Crowdsource by Google: A Platform for Collecting Inclusive and Representative Machine Learning Data

Supheakmungkol Sarin, Knot Pipatsrisawat, Khiem Pham, Anurag Batra, Luís Valente
Google AI
{mungkol, thammaknot, khiem, pocketaces, lvalente}@google.com

Abstract

This demo paper presents *Crowdsource by Google*, a platform for collecting inclusive and representative machine learning data to build AI products for everyone. *Crowdsource by Google* is enjoyed by our global community of passionate individuals who care about their languages and cultures and understand the need for diversity in machine learning and AI. In this paper, we discuss our design principles when it comes to our users: delightful experience, respect, and transparency. These principles make contributing data to *Crowdsource by Google* an open, trusting, and enjoyable experience for users. One of our early impacts includes creating an open-source dataset called Open Images Extended with over 478,000 images across 6,000+ categories from 70+ countries. This dataset increased representation of the Indian subcontinent by an estimated 250% in the original Open Images dataset.

Introduction

Collecting and labeling data for machine learning at scale is a challenge. Among many issues, two critical ones are (1) collecting data for training and evaluating representative and inclusive AIs to make sure that the end-products will work for everyone without bias, and (2) collecting data for long-tailed use cases which often require setting up expensive in-country operations.

We can tackle these two issues with crowdsourcing by partnering with the global user community. *Crowdsource by Google*, presented herein, is our attempt at this approach. Many existing crowdsourcing systems focus on collecting data, but operate on a much smaller scale (Phuttharak and Loke 2018; Mahmud and Aris 2015).

Our platform focuses on the users instead. It is designed to be delightful, respectful, and transparent to the user. Thus far, users have completed 300 million tasks and have contributed over 1 million images.

Crowdsource by Google

System Description

Crowdsource by Google is a web [1] and Android app [2] which Google launched in 2016. When users go to our app or web application's home screen, they are presented with different task categories, such as Image Label Verification, Sentiment Evaluation, Handwriting Recognition, and Translation. For each category, users are asked a series of simple tasks. For Image Label Verification tasks, for example, users are asked to check if the image is labelled correctly. The app and web versions allow us to serve both mobile and web users, depending on the task design and data requirements. Figure 1 (A) shows the home screen UI of our Android app and web applications. Each tile represents a category of tasks. Figure 1 (B) shows an example of Image Label Verification task.

Community of Users

Our community of passionate individuals spans the globe. They are students, working professionals, educators, homemakers, technology enthusiasts, and more. They are as diverse as the homelands, cultures and languages that they represent. There are 3 million global users representing 190 countries. All of these users are our invaluable partners in building a more inclusive AI.

Design

Crowdsource by Google adheres to three design principles namely, delightful experience, respect, and transparency.

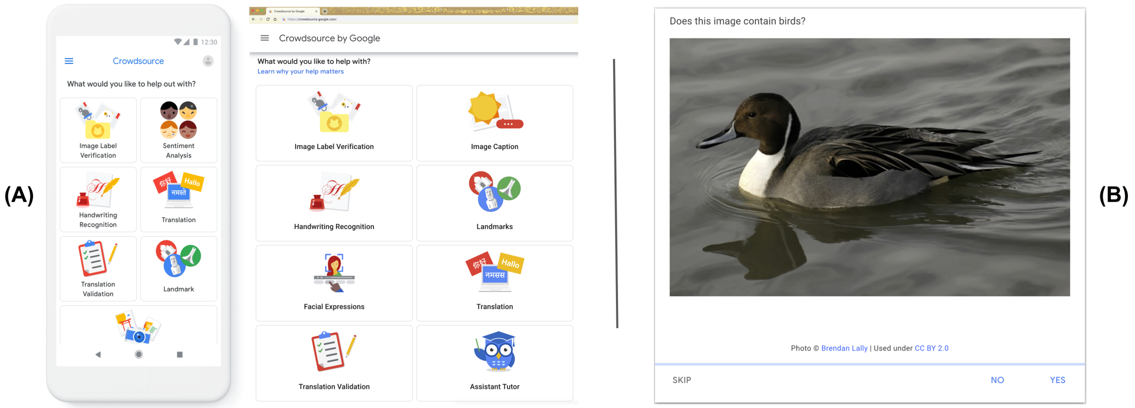
Delightful experience

(a) Gamification
We provide users with incentives and challenges to motivate them to make more contributions. Levels, Badges and Leaderboards make up our reward system.

- *Levels*: Users are promoted to the next level when they meet the required number of contributions. There are 18 levels, and different benefits can be unlocked when a user achieves these milestones. For instance, users who reach

Presented in the Work in Progress and Demo track, HCOMP 2019.

[1] Web: https://crowdsource.google.com
[2] Android App: https://crowdsource.app



Does this image contain birds?

Photo © Brendan Lally | Used under CC BY 2.0

Figure 1: *(A): Crowdsource by Google* Android App (left) and Web (right), (B): An Example of Image Label Verification Task

Level 6 are able to view the accuracy of their contributions and the number of times their answers have been upvoted by other users.

- *Badges*: Badges are tied to a variety of triggers such as volume of tasks, streaks, versatility, and location. Each badge serves as a mission for users to accomplish. There are 18 badges offered as awards. For instance, the Explorer Badge is awarded for completing 10 tasks in 3 different categories.
- *Leaderboards*: Leaderboards allow users to engage in some friendly competition with our global community.

(b) Task Simplicity
Tasks are designed to be short and simple with a clear focus. They shouldn't require user training. Below is a list of the current task categories:

- *Handwriting Recognition*: Identify and type out a word from a handwriting sample.
- *Image Capture*: Capture and share images from the surrounding environment.
- *Image Label Verification*: Verify whether an image matches a given label.
- *Image Caption*: Verify whether an image matches the caption.
- *Landmarks*: Find a specific landmark in a photo.
- *Sentiment Evaluation*: Identify the author's sentiment in a segment of text.
- *Translation*: Translate a sentence.
- *Translation Validation*: Validate translations of a given sentence.
- *Facial Expressions*: Identify the emotion(s) of the expression shown in a video.
- *Assistant Tutor*: Validate the naturalness and accuracy of a given phrase.

Respect and Transparency

Crowdsource by Google respects users' privacy and wants to be transparent throughout their entire journey. Users will always be asked for their explicit consent. For example, for the Image Capture task, users need to give separate consents on (1) allowing *Crowdsource by Google* to use their phone camera to take photos, (2) allowing Google to use the uploaded photos, and (3) allowing the photos to be open-sourced. To protect users' privacy, *Crowdsource by Google* removes all detailed location information stored in images before uploading them to its servers. We only store IP-based country information where the photo is uploaded, which helps us ensure the photo database shows representation of a global population. In addition, to protect the privacy of people depicted in photos, *Crowdsource by Google* blurs faces before uploading them. Moreover, users can always review images they have contributed and have them deleted from Google's database.

Impact

Data gathered through *Crowdsource by Google* has been used across many of Google's AI products, from Google Lens to Google Photos and Google Translate. We have open-sourced a subset of these images under Creative Commons CC-BY 4.0 licenses. The dataset is called "Open Images Extended (OIX)." It is meant to complement the core Open Images dataset (Krasin et al. 2017). The OIX is a dataset composed of over 478,000 images across 6,000+ categories from 70+ countries (Chi et al. 2019; Google Crowdsource app users 2018). The images focus on key categories like household objects, plants and animals, food, and people in various professions. The addition of the OIX dataset resulted in an estimated increase by 250% of Indian subcontinent representation in the Open Images dataset.

Conclusion

We presented *Crowdsource by Google* - our crowdsourcing platform designed to be delightful, respectful, and transparent to all users. It is a solution for more representative and inclusive machine learning data. In addition, it can solve for data for long-tailed use cases, where it's hard to collect or label the data. By partnering with the global community, *Crowdsource by Google* helps bring out every individual's passion for language, culture, and machine learning while leveraging their contributions to create AI products that work well and represent people everywhere.

References

Chi, P.; Long, M.; Gaur, A.; Deora, A. K.; Batra, A.; and Luong, D. 2019. Crowdsourcing images for global diversity.

Google Crowdsource app users. 2018. Dataset: Open Images Extended - Crowdsourced. https://ai.google/tools/datasets/open-images-extended-crowdsourced/. Accessed: 2019-09-25.

Krasin, I.; Duerig, T.; Alldrin, N.; Ferrari, V.; Abu-El-Haija, S.; Kuznetsova, A.; Rom, H.; Uijlings, J.; Popov, S.; Kamali, S.; Malloci, M.; Pont-Tuset, J.; Veit, A.; Belongie, S.; Gomes, V.; Gupta, A.; Sun, C.; Chechik, G.; Cai, D.; Feng, Z.; Narayanan, D.; and Murphy, K. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from* https://storage.googleapis.com/openimages/web/index.html.

Mahmud, F., and Aris, H. 2015. State of mobile crowdsourcing applications: A review. In *2015 4th International Conference on Software Engineering and Computer Systems (ICSECS)*, 27–32. IEEE.

Phuttharak, J., and Loke, S. W. 2018. A review of mobile crowdsourcing architectures and challenges: Toward crowd-empowered internet-of-things. *IEEE Access* 7:304–324.