Korean Localization of Visual Question Answering for Blind People

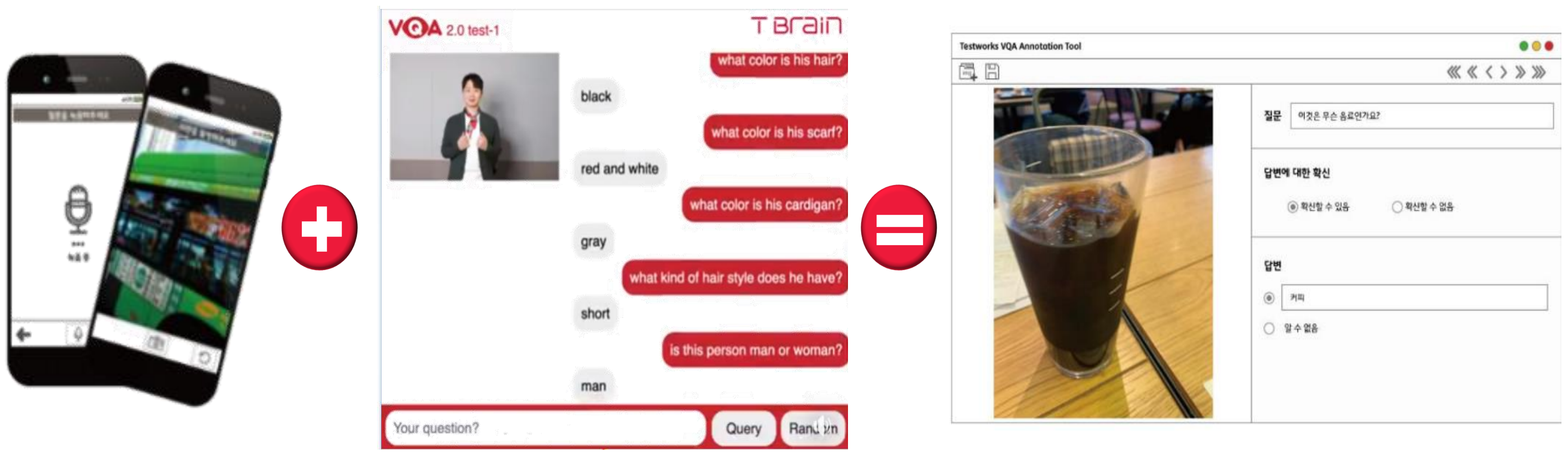
Jin-Hwa Kim* Soohyun Lim* Jaesun Park Hansu Cho

SK T-Brain, Republic of Korea

Introduction

The United Nations adopted Sustainable Development Goals (SDGs) to be achieved by all member states until 2030. One of the principles underlying SDGs originated from the issue of human rights. According to the Universal Declaration of Human Rights, 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) technology to assist people with disabilities. At 2018 ECCV, we participated in the VizWiz Grand Challenge: Answering Visual Questions from Blind People and achieved 2nd place with our Bilinear Attention Networks (BAN). In 2019, we initiated a project to collect VQA Dataset created by Korean Blind.



Methodology

We benchmark the VizWiz dataset creation process. Whereas, our target period for data collection is six months while maintaining the quality of data as well as the diversity of answer types. Our protocol used in data collection was reviewed by a law firm to avoid infringement of privacy.

Subject

We 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. We reward each participant for a valid pair of a captured image and corresponding questions.

Collecting device and application

We develop a data collection tool for Android and iOS to take advantage of the camera-equipped mobile devices.

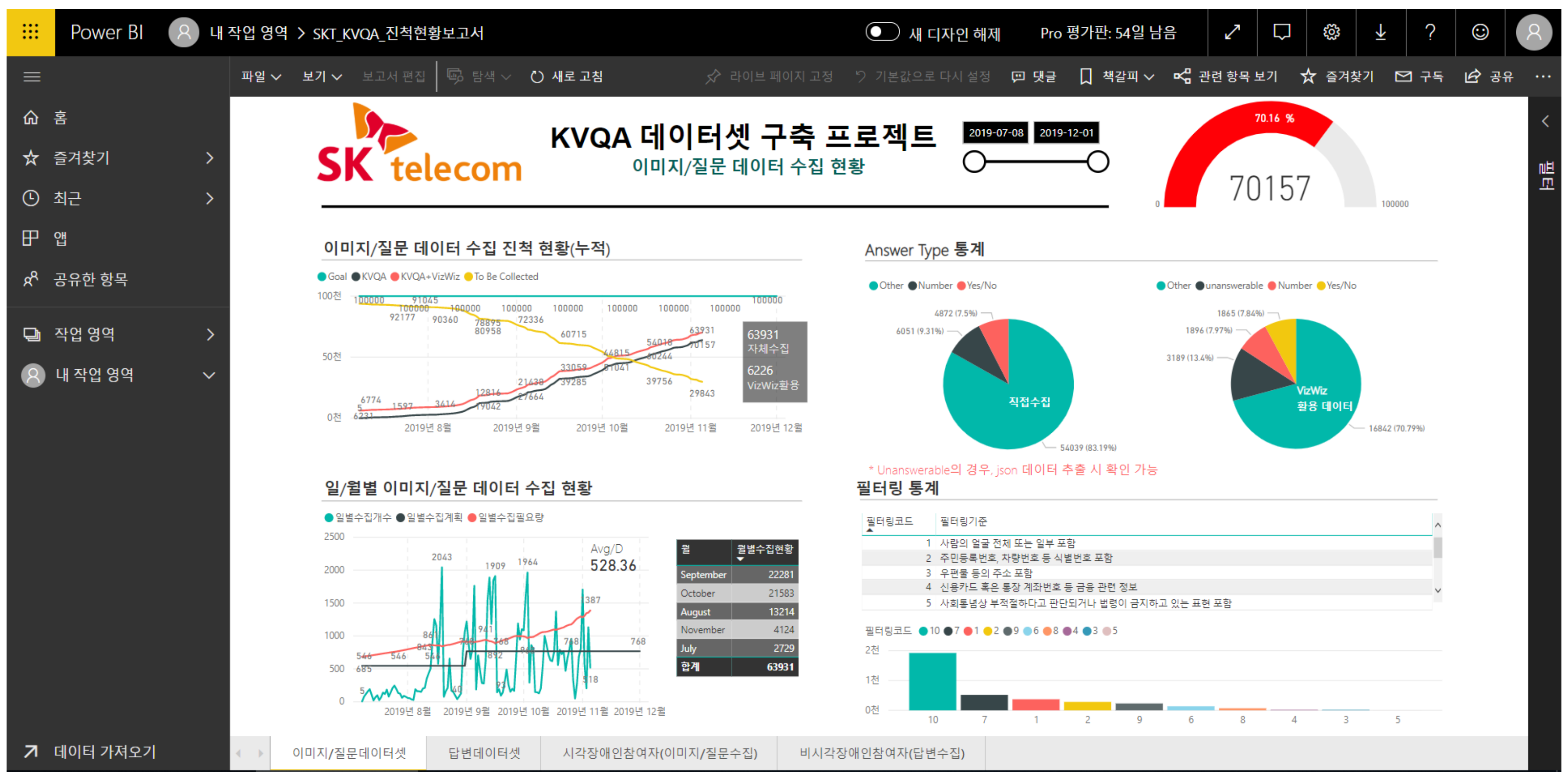
Participants require to input personal information and agree to our consent to proceed. We request to have a single answer for each question in order to avoid ambiguity.

Annotation

We collect 10 answers from 10 different annotators per question, using a web-based annotation system.

Anonymizing and filtering

We notify that submissions will be excluded if the image contains specific individual or location, adult content, or any personal information. We also eliminate the metadata from collected images to prevent privacy exposure.



KVQA 데이터셋 구축 프로젝트

이미지/질문 데이터 수집 현황

이미지/질문 데이터 수집 진척 현황(누적)

Answer Type 통계

* Unanswerable의 경우, json 데이터 추출 시 확인 가능

일/월별 이미지/질문 데이터 수집 현황

월	월별수집현황
September	22281
October	21583
August	13214
November	4124
July	2729
합계	63931

필터링 통계

필터링코드	필터링기준
1	사람의 얼굴 전체 또는 일부 포함
2	주민등록번호, 차량번호 등 식별번호 포함
3	우편물 등의 주소 포함
4	신용카드 혹은 통장 계좌번호 등 금융 관련 정보
5	사회통념상 부적절하다고 판단되거나 법령이 금지하고 있는 표현 포함

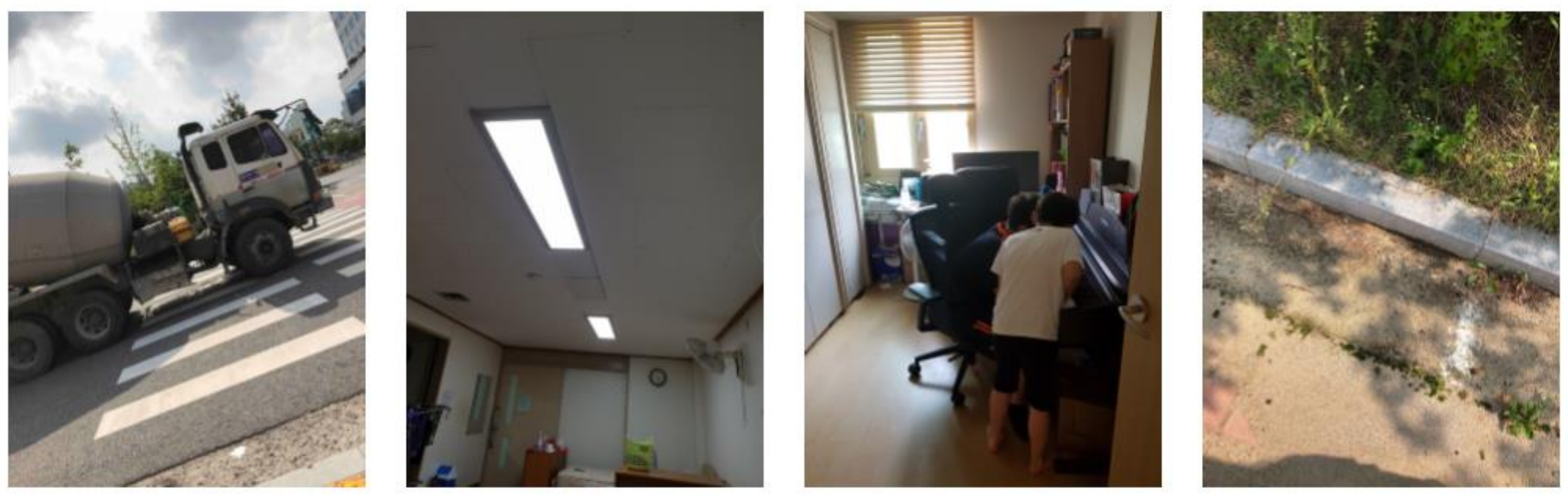
KVQA Dataset Collection Statistics Dashboard (As of Nov 14)

Post processing

We correct simple syntax errors and leave the polite expression in Korean for the diversity of question forms.

Statistics as of October 2019

We have collected 30,031 pairs of image and question, and 300,310 answers. The dataset consists of four types of answers, Yes/No (6.74%), Number (6.76%), Other (68.71%) and Unanswerable (18.33%).



(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.

Model

We use BAN as our baseline model. The hidden size and the number of glimpses are 512 and 8 respectively. To evaluate the VQA performance, we use 5-fold cross validation.

Results

The table shows the performance of VQA models for different word embeddings.

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

Among the word embeddings, fastText (Piotr et al., 2017) achieved the best performance.

Discussions

1. KVQA dataset was solely produced by blind; isn't there any subject bias or data imbalance within the dataset that can possibly limit the proper training of VQA model?

2. If the dataset is opened for commercial use, how far the original purpose of sharing social good to people with disabilities be tainted?

3. What level of AI technological advancement is expected, so that people with disabilities use the technology in spite of privacy tradeoff?

References

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):4–31,127 2017.

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.

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

UN. Universal Declaration of Human Rights

UN. United Nations Sustainable Development Goals.

AI for Social Good Workshop at Neural Information Processing Systems (NeurIPS), December 14th 2019, Vancouver Convention Center, Vancouver, Canada