Korean Localization of Visual Question Answering for Blind People

Jin-Hwa Kim * **Soohyun Lim** * **Jaesun Park** **Hansu Cho**
SK T-Brain
jnhwkim,kathylim05,jayden_park,hansu.cho@sktbrain.com

https://sktbrain.github.io/KVQA

Abstract

The advancement of computer vision and natural language processing increases the possibility of assisting people with visual impairment upon their verbal commands. At 2018 ECCV, VizWiz Grand Challenge: Answering Visual Questions from Blind People initiated this effort by questioning how advanced technology can contribute to a higher standard of human rights by providing an inclusive platform. Here, we extend its effort by inviting Korean blind people to create a Korean version of the VizWiz dataset. We collect the images taken by Korean blind people and the corresponding questions reflecting the local context. In contrast to the 7 years collection of 31K VizWiz, we expect to collect 100K dataset within 6 months and benchmark the dataset using one of state-of-the-art visual question answering models. We hope that this project will contribute to social welfare by providing an appropriate solution to make our society more equitable and sustainable for all.

1 Introduction

The United Nations adopted Sustainable Development Goals (SDGs) [13] defining the 17 global agenda to be achieved by all member states until 2030. Within SDGs, each goal touches upon different issues including education, environmental sustainability, and partnership among diverse sectors that the international community should foremost consider. Although many goals are interrelated and inter-explainable by supporting each goal's purpose, one of the principles underlying SDGs originated from the issue of human rights [12].

Oftentimes, surrounding circumstances can affect not only the persons' way of thinking but also the attitude in reaction to the specific situation. For this reason, it is not always easy to consider the situation from other perspectives, especially the view from those who are socially marginalized or have a disadvantage in society. Although human rights do not necessarily imply the same state for every person, according to the Universal Declaration of Human Rights [11], every human being is born free and entitled to all rights regardless of his or her race, color, sex or another status. In this vein, we approached to use an existing Visual Question Answering (VQA) [1] technology to assist people with disabilities, so they can experience the convenience as many other people enjoy. VQA understands the provided pictures and if a person asks questions about them, it provides an answer after understanding the image via natural language.

At 2018 ECCV, we participated in the VizWiz Grand Challenge: Answering Visual Questions from Blind People [5]. VizWiz is the first goal-oriented VQA dataset collected from natural VQA settings from the blind. The dataset consists of more than 31,000 visual questions created by blind people who took photos and asked about the pictures as well as crowdsourced 10 answers per each visual

[*] equal contribution, a coin toss determined the order

AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

question. Since the images were taken by blind people, many of them have blurred with poor quality, so provide a reason to carefully design and revise existing VQA models. For this situation with a poor quality image taken by blind people, we used Bilinear Attention Networks [6] which consider every interaction of detected objects and questioning words, and achieved the 2nd place in this challenge.

Building upon this achievement, in 2019 we initiated a project to collect VQA dataset created by Korean blind people. Starting from this April, we have been collaborating with a Korean social enterprise to collect pictures taken within Korean setting and questions related to those images. So far, more than 7K sets of image and answer were collected and by the end of this year, we expect to collect up to 100,000 sets to train our Korean VQA (KVQA) model. From this project, we expect further engagement by blind people in Korea to achieve a higher standard of human rights as well as create a stepping stone for disseminating AI technology in a friendly way to Korean society.

2 Method

We benchmark the VizWiz dataset [5] creation process; however, one of the challenging criteria is time. In the case of VizWiz, they first released VizWiz application in May 2011 and collected a final set of 31,173 visual questions from whom agreed to the terms of anonymous sharing as of June 2018 (publication date). Whereas, our target period for data collection is only six months while maintaining the quality of data as well as the diversity of answer types. In the following section, we share our strategy for and speculation from data collection process and the details of KVQA focusing on the difference from VizWiz ,which is in English. We clarify that our protocol used in data collection was reviewed by a law firm and tried our best to follow a code of ethics to avoid any possible infringement of privacy.

2.1 Subject

We have recruited 143 blind people (58% male, 41% female, as of August 2019) from regional welfare centers, schools, and unions for the blind, research institute and braille libraries, and keep recruiting more participants. We reward each participant, who consents to our protocol, with $0.083 (100 Won) for a valid pair of a captured image and a corresponding question by limiting up to 5,000 questions per participant and observe a long-tail distribution from the number of submissions. Currently, the maximum number of pairs collected from a participant is 500.

2.2 Collecting device and application

We develop a data collection tool for Android and iOS to take advantage of the camera-equipped mobile devices and observe that 60% participants use iOS and 40% participants use Android. We confirm that the average resolution of camera used by our participants is higher than the case from VizWiz (4.5MP vs. 2MP).

Participants require to input a limited set of personal information and agree to our consent [2] to proceed. After capturing an image using our application, participants should record a verbal question and send the pair of image and recorded verbal question to our server. Then, the verbal question is transcribed by qualified representatives. We request to have a single answer for each question in order to avoid ambiguity.

2.3 Annotation

We collect 10 answers from 10 different annotators per question following the previous works [1, 5]. In contrast to VQA [1] and VizWiz [5], which used Amazon Mechanical Turk (AMT) interface to collect answers, we develop a web-based annotation system. We diversify the annotators considering gender and age distributions and limit 1,000 annotations per annotator.

[2] The consent is subject to *Privacy Act Article 18*, restrictions on use and provision of personal information other than the purpose of personal information by Ministry of Public Administration and Security, Republic of Korea.

Table 1: The scores using 5-fold cross validation on KVQA dataset for the Korean word embedding models. The standard deviations are reported after ± using five times of 5-fold cross validation. We followed [1] to evaluate the models.

Embedding	Dimension	All	Yes/No	Number	Other	Unanswerable
Word2vec [8, 9]	200	37.23 ± 0.11	**66.95**	20.47	20.08	**93.57**
GloVe [10, 7]	100	37.91 ± 0.08	65.98	20.76	21.97	93.18
fastText [3, 9]	200	**38.16 ± 0.13**	66.05	**20.79**	**22.45**	92.72
BERT [4]	768	37.95 ± 0.10	63.77	20.46	22.35	92.92

2.4 Anonymizing and filtering

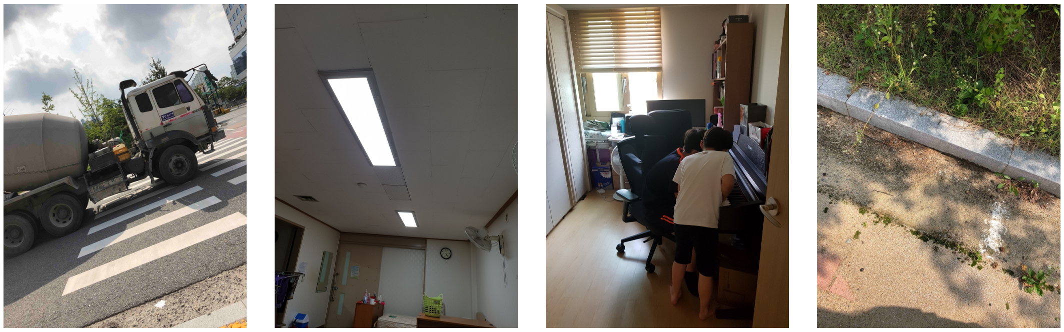
Protecting privacy and safety of participants is especially important, since "they often make the trade-off to reveal personal information to a stranger in exchange for assistance" [2, 5]. We notify that submissions will be excluded if the image contains specific individual or location, adult content, or any personal information. We eliminate the metadata from collected images to preempt privacy exposure but leave the rotating information, so we can track down the characteristics of original image if we need them.

2.5 Post processing

We correct simple syntax errors. We leave the natural appearances of polite expression in Korean, which enrich the diversity of question forms.

2.6 Statistics as of October 2019

So far, we have collected 30,031 pairs of image and question, and 300,310 answers. Notice that for our experiment, we use the current set of data. KVQA dataset consists of four types of answers, *Yes/No* (6.74%), *Number* (6.76%), *Other* (68.17%), and *Unanswerable* (18.33%). We also consider the number of examples for each answer type to be even, so that the distribution is more regularized than the one observed in Gurari et al. [5].



(a) **Q**: 지금 횡단보도를 건너도 될까? (Can I cross the crosswalk now?) **A**: 아니오 (No)

(b) **Q**: 이 방에는 몇 개의 형광등이 있나요? (How many lights in this room?) **A**: 2

(c) **Q**: 방에 있는 사람은 지금 뭘하고 있지? (What is the person doing in this room?) **A**: 피아노 (Piano)

(d) **Q**: 무슨 꽃이 피어있지? (What kind of flower is this?) **A**: Unanswerable

Figure 1: Examples of KVQA dataset. The most frequent answers are shown for each question. The above examples are image-question pairs of *Yes/No*, *Number*, *Other*, and *Unanswerable* type from left to right.

2.7 Model

We use one of the off-the-shelf state-of-the-art VQA models, BAN [6], as our baseline model. The hidden size and the number of glimpses are 512 and 8 respectively. We fix the word embeddings to increase the performance. We follow [6] for the hyperparameters, such as learning rate scheduling. To evaluate the VQA performance, we use 5-fold cross validation.

3 Results

Table 1 shows the performance of VQA models for different word embeddings. Among the word embeddings, fastText [3, 9] achieved the best performance. Especially, the fastText-based model shows the lowest accuracy in unanswerable type. The accuracy for each answer type is variant depending on the question embedding modules.

4 Discussions

By the end of this year, we plan to publish the collected KVQA dataset for future research work. We will first allow the license of using the dataset for research purpose only and then monitor the impact afterward since it is not yet fully tested whether the dataset can cause major concerns for public use. In 2020, we plan to organize a challenge with the dataset in Korea to find out innovative approaches for further enhancement of VQA technology. Through this opportunity, we can not only improve the performance of KVQA model to assist blind people but also publish AI technology in Korean society as a friendly tool, so people can perceive that AI, instead of replacing their jobs and threatening future life, can be a mechanism to provide an inclusive platform especially for blind people.

However, we still need to consider potential risks when publishing the dataset for diverse purpose. First of all, since this dataset was solely produced by blind people, unexpected bias and imbalance within the dataset can limit the proper training of VQA model and affect its accuracy. In this case, we suggest collecting more data applying the technique of incremental learning, as our group has a solid foundation to further engage with more blind people through diverse platforms. Second, if the dataset will be opened for commercial use, the original purpose of providing an opportunity to people with disabilities for convenient lives can be tainted, as many enterprises can target them as business customers.

5 Conclusions

We acknowledge that this project can lead both positive as well as negative effects. However, we strongly believe that the positive impact of KVQA dataset will surpass the negative implication by reaching forward to assist the lives of people with disabilities. For this reason, we sincerely hope that this project can contribute to the achievement of SDGs by providing the necessary solution to make our society more equitable and sustainable for all.

Acknowledgments

We would like to thank TestWorks for the collaboration in collecting Korean VQA dataset and Jaehyun Jun for the assistance in our KVQA model experiment.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2017.

[2] Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3523–3532. ACM, 2015.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[5] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *IEEE Computer Vision and Pattern Recognition*, 2018. URL `http://arxiv.org/abs/1802.08218`.

[6] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31*, pages 1571–1581, 2018.

[7] Gichang Lee. Embedding for Korean. URL `https://ratsgo.github.io/embedding`.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

[9] Kyubyong Park. Pre-trained word vectors of 30+ languages. URL `https://github.com/Kyubyong/wordvectors`.

[10] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

[11] UN. Universal Declaration of Human Rights, . URL `https://www.un.org/en/universal-declaration-human-rights/`.

[12] UN. United Nations Human Rights, . URL `https://www.un.org/en/sections/issues-depth/human-rights/`.

[13] UN. United Nations Sustainable Development Goals, . URL `https://sustainabledevelopment.un.org/?menu=1300`.