Improving Noisy Fine-Grained Datasets using Active Label Cleaning Framework

Avik Pal
University of Amsterdam
avik.pal@student.uva.nl

Abstract

In this work, we address the escalating data labeling challenge in deep learning, focusing on the effectiveness of the Active Label Cleaning (ALC) framework in Fine-grained Visual Categorization (FGVC) tasks. With the rising complexity of models, crowdsourcing becomes crucial, but it often introduces noise. ALC, leveraging Active Learning, proves to be a cost-effective solution for relabeling, specifically in FGVC datasets. The study explores acquisition functions for efficient sample prioritization and evaluates ALC's suitability in cleaning noisy FGVC data. Contributions made in this paper include simulating crowd-generated labels, demonstrating ALC's efficacy in FGVC scenarios, and highlighting its synergy with noise-robust learning methods. Prioritizing samples based on model posteriors and entropy emerges as a promising acquisition strategy.

1. Introduction

With the increasing complexity and scale of deep learning models, the data requirements to effectively train them have analogously increased. In this regard, obtaining data in this digital age is not as much of a concern as taking up the painstaking effort of labeling them. Crowd-sourcing [2, 15, 21] hence becomes a desirable option to obtain labels at large scale, often leading to noisy datasets [7]. This noise could have emerged from certain ambiguities in input/output spaces such as semantics, wrongful automation in the process, or the lack of expertise of the crowd for the particular task. Directly learning a model using a supervised approach on such noisy datasets could harm its generalization capability as the model could memorize errors. This would also adversely impact the validity of models during evaluation. If wrongfully deployed, it could potentially have dangerous consequences for sensitive learning tasks such as in medicine or autonomous driving. The noise scenarios mentioned above are further exacerbated in Fine-grained Visual Categorization (FGVC) task where a fair level of expertise is expected to properly classify images, for example, dog breeds or fine-grained bird categorization [6, 11, 27]. This is because of the presence of many categories and only subtle differences between these categories. Thus, label cleaning through relabelling the data becomes critical to improve the dataset quality and model performance.

Relabelling a large dataset manually through experts would be time-intensive and arduous, in most cases infeasible. Bernhardt et al. [4] introduce the framework of Active Label Cleaning (ALC) for the task of (automated) relabelling of a noisy dataset in a cost-effective manner using simple data-driven approaches. They use Active Learning methods to prioritize samples for relabelling considering the labeling difficulty for the sample (for example, consensus taken from multiple experts to form an opinion for an ambiguous image) and the total budget. However, they limit their experiments to CIFAR10H [20] with 10 generic labels and a noisy version of NIH's ChestX-Ray8 medical image dataset [28] with only binary labels. We hypothesize that non-expert human annotators would likely misjudge labels that have closer semantic connections and thus, FGVC datasets form a good proving ground. In the absence of noisy labels, we attempt to *generate noisy annotations from the crowd* by utilizing information on semantic and taxonomic relations between categories (see Figures 1a and 1b). Hence, this work aims to probe the effectiveness of the ALC framework on FGVC tasks in a more real-world setting.

To achieve our goals, we set forth the following research questions:

RQ1 Does the Active Label Cleaning framework form a cost-effective option for relabelling noisy samples for FGVC datasets?

RQ2 Which acquisition functions are suitable for prioritizing samples so that the relabelling procedure is efficient?

Our contributions are: **(1)** We simulated real-world label counts from the crowd for the samples of the FGVC datasets using semantic relations between the categories. **(2)** We show that the ALC framework can cost-effectively clean the noisy FGVC dataset. **(3)** We show that noise-robust learning methods complement the label-cleaning procedure. **(4)** Finally, we also show that prioritizing samples based on

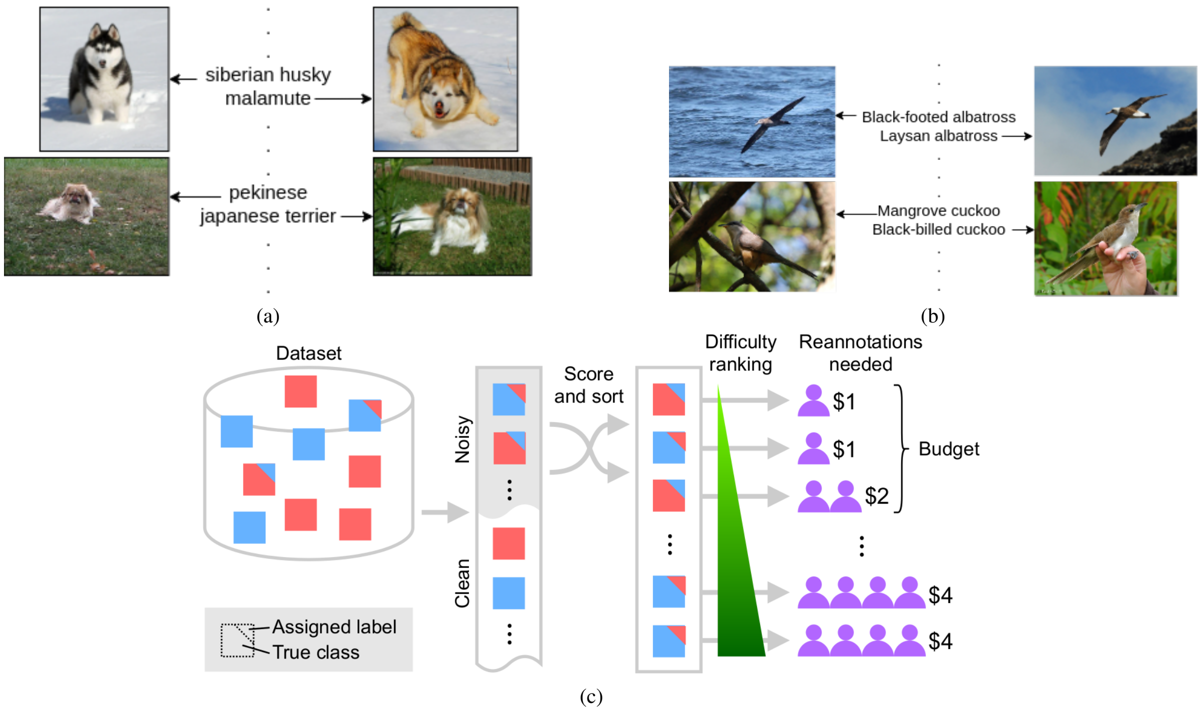


Figure 1. Labels in Fine-grained Visual Categorization (FGVC) datasets usually have subtle differences between semantically or taxonomically related labels. Such minute differences might not be picked up by human annotators during crowd-sourcing which results in noisy labels. **(a)** shows similar-looking images from the Stanford Dogs dataset but are categorized by different dog breeds that share the same parent in the WordNet hierarchy. **(b)** shows images of different categories of birds from CUB-200-2011 dataset that have similar features and are related by the same type (for example, sub-species of albatross bird-type). **(c)** provides an overview of the Active Label Cleaning (ALC) algorithm proposed by Bernhardt et al. [4] which sorts and prioritizes easier noisy samples to relabel to efficiently utilize budget. (Diagram image courtesy of [4])

their model posteriors and corresponding entropy forms a good choice for the acquisition function.

2. Related Work

In this section, we briefly review related literature on FGVC, existing methods tackling learning with noisy labels, and Active Learning (AL).

2.1. Fine-grained Visual Categorization (FGVC)

FGVC involves categorizing images into subgroups within a broader category, such as distinguishing between various bird species (images in CUB-200-2011 dataset [27]) or dog breeds (Stanford Dogs Dataset [11]). This task is labeled as "fine-grained" due to its demand for the model to discern nuanced disparities in visual characteristics and patterns, presenting a greater challenge compared to standard image classification tasks. Many works try to learn the discriminative features in local regions of the image since the global structure for many categories is similar [1, 13, 34]. In reality, obtaining accurate annotations for numerous fine-grained categories is hard and requires domain experts. Hence, it becomes important to explore methods that could utilize such cheaper information (noisy labels) to improve accuracy but is still rarely studied in the literature. Tan et al. [26] use multi-branch attention to learn fine-grained features from different scales of images to achieve robust predictions. Wei et al. [33] further show that existing methods of learning with noisy labels do not achieve satisfying performance for fine-grained datasets and propose stochastic noise-tolerated supervised contrastive learning to extract distinguishable features for the categories. Our work differs in the sense that we first intend to clean the noisy fine-grained dataset and then utilize it for learning tasks.

2.2. Robust methods in Learning with noisy labels

Methods existing for learning with noisy labels could be categorized into robust loss function, sample selection, sample reweight, and label cleaning. Creating a robust loss function has been studied more in earlier works [14, 16, 30, 35] which intend to provide more generalization capability over the simple cross-entropy loss. In sample selection, correctly labeled points are sampled in the learning process using some selection criteria. Small empirical loss criteria for selection have been studied in [8, 31], and recent works such as [3, 19, 32] focus more on the history of predictions that provide more information for selection. Notably, the co-teaching [8] paradigm simultaneously trains two deep neural networks where each network selects samples with clean labels from the mini-batch that are then used to train the other network. Sample reweighting methods [22, 24] is a sub-category for sample selection in which samples are weighted such as with the obtained loss. In label cleaning methods, noisy labels are sampled based on self-prediction from the model's outputs [4, 25, 29]. Active Label Cleaning (ALC) framework proposed by Bernhardt et al. [4] uses Active Learning methods for designing relabelling strategies that consider both resource constraints and individual sample difficulty to simulate limited expert interactions. The ALC framework and co-teaching method for learning with noisy labels are a primary focus of this work and are hence covered more thoroughly in Section 3.

2.3. Active Learning (AL)

AL is a machine learning paradigm that emphasizes the importance of selecting informative data points for model training. Unlike traditional learning, where the algorithm is trained on a fixed dataset, active learning allows the model to choose which examples from the dataset it wants to learn, actively querying the most valuable examples for improvement. This iterative process of selecting and labeling instances helps the model achieve better performance with fewer labeled examples, making it particularly beneficial in scenarios where labeled data is scarce or expensive to obtain. Settles [23] extensively covers this topic in his survey - *"Active Learning Literature Survey"*. He additionally provides an overview of the different active learning settings, the key amongst which is the Pool-Based AL [12] extensively used in this work. The *pool set* takes a different notion in our case, in which all samples are initially present and criteria are set to pick out noisy samples (easier first) for relabelling. The criteria for the model to query examples is usually defined using Acquisition Functions described further in Section 4.2. The topic of *Noisy Oracles* is covered in the survey paper Section 6.2 which highly relates to this paper's goals.

3. Background

In this section, we present some preliminaries for the paper.

3.1. Co-teaching for noise robust learning

This method [8] uses the *memorization* effect of deep networks, where it learns clean labels from easier patterns in the initial epochs and eventually becomes robust enough to filter out noisy instances using their loss values assuming the loss would be less for correctly labeled data (see Algorithm 1). Specifically, two networks f with parameters w_f and g with parameters w_g are trained using mini-batches. Each mini-batch $\bar{D}$ is passed through f (and respectively g), which selects a small proportion $R(T)$ amount of instances with small training loss ℓ forming new mini-batch $\bar{D}_f$ (respectively $\bar{D}_g$). This is used to train the corresponding peer network for parameter updates. The overfitting on noisy labels in later stages of training is regularized through $R(T)$, i.e., $R(T)$ is kept larger at the start to select more instances and is gradually reduced so that only clean instances are selected later on.

Input: w_f and w_g, learning rate η, fixed τ, epoch T_k and T_{max}, iteration N_{max}

for $T = 1, 2, ..., T_{max}$ **do**

  Shuffle: training set D

  for $N = 1, ..., N_{max}$ **do**

    Fetch: mini-batch $\bar{D}$ from D

    Obtain: $\bar{D}_f = \arg\min_{D':|D'|\geq R(T)|\bar{D}|} \ell(f, D')$

    Obtain: $\bar{D}_g = \arg\min_{D':|D'|\geq R(T)|\bar{D}|} \ell(g, D')$

    Update: $w_f = w_f - \eta\nabla\ell(f, \bar{D}_g)$

    Update: $w_g = w_g - \eta\nabla\ell(g, \bar{D}_f)$

  end

  Update: $R(T) = 1 - \min\left\{\frac{T}{T_k}\tau, \tau\right\}$

end

Output: w_f, w_g

Algorithm 1: Co-teaching algorithm as mentioned in [8].

3.2. Active Label Cleaning (ALC)

The ALC framework [4] is a sequential label-cleaning procedure that maximizes the total number of corrected samples given some resource budget $B \in \mathbb{N}$. Suppose, a dataset $D = \{(\mathbf{x}_i, \hat{\mathbf{L}}_i)\}_{i=1}^N$ is given where $\mathbf{x}_i$ is the i^{th} image and $\hat{\mathbf{L}}_i \in \mathbb{N}^C$ is the corresponding label counts vector with C classes. The initial (majority) label $\hat{y}_i = \arg\max_{c\in\{1,...,C\}} \hat{\mathbf{L}}_i$ could be mislabeled in some instances through wrong majority or simulation, and the true class is y. Unlike conventional AL objectives, the framework's primary objective is to obtain a clean set of labels that could further be used for model training and evaluation.

In ALC (see Algorithm 2), a selector model which is a classifier neural network model, $p_\theta(\hat{y}|\mathbf{x})$ parameterized by θ, is initially trained using the noisy dataset $\{\mathbf{x}_i, \hat{y}_i\}$. The ALC takes place over multiple iterations. In each iteration of the cleaning procedure, samples are ranked according to the corresponding ambiguity and predicted label's accuracy using acquisition function Φ (detailed in Section 4.2). The highly ranked sample or a batch of highly ranked samples is selected for relabelling. In a real-world setting, differ-

Given: $Y = \{\mathbf{L}_i\}_{i=1}^N$: True label distributions
Input: $D = \{(\mathbf{x}_i, \hat{\mathbf{L}}_i)\}_{i=1}^N$: Noisy dataset
Input: $B \in \mathbb{N}$: Budget, $r \in \mathbb{N}$: Frequency of weight updates
$\theta \leftarrow TrainRobustModel(D)$
$\mathcal{I}_{avail} \leftarrow \{1, ..., N\}$, $\quad \mathcal{I}_{cleaned} \leftarrow \emptyset$
$count \leftarrow 0$
while $count < B$ **do**
  $j \leftarrow \arg\max_{i \in \mathcal{I}_{avail}} \Phi(\mathbf{x}_i, \hat{\mathbf{L}}_i; \theta)$
  repeat
    $\hat{\mathbf{L}}_j \leftarrow \hat{\mathbf{L}}_j + Sample(\mathbf{L}_j)$
    $count \leftarrow count + 1$
  until *majority formed in* $\hat{\mathbf{L}}_j$;
  $\mathcal{I}_{avail} \leftarrow \mathcal{I}_{avail} \setminus \{j\}$,
  $\mathcal{I}_{cleaned} \leftarrow \mathcal{I}_{cleaned} \cup \{j\}$
  $D \leftarrow \{(\mathbf{x}_i, \hat{\mathbf{L}}_i) : i \in \mathcal{I}_{avail} \cup \mathcal{I}_{cleaned}\}$
  if $count \% r == 0$ **then**
    $\theta \leftarrow Update(\theta, D)$
end
Output: D

Algorithm 2: Active label cleaning algorithm as mentioned in [4].

ent annotators could review the samples until a majority is formed. The number of annotations required to form a majority shows the difficulty of the sample and is extracted from the budget. To automate the annotation process, new labels are sampled from corresponding label noise distribution formed by $\hat{\mathbf{L}}$, which could additionally be distorted (for simulation purposes) using some temperature value. The remaining samples are again re-prioritized and the process repeats until the budget B is exhausted. Finally, the selector model is also fine-tuned at regular intervals using the corrected labels which improves cleaning performance.

4. Methods

4.1. Creating noisy annotations for the fine-grained datasets

Stanford Dogs with parent symmetric noise The Stanford Dogs dataset [11] is a large-scale FGVC dataset that has 20580 annotated images of dogs belonging to 120 species. The dataset is challenging not only because of its small inter-class differences (see Figure 1a) but also large intra-class variations originating from different poses, colors, occlusions, and background settings. This dataset is a subset of ImageNet [6] and hence, the labels form a semantic hierarchy or taxonomy derived from WordNet [18]. We utilize this hierarchical information to simulate noisy label counts from the crowd as realistically a non-expert human would be most confused between taxonomically similar breeds. While the entire dog breed hierarchy is provided in supplementary material, a sub-section of the hierarchy is illustrated in Figure 2. We term similar categories for a given category to belong in the set of sibling labels ($SiblingDict$) for the category. For Stanford Dogs, we create sibling labels for a dog breed by selecting the breeds that share the same parent node in the hierarchy tree and have no further children nodes. Suppose we are given a noise rate $\epsilon \in [0, 1]$, label counts for each sample $A \in \mathbb{N}$, and a list of sibling labels for all breeds in the dataset, we generate annotations using Algorithm 3.

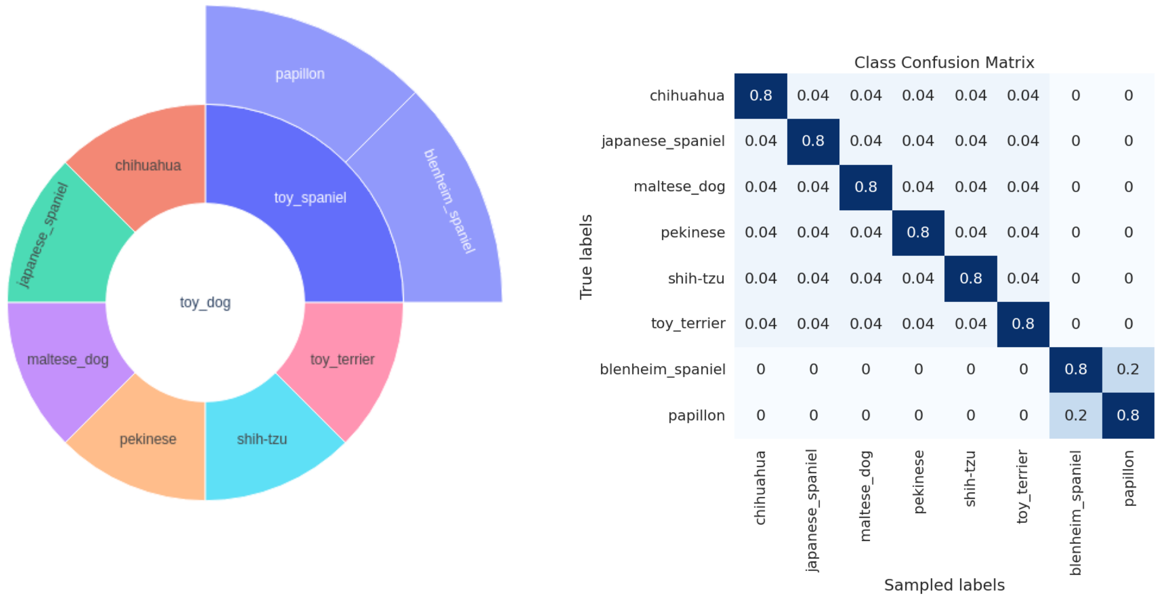
Caltech-UCSD Birds with type symmetric noise The CUB-200-2011 dataset [27] is another FGVC dataset containing 11788 images of 200 bird species. The species label here is associated with a Wikipedia article and arranged by scientific classification *(order, family, genus, species)*. In the absence of a taxonomy graph to identify sibling labels, we choose the different varieties in the type of bird species as the corresponding label siblings. For example, the sibling labels for *Black-footed Albatross* would be $\{$*Laysan Albatross, Sooty Albatross*$\}$, similarly for *Black-billed Cuckoo* the sibling labels would be $\{$*Mangrove Cuckoo, Yellow-billed Cuckoo*$\}$. The label counts are again similarly generated using Algorithm 3.

Input: D : Dataset, N_c : No. of classes
Input: SiblingDict, $\epsilon \in \mathbb{N}$: noise rate, $A \in \mathbb{N}$: total annotators
AllLabelCounts $\leftarrow list()$
for $n = 1, 2, ..., len(D)$ **do**
  label $\leftarrow D[n].label$
  LabelCounts $\leftarrow np.zeros(N_c)$
  for $a = 1, ..., A$ **do**
    if $ChooseNoise(\epsilon, 1 - \epsilon)$ **then**
      annotation $\leftarrow$ $ChooseRandomLabel$(SiblingDict[label])
    else
      annotation $\leftarrow$ label
    LabelCounts[annotation] $\leftarrow$ LabelCounts[annotation] $+1$
  end
  AllLabelCounts.append(LabelCounts)
end
Output: AllLabelCounts

Algorithm 3: Generating label counts for fine-grained categories.

4.2. Sample selection algorithms

Network trained on noisy datasets for sample selection As summarized in Algorithm 2, we initially train a deep neural network to obtain class posteriors, $p_\theta(\hat{y}|\mathbf{x})$, and then use it to identify the noisy labels. To test the effectiveness of the ALC algorithm, we experiment with two types of training methods for the classifier.



(a) (b)

Figure 2. **(a)** shows a sub-section of the intermediate *toy dog* node from the WordNet hierarchy tree for dog breeds. The sibling labels for a breed are chosen to be the labels that share a common parent node and no further children. For example, sibling labels for *pekinese* are $\{$*chihuaua, Japanese spaniel, Maltese dog, shih-tzu, toy terrier*$\}$ **(b)** shows the class confusion matrix used to generate annotations with noise rate $\epsilon = 0.2$, where the noisy label is uniformly sampled from corresponding sibling labels.

First, we train the CNN network with noisy labels and augmented images by minimizing the negative log-likelihood loss. It is expected to perform sub-optimally when prioritizing samples as it would overfit the noise. We call this selection approach *vanilla*. Secondly, we use the *co-teaching* scheme for noise-robust learning. Since the two networks when co-teaching learn the easier cases initially, images that get high-loss values would indicate disagreement with learned knowledge and might have corrupted labels. Training two networks instead of one also prevents a self-confirmation loop which reduces overfitting. At prediction time, the class posteriors are obtained by simply taking a mean of the output logits of the two networks.

Additionally, we compare the results of above mentioned approaches with two baselines as taken in [4] - *oracle* and *random*. The oracle selector simulates perfect ranking in each iteration by accessing the true label distribution, forming an upper bound. The random selector chooses the next label from a uniform distribution and forms a lower bound to the methods.

Acquisition function to prioritize samples for relabelling When under budget constraints, the acquisition function needs to prioritize easier mislabeled samples over the difficult ones while correctly labeled samples have to be ranked the lowest. To this end, we experiment with three variants of the acquisition function. Firstly, we take the cross-entropy from the normalized label counts of the predicted posteriors which corresponds to the estimated noise of the labels. We refer to this method as *Posterior*.

$$\begin{aligned}\Phi_1(\mathbf{x}, \hat{\mathbf{L}}; \theta) &= CE(\hat{\mathbf{L}}, p_\theta) \\ &= -\mathbb{E}_{\hat{\mathbf{L}}/\|\hat{\mathbf{L}}\|_1}[\log p_\theta(\hat{y}|\mathbf{x})] \end{aligned} \tag{1}$$

Secondly, we need to account for how difficult the image is for the prediction task. Hence, we also want to deprioritize ambiguous cases so that easier case gets relabelled first to maximally utilize the budget. This could be included by subtracting the entropy of the sample from the cross-entropy (Equation 1) as we want to reduce the scores of difficult samples. We call this method *Posterior-Entropy*. This formulation is similar to the Expected Information Gain (EIG) in [5].

$$\begin{aligned}\Phi_2(\mathbf{x}, \hat{\mathbf{L}}; \theta) &= CE(\hat{\mathbf{L}}, p_\theta) - \mathbb{H}(p_\theta(\hat{y}|\mathbf{x})) \\ &= -\mathbb{E}_{\hat{\mathbf{L}}/\|\hat{\mathbf{L}}\|_1}[\log p_\theta(\hat{y}|\mathbf{x})] + \mathbb{E}_{p_\theta(\hat{y},\mathbf{x})}[\log p_\theta(\hat{y}|\mathbf{x})] \end{aligned} \tag{2}$$

Finally, since we are working in an AL setting, we also implement a typical acquisition function that selects the most informative samples which is the Bayesian Active Learning by Disagreement (*BALD*) [10, 17]. BALD checks for the mutual information between the sample's label and the model parameters. Hence, it would rather prioritize samples that are not frequently seen during training which

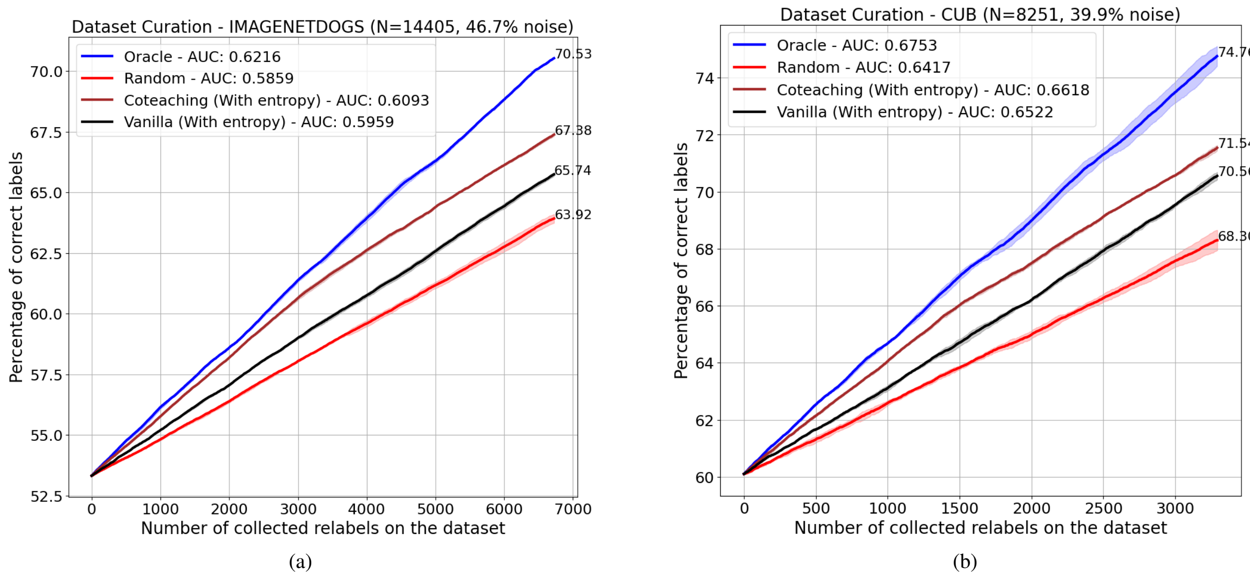


Figure 3. Results of the Active Label Cleaning simulation on the noisy training datasets are plotted for **(a)** Stanford Dogs dataset, and **(b)** CUB-200 bird categorization dataset. Cost-efficient algorithms should be able to maximize the label accuracy (y-axis) for the number of relabels (x-axis) constrained by budget. The cleaning efficiency of the selectors is also reported as the area under the curve (AUC) of each plot. The upper bound is set by oracle sampling (in blue) whereas the lower bound is set by random sampling (in red). The standard deviation over 3 random seeds is shown as a shaded region.

might not be the noisy samples that are easy to relabel.

$$\Phi_3(\mathbf{x}, \hat{\mathbf{L}}; \theta) = \mathcal{I}(\hat{y}, \theta | \mathbf{x}, \hat{\mathbf{L}}) \\ = \mathbb{H}[p(\hat{y}|\mathbf{x}, \hat{\mathbf{L}})] - \mathbb{E}_{\theta|\hat{\mathbf{L}}}[\mathbb{H}[p_\theta(\hat{y}|\mathbf{x})]] \quad (3)$$

4.3. Evaluation metrics

Label Accuracy Since we intend to maximize the number of correctly labeled samples, we use the label accuracy of the dataset as our primary evaluation metric. It is denoted as the percentage of correctly labeled samples in the dataset.

Area-under-the-curve (AUC) The AUC for the label accuracy curve from the ALC procedure provides an overview of the cleaning efficiency for the various selector algorithms and datasets using different relabelling budgets.

Classification accuracy This is simply the top-1 classification accuracy of the classifier models useful in evaluating their performance.

5. Experimental Setup and Results

We closely follow the setup of the ALC framework of [4] using their provided codes. For both of the FGVC datasets, we take a ratio of $7:3$ train-validation split and keep a noise rate of $\epsilon = 0.2$ while generating $A = 50$ label counts for all samples in the dataset. In this way, we obtain true label distribution for each sample. Additionally, to add more ambiguity and better simulate crowd noise, we scale all label distributions with a temperature value of 2.0 which results in a more noise-skewed distribution. From this distribution, we sample our initial labels for all images. Table 1 summarizes the final dataset statistics.

We use the same type of image encoder ResNet50 [9] as well as the same optimizer type and augmentations for both standard vanilla CNN and co-teaching CNNs when training on the initial noisy data. All hyperparameters used for training and the convergence plots are provided in the supplementary material. The final model performance on validation data is summarized in Table 2. The budget (B) for relabelling in the simulation is kept as the expected number of noisy samples in the dataset which assumes that annotators can correctly re-label all noisy samples on their first try ($AUC = 1.0$) which is practically not possible since all our approaches sample the label from a distribution with some randomness (see AUC values in Figure 3). The selector model is fine-tuned every 1000 iteration for 10 epochs with a static learning rate of 10^{-6}. We additionally run the ALC simulation using 3 seeds to check for any significant deviations. All codes and experimental setups are available at - https://github.com/PalAvik/alclean.

5.1. RQ1: Cost-effective Label Cleaning

The results of the sequential relabelling process using the discussed sample selection methods on the training splits of both FGVC datasets using the Posterior-Entropy acquisition function (Equation 2) are plotted in Figure 3. We observe that both vanilla and co-teaching methods perform better

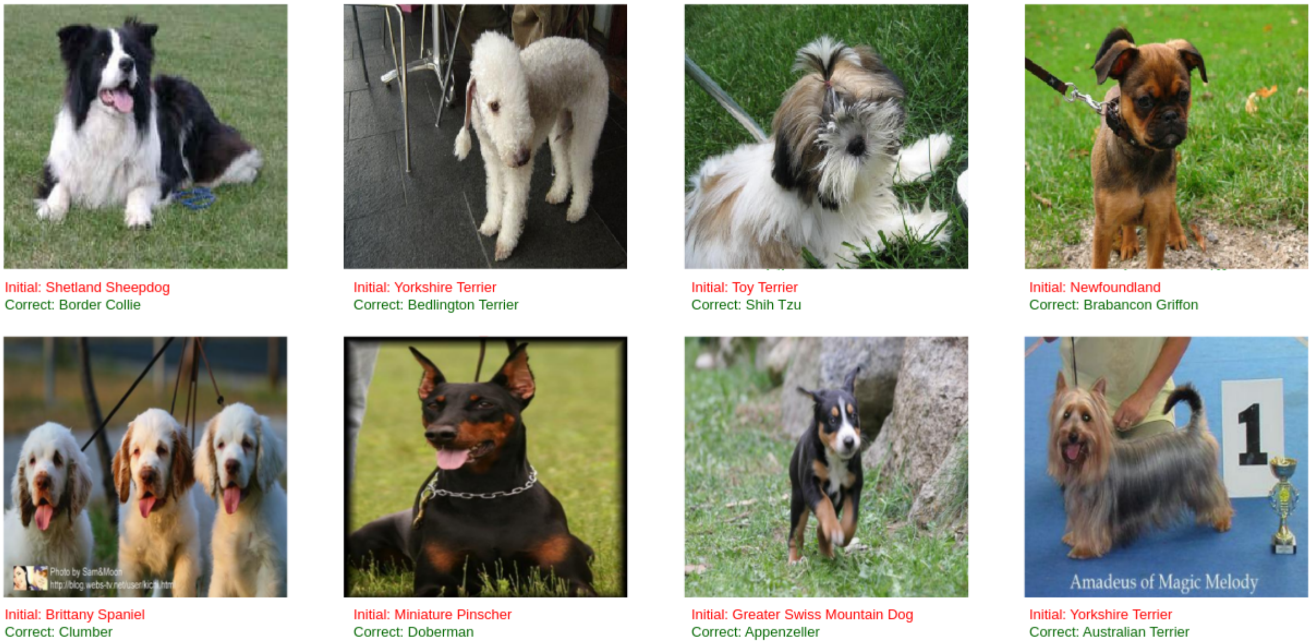


Figure 4. Few images with noisy labels picked from top-10 (in the first row) and bottom-10 (in the second row) when ranked for relabelling during the first iteration of ALC. The high-prioritized images share very different features between the correct and the incorrect initial dog breed labels. Contrarily, low-prioritized images share very similar features making them difficult to re-annotate.

Table 1. Statistics of FGVC datasets with noise.

	Train		Val	
Dataset	Size	Noise %	Size	Noise %
Stanford Dogs	14405	46.7%	6175	46.9%
CUB-200	8251	39.9%	3537	39.5%

Table 2. Classification accuracy (%) of the vanilla and co-teaching classifiers on both noisy and clean versions of the validation set.

Dataset	Classifier	Noisy val	Clean val
Stanford Dogs	vanilla	18.494	27.385
	co-teaching	21.781	33.23
CUB-200	vanilla	24.682	33.051
	co-teaching	26.096	34.295

than the random selector which shows that prioritizing easier labels with noise forms a cost-effective way for label cleaning under budget constraints. For example in the plot for Stanford Dogs (Figure 3a), the vanilla and co-teaching approaches can reach a label accuracy of 62.5% using 1.2× and 1.5× fewer re-annotations respectively. Similarly for CUB (Figure 3b), the vanilla and co-teaching approaches can reach a label accuracy of 68% using 1.4× and 1.5× fewer re-annotations respectively. We also see that the co-teaching approach performs better than vanilla in prioritizing noisy labels, showcasing that some noise-robust learning complements the ALC procedure.

For qualitative analysis, we also plot some images from the Stanford Dogs dataset which is at the top of the priority for relabelling in the first iteration of ALC in row 1 of Figure 4 and some bottom-ranked images in row 2 of Figure 4. We can observe that the top-ranked images have initial labels of dog breeds that have very different features and probably could be easily re-annotated, for example, the noisy initial label - Shetland sheepdog has very different features from the observed image of Border Collie. Similarly, we observe that the bottom-ranked images are indeed difficult cases, for example, the breed Yorkshire Terrier shares many similar features with an Australian Terrier and might need more re-annotations from experts which utilizes resources from the budget.

Hence, both quantitatively and qualitatively we note that the **ALC framework is a cost-effective method for relabelling FGVC datasets**.

5.2. RQ2: Acquisition function better at the label cleaning task

We experiment with the acquisition functions (AFs) described in Section 4.2 for scoring and prioritizing samples for relabelling. We first run ALC applying the AFs using both vanilla and co-teaching selectors to clean the val-

Table 3. Classification accuracy (%) before and after label cleaning using different acquisition functions. The best accuracy and approach are highlighted in bold for each dataset-selector combination.

Dataset	Selector/Classifier	Before cleaning	Acquisition Function	After cleaning
Stanford Dogs	vanilla	18.494	BALD	20.664
			Posterior	22.121
			Posterior-Entropy	**22.219**
	co-teaching	21.781	BALD	23.158
			Posterior	26.316
			Posterior-Entropy	**26.591**
CUB-200	vanilla	24.682	BALD	26.378
			Posterior	27.622
			Posterior-Entropy	**27.735**
	co-teaching	26.096	BALD	27.113
			Posterior	29.658
			Posterior-Entropy	**30.054**

idation set of the corresponding FGVC dataset using the same relabelling budget. We then evaluate the classifier model (same as the selector) using the cleaned validation set. The results of this experiment are summarized in Table 3. **The Posterior-Entropy AF shows the best classification performance in all experiments indicating its capability to prioritize samples for better budget efficiency**. It marginally improves upon the Posterior method proving that adding the entropy term to discern between ambiguous and simple noise is helpful. We expect the margin of improvement to increase further when there is more ambiguous noise in the data. The BALD AF performs poorly which shows that prioritizing samples based on their disagreement does not necessarily correspond to noisy labels and is hence not suitable for the label cleaning task.

6. Conclusions, Limitation, & Future Work

This work investigated the effectiveness of the Active Label Cleaning framework proposed by Bernhardt et al. [4] when we have a noisy Fine-grained Visual Categorization (FGVC) dataset. We experimented with Stanford Dogs and the Caltech-UCSD Birds with artificially generated annotations from the crowd which simulates noise based on semantic (taxonomical) connection between labels with similar image features. Based on our experimental results, we can conclude that the framework can efficiently clean noisy samples in FGVC datasets under budget constraints. We also show that typical acquisition functions used in Active Learning such as BALD are not well suited for the label-cleaning task. An acquisition function that apprehends the noisiness of sample from model posteriors along with the penalty of corresponding sample ambiguity captured from entropy is better suited for scoring and ranking samples for the task.

Due to system memory limitations, we were not able to experiment with larger budgets when relabelling and hence could not reach a point of a fully cleaned dataset. This would have provided clearer demarcation between the performance of the various selection algorithms. Additionally, due to time constraints, we could not experiment with more noise-robust learning methods or even self-supervised methods and leave this for future work. Furthermore, it would also be interesting to include the hierarchical/taxonomical information for ranking noisy samples.

References

[1] Anelia Angelova and Shenghuo Zhu. Efficient object detection and segmentation for fine-grained recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 811–818. IEEE Computer Society, 2013. 2

[2] Enrique Estellés Arolas and Fernando González-Ladrón-de-Guevara. Towards an integrated crowdsourcing definition. *J. Inf. Sci.*, 38(2):189–200, 2012. 1

[3] Yingbin Bai and Tongliang Liu. Me-momentum: Extracting hard confident examples from noisily labeled data. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9292–9301. IEEE, 2021. 3

[4] Mélanie Bernhardt, Daniel C. Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C. Tezcan, Miguel Monteiro, Shruthi

Bannur, Matthew P. Lungren, Aditya Nori, Ben Glocker, Javier Alvarez-Valle, and Ozan Oktay. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications*, 13(1), 2022. 1, 2, 3, 4, 5, 6, 8

[5] Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented Bayesian active learning. *International Conference on Artificial Intelligence and Statistics*, 2023. 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 1, 4

[7] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, 25(5):845–869, 2014. 1

[8] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8536–8546, 2018. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 6

[10] Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745, 2011. 5

[11] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011. 1, 2, 4

[12] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 3–12. ACM/Springer, 1994. 3

[13] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep LAC: deep localization, alignment and classification for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1666–1674. IEEE Computer Society, 2015. 2

[14] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 6226–6236. PMLR, 2020. 2

[15] Yang Liu and Mingyan Liu. An online learning approach to improving the quality of crowd-sourcing. *IEEE/ACM Trans. Netw.*, 25(4):2166–2179, 2017. 1

[16] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah M. Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 6543–6553. PMLR, 2020. 2

[17] David J. C. MacKay. Information-based objective functions for active data selection. *Neural Comput.*, 4(4):590–604, 1992. 5

[18] George A. Miller. WORDNET: A lexical database for english. In *Human Language Technology, Proceedings of a Workshop held at Plainsboro, New Jerey, USA, March 8-11, 1994*. Morgan Kaufmann, 1994. 4

[19] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi-Phuong-Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. SELF: learning to filter noisy labels with self-ensembling. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 3

[20] Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9616–9625. IEEE, 2019. 1

[21] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322, 2010. 1

[22] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 4331–4340. PMLR, 2018. 3

[23] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 3

[24] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1917–1928, 2019. 3

[25] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: refurbishing unclean samples for robust deep learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 5907–5915. PMLR, 2019. 3

[26] Xinxing Tan, Zemin Dong, and Hualing Zhao. Robust fine-grained image classification with noisy labels. *Vis. Comput.*, 39(11):5637–5650, 2023. 2

[27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2023. 1, 2, 4

[28] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-

ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3462–3471. IEEE Computer Society, 2017. 1

[29] Xinshao Wang, Yang Hua, Elyor Kodirov, David A. Clifton, and Neil M. Robertson. Proselflc: Progressive self label correction for training robust deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 752–761. Computer Vision Foundation / IEEE, 2021. 3

[30] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 322–330. IEEE, 2019. 2

[31] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13723–13732. Computer Vision Foundation / IEEE, 2020. 3

[32] Qi Wei, Haoliang Sun, Xiankai Lu, and Yilong Yin. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXX*, pages 516–532. Springer, 2022. 3

[33] Qi Wei, Lei Feng, Haoliang Sun, Ren Wang, Chenhui Guo, and Yilong Yin. Fine-grained classification with noisy labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11651–11660. IEEE, 2023. 2

[34] Ning Zhang, Jeff Donahue, Ross B. Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 834–849. Springer, 2014. 2

[35] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8792–8802, 2018. 2