Training with Less: How People Select Data with Higher Value for AI

Farnaz Zamiri Zeraati,[1] **Jonggi Hong,** [1] **Kyungjun Lee,** [1] **Hernisa Kacorri** [1]

[1] University of Maryland, College Park

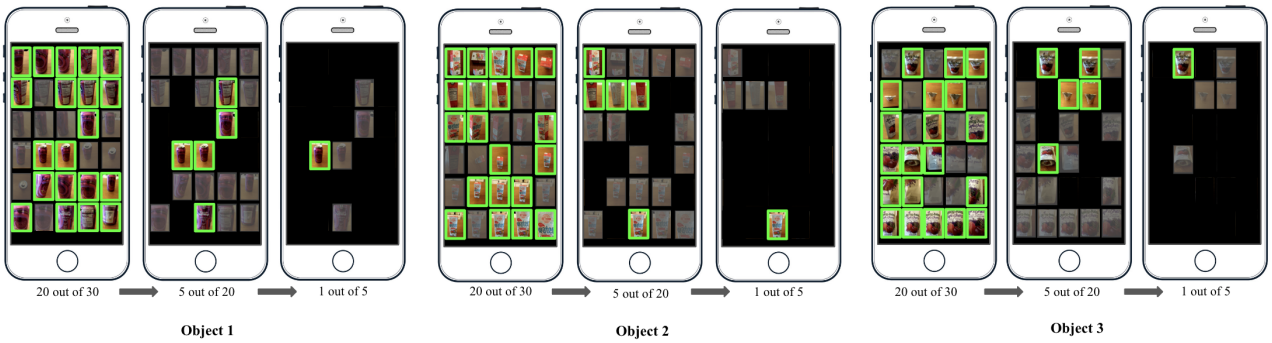

Figure 1: A subset selection activity illustrating how a participant (P41) in our study subsequently chose their "best" subsets of 20, 5, and 1 images across their 3 objects of choice for training a 3-way image classifier to be "robust".

Abstract

People are increasingly made aware of the importance of data for AI. They are often called to make conscious decisions around the use of their own photos, text, and interactions for improving models overall or fine-tuning them to their needs. Understanding the value of their data can play a critical role in these decisions. Yet, it is unclear how those who may not have machine learning expertise can tell which of their data are of value for training models. We conduct a crowdsourcing study and publicly share a dataset called CrowdTeaMa, where participants ($N = 100$) are called to fine-tune a 3-way classification model with their own image data and select consecutively smaller data subsets that they deem to be of higher value to the model. Our results highlight six unique patterns in participants' subset selection strategies, which outperform random. However, when comparing them to other computational methods for selection that employ submodular functions and generative AI, it isn't always a clear win. These findings demonstrate the potential of computational methods for supporting people in decisions around their data and their value. We examine this promise in a new context, machine teaching in accessibility, where blind people are called to fine-tune image recognition models with their own photos without being able to visually inspect their data.

Introduction

The increasing recognition of the pivotal role that data plays in artificial intelligence (AI) is evident as it becomes integral to training AI models. Individuals are now conscious of the fact that they leave valuable data traces when interacting with technology, that can be stored and repurposed for AI training. This awareness has made it to the public consciousness, emphasizing that individuals, not just companies, can leverage their data to train, fine-tune, and enhance AI models. Examples of this are data privacy concerns in health and fitness apps (Mink et al. 2022), virtual assistants (Chen et al. 2023), and social media algorithms (McHatton and Ghazinour 2023). This shift in consciousness spans from early machine teaching endeavors like object recognizers (Hong et al. 2020) to contemporary applications such as personalization in generative AI, showcasing a growing understanding of the potential inherent in personal data. Despite this increased awareness, a notable gap persists in comprehending how individuals lacking expertise in AI and machine learning (ML) value their data for training these models. The conventional understanding of AI development involving offline training phases remains opaque to end users, limiting their insight into the system's behavior. Bridging this knowledge gap becomes crucial in ensuring responsible AI usage, particularly as AI becomes more pervasive in user-centered applications. Integrating Interactive Machine Learning methodologies, which empower users in the ML model training process, emerges as a potential solution to democratize access and enhance understanding, emphasizing the need for a holistic approach to AI development that involves and educates the general public. Given the importance of fine-grained object recognition and the need to un-

derstand how people who have no expertise in ML value data for training AI, we propose a solution that leverages crowd workers as a proxy to control the study and get a larger sample. Our approach involves an iterative data subset selection task for machine teaching that allows us to identify the most important features of selected data for an object recognition model. **Why machine teaching?**: Machine teaching aims to reduce the number of examples a teacher needs to provide to a learner (Goldman and Kearns 1995; Shinohara and Miyano 1991). In certain contexts, this can be viewed as the task of optimal data subset selection. Aligned with our research questions, machine teaching serves as an effective proxy by providing an opportunity to gain deeper insights into how people perceive and interact with these systems, and what types of data they consider most valuable for training models. We developed a web-based testbed for a mobile teachable object recognizer, inviting participants to train and evaluate it using three objects of their choice within a specific object category (*e.g.*, cereal, drink, snack, spice, etc.). After testing their model, we asked them to consecutively select smaller data subsets of 20, 5, and 1 from the 30 images that they took of each object. We asked them to select the subset of photos that make the model more robust. We also implemented 3 different algorithms for data selection: Random selection, Facility Location-based selection, and Copilot selection. By calculating image similarity metrics across these different subsets, and model's performance across different subsets, we gain insights into how humans select data with higher value for AI. We also explore human data subset selection in a novel context: machine teaching for accessibility. Here, blind individuals are tasked with fine-tuning image recognition models using their own photos, despite being unable to visually inspect their data.

This paper presents three main contributions. First, the primary contribution is empirical which arises from an examination of how people select subset of data that they consider of higher value for training machine learning algorithms. As another empirical contribution, we present a collection of qualitative findings about different user groups' strategies in selecting data and their comprehension (or lack of comprehension) of the data value. From these insights, we derive a set of implications for the design of Interactive Machine Learning systems tailored to novice users. Last, we have compiled and publicly released the CrowdTeaMa dataset, which we collected from crowd workers during our study.

Related Work

There is a rich literature in subset selection. We review a sample of prior work over the last decade through the lenses of (selector $\rightarrow$ consumer) pairs with the data *selector* being a human or an algorithm that chooses the subset and data *consumer* a human or an algorithm that uses the subset. As shown in Table 1, we identify four distinct patterns.

Prior work under the first pattern includes studies where both the selector and consumer is an individual (🧍 $\rightarrow$ 🧍). The focus here is in understanding how people select subsets of data so that the resulting selection is more meaningful to them. This work is typically in the context of photo and video albums where people select photos that capture interest (Walber, Scherp, and Staab 2014; Payne 2017), photos that best represent a collection (Kuzovkin et al. 2018), or human-interpretable concepts about images that supports them to reason about classification models (Barker et al. 2023). The ultimate goal for all is to learn how to automate this process to generate smaller visuals that allow for more efficient and pleasurable experience. We also observe this pattern in the context of data analysis tools, where users select subsets of tabular data from large datasets to generate meaningful visualizations (Narechania et al. 2023). The goal here differs, as the work aims to support users in better valuing parts of a dataset for increasing validity, appropriateness, and utilization from others for analytical tasks. The majority of this prior work incorporates user studies with 30 to 36 participants sometimes recruited from crowdsourcing platforms like Prolific (Barker et al. 2023). In all studies the total number of options was bound but the subset selection size was usually up to the participants (Kuzovkin et al. 2018; Narechania et al. 2023; Barker et al. 2023) with one exception (Walber, Scherp, and Staab 2014). Interestingly, the level of familiarity that the selectors had with the data differs some being familiar (Walber, Scherp, and Staab 2014; Payne 2017; Barker et al. 2023) and others not (Kuzovkin et al. 2018; Narechania et al. 2023). Contrary to our study, participants were not the ones collecting the sets.

A second pattern seen in prior work that expands on the first (hence some of the overlap), leverages algorithmic approaches as the selector with people as consumers (</> $\rightarrow$ 🧍). The focus here is to select subsets for supporting humans under specific constraints. At an individual level, cases include selecting a single representative sign language facial expression from a corpus for generating understandable and natural animations (Kacorri et al. 2016), selecting a subset of human-interpretable concepts to minimize cognitive load (Barker et al. 2023), and capturing meaningful moments with a few video clips (Payne 2017). At an organization level, efforts focus on selecting subsets for archival purposes that aim to optimize storage efficiency given based on space constraints (Davidson et al. 2022). Often, these efforts do not report user studies with participants evaluating the final subsets. When evaluated, they use similarity and performance metrics, an approach that informs our work.

The most common pattern involves prior work where both the selector and consumer are algorithmic approaches (</> $\rightarrow$ </>). Here, efforts typically fall under the broader data-centric AI paradigm, where the focus is to select suitable data to impact model effectiveness and efficiency (Jakubik et al. 2024). These studies span a spectrum of selection objectives and data types. Few focus on selecting a representative subset to enable faster experimental iterations for instance, to improve image classification*e.g.* (Singh, Virmani, and Subramanyam 2019). Typically, the goal is to capture more relevant and high quality data points such as video frames for improving semantic segmentation in autonomous driving (Das et al. 2020), utterances from non-native speakers that better support ASR personalization (Kothyari et al. 2021), and utterances of higher quality for automatic speech recognition (Raza Syed and Mandel 2023). Other contexts

	🧍→🧍	</>→🧍	</>→</>	🧍→</>	Data Type	Selection Purpose
(Walber, Scherp, and Staab 2014)	•				image	interestingness
(Kacorri et al. 2016)		•			facial expressions	representativeness
(Payne 2017)	•	•			video	interestingness
(Bullard, Chernova, and Thomaz 2018)				•	object	classification performance
(Kuzovkin et al. 2018)	•				image	representativeness
(Singh, Virmani, and Subramanyam 2019)			•		image	representativeness
(Kaushal et al. 2019)			•		image	experimentation efficiency
(Das et al. 2020)			•		video	semantic segmentation performance
(Kothyari et al. 2021)			•		speech	relevance for personalization
(Killamsetty et al. 2021)			•		text, image	robustness & efficiency
(Davidson et al. 2022)		•			image	storage efficiency
(Narechania et al. 2023)	•	•			tabular	quality filtering
(Barker et al. 2023)		•			concept	interpretability
(Gong et al. 2023)			•		network data	storage efficiency
(Raza Syed and Mandel 2023)			•		audio	quality filtering
This study				•	image	classification performance

Table 1: Characteristics of prior work on subset selection divided into four categories based on data selector and data consumer: human for human (🧍→🧍), machine for human (</>→🧍), machine for machine (</>→</>), and human for machine (🧍→</>).

include online data selection for federated learning *e.g.*, in mobile networks (Gong et al. 2023) and data selection more broadly *e.g.*, for efficient learning (Killamsetty et al. 2021; Kaushal et al. 2019). The subset selection approaches employed also vary including generative adversarial networks (Singh, Virmani, and Subramanyam 2019) and data valuation (Gong et al. 2023; Raza Syed and Mandel 2023) to rank data points, pairwise similarity to remove redundant data points (Das et al. 2020), and submodular functions to optimize for relevance (Kothyari et al. 2021; Killamsetty et al. 2021; Kaushal et al. 2019). Approaches are contrasted to random as a lower baseline. Informed by this prior work, we contrast our participants strategies against a series of subset selection approaches employing generative models, and submodular functions along random.

A fourth pattern that has received less attention involves humans as the selector with algorithmic approaches as consumers (🧍→</>), the focus of this paper. The goal is typically to understand humans and their interactions with machines either for the purpose of building AI literacy or supporting efforts in human control or personalization. In one of closest studies to our work, participants are asked to select important objects from different groups to teach a grocery classification task to a robot learner (Bullard, Chernova, and Thomaz 2018). Similar to our study, overall set and selection sizes were fixed across participants; participants selected 3 objects out of 15 per category for a total of 4 categories. In contrast to our study, the larger set of objects were defined by the researchers. In our study, participants have more degrees of freedom. They can choose both the objects they want to train the algorithm on as well as define the larger set of pictures for training from which they are tasked to select subsets. Another difference lays in the participants. Participants (N=30) were recruited locally as people who had some expertise in machine learning (Bullard, Chernova, and Thomaz 2018). In our study, participants (N=100), recruited via Amazon Mechanical Turk, vary in their level of exposure to machine learning with the majority having no expertise.

Crowdsourcing Study

To examine how non-experts select data that they deem to have higher value for training a machine learning model, we conduct a crowdsourcing study (IRB # <*anonymized*>), where participants are recruited via Amazon Mechanical Turk. Our methods build on prior work leveraging crowdsourcing for behavioral and perception studies such as (Heer and Bostock 2010; Buhrmester, Kwang, and Gosling 2011; Jacques and Kristensson 2021) and those exploring human interactions with machine learning in (Vaughan 2018).

Specifically, we incorporate subset selection tasks in a web-based object recognizer for mobile phones (our testbed), where participants are asked to train, test, and retrain an image classification model to identify three objects of their choosing. For each user, the testbed creates a new convolutional neural network based on Google Inception V3 (Szegedy et al. 2016), pre-trained on ImageNet (Deng et al. 2009). Each time a user provides a teaching set, the last layer of the pre-trained model is replaced with a new softmax layer and re-trained with the user's images for 500 steps using a gradient descent learning rate of 10^{-2}. Models are trained in real-time on our 8-GPU server asynchronously, allowing the app to continue running and collecting open-ended feedback from participants during the training process. The web interface communicates with the server through the Flask API (Pallets 2024).

Participants

We recruited 143 participants over 10 days. However, data from 43 participants were excluded – 7 helped in piloting, 1 used the same object for all classes, 3 took photos of objects in display screens, 2 took photos with no objects. The other 30 had technical problems by attempting the task simultaneously with our system failing to distribute them across the GPUs, losing data from 12 and interrupting the task for other 18; all were compensated and the bug was fixed. The 100 participants who were included in this paper were 20 to 60 years old ($\mu = 32.6$, $\sigma = 8.3$). Almost half (49) identified as

man, 50 as women, and 1 as non-binary. The majority (90) reported being right-handed, and no one reported visual or motor impairments. Most participants used mobile devices to take photos on a weekly basis, but only a few used object, food, or plant recognition applications. Regarding participants' familiarity with machine learning, 6 participants had never heard of it, 45 had heard of it but didn't know what it does, 48 had a general understanding of its purpose, and only one reported having extensive knowledge.

Procedure

The subset selection task in this study is part of a broader series of tasks where participants trained and tested an object recognition model, described here for more context.

Initially, the testbed gathers background information, technology experience, and familiarity with machine learning from participants. It then presents five object category options: bottle, cereal, drink, snack, and spice, with three sample icons for each category representing the preferred shape. These categories are inspired by previous research on personal object recognizers (Kacorri et al. 2017) and are designed to prompt the selection of everyday objects that vary in size, shape, color, material, and function. To ensure that object shape or size does not contribute to any observed inconsistencies between classes, participants are instructed to use three objects within the same category.

After that, participants are taken through a guided experience with the web-based teachable object recognizer on their mobile phones, including 5 tasks. In **test 0**, participants are asked to take photos of their objects to see if the existing non-personalized model can recognize them. This task helps to familiarize participants with the interface and to collect evaluation examples unbiased from one's teaching experience that is to follow. In **train 1**, participants are given instructions to train the object recognizer with the aim of making it robust enough to identify their objects '*in any location and for anyone*'. They take 30 photos of each object for training. This process is repeated three times, once for each of the randomly assigned objects. As a result, the first teaching set involves a total of 90 photos (30 photos per object). While the model was being trained on their photos, participants were asked to review their training sets and select the *best* 20 out of 30 training photos per object, 5 out of 20, and 1 out of 5, as shown in Figure 1. In **test 1**, participants are instructed to take photos of their object and test the trained object recognizer to evaluate its robustness. In the subsequent tasks (**train 2** and **test 2**), participants are given an opportunity to re-train and test the model from scratch.

During our pilot sessions, we estimated that participants could successfully complete the study in 30 to 40 minutes. Based on a compensation rate of $15 per hour (Hara et al. 2018), all participants received a total of $10 upon completing the data collection. To motivate participants, we implemented a performance-based payment system (Ho et al. 2015), where the total of $10 was divided into a $5 flat participation fee, a $2 bonus for first training, and an additional $3 bonus for improved performance in the second training. Closely aligning with our estimates, participants spent on average 35.57 minutes ($\sigma = 12.85$) on the testbed.

Subset Selection Approaches

To better characterize our participants subset selection strategies we contrast them to random, a baseline approach, as well as computational methods for selection that employ submodular functions and generative AI. The choice behind these methods is informed from recent literature surveyed in Related Work.

Submodular Functions: Facility Location

A submodular function (Edmonds 2003), $f : 2^V \rightarrow \mathbb{R}$, assigns values to a subset from a finite ground set V and satisfies the following property for any $S \subseteq S' \subseteq V$, $u \notin S'$:

$$f(S \cup \{u\}) - f(S) \geq f(S' \cup \{u\}) - f(S') \quad (1)$$

In our scenario, V is the pool of photos taken by a participant, S and S' are candidate subsets, u is a photo under consideration, and f is valuation function measuring quality of the subsets. The submodular property 1 states that the marginal benefit of adding a photo to S is at least as high as the marginal benefit of adding u to the superset S'. Intuitively, it captures the notion of diminishing returns and allows us to find a maximally valued subset for training given an appropriate submodular function, in our case, the Facility Location function. This particular submodular function aims to choose examples that effectively represent the data space by maximizing the pairwise similarities between data points. In our study, this translates to the selection of diverse images within a set, taking into account pixel-by-pixel data. For this approach, we utilized Apricot (Schreiber, Bilmes, and Noble 2019), a submodular optimization package.

Generative AI: Copilot

Since the emergence of large language models (LLMs), there has been significant interest in comparing their performance to that of humans across various tasks (Møller et al. 2024; Li et al. 2024; Martin et al. 2024). For our comparison, we utilized Microsoft Copilot[1] in Bing chat (Microsoft 2024). Specifically, to maintain consistency with the instructions given to participants and to accurately replicate the task, we provided a photo grid containing 30 images that each participant took of a specific object. Each image was labeled with a number from 1 to 30. We then gave the following prompt in the chat: *"In this grid of images, there are 30 images of an object, each labeled with a number. Imagine we want to train an object recognizer with 20 out of these 30 images to recognize this object later. Which image do you think is the best for this task? Please provide the number of the image you choose in brackets '[]' and explain why you chose it."* We used SydneyClient (vsakkas 2023), a Python client for Copilot, to retrieve responses from the chat. After collecting the numbers of the selected images, we created a new image plotting these selected images in a grid. We then input this newly generated image to Copilot with the same prompt but asked it to select a smaller subset (5 out of 20), and similarly, for a final selection (1 out of 5).

[1]In contrast to other open-source LLMs we experimented with at the time, Copilot could handle a grid of images similar to those shown to our study participants. Additionally, its implementation in Bing Chat supported some privacy options.

Analysis

Our analysis includes two datasets. The first, called **CrowdTeaMa**, includes a total of 22500 photos from our crowdsourcing study with 100 participants and is used to explore how people who may not have expertise in machine learning select subsets of data for training AI and how their strategies compare to computational methods. We are now publicly sharing this dataset at *[anonymous link]*.

To gain insights into selection patterns in this dataset, we first analyzed performance trends. We trained the testbed model on each of the human/algorithm-selected subsets of images from 100 participants. We utilized data from our four subset selection approaches: Human, Facility Location, Copilot, and Random. Each approach was applied to three different subset sizes—20, 5, and 1—plus the original set of 30 images taken by each participant, resulting in a total of 1300 models. For each model, we measured the F1 score using two different test sets. The first set, Test 1, included photos that participants took immediately after selecting their subsets. The second set encompassed all of Test 0, 1, and 2, which consisted of a larger set of photos taken before, immediately after, and long after the subset selection.

Another aspect employed in our analysis that provided more context for the observed patterns was the presence of diversity among the images taken by the participants. As a subjective approach, one researcher from our team annotated the visual distinctiveness within each of the participants train and test sets. The researcher assessed variation across three categories: the background behind the object, the object's size, and object viewpoint, and calculated an overall variation score within each train and test sets as a percentage. For a more objective approach, we employ the Structural Similarity Index (SSIM) (Wang et al. 2004), a metric utilized for quantifying image similarity. We computed SSIM values between the training and testing sets to evaluate the degree of similarity between the two per participant.

We also examine the promise of computational methods to support non-experts in machine teaching within a new context: accessibility. Here, blind people are called to fine-tune image recognition models with their own photos without being able to visually inspect their data. For this analysis, we leverage an existing dataset called **TEgO** (Lee and Kacorri 2019). It includes a total of 10260 photos from 19 distinct objects, captured by two individuals–a blind person and a sighted person–using a smartphone camera. The data collection task is similar to our study, where the photos are used to train and test a teachable object recognizer, with each individual creating their own training (30 photos per object) and testing (5 photos per object) sets. The data collection task is repeated under different settings and environments: *'in-the-vanilla'*, where an object is placed against a plain background such as an empty desk, and *'in-the-wild'*, where an object is placed in a cluttered background among others. We conducted a similar performance analysis on the TeGO dataset. We created models for both sighted and blind datasets in two different environments and tested them with the corresponding test sets.

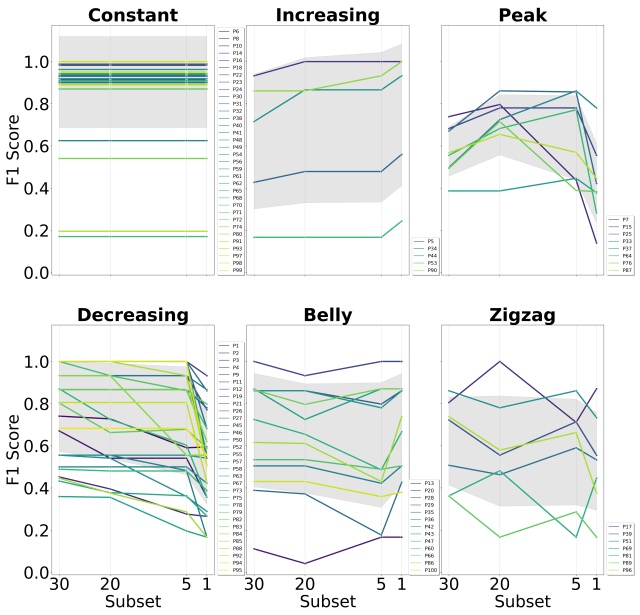

Figure 2: F1 score trends across clusters of participants

Results on the CrowdTeaMa Dataset

In this section, we first examine the subset selection strategies employed among our participants by looking at different trends in model performance as subset size decreases. We then compare these strategies against a Random baseline and other computational approaches.

How do Humans Select Subsets

As shown in Figure 2, we observe six distinct trends on model performance as participants decrease subset size.

Constant Trend: For the majority (N=34), their models performed roughly the same when trained with 30, 20, 5, or 1 examples. This was not a surprise as training and testing sets inhibited very little variation (<3%) and were quite similar to each other (SSIM=79%). Thus, whatever the selection strategy, it had little effect. This is the case both for those who reported having a broad understanding of what machine learning is and what it does (N=17) and those who either had heard of it but don't know what it does (N=15) or had never heard of it (N=2).

Increasing Trend: For a few (N=5), the trend was counterintuitive: smaller training size led to a better model performance. What is notable about this group is that they all trained on fine-grained objects, differing primarily in small text written on the containers. They also incorporated some variation among their training (17%) and test set (8%). All selected as a single image one where the main side and logo of the object were clearly visible.

Peak Trend: For some (N=8), we observe that performance peaks around 20 and 5 photos. This group demonstrated the highest variation in training and test sets (40% and 20%, respectively) mostly changing the background *e.g.*, placing the object among others in a cluttered setting.

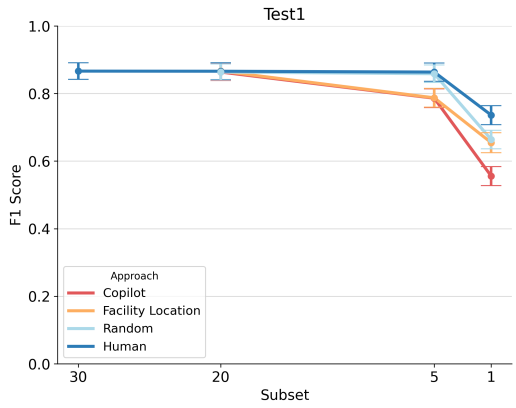

Figure 3: Median F1 scores in Train 1/Test 1.

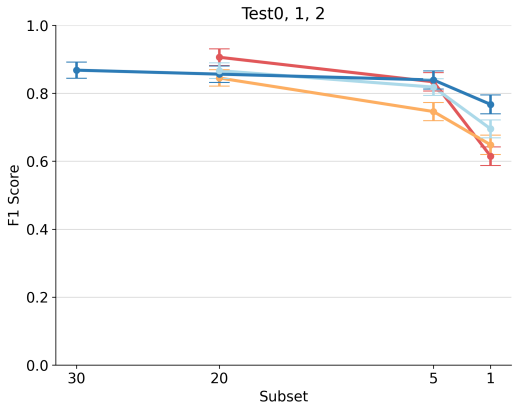

Figure 4: Median F1 scores in Train 1/Test 1, 2 & 3.

Decreasing Trend: For about a third (N=32), model performance decreased with size; a trend we would expect. Among this group, there was a notable difference between training and testing sets in variation (17% versus 6%). This difference in variation was primarily due to view variation. Participants in this group incorporated more view variation in their training set by capturing the object from different angles, but exhibited minimal view variation in their test set.

Belly Trend: Approximately one-tenth of the participants (N=13) experienced a decline in their performance trend followed by recovery on a smaller subset, primarily from subset 5 to 1. Visual inspection of the images revealed that most participants in this category faced challenges related to image quality. For instance, certain participants presented images captured under poor lighting conditions or where the objects were not clearly visible, such as transparent bottles, or where the object was cropped within the image. In such instances, a single high-quality image provides better learning opportunities than multiple low-quality images.

Zigzag Trend: A small group of participants (N=7) exhibited a non-linear performance trend characterized by a zigzag pattern. Upon visual inspection, we discovered that this category encompasses participants who employed mixed strategies in their approach. One particularly interesting approach was observed in one of the participants. This participant claimed to possess a broad understanding of machine learning concepts and its functionalities. They adopted a multi-faceted approach by capturing images of the packaging up close, photographing the emptied package, and taking pictures of the contents inside the package. This varied approach resulted in the fluctuating trend, resembling the model being trained with multiple objects simultaneously rather than focusing on one object at a time.

How do Humans' Subset Selection Strategies Compare to Computational Methods

To see how the subset selection strategies of our participants compare to different computational methods, we contrast the performance of their models (F1 scores) across the same test sets. We first compare performances on the 15 test images (5 per object) that participants took immediately after the first attempt at training and selecting subsets. Surprisingly, we observe that they tend to perform equally or better than other algorithmic approaches even though almost all of them have no machine learning expertise. As shown in Figure 3, the overall trend for our participants' models (Human) is a barely dropping performance for 20 out of 30 (median F1 score: 0.866 for both 20 and 30) and 5 out 20 (median F1 score: 0.863) with a more prominent drop for 1 out of 5 (median F1 score: 0.736). However, as we saw in the previous section this trend in not consistent across participants.

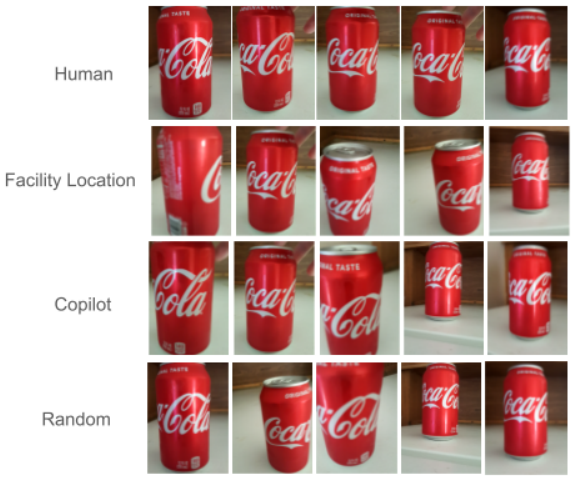

Figure 5: Contrasting approaches for 5 of 20 (P75).

At 20 out of 30, the comparison is less compelling as all selection approaches, those adopted by our participants and those by the algorithms tend to be similar to Random in median F1 performance. Visual inspection of these subsets indicate a lower variation being incorporated across the Human selected subsets and higher variation across subsets selected by Facility Location and Copilot. This is expected since Copilot stated in all of its responses that diversity is prioritized in its selection (Figure 5, 6). At 5 out of 20, results get more interesting. We see that the median F1 score for Human remains relatively unchanged, as it does for Random (median F1 score: 0.858). However, this is not the case

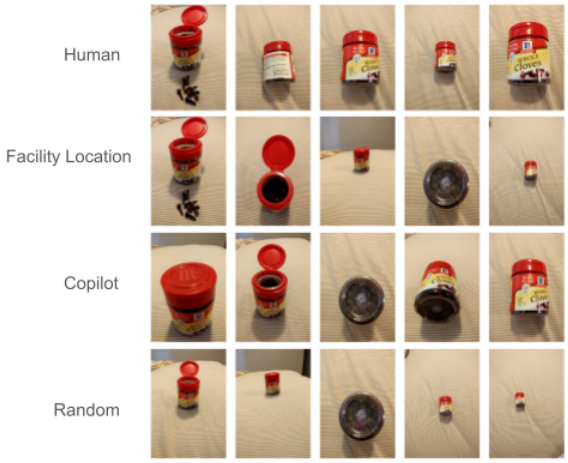

Figure 6: Contrasting approaches for 5 of 20 (P1).

Figure 7: Contrasting approaches for 1 of 5 (P1).

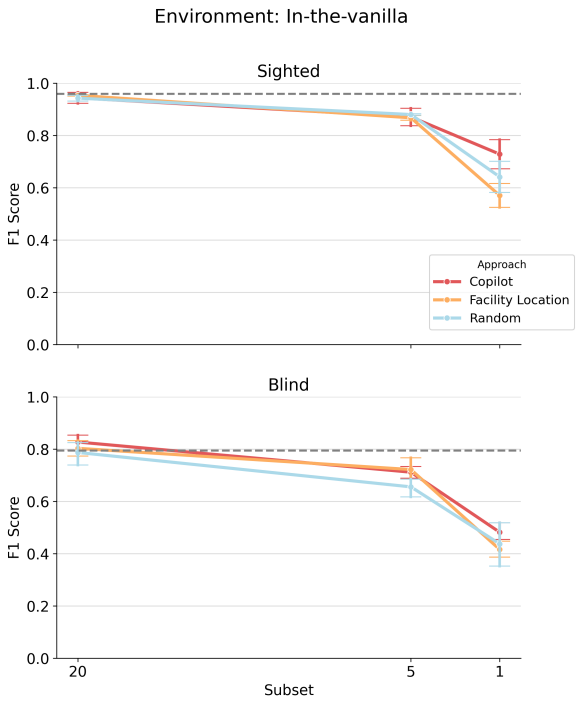

Figure 8: Comparing subset selection algorithms F1 score for the Tego Dataset, in-the-vanilla environments. The gray line indicates the F1 score for subset=30

for Facility Location (median F1 score: 0.787) and CoPilot approaches (median F1 score: 0.786). As shown in figure 5, visual inspection of the subsets revealed that, particularly in smaller subsets (5 and 1), humans tend to filter out lower-quality images, such as blurred or cropped ones. In contrast, the Facility prioritizing diversity does not always filter out images based on these criteria, same as the copilot. However, these patterns were not consistent in the images selected by Random.

At 1 out of 5, these splits in performance are more striking though the order is preserved. Human withstand at the top (median F1 score: 0.736), Random (median F1 score: 0.663) approximates Facility Location (median F1 score: 0.654) and at the bottom stands CoPilot (median F1 score: 0.555). As shown in Figure 7, the top images that humans selected tend to be images where the brand name and primary features are clearly visible to the camera.

These results are surprising. *Are people, even without machine learning expertise, much better at valuing data of importance for models to learn?* Given that Random is the second best approach, we hypothesize that this observation is instead explained by the testing images, testing with similar photos to those participants selected as being the "best".

To explore this hypothesis, we expand the testing set to also include images that participants took even before training or after a subsequent training attempt. As shown in Figure 4, the trends remain roughly the same for Human and Random. However, given a more challenging testing set, CoPilot outperforms them both for 20 out of 30 and 5 out of 20 but quickly drops to being last for 1 out of 5, and Human performing the best. Our hypothesis is supported by the observation that subsets with greater diversity perform better when tested against a more varied test set. However, when it comes to selecting a single best image, humans remain at the top. They excel at choosing images that clearly depict the object from an optimal perspective.

Results on the TEgO Dataset

Technologies such as teachable object recognizers, enable blind individuals to personalize algorithms and train them to locate or recognize objects of interest. This presents a compelling case study where subset selection becomes important. In prior research (*e.g.*, (Kacorri 2017; Morrison et al. 2023; Lee et al. 2019), researchers have explored various methodologies within this domain. Our analysis aims to extend this inquiry by examining how the relationships observed between the subset selection methods in the previous section, generalize to the realm of accessibility.

In-the-vanilla environment, where objects are placed against a plain background: For subsets 20 and 5, we observed that algorithmic approaches had comparable performance levels. However, in subset 1, Copilot outperformed Facility and Random, contrasting with our observations in CrowdTeama tested by both test 1 and test 0, 1, 2, where Copilot displayed the lowest performance among the existing approaches. The same pattern is seen for the photos taken by the sighted and Blind.

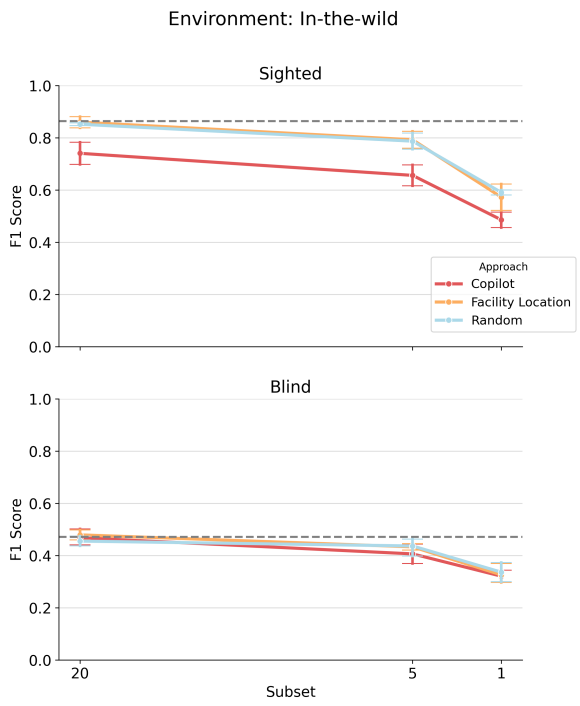

Figure 9: Comparing subset selection algorithms F1 score for the Tego Dataset, in-the-vanilla environments. The gray line indicates the F1 score for subset=30

In-the-wild, where objects are placed against a cluttered background: Performance patterns in cluttered environments shows Copilot declining performance. Upon closer examination of the images, we observed that in this set, Copilot frequently chose images from non-prominent areas of objects, such as the top, where significant brand elements were lacking. This trend was particularly pronounced in wild environments. Also, Copilot seems to favor images where the object is closer to the camera, yet it falls short in effectively filtering out excessively cropped ones.

Conclusion

We conducted a crowdsourcing study where MTurk participants selected three objects from their environment and trained a model to distinguish between them in real-time using the camera on their mobile phones, and finally selecting smaller subsets from their data that they deem to be of higher value for AI. This allowed us to investigate a subset selection problem with a large participant pool (N = 100), where many non-experts acted as the oracle. After analyzing the F1 scores of our teachable object recognition model, which was trained separately on different subsets selected by both humans and algorithms, we made several observations on how humans value data for AI, and expanded these findings to the accessibility context. We have also laid out a thorough description of our test bed and made the CrowdTeaMa dataset publicly available to allow for study replicability and future comparisons.

References

Barker, M.; Collins, K. M.; Dvijotham, K.; Weller, A.; and Bhatt, U. 2023. Selective Concept Models: Permitting Stakeholder Customisation at Test-Time. *ArXiv*, abs/2306.08424.

Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1): 3–5.

Bullard, K.; Chernova, S.; and Thomaz, A. L. 2018. Human-Driven Feature Selection for a Robotic Agent Learning Classification Tasks from Demonstration. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 6923–6930.

Chen, B.; Wu, T.; Zhang, Y.; Chhetri, M. B.; and Bai, G. 2023. Investigating Users' Understanding of Privacy Policies of Virtual Personal Assistant Applications. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, ASIA CCS '23, 65–79. New York, NY, USA: Association for Computing Machinery. ISBN 9798400700989.

Das, S.; Mandal, S.; Bhoyar, A.; Bharde, M.; Ganguly, N.; Bhattacharya, S.; and Bhattacharya, S. 2020. Multi-criteria online frame-subset selection for autonomous vehicle videos. *Pattern Recognition Letters*, 133: 349–355.

Davidson, S. B.; Gershtein, S.; Milo, T.; Novgorodov, S.; and Shoshan, M. 2022. PHOcus: efficiently archiving photos. *Proceedings of the VLDB Endowment*, 15(12): 3630–3633.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Edmonds, J. 2003. *Submodular functions, matroids, and certain polyhedra*, 11–26. Berlin, Heidelberg: Springer-Verlag. ISBN 3540005803.

Goldman, S. A.; and Kearns, M. J. 1995. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1): 20–31.

Gong, C.; Zheng, Z.; Wu, F.; Shao, Y.; Li, B.; and Chen, G. 2023. To Store or Not? Online Data Selection for Federated Learning with Limited Storage. In *Proceedings of the ACM Web Conference 2023*, WWW '23, 3044–3055. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394161.

Hara, K.; Adams, A.; Milland, K.; Savage, S.; Callison-Burch, C.; and Bigham, J. P. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356206.

Heer, J.; and Bostock, M. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, 203–212.

New York, NY, USA: Association for Computing Machinery. ISBN 9781605589299.

Ho, C.-J.; Slivkins, A.; Suri, S.; and Vaughan, J. W. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, 419–429. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450334693.

Hong, J.; Lee, K.; Xu, J.; and Kacorri, H. 2020. Crowdsourcing the perception of machine teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Jacques, J. T.; and Kristensson, P. O. 2021. Studying Programmer Behaviour at Scale: Anbsp;Casenbsp;Studynbsp;usingnbsp;Amazon Mechanical Turk. In *Companion Proceedings of the 5th International Conference on the Art, Science, and Engineering of Programming*, Programming '21, 36–48. New York, NY, USA: Association for Computing Machinery. ISBN 9781450389860.

Jakubik, J.; Vössing, M.; Kühl, N.; Walk, J.; and Satzger, G. 2024. Data-centric artificial intelligence. *Business & Information Systems Engineering*, 1–9.

Kacorri, H. 2017. Teachable machines for accessibility. *ACM SIGACCESS Accessibility and Computing*, (119): 10–18.

Kacorri, H.; Kitani, K. M.; Bigham, J. P.; and Asakawa, C. 2017. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 5839–5849. New York, NY, USA: Association for Computing Machinery. ISBN 9781450346559.

Kacorri, H.; Syed, A. R.; Huenerfauth, M.; and Neidle, C. 2016. Centroid-based exemplar selection of asl non-manual expressions using multidimensional dynamic time warping and mpeg4 features. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages: Corpus Mining, Language Resources and Evaluation Conference 2016*. European Language Resources Association (ELRA).

Kaushal, V.; Iyer, R.; Kothawade, S.; Mahadev, R.; Doctor, K.; and Ramakrishnan, G. 2019. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1289–1299. IEEE.

Killamsetty, K.; Sivasubramanian, D.; Ramakrishnan, G.; and Iyer, R. 2021. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8110–8118.

Kothyari, M.; Mekala, A. R.; Iyer, R. K.; Ramakrishnan, G.; and Jyothi, P. 2021. Personalizing ASR with limited data using targeted subset selection. *ArXiv*, abs/2110.04908.

Kuzovkin, D.; Pouli, T.; Cozot, R.; Le Meur, O.; Kervec, J.; and Bouatouch, K. 2018. Image Selection in Photo Albums. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, 397–404. New York, NY, USA: Association for Computing Machinery. ISBN 9781450350464.

Lee, K.; Hong, J.; Pimento, S.; Jarjue, E.; and Kacorri, H. 2019. Revisiting blind photography in the context of teachable object recognizers. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 83–95.

Lee, K.; and Kacorri, H. 2019. Hands Holding Clues for Object Recognition in Teachable Machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Li, J.; Zhou, F.; Sun, S.; Zhang, Y.; Zhao, H.; and Liu, P. 2024. Dissecting Human and LLM Preferences. arXiv:2402.11296.

Martin, L.; Whitehouse, N.; Yiu, S.; Catterson, L.; and Perera, R. 2024. Better Call GPT, Comparing Large Language Models Against Lawyers. arXiv:2401.16212.

McHatton, J.; and Ghazinour, K. 2023. Mitigating Social Media Privacy Concerns - A Comprehensive Study. In *Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics*, IWSPA '23, 27–32. New York, NY, USA: Association for Computing Machinery. ISBN 9798400700996.

Microsoft. 2024. Microsoft Copilot in Bing. Accessed: 2024-06-10.

Mink, J.; Yuile, A. R.; Pal, U.; Aviv, A. J.; and Bates, A. 2022. Users Can Deduce Sensitive Locations Protected by Privacy Zones on Fitness Tracking Apps. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.

Møller, A. G.; Pera, A.; Dalsgaard, J.; and Aiello, L. 2024. The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 179–192.

Morrison, C.; Grayson, M.; Marques, R. F.; Massiceti, D.; Longden, C.; Wen, L.; and Cutrell, E. 2023. Understanding Personalized Accessibility through Teachable AI: Designing and Evaluating Find My Things for People who are Blind or Low Vision. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702204.

Narechania, A.; Du, F.; Sinha, A. R.; Rossi, R.; Hoffswell, J.; Guo, S.; Koh, E.; Navathe, S.; and Endert, A. 2023. DataPilot: Utilizing Quality and Usage Information for Subset Selection during Visual Data Preparation. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Pallets. 2024. *Flask Documentation*. Accessed: 2024-06-11.

Payne, J. 2017. A New Angle on Your Favorite Moments with Google Clips. https://blog.google/products/devices-services/google-clips/. Accessed: 2024-06-11.

Raza Syed, A.; and Mandel, M. I. 2023. Estimating Shapley Values of Training Utterances for Automatic Speech Recognition Models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Schreiber, J. M.; Bilmes, J. A.; and Noble, W. S. 2019. apricot: Submodular selection for data summarization in Python. *CoRR*, abs/1906.03543.

Shinohara, A.; and Miyano, S. 1991. Teachability in computational learning. *New Generation Computing*, 8: 337–347.

Singh, A.; Virmani, L.; and Subramanyam, A. 2019. Image Corpus Representative Summarization. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 21–29.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826.

Vaughan, J. W. 2018. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *Journal of Machine Learning Research*, 18(193): 1–46.

vsakkas. 2023. sydney-py · PyPI. Accessed: 2024-06-10.

Walber, T. C.; Scherp, A.; and Staab, S. 2014. Smart Photo Selection: Interpret Gaze as Personal Interest. CHI '14, 2065–2074. New York, NY, USA: Association for Computing Machinery. ISBN 9781450324731.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.