotated by 4 workers.

Experiment 1: Observe the difference of elements highlighted or ranked by workers who have (not) been informed that the captions are used for helping BVIP in understanding the image. The functionality of image captions makes them elastic to the purposes of usage. The purpose of this experiment is to explore whether there will be any difference in the elements highlighted and ranked if we specially inform workers the image was taken by and aimed to help BVIP to understand images. For the same set of captions, 2 workers (Group A) will be asked to perform tasks with instructions on the dataset usage and background while the rest 2 (Group B) will not. The work done by workers who hadn't been informed of the usage will serve as the baseline. As there is little literature carrying out the same experiment design before, this paper uses the quantitative method and qualitative method as evaluation metrics, i.e. measuring the number of words in the datasets of 2 groups, and using recall to measure the overlapping collection and differences annotated by Group A and B.

Experiment 2: Observe the value of subjective description in image captions. The purpose of this experiment is to examine the subjective description in image captions and whether information alike should be considered in image caption generation. Workers will be asked to highlight the elements deemed important from image captions with a mixture of subjective and objective descriptions, and analyze whether subjective descriptions are included in the elements limited to be highlighted. The evaluation metrics is the ratio of data entries with subjective words highlighted to the overall data entries under the image category of "subjective description". Image caption dataset collected previously, such as MSCOCO [4], requires workers to not include any speculative or subjective words when annotating the captions. Subjective descriptions are generally neglected in image captions. In this scenario, the baseline could be regarded as 0.



Welcome

Thank you for joining our research! Your involvement will help to improve automatically-generated image captions and assit blind and visually impaired people learn about image contents.

To guarantee the quality of data collected, please read carefully the following instructions and pass a *test* for getting started with the HIT.

Instruction

Guidance

Task 1

1. Look at the image.

2. You will be shown 5 captions. Please highlight the words/phrases you think are most important to understand the image and hope to appear in the caption.

3. You need to highlight 5 words/phrases. The words/phrases highlighted should have substantial & single meanings, such as "dog", "happy", "run" etc. instead of "at/in/on" "a/an/the" or "kind of".

4. The words/phrases highlighted need to be complete.

5. Click EXAMPLES below for reference.

Task 2

1. Rank the elements according to their importance in your opinion.

2. From 1 to 5 it indicates "the most important" to "the least important"

Task 3

1. Choose your favorite caption.

2. Input reasons for your selections above.

Here goes the test. Select the important things

Task 1: Highlight the 5 important elements

A black dog is poking his head from under a desk or table, and a person is holding its red collar.

A black dog is trying to get attention from under a desk.

An image showing a person hand caressing a black dog.

Black dog under the table wanting to get petted.

I see a black dog getting his face rubbed.

Task 2: Rank the elements highlighted

I think the most important element is dog

I think the second important element is attention

I think the third important element is black dog

I think the forth important element is head

I think the fifth important element is desk

Task 3: Choose the best caption

○ A black dog is poking his head from under a desk or table, and a person is holding its red collar.

○ A black dog is trying to get attention from under a desk.

● An image showing a person hand caressing a black dog.

○ Black dog under the table wanting to get petted.

○ I see a black dog getting his face rubbed.

Explain the reason

Figure 4. AMT task interface

5. Experimental Results

Main Finding. By tokenizing the words and phrases collected according to their word or phrase classes (noun, adjective, verb, adjective+noun, adverb and other) as well as the contents they denote (object, color, number, status, location, description and text information), we calculated the overall ratio of different word classes and content classes (see Figure 5 and Figure 6).

Based on the results shown in Figure 5, among the priority 5 elements in image captions, the majority is composed by nouns, followed by adverbs and adjectives. In terms of the contents, in a corresponding manner, objects accounted for the majority votes in the priority elements, followed by words/phrases indicating the current status, colors and text information. As the prevalent method of image caption auto-generation depends on object recognition and detection, the statistical results could shed some light on the significance of contents that object recognition paid inadequate attention (such as the status of objects) or failed to capture (like descriptive adjectives, or vibe) in image captions.

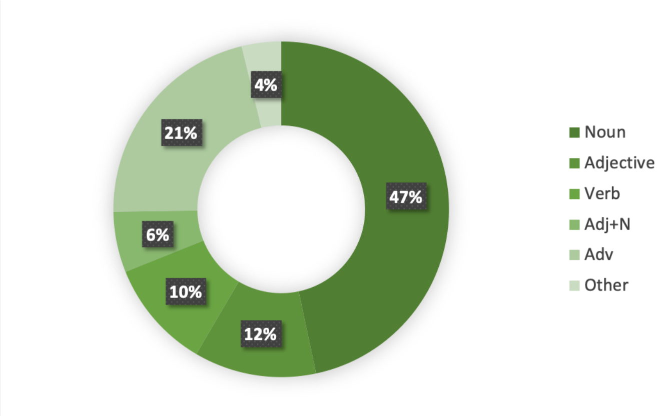


Figure 5. Word class analysis based on the data collected

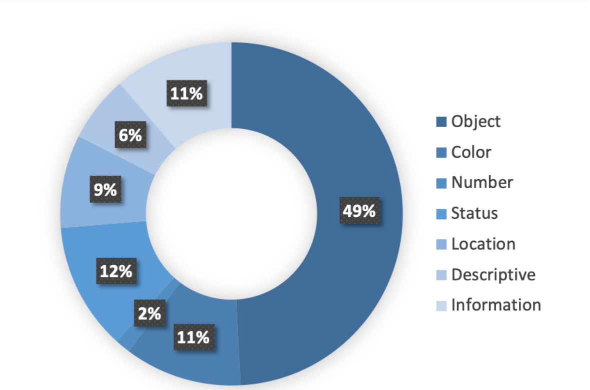


Figure 6. Content class analysis based on the data collected

To further improve the experiment, work could be done to compare the verbal saliency of annotated captions and the image saliency generated from the highlight of elements from crowdsourcing workers. It could serve as a future consideration for the evaluation metrics of image captioning after more studies carried out on the combination of elements highlighted by workers and adjusting the captions generated by RNN.

Element 1 to 5 stand for the most important element to the least important element highlighted by workers. Based on the tokenization above, it can be observed that for the most important element, nouns and objects are the primary category. Since the 2nd element, the significance of adjectives, verbs and adverbs increased. Correspondingly, in terms of contents, objects decreased compared to Element 1, whilst words denoting color, location and status increased from Element 2 to 5 (Figure 7 and Figure 8).

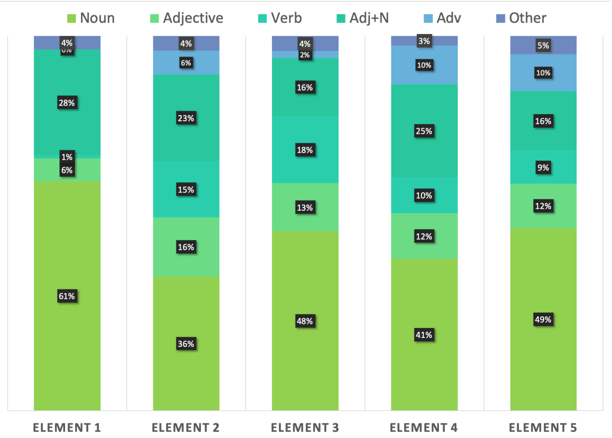


Figure 7. Ratio of word classes from Element 1 to 5

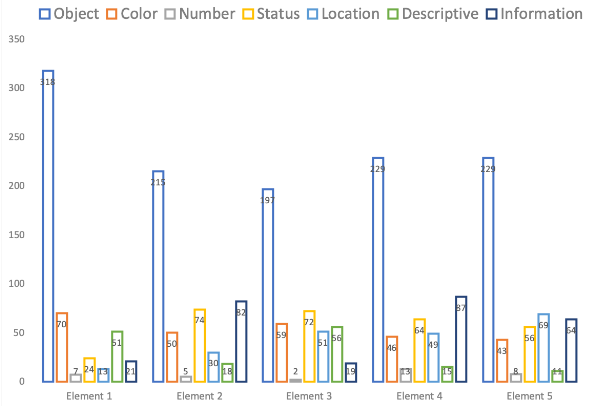


Figure 8. Content classes from Element 1 to 5

Experiment 1. We hypothesized that the data collected from workers informed that the captions were used for assisting BVIP to understand images would differ from those not informed, in ways that the former put more value on the text information, status of the objects (e.g. if it is a picture about a bottle of water, the elements highlighted might include the brand of water and remaining water status) and color among other information deemed specially important in the context of assisting BVIP. As mentioned above, two groups of workers separately added annotations on the same image. Workers in Group A are well informed that the pictures were taken by BVIP and this study would help them to understand images. Workers in Group B were shown no more information than operation instructions.

For quantitative analysis, to testify whether there is any variance, this paper calculated the number of words (N) in the elements annotated by Group A and B. Although we get N_{GroupA}/N_{GroupB} = 1.16, indicating that the annotations by Group A workers contain more information than Group B. Nonetheless, the t-test demonstrated no significant difference (p-value = 0.7376, >0.05).

For qualitative analysis, this paper used recall score, i.e. calculating the number of words that are both highlighted by Group A and B versus those annotated by any single group. As shown in the following lines, the recall values are 0.61 and 0.63 respectively for Group A and B, indicating variances in the elements annotated by 2 groups.

$$N_{GroupA \cap GroupB}/N_{GroupA} = 0.61 \quad (1)$$

$$N_{GroupA \cap GroupB}/N_{GroupB} = 0.63 \quad (2)$$

To explore what are the different elements between Group A and Group B, this paper carried out word class analysis and content analysis of annotations by each group after subtracting their overlapping group of words (Figure 9 and Figure 10). The word class analysis result shows that there are more nouns and adverbs in annotations from Group A. In terms of content analysis, the annotations from Group A had more information related to objects, numbers and locations.

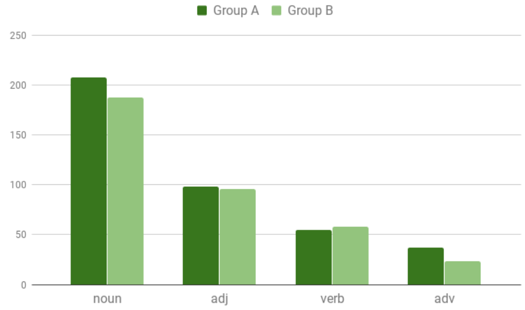


Figure 9. Analysis of word classes in Group A B annotations

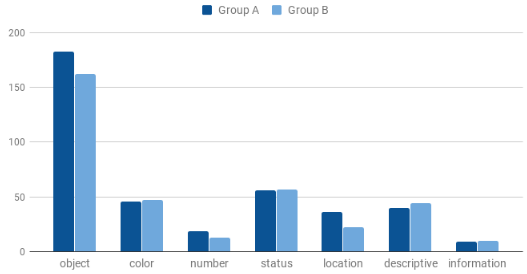


Figure 10. Analysis of content categories in Group A B annotations

In other words, there are variances in the annotations from workers who have been informed the task purpose and those not, but the variances are not quite prominent. In the case of captions for BVIP, generating image captions could pay more attention to providing information on number of objects and their locations. Though the findings failed to vigorously support the previous hypothesis, this paper holds that it is still important to inform workers about the study purpose as it found that the rate of qualified data from Group A was much higher than Group B during data collection.

To further improve the experiment, other variables could be added such as informing the caption is used for people with no visual impairments, or the caption is used for social media or news. The number of elements for highlighting could also be adjusted, to testify whether there will be more prominent differences if more elements were asked to be highlighted. Highlights might vary across different functions and provide further reference on the generation of appropriate and humane captions by algorithms.

Experiment 2. The results showed that among all the annotated work on images featuring different captions in terms of containing subjective or speculative description, about 38% highlighted elements containing subjective or speculative description as shown in Figure 11. Even in cases where people are required to only highlight limited elements, less than half of workers chose those with subjective descriptions. In terms of data collection in visual datasets, adding and collecting subjective description in image captions are also quite important and could help to generate more satisfying description of pictures. How to make sure subjective description could be "recognized" through machine learning accounts for an important issue to be addressed.

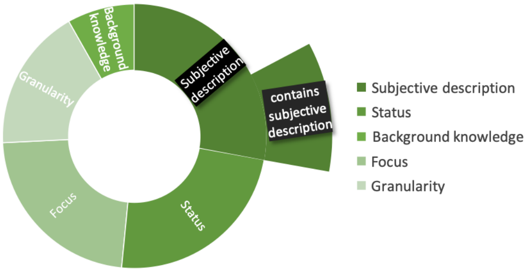


Figure 11. The ratio of annotations which highlighted subjective and speculative description as important elements

To further improve the experiment, there can be a research into what kinds of images are more easily annotated with subjective or speculative captions than others, and researchers could make a comparison between the findings above and whether human evaluators agree that the subjective terms are important.

6. Conclusion

Through our work, we hope to contribute to image caption generation in the following ways: 1) providing a sample dataset of key elements in image captions, in a bid to highlight visually salient elements in VizWiz pictures by words, 2) categorizing the variances in image captions and reframing it into the disconnection in visual saliency and verbal saliency, 3) exploring a possible way to address this issue, i.e. using crowdsourcing platforms to prioritize elements in captions they considered to be important after looking at images, 4) shedding light into what information needs to be paid attention to in object recognition and detection through analysis into the data collected, 5) reviewing whether subjective and speculative descriptions are needed in caption generation, and 6) sharing thoughts on factors that influence caption generation, including usage, and targeted audience.

By categorizing the elements highlighted by workers according to their word classes and content denotation, this paper finds that nouns and objects occupied the dominant significance in image caption generation, but the importance of verbs, adjectives and adverbs, or information related to status, location and description increased gradually from Element 2. While informing workers about the study purpose might not indicate prominent difference in their annotation, this paper suggests to provide workers with adequate information as deemed appropriate based on the annotation contents and data quality from 2 groups of workers. Last but not the least, for the development of image caption generation, attention should also be paid to include subjective description, which was either neglected in image caption collection for visual dataset, or could not be included due to the limitations of object recognition and detection in previous work.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

[2] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016.

[3] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342. ACM, 2010.

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[5] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

[6] D. Gurari, K. He, B. Xiong, J. Zhang, M. Sameki, S. D. Jain, S. Sclaroff, M. Betke, and K. Grauman. Predicting foreground object ambiguity and efficiently crowdsourcing the segmentation (s). *International Journal of Computer Vision*, 126(7):714–730, 2018.

[7] M. Jas and D. Parikh. Image specificity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2727–2736, 2015.

[8] H. Liang, M. Jiang, R. Liang, and Q. Zhao. Visual-verbal consistency of image saliency. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3489–3494. IEEE, 2017.

[9] H. Liang, M. Jiang, R. Liang, and Q. Zhao. Capvis: Toward better understanding of visual-verbal saliency consistency. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(1):10, 2018.

[10] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2657–2664, 2014.

[11] H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999. ACM, 2017.

[12] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.

[13] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston. Engaging image captioning via personality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12516–12526, 2019.

[14] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[15] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*, 2017.