A Crowd-AI Hybrid System for Deep Learning-based Applications

Akash Govind Kuttikad
NetID: agk4

December 2022

1 Introduction

Over the last few years, machine learning has tremendously eased our day-to-day lives with models that provide movie recommendations, language translations, healthcare predictions, face recognition, and many more applications. A Forbes 2022 report predicts that AI has achieved an inflection point and is poised to transform every industry - including healthcare, consumer experience, foreign policy and climate crisis [1]. The key advantage of such systems is its capability to identify, extract and predict crucial inferences from a data-set with minimal or no human intervention.

To train these models, large amounts of data is collected from the users, since ML systems are only as good as the quality of the data that informs the training of ML models. Data required is more than what a single individual or organization can contribute. For instance, Google Ngram (an online search engine for text) uses 486 billion data records, and Google translate roughly uses 1 trillion data samples for training their models [2]. But, Google makes use of a very pragmatic solution — the task of data labeling and validation for their machine learning models are outsourced to all those who are Google users. However, not all organizations would have this luxury of collecting training data for free. Labeling data in-house requires hiring skilled employees and gives the advantage to have a transparent labeling process by knowing the people who perform the labeling. Rather than doing such in-house labeling, crowdsourcing platforms allow companies to distribute thousands of tasks and easily maximize the return on investment by having operational expenditure based on the required demand.

The crowdsourcing platform (Figure 1) targets the solutions of human tasks by adopting an internet crowd which is a scaling workforce and is flexible regarding required qualifications. In exchange for its services a contributor receives a small reward per task. Businesses pay based on the work done by individuals rather than agreeing on a contract with fixed terms. This business model help organizations save money while encouraging the crowd to provide quality work. Data collection from diverse backgrounds also enables businesses to eventually help reduce bias in AI solutions.

In this work, we make use of Amazon Mechanical Turk as the crowdsourcing platform, which was the first to enter the market of automating human intelligent tasks. It is part of Amazon's Web Service offerings and is commonly used for text classification, transcriptions, surveys, and data labeling.

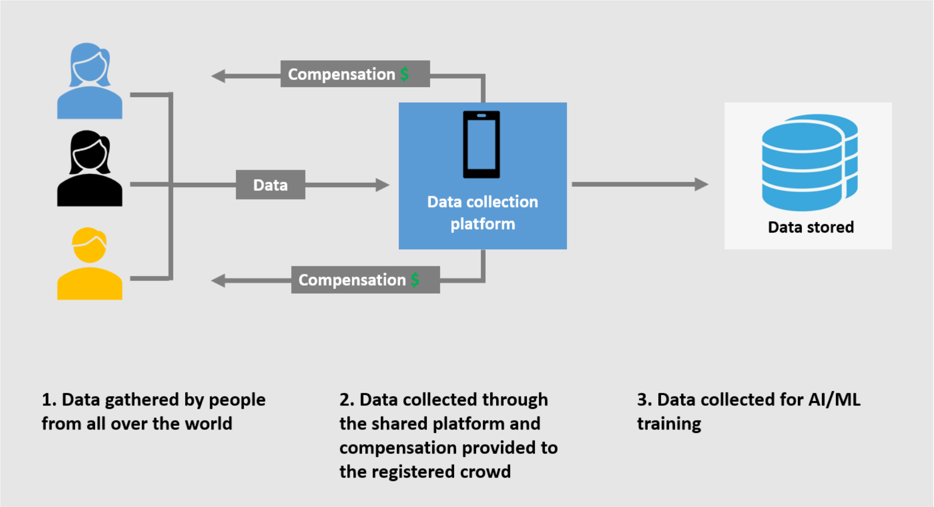

Compensation $
Data
Data collection platform
Data stored
Compensation $
1. Data gathered by people from all over the world
2. Data collected through the shared platform and compensation provided to the registered crowd
3. Data collected for AI/ML training

Figure 1: Crowdsourcing layout

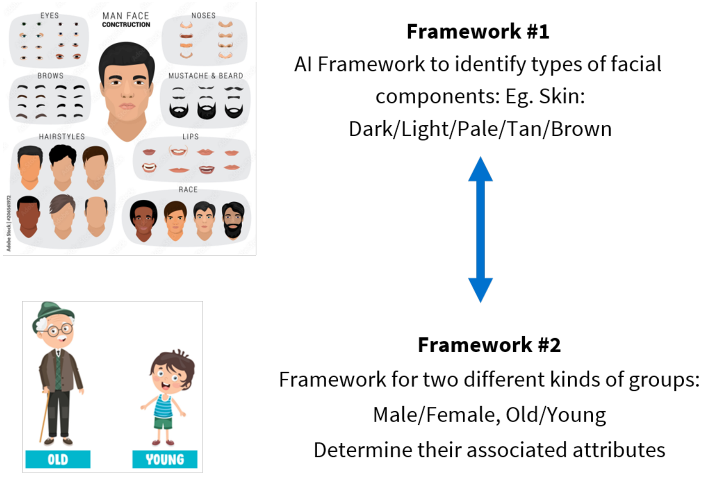
1.1 Problem Statement

Increasingly, AI is being used in a variety of sensitive applications that affect human lives. These range from recruitment and compensation, to consumer lending, to healthcare and criminal justice. This proliferation of AI use cases has led to the novel and emerging field of Responsible AI – the practice of designing, building, and deploying AI in a manner that empowers organizations, while treating people fairly without prejudice. Bias in AI can exist in many shapes and forms, and can be introduced at any stage in the model development pipeline.

AI, while having empowered organizations to make informed decisions, has also created unique challenges because when AI models are being trained, the feedback process is automatic. Any issues/biases in the training data, quickly translate into outcomes. Because AI learns so efficiently, it risks scaling and magnifying biases to a greater degree than manual systems. The use of AI may suggest that data-driven outcomes are inherently more objective and reliable than human decision-making. This is only partly true. AI systems formalize the decision-making process, and they tend to be more objective and less prone to human prejudice. However, they are not immune from bias [3].

Representation bias is a notion of fairness which happens from the way we define and sample a population to create a dataset. For example, the data used to train Amazon's facial recognition [4] was mostly based on white faces, leading to issues detecting darker-skinned faces. Another example of representation bias is data sets collected through smartphone apps, which can end up underrepresented lower-income or older demographics. To address this issue with fairness, the best practice is to make sure that your predictions are calibrated for each group - in other words, to ensure the representation of each group is stratified with respect to the sensitive attribute (race, color, or gender).

This work aims to learn facial attributes of two groups and samples images based on these learned attributes. We introduce two frameworks - #1 and #2 (Figure 2. Framework #1 aims to learn classes (adjectives) within facial components like skin, lips, eyes, forehead and many more.



Framework #1
AI Framework to identify types of facial components: Eg. Skin: Dark/Light/Pale/Tan/Brown

Framework #2
Framework for two different kinds of groups:
Male/Female, Old/Young
Determine their associated attributes

Figure 2: Frameworks #1 and #2

Crowdsource workers are required to provide adjectives that help to differentiate facial components amongst images. Framework #2 is used to define n groups of images, which has to be sampled ultimately. Crowdsource workers are required to label adjectives that help to differentiate facial components between these groups. The goal of this framework is to provide adjectives that could be used for sampling, which would provide a fair means for sampling rather than the group name itself. Such a framework has not been implemented previously to our knowledge. Moreover, this tool can be used to sample any set of groups once data for Framework #2 is available, and hence it is highly flexible in terms of application.

1.2 Related works

Crowdsourcing has been widely in literature used in several applications lacking labelled facial data. Washington et al. (2021) [5] explore the feasibility of using crowdsourcing to acquire reliable soft-target labels and evaluate an emotion detection classifier trained with these labels. They recognize crowdsourcing with a sufficient filtering mechanism as a feasible solution for acquiring softtarget labels. Chen et al. (2011) [6] studied identification of facial attributes like gender, race, age, hair style through social media platform data for locating designated persons and profiling the communities from image/video collections. A model to detect facial expressions of video-game players as: happy, anger, disgust, contempt, sad, fear, surprise, or neutral to link with the outcome of the game was developed by Tavares et al. (2016) [7]. They tune their crowdsourcing parameters by examining the effect of multiple variables (reward, judgment limits, golden questions) on annotators' performance, and compare ground-truth (expert) labels and crowdsourcing labels.

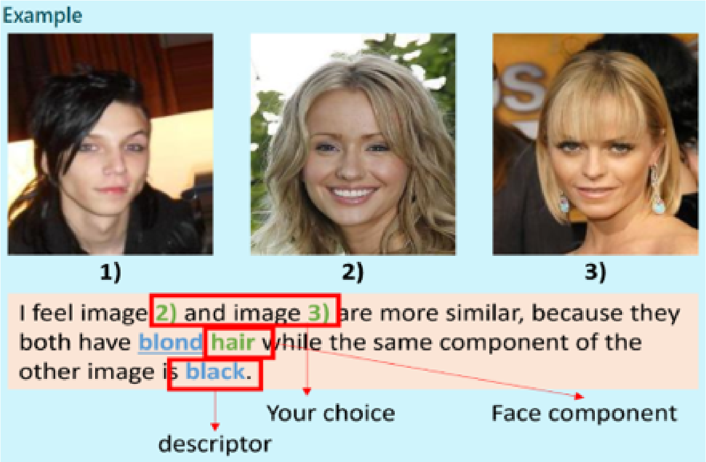
2 Theory and Modelling

2.1 Data Collection

The CelebFaces Attributes Dataset (CelebA) is used for training the models in this work, which is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. It has large diversities, large quantities, and rich annotations, including 10,177 number of identities.

For Framework #1, the crowdsource worker is asked the following set of questions (Numbered based on responses given in Figure 5):

1. Which of the two images (a b) are similar?
2. Which of their component is similar?
3. What is the adjective that describes the similar components?
4. What is the adjective that differentiates the same component in the third image?



I feel image 2) and image 3) are more similar, because they both have blond hair while the same component of the other image is black.

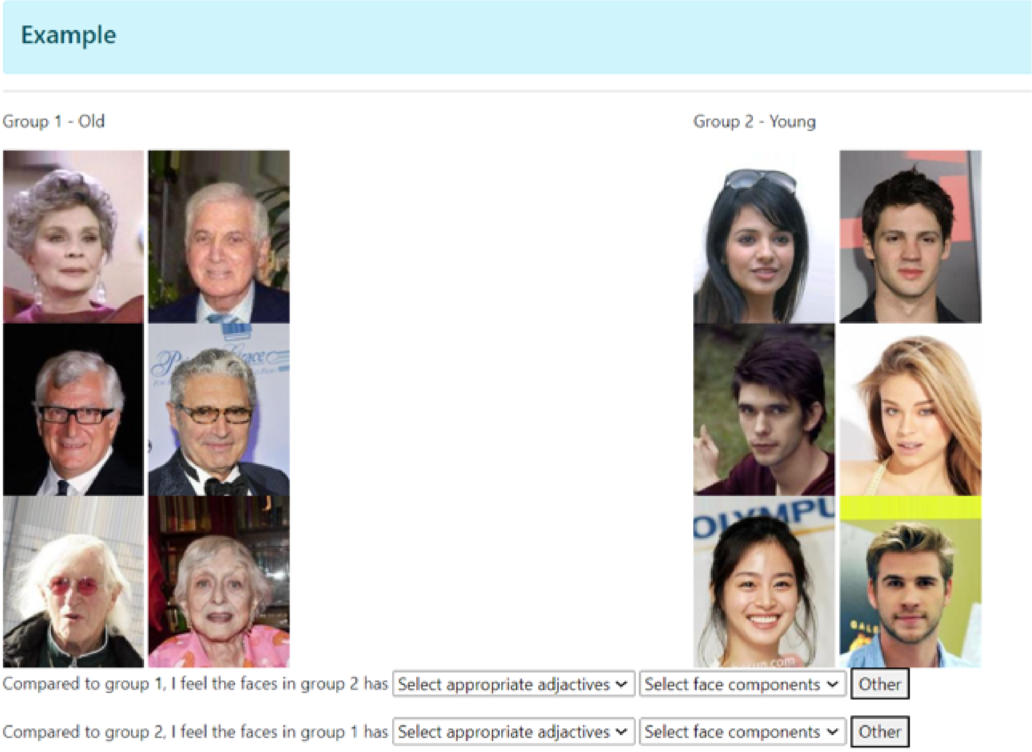
Figure 3: Crowdsoruce Question example: Framework #1

An example of a response is given to the worker to guide them in the right direction (Figure 3). According to the example, the worker is expected to select the 'hair' component where images (2) and (3) have blond hair whereas image (1) has black hair. Crowdsource workers are also asked a security question to ensure reliability of their answers.

For Framework #2, the workers are asked to identify the components/adjectives associated with two different groups: in this case Old vs. Young. The crowdsource worker also has the prerogative to enter a custom component/adjective - which could be later used/added to Framework 1 to learn a new class.

The MTurk crowdsourcing parameters used are as follows:

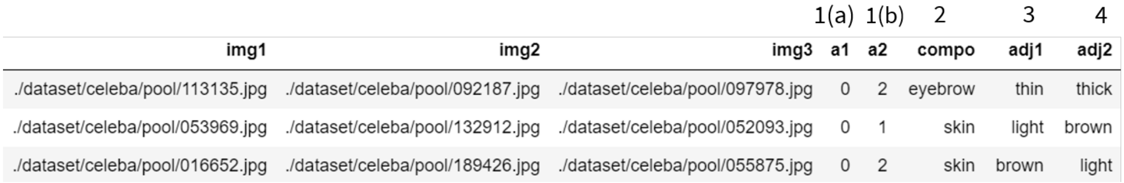
- Number of assignments per task = 10
- Reward per assignment = \$ 0.05
- HIT approval rate > 95%
- Number of HITs approved > 1000



Compared to group 1, I feel the faces in group 2 has

Compared to group 2, I feel the faces in group 1 has

Figure 4: Crowdsoruce Question: Framework #2



			1(a)	1(b)	2	3	4
img1	img2	img3	a1	a2	compo	adj1	adj2
./dataset/celeba/pool/113135.jpg	./dataset/celeba/pool/092187.jpg	./dataset/celeba/pool/097978.jpg	0	2	eyebrow	thin	thick
./dataset/celeba/pool/053969.jpg	./dataset/celeba/pool/132912.jpg	./dataset/celeba/pool/052093.jpg	0	1	skin	light	brown
./dataset/celeba/pool/016652.jpg	./dataset/celeba/pool/189426.jpg	./dataset/celeba/pool/055875.jpg	0	2	skin	brown	light

Figure 5: Sample Response: Framework #1

2.2 Pre-Processing and Cleaning

As a part of the initial cleaning process, invalid adjectives are removed from the dataset that is not of interest to our modelling (eg. attractive/beautiful). Later, rows that have a repsonse time of less than 20 seconds are eliminated, which might signify trivial/misleading responses. For the first batch obtained, the above steps eliminate 30% of the data.

Later, in order to convert the responses in Figure 5, we normalize the data to match synonyms (dark-black, white-light, large-big). We retain only valid responses for each image through majority voting - adjectives which appear at least 10 times for each image. Thereafter, we convert to label for each image for a chosen target attribute and perform the training.

2.3 Modelling

We make use of the well-established VGG-16 (Figure 6) model to train our models. If we take the instance of the 'skin' component, we have 8 different classes. A train-test split of 80:20 is maintained for all models which is stratified over the class label. We train the model for approx. 50 epochs, and any further training results in overfitting.

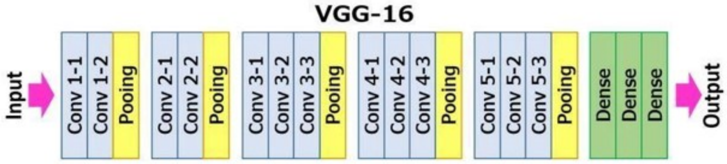


Figure 6: VGG-16 Architecture

3 Results and Future Work

After training for atmost 50 epochs for the skin /eyebrow component and 25 epochs for the eye components, we arrive at the test-train losses as shown in Figure 7. We calculate the top1 and top2 accuracy for these models (Table 1). The top-k score is defined as follows: we check if the top class is one of your top k predictions (the top k ones with the highest probabilities).

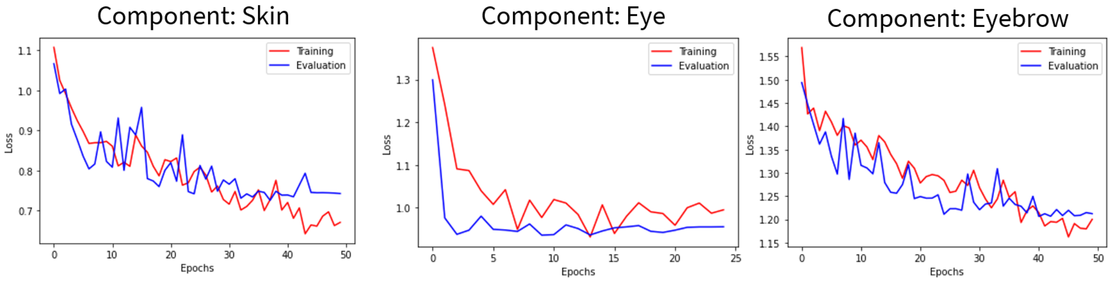


Figure 7: Test-train losses

We observe that the top-k score for the eyebrow component seems to be pretty low in comparison with other components. One of the primary reasons could be the lack of training data: majority of crowdsource workers do not find the eyebrow to be a distinguishing component among images. However, it can be trained later if it seems to be an integral identifier for any group in Framework #2.

Moving forward, the next steps in the project is to obtain an initial batch for Framework #2 for the two groups: Old and Young and identify the key components-adjectives that discern these groups. Later, based on the responses from Framework #2 can include more facial components in Framework #1 that helps to represent the two groups through fair sampling.

Component	Top1 Accuracy (%)	Top2 Accuracy (%)
Skin	85.96	92.98
Eye	75.84	86.93
Eyebrow	62.42	78.18

Table 1: Top1 and Top2 Accuracy

4 Acknowledgements

I would like to express my deep gratitude to Dr. Dong Wang for providing me with this opportunity to work on this project. I also thank Yang Zhang PhD. for his constant guidance, mentorship and support throughout the course of this project. I would also like to extend my regards all members of the Social Sensing and Intelligence Lab at the School of Information Sciences, UIUC for their resources and assistance throughout this independent study.

References

1. https://www.forbes.com/sites/forbesbusinesscouncil/2022/05/05/the-future-of-ai-5-things-to-expect-in-the-next-10-years/?sh=65476ba7422b

2. https://developers.google.com/machine-learning/data-prep/construct/collect/data-size-quality

3. Bateni, A., Chan, M., & Eitel-Porter, R. (2022). AI Fairness: from Principles to Practice.

4. https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender

5. Washington P, Kalantarian H, Kent J, Husic A, Kline A, Leblanc E, Hou C, Mutlu C, Dunlap K, Penev Y, Stockham N, Chrisman B, Paskov K, Jung JY, Voss C, Haber N, Wall DP. Training Affective Computer Vision Models by Crowdsourcing Soft-Target Labels. Cognit Comput. 2021 Sep;13(5):1363-1373. doi: 10.1007/s12559-021-09936-4. Epub 2021 Sep 27. PMID: 35669554; PMCID: PMC9165031

6. Yan-Ying Chen, Winston H. Hsu, and Hong-Yuan Mark Liao. 2011. Learning facial attributes by crowdsourcing in social media. In Proceedings of the 20th international conference companion on World wide web (WWW '11). Association for Computing Machinery, New York, NY, USA, 25–26.

7. Gonçalo Tavares, André Mourão, and João Magalhães. 2016. Crowdsourcing facial expressions for affective-interaction. Comput. Vis. Image Underst. 147, C (June 2016), 102–113.