Supplementary Material for Beyond Active Learning: Leveraging the Full Potential of Human Interaction via Auto-Labeling, Human Correction, and Human Verification

Table of Contents

A. Reproducibility

While we mention the experimental setting in Section 5, we discuss a couple other aspects of reproducibility in this section. We provide an experiment script[2] that executes a CLARIFIER configuration given a set number of command-line arguments:

- **al_strategy**: One of *badge, entropy*
- **auto_assign_strategy**: *highest_confidence*
- **b1**: AL selection budget
- **b2**: SMI selection budget
- **b3**: Auto-assign budget
- **dataset**: One of *cifar10, cifar100, birds, dogs*
- **device**: CUDA device ID
- **human_correct_strategy**: *logdetmi*
- **num_partitions_human**: Number of unlabeled set partitions (see Section 5)
- **rounds**: Number of selection rounds
- **runs**: Number of repeated experiment runs to execute
- **seed_size**: Size of the initial labeled seed set
- **thread_count**: Number of threads in SMI selection

The script utilizes the DISTIL toolkit presented in [4] to produce result JSON files that contain information about each selection round. We utilize a separate Jupyter notebook to analyze and plot our results, which we provide with the experiment script. Lastly, we detail the licenses of each repository and dataset used in this work:

[2] https://github.com/nab170130/auto_label_mp

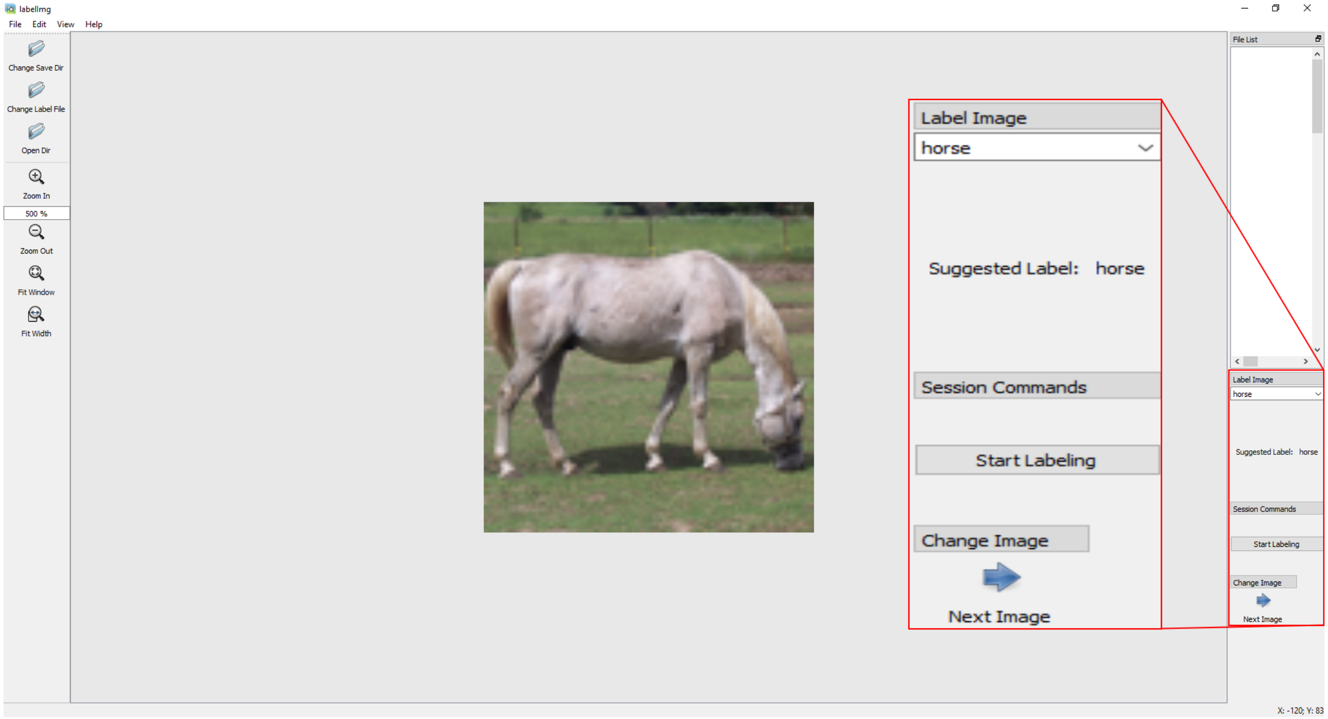

Figure 7. The tool interface used in our labeling experiment. The subject presses the "Start Labeling" button to initiate the timing experiment. The tool provides a suggested label that is correct 50% of the time. The subject fixes the suggested label by either selecting the correct label from the drop-down menu or by typing the label in the drop-down menu's field. The subject clicks the "Next Image" button once he or she has verified the correctness of the suggested label or has fixed an incorrect suggested label.

- Caltech-UCSD Birds-200-2011 [19]: Non-commercial Research
- CIFAR-10(0) [12]: MIT License
- DISTIL [4]: MIT License
- labelImg [18]: MIT License
- PyTorch [15]: Modified BSD
- Stanford Dogs [8]: Non-commercial Research
- STL-10 [5]: Non-commercial Research
- SVHN [14]: CC0 1.0 Public Domain
- UC Merced Land Use [20]: CC0: Public Domain

B. Additional Labeling Experiment Details

Here, we present more details on the labeling tool used in our labeling experiment.

B.1. Labeling Tool

To compute estimates of the cost to fix incorrectly suggested labels (c_a) and the cost to verify correctly suggested labels (c_v) for each of the datasets discussed in Section 3, we utilize a labeling tool based on [18] that records the time spent to assign a final label for a series of images with potentially incorrect suggestions. A snapshot of the tool is given in Figure 7. Once a dataset has been chosen, the tool sequentially presents a random sample of images in that dataset. Timing information starts to be collected when the user presses the "Start Labeling" button. From that point, the subject assigns the final label for the

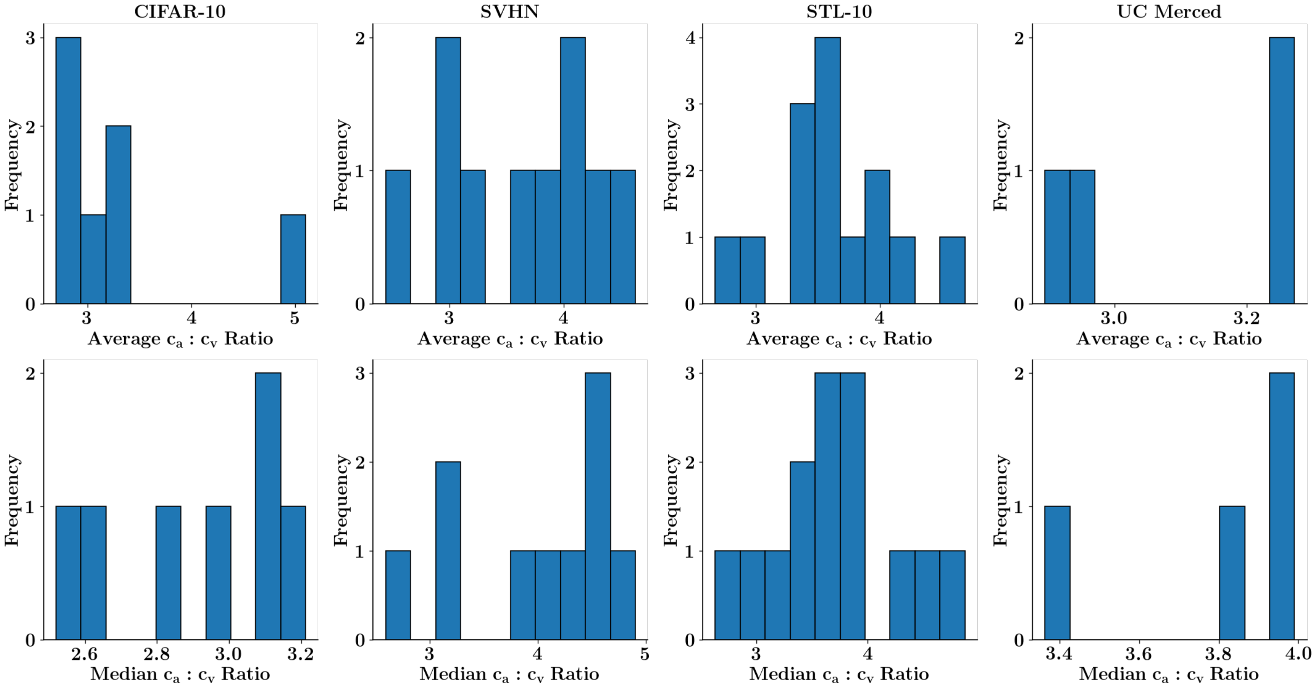

Figure 8. Histograms of average / median ratios across each dataset. As shown, the $c_a : c_v$ ratio for most subjects falls between 3 and 4.

image by pressing the "Next Image" button, which immediately presents the next image. The subject is provided a suggested label for each image to assist in the labeling effort. If the suggestion is wrong (which occurs 50% of the time as mentioned in Section 3), then the subject can correct the suggestion by selecting the correct label from the drop-down menu or by typing the label in the drop-down menu's field. Once the subject has finished labeling the image sample for a dataset, the tool saves the timing information for each image along with the final label given by the subject.

When computing the average and median values for c_a and c_v for each subject's performance on a dataset, only the images whose final label matches the ground truth label of the dataset are considered. These average and median values are also saved and are used to produce average and median ratios between c_a and c_v for a subject's performance on a dataset. In Section 3, the average of the average ratios and the average of the median ratios are presented in Figure 2. We present the distribution of average and median $c_a : c_v$ ratios for each dataset in Figure 8.

C. SMI-Based Selection

In this section, we highlight the optimization background and the methodology used in selecting the intermediate-hardness instances mentioned in Figure 1 and in Section 4.

C.1. Submodular Mutual Information

To target the intermediate-hardness instances, we apply the recently proposed submodular mutual information [7], which we introduce and motivate here. A set function $f : 2^\Omega \to \mathbb{R}$ is submodular if $f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)$ for any $A \subseteq \Omega, B \subseteq \Omega, x \in \Omega$ satisfying $A \subseteq B, x \notin B$. Furthermore, f is monotone if $f(A) \leq f(B)$ for any $A \subseteq B \subseteq \Omega$. If f is both monotone and submodular, then the function can be maximized under a cardinality constraint on its domain with a $(1 - \frac{1}{e})$ approximation guarantee using a simple greedy algorithm [13]. In many instances, real-valued set functions are used to assign scores to subsets that correspond to ratings of set coverage, set diversity, and other metrics of interest; furthermore, many of these set functions are monotone submodular, giving a very sound method of selecting cardinality-constrained subsets from large amounts of data that appeal to some desired measure.

Submodular functions have also been used to model information-theoretic quantities. [7] define the submodular mutual information between sets $A, Q \subseteq \Omega$ under f as $I_f(A; Q) = f(A) + f(Q) - f(A \cup Q)$. In addition to being monotone, [7] show that $I_f(A; Q)$ is also submodular in A or Q when fixing the other argument under further conditions on f. Hence, $I_f(A; Q)$ can be fixed in Q and maximized by the aforementioned greedy algorithm [13] for sufficient f, giving cardinality-constrained subsets that are similar to Q under f that also appeal to the measure modeled by f. As such, Q can be treated as a query set that represents a target for the selected subset. This strategy has been used in prior work to perform targeted AL; namely, [9] use various submodular mutual information instantiations as described in [11] to mine rare-class, non-redundant, and in-distribution instances for image classification. [10] use these functions for mining rare objects and slices in object detection problems.

C.2. Choice of SMI Function

As detailed in [9] and [11], many different instantiations of submodular mutual information can be chosen. In this work, we opt to instantiate I_f as the Log-Determinant Mutual Information (LOGDETMI) function used in [9] and [11]:

$$I_f(A;Q) = \log\det \mathcal{S}_A - \log\det\left(\mathcal{S}_A - \mathcal{S}_{A,Q} S_Q^{-1} S_{A,Q}^T\right) \tag{2}$$

where $\mathcal{S}_A$ is a matrix of similarity scores between the instances in A and $\mathcal{S}_{A,Q}$ is a matrix of similarity scores between the instances in A to the instances in Q. To compute each $(S_{A,Q})_{ij}$, we take the cosine similarity between their vector representations v_i, v_j. We formulate v_i, v_j in the same manner done by [3] and [9], where the loss gradient of the last-layer parameters is calculated and used as the vector representation. In both works, the loss for unlabeled instances is computed using the hypothesized label. For labeled instances, the loss is computed using the ground-truth label.

C.3. Semi-Hard Selection

Algorithm 2 SMI Selection

```
**Input:** Labeled set $\mathcal{L}$, Unlabeled set $\mathcal{U}$, Model $M$, Model parameters $\theta$, Budget $b$, Class set $\mathcal{C}$
$\hat{G} \leftarrow \{\nabla_{\theta_{last}} L_\theta(x, \hat{y}) | x \in \mathcal{U}, \hat{y} = \text{argmax}_i M_\theta(x)\}$
$k \leftarrow \frac{b}{|\mathcal{C}|}$
**for** $c$ **in** $\mathcal{C}$ **do**
    $\mathcal{L}_c \leftarrow \{(x, y) \in \mathcal{L} | y = c\}$
    $G_c \leftarrow \{\nabla_{\theta_{last}} L_\theta(x, y) | (x, y) \in \mathcal{L}_c\}$
    $\mathcal{S}_{\mathcal{L}_c} \leftarrow \text{CosSimKernel}(G_c, G_c)$
    $\mathcal{S}_{\mathcal{U},\mathcal{L}_c} \leftarrow \text{CosSimKernel}(\hat{G}, G_c)$
    $\mathcal{S}_{\mathcal{U}} \leftarrow \text{CosSimKernel}(\hat{G}, \hat{G})$
    $I_f \leftarrow \text{LogDetMI}(\mathcal{S}_{\mathcal{L}_c}, \mathcal{S}_{\mathcal{U},\mathcal{L}_c}, \mathcal{S}_{\mathcal{U}})$
    $A_c \leftarrow \text{argmax}_{X \subseteq \mathcal{U}, |X| \leq k} I_f(X; \mathcal{L}_c)$
**end for**
$A \leftarrow \cup_{c \in \mathcal{C}} A_c$
$A_{suggested} \leftarrow \emptyset$
**for** $x$ **in** $A$ **do**
    $\hat{c} \leftarrow \text{MaxMarginal}(x, \{A_c | c \in \mathcal{C}\})$
    $A_{suggested} \leftarrow A_{suggested} \cup \{(x, \hat{c})\}$
**end for**
**return** $A_{suggested}$
```

We detail our selection of the intermediate unlabeled points in Algorithm 2. As mentioned in Section 4, selecting instances of intermediate hardness is achieved on a per-class basis using simple greedy submodular maximization [13]. For an intermediate hardness selection budget of b, $b/|\mathcal{C}|$ instances are chosen for each class c in $\mathcal{C}$ by maximizing the submodular mutual information (LOGDETMI as mentioned previously) between the unlabeled data $\mathcal{U}$ and labeled class exemplars $\mathcal{L}_c$. Notably, all selected samples A_c for class c can be suggested as having a potential label c due to the query relevance with $\mathcal{L}_c$ as mentioned in Section 4. However, in scenarios with high class confusion (such as fine-grained classification scenarios such as those presented in Section 5), there may be instances in $\mathcal{U}$ that get selected more than once (*e.g.*, are present in multiple A_c) and thus have multiple class suggestions. To remedy this, we assign such an instance with a suggested label that reflects the A_c to which it contributes the largest marginal gain. In doing so, the instance is assigned the suggested label that most heavily matches the corresponding $\mathcal{L}_c$.