DualLabel: Secondary Labels for Challenging Image Annotation

Author 1*
Organization or School

Author 1*
Organization or School

Author 1*
Organization or School

Abstract

It is demanding for non-expert annotators to select an appropriate label for an image when the annotation task is difficult. In such cases, it might be easier for an annotator to select multiple "likely" labels than to be forced to select a single label. Multiple labels might be more informative in the training of a classifier because multiple labels can contain a correct one even when a single label is wrong. We present *DualLabel*, an annotation tool that allows annotators to assign secondary labels for an image to make the annotation process easier and also improve the classification accuracy of a trained model. We ran a user study to compare our proposed dual-label and a traditional single-label approach for an image annotation task. The results show that our dual-label approach not only reduces the task completion time but also improves the accuracy of a classifier trained with the given labels.

Keywords: Annotation Tool, Challenging Image Annotation, Non-Expert Annotator, Secondary Label, Machine Learning.

Index Terms: • Human-centered computing~Human computer interaction (HCI)~Interactive systems and tools

1 Introduction

Annotations are usually expected to be accurate. Classifiers are trained and evaluated based on annotations assuming that they represent the ground truth. However, in practice, annotations are given by humans, but humans are not perfect. This is especially true when the budget is limited, and one cannot recruit a sufficient number of domain experts. One needs to use a limited number of annotators without sufficient domain knowledge (usually via crowdsourcing) [2, 10, 37] and train a classifier based on inaccurate training data. Multi-class labeling is particularly challenging for non-expert annotators where an annotator is required to select a single label among multiple "likely" options for a target image. Figure 1 shows an example of the difficulty faced by a non-expert annotator in a multi-class image classification task (i.e., a non-expert worker may not know which label is correct, but he/she may know either label A or label B). In such cases, it is desirable to have a method to reduce the burden of annotations (make it easier for non-expert annotators to do the task) while maintaining (or even improving) the accuracy of classifiers trained with the resulting annotations without increasing annotation cost.

We present a novel annotation tool, *DualLabel*, as a solution to the above problem. The basic idea is simple. When it is difficult for an annotator to select an appropriate label for a target image, we allow the annotator to select a second choice. We expect that this would reduce the psychological burden to make a final decision, and thus expedite the annotation process. Furthermore, this process produces more annotations for a set of images compared to a traditional single-label method, so we expect that this method would improve the accuracy of a classifier trained with the same amount of images.

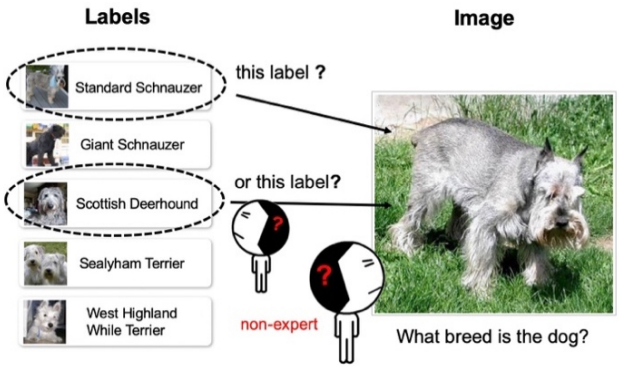


Figure 1: Problem of multi-class labeling for non-expert annotators.

We ran a user study to compare the proposed dual-label approach and a traditional single-label approach for a multi-class image annotation task. We measured the time required for annotation and collected subjective feedback. We also trained classifiers using the annotation results and measured the accuracy of the classifiers. The results showed that the dual-label approach requires less time to complete the image annotation task than the single-label approach in a challenging image annotation task. In addition, the accuracy of a classifier trained with dual labels was better than that of a classifier trained with single labels. Our questionnaire results showed that the participants felt that the dual-label approach is easier, more helpful, and more efficient than the single-label approach when they were not confident with their label selection during annotation. Additionally, the dual-label approach increased annotators' confidence in challenging image annotation. This study makes the following contributions:

- *DualLabel*, a novel annotation tool that allows users to assign a secondary label in a difficult multi-class image classification task.
- A user study comparing *DualLabel* to a traditional single-label approach, demonstrated the benefits of the dual-label approach in a challenging image annotation task.
- A machine learning experiment training classifiers by using the two sets of annotations, showed that the dual label approach could improve the accuracy of classifiers.

* email address

2 Related Work

2.1 Crowdsourcing Data Annotation

Crowdsourcing is a popular method for conducting data annotation tasks. However, data quality in crowdsourcing often contains numerous errors [3, 4, 5]. In order to increase data quality, crowdsourcing data annotation tasks are often designed relatively simple and easy. For example, various types of binary tasks (i.e., users select one option from two) are used. In a binary task, crowd workers are requested to answer a simple yes-or-no question such as "does this image contain a dog?" ImageNet [1] is a large-scale image dataset that was manually verified by humans via a binary judgment. Pairwise HITS [19] is an annotation workflow that allows annotators to compare a pair of labeled data and select a better one. A binary labeling task is easy when the data is clear, or the task is simple (e.g., select a dog or a cat label). However, it becomes difficult for annotators when the data is difficult or the task is complex (e.g., select a dog breed label for a dog image). In such cases, the data quality might be unstable and low and this brings problems into the machine learning process [22].

In order to avoid errors during annotation, an "unsure" function is conventionally provided during annotation that allows annotators to give up (skip) a labeling task when it is difficult [6, 7]. For example, Audio Set [17] and Revolt [18] provide an "unsure" feature in their annotation tasks. This "unsure" option can avoid errors made by the annotators when they are not sure about the data. In addition, "user confidence" is another information that can be collected during annotation for measuring label quality. For example, annotators indicate their confidence scores to a label they select to an image. Oyama et al. [21] proposed a labeling approach that allows annotators to give their confidence scores in a binary labeling task. Jinhua et al. [21] proposed a method that shares the same concept in a data annotation task. In the task, annotators were asked to judge their answers by giving their confidence via a sliding bar.

These approaches mainly focus on a single-label annotation task. However, it may be insufficient for a challenging data annotation task. For example, it can be very difficult for annotators to select an appropriate one label from two possible labels when they are not confident with the labels. We believe that our dual-label approach can open new opportunities to address this issue, which is different from other existing approaches.

2.2 Utilizing Different Data in Machine Learning

An unsure function has been wildly used in various data annotation tasks. Several studies have proved that unsure data could help machine learning results. Takeoka et al. [8] indicated that the performance of unsure loss is better than conventional models (no unsure labels) because unsure labels tend to be located close to the ground truth decision boundary. Zhong et al. [6] indicated that providing annotators unsure option would benefit Active Learning from Crowds (ALC) significantly. Cui et al. [51] proposed an uncertainty pairwise comparison oracle to assist interactive labeling by that comparing the uncertainty of two unlabeled data points. In addition, Wu et al. [9] proposed a unified end-to-end learning framework by using "unsure data" in medical image analysis. This framework shows the benefits of learning with unsure data, and the validity of our models is demonstrated on the prediction of Alzheimer's Disease and lung nodules.

On the other hand, Ishida et al. [11] introduced a learning approach learning from complementary labels, for optimizing algorithms. A complementary label specifies a class that a pattern does not belong to. It can reduce human workload in manual data annotation because selecting an incorrect label is easier than selecting a correct label when data is difficult. Yu et al. [12] shared the same concept and proposed a framework: learning with biased complementary labels. Moreover, partial label learning is an approach for training a set of possible labeled data where each instance is tagged with more than one label, only one of which is correct [13, 23, 24, 52, 53]. These studies focus on partial multi-label for machine learning, while our study focuses on the annotation process. Cui and Sato [50] introduced a method for learning from triplet comparison data by using only passively obtained triplet comparison data. In addition, Whitehil et al. [14] induced GLAD, a learning model for optimizing the integration of labels from annotators of unknown expertise. This model refers to the concept "whose vote should count more" that can recover the true data labels more accurately than the Majority Vote heuristic.

Various systems/algorithms have been developed to use different types of collected data for improving machine learning results. Developing a better system/algorithm to improve machine learning accuracy is a popular (common) approach that most studies work on it. However, there is also another approach which is to provide better data. In this study, we focus on the latter one and aim to expedite the data annotation process from a human aspect.

2.3 Expediting Data Annotation Process

Expediting data annotation process can be discussed in two aspects: (a) data quality and labeling efficiency (i.e., objective data analysis) and (b) annotator's perception during annotation (i.e., subjective data analysis). Many studies focused on the former aspect and proposed efficient and supportive approaches to expedite the data annotation process (i.e., decrease labeling time and increase label quality). For example, a hierarchical task assignment was proposed to expedite manual image annotation [28]. Moreover, many semi-automatic annotation systems have been proposed to assist manual image labeling using collaborative filtering and computer vision techniques [29, 30, 31]; and interactive concept learning guides annotators to assign labels to images that are most informative for classifiers [32, 33, 34].

On the other hand, some studies focused on annotators' perception. For example, Von Ahn et al. [35] proposed an image annotation tool combined with a computer game. Chang et al. [36] designed a spatial layout labeling interface that can increase non-expert annotators' confidence in an image annotation task. In addition, some studies have indicated that the human aspect is an important part of a crowdsourcing annotation task. For example, Zhuang and Gadiraju [38] analyzed crowd workers' mood, performance, and engagement during annotation, LaPlante et al. [39] investigated the trust issue between crowd workers and task requesters, and Durward et al. [40] identified ethical issues in crowdsourcing, especially with a focus on the crowd workers' perspective. In addition, Chang et al. [56] investigated the effect of quick and careful labeling styles in an image annotation task.

These studies have shown that the human aspect plays an important role in manual data annotation. However, there is no detailed study exploring the relationship between the data quality/labeling efficiency (i.e., objective data analysis) and the annotator's perceived usability (i.e., subjective data analysis). We believe this can be a significant point in expediting the data annotation process, which we investigate in this study.

3 Dual Label

We propose *DualLabel*, a dual-label image annotation tool that allows annotators to assign both a high-confident and a low-confident (secondary) label for an image when they are feeling difficult to make a label decision during annotation (i.e., not confident with the target image and labels). Figure 2 shows a screenshot of the dual-label annotation interface taking dog breed

labeling as an example. The lower part of the left-side interface shows the labels to assign with corresponding representative images, and the upper part presents a magnified view of an activated label. The user activates a label by clicking one of the labels below. The upper part of the right-side interface presents a target image to be labeled. The lower part of the right-side interface is the selected label(s) for the target image, which includes a high-confident label box (left) and a low-confident label box (right). The annotator drags a label on the left bottom to the label box on the right bottom to assign a label. High-confident label is required, but a low-confident label is optional (annotators are allowed to select only one label if they are confident with it). In addition, annotators are allowed to swap the high-confident label and the low-confident label by clicking the arrow icon between the two boxes.

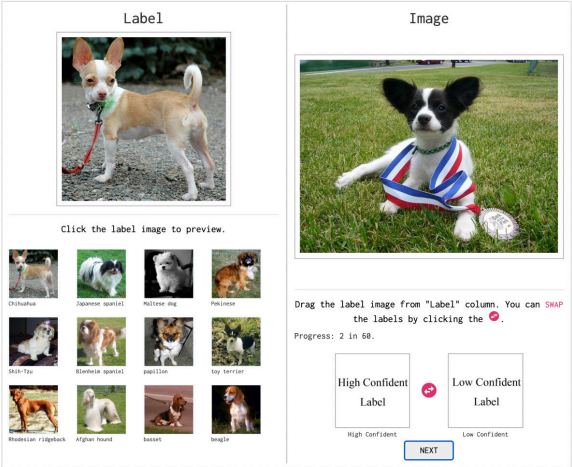


Figure 2: *DualLabel* user interface.

Figure 3 shows an example of a usage scenario in a challenging image annotation task via the dual-label image annotation tool. In this scenario, the annotator is not confident with the target image and labels (i.e., feels difficult to make a label decision). In such cases, the annotator is not sure which dog breed (label) is the most appropriate label for the target dog image, but they think that the answer is probably either "*Chih-Tzu*" or "*Maltese dog*". So, the annotator selects "*Chih-Tzu*" as a primary label (high-confident label) and "*Maltese dog*" as a secondary label (low-confident label) for the image.

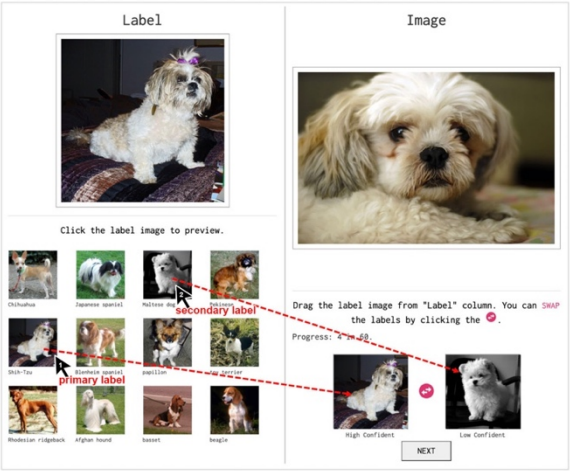


Figure 3: An example of a usage scenario.

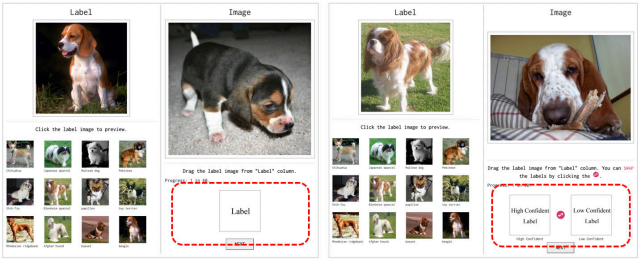
The main purpose of this dual-label annotation tool is to help annotators to make a label decision more easily and quickly when they face difficulties making an appropriate label decision during annotation. We assume that, in most cases, the selected two labels may contain one "correct" label and one "likely" label. We expect that this "likely" label is a useful resource and can contribute to the improvement of the classification accuracy of a trained model.

4 User Study

A user study was conducted to compare a traditional single-label annotation approach and our proposed dual-label annotation approach for a manual image annotation task. Our hypothesis is that the dual-label approach can expedite the annotation process in challenging image classification, especially when the annotators are not confident with label assignment.

4.1 Image Annotation System

Two online image annotation systems (single-label and dual-label) were developed for the user study. We used Html and JavaScript to develop the user interface and Node.js on the server-side to collect and process the data. The system was deployed on an Ubuntu system with 2GB of RAM provided as a server from Sakura VPS. The image annotation interfaces as shown in Figure 4. The only difference between the two interfaces is that the single-label interface has only one label box (participants select only one label for an image), and the dual-label interface has two label boxes (participants can select a low-confident label in addition to a high-confident label for an image).



(a) single-label interface (b) dual-label interface

Figure 4: Screenshots of single-label and dual-label image annotation interfaces.

4.2 Participants

24 participants (16 men and 8 women, in the age range of 18 to 59 years) were recruited by using Amazon Mechanical Turk (MTurk) [25]. All of the participants are MTurk Master Workers and have 98% HIT Approval Rate [26]. 18 of the 24 participants (75%) have (or had) dogs as a pet. However, none of the participants have professional knowledge of dog breeds. We paid a bit higher compensation ($15.75/h) to the participants than the average wage in Amazon MTurk ($11.58/h) [27].

4.3 Dataset

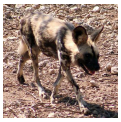
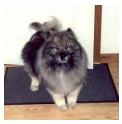
We used a dog image dataset from ImageNet (ILSVRC 2012), which contains 120 dog labels (breeds) [15]. For the image annotation tasks, we manually created two datasets (Dataset A and Dataset B) with different difficulty levels. The difficulty levels of selected dog labels (breeds) are based on the results of the Stanford Dogs images classification [16]. The Stanford Dogs Dataset also uses the 120 dog labels from ImageNet. Their machine learning experiment follows a similar training/testing methodology as Caltech-101 [49]. Dataset A includes the most difficult 12 dog breeds (among the 120 dog breeds) and Dataset B includes the easiest 12 dog breeds. A difficult dog breed means

that the classification accuracy was low for the dog breed [16]. Tables 1 and 2 show the 12 dog breeds in Dataset A (difficult) and Dataset B (easy). For each dog breed (label), 5 images were randomly selected from 1,300 images, from ImageNet (ILSVRC 2012) [15] for the image annotation task in the user study.

Table 1. Dataset A: 12 difficult dog breeds.

Chihuahua	Japanese Spaniel	Maltese Dog	Pekinese
Shih-Tzu	Blenheim Spaniel	Papillon	Toy Terrier
Rhodesian Ridgeback	Afghan Hound	Basset	Beagle

Table 2. Dataset B: 12 easy dog breeds.

 African Hunting Dog	Dhole, Cuon Alpinus	Dingo, Warrigal	Mexican Hairless
Standard Poodle	Miniature Poodle	Toy Poodle	Cardigan
Pembroke	Brabancon Griffon	 Keeshond	Chow

4.4 Task and Condition

The image annotation tasks for each participant involved labeling 120 dog images (60 = 12 x 5 images for each dataset). In the annotation task, the participants were requested to select an appropriate dog breed (label) for each dog image, from a 12-dog breeds list (12 labels). The method of within-subjects was used, where 24 participants were asked to complete two annotation tasks via the single-label and dual-label interfaces (different datasets were used for the tasks conducted by different interfaces).

Single-label Condition. This is the baseline. Figure 4 (a) shows a screenshot of the single-label annotation interface. An instruction of the single-label annotation task is as below:

"If you are confident with the target image and the label you intend to select, assign it to the image directly. However, if you are not confident with the target image and the label, please consider carefully and try to assign the most appropriate one."

Dual-label Condition. This is the proposed approach. Using the dual-label annotation interface, see Figure 4 (b). An instruction of the dual-label annotation task is as below:

"If you are confident with the target image and the label you intend to select, assign it to the image as the 'high-confident label'. However, if you are not confident with the target image and the label, assign the label you are more confident with to the high-confident label box, and the label you are less confident with to the low-confident label box."

In order to ensure the balance of the two conditions and the two datasets, we balanced the order of the conditions and dataset among the 24 participants. Table 3 shows the distribution of the conditions and datasets.

Table 3. Distribution of the training data.

Participants	1st Task	2nd Task
P01 - P06	Single-Label + Dataset A	Dual-Label + Dataset B
P07 - P12	Single-Label + Dataset B	Dual-Label + Dataset A
P13 - P18	Dual-Label + Dataset A	Single-Label + Dataset B
P19 - P24	Dual-Label + Dataset B	Single-Label + Dataset A

4.5 Procedure

A 3-minutes video instruction was presented at the beginning to the participants, to explain the details of the user evaluation process. It includes a demonstration (step by step) of how to use the annotation interfaces to complete the given image annotation tasks. After the two annotation tasks, the participants were asked to fill a questionnaire regarding the annotation tasks they completed.

4.6 Measurement

Our annotation system recorded and measured the time and accuracy rate of the image annotation tasks completed by the participants. The system also recorded the time spent by the participants for each image labeling. Furthermore, the system recorded the number of the secondary labels collected via the dual label approach, in the cases (images) when the participants were not confident with a single label selection. In addition, a questionnaire contained three Likert scale questions and one open-ended question regarding the participants' perception and preference of the single-label and dual-label annotation approaches.

5 Machine Learning Experiment

In addition to the evaluation of the manual image annotation process, we conducted a small-scale machine learning experiment to compare the accuracy of classifiers trained with the labeled data created using the single-label and the dual-label approaches. Our hypothesis is that the labeled data collected using the dual-label approach achieves better classification accuracy than the data collected via the traditional single-label approach.

Training Data. We used the data collected in the two annotation approaches in the training of classifiers. A total of 2,880 labeled images were collected from the user study. Table 4 shows the distribution of the training data. 720 labeled images (60 labeled images for each dog breed) were collected in each condition. The number of images is the same (720 images), but the dual-label task contains more labels for training. The training data collected contain errors made by the participants (i.e., the accuracy rate of the training data is not 100%).

Table 4. Distribution of the training data.

Tasks	Dataset	Amount
Single-label	Dataset A (difficult)	60x12=720
Single-label	Dataset B (easy)	60x12=720
Dual-label	Dataset A (difficult)	60x12=720
Dual-label	Dataset B (easy)	60x12=720

Testing Data. The test data used in the machine learning experiment were total 2,880 images (60x12=720 images in each condition). These images were not included in the training data.

Classification Algorithm. We used unpre-trained AlexNet [46] implemented by PyTorch [47]. We trained four classifiers, each of which were trained using 720 images as shown in table 4. The accuracy of each classifier was measured with 720 test images (60 images from each class). We reshaped the input images to a size of 224×224 and trained the model with a learning rate of 0.01 and a weight decay of 0.0005 as parameters. We used SGD (Stochastic Gradient Descent) optimizer and passed the output of the model to a sigmoid layer, then obtained Binary Cross Entropy Loss. The epoch of model training is 200. For the data collected via the dual-label approach, we encoded all the categories into 0-1 vectors while assigning a weight of 1 to the high-confidence label and 0.5 to the low-confidence label. The sum of the vector is normalized to 1. The output is consistent with the single-label approach; both are single result output. Then we compared the accuracy of the labeled data obtained from the two annotation approaches on the test data.

6 Results

6.1 Task Completion Time

Overall Task Completion Time. Figure 5 shows that the participants spent an average of 10 min 6 s and 10 min 37 s to label the 60 images using the single-label and dual-label approaches, respectively. The result of paired t-test on task completion time shows that the difference was not significant ($p > 0.05$) between the single-label and dual-label approaches. This indicates that the dual-label task does not require more time to complete the image annotation task.

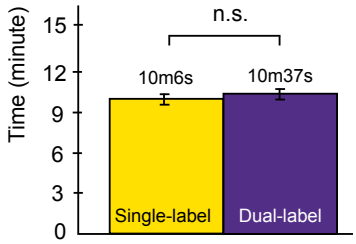


Figure 5: Overall task completion time. Single: mean = 10.06; SD = 3.76; Dual: mean = 10.37; SD = 3.90.

Task Completion Time in Easy and Difficult Datasets. Figure 6 presents the task completion time in the easy and difficult datasets. In the easy dataset, the result shows that the participants spent an average of 7 min 49 s and 11 min 39 s to label the 60 images via the single-label and dual-label approaches respectively. Although the time looks different between the two approaches, the result of paired t-test shows that the difference is not significant ($p > 0.05$). It is because the standard deviation (SD) in the dual-label task was high. In the difficult dataset, the participants spent an average of 12 min 25 s and 9 min 36 s via the single-label and dual-label approaches respectively. The result of paired t-test shows that the difference is significant ($p < 0.05$) between the two approaches. This indicates that the dual-label approach is more efficient (i.e., requires less time to complete the task) than the single-label approach when the target images are difficult.

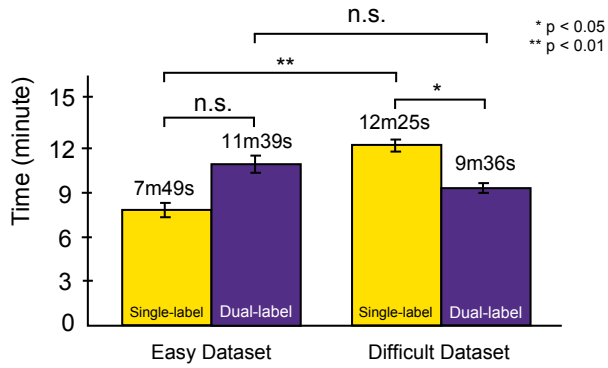


Figure 6: Task completion time in the easy and difficult datasets. Easy Dataset. Single: mean = 7.49; SD = 3.24; Dual: mean = 11.39; SD = 4.54. Difficult Dataset. Single: mean = 12.25; SD = 2.70; Dual: mean = 9.36; SD = 2.76.

In addition, the results show that, in the single-label task, the participants spent significantly ($p < 0.01$) longer time to complete the difficult task (12m25s) and the easy task (7m49s). However, the difference was not significant ($p > 0.05$) between the easy (11m39s) and difficult (9m36s) tasks in the dual-label task. This indicates that the dual-label approach is robust against task difficulty in terms of task completion time.

6.2 Annotation Accuracy

Figure 7 displays the accuracy of annotations given by the participants using the single-label and dual-label approaches. The results show that the annotation accuracy was 75% for the single-label task, while the annotation accuracy was 71.12% for the high-confident label and 12% for the low-confident label in the dual-label task. The result of paired t-test shows that the difference was not significant ($p > 0.05$) between the labels in the single-label task and the high-confident labels in the dual-label task. This indicates that the dual-label approach cannot significantly improve label quality, but it can collect more data (i.e., low-confident label) during annotation. Additionally, the result shows that the accuracy of the low-confident label in the dual-label task was only 12%. This indicates that the low-confident labels themselves are not very reliable.

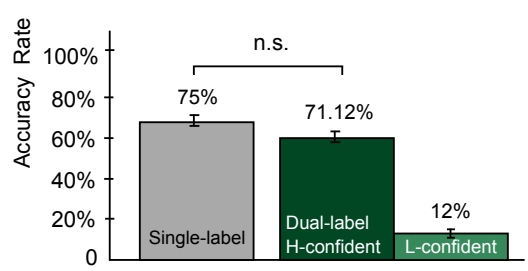


Figure 7: Accuracy rate. Single: mean = 75.00; SD = 9.65; Dual-label H-confident: mean = 71.12; SD = 8.98 and L-confident: mean = 12.00; SD = 6.26.

Figure 8 displays the annotation accuracy of the annotation tasks in the easy and difficult datasets. In the easy dataset, the results show that the accuracy rate was 78.42% for the single-label task, while the accuracy rate was 70.50% for the high-

confident label and 13.39% for the low-confident label in the dual-label task. In the difficult dataset, the results show that the accuracy rate was 71.58% for the single-label task, while the accuracy rate was 71.75% for the high-confident label and 10.08% for the low-confident label in the dual-label task. The result of paired t-test shows that the difference was not significant ($p > 0.05$) between the annotation accuracy in the single-label task and the dual-label task (high-confident label) for both easy and difficult datasets.

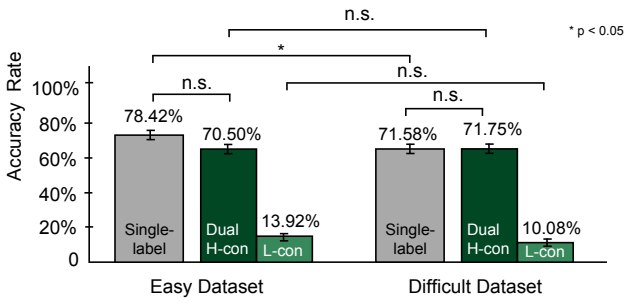


Figure 8: Accuracy rates in the easy and difficult datasets. Easy Dataset. Single: mean = 78.42; SD = 6.86; Dual_H-con: mean = 70.50; SD = 8.61 and L-con: mean = 13.92; SD = 7.61. Difficult Dataset. Single: mean = 71.58; SD = 11.07; Dual_H-con: mean = 71.75; SD = 9.69 and L-con: mean = 10.08; SD = 4.01.

Additionally, the results show that the difference was significant ($p < 0.05$) between the two datasets in the single-label task. However, the difference was not significant ($p > 0.05$) between the two datasets in the dual-label task. This indicates that the data difficulty only affects the annotation accuracy in the single-label task. It does not affect the label quality in the dual-label task.

6.3 Number of Assigned Labels

Table 5 shows the number of assigned labels collected via the single-label and dual-label annotation tasks in the user study. In the single-label annotation task, the total number of labels assigned to images was 1440 labels (720 for Dataset A and 720 for Dataset B). In the dual-label annotation task, a total number of the labels assigned to images was 2397 (66.45% more than the single-label task), which includes 1440 high-confidence labels (720 for Dataset A and 720 for Dataset B), and 957 low-confidence labels (481 for Dataset A and 476 for Dataset B).

Table 5. Distribution of the training data.

Tasks	Datasets	Number of Assigned Labels
Single-label	Dataset A	720
	Dataset B	720
Dual-label	Dataset A	720 (h-con) and 481 (l-con)
	Dataset B	720 (h-con) and 476 (l-con)

(h-con = high-confident label; l-con = low-confident label)

The results show that the dual-label annotation task can collect 66.8% (Dataset A) and 66.11% (Dataset B) more labeled images (i.e., low-confident labeled images) than the single-label annotation task. Originally, we expected that the dual-label approach would collect more data (i.e., low-confident labeled images) in the difficult dataset than the easy dataset because the annotators might feel difficult to make a single label decision when the images are difficult. However, the results showed that there was almost no difference between the easy (481 images) and difficult (476 images) datasets. This indicates that data difficulty does not affect the amount of collected data in the dual-label annotation task.

6.4 Learning Effect

Figure 9 shows the average time for the annotation process for the first half (1-30 images) and the second half (31-60 images) via the single-label and dual-label approaches, in the easy dataset. The results showed that the participants spent an average of 4 min 11 s and 3 min 38 s to complete the first half and second half using the single-label approach, while they spent an average of 5 min 21 s and 6 min 18 s via the dual-label approach. The results of paired t-test showed that there was no significant difference ($p > 0.05$) between the first half and second half of the annotation process in both single-label and dual-label annotation tasks. This indicates that there is no learning effect when the images are easy.

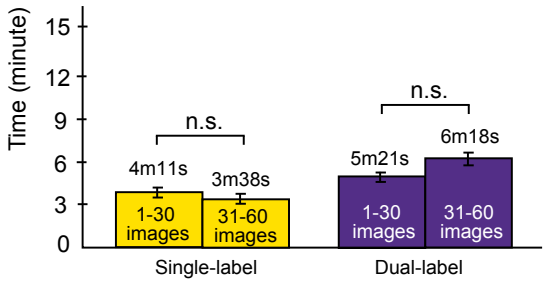


Figure 9: Average time for the annotation process in the easy dataset. Single (1-30 images): mean = 4.11; SD = 1.95; Single (31-60 images): mean = 3.38; SD = 2.12. Dual (1-30 images): mean = 5.21; SD = 3.43; Dual (31-60 images): mean = 6.18; SD = 4.13.

Figure 10 shows the average time for the annotation process for the first half (1-30 images) and the second half (31-60 images) via the single-label and dual-label approaches, in the difficult dataset. The results showed that the participants spent an average of 6 min 3 s and 6 min 22 s to complete the first half and second half using the single-label approach, while they spent an average of 5 min 39 s and 3 min 57 s via the dual-label approach. The results of paired t-test showed that there was no significant difference ($p > 0.05$) between the first half and second half of the annotation process in the single-label annotation task. However, it showed that there was a significant difference ($p < 0.05$) in the dual-label annotation task. This indicates that the time per image decreased more clearly in the second half of the annotation process. This implies that annotators are able to learn and improve the efficiency of the annotation process using the dual-label approach when the images are difficult.

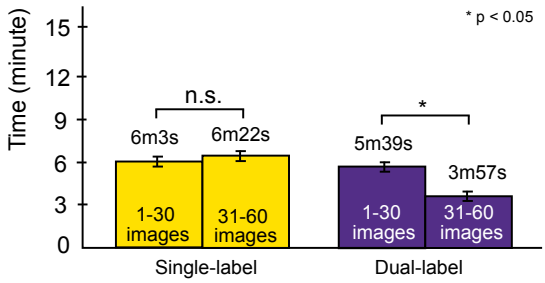


Figure 10: Average time for the annotation process in the difficult dataset. Single (1-30 images): mean = 6.03; SD = 2.52; Single (31-60 images): mean = 6.22; SD = 2.37. Dual (1-30 images): mean = 5.39; SD = 1.01; Dual (31-60 images): mean = 3.57; SD = 1.39.

6.5 Questionnaire

Figure 11 shows questionnaire results. Fig 11 (a) shows that the dual-label approach can reduce annotators' subjective feeling of difficulty when the images are difficult in the annotation task. Fig 11 (b) shows that the dual-label approach was perceived to be more helpful than the single-label approach when the images are difficult in the annotation task. Fig 11 (c) shows that the dual-label approach was perceived to be more efficient than the single-label approach when the images are difficult in the annotation task.

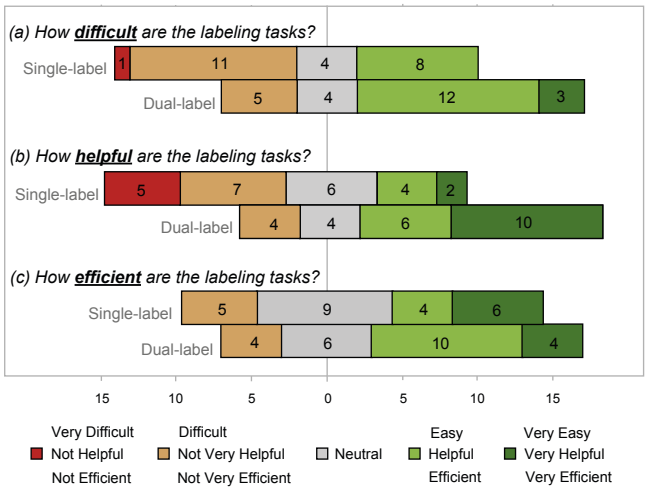


Figure 11: Questionnaire results.

The results also showed that 66.67% of the participants (n = 16) preferred the dual-label approach, while 33.33% preferred the single-label approach (n = 8). The participants who preferred the dual-label approach indicated that the dual-label annotation interface was supportive and helpful. For example, one participant indicated that "*Even if I got my answer wrong on the most confident, between the two of them I felt I was probably right on one of them*"; one participant indicated that "*having a more confident/less confident choice gave more leeway into picking the right breed, especially the ones that were hard to tell/could have been more than one*"; and another participant indicated that "*I felt like using two labels was insurance against being wrong so I felt less anxiety about my choices. I just felt more comfortable working on the task when I could choose two labels instead of one*." In contrast, the participants who preferred the single-label approach indicated that the single-label annotation interface was easy and simple. For example, one participant indicated that "*I feel like having two options made me second-guess my instincts a little bit*"; one participant indicated that "*most of the dog images were clear for me, I only need the single-label interface*", and another participant indicated that "*it's an easy mode for the annotation task*".

6.6 Results of Machine Learning Experiment

Figure 12 shows the accuracy of the classifiers trained with the labeled data collected in each annotation condition (each bar corresponds to a classifier trained with 720 images). The results show that the dual-label approach can increase the accuracy of classifiers by 6.95% (from 78.19% to 84.03%) and 9.86% (from 71.53% to 81.94%) in the easy and difficult datasets, respectively. The result of paired t-test (for each dog breed, we compare the accuracy of single-label and dual-label as a pair) on classification accuracy shows that the difference was significant ($p < 0.01$) between the single-label and dual-label approaches in the difficult dataset, but the difference was not significant ($p > 0.05$) between the two approaches in the easy dataset. This indicates that the labeled data collected via the dual-label approach can improve the accuracy of classifiers in the difficult dataset (i.e., challenging image annotation).

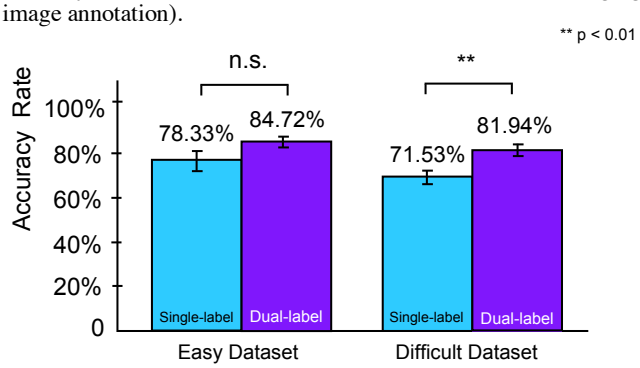


Figure 12: Accuracy rates of machine learning performance. Easy Dataset. Single: mean = 78.33; SD = 19.38; Dual: mean = 84.72; SD = 9.27. Difficult Dataset. Single: mean = 71.53; SD = 17.47; Dual: mean = 81.94; SD = 14.78.

7 Discussion

7.1 Dual-label Approach Expedites Annotation Process in Challenging Image Classification

Manual image annotation is a tedious and time-consuming process that often relies on crowd workers. It can be very demanding for non-expert crowd workers (who have insufficient domain knowledge) to select a correct label when the task is difficult [3, 4, 5]. The dual-label approach allows annotators to select two labels for an image when it is difficult to make a single-label decision. Our original concern was that selection of two labels for an image might take more time than just selecting one label because annotators need to repeat the same operation twice (i.e., selecting a label for an image). However, the results show that the dual-label approach performed even better than our expectation in the user study. It shows that the participants spent significantly ($p < 0.05$) less time to complete the image annotation task in the dual-label approach than the single-label approach when the image annotation task is difficult. This result implies that the dual-label approach expedites the manual image annotation process (i.e., reduces the task completion time) in challenging image classification. We believe that the reduction of image annotation time brings significant benefits to manual data annotation because it means that we can reduce the cost of conducting a manual data annotation task (normally, labor-intensive and costly).

7.2 Dual-label Approach Shows Learning Effect During Annotation in Challenging Image Annotation

A learning effect describes how people learn during a process. It is a phenomenon that we can use to analyze the efficiency of an activity or a study. The learning effect has been studied and discussed in various areas [44, 45, 48]. It is also used as a method in data annotation to analyze task performance (i.e., efficiency) during a labeling process [28]. In the user study, we analyzed the learning effect during the given image annotation tasks. The results showed that there was a significant learning effect ($p < 0.05$) during the annotation process in the dual-label annotation task when the images are difficult. The dual-label approach allows the annotators to "learn" during annotation and increase the annotation efficiency (i.e., reduce the task completion time in the

second half of the image annotation task) in difficult image classification. The reason for this effect is not clear. One possible reason is that it helps annotators to build domain knowledge during the process. Another possible reason is that it helps annotators to learn how to use the user interface. We conjecture that the second one may not be the reason because we did not see a similar learning effect in the easy image labeling task. These results have shown that the learning effect is an important factor during annotation and it has brought potential insights for future development of annotation tools such as designing a learnable annotation tool (e.g., learning domain knowledge during annotation).

7.3 Dual-label Approach Collects More Data and Improves Classifier Accuracy

More data gives better classification accuracy in general. Usually, to collect more training data, more human effort and money are necessary in a manual data annotation task. Our results showed that the dual-label approach could collect more data (labeled images) without significantly increasing the time (even requiring less time when the images are difficult). More specifically, the dual-label approach collected 481 (66.8%) more labeled images (i.e., low-confident label) than the single-label approach when the annotation task is difficult; and 476 (66.1%) more labeled images when the annotation task is easy. These additional data collected via the dual-label approach is the "low-confident" labeled images. The results of our machine learning experiment show that these 66.8% and 66.1% additional data improved classification accuracy by 9.86% (difficult dataset) and 6.95% (easy dataset). This improvement was affected by the low-confident labeled images collected via the dual-label approach. This result implies that the low-confident data plays a significant role in the improvement of classification accuracy.

The dual-label approach expedites the image annotation process and improves classification accuracy. We believe that this brings a significant benefit to difficult image annotation and valuable insight for conducting a crowdsourcing annotation task. It is important to note that the dual-label approach collects more labels without needing more annotation time; and increases the number of annotations (label assignments) without increasing the number of images to be annotated. This is much more efficient than increasing the number of annotations by increasing the number of images to be annotated. An annotator spends a significant amount of time observing the target image to assign the first label, and dual label approach "reuses" the effort in assigning the second label.

7.4 Dual-label Approach Increases Perceived Usability and Worker's Confidence in Challenging Image Classification

The analysis of subjective feedback from the questionnaires shows that the dual-label approach can increase crowd workers' perceived usability in challenging image annotation. More than half the participants felt easier (n = 15), more helpful (n = 16) and more efficient (n = 14) when using the dual-label approach than the single-label approach. A participant commented "*The dual-label interface is more preferable since it allows one to enter their top two choices, but doesn't mandate them to do so. This is both efficient as it doesn't require easy judgments to be more complicated than they are yet simultaneously allowing one to put their top two selections without spending an inordinate amount of time internally debating about which answer they should select*." This shows the benefits of the dual-label approach from a human aspect (e.g., annotators' perception). This is an important result as several studies [38, 39, 40] have indicated that the human aspect is a critical issue that should be considered when conducting crowdsourcing tasks. In addition, the dual-label approach can increase crowd workers' confidence during annotation. For example, a participant indicated, "*The dual label interface leads to much higher confidence when selecting an answer, especially for breeds that are very similar in appearance*." Another participant indicated, "*I feel bad when I make mistakes. With the dual-label interface, I feel less likely to give a wrong answer. Also, I spent less time choosing a label because I felt more confident when I could pick two labels*." We believe that confidence is a factor that affects the efficiency of manual data annotation in challenging image annotation, as confidence is an important factor that improves people's motivation for learning and its efficiency [41, 42, 43].

8 Limitation and Future Work

limitation of this study is the training data used in the machine learning experiment was small (60 images for each dog breed) and only tested in a limited setting. However, the main purpose of this study is not to pursue high accuracy rates of machine learning performance, and instead to focus on the effects of the dual-label approach during annotation in challenging image classification. Our results have shown that the dual-label approach can collect more data, and the data can increase machine learning accuracy. We believe that the dual-label approach might have an even greater effect in a large-scale data annotation task. Additionally, this study only evaluated the dual-label approach with a type of data and task (dog image classification task). We believe that the dual-label approach also can be used and bring benefits in various data types and annotation tasks.

In the future, we would like to run a large-scale user study and test more machine learning conditions (e.g., soft labeling [54] and Self-knowledge distillation [55]). There are many possible variations for using the collected data vis the dual-label approach. For example, more than two (three or four) labels to an image or forcing two labels to an image. It might be possible to apply dual-label to binary classification (no label to an image might be different from two labels to an image). In addition, we would like to investigate the real reason for the learning effect (the improvement of annotation efficiency). One possible direction is to observe annotators' behaviors during the whole annotation process. For example, how do workers spend their time during an annotation task (i.e., what do they exactly do?). Another possible direction is to explore self-learning during annotation and design a learnable annotation interface to help non-expert annotators to build domain knowledge during annotation.

9 Conclusion

In this paper, we presented *DualLabel*, a new manual image annotation tool for expediting the manual data annotation process as well as improving the classification accuracy of the trained model in challenging image annotation. This tool allows annotators to assign a high-confident and a low-confident label for an image when the annotators feel difficult to make a label decision. Our user study compared our proposed dual-label approach to a traditional single-label annotation approach for a manual multi-class image annotation task. The results showed that our dual-label approach could collect more data with less task completion time when the images are difficult. In addition, the dual-label approach showed a positive learning effect during the annotation process. Our machine learning results demonstrated that the labeled images using the dual-label approach improve the accuracy of a classifier trained with the annotation results. Additionally, we discussed perceived usability and workers'

confidence in challenging image annotation. The findings of this study have presented significant insights that can be used in the future development of crowdsourcing annotation tools in challenging image annotation.

ACKNOWLEDGMENTS

This work was supported by ******** (removed for the review).

REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. DOI: http://dx.doi.org/10.1109/CVPR.2009.5206848

[2] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 139-147. https://dl.acm.org/doi/10.5555/1866696.1866717

[3] Jiyin He, Jacco van Ossenbruggen, and Arjen P. de Vries. 2013. Do You Need Experts in the Crowd? A Case Study in Image Annotation for Marine Biology. In Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, pp. 57-60. https://dl.acm.org/doi/10.5555/2491748.2491763

[4] Donghui Feng, Sveva Besana, and Remi Zajac. 2009. Acquiring High Quality Non-Expert Knowledge from On-demand Workforce. In Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web), pp. 51-56. https://dl.acm.org/doi/10.5555/1699765.1699773

[5] Wei Wang, and Zhi-Hua Zhou. 2015. Crowdsourcing label quality: a theoretical analysis. Science China Information Sciences 58, no. 11: 1-12. DOI: http://doi.org/10.1007/s11432-015-5391-x

[6] Jinhong Zhong, Ke Tang, and Zhi-Hua Zhou. 2015. Active Learning from Crowds with Unsure Option. In IJCAI, pp. 1061-1068. http://dl.acm.org/doi/10.5555/2832249.2832396

[7] Yao-Xiang Ding, and Zhi-Hua Zhou. 2018. Crowdsourcing with unsure option. Machine Learning 107, no. 4: 749-766. DOI: http://doi.org/10.1007/s10994-017-5677-x

[8] Kunihiro Takeoka, Yuyang Dong, and Masafumi Oyamada. 2020. Learning with unsure responses. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 01, pp. 230-237. DOI: http://doi.org/10.1609/aaai.v34i01.5355

[9] Botong Wu, Xinwei Sun, Lingjing Hu, and Yizhou Wang. 2019. Learning with unsure data for medical image diagnosis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10590-10599. DOI: http://doi.org/10.1109/ICCV.2019.01069

[10] Jing Zhang, Xindong Wu, and Victor S. Sheng. 2016. Learning from crowdsourced labeled data: a survey. Artificial Intelligence Review 46, no. 4: 543-576. DOI: http://doi.org/10.1007/s10462-016-9491-9

[11] Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. 2017. Learning from complementary labels. arXiv preprint arXiv:1705.07541. https://arxiv.org/abs/1705.07541

[12] Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. 2018. Learning with biased complementary labels. In Proceedings of the European conference on computer vision (ECCV), pp. 68-83. https://openaccess.thecvf.com/content_ECCV_2018/papers/Xiyu_Yu_Learning_with_Biased_ECCV_2018_paper.pdf

[13] Timothee Cour, Ben Sapp, and Ben Taskar. 2011. Learning from partial labels. The Journal of Machine Learning Research 12: 1501-1536. https://dl.acm.org/doi/10.5555/1953048.2021049

[14] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Advances in neural information processing systems 22: 2035-2043. DOI: http://dl.acm.org/doi/10.5555/2984093.2984321

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. 2015. ImageNet Large Scale Visual Recognition Challenge. International journal of computer vision 115, no. 3: 211-252. DOI: http://dx.doi.org/10.1007/s11263-015-0816-y

[16] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC), vol. 2, no. 1. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.255.6394&rep=rep1&type=pdf

[17] Jort F. Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776-780. IEEE. DOI: http://dx.doi.org/10.1109/ICASSP.2017.7952261

[18] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 2334-2346. DOI: http://dx.doi.org/10.1145/3025453.3026044

[19] Takeru Sunahase, Yukino Baba, and Hisashi Kashima. 2017. Pairwise HITS: Quality Estimation from Pairwise Comparisons in Creator-Evaluator Crowdsourcing Process. In Thirty-First AAAI Conference on Artificial Intelligence. https://dl.acm.org/doi/abs/10.5555/3298239.3298383

[20] Satoshi Oyama, Yukino Baba, Yuko Sakurai, and Hisashi Kashima. 2013. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In Twenty-Third International Joint Conference on Artificial Intelligence. https://dl.acm.org/doi/10.5555/2540128.2540496

[21] Jinhua Song, Hao Wang, Yang Gao, and Bo An. 2018. Active learning with confidence-based answers for crowdsourcing labeling tasks. Knowledge-Based Systems 159: 244-258. DOI: http://doi.org/10.1016/j.knosys.2018.07.010

[22] Alkida Balliu, Sebastian Brandt, Yuval Efron, Juho Hirvonen, Yannic Maus, Dennis Olivetti, and Jukka Suomela. 2019. Classification of distributed binary labeling problems. arXiv preprint arXiv:1911.13294. https://arxiv.org/abs/1911.13294

[23] Nam Nguyen, and Rich Caruana. 2008. Classification with partial labels. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 551-559. DOI: https://doi.org/10.1145/1401890.1401958

[24] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. 2019. Learning a deep convnet for multi-label classification with partial labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 647-657. https://arxiv.org/abs/1902.09720

[25] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science 6, 1, 3–5. DOI: http://dx.doi.org/10.1177/1745691610393980

[26] Peer E, Vosgerau J, Acquisti A. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. Behavior Res Methods. December 1;46(4):1023–31.

[27] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 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, pp. 1-14. DOI: https://doi.org/10.1145/3173574.3174023

[28] Chia-Ming Chang, Siddharth Deepak Mishra, and Takeo Igarashi.2019. AHierarchical Task Assignment for Manual Image Labeling. In 2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pp. 139-143. DOI: http://dx.doi.org/10.1109/VLHCC.2019.8818828

[29] Shingo Uchihashi, and Takeo Kanade. 2005.Content-Free Image Retrieval by Combinations of Keywords and User Feedbacks. In International Conference on Image and Video Retrieval, pp. 650-659. Springer, Berlin, Heidelberg. DOI: http://dx.doi.org/10.1007/11526346_68

[30] Wenyin Liu, Susan T. Dumais, Yanfeng Sun, HongJiang Zhang, Mary Czerwinski, and Brent A. Field. 2001. Semi-Automatic Image Annotation. In Interact, vol. 1, pp. 326-333. https://www.researchgate.net/profle/Liu_Wenyin/publication/2328523_ Semi-Automatic_Image_Annotation/links/5650863608aeafc2aab71e41/Semi- Automatic- Image- Annotation.pdf

[31] Suh Bongwon and Benjamin B. Bederson. 2004. Semi-Automatic Image Annota- tion Using Event and Torso Identifcation. Human Computer Interaction Labora- tory, University of Maryland, College Park, Maryland, USA. http://www.cs.umd. edu/hcil/trs/2004-15/2004- 15.pdf

[32] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: Interactive Concept Learning in Image Search. In Proceedings of the sigchi conference on human factors in computing systems, pp. 29-38. DOI: http://dx. doi.org/10.1145/1357054.1357061

[33] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2009. Overview-Based Example Selection in End-User Interactive Concept Learning. In Proceedings of the 22nd annual ACM symposium on User interface software and technology, pp. 247-256. DOI: http://dx.doi.org/10.1145/1622176.1622222

[34] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2010. Exam- ining Multiple Potential Models in End-User Interactive Concept Learning. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1357-1360. DOI: http://dx.doi.org/10.1145/1753326.1753531

[35] Luis Von Ahn, and Laura Dabbish. 2004. Labeling Images with a Computer Game. In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 319-326. DOI: http://dx.doi.org/10.1145/985692.985733

[36] Chia-Ming Chang, Chia-Hsien Lee, and Takeo Igarashi. 2021. Spatial Labeling: Leveraging Spatial Layout for Improving Label Quality in Non-Expert Image Annotation. In CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3411764.3445165

[37] Antonio Ghezzi, Donata Gabelloni, Antonella Martini, and Angelo Natalicchio. 2018. Crowdsourcing: a review and suggestions for future research. International Journal of Management Reviews 20, no. 2: 343-363.

[38] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In Proceedings of the 10th ACM Conference on Web Science, pp. 373-382. https://doi.org/10.1145/3292522.3326010

[39] Rochelle LaPlante, M. Six Silberman, and Industriegewerkschaft Metall. 2016. Building trust in crowd worker forums: Worker ownership, governance, and work outcomes. Proceedings of WebSci16. ACM: 35-63.

[40] David Durward, Ivo Blohm, and Jan Marco Leimeister. 2016. Is there papa in crowd work?: a literature review on ethical dimensions in crowdsourcing. In 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp. 823-832. IEEE.

[41] Mei-Mei Chang. 2005. Applying Self-Regulated Learning Strategies in a Web-Based Instruction—An Investigation of Motivation Perception. Computer Assisted Language Learning 18, no. 3: 217-230. DOI: http://dx.doi.org/10.1080/09588220500178939

[42] Robert M Klassen. 2010. Confidence to Manage Learning: The Self-Efficacy for Self-Regulated Learning of Early Adolescents with Learning Disabilities. Learning Disability Quarterly 33, no. 1: 19-30. DOI: http://dx.doi.org/10.1177/073194871003300102

[43] Juan Carlos Ortiz-Ordoñez, Friederike Stoller, and Bernd Remmele. 2015. Promoting Self-Confidence, Motivation and Sustainable Learning Skills in Basic Education. Procedia-Social and Behavioral Sciences 171: 982-986. DOI: http://dx.doi.org/10.1016/j.sbspro.2015.01.205

[44] Gur Mosheiov. 2001. Parallel machine scheduling with a learning effect. Journal of the Operational Research Society 52, no. 10: 1165-1169. DOI: https://doi.org/10.1057/palgrave.jors.2601215

[45] Koun-tem Sun, Yuan-cheng Lin, and Chia-jui Yu. 2008. A study on learning effect among different learning styles in a Web-based lab of science for elementary school students. Computers & Education 50, no. 4: 1411-1422. DOI: https://doi.org/10.1016/j.compedu.2007.01.003

[46] [46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12). Curran Associates Inc., Red Hook, NY, USA, 1097–1105. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen et al. 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32: 8026-8037. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[48] Yvonne Kammerer, Rowan Nairn, Peter Pirolli, and Ed H. Chi. 2009. Signpost from the masses: learning effects in an exploratory social tag search browser. In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 625-634. DOI: https://doi.org/10.1145/1518701.1518797

[49] Fei-Fei Li, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In 2004 conference on computer vision and pattern recognition workshop, pp. 178-178. IEEE. DOI: https://doi.org/10.1109/CVPR.2004.383

[50] Zhenghang Cui, Nontawat Charoenphakdee, Issei Sato, and Masashi Sugiyama. 2020. Classification from triplet comparison data. Neural computation 32, no. 3: 659-681. DOI: https://doi.org/10.1162/neco_a_01262

[51] Zhenghang Cui, and Issei Sato. 2020. Active classification with uncertainty comparison queries. arXiv preprint arXiv:2008.00645. https://arxiv.org/abs/2008.00645

[52] Jun-Peng Fang,, and Min-Ling Zhang. 2019. Partial multi-label learning via credible label elicitation. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 3518-3525. DOI: https://doi.org/10.1609/aaai.v33i01.33013518

[53] Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. 2019. Partial multi-label learning by low-rank and sparse decomposition. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 5016-5023. DOI: https://doi.org/10.1609/aaai.v33i01.33015016

[54] Jianlin Cheng, Amanda Jones, Caroline Privault, and Jean-Michel Renders. 2013. Soft labeling for multi-pass document review. In ICAIL 2013 DESI V Workshop. http://legacydirs.umiacs.umd.edu/~oard/desi5/research/Cheng-final.pdf

[55] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. 2020. Self-knowledge distillation: A simple way for better generalization. arXiv preprint arXiv:2006.12000. https://arxiv.org/abs/2006.12000

[56] Chia-Ming Chang, Xi Yang, and Takeo Igarashi. 2022. An Empirical Study on the Effect of Quick and Careful Labeling Styles in Image Annotation. The 48th International Conference on Graphics Interface and Human-Computer Interaction (Gl 2022), Virtual Conference, 17-19 May 2022 https://openreview.net/pdf?id=SDyj8aZBPrs