WHY AI ALT TEXT GENERATOR FAIL

A Spec Project

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

MSc. Information Science

by

Jiahui Zhou Jiaxin Guo

May 2020

ABSTRACT

We did the research on the reason why AI Alt generator fail. We collect the data from Twitter and label the data manually. Based on the data set we got from Twitter, we defined what is failure and formed categories that explains why AI alt generator could fail. The result could be used to provide the direction of improvement for AI Alt generator. This is not a thesis, but the specialization project in the thesis form required by the Cornell Tech curriculum.

ACKNOWLEDGEMENT

It is a great time to work with my classmate Jiaxin and my advisor Prof. Shiri at this project. This is my last project at school. A lot of thanks to all my teachers and people who helped me.

Thanks to Cornell Tech. Time passing is really fast here, especially when I was starting my own business. I kept coding all the time. I would say thanks to all the people who have teamed up with me, who were not mad at my absence all the time.

CHAPTER 1

INTRODUCTION

Alt text is used to explain the image when the image is absent in the website because of network or some other reasons. Alt text is also read to visually disabled people when they browse the website. For this reason, there are a lot of alt text to generate in the website. Scientist create AI alt text generator to do the work. However most of the time, it cannot convey the real meaning of the image. We are trying to figure out the reason why AI alt generator fail to convey the right meaning of images and provide directions for the AI alt text generator to improve. Based on the improvement, people can better understand what is supposed to be in the website and visually disabled people can better understand the web or mobile text content.

CHAPTER 2

RELATED WORK

Image analysis algorithms is currently one of the most widely used tools in computer vision field. However, most of previous work[2][4] in visual recognition is about labeling images with a pre-trained set of visual categories. The task of dense image annotations[5] and generating descriptions for images [3] has also been explored.

Unlike this prior work on improving precise of object recognition, we focus on generating descriptions based on context to help especially visually impaired people better understand the images. Some prior research worked in generating alt text automatically to help people consume photo based information. For

example, CaptionBot[1] is a tool developed by Microsoft which can understand the content of images and generate descriptions as well as human. To help visually impaired people select and share photographs from their albums on social networking services, Yuhang et al. [6] incorporated state-of-the-art computer-generated alt text into Facebook's photo-sharing feature.

CHAPTER 3
[METHOD]

We first signed up a new Twitter account. We pick the account we want to follow arbitrarily by using random number, which is generated by random function, as our stride. After we got the feed from the account we chose to follow. We used the same method to choose the tweet we want to use for the data set. We only pick the tweet with images. If the chose tweet do not contain image according to the stride, we would choose the next tweet. This process keep rolling until we get the tweet with image.

We set the image as the data source and the text of the tweet would be the correct label. We then manually generate our label using the Caption bot provided by Microsoft.

Then we manually compared the label generated by us with the correct label, which is the text of the tweet. We then do the inductive analysis, which is to generate themes and use them to create theories/narratives and draw conclusions in the end.

CHAPTER 4

RESULTS

After the inductive analysis, the categories we defined are shown in the following picture. For 100 alt text generated, 5 percent are categorized as totally incorrect and 95 percent are categorized as partially correct.

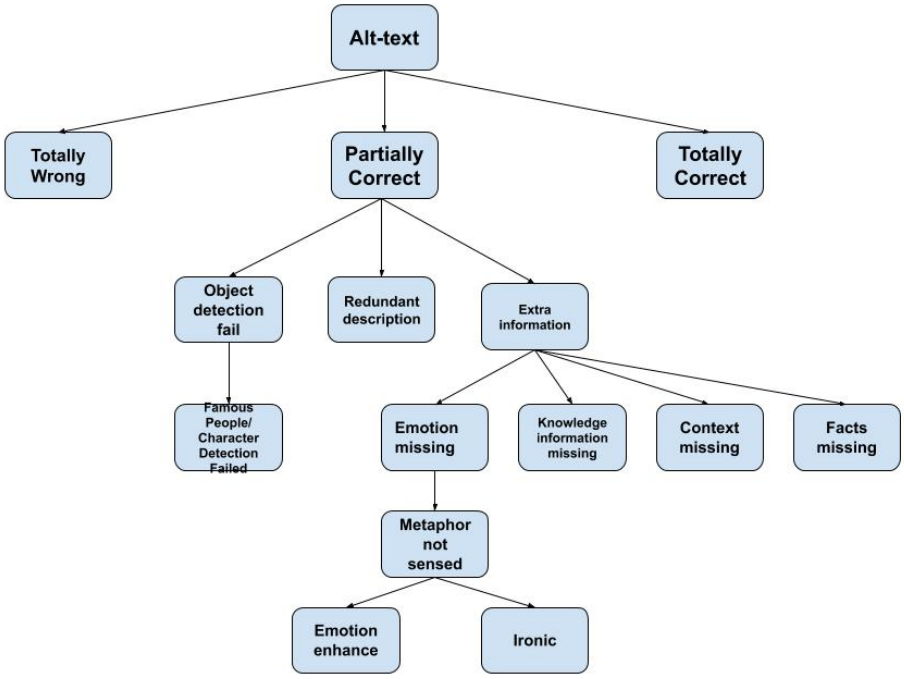

Figure 4.1: Categories defined according to the results

In this part, we will include examples for each sub-category.

For object detection failure, the example tweet is ”Rare deer-like animal rediscovered in Vietnam after 30 years” and the alt text generated is ”It’s a brown bear walking through a forest”.

Figure 4.2: Example image for object detection failure

For famous people/character detection failure, the example tweet is ”In December, Michelle and I will head to Malaysia for the first @ObamaFoundation Leaders: Asia-Pacific gathering. The region plays an important role in my own story, and today.” and the alt text generated is ” It’s a group of people sitting at a table.”

Figure 4.3: Example image for famous people/character detection failure

For redundant description category, the example tweet is ”Here’s a close up of Bei Bei’s ice cake! It featured some of his fav treats like sugar cane and sweet potato!” and the alt text generated is ”A bunch of stuffed animals that are on a table”. For facts information missing category, the example tweet is ”In

Figure 4.4: Example image for redundant description

December, Michelle and I will head to Malaysia for the first @ObamaFoundation Leaders: Asia-Pacific gathering. The region plays an important role in my own story, and today.” and the alt text generated is ” It’s a group of people sitting at a table.”

Figure 4.5: Example image for facts information missing category

For knowledge information missing category, the example tweet is ”BeiBei is the 3rd giant panda born at the Zoo to move to China. Tai Shan moved in 2010 and Bao Bao moved in 2017. All of them will participate in the giant panda breeding program.” and the alt text generated is ”A close up of a stuffed animal on a table”.

Figure 4.6: Example image for knowledge information missing category

For context information missing category, the example tweet is ”Happy birthday to my favorite source of inspiration, Mickey Mouse! 91 has never looked so good.” and the alt text generated is ”A woman standing in front of a mirror holding a teddy bear”.

Figure 4.7: Example image for context information missing category

For positive emotion missing category, the example tweet is ”Puppies are welcome to join the @dogagingproject CitizenScience Dogs OpenScience OpenData” and the alt text generated is ”A person holding a dog.”

Figure 4.8: Example image for positive emotion missing category

For ironic emotion missing category, the example tweet is ”We got a new video baby monitor and I think that was a mistake” and the alt text generated is ”It’s a screenshot of a computer monitor sitting on top of a television.”

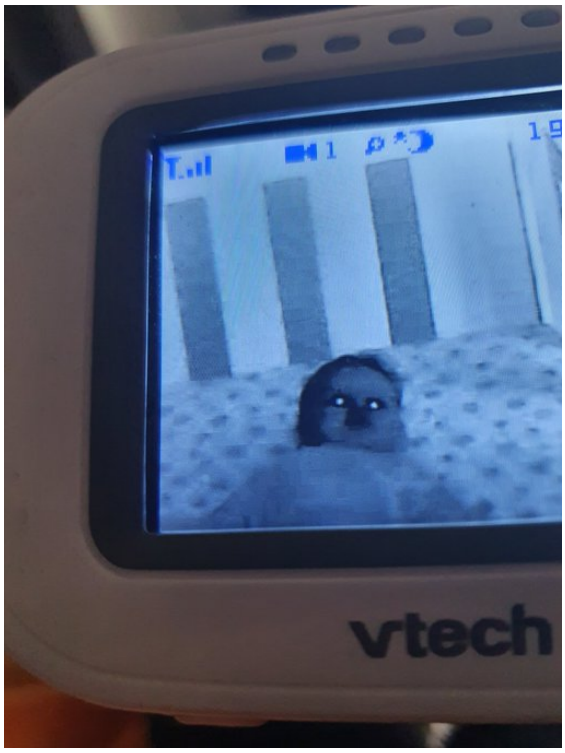

Figure 4.9: Example image for ironic emotion missing category

CHAPTER 5

DISCUSSION

We defined totally incorrect as all the elements related to expressing the meaning of image is missing, which includes object detection failure, objects spatial relationship failure, context missing, facts missing and emotion missing etc. This case is rare because the AI bot can always recognize some objects in the image. Another extreme case is that the AI bot get all the things right, which is also rare according to our observation. As a result, we focus on analyzing the partially correct category. In this category we divide the subcategory to be information missing, objects detection failure, spatial detection failure. Each of these categories has been explained in the results. After observation, we think these three are the most common mistakes that made by the AI bot. Under the in-

formation missing category, we further divide this as fact, knowledge, context and emotion missing. Fact is the general common things people agree with, knowledge is the things people learned from the facts. Context is the relationship between everything including the knowledge and facts. Emotion is the feeling of people, which can be further divided into ironic emotion and positive emotion.

After comparison our label with the correct label, we think it is very hard for the caption bot to produce the correct label because people put a lot of subjective things in the tweet. Emotion towards the same image can be different. People are using metaphor and ironic words as well. Even human self cannot really guess the meaning of the image. Further work should be done on improving the results.

5.1 Limitations

As we have mentioned, although all the data can be included in our formed categories, we think it is hard for the AI bot to improve by following our directions right now. We know nowadays the AI bots can only detect objects. Sentiment analysis is done with other AI bots. However according to the technical limitation, the sentiment analysis now is very basic. Context, fact, knowledge and complex emotion is not possible to be analyzed using the current AI technology. However, our work is still useful to provide instructive improvement direction for the development of AI alt text generator in the future.

5.2 Future work

Further work can be done in automatically scraping the tweet and forming the data to usable data set. Right now, we did the data collection manually due to time limitation. Also we manually label the data by using Caption bot from Microsoft. This could also be done automatically by scripts.

As we have mentioned, now the tweets contain a lot of subjective feelings. Considering the current technology we have in machine learning, we think it is better to use the human description of the image as the correct label, which contains the objects and the spatial relationship of the objects most of the time.

CHAPTER 6
CONCLUSION

We used Twitter and AI alt text generator as sources and explored the performance of the alt text generation. We did inductive analysis with these data and coded them into several categorises. As a result, the majority alt text generated turned to be partially correct with object detection failure, redundant description, context missing, facts missing and emotion missing etc. We think it is very hard for the AI generator to produce the correct alt text since people put a lot of subjective things in the tweet. Context, fact, knowledge and complex emotion, which is hard to guess and explore by even human, is not possible to be analyzed using the current computer vision technology. However, our work is still useful to provide instructive improvement direction for the development of AI alt text generator in the future.

BIBLIOGRAPHY

[1] Captionbot. https://www.captionbot.ai/. Accessed: 2019-10-08.

[2] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88:303–338, 2010.

[3] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[5] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 966–973. IEEE, 2010.

[6] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):121, 2017.