Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Demonstrations Track

SaGol: Using MiniGPT-4 to Generate Alt Text for Improving Image Accessibility

Yunseo Moon , **Hyunmin Lee** , **SeungYoung Oh** and **Hyunggu Jung**
University of Seoul, Seoul, Republic of Korea
{copel0317, yunoa64, dhtmddud08, hjung}@uos.ac.kr

Abstract

SaGol is an AI-powered application to improve image accessibility for people with visual impairments (PVI) users. Alternative (alt) text, a general method of web accessibility for PVI users, is text or phrases that describe images on a website in an understandable way. SaGol generates alt text with the images on the user's smartphone using a vision large language model called MiniGPT-4. SaGol searches for similar images based on the generated alt text. We evaluated the length of alt text and the search accuracy. This paper shows a potential opportunity to improve image accessibility for PVI users.

1 Introduction

Alternative (alt) text, a general method of web accessibility for people with visual impairments (PVI) users, is text or phrases that describe images on a website in an understandable way [James Edwards *et al.*, 2021]. According to the W3C's Web Content Accessibility Guidelines 2.1, alt text "serves the equivalent purpose" of the non-text content [Kirkpatrick *et al.*, 2018]. However, many content creators tend not to create alt text for their content because they forgot to write [Jung *et al.*, 2022; James Edwards *et al.*, 2021]. Therefore, it is important to generate alt text automatically.

Previous studies employed various methods to create alt text, including the human-based method, image-search-based method, and AI-based method. Several studies utilized human-based methods to create alt text [Zhang *et al.*, 2022; Gleason *et al.*, 2019]. For instance, Zhang et al. developed a tool to generate alt text for GIFs by crowdsourcing alt text written by several people and choosing the best [Zhang *et al.*, 2022]. This method has the advantage of producing alt text easily understood by humans. Other studies searched the image for the alt text that had already been made and crawled the text [Guinness *et al.*, 2018; Pareddy *et al.*, 2019]. For instance, Guinness et al. developed a tool that generates alt text for users by searching for a given image to look for existing alt text [Guinness *et al.*, 2018]. This method has the advantage of providing quality-assured alt text without generating a new alt text. Additionally, some studies employed artificial intelligence (AI) models to generate alt text [Cho and Kim, 2023; Gleason *et al.*, 2020; Jobin *et al.*, 2022]. For example, Jobin K. V. et al. developed a tool that creates alt text on lecture slides using an optical character recognition module [Jobin *et al.*, 2022]. This method has the advantage of being cheaper than having a human-generated alt text. However, all three methods have the disadvantage of being unable to reduce costs and maintain quality simultaneously. Also, these methods did not use a large language model (LLM) to generate alt text In addition, little is known about how to develop and evaluate tools for generating alt text in smartphone images and searching those images to assist PVI users. For example, images captured or downloaded with Android smartphones often lack alt text. Also, accessing images via Talkback only offers date information, rendering them inaccessible to blind individuals on Android smartphones.

To address this gap, we developed SaGol with two major functions. First, SaGol generates detailed alt text as it helps PVI users understand images specifically [Jung *et al.*, 2022]. Second, SaGol searches for images using the generated alt text. This is because existing studies did not create a tool which generates alt text for images on smartphones. With these two features, SaGol helps PVI users easily understand and select images on their smartphones.

2 System Overview

We developed SaGol, an AI-powered image alt text generation and search application, to improve image accessibility for PVI users. SaGol 1) uses an LLM, MiniGPT-4, to generate alt text of images, and 2) searches for images based on the generated alt text [Zhu *et al.*, 2023] (see Figure 1). SaGol helps users access the images on their smartphones.

2.1 Generate Alt text

For PVI users who have a challenge understanding images on their smartphones, We created a tool to generate alt text for images. First, SaGol loads images from the user's smartphone and uploads them to the AI server. Second, SaGol sends the loaded images to the AI server. The AI server generates the alt text for each image by inputting the question "compare similarity of vectors" along with the image to the MiniGPT-4 model. Third, the AI server sends the generated alt text to SaGol. SaGol stores the alt text on the smartphone.

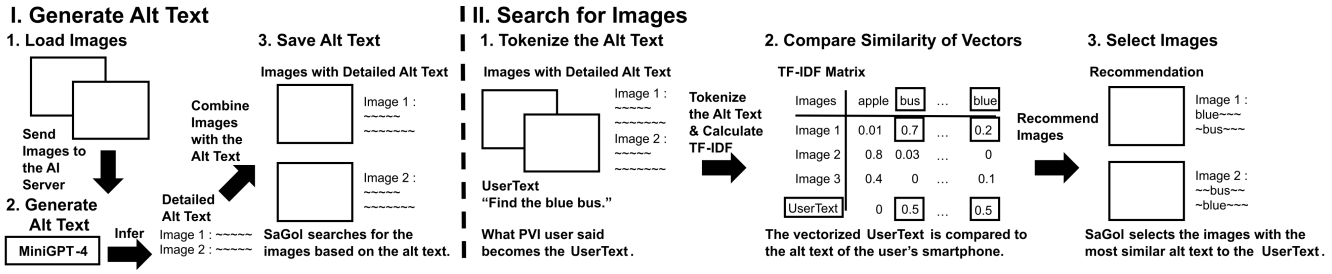

Figure 1: System Overview. SaGol 1) generates the alt text of images by loading images and saving alt text on the user's smartphone. SaGol 2) searches for images through tokenization and comparing similarity based on the alt text generated in 1).

2.2 Search for Images

For PVI users facing a challenge of selecting images on smartphones, we offer image search by alt text (see Figure 1).

Tokenize the Alt Text SaGol uses the NLTK module [Loper and Bird, 2002] for alt text tokenization and creates a Term Frequency-Inverse Document Frequency (TF-IDF) matrix from the tokenized alt text and the search term.

Compare Similarity of Vectors SaGol vectorizes the user's search term of the user into a vector regarding the TF-IDF matrix. SaGol compares the vector with other alt text vectors in the TF-IDF matrix using cosine similarity. Alt text vectors with high similarity contain many words in the user's question, so the more likely it is to answer the user's question.

Select Images Based on the similarity comparing results, SaGol selects images with alt text similar to the search term. The results screen shows the user with images with the most similar alt text to the search term. The user is allowed to get a detailed alt text of the image in voice.

3 Demonstration

SaGol, a smartphone image search application, improves image accessibility for PVI users by generating alt text and searching images based on the generated alt text. In addition, SaGol is designed for use with TalkBack, an accessibility tool for VBI users. SaGol works as follows (see Figure 2). Users initiate SaGol by clicking "Upload Images," which triggers an AI model to generate alt text for images on their smartphone. This step is only required the first time. After alt text are generated and saved, users enter search terms via voice recognition or manual typing. After clicking the search button, SaGol displays the 10 most similar images based on the generated alt text. Users can select an image using accessibility tools or direct clicks, prompting SaGol to provide a detailed alt text. By listening to the alt text of the images in the search results, users can confirm that they have found the image they are looking for. Both voice and alt text are accessible via TalkBack. SaGol not only benefits PVI users but also extends support to those searching for alt text or performing smartphone image search.

Usage Scenario 1 James, a 32-year-old visually impaired novelist known for his intricate descriptions, was looking for a suitable image of a white bicycle for his new novel cover. To overcome his PVI, he used SaGol to perform the voice image search for "white bicycle." SaGol provided detailed alt text for each image, including elements, such as surrounding houses, dogs, and trees. This allowed James to gather information and moods associated with the images, effectively overcoming his PVI and achieving success in his work.

	Word counts	t-value	p-value
MiniGPT-4	110.13±50.23	33.86	3.31×10^{-141}
BLIP	11.85±1.51		

Table 1: Average alt text word counts with two vision LLMs.

Usage Scenario 2 Aliyah, a 35-year-old blind teacher who supports students with special needs, shares images with parents to show their children's learning activities and achievements. In particular, she highlights the images of cultural events or school festivals. To search for the images she wants, Aliyah says "school events" in SaGol. The application then searches for and delivers detailed alt text of various cultural events or school festivals, allowing Aliyah to vividly share her students' activities with parents, despite her PVI.

4 System Evaluation

The evaluation of our system was based on the length of alt text and the accuracy of the search. To perform the evaluation, we used the COCO dataset, which is a commonly used image dataset for object detection, segmentation, and captioning [Veit *et al.*, 2016]. We randomly selected 300 images without duplicates to create a test set. The selection process was based on a normal distribution derived from the label distribution of the COCO dataset. The test set was used as input for several models to generate alt text for the images. We then measured the length of alt text and the accuracy of the image search using our test set.

4.1 Alt Text Word Count

To measure the alt text word count, we used MiniGPT-4 and BLIP to describe our sampled image dataset [Li *et al.*, 2022; Zhu *et al.*, 2023]. BLIP, a high-performance Vision-Language Pre-training model, excels in image-text tasks using noisy web data, while MiniGPT-4 demonstrates advanced multimodal capabilities by generating detailed alt text. The text generated by each model was statistically analyzed for length differences (see Table 1). MiniGPT-4 had an average

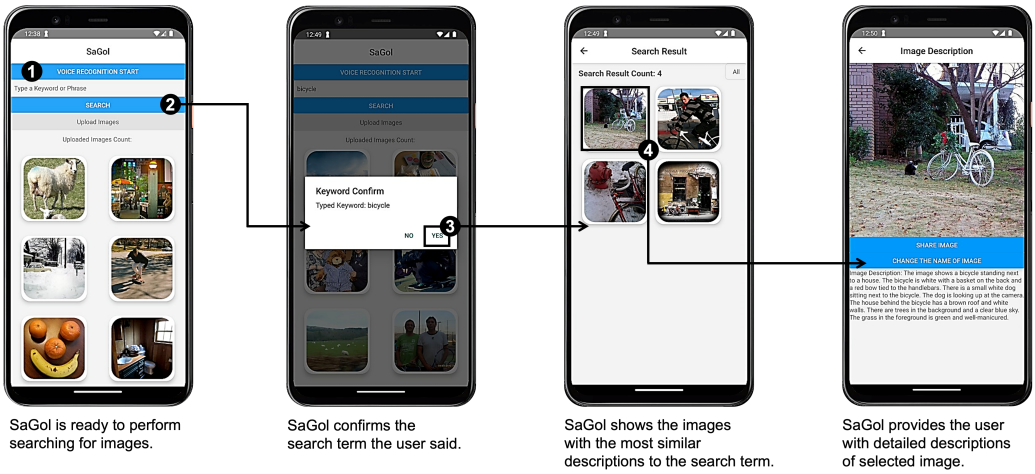

Figure 2: Prototype of the SaGol application. 1) The user selects the voice recognition start button (using an accessibility tool or by pressing it directly) and say the search keyword. 2) When the user presses the search button, SaGol asks the user to confirm the search term before searching. 3) If the user confirms the keywords and continues with the search, SaGol shows the user a list of images that are similar to the keywords. 4) The user can select the images and get detailed alt text.

Model	Accuracy (%)
VGG16 [Simonyan and Zisserman, 2015]	13.3
ResNet152 [He *et al.*, 2016]	16.3
MobileNetV3Small [Howard *et al.*, 2017]	18.0
InceptionV3Small [Szegedy *et al.*, 2016]	24.3
BLIP [Li *et al.*, 2022]	64.3
MiniGPT-4 [Zhu *et al.*, 2023]	68.7

Table 2: Four object recognition models and two vision LLMs were utilized to search for the image. Accuracy calculation of the searching was done based on top-10.

word count of 110.13, compared to BLIP's 11.85. A t-test yielded a t-value of 33.86 and a p-value of 3.31×10^{-141}. Since the p-value was less than 0.05, we rejected the null hypothesis and confirmed that MiniGPT-4's average word count was greater than BLIP's average word count.

4.2 Image Search Accuracy

Based on the TF-IDF, we performed the image search of alt text generated by several object recognition models and LLMs and compared their accuracy to determine the alt text generation model for SaGol (see Table 2) [Chollet and others, 2015]. To perform the image search, we used 300 images from the COCO dataset mentioned in subsection 4.1. Alt text for these images was generated using four object recognition models and two vision LLMs. We used the same 300 images as our evaluation dataset, and the evaluation process included the following steps: 1) select images and labels from the dataset; 2) tokenize the generated alt text and the search term; 3) create a TF-IDF matrix using the tokens and the search term [Aizawa, 2003]; 4) identify the 10 images with the highest cosine similarity to the search term in the TF-IDF matrix; 5) count as "found" if any of the 10 images is identical to the image used in the search, for accuracy calculation; 6) repeat steps 1) through 5) for all 300 images.

For evaluation, we included four object detection models: VGG16, ResNet152, MobileNetV3Small, and InceptionV3Small with the first two models [Simonyan and Zisserman, 2015; He *et al.*, 2016; Howard *et al.*, 2017; Li *et al.*, 2022]. MiniGPT-4 showed the highest accuracy (see Table 2). Out of 300 images, 206 images were found with MiniGPT-4's alt text, while 193 were found with BLIP's alt text.

5 Conclusion

In this study, we developed and evaluated a smartphone application that allows users to search and understand images on their smartphones. However, our study has several limitations. First, while we evaluated the application using 300 images from the COCO dataset, we did not assess the usability and the alt text with the user's images. Moreover, despite the fact that we utilized the MiniGPT-4 model to generate alt text, we only evaluated two LLMs to select a model for generating alt text. Additionally, although we performed the image search using the alt text generated by the models, we did not perform any accuracy enhancements, such as fine-tuning and setting a threshold for the search. Furthermore, we used the prompt "Describe the image" as the default in the model, but no evaluations were conducted with different prompts.

Accordingly, we propose the following recommendations for future research. First, we should assess the usability and the accuracy of the image search with PVI users to identify areas for improvement. Moreover, we need to evaluate other LLMs to find more appropriate ways of generating alt text, thereby improving SaGol's performance. Plus, we ought to improve the alt text and search accuracy by fine-tuning, setting thresholds and exchanging the model with the existing LLM API. In addition, it is important to provide different prompts to improve accessibility for PVI users of different ages, nationalities, genders and education levels. Our research aims to enhance image accessibility for PVI users by generating detailed and low-cost alt text through SaGol.

References

[Aizawa, 2003] Akiko N. Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing Management*, 39(1):45–65, 2003.

[Cho and Kim, 2023] Jaemin Cho and Hee Jae Kim. Dimensional alt text: Enhancing spatial understanding through dimensional layering of image descriptions for screen reader users. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery.

[Chollet and others, 2015] François Chollet et al. Keras. https://keras.io, 2015. Accessed: 2024-02-07.

[Gleason *et al.*, 2019] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. Making memes accessible. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 367–376, New York, NY, USA, 2019. Association for Computing Machinery.

[Gleason *et al.*, 2020] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. Twitter a11y: A browser extension to make twitter images accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[Guinness *et al.*, 2018] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. Caption crawler: Enabling reusable alternative text descriptions using reverse image search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–11, New York, NY, USA, 2018. Association for Computing Machinery.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[Howard *et al.*, 2017] Andrew Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[James Edwards *et al.*, 2021] Emory James Edwards, Kyle Lewis Polster, Isabel Tuason, Emily Blank, Michael Gilbert, and Stacy Branham. "that's in the eye of the beholder": Layers of interpretation in image descriptions for fictional representations of people with disabilities. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA, 2021. Association for Computing Machinery.

[Jobin *et al.*, 2022] K. V. Jobin, Ajoy Mondal, and C. V. Jawahar. Classroom slide narration system. In Balasubramanian Raman, Subrahmanyam Murala, Ananda Chowdhury, Abhinav Dhall, and Puneet Goyal, editors, *Computer Vision and Image Processing*, pages 135–146, Cham, 2022. Springer International Publishing.

[Jung *et al.*, 2022] Crescentia Jung, Shubham Mehta, Atharva Kulkarni, Yuhang Zhao, and Yea-Seul Kim. Communicating visualizations without visuals: Investigation of visualization alternative text for people with visual impairments. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1095–1105, 2022.

[Kirkpatrick *et al.*, 2018] Andrew Kirkpatrick, Joshue O. Connor, Alastair Campbell, and Michael Cooper. Web content accessibility guidelines (wcag) 2.1. https://www.w3.org/TR/WCAG21/, 2018. Accessed: 2024-02-04.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 17–23 Jul 2022.

[Loper and Bird, 2002] Edward Loper and Steven Bird. Nltk: The natural language toolkit, 2002.

[Pareddy *et al.*, 2019] Sujeath Pareddy, Anhong Guo, and Jeffrey P. Bigham. X-ray: Screenshot accessibility via embedded metadata. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 389–395, New York, NY, USA, 2019. Association for Computing Machinery.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[Veit *et al.*, 2016] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images, 2016.

[Zhang *et al.*, 2022] Mingrui Ray Zhang, Mingyuan Zhong, and Jacob O. Wobbrock. Ga11y: An automated gif annotation system for visually impaired users. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.