Human Computation (2021) 8:2:76-106

ISSN: 2330-8001, DOI: 10.15346/hc.v8i2.121

Exploring the Use of Deep Learning with Crowdsourcing to Annotate Images

SAMREEN ANJUM, UNIVERSITY OF TEXAS AT AUSTIN

AMBIKA VERMA, COGNEX CORPORATION

BRANDON DANG, AMAZON

DANNA GURARI, UNIVERSITY OF TEXAS AT AUSTIN

ABSTRACT

We investigate what, if any, benefits arise from employing hybrid algorithm-crowdsourcing approaches over conventional approaches of relying exclusively on algorithms or crowds to annotate images. We introduce a framework that enables users to investigate different hybrid workflows for three popular image analysis tasks: image classification, object detection, and image captioning. Three hybrid approaches are included that are based on having workers: (i) verify predicted labels, (ii) correct predicted labels, and (iii) annotate images for which algorithms have low confidence in their predictions. Deep learning algorithms are employed in these workflows since they offer high performance for image annotation tasks. Each workflow is evaluated with respect to annotation quality and worker time to completion on images coming from three diverse datasets (i.e., VOC, MSCOCO, VizWiz). Inspired by our findings, we offer recommendations regarding when and how to employ deep learning with crowdsourcing to achieve desired quality and efficiency for image annotation.

1. INTRODUCTION

Recent successes of scalable image analysis tools are triggering an exciting rise in the number of services that benefit society. For example, systems that create image captions are empowering people with visual impairments to quickly discover what contents are shown in online images (MacLeod et al., 2017). Systems that automatically recognize what objects are in images (e.g., horses, cars, chairs) are enabling online search algorithms to more quickly and accurately return images that show the object a user is searching for. Systems that can localize objects assist people monitoring surveillance videos to spot anomalous (and so suspicious) content (Sodemann et al., 2012). Such examples merely represent the tip of the iceberg of innovative intelligent services that could be deployed with access to accurate, scalable image annotation systems.

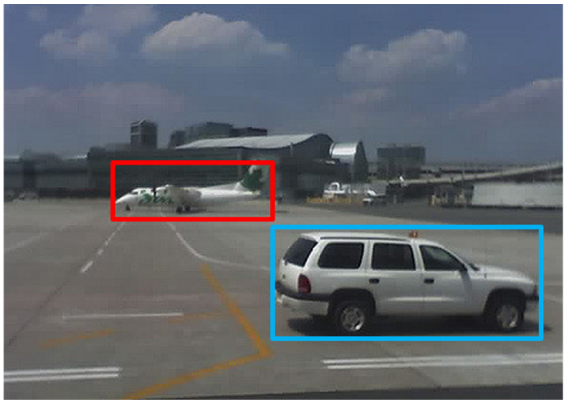


Figure 1. ***Given an image, valuable information to learn about it includes: (i) what objects are present (image classification), (ii) where each object is located (object detection), and (iii) a textual description of it (image captioning). Our goal is to examine what, if any, benefit arises by completing these tasks using hybrid systems that employ the efforts of deep learning algorithms and crowd workers together rather than the status quo of relying on human or algorithms alone.***

A challenge for designing image annotation systems is building them to be inexpensive, fast, and accurate for all image content (i.e., generalizable). While automated options have been preferred for their speed, scalability, and inexpense, they often fail to offer the consistently accurate annotations humans can deliver. Accordingly, system designers have experimented with numerous workflows that demonstrate how to decompose an image analysis task to take advantage of the aforementioned benefits of algorithms while embracing crowdsourced workers to ensure accuracy (Cheng and Bernstein, 2015; Guo, 2018; Gurari et al., 2016; Hara et al., 2014; Laput et al., 2015; Wigness et al., 2015; Zhang et al., 2013). However, such workflows have only been studied for specific applications. The limited reports of success for such workflows begs a question of how well such hybrid algorithm-crowd workflows would generalize to different image analysis tasks and images.

The goal for this work is to establish generalizable best practices for employing hybrid algorithm-crowd partnerships to analyze images. We propose a framework to examine the effectiveness of multiple hybrid algorithm - crowd workflows to annotate images across different tasks, datasets and algorithms. To do so, we employ this framework to benchmark these workflows for three distinct tasks, specifically image captioning, image classification, and object detection (as exemplified in Figure 1). We evaluate various implementations of this framework on images taken by sighted and blind people. We use deep learning algorithms in these workflows since they offer state-of-art performance for image annotation. Deep learning is a form of artificial intelligence that learns to recognize from a large number of examples the meaningful patterns for making predictions. We report the trade-offs between annotation quality and required human effort (i.e., time spent) for each workflow in order to offer generalizable conclusions about the strengths and weaknesses of the various workflows. Such information is critical to enable system designers to make informed decisions,

based on their in-house demands (e.g., time constraints, human effort constraints, computational constraints), regarding what is the best approach for their data analysis needs.

2. RELATED WORKS

Image Analysis Methods The pioneering crowdsourcing work demonstrated it is possible to efficiently annotate massive datasets with anonymous crowd workers, including to classify what objects are present (Von Ahn and Dabbish, 2004), locate the objects in images (Von Ahn et al., 2006b), and describe images (Von Ahn et al., 2006a). More recently, deep learning methods have emerged that can achieve human performance for a small set of tasks such as classifying and detecting pre-specified objects (e.g., 1000 categories chosen for the ImageNet dataset (Russakovsky et al., 2015)). However, little is known about the general-purpose value of solutions that lie in between these two extremes of relying exclusively on either crowds or algorithms. Accordingly, we benchmark methods across different tasks, datasets, and algorithms to provide generalizable design recommendations for when and how to employ deep learning with crowdsourcing to achieve desired quality-efficiency trade-offs.

Design-Time Algorithm-Crowd Partnerships Numerous papers have proposed algorithm-crowd partnerships to improve the utility of machine learning algorithms with less human effort. This is because large training datasets are often needed to achieve strong predictive performance. Solutions show how to intelligently solicit as few annotations from humans as possible to yield a desired performance from machine learning solutions (Laput et al., 2015; Wigness et al., 2015). Our work differs in that we aim to examine effective partnerships that minimize human effort at run-time rather than focus on how to more efficiently train fully-automated solutions.

Run-Time Algorithm-Crowd Partnerships Numerous papers examine algorithm-crowd partnerships which supply human involvement to compensate for algorithms' weaknesses at run-time. Solutions have been explored for specific tasks such as finding answers to questions (Bernstein et al., 2012), detecting things in images (Hara et al., 2013), assistive navigation (Guo, 2018), image captioning (Guinness et al., 2018), and more (Rzeszotarski and Kittur, 2012; Song et al., 2019; Cheng and Bernstein, 2015; Gurari et al., 2016; Hara et al., 2014; Konyushkova et al., 2017; Lofi and El Maarry, 2014; Sabou et al., 2013; Salisbury et al., 2017). Despite that such solutions can successfully leverage the independent strengths of algorithms and crowds together, a person must typically invest substantial effort to establish which hybrid solution is suitable for their purposes. Accordingly, we analyze the general benefit of several hybrid workflows that are popular for image annotation: (1) algorithm verification, which was previously used for object detection (Hara et al., 2014; Konyushkova et al., 2017), (2) algorithm correction, which was previously used for image captioning (Salisbury et al., 2018), and (3) replacing low confidence results from algorithms, which was previously used for object detection (Gurari et al., 2016; Hara et al., 2014). We benchmark these workflows for three tasks using multiple datasets and deep learning algorithms to better understand the general (dis)advantages of each.

Crowdsourcing Workflows While few papers explicitly define the term workflow, the predominant implicit assumption is that it refers to a multi-stage process of decomposing a complex task into simpler micro-tasks, as explicitly defined by Kittur et al. (Kittur et al., 2013). Some work has offered generalized principles for designing a workflow; e.g., discussions of specific patterns such as generate-apply-edit (Chilton et al., 2013) or summaries of many patterns including fix-verify-pattern (Quinn and Bederson, 2011; Sabou et al., 2013) . More recently, some works have designed workflows that integrate both human and algorithm effort into hybrid architectures (Lasecki et al.,

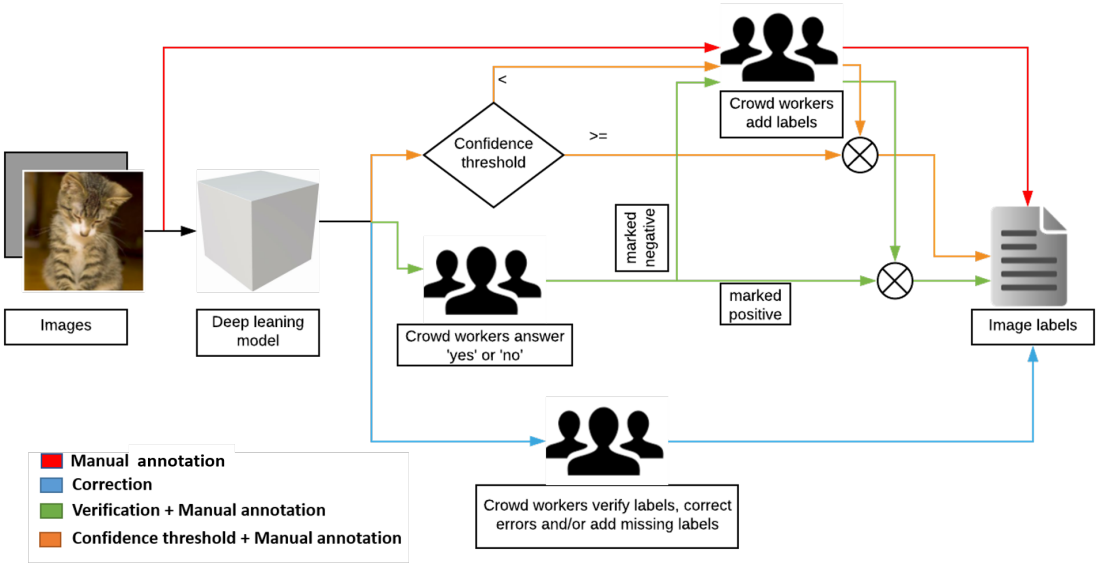


Figure 2. ***Summary of the image annotation workflows in our framework. Given an image, either (i) workers produce annotations (Manual annotation), (ii) workers verify labels predicted by algorithms for correctness with negative outcomes sent for manual annotation by workers (Verification + Manual annotation), (iii) workers verify and then correct any inaccurate algorithm-predicted labels (Correction), or (iv) algorithms predict labels with confidence scores in their predictions and images with low confidence predictions (below a pre-defined threshold) are sent for manual annotation by workers (Confidence threshold + Manual annotation).***

2014; Lofi and El Maarry, 2014; Lundgard et al., 2018). Our work most closely relates to works that aim to offer concrete guidance to users on which is the most appropriate workflow for their purposes (Weld and Dai, 2011; Lin et al., 2012; Gurari et al., 2016). We explore three different hybrid workflows that integrate the efforts of crowds with algorithms on three significant image annotation tasks, and discuss the impact of these workflows in terms of speed, quality, and costs involved.

3. METHODS

We propose a modular framework that supports three distinct hybrid workflows for partnering crowds and algorithms to collaboratively annotate images. We designed this framework to support three popular image annotation tasks: image classification, object detection, and image captioning. We first describe the three tasks and then define each hybrid workflow. Finally, we explain our implementation of the crowdsourcing systems and deep learning systems used for each task.

3.1. Annotation Tasks

Our framework enables users to investigate different workflows with respect to three popular image analysis tasks: image classification, object detection, and image captioning. In what follows, we describe these tasks and exemplify the output from executing each task in Figure 3:

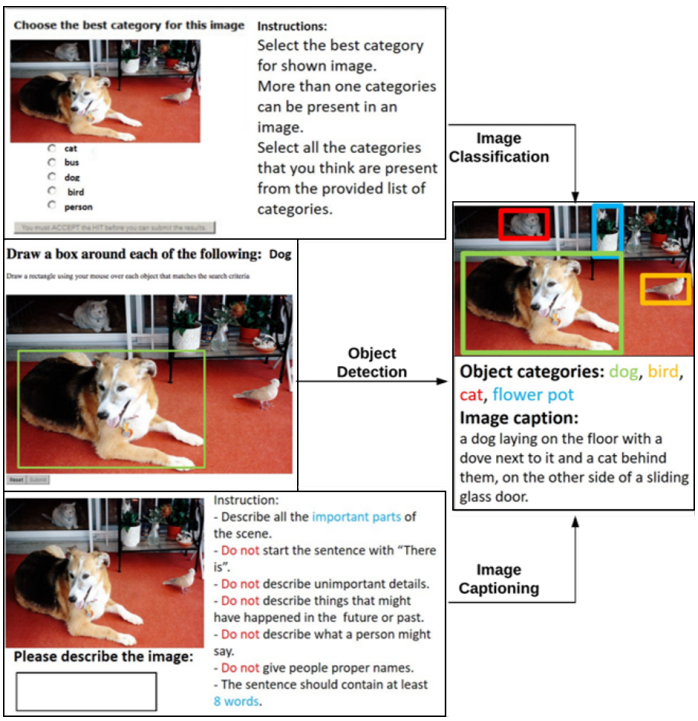


Figure 3. ***Illustration of the three image annotation tasks - image classification, object detection and image captioning. As shown, object classification yields a list of class labels, object detection yields rectangular boundaries (i.e., bounding boxes) demarcating where objects are located, and image captioning yields a description of the image. Also shown are screen-shots of the user interfaces we employed to collect manual annotations from crowd workers.***

Image classification**:** the task is to list labels that indicate what objects are present in a given image. This is valuable for many applications including content based image search and retrieval (e.g. Google Images), assisting visually-impaired individuals to recognize objects in their visual surroundings (Kacorri et al., 2017), as well as fine-grained identification of cars and food items (Zhou and Lin, 2016).

Object detection**:** the task is to locate each object present in a given image with a bounding rectangle. This is valuable for many applications including self-driving cars for environment understanding (e.g. Waymo, Uber), surveillance applications (Cohen and Medioni, 1999), face detection to automatically tag pictures (e.g. Facebook), and activity recognition scenarios such as tele-rehabilitation (Pirsiavash and Ramanan, 2012).

Image captioning**:** the task is to generate a text description of an image. Valuable applications include correlating varied multimedia sources (Pan et al., 2004) and assisting people with visual impairments to interpret images (Kacorri et al., 2017; MacLeod et al., 2017).

3.2. Annotation Workflows

For each task, we implement the following four annotation workflows that employ different levels of effort from crowd workers and deep learning models:

***Manual annotation*:** workers label images from scratch without any input or assistance from a deep learning algorithm. This approach is in line with conventional approaches used previously to create large-scale datasets that describe images (Chen et al., 2015; Von Ahn et al., 2006a), classify objects (Deng et al., 2009; Von Ahn and Dabbish, 2004), and locate objects in images (Lin et al., 2014; Von Ahn et al., 2006b).

***Verification*:** a deep learning system generates labels which are then verified by workers as being correct or not. Specifically, workers answer simple *yes* or *no* questions. Verification methods have been used in numerous machine learning pipelines (Hara et al., 2014; Konyushkova et al., 2017; Krasin et al., 2017; Papadopoulos et al., 2016). It is important to note that, by employing verification, we do not obtain a fully annotated dataset since negatively verified images still require correct labels. Images which are negatively marked after verification are fed back through the *manual annotation* pipeline.

***Correction*:** labels are again predicted by a deep learning system. However, in addition to verifying a prediction, a worker is also tasked with fixing inaccurate annotations by modifying the algorithm-generated results. This approach is explored by (Gaur et al., 2016) and implemented in (Harrington and Vanderheiden, 2013; Huang et al., 2017; Salisbury et al., 2018).

***Confidence threshold*:** a deep learning system that generates results also outputs confidences value (between 0 and 1) in its predictions. Predictions for which the confidence value is above a predefined threshold value are kept, and all remaining images are fed back through the *manual annotation* pipeline. This approach is explored in (Gurari et al., 2016; Hara et al., 2014).

The aforementioned workflows are visualized in Figure 2. As illustrated, these workflows offer various ways to decompose data annotation into micro-tasks that can be distributed between algorithms and crowd workers.

3.3. Crowdsourcing Systems

We design a "Human Intelligence Task" (i.e., HIT) to support the crowdsourcing effort involved with each workflow for each of the three image annotation tasks. In total, we designed nine crowdsourcing systems. All code will be shared publicly upon publication (*https://anonymous-link*). We summarize the system designs below.

***Image classification*:** In the *manual annotation* HIT, a worker identifies what objects are present in a given image by selecting choices from a pre-set list of 20 object categories. In the *verification* HIT, a worker is shown the top object category predictions from a deep learning model for a given image and must indicate whether each prediction is correct or not. The *correction* HIT is identical to the verification HIT, except a worker can additionally select from the list which categories are present in the image but were missed by the model as well as remove erroneous labels predicted by the deep learning model.

***Object detection*:** In the *manual annotation* HIT, a worker is given an image and an object type and must draw a tight bounding box over an instance of that object. In the *verification* HIT, the worker is shown both a bounding box and corresponding object category prediction from a deep learning model and must indicate whether the bounding box indeed covers an instance of

Figure 4. ***Example images from the three datasets used in our experiments to illustrate the diversity of content: (a) VOC2012, (b) MSCOCO, and (c) VizWiz. These images were taken by both people who are sighted (a, b) and blind (c). Each dataset has distinct properties, varying in terms of scene complexity (e.g., typically VOC2012 is simple and MSCOCO is complex), location of objects (e.g., objects often are centered in VOC2012), and image quality (e.g., many images in VizWiz are lower quality).***

the predicted object category. The *correction* HIT is identical to the verification HIT, except workers can additionally modify the bounding box or modify the category label.

***Image captioning*:** In the *manual annotation* HIT, workers must manually annotate the image from scratch by typing in a textual description. In the *verification* HIT, the worker is shown a predicted caption/description for a given image and answers a simple yes or no question. The *correction* HIT is identical to the verification HIT, though workers can additionally provide an updated caption if applicable.

3.4. Deep Learning Systems

We utilize modern deep learning models pre-trained on large-scale, public datasets in our workflows. Such models learned complex representations by training from many labeled examples. We describe our model choices below.

***Image classification*:** We employ pre-trained models which have achieved high performance in benchmarking challenges such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015) and MSCOCO Challenge (Lin et al., 2014). We include the following three architectures - **VGG** (ILSVRC 2014 localization winner and classification runner-up), **ResNet** (ILSVRC 2015 and MSCOCO 2015 winner), and **Google Inception** (ILSVRC 2015 top ranked unofficial). All implementations are pre-trained on the ILSVRC 2012 training dataset and are available in TensorFlow[1].

***Object detection*:** We incorporate support for the recent high performing architecture **Faster R-CNN** (ILSCRC 2015 and MSCOCO 2015 top ranked), which is open source and pre-trained on ILSVRC dataset (with fine-tuning on VOC2007 dataset). VOC2012 and MSCOCO datasets have no overlap with VOC2007, hence this model can be used. We also employ **SSD** (Single

[1]Though object detection models provide category label outputs as well, we choose to include models designed specifically for object classification to avoid accruing the extra computational complexity required for object detection.

Shot Detector), however we only use it with the MSCOCO dataset since it is trained with fine-tuning on the VOC 2007 and 2012 datasets.

***Image captioning*:** We integrate support for Microsoft (MS) Azure's computer vision API, a cloud based service that allows users to run complex algorithms without worrying about computational resources required. We employ its API support for image captioning.

4. EXPERIMENTAL DESIGN

We now describe our studies for examining what, if any, advantages arise by employing deep learning as a part of the crowdsourcing pipeline for image annotation. We focus on the following research questions: (1) Can the annotation quality from hybrid approaches match the quality from human annotations? and (2) What are the trade-offs between quality and human effort for different annotation protocols?

4.1. Datasets

We conduct our studies on 500 images coming from three diverse publicly-available datasets. We chose images taken by both sighted (Everingham et al., 2010; Lin et al., 2014) and blind people (Gurari et al., 2018) in order to address the interests from different plausible users of image annotation systems. We also chose images from well-established computer vision datasets so that we could have access to ground truth annotations for evaluation for the various tasks we are studying in this paper. We describe each dataset below and show examples from each in Figure 4.

Pascal VOC Challenge dataset (VOC2012) (Everingham et al., 2010): This dataset emerged as the first widely-accepted benchmarking dataset in the computer vision community for object classification and detection. The dataset consists of images showing 20 object categories with some variability in object size, orientation, pose, illumination, position, and occlusion. We randomly sample 10 images from each of the 20 categories present in VOC2012: airplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dog, horse, motorcycle, person, potted plant/flower pot, sheep, sofa/couch, table, train, and TV/television. Images are sampled from the training set, resulting in a total of 200 instances. This dataset comes with ground truth annotations for the image classification and object detection tasks.

MSCOCO dataset (MSCOCO) (Lin et al., 2014): This dataset was designed to overcome the simplicity of VOC2012 by providing more complex scenes that show multiple objects. Objects typically vary in scale and location (i.e., non-centered), and may be partially obscured by other objects or the environment. While the total dataset consists of 80 categories, in our experiments, we use a subset of the dataset that contains the same object categories as VOC2012. We again randomly sample 10 images from each of the 20 categories listed above, resulting in another set of 200 images. Because these images depict complex scenes, objects can co-occur, potentially resulting in some categories having more than 10 instances. For instance, an image sampled from the "horse" category may also contain a "person" riding the horse. This dataset comes with ground truth for the image classification, object detection, and image captioning tasks.

VizWiz dataset (Gurari et al., 2018): This dataset contains images taken by blind photographers who used the cameras on their mobile phones. It is unique in that each image is captured by a blind person who also recorded a spoken question and submitted those to crowd workers to receive an answer. As such, images typically represent familiar everyday objects the people want to learn about. Furthermore, images may be poorly framed, improperly focused, or blurry since blind people

cannot verify the quality of the images. Because VizWiz is not designed with ground truth for the three tasks we examine, we created it in-house. To do so, we first selected images from the training set which had sufficiently high quality to recognize the content, and then manually selected 100 images which contained objects from the selected 20 object categories used for VOC2012 and MSCOCO. Next, to establish ground truth for image classification, we followed the manual annotation and correction methods described previously. For image captioning, since blind users commonly asked generic questions to learn what is in an image (Brady et al., 2013), we used the 10 answers provided per image in the VizWiz dataset as a proxy for image captions.

4.2. Crowdsourcing Set-Up

We employ crowd workers from Amazon Mechanical Turk (AMT), an online platform where requesters post tasks for workers to complete in exchange for pay. To reduce concerns about quality, we only accepted workers who already completed more than 500 Human Intelligence Tasks (i.e., HITs) and had an acceptance rate of greater than 92%. Towards collecting high quality annotations, we collected redundant annotations. Specifically, we assigned each HIT to three workers. We paid \$0.20 for completion of each HIT, where each HIT included 10 images.[2]

We measured **human effort** by adapting the freely-available *Mmm*Turkey framework (Dang et al., 2016) to collect on-focus task completion time, total number of clicks, and mouse movements. While all behavioral information will be publicly-shared upon publication to support future work, we analyze task completion time in this work.

To evaluate the **quality** of results obtained using the different hybrid workflows as well as the status quos of manual annotation and algorithms, we employ traditional evaluation metrics for each task as described below.

4.3. Study 1: Image Classification

We evaluate *classification quality* using the standard technique (Deng et al., 2009) of computing the percentage of exact matches between the output of an annotation protocol and the ground truth classifications. The annotation is correct when the aggregated object categories match all those present in the corresponding ground truth labels. For crowdsourced annotations, we aggregate the three workers' submissions into a single annotation using majority voting, i.e., a category is aggregated if at least two of the three workers agreed it was present. For algorithms, to minimize concerns that our analysis is over-fitting to the behaviors of a particular algorithm, we evaluate each hybrid workflow with three different deep learning models—VGG, ResNet, and Inception. Consequently, we evaluate nine hybrid algorithm-crowdsourcing workflows (i.e., 3 workflows x 3 deep learning models). For the *confidence threshold* workflow, we set the threshold for model confidence to 0.65 (0 being least and 1 being most confident) to emulate the majority vote confidence required from crowdsourcing (i.e., agreement by 2 of 3 people). We apply a distribution-free bootstrap test (Efron and Tibshirani, 1994) to compare the results for each benchmarked method to those for manual annotation with respect to quality as well as the time taken to complete the annotations. This statistical test resamples the results with replacement multiple times to create simulated results. We resample 10,000 times and deem a method to be significantly different when the p-value < 0.05.

[2]This payment yields a \$9 per hour wage, as the median completion time for all HITs was 80 seconds.

4.4. Study 2: Object Detection

We evaluate *detection quality* with mean IOU (intersection over union) of the bounding boxes, i.e., the fraction of pixel overlap between the bounding boxes of the same object category produced by the workflow and the ground truth. Following prior work (Everingham et al., 2010), positive detections have IOU scores greater than 0.5. To avoid inadvertently biasing our findings to the performance of a single algorithm, we evaluate each hybrid workflow with two deep learning models—Faster RCNN and SSD. Consequently, we evaluate six hybrid algorithm-crowdsourcing workflows (i.e., 3 workflows x 2 deep learning models). For the *confidence threshold* workflow, we chose as the threshold the average value for all confidence values reported by prediction models across all images. We again apply a distribution-free bootstrap test to compare the results for each benchmarked method to those for manual annotation with respect to quality as well as the time taken to complete the annotations, with p-values < 0.05 indicating a significant difference. We compute significance values only on algorithms that could be applied to all the images.[3]

4.5. Study 3: Image Captioning

Following the classical technique used for *image captioning* evaluation, quality is measured using the average BLEU (Bilingual Evaluation understudy) scores (Papineni et al., 2002). This measure indicates the number of words/phrases in the predicted caption that are also found in the ground truth captions, with 0 being an incorrect match and 1 being the best possible match. Since each image is annotated thrice by different crowd workers, we average the BLEU score computed in each case for the final score. For the algorithm, we use the off-the-shelf MS Azure captioning API, resulting in us evaluating three hybrid algorithm-crowdsourcing workflows (i.e., 3 workflows x 1 deep learning model). For the workflows, we chose the same threshold for *confidence threshold* as was employed for the object detection task. We again apply a distribution-free bootstrap test to compare the results for each benchmarked method to those for manual annotation with respect to quality and the time taken to complete the annotations, with p-values < 0.05 indicating a significant difference.

5. EXPERIMENTAL RESULTS

We now report our findings for the quality and human effort resulting from evaluating the three hybrid algorithm-crowd workflows, manual annotation, and algorithms.

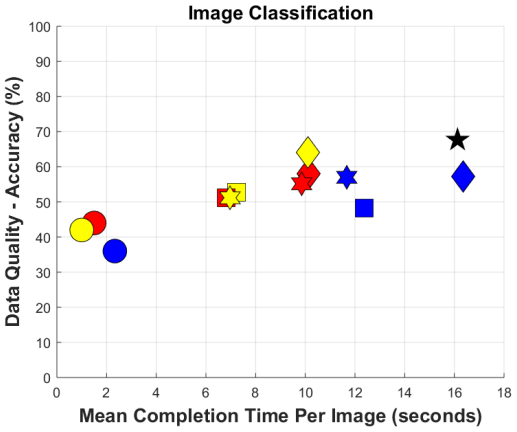
5.1. Study 1: Image Classification

The trade-offs between annotation quality and human effort for each workflow implementation for all images in the three datasets (VOC2012, MSCOCO, VizWiz) are shown in Figure 5. Significance testing results are shown in Table 1.

While the **highest quality** labels come from the *manual annotation* protocol, significance testing shows that most of the hybrid workflows produce comparable results to manual annotation. Specifically, all three algorithms lead to comparable results for the *correction* workflow whereas two of the three algorithms lead to comparable results for the *verification* and *confidence threshold* workflows.

Our findings also highlight what one may expect when there are **practical constraints such as limited time**. The fastest option is *algorithm only* followed by *verification* and *confidence threshold*. Recall that for both of the hybrid workflows a different worker is recruited to manually annotate

[3]Recall from the description of “Deep Learning" systems that SSD was trained on VOC and so could not be tested on that dataset"



Manual annotation			
	VGG	ResNet	Inception
Algorithm only			
Correction			
Verification + Manual annotation			
Confidence threshold + Manual annotation			

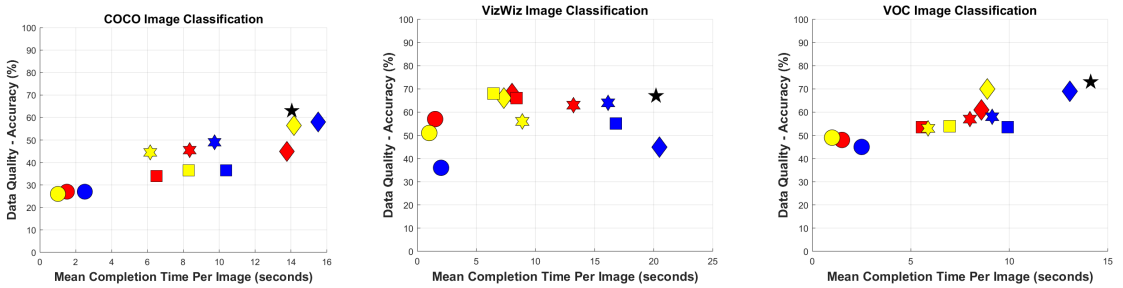
Figure 5. ***Shown is the performance of three workflows for distributing varying levels of human involvement with deep learning algorithms to classify images. Each plot shows the mean quality resulting from all pairings of a workflow with a deep learning algorithm and the corresponding average time to annotate each image for all images in VOC2012, MSCOCO, and VizWiz. The data quality refers to the percentage of exact matches between the output of an annotation protocol and the ground truth classifications. (best viewed in color)***

	Quality			Time		
	VGG	ResNet	Incep	VGG	ResNet	Incep
Algorithm	**0.02**	0.05	**0.04**	**0.02**	**0.01**	**0.01**
Correction	0.16	0.17	0.4	0.9	0.07	0.1
Verification + Manual ann.	**0.04**	0.14	0.11	0.18	**0.03**	**0.03**
Confidence threshold + Manual ann.	0.08	0.07	**0.03**	0.14	0.05	**0.03**

Table 1. ***p-values obtained from significance tests for how each method compares to manual annotation with respect to quality and time to annotate for the image classification task. Values shown in bold are statistically significant. As shown, most hybrid workflows are statistically similar in terms of quality to manual annotation. Additionally, most hybrid workflows are statistically different (in terms of being faster) than manual annotation. (Incep = Inception; ann = annotation)***

images deemed poor quality by an algorithm (i.e., *confidence threshold*) or crowd worker (i.e., *verification*). This suggests that decomposing the task to treat verification and re-annotation separately requires less human effort than relying on a single worker to perform both subtasks (i.e., *correction*).

Overall, an effective tool for achieving a desirable **quality-effort trade-off** is the *correction* protocol (marked by diamonds). This approach achieved quality statistically similar to *manual annotation* while considerably reducing human effort; i.e., by roughly 6 seconds per image from 16 seconds to



Manual annotation			
	VGG	ResNet	Inception
Algorithm only			
Correction			
Verification + Manual annotation			
Confidence threshold + Manual annotation			

Figure 6. ***Shown is the performance of different workflows for distributing varying levels of human involvement with algorithms to classify images in three datasets: MSCOCO, VizWiz, and VOC2012.***

10 seconds per image for two of the three algorithms (ResNet and Inception).

We performed qualitative thematic coding (Clarke and Braun, 2014) to identify common reasons behind errors. We compared the best performing workflow (correction) with the worst performing workflow (verification) for image classification, since image classification is a key step for all three studied image annotation tasks (i.e., it is a precursor for completing object detection and image captioning). We conducted our analysis on a sample of 200 images from the MSCOCO dataset using the initial results obtained from the VGG model. We used the ground truth of the datasets that shows segmentations of objects in each image in order to analyze the frequency of errors based on whether an image contains a single object versus multiple objects. We then coded errors across all images into three categories: (a) *incomplete* classification, where the resulting output of the hybrid workflow contained only a subset of objects in the image, (b) *misclassification*, where the resulting output of the hybrid workflow contained objects not present in the image, (c) *unclassified*, where the resulting output of the hybrid workflow was empty.

Across both workflows, errors were more common in images that show multiple objects than those with a single object; i.e., over 68% and 96% of the 84 images with multiple objects were incorrectly classified for *correction* and *verification* respectively versus 8% and 21% of the 116 images with single objects for *correction* and *verification* respectively. In terms of types of errors, of the 66 images incorrectly classified with *correction*, the majority of the errors (i.e., 79%) arose from *incomplete* classification, 15% from images being *unclassified* and the rest (i.e., 6%) were a result of

misclassification. With regards to the types of errors observed in the 105 images incorrectly classified with *verification*, while *incomplete* classification still was the major reason for errors (i.e., 55%), a considerably large portion of the errors (i.e., 44%) also emerged from images being *unclassified*. Upon further observation, we found the latter group to be dominated with images showing single objects. These findings underscore that *correction* considerably outperforms *verification* because it can overcome the issue that the initial results provided by the algorithms were insufficient; i.e., correction could correctly classify the many images with single objects that were left *unclassified* and the many images with multiple objects that were *incompletely* classified. We suggest in Section 6 a potential different workflow that may achieve the faster speeds of the verification workflow while overcoming this issue that the initial results provided by the algorithms are insufficient.

Different **deep learning models** were used to not only examine the stability of our findings for various protocols but also to offer practical guidance. In terms of computation costs, the run-time costs for generating predictions are negligible when using GPU resources and much more when relying on CPUs (i.e., $\sim$2 seconds per image for VGG and 1 to 1.5 seconds for ResNet and Inception on a standard Intel 7th generation 4 core CPU). Consequently, Inception emerged as the ideal deep learning model across all the datasets not only in terms of quality but also computational time. On the other hand, workflows involving the VGG model took the most amount of time across all datasets, in most cases comparable to the time taken in *manual annotation*.

Figure 6 shows the trade-offs between annotation quality and human effort for each workflow implementation with respect to each of the three datasets (VOC2012, MSCOCO, VizWiz). These plots parallel those shown in Figure 5, while offering finer grained insights.

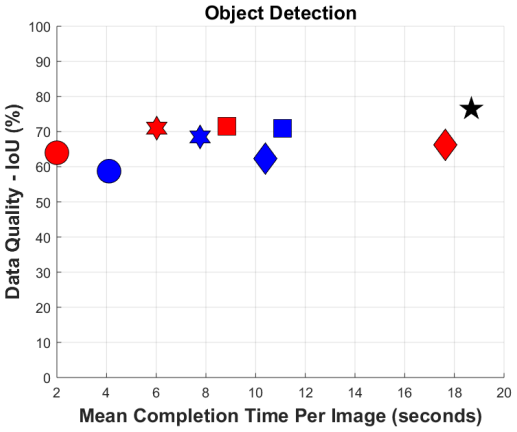
One difference between datasets that we observe is that both *correction* and *verification* yield slightly higher quality results than *manual annotation* on the VizWiz dataset, while also eliminating roughly 67% and 65% of human effort respectively. We hypothesize this is because more concentration (and so time) is needed to identify objects when reviewing the lower quality images that common in VizWiz since the images are taken by people who are blind. In this case, we suspect the algorithms effectively shortcut the time needed to locate the plausible regions in the images.

Another difference that we observe between datasets is that human effort speed ups occur only for the images in VOC2012 and VizWiz, and not for the images in MSCOCO. This observation is consistent with previous revelation that *correction* yields better results for images with a single object (common for VOC2012 and VizWiz) than for complex images containing multiple objects (common in MSCOCO). Accordingly, we hypothesize the *correction* workflow is a poorer fit for complex images because it takes more time for workers to locate the object of interest in complex scenarios where there can be great variability in an object's size, orientation, and pose.

5.2. Study 2: Object Detection

Figure 7 shows the trade-offs between annotation quality and human effort for each workflow for the two datasets which include ground truth annotations (VOC2012, MSCOCO). For significance testing, we found that all hybrid workflows are statistically similar in terms of quality to manual annotation and most hybrid workflows are statistically different (in terms of being faster) than manual annotation.

Overall, *manual annotation* again results in the **highest quality** labels. Interestingly though, quality of most other hybrid protocols are within 3-15 percentage points of that obtained from *manual anno-*



Manual annotation		
	Faster RCNN	SSD
Algorithm only		
Correction		
Verification + Manual annotation		
Confidence threshold + Manual annotation		

Figure 7. ***Shown is the performance of three workflows for distributing varying levels of human involvement with deep learning algorithms to detect objects in images. Each plot shows the mean quality resulting from all pairings of a workflow with an algorithm and the corresponding time per image required to annotate all images in two datasets: VOC2012 and MSCOCO. The data quality is evaluated using IoU (intersection over union), i.e., the fraction of pixel overlap between the bounding boxes of the same object category produced by the workflow and the ground truth.***

tation, while requiring around half the amount of human effort. This observation is supported by the significance tests which shows that the results obtained using all hybrid workflows are statistically similar to the ones obtained from *manual annotation*.

For **time-constrained** applications, the fastest option is *algorithm only* followed by *confidence threshold* and then *verification*. This trend largely resembles that observed for the image classification task. In addition, the significance tests for this task show that the time taken by all hybrid workflows are similar to that of the fastest option (*algorithm only*).

Overall, the most promising option for achieving a desirable **quality-effort trade-off** is *confidence threshold*. It achieves 71% (for SSD) and 69% (for Faster RCNN) in data quality while requiring less than one half of the human effort required for *manual annotation*. The results obtained are significantly better than those obtained using the algorithmic workflows. Moreover, *confidence threshold* skips the verification crowdsourcing step needed in the *verification* workflow, leading to additional savings in cost and human effort with almost no loss to quality. As for the *correction* workflow, it typically offers no benefits in terms of improving quality or reducing human effort. We attribute this to the complexity of the task where we observed crowd workers either corrected the predicted object category or fixed the bounding box boundaries but not both.

Again, we employed different **deep learning models** to not only examine the robustness of our findings but also to provide practical guidance. Overall, workflows using SSD resulted in higher quality and faster results than Faster-RCNN. Faster-RCNN is an older model, and we attribute its poorer performance in part to limitations in its architecture and training data (Faster-RCNN is

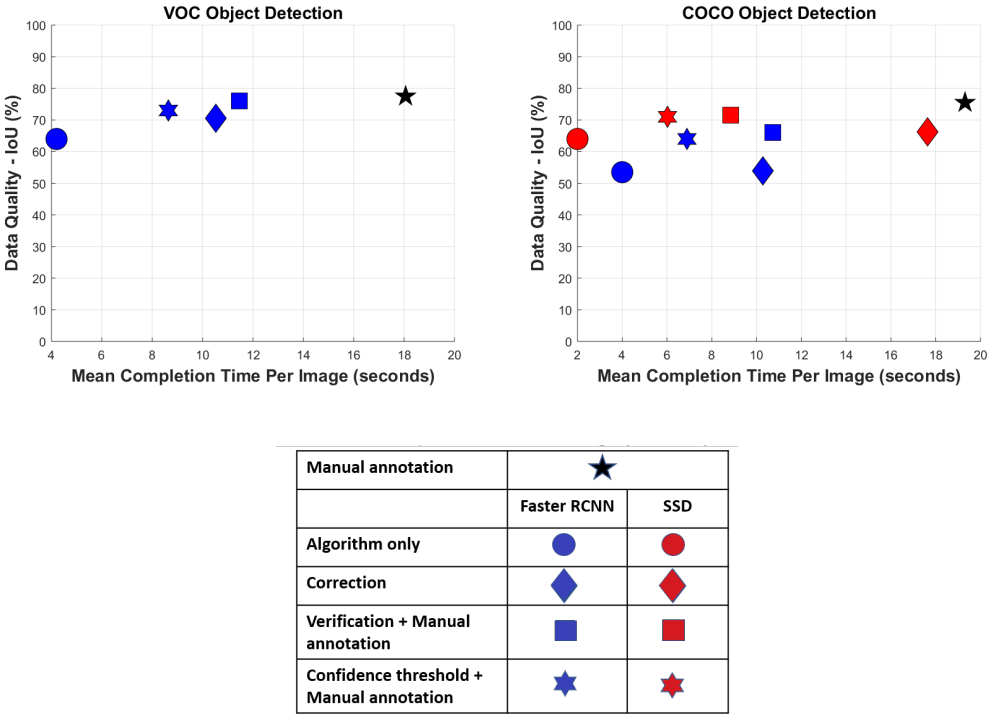


Figure 8. ***Shown is the performance of several workflows for distributing varying levels of human involvement with deep learning algorithms to detect objects in images from VOC2012 and MSCOCO.***

trained on VOC 2007 whereas SSD is trained on both VOC 2007 and VOC 2012).

	Quality	**Time**
	Microsoft Azure API	Microsoft Azure API
Algorithm	0.05	**0.0345**
Correction	0.86	0.08
Verification + Manual annotation	0.25	**0.0462**
Confidence threshold + Manual annotation	0.12	**0.0478**

Table 2. ***p-values obtained from significance tests comparing each method to manual annotation with respect to quality and time to annotate for image captioning. Values shown in bold are statistically significant. All hybrid workflows are statistically similar in terms of quality and most hybrid workflows are statistically different (in terms of requiring less human effort) compared to manual annotation.***

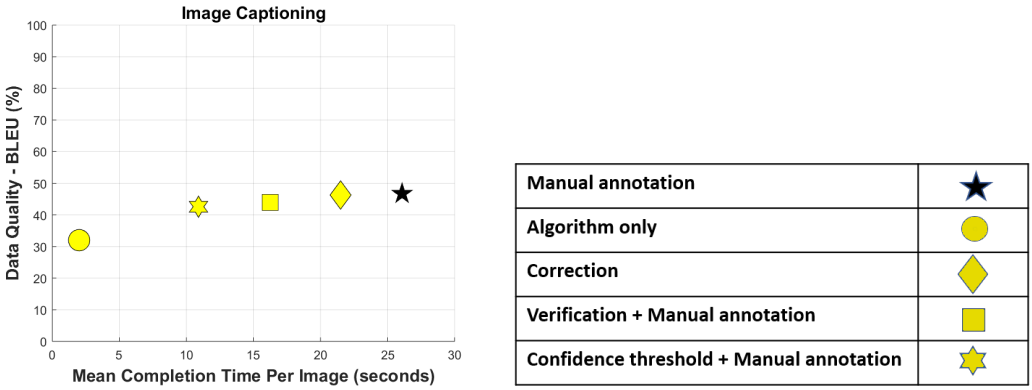


Figure 9. ***Shown is the performance of three workflows for distributing varying levels of human involvement with a deep learning algorithm to caption images. Each plot shows the mean quality and the corresponding time per image required to complete annotating all images in two diverse datasets: MSCOCO and VizWiz. The data quality is evaluated using BLEU scores which computes the number of words/phrases in the predicted caption that are also found in the ground truth captions.***

Figure 8 shows the trade-offs between annotation quality and human effort for each workflow for both datasets—VOC2012 and MSCOCO, offering finer grained insights than shown in Figure 7. One key difference that we observe between datasets is for VOC2012 on which all hybrid workflows reduce human effort by more than half the time taken in *manual annotation*, and *verification* yields higher quality results than other workflows. In contrast, for MSCOCO, the *correction* workflow takes longer, while both *verification* and *confidence threshold* workflows yield better results in terms of quality as well. We hypothesize this difference is due to the greater complexity of the images in MSCOCO than contained in VOC2012.

5.3. Study 3: Image Captioning

Figure 9 shows the trade-offs between annotation quality and human effort for each workflow for the two datasets for which we have ground truth (MSCOCO and VizWiz). Table 2 shows the significance testing results.

While *manual annotation* results in the **highest quality** results, it also requires extensive human effort. We found that quality of all considered workflows produce results within five percentage points of the quality obtained from *manual annotation*, with human effort reduced by roughly 20-60% for both datasets. The *correction* workflow offers the highest quality results among all hybrid approaches, achieving quality within one percentage point of that obtained from solely relying on human effort, and so offering the best **quality-effort trade-off**. Significant tests also confirmed that the quality of all three hybrid workflows are statistically similar to that of manual annotation and that *correction* workflow performed better than the other two hybrid workflows.

When applications impose **practical constraints such as limited time**, other hybrid workflows yield promising results in terms of reducing effort while yielding only slightly lower quality than

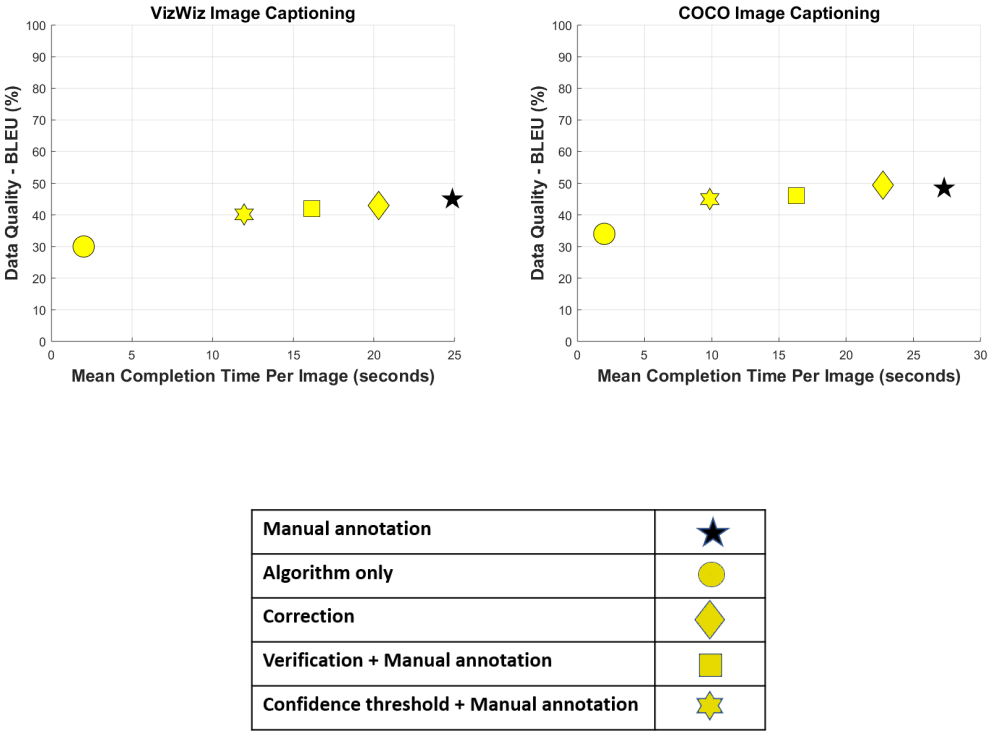


Figure 10. Shown is the performance of three workflows for distributing varying levels of human involvement with a deep learning algorithm to caption images coming from VizWiz and MSCOCO.

that of the top-performing *correction* workflow. Based on the significance tests, both *verification* and *confidence threshold* perform significantly faster than *manual annotation*. For example, *confidence threshold* returns results the fastest, while reducing effort by approximately 60% compared to relying exclusively on humans. This is exciting since it still yields quality within five percentage points of relying exclusively on humans. The effort reduction is attributed to fewer images being crowdsourced, since the verification step is skipped and instead decided automatically based on the algorithm's confidence. Also, computational costs for obtaining labels from algorithms are small for cloud computing services such as Microsoft Azure API.

Figure 10 shows the trade-offs between annotation quality and human effort for each workflow implementation for all images in two datasets (MSCOCO, VizWiz). Our findings show that all hybrid workflows follow the same trend where the quality of all workflows is similar or better than that of (*manual annotation*). In terms of time, *confidence threshold* was the fastest of all three hybrid workflows, followed by *verification* and *correction*. Moreover, unlike the observations for previous tasks (image classification, object detection), the individual performance of all three hybrid workflows in this task across both VizWiz and MSCOCO datasets was consistent in terms of both quality and time.

6. DISCUSSION

From the trends across all tasks, datasets, and algorithms, we identify several recommendations for system designers facing different constraints. Our experimental results highlight which hybrid workflows may generalize well across various tasks and demonstrate the relationship of human-time with the quality of results. We offer our findings as valuable context to inform future design choices as well as reflect on past hybrid design choices employed for applications such as detecting inaccessible sidewalks (i.e., used *verification*) (Hara et al., 2014), describing images to people who are blind (i.e., used *correction*) (Salisbury et al., 2018), and more (Lofi and El Maarry, 2014).

For applications where **speed is critical** (e.g., image search/retrieval), *confidence threshold* is the best choice. This approach also offers a further advantage in that it reduces the number of images needing manual annotation, and so is advantageous for applications with **constrained crowdsourcing budgets or access to large crowds**.

For applications where **quality is critical**, our overall recommendation is to use *correction*. We recommend restricting its use to simple tasks, such as image classification and captioning, rather than more complex tasks such as object detection, which requires both object categories and bounding boxes. This is particularly valuable when we scale the size of the image collection. In real-world scenarios, the size of image databases used are significantly larger than the size used in these experiments. For instance, the popular ImageNet database contains 14 million images, iStock offers 30 million photos, and Flickr hosts more than 6 billion images. If we consider even a fraction of this size, annotating a dataset of 1 million images using the *correction* workflow would save approximately 1,667 hours of human effort while producing similar results to *manual annotation*. This is equivalent to saving 41 40-hour work weeks in time, and consequently $13,000 in cost (considering a wage of $8/hour). This highlights the immense potential of employing hybrid workflows in terms of achieving high quality results while making great savings on time and money.

Our findings highlight that employing algorithmic predictions with crowdsourcing can be beneficial. This is especially exciting given that deep learning models have recently become more accessible to a larger base of users through cloud-based services (e.g., Google Cloud Platform, Microsoft Azure, Amazon Web Services) and user-friendly software platforms (e.g., Tensorflow, Keras). Until the past few years, access to deep learning models was limited because domain expertise was necessary to configure complex deep learning software environments (e.g., Caffe) and it was difficult to access necessary computational resources (e.g., GPUs or CPU clusters). A clear, necessary stepping stone for progress in hybrid workflows is the availability of well-performing algorithms for target tasks, which is not yet available for many tasks (e.g., video analysis, question answering, robot navigation).

This work is motivated by our desire to inspire innovation in maximizing the strengths of hybrid systems in solving tasks much more efficiently as opposed to the traditional way of relying solely on humans or machines. We offer our framework and implementation as a starting point to simplify the process of efficiently designing and comparing hybrid workflows for a specific task. To support these purposes, we will publicly release all code and material. We hope users will leverage and extend this framework to establish best practices and design choices for their own data annotations tasks and datasets.

A valuable direction for future work is to consider other ways to combine algorithmic and crowd-

sourcing efforts to overcome the limitations of the workflows we benchmarked. For instance, observing from our qualitative analysis of the results for the image classification task that a major reason for errors from the fastest task (i.e., *verification*) is insufficient initial results from algorithms due to images being unclassified or incompletely classified, a modified version could instead offer a ranked list of proposed initial results from multiple algorithmic models from which humans can select (Jain and Grauman, 2016). This may enable simultaneously achieving the fast speed of *verification* with the high quality that is possible with *correction*.

While our work focuses on hybrid workflows at run-time, another valuable direction for future work is to expand upon our analysis to also consider hybrid workflows at design-time. This includes consideration of active learners (Gilyazev and Turdakov, 2018), which are algorithms that minimize human involvement by directing their annotation efforts only to those examples that will lead to the greatest boost in model performance during model training. A key distinction between analyzing run-time and design-time methods is which evaluation metrics matter the most. For instance, annotation time is a critical factor that matters at run-time, while it is less so at design-time where the sole focus centers on capturing high quality annotations. A promising design-time workflow is to iteratively retrain a model using corrections provided by workers for images that an algorithm has low confidence in its predictions. Already, the promise of this hybrid workflow has been shown for annotating the LSUN dataset (Yu et al., 2015), which contains 69 million annotated images.

Another interesting future work would be to integrate interpretability (i.e. explanation) of the deep learning models. Understanding the reasons why a model produces a certain prediction can further enable us to enhance the hybrid workflow by leveraging the model's strengths while compensating its flaws with the crowd workers' support to improve the overall accuracy.

7. CONCLUSION

We introduce a framework that allows users to employ different hybrid methods which partner crowds and algorithms to collaboratively annotate images. Our evaluation of how three popular hybrid algorithm-crowdsourcing workflows compare to manual annotation and automated annotation for three tasks (image classification, object detection, image captioning) reveal the general (dis)advantages of each approach. All code and data will be publicly shared at `http://anonymous.com` to facilitate extensions to this work.

8. REFERENCES

Bernstein, M. S, Teevan, J, Dumais, S, Liebling, D, and Horvitz, E. (2012). Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 237–246.

Brady, E, Morris, M. R, Zhong, Y, White, S, and Bigham, J. P. (2013). Visual Challenges in the Everyday Lives of Blind People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2117–2126.

Chen, X, Fang, H, Lin, T.-Y, Vedantam, R, Gupta, S, Dollár, P, and Zitnick, C. L. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325* (2015).

Cheng, J and Bernstein, M. S. (2015). Flock: Hybrid Crowd-Machine Learning Classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 600–611.

Chilton, L. B, Little, G, Edge, D, Weld, D. S, and Landay, J. A. (2013). Cascade: Crowdsourcing Taxonomy Creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1999–2008.

Clarke, V and Braun, V. (2014). Thematic analysis. In *Encyclopedia of critical psychology*. Springer, 1947–1952.

Cohen, I and Medioni, G. (1999). Detecting and tracking moving objects for video surveillance. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, Vol. 2. IEEE, 319–325.

Dang, B, Hutson, M, and Lease, M. (2016). MmmTurkey: A crowdsourcing framework for deploying tasks and recording worker behavior on Amazon Mechanical Turk. *arXiv preprint arXiv:1609.00945* (2016).

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

Efron, B and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Everingham, M, Van Gool, L, Williams, C. K, Winn, J, and Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (2010), 303–338. DOI:http://dx.doi.org/10.1007/s11263-009-0275-4

Gaur, Y, Lasecki, W. S, Metze, F, and Bigham, J. P. (2016). The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th Web for All Conference*. ACM, 23.

Gilyazev, R and Turdakov, D. Y. (2018). Active learning and crowdsourcing: A survey of optimization methods for data labeling. *Programming and Computer Software* 44, 6 (2018), 476–491.

Guinness, D, Cutrell, E, and Morris, M. R. (2018). Caption Crawler: Enabling Reusable Alternative Text Descriptions using Reverse Image Search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 518.

Guo, A. (2018). Crowd-AI Systems for Non-Visual Information Access in the Real World. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, DC09.

Gurari, D, Jain, S, Betke, M, and Grauman, K. (2016). Pull the Plug? Predicting If Computers or Humans Should Segment Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 382–391.

Gurari, D, Li, Q, Stangl, A. J, Guo, A, Lin, C, Grauman, K, Luo, J, and Bigham, J. P. (2018). VizWiz grand challenge: Answering visual questions from blind people. *arXiv preprint arXiv:1802.08218* (2018).

Gurari, D, Sameki, M, Wu, Z, and Betke, M. (2016). Mixing Crowd and Algorithm Efforts to Segment Objects in Biomedical Images. In *Medical Image Computing and Computer Assisted Intervention Interactive Medical Image Computation Workshop (2016)*. 1–8.

Hara, K, Le, V, and Froehlich, J. (2013). Combining crowdsourcing and google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 631–640.

Hara, K, Sun, J, Moore, R, Jacobs, D, and Froehlich, J. (2014). Tohme: Detecting Curb Ramps in Google Street View Using Crowdsourcing, Computer Vision, and Machine Learning. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. ACM, 189–204.

Harrington, R. P and Vanderheiden, G. C. (2013). Crowd caption correction (CCC). In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 45.

Huang, Y, Huang, Y, Xue, N, and Bigham, J. P. (2017). Leveraging complementary contributions of different workers for efficient crowdsourcing of video captions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 4617–4626.

Jain, S and Grauman, K. (2016). Click carving: Segmenting objects in video with point clicks. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4.

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*. ACM, 5839–5849.

Kittur, A, Nickerson, J. V, Bernstein, M, Gerber, E, Shaw, A, Zimmerman, J, Lease, M, and Horton, J. (2013). The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1301–1318.

Konyushkova, K, Uijlings, J, Lampert, C. H, and Ferrari, V. (2017). Learning Intelligent Dialogs for Bounding Box Annotation. *arXiv preprint arXiv:1712.08087* (2017).

Krasin, I, Duerig, T, Alldrin, N, Ferrari, V, Abu-El-Haija, S, Kuznetsova, A, Rom, H, Uijlings, J, Popov, S, 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://github.com/openimages* (2017).

Laput, G, Lasecki, W. S, Wiese, J, Xiao, R, Bigham, J. P, and Harrison, C. (2015). Zensors: Adaptive, Rapidly Deployable, Human-Intelligent Sensor Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1935–1944.

Lasecki, W. S, Homan, C, and Bigham, J. P. (2014). Architecting Real-Time Crowd-Powered Systems. *Human Computation* 1, 1 (2014).

Lin, C. H, Mausam, M, and Weld, D. S. (2012). Dynamically Switching between Synergistic Workflows for Crowdsourcing. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Lin, T.-Y, Maire, M, Belongie, S, Hays, J, Perona, P, Ramanan, D, Dollár, P, and Zitnick, C. L. (2014). Microsoft COCO: Common

Objects in Context. In *European Conference on Computer Vision*. Springer, 740–755.

Lofi, C and El Maarry, K. (2014). Design Patterns for Hybrid Algorithmic-Crowdsourcing Workflows.. In *CBI (1)*. 1–8.

Lundgard, A, Yang, Y, Foster, M. L, and Lasecki, W. S. (2018). Bolt: Instantaneous crowdsourcing via just-in-time training. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 467.

MacLeod, H, Bennett, C. L, Morris, M. R, and Cutrell, E. (2017). Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5988–5999.

Pan, J.-Y, Yang, H.-J, Faloutsos, C, and Duygulu, P. (2004). Automatic multimedia cross-modal correlation discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 653–658.

Papadopoulos, D. P, Uijlings, J. R, Keller, F, and Ferrari, V. (2016). We don't need no bounding-boxes: Training object class detectors using only human verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 854–863.

Papineni, K, Roukos, S, Ward, T, and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

Pirsiavash, H and Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2847–2854.

Quinn, A. J and Bederson, B. B. (2011). Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1403–1412.

Russakovsky, O, Deng, J, Su, H, Krause, J, Satheesh, S, Ma, S, Huang, Z, Karpathy, A, Khosla, A, and Bernstein, M. (2015). Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

Rzeszotarski, J and Kittur, A. (2012). CrowdScape: interactively visualizing user behavior and output. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 55–62.

Sabou, M, Scharl, A, and Föls, M. (2013). Crowdsourced Knowledge Acquisition: Towards Hybrid-Genre Workflows. *International Journal on Semantic Web and Information Systems (IJSWIS)* 9, 3 (2013), 14–41.

Salisbury, E, Kamar, E, and Morris, M. R. (2017). Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind. *Proceedings of HCOMP 2017* (2017).

Salisbury, E, Kamar, E, and Morris, M. R. (2018). Evaluating and Complementing Vision-to-Language Technology for People who are Blind with Conversational Crowdsourcing.. In *IJCAI*. 5349–5353.

Sodemann, A. A, Ross, M. P, and Borghetti, B. J. (2012). A Review of Anomaly Detection in Automated Surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 1257–1272.

Song, J. Y, Lemmer, S. J, Liu, M. X, Yan, S, Kim, J, Corso, J. J, and Lasecki, W. S. (2019). Popup: reconstructing 3D video using particle filtering to aggregate crowd responses. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 558–569.

Von Ahn, L and Dabbish, L. (2004). Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 319–326.

Von Ahn, L, Ginosar, S, Kedia, M, Liu, R, and Blum, M. (2006)a. Improving Accessibility of the Web with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 79–82.

Von Ahn, L, Liu, R, and Blum, M. (2006)b. Peekaboom: A Game for Locating Objects in Images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 55–64.

Weld, D. S and Dai, P. (2011). Human Intelligence Needs Artificial Intelligence. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Wigness, M, Draper, B. A, and Ross Beveridge, J. (2015). Efficient Label Collection for Unlabeled Image Datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4594–4602.

Yu, F, Seff, A, Zhang, Y, Song, S, Funkhouser, T, and Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365* (2015).

Zhang, H, Horvitz, E, and Parkes, D. C. (2013). Automated Workflow Synthesis.. In *AAAI*.

Zhou, F and Lin, Y. (2016). Fine-grained image classification by exploring bipartite-graph labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1124–1133.