Crowd Teaching with Imperfect Labels

Yao Zhou*
University of Illinois at Urbana Champaign
yaozhou3@illinois.edu

Arun Reddy Nelakurthi
Samsung Research America
arunreddy.nelakurthi@gmail.com

Ross Maciejewski
Arizona State University
rmacieje@asu.edu

Wei Fan
Tencent America
wei.fan@gmail.com

Jingrui He
University of Illinois at Urbana Champaign
jingrui@illinois.edu

ABSTRACT

The need for annotated labels to train machine learning models led to a surge in crowdsourcing - *collecting labels from non-experts*. Instead of annotating from scratch, given an imperfect labeled set, how can we leverage the label information obtained from amateur crowd workers to improve the data quality? Furthermore, is there a way to teach the amateur crowd workers using this imperfect labeled set in order to improve their labeling performance? In this paper, we aim to answer both questions via a novel interactive teaching framework, which uses visual explanations to simultaneously teach and gauge the confidence level of the crowd workers.

Motivated by the huge demand for fine-grained label information in real-world applications, we start from the realistic and yet challenging assumption that neither the teacher nor the crowd workers are perfect. Then, we propose an adaptive scheme that could improve both of them through a sequence of interactions: the teacher teaches the workers using labeled data, and in return, the workers provide labels and the associated confidence level based on their own expertise. In particular, the teacher performs teaching using an empirical risk minimizer learned from an imperfect labeled set; the workers are assumed to have a forgetting behavior during learning and their learning rate depends on the interpretation difficulty of the teaching item. Furthermore, depending on the level of confidence when the workers perform labeling, we also show that the empirical risk minimizer used by the teacher is a reliable and realistic substitute of the unknown target concept by utilizing the unbiased surrogate loss. Finally, the performance of the proposed framework is demonstrated through experiments on multiple real-world image and text data sets.

KEYWORDS

Interactive teaching, Personalized crowdsourcing, Explanation.

ACM Reference Format:
Yao Zhou, Arun Reddy Nelakurthi, Ross Maciejewski, Wei Fan, and Jingrui He. 2020. Crowd Teaching with Imperfect Labels. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3366423.3380099

*Yao Zhou and Arun Reddy Nelakuithi are equally contributed.

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380099

1 INTRODUCTION

The prevalence of complex models, such as deep neural networks, has enabled unprecedented breakthroughs in a variety of real-world applications, including data mining, computer vision, natural language processing, etc. Many of these models are supervised in nature, which assume abundant labeled data for model training obtained from the annotation process [5, 10, 44, 45]. However, current annotations may be insufficient for fine-grained categorization tasks. These tasks focus on discriminating items at the subordinate level. Examples of fine-grained categorization tasks include: distinguishing images showing different breeds of cats or dogs (as opposed to simply distinguishing if the image contains a cat or a dog); sorting various topics of documents, or categories of emails for the tasks in text classification; analyzing multiple types of cancer tissue images, types of arrhythmia ECG signals for the tasks in medical diagnosis. As the demand for such fine-grained categorization tasks has increased [6], the qualitative and quantitative requirements for annotations have also significantly increased.

In fine-grained categorization, the distinctions between categories are usually subtle and highly local. As such, fine-grained categorization is more challenging than the common high-level classification tasks. Fine-grained categorization has fewer discriminative features and often suffers from lower quality labels. In terms of features, distinguishing between the image of a house and the image of a dog is easy since highly distinct visual features can be extracted to classify these two categories. However, in fine-grained categorization tasks, such as classifying a domestic cat and a wildcat, visually discriminative features are often scarce and tend to be more locally distributed. For example, the Lynx (one species of the wildcat, Figure 1) usually has larger paws, longer chest fur, and pointed ear tufts. However, the face of a lynx can easily be mistaken as a common domestic cat. If only the facial features are considered, the fine-grained classification can fail. Thus, the task of categorizing fine-grained items requires human labelers to have a strong domain knowledge about the subject area, so that s/he can correctly identify visual cues and perform annotations. Unfortunately, the most popular means of collecting the labeled data is through crowdsourcing platforms (e.g., Amazon Mechanical Turk, Figure Eight, etc.) where annotations are outsourced to a group of mostly unskilled online workers. While such crowdsourcing may

Figure 1: An example of fine-grained cat categorization. Left: original input image of a Lynx. Middle & Right: visual cues of predicting it as domestic cat and wildcat generated using Grad-CAM [30].

be adequate for high level categories (such as the house versus dog example), labels obtained from these crowdsourcing platforms may be insufficient for training fine-grained categorization models even after cleaning and aggregation using state-of-the-art techniques.

For the sake of collecting and post-processing labels that will be suitable for fine-grained categorization, several approaches have been proposed to identify high-quality crowdsourcing workers who have a good understanding of a specific domain [28, 44]. (1) Providing general instructions to all workers regarding the specific labeling concept. (2) Mixing the "gold standard" questions with true crowdsourcing questions and filtering out the unqualified workers who continuously provide incorrect and random answers on the "gold standard" questions. (3) Modeling and estimating the expertise of these workers, then down-weighting the votes of the weak workers. These approaches can potentially reduce the error within the labels, but they also suffer from several problems such as low annotation rates and limited numbers of qualified workers, all of which place a huge burden on the mechanism designer when specifying the annotation tasks.

Therefore, in this paper, we focus on improving crowdsourcing annotations for fine-grained classification tasks by introducing a novel interactive learning and teaching framework. It is designed to increase the model performance by relabeling and expanding the existing labeled set. It is also able to improve the labeling expertise of the crowd workers or learners by providing explanations for the visual cues. The major contributions of this work include:

- **Framework**: The proposed adaptive crowd teaching framework is designed to have three-step interactions between the teacher and the learners. The learners could provide confident annotations to either relabel the labeled items or give initial label to the unlabeled ones.
- **Adaptation**: The main objective of teaching considers both the influence of items by incorporating the principle of curriculum learning and the personalized learning progress of the learner by balancing between the usefulness and diversity of the teaching sequence.
- **Analysis**: A theoretical bound of this interactive teaching scheme has been provided to fit the realistic teaching scenario of using empirical cost minimizer as the target concept. The confident labels provided by the workers can also reduce the error rate of the labeled items.
- **Experiments**: We have built a web-based platform to conduct the experiments. The teaching effectiveness and teaching reliability of the framework have been validated on three real data sets of images and text documents.

The remainder of the paper is organized as follows. We first introduce the problem definition in Section 2. Then, we formally present the proposed interactive teaching framework in Section 3 followed by descriptions of the algorithm and the discussions of the model in Section 4. The experiments and results are presented in Section 5 and the related work is discussed in Section 6. We conclude the paper in Section 7.

2 PROBLEM DEFINITION

In this section, we provide the preliminary regarding the interpretable explanations. Next, the notation and the definition of adaptive crowd teaching are formally presented.

2.1 Preliminary: Interpretable explanations

We denote the original prediction model $\psi(\cdot;\theta)$, parameterized by θ, as the model that we want to explain. Understanding the way that a model (e.g., random forest, convolutional neural network, etc.) makes decisions is crucial in understanding the model's behaviors. Proper explanations can engender trust from the users and provide insights into model properties [21, 29, 32]. The essential criterion of explanations is that they should be interpretable by humans.

2.1.1 Feature-additive explanation. We denote $g(\cdot)$ as the feature-additive explanation model which uses a simplified feature $\boldsymbol{x'}$ (which is treated as the interpretable input due to its low dimensionality comparing with the original feature $\boldsymbol{x}$) as its input. This interpretable input $\boldsymbol{x'}$ is usually mapped from the original input $\boldsymbol{x}$ through a mapping function as: $\boldsymbol{x} = h_x(\boldsymbol{x'})$. The additive feature explanation models are local methods that are designed to ensure $g(\boldsymbol{x'}) \approx \psi\big(h_x(\boldsymbol{x'});\theta\big)$. Overall, the additive feature method will have a linear interpretable model:

$$g(\boldsymbol{x'}) = \phi_0 + \sum_{j=1}^{d} \phi_j \boldsymbol{x}'_j \tag{1}$$

where d is the number of simplified features. One of the most prominent text explanation models using additive features is LIME [29], which minimizes the following objective:

$$\min_{g} l(\psi, g, \pi_{x'}) + \Omega(g) \tag{2}$$

where Ω is the complexity measurement used to penalize the explanation model $g(\cdot)$. LIME imposes the faithfulness of the explanation model $g(\cdot)$ towards the original model $\psi\big(h_x(\boldsymbol{x'});\theta\big)$ by enforcing through a cost $l(\psi, g, \pi_{x'})$ over a set of perturbed samples weighted by a local distance kernel $\pi_{x'}$.

2.1.2 Saliency-based Explanation. Saliency maps are often considered to be explanatory and they are extremely useful in categorization tasks to determine which area of the observation is relevant. One of the most popular approaches for generating saliency maps is Class Activation Mapping (CAM) [30, 42]. Utilizing Gradient-weighted CAM (Grad-CAM) [30], the saliency explanation is generated using gradient information flowing into the last convolution layer of the neural network to understand the importance of each neuron for a decision of interest.

Given an input image, let o^c be the score for class c after the forward pass of the neural network. Then, the gradient of o^c with respect to the k-th activation map A^k of the last convolutional layer

Table 1: Summary of symbols

Category	Symbol	Definition
General setting	$\mathcal{D}_L$	Imperfect labeled data set
	$\mathcal{D}_U$	Unlabeled data set
	$\boldsymbol{x}_i, y_i$	Feature vector and aggregated imperfect label of the $i-$thitem
	$z = (\boldsymbol{x}, y)$	Item tuple of feature and label
	L_{x_i}	Setofcrowdsourcinglabels foritem $\boldsymbol{x}_i$
	N, n, m	# ofallitems, # ofitemswith initiallabels, # offeatures
	$\mathbb{I}(\cdot)$	Indicator function
Teacher Model	$\psi(\cdot;\theta), \hat{y}$	Predictionmodeloftheteacher anditsrealvalueprediction
	$C(\cdot,\cdot), \tilde{C}(\cdot,\cdot)$	Costfunctionandthesurrogate costfunctionoftheteacher
	$\mathcal{I}_{up,params}(z)$	Influence of upweighting z
	$\mathcal{I}_{pert,loss}(z, z_{test})$	Influenceofperturbingthe labelof z onatestitem z_{test}
	$p_1, p_2, \ldots, p_N$	Influence scores
Learner Model	$\boldsymbol{v}_t$	Memorymomentumatthe $t-$thteachingstep
	$\mathcal{L}(\cdot,\cdot)$	Learning loss of the learner
	w_{I_t}	Rescalingcoefficientoftheselected item(index I_t)in $t-$thteachingstep
	$D(\boldsymbol{e})$	Interpretabledifficultyof anitem′sexplanation $\boldsymbol{e}$

is calculated as $\frac{\partial o^c}{\partial A^k}$. The overall importance weight α_k^c of the k-th activation map for c-th class is global average pooled as:

$$\alpha_k^c = \frac{1}{Z}\sum_{h=1}\sum_{v=1}\frac{\partial o^c}{\partial A_{h,v}^k} \tag{3}$$

where Z is the norm of the gradient for normalization purposes. Next, the forward activation maps of the last convolutional layer are added up using this importance weights and a ReLU operation is followed to obtain the saliency map for each class as $\text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$.

2.2 Notation

All frequently used notations in the remaining of this paper are summarized in Table 1. We denote $\mathcal{X} \subset \mathrm{R}^m$ as the m-dimensional feature representation of all examples (e.g., images or documents) and $\mathcal{Y} = \{+1, -1\}$ as the collection of labels[1]. In crowdsourcing, each item usually has one or more labels collected from the workers. For each $\boldsymbol{x}_i \in \mathcal{X}$, we let $(\boldsymbol{x}_i, L_{x_i})$ be the corresponding input pair, where $L_{x_i} = \{y_i^1, y_i^2, \ldots, y_i^{\tau_i}\} \subset \mathcal{Y}$ is the set of imperfect labels for $\boldsymbol{x}_i$, and τ_i is the number of such labels for $\boldsymbol{x}_i$. We let $\mathcal{D}_L = \{(\boldsymbol{x}_i, L_{x_i})\}_{i=1}^n$ be the imperfect labeled data set collected from the crowdsourcing workers and $\mathcal{D}_U = \{(\boldsymbol{x}_i, L_{x_i})\}_{i=n+1}^N$ be the unlabeled data set. Each item in $\mathcal{D}_L$ has $\tau_i \geq 1$ and each item in $\mathcal{D}_U$ has $L_{x_i} = \emptyset$. Notice that in conventional supervised learning setting, the collected set of labels L_{x_i} for each item $\boldsymbol{x}_i$ are usually aggregated as one label y_i to train the model. After aggregation, following [25], we also define the class-conditional error rate ρ_{+1}, ρ_{-1}

[1]Although this paper focuses on the binary classification setting, the proposed techniques can be naturally generalized to multi-class setting.

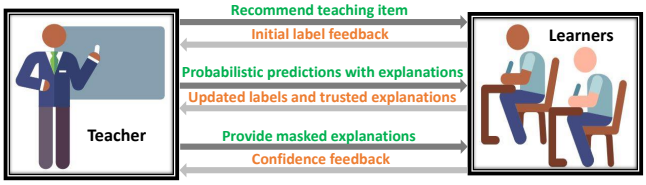


Figure 2: An overview of the interactive learning and teaching framework.

of the labels in $\mathcal{D}_L$ as:

$$\rho_{+1} = P(y = -1|y_{gt} = +1),\ \rho_{-1} = P(y = +1|y_{gt} = -1) \tag{4}$$

where y_{gt} is the underlying ground truth label and $\rho_{+1} + \rho_{-1} < 1$. It should be noticed that in remaining context, the phrases of worker and learner are used interchangeably.

2.3 Problem Setting

Given an imperfect labeled data set $\mathcal{D}_L$ that has a mixture of mostly correct labels and possible a small fraction of incorrect labels, an unlabeled data set $\mathcal{D}_U$, a real-valued prediction model $\psi(\cdot;\theta)$, and a group of crowd workers, the proposed framework is designed to simultaneously target the following two objectives:

- **Objective I**: The framework aims to improve the prediction model's performance by providing a new labeled data set $\mathcal{D}_{new}$ with better label qualities. $\mathcal{D}_{new}$ includes three groups of data items: (**a**). A group of items that originally belong to $\mathcal{D}_L$ but have been re-labeled and verified by these crowd workers; (**b**). A second group of items that also belong to $\mathcal{D}_L$, but their labels stay untouched; (**c**). Another group of items that originally belong to $\mathcal{D}_U$ but eventually get labels from crowd workers. With the new data set $\mathcal{D}_{new}$, the prediction model $\psi(\cdot;\theta)$ is expected to have improved performance.
- **Objective II**: The framework uses the prediction model $\psi(\cdot;\theta)$, which is not perfectly trained because it is optimized on the imperfect labeled set $\mathcal{D}_L$, as the teacher to interactively teach the crowd workers by showing them a personalized sequence of items with probabilistic prediction labels and fine-grained visual explanations. In this way, the crowd workers would have improved labeling abilities for the annotation tasks after teaching.

3 ADAPTIVE TEACHING AND LEARNING

Let the prediction model $\psi(\cdot;\theta)$ be the teacher, where the parameter vector θ is obtained by minimizing a cost function $C(\cdot,\cdot)$ over the labeled training data. To teach the crowd workers, we assume that the teacher knows the empirical target concept[2] $\hat{\theta}$ beforehand, which is defined as:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} C\big(\psi(\boldsymbol{x}_i;\theta), y_i\big) \tag{5}$$

where n represents the number of training examples in $\mathcal{D}_L$.

[2]In practice, the true target concept θ_* is usually unknown. In Section 4, we will demonstrate that the risk of the empirical target concept is bounded, and it is reasonable to perform teaching with $\hat{\theta}$.

Notice that in our setting, the aggregated labels y_i, which are obtained from the imperfect label sets, still include errors in nature. In order to obtain effective prediction models in the presence of labeling error, we adopt the surrogate cost function proposed in [25], which is defined as follows:

$$\tilde{C}(\hat{y}, y) := \frac{(1-\rho_{-y})C(\hat{y}, y) - \rho_y C(\hat{y}, -y)}{1-\rho_{+1}-\rho_{-1}} \tag{6}$$

where $\hat{y} = \psi(\boldsymbol{x}; \theta)$ and the class dependent error rate of the surrogate cost is defined as: $\rho_y = \rho_{+1}^{\mathbb{I}(y=+1)}\rho_{-1}^{\mathbb{I}(y=-1)}$. Lemma of [25] shows that this surrogate cost on the imperfect labels is an unbiased estimator of the original cost on the underlying true label y_{gt} as:

Lemma 3.1. *If $C(\hat{y}, y_{gt})$ is a bounded cost function and $\hat{y}$ is a real-valued prediction. Then the surrogate cost $\tilde{C}(\hat{y}, y)$ constructed in Eqn. (6) has $\mathbb{E}_y\left[\tilde{C}(\hat{y}, y)\right] = C(\hat{y}, y_{gt})$ for any $\hat{y}$.*

3.1 Interactions Between Teacher and Learners

An overview of the interactive learning and teaching between learners and the teacher is provided in Figure 2. Each teaching iteration includes the following steps:

- **First**, the teacher recommends and shows an item to the learner based on the ranking score, which considers both the learner's learning progress and the item's influence on the prediction model. The learner is then asked to provide his/her initial label for item.
- **Second**, the teacher shows the labels made by its prediction model along with their probability scores and the visual explanations. The learner is then asked to give his/her updated labels and trusted explanations.
- **Third**, the teacher evaluates the learner's confidence regarding his/her annotations using the masked explanations. If learner's confidence is high, the provided label will be added into the label set of the item. Otherwise, the label will not be recorded.

3.2 Learner Model

In the process of interactive teaching, each learner observes a sequence of items and their corresponding label information as well as the visual explanations provided by the teacher. The most prevalent modeling [19, 20, 40, 47] of the learners is assuming that they made linear decisions, i.e., $\langle \theta, \boldsymbol{x} \rangle$, and have an iterative gradient-based procedure for concept learning. The adaptivity of the learner is reflected in three aspects: each learner has a personalized memory decay regarding the learned concepts with an exponential rate; the learning rate is adjustable in terms of the item interpretation difficulty; and the learning concept of each learner is estimated from his/her previous labels.

Similar to iterative crowd teaching [47] , we also treat the learner as having an exponential decayed retrievability of memory:

$$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + \nabla_{\theta_{t-1}} \mathcal{L}(\theta_{t-1}^\top \boldsymbol{x}_{I_t}, y_{I_t}) \tag{7}$$

where we denote I_t as the index of the selected teaching item in the t-th teaching step, and $\mathcal{L}(\cdot, \cdot)$ is the learning loss of the learner. Usually, the initial memory momentum $\boldsymbol{v}_0$ is set to $\mathbf{0}$, then, the memory momentum of the learner for the t-th learning step is:

$$\boldsymbol{v}_t = \sum_{r=1}^{t} \beta^{t-r} \nabla_{\theta_{r-1}} \mathcal{L}(\theta_{r-1}^\top \boldsymbol{x}_{I_r}, y_{I_r}) \tag{8}$$

Here, $\beta \in (0, 1)$ is the personalized memory decay rate of the learner and r is the index of teaching iteration. The learners use the gradient based learning procedure to improve their concepts in an iterative way with learning rate η_t:

$$\theta_t = \theta_{t-1} - \eta_t w_{I_t} \boldsymbol{v}_t \tag{9}$$

where the learning rate also depends on a re-scaling coefficient w_{I_t} that has a connection with the interpretable difficulty of the item. Intuitively, the teaching items that are easier to interpret should have a larger learning rate (by setting w_{I_t} larger) and the difficult items should have relatively smaller learning rate (with smaller w_{I_t}). After the t-th teaching step, the learner applies a linear model, i.e., $\langle \theta_t, \boldsymbol{x} \rangle$, to make predictions using the learned concept θ_t.

3.3 Teacher Model

The teacher observes all items (e.g., images, text, etc.) which include their feature representation and labels. The teacher is also assumed to have access to the learner's learning procedure. Given the empirical target concept $\hat{\theta}$, the teacher aims to maximize the learner's speed of convergence [11] in terms of learning by minimizing the distance of the learner's current concept from the empirical target concept in two consecutive iterations t and $t-1$. Then, the teaching objective can be decomposed as:

$$\begin{aligned} &\left\|\theta_t - \hat{\theta}\right\|_2^2 - \left\|\theta_{t-1} - \hat{\theta}\right\|_2^2 \\ =&\eta_t^2 w_{I_t}^2 \left\|\sum_{r=1}^{t} \beta^{t-r} \nabla_{\theta_{r-1}} \mathcal{L}(\theta_{r-1}^\top \boldsymbol{x}_{I_r}, y_{I_r})\right\|_2^2 \\ &-2\eta_t w_{I_t} \left\langle \theta_{t-1} - \hat{\theta}, \sum_{r=1}^{t} \beta^{t-r} \nabla_{\theta_{r-1}} \mathcal{L}(\theta_{r-1}^\top \boldsymbol{x}_{I_r}, y_{I_r}) \right\rangle \end{aligned} \tag{10}$$

For t-th teaching iteration, the teacher only aims to recommend the teaching item $(\boldsymbol{x}_{I_t}, y_{I_t})$ with its explanation $\boldsymbol{e}_{I_t}$. Therefore, all the terms in Eqn. (10) that don't include t-th teaching item can be excluded from the objective. With some simplifications, the goal of recommending the next teaching item is formulated as a pool-based searching problem[3]:

$$\begin{aligned} (\boldsymbol{x}_{I_t}, y_{I_t}, \boldsymbol{e}_{I_t}) = \arg\min_{\boldsymbol{x}, y, \boldsymbol{e}} \; &\eta_t^2 w_{I_t}^2 \left\|\nabla_{\theta_{t-1}} \mathcal{L}(\theta_{t-1}^\top \boldsymbol{x}, y) + \boldsymbol{v}_{t-1}\right\|_2^2 \\ &-2\eta_t w_{I_t} \left\langle \hat{\theta} - \theta_{t-1}, -\nabla_{\theta_{t-1}} \mathcal{L}(\theta_{t-1}^\top \boldsymbol{x}, y) \right\rangle \end{aligned} \tag{11}$$

From the above objective, we know that the learning direction of a learner is $-\nabla_{\theta_{t-1}} \mathcal{L}(\theta_{t-1}^\top \boldsymbol{x}, y)$. The teacher prefers the negative gradient of the t-th teaching item to be similar as the concept momentum $\boldsymbol{v}_{t-1}$ in order to increase the teaching sequence diversity. At the same time, the second term of the objective also suggests that the negative gradient of the teaching item has a large correlation (i.e., large inner product) with the optimal learning direction

[3]The item $\boldsymbol{x}$ is either from $\mathcal{D}_L$ or $\mathcal{D}_U$, its label y is predicted using teacher's model $\psi(\cdot; \theta)$, and the corresponding explanation $\boldsymbol{e}$ is generated utilizing the methods described in Subsection 2.1.

$\hat{\theta} - \theta_{t-1}$. Overall, this teaching scheme aims to simultaneously maximize the *teaching diversity* and *teaching usefulness*.

It should be noticed that the candidate pool of teaching items includes both $\mathcal{D}_L$ and $\mathcal{D}_U$, which means that the framework either allows the worker to re-label current teaching item $\boldsymbol{x}_{I_t}$ (if it belongs to $\mathcal{D}_L$) by expanding $L_{x_{I_t}}$ with one more label or annotate $\boldsymbol{x}_{I_t}$ (if it belongs to $\mathcal{D}_U$) by adding the first label to its empty label set.

3.4 Interpretation Difficulty

An item (e.g., image or text) with a smaller attention region (e.g., a small area of pixels or a few keywords) being highlighted would indicate that less effort is needed for a worker to interpret the visual explanations. For example, if the breed of a cat in one image has been correctly predicted by the model, the explanation which highlights the attention area of the head & the body would be much easier to understand than the explanation that highlights the whole image that includes the cluttered background because the background does not matter during predictions. Similarly, the explanation that highlights a few representative words, e.g., "key", "security", for an encryption-related document, will be easier to interpret than the document with too many words being highlighted. We define the interpretation difficulty of an explanation $\boldsymbol{e}$ as the entropy of all entries on a saliency map (e.g. for image items) or an additive feature vector (e.g. for text items):

$$D(\boldsymbol{e}) = \begin{cases} -\sum_j \boldsymbol{e}_j \log(\boldsymbol{e}_j), & \Leftrightarrow \text{Feature-additive explanation} \\ -\sum_{h,v} \boldsymbol{e}_{h,v} \log(\boldsymbol{e}_{h,v}), & \Leftrightarrow \text{Saliency-based explanation} \end{cases} \tag{12}$$

where the $\boldsymbol{e}_j$ is the j-th interpretable feature of the additive feature vector and $\boldsymbol{e}_{h,v}$ is the visual cue at the (u, v) location on the saliency map. Furthermore, no matter if it is the explanation of the saliency map or the additive feature vector, all entries could be normalized to span the range of $[0, 1]$ and satisfy the property as a probability distribution. Since the entropy of explanation is the proxy of the interpretation difficulty, then we can properly encode the re-scaling coefficient of a selected teaching item (with index I_t at t-th teaching iteration) as $w_{I_t} = e^{1-\frac{D(\boldsymbol{e}_{I_t})}{\kappa}}$. Usually we set $\kappa = \max_{\boldsymbol{e}} D(\boldsymbol{e})$ to be the maximum interpretation difficulty of all possible explanations (i.e., $\boldsymbol{e}$ is an uniformly distributed explanation). The re-scaling coefficient will always have a value of 1 when the framework is reduced to the setting of standard iterative crowd teaching (by demonstrating the whole image or document without providing visual explanations).

3.5 Item Influence

The existing teaching scheme usually assumes all items have an equal influence on the prediction model. However, in real-world teaching, a more effective strategy is following the principle of curriculum learning [2], which encourages the recommended teaching sequence to have tasks that range from easy to difficult. Intuitively, if the learners are making linear decisions during learning, the easy items should be demonstrated to them in their earlier learning stage so that they can make fewer mistakes. These easy items should have some small influence on the prediction model as if they are usually the data points with a large marginal distance in the feature space. Gradually, with the more teaching iterations, the more difficult items should be shown to the workers who should already have improved their labeling expertise. These difficult items usually have large influences (i.e., changing their labels would impact the model behaviors significantly) on the prediction model because they are the data points with incorrect labels or the data points that have small marginal distances to the target concept.

Then, the overall goal becomes teaching the crowd workers by demonstrating a personalized sequence of items that are effective for the concept learning as well as having increasing influences. To begin with, we compute the influence scores $p_1, p_2, \ldots, p_N$ over all items, which is defined as the **model prediction's change w.r.t. the label perturbations**. We denote the item with its label as $z = (\boldsymbol{x}, y)$, and we further define its label perturbed version as $z_\delta = (\boldsymbol{x}, y + \delta)$. For simplicity, we denote $\tilde{C}(\psi(\boldsymbol{x}; \theta), y)$ as $\tilde{C}(z, \theta)$. Consider the perturbation $z \rightarrow z_\delta$, and let $\hat{\theta}_{z_\delta, -z}$ be the empirical risk minimizer on the training items with z_δ in place of z. As an approximation of its influence when moving a small mass ϵ from z to z_δ, we have:

$$\hat{\theta}_{\epsilon, z_\delta, -z} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \tilde{C}(z_i, \theta) + \epsilon \tilde{C}(z_\delta, \theta) - \epsilon \tilde{C}(z, \theta) \tag{13}$$

The influence functions introduced in [12] provide an efficient approximation for the computation of upweighting z as: $\mathcal{I}_{up.params}(z) = \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\Big|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_\theta \tilde{C}(z, \hat{\theta})$. Then, the parameter change of perturbing the label y should be given as the difference of the influences between upweighting z_δ and upweighting z:

$$\begin{aligned} \frac{d\hat{\theta}_{\epsilon, z_\delta, -z}}{d\epsilon}\Big|_{\epsilon=0} &= \mathcal{I}_{up,params}(z_\delta) - \mathcal{I}_{up,params}(z) \\ &= -H_{\hat{\theta}}^{-1} \Big(\nabla_\theta \tilde{C}(z_\delta, \hat{\theta}) - \nabla_\theta \tilde{C}(z, \hat{\theta}) \Big) \end{aligned} \tag{14}$$

where the Hessian is given as $H_{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta^2 \tilde{C}(z_i, \hat{\theta})$. If $\tilde{C}$ is in differentiable in θ and y. Also, the soft label y is continuous and the amount of perturbation δ is small, the above equation can be approximated using the first-order derivative approximation by treating $\nabla_\theta \tilde{C}(z, \hat{\theta})$ as a function of y. Then, we obtain:

$$\frac{d\hat{\theta}_{\epsilon, z_\delta, -z}}{d\epsilon}\Big|_{\epsilon=0} \approx -H_{\hat{\theta}}^{-1} \nabla_y \nabla_\theta \tilde{C}(z, \hat{\theta}) \delta \tag{15}$$

If the upweighting mass is set as $\epsilon = \frac{1}{n}$, we can have the linear approximation: $\hat{\theta}_{z_\delta, -z} - \hat{\theta} \approx -\frac{1}{n} H_{\hat{\theta}}^{-1} \nabla_y \nabla_\theta \tilde{C}(z, \hat{\theta}) \delta$. Next, we apply the chain rule[4] to measure the influence of perturbing y on a test item:

$$\begin{aligned} \mathcal{I}_{pert,loss}(z, z_{test}) &= \nabla_\delta \tilde{C}(z_{test}, \hat{\theta}_{z_\delta, -z})\Big|_{\delta=0} \\ &= -\frac{1}{n} \nabla_\theta \tilde{C}(z_{test}, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_y \nabla_\theta \tilde{C}(z, \hat{\theta}) \end{aligned} \tag{16}$$

Example 3.2 **Influence calculation for surrogate logistic cost.** Considering the binary classification problem on a logistic regression model where $\sigma(t) = \frac{1}{1+\exp(-t)}$ is the sigmoid function. The logistic cost of any item $z = (\boldsymbol{x}, y)$ is given as $C(z, \theta) = \log\big(1 + \exp(-y\theta^T \boldsymbol{x})\big)$ and the surrogate cost $\tilde{C}(z, \theta)$ is given in Eqn. (6). The

[4] It is implicitly assumed that $\frac{\partial \tilde{C}(z_{test}, \hat{\theta}_{z_\delta, -z})}{\partial \hat{\theta}_{z_\delta, -z}} = \frac{\partial \tilde{C}(z_{test}, \hat{\theta})}{\partial \hat{\theta}}$ is satisfied.

gradient of the surrogate logistic cost $\nabla\tilde{C}(z,\theta)$ w.r.t θ respectively are given as:

$$\nabla_\theta \tilde{C}(z,\theta) = \frac{\Big[(\rho_{-y}-1)\sigma(-y\theta^T\boldsymbol{x}) - \rho_y\sigma(-y\theta^T\boldsymbol{x})\Big]y\boldsymbol{x}}{1-\rho_{+1}-\rho_{-1}} \tag{17}$$

Following the above derivation, the gradient w.r.t the variables y and θ is given as:

$$\nabla_y\nabla_\theta \tilde{C}(z,\theta) = \frac{(\rho_{-y}-1)\sigma(-y\theta^T\boldsymbol{x})\Big[1-(\theta^T\boldsymbol{x})\exp(y\theta^T\boldsymbol{x})\sigma(-y\theta^T\boldsymbol{x})y\Big]\boldsymbol{x}}{1-\rho_{+1}-\rho_{-1}} - \frac{\rho_y\sigma(y\theta^T\boldsymbol{x})\Big[1+(\theta^T\boldsymbol{x})\exp(-y\theta^T\boldsymbol{x})\sigma(y\theta^T\boldsymbol{x})y\Big]\boldsymbol{x}}{1-\rho_{+1}-\rho_{-1}} \tag{18}$$

Interestingly, the Hessian of the surrogate logistic cost w.r.t θ is the same as the Hessian of the logistic cost, it is given as:

$$\nabla^2_\theta \tilde{C}(z,\theta) = y^2\cdot\sigma(y\theta^T\boldsymbol{x})\cdot\sigma(-y\theta^T\boldsymbol{x})\cdot\boldsymbol{x}\boldsymbol{x}^T = \nabla^2_\theta C(z,\theta) \tag{19}$$

Overall, the label perturbation influence on any item z_{test} for the unbiased surrogate logistic cost can be computed by substituting Eqn. (17) - (19) into Eqn. (16). ▲

The influence of label perturbation could be either positive (perturb labels from −1 to +1) or negative (perturb labels from +1 to −1). Therefore, we compute the influence score of any item z_{test} using its absolute influence value on each labeled item as:

$$p_{test} = \sum_{i=1}^{n} \left|\mathcal{I}_{pert,loss}(z_i, z_{test})\right| \tag{20}$$

4 ALGORITHM

Based on this overall objective (Subsection 3.3 and 3.5) which considers both the learner's learning progress as well as the item's influence, we propose the adaptive interactive teaching algorithm **VADER** (**V**isually Expl**A**inable A**D**aptive T**E**aching with Human Learne**R**). The VADER teaching algorithm is shown in Algorithm 1. The given input includes the labeled and the unlabeled data set, the teacher's prediction model, the learner's memory decay rate, the empirical target concept, the learning rate, and the influence intensity. The algorithm will output the updated label set of all items. In each worker's teaching session, VADER works as follows. We first initialize the teaching iterator $t = 1$, teaching momentum $\boldsymbol{v}_0 = \mathbf{0}$ and the initial influence scores. Then, in each teaching iteration, the VADER teacher estimates the learner's progress and computes all items' teaching scores based on their teaching usefulness and teaching diversity. Next, the teaching score is combined with the complementary influence score weighted by the influence intensity and the item with the maximum combined score would be recommended to show to the learner. Finally, the interactions introduced in Subsection 3.1 will be performed and the labels provided by the learner **will only be added to the current label set if the learner is confident.** The confident gauging has theoretical guarantee from Theorem 4.1 and it is reflected on the teacher-learner's third interaction step. More details are shown in Figure 6in the Supplementary Material.

Algorithm 1 VADER (one session)

1: **Input:** Imperfect labeled data set $\mathcal{D}_L$, unlabeled data set $\mathcal{D}_U$, prediction model ψ, learner's memory decay rate β, empirical target concept $\hat{\theta}$, initial learning rate η_0, influence intensity ξ, MaxIter.
2: **Initialization:**
3: Set $t = 1$ and $\boldsymbol{v}_0 = \mathbf{0}$. Compute the initial influence scores $p_1, p_2, \ldots, p_N$ using Eqn. (20)
4: **Repeat:**
5: (i). Compute the teaching scores $s_1, s_2, \ldots, s_N$ of all items in $\mathcal{D}_L \cup \mathcal{D}_U$ using the objective in Prob. (11).
6: (ii). Combine the teaching score with the complementary influence scores. The recommended teaching item has an index of I_t:
$$I_t = \arg\min_{i\in N} \; s_i + \xi_i p_i$$
7: (iii). The teacher and learner perform the first two interactions.
8: (iv). Based on the learner's selections in the second interaction, the teacher shows the masked explanation and the learner provides the confidence feedback. If the learner is confident, add his/her label to the teaching item's label set.
9: (v). $t \leftarrow t + 1$
10: **Until** $t >$ MaxIter
11: **Output:** The updated label set $L_{x_1}, L_{x_2}, \ldots, L_{x_N}$

4.1 Upper Bound

In this paper, the class-conditional random error ρ_{+1} and ρ_{-1} exist within the initial imperfect labeled set $\mathcal{D}_L$. Furthermore, we assume that in each worker's teaching session, the teacher works with a single learner, whose initial classification performance is not as good as the teacher. In other words, if $\bar{y}$ is the label provided by the worker, the worker will also has class-conditional random error as:

$$P(\bar{y}=-1|y_{gt}=+1) = \rho'_{+1}, P(\bar{y}=+1|y_{gt}=-1) = \rho'_{-1}$$
$$\rho'_{+1} + \rho'_{-1} > \rho_{+1} + \rho_{-1}$$

As shown earlier, the surrogate cost $\tilde{C}\big(\psi(\boldsymbol{x};\theta), y\big)$ defined in Eqn. (6) is an unbiased estimator of $C\big(\psi(\boldsymbol{x};\theta), y_{gt}\big)$ and the empirical target concept $\hat{\theta}$ is learned using the surrogate cost. Furthermore, it is proven in [25], the generalization performance of the surrogate cost has the following bound with probability at least $1-\varepsilon$,

$$R_C(\hat{\theta}) \le R_C(\theta_*) + \frac{8K}{1-\rho_{+1}-\rho_{-1}}\mathcal{R}(\Psi) + 2\sqrt{\frac{\log(1/\varepsilon)}{2n}} \tag{21}$$

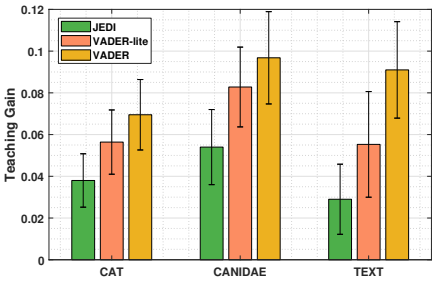
where K is the Lipschitz constant of $C(\cdot,\cdot)$ and $\mathcal{R}(\Psi)$ is the Rademacher complexity of the prediction function class Ψ. The risk $R_C(\hat{\theta}) := \mathbb{E}[C\big(\psi(\boldsymbol{x};\hat{\theta}), y\big)]$ is defined as the expectation over the unknown ground truth distribution.

In the proposed interactive process, we assume that the labels provided by the learner will only be added to the current label set if the learner is confident. Furthermore, the probability of the learner being confident depends on the true class label as follows:

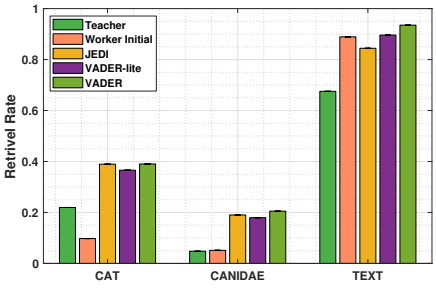
$$P(Conf|y_{gt}=+1) = c_{+1}, P(Conf|y_{gt}=-1) = c_{-1}$$

The following theorem shows the condition under which the upper bound in Eqn. (21) can be improved.

(a) The average teaching gain of the crowd workers.



(b) The label retrieval rate of the teacher and the crowd workers.



(c) The prediction model performance comparison.

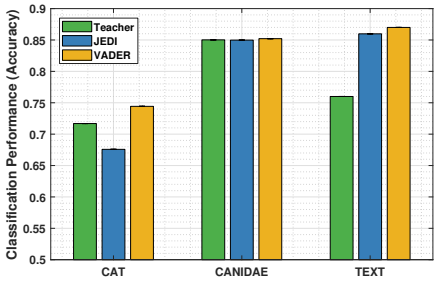


Figure 3: Experimental results w.r.t. various evaluation metrics. Bar plots in (a) has standard deviations among group of learners per baseline per data set. Bar plots in (b), (c) are evaluated with aggregated labels, therefore, no standard deviations in them.

Theorem 4.1. *If the learner satisfies the following condition:*

$$c_{+1}(\rho'_{+1} - \rho_{+1}) + c_{-1}(\rho'_{-1} - \rho_{-1}) < 0 \tag{22}$$

then the upper bound shown in Eqn. (21) is reduced after taking into consideration the confident labels provided by the learner.

Proof. Notice that the first and the third terms on the right hand side of Eqn. (21) remain unchanged before and after taking into consideration the confident labels provided by the learner. As the complexity of the function class Ψ is fixed, the middle term is negatively correlated to the change in terms of the sum over the class conditional label error ρ_{+1} and ρ_{-1}. It can be easily verified by solving the inequality below and show that when Eqn. (22) is satisfied, $\rho_{+1} + \rho_{-1}$ is reduced after taking into consideration the confident labels provided by the learner.

$$(1-c_{+1})\rho_{+1} + c_{+1}\rho'_{+1} + (1-c_{-1})\rho_{-1} + c_{-1}\rho'_{-1} - \rho_{+1} - \rho_{-1} < 0 \tag{23}$$

□

This theorem shows that when the learners are confident about their labels, the error rate of the labeled data set will decrease after including these confidence labels. Followed with it, the risk under the unknown ground truth distribution using empirical target concept $\hat{\theta}$ will gradually have smaller upper bound and getting close to the risk using true target concept θ_*. Therefore, the empirical target concept $\hat{\theta}$ will gradually become a reliable substitute of the true target concept θ_* with more and more teaching iterations.

4.2 Discussion and Extensions

Case #1: Worker only has confidence on one category. Assume that there is a worker who is only confident about the positive class such that ρ'_{+1} is smaller than ρ_{+1}, and she only provides the confident positive labels. In this case, $c_{+1} > 0$ and $c_{-1} = 0$, and the condition in Eqn. (22) is satisfied. Therefore, the updated labeled set can still lead to a better prediction model.

Case #2: Teaching with starving prevention. When the repeated labeling is allowed, the overall teaching score could be high for certain items in order to favor the teaching objective. Some low-score items could be starved and never be recommended. Then, in each teaching session, the influence intensity ξ_i of i-th item could be adaptively updated as the entropy of its label set L_{x_i}. Intuitively, the low-entropy label set (e.g., $L_{x_i} = \{+1, +1, +1, +1, -1\}$ of five confident labels) will downgrade the influence score faster with smaller ξ_i because the label aggregation (e.g., majority voting) for this item is no longer debatable. The score of the high-entropy label set (e.g., $L_{x_i} = \{+1, -1, +1, -1\}$ of four confident labels) will be downgraded slower with larger ξ_i so that this item could still be recommenced to break the tie in the later teaching iterations.

Table 2: Statistics of three data sets.

Data set	# Items($\mathcal{D}_L$)	# Items($\mathcal{D}_U$)	# Features	# Workers	Error rate
CAT	220	219	512	21	0.15
CANIDAE	257	257	512	19	0.15
TEXT	300	200	120	21	0.25

5 EXPERIMENTS

5.1 Details of the Data Sets

We conduct experiments on three real data sets which include two image data sets [47] (classify domestic/wild animals) and one text data set which belongs to a subset of 20 Newsgroup data sets and has subject categories: *comp.os.ms-windows.misc* and *sci.crypt.* [27] (classify encryption/operating-system documents). The details of these real-world data sets are provided in Table 2. Regarding the feature extractions of images, we first fine-tune and transfer the ResNet34 [8] model into our binary teaching scenario on the imperfect data set $\mathcal{D}_L$. Next, the features are extracted from the penultimate layer before the average pooling layer. The explanations for images are the saliency maps generated by Grad-CAM [30] on the finetuned ResNet34 architecture. Regarding the feature extractions of text, we remove the stopwords, the footers, and the quotes in the documents to prevent overfitting on irrelevant metadata. Then, the top TF-IDF features are extracted. The explanations of the text are the additive features extracted using LIME [29] with ten interpretable simplified features. To inject error into the labeled data set, we randomly flipped the labels of the labeled set with an error rate of $\rho_{+1} = \rho_{-1} = 0.15$ for images data set and $\rho_{+1} = \rho_{-1} = 0.25$ for text data set. As for the teacher, the empirical target concept is obtained by minimizing the unbiased surrogate logistic cost. We have hired 61 graduate student workers to perform learning and each worker is assigned with one exclusive teaching algorithm using Round-robin scheduling. The worker's learning loss is also set to the logistic loss. However, the worker's learning concept is not observable in practice. Following the convention [47], we estimate the worker's learning progress using the harmonic function [49].

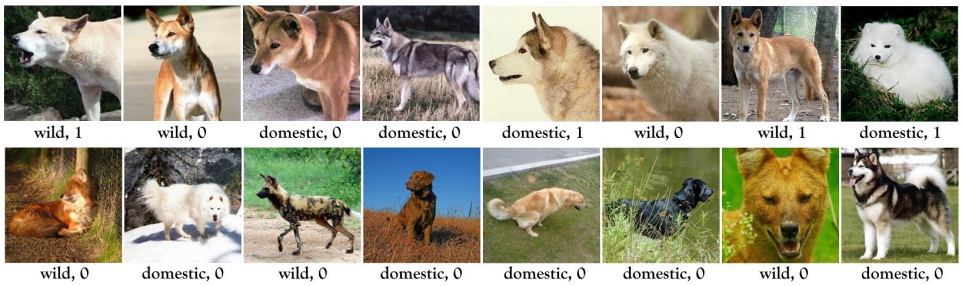


Figure 4: Visualization of the top eight high influence images (first row) and bottom eight low influence images (second row) on canidae. Each image is described by a tuple composed of its true category and indicator of error (1 means flipped label, 0 mean no error on its label)

The experiments were conducted remotely through a web interface. The teaching procedure is nearly real-time. But the model training and predictions can be computationally intensive, and therefore are only performed twice, before and after the teaching.

5.2 Quantitative Results

In order to evaluate the effectiveness and reliability of interactive teaching we have utilized three metrics: (1) **Teaching gain** is designed to evaluate the overall teaching performance of all workers. (2) **Retrieval rate** is used to evaluate the reliability of teaching on these items with incorrect labels. (3) **Model performance comparison** aims to evaluate the model performance before and after teaching.

5.2.1 Teaching Gain of the Workers. The purpose of teaching is to help the human workers learn how to improve labeling. In order to evaluate and confirm that these workers have made progress towards their annotations tasks, we propose to use the *Teaching Gain*, which is defined as the labeling accuracy after teaching minus the labeling accuracy before teaching. In the third step of interactions, the confidence labels are also recorded, we also compute the teaching gain that only takes the confidence labels into consideration. As a comparison, JEDI [47] is an interactive teaching framework without any explanation, VADER-lite removes the confidence gauging (i.e., JEDI with explanation) and it is a simplified version of VADER. As shown in Figure 3a, the teaching gains of learners with explanation outperform the baseline learners significantly, which shows that explainable teaching is more effective. We also observe that the confidence gauging further improves the performance of these VADER learners on all data sets.

5.2.2 Label Retrieval Rate. We also value the reliability of teaching in terms of the *Retrieval Rate*, which is defined as the fraction of items with incorrect labels that have been corrected. As a comparison, we use the empirical teacher $\hat{\theta}$ as the baseline and compare it with the initial crowd labels, the updated crowd labels, and the confident crowd labels. The label aggregation is performed by first using the minimax conditional entropy approach [44] to estimate the worker expertise and then take the weighted average of their labels. The results are shown in Figure 3b, we observe that the cat and canidae data sets are relatively difficult compared with the text data set because the former two have a lower retrieval rate. We also conclude that fixing the mislabeled items using the taught workers is more reliable using the original teacher. The retrieval rate results in Figure 3b show that VADER is better than or comparable to the state-of-the-art teaching models. As the analysis shown in Section 4.1, explanation is surely helpful when the confidence gauging is performed. Theorem 4.1 guarantees that if the initial labeling abilities of the workers are lower than the teacher, the proposed teaching model could guarantee their improvement after teaching. If this assumption is not satisfied, Figure 3b shows empirically these workers could still eventually benefit from the teaching.

5.2.3 Model Performance Comparison. Using the aggregated crowd labels, we retrained the prediction model and compared the performance of the retrained model with the teacher's performance. As shown in Figure 3c, comparing with the teacher, the performances of the workers regarding cat images and text have a clear improvement in terms of accuracy. The performance difference on canidae images is not very obvious. The reason is that the retrieval rate of the items with incorrect labels in canidae is very low. Even among the confident labels, most of these high influence items that have incorrect labels did not get fixed. Therefore, the prediction models of the teacher and the workers are about the same on canidae. Figure 3a shows the self-improvement of these learners before and after teaching. However, the model performance, shown in Figure 3c, is mostly influenced by the difficult items, e.g., the items near the decision boundary. Learners with a large teaching gain could improve their labeling abilities on the easy items, which do not necessarily have an impact on the model predictions.

5.3 Qualitative Results

We qualitatively checked the results of the influence scores. The top eight high influence canidae images and bottom eight low influence canidae images are shown in Figure. 4. The highly influenced images are actually the ones with flipped labels and those of easily confused breeds (e.g., Dingo and Akita, Samoyed and Arctic Fox, etc.). The lowest influenced images did not have a trend on selecting certain types of images and the label flipped images are rare in them. It is straightforward to know that perturbing the labels of these highly influenced images would have a large impact on the prediction model. Teaching the workers should follow the curriculum principle by recommending items from low to high influence.

We also compare influence and marginal distance of the items for all three data sets. The marginal distance of an item is its geometrical distance from the discriminatory boundary and it has widely used in standard active learning strategies like uncertainty sampling and expected error reduction. We hypothesize that those items near to discriminatory boundary are crucial to classification,

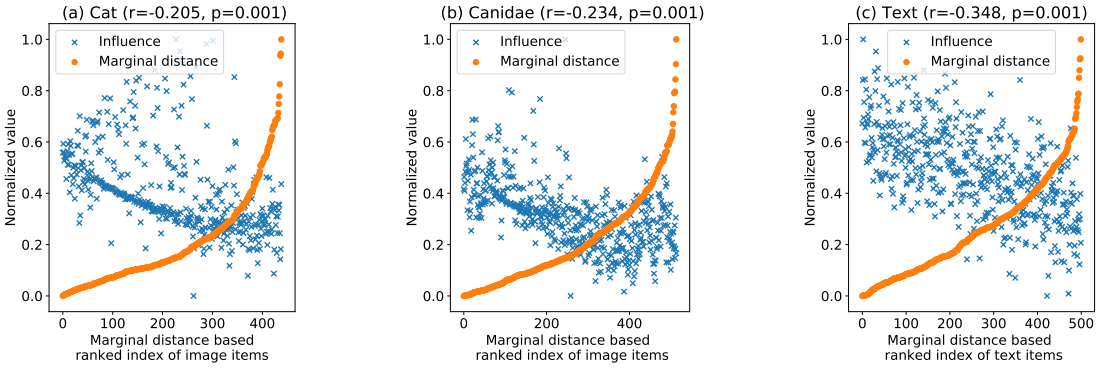


Figure 5: Influence vs. marginal distance on three data sets: (a) Cat, (b) Canidae, and (c) Text. Pearson correlation coefficient r and its significance value p are reported. Y-axis represents normalized value for influence and marginal distance, normalized to the scale $[0, 1]$.

there by given high values of influence. Pearson correlation coefficient between marginal distance and influence for all the examples supported our hypothesis. The plots for Cat, Canidae, and Text data sets are shown in Figure 5. For each plot, r and p values represent correlation coefficient and significance respectively. From the figures, it can be observed that, those items that are close to boundary are commonly given high influence values. However, with the help of the visualization results from Figure 4, we also observe that unlike marginal distance, the influence computing scheme used in the proposed VADER framework also takes difficulty of the labeling the image into consideration. These visually hard-to-distinguish items usually have high influences and this observation matches with the principle of curriculum learning.

6 RELATED WORK

6.1 Explanation Models

Understanding why a model makes certain predictions is as important as the model performance on many occasions [1, 7, 13, 14, 43], especially in high-stakes decision-making applications. Starting from deconvolution [26] and guided backpropogation [34], multiple efforts have been made to explain deep models. Inspired by the global average pooling architecture [18], one of the most popular and effective explanation models for images is class activating mapping (CAM) [30, 42]. In the original CAM model, the activation weights need to be learned as part of the architecture. However, the Grad-CAM generalizes CAM by enabling these weights to be learned using the gradients on the activation maps without retraining the model. Another branch of explanation models is additive feature based models, which focus on using simplified features to approximate the original model. LIME [29] locally explains the model using perturbed examples. DeepLIFT [32] decomposes the output on a simple input by backpropagating contributions to every input feature. Shapley value estimation [21] assigns each feature an importance weight for a particular prediction by guaranteeing local accuracy, missingness, and consistency. However, none of these approaches provide visual explanations to guide the learning procedure of crowd workers with theoretical connections between worker and learners.

6.2 Crowd Teaching

In the context of crowdsourcing [22, 23, 28, 37, 44, 46, 48], crowd teaching is a sub-area of machine teaching[50] where the learners are crowd workers and the teacher is the machine that guides the learners towards a specific labeling concept. It supervises the labeling process of crowdsourced workers in the form of teaching in order to improve the workers' annotation expertise and collect data sets with higher label quality. Previous research [3, 47] has shown that non-experts can be trained to perform accurate and complex tasks. Based on the teaching styles, the methodology of performing crowd teaching has two branches. The first branch of approaches [4, 24, 33] treats the workers as global learners who learn things in large jumps holistically. The hypothesis transition model STRICT [33] assumes the worker's concept are randomly switched in the pre-given hypothesis space which is computed on observed workers' feedback. The model in [24] extends STRICT by considering both the item explanation and modeling representativeness. Another branch of crowd teaching [19, 38, 40, 47] treats the workers as sequential learners who learn concepts in continuous steps. Starting from the iterative machine teaching (IMT) [19], multiple efforts have been made in this direction. JEDI [47] assumes learners have forgetting behavior and it extends IMT by guaranteeing the teaching usefulness and teaching diversity. Instead of teaching a single learner at a time, the model in [40] extends IMT to the scenario of classroom teaching by teaching a diverse group of learners. In [9], the authors have proposed a greedy approach to teach a forgetful learner multiple learning concepts by assuming every single concept decays over time. Our work belongs to the general framework of sequential teaching with adaptive learning rate and confidence-gauged label feedback. However, global teaching requires a heuristic pre-defined hypothesis space and sequential teaching requires the unknown target concept beforehand. As a comparison, the unbiased empirical risk minimizer used in this paper is a reliable and realistic substitute of the optimum target concept with a bounded performance guarantee.

6.3 Learning with imperfect Labels and Imperfect Labelers

Learning from imperfect labels is very useful in many applications [17, 31, 36, 39, 41], as a large number of imperfect labels are relatively easy to collect using crowdsourcing. In the traditional setting

Figure 6: Interactions between the teacher and the worker on one canidae image.

of supervised learning, these imperfect labels are usually treated as outliers or label flips. However, imperfect items could be beneficial for learning because they have a certain level of significant mass. The work of [25] proposes to modify the loss term with pregiven class-conditional error and ends up with a error-tolerable learning model. A system of learning from imperfect labels by leveraging a knowledge graph has been used in [15]. The modified deep model [35] could also be adapted to learn using flipped label error and outlier error by introducing an extra error layer into the network. From another perspective, researchers also devote effort on learning with imperfect labelers. Existing work [31, 39] addresses the repeated acquisition of labels from multiple imperfect labelers for every single item in active learning. The re-active learning model [17] extends the concept of active learning with crowdsourcing by allowing the labeled item being re-labeled using impact sampling. The investigation conducted by [16] further pointed out that item re-labeling would be helpful when the learning problems have high model expressiveness. For all that, impact sampling requires multiple model retraining which is computationally prohibitive and uncertainty sampling could easily starve items [17] and has difficulty handling item re-labeling. Our model estimates each item's influence on the prediction model using label perturbations without model retraining and yet has the ability to either perform re-labeling or assign new labels to the unlabeled items.

7 CONCLUSION

In this paper, motivated by the huge demand for fine-grained label information from real applications, we propose a novel framework for interactive teaching and learning between the teacher and the crowd workers. It utilizes the empirical minimizer as the target concept of the teacher, and instructs the learners to focus on the informative visual cues during learning. This framework benefits both the teacher and the workers in terms of the performance of the predictive model for the teacher and workers' expertise. Compared with state-of-the-art techniques, this framework takes one step further towards the real-world crowd teaching with explanations, as it is designed for imperfect teacher and workers. Furthermore, our analysis of the proposed framework is verified by the experiments on various data sets.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation under Grant No. IIS-1552654 and Grant No. IIS-1813464, the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-02-00 and Ordering Agreement Number HSHQDC-16-A-B0001, and an IBM Faculty Award. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

SUPPLEMENTARY MATERIAL

A. Interactions on Images and Texts

The three-step interactions between the teacher and the worker for an image item are shown in Figure 6. In Step-1, the teacher will recommend one image and ask for the initial label from the worker. In Step-2, the teacher will demonstrate its probabilistic soft labels regarding this image as well as the visual explanations (saliency map for images). Then, the worker could provide the updated class label and the preferred visual explanations. For instance, let's say the worker has chosen "Wild" and "Right" for the given image. At last, the third step is designed to collect the confidence information regarding his/her choices made in the second step. The masked image will be generated by applying a thresholded saliency map on the original input image. The threshold would be the teacher's probability 0.382 and any pixels with its grayscale value higher than $\lfloor 255 \times 0.382 \rfloor = 97$ will remain visible. If the worker selects "Yes" in Step-3, then we assume s/he is confident regarding the updated label in step-2 and this label will be added to the item's label set. One important thing should be noticed is that when the teacher's confidences regarding two classes are approximately equal (~ 0.5), the masked explanation for confidence evaluation will be almost the same no matter which class label the worker chooses.

B. Reproducibility

To better reproduce the empirical analysis presented in the paper, we provide additional implementation details on the proposed VADER algorithm. The implementation and data sets will be released upon acceptance. All three data sets are randomly split to have 50% as $\mathcal{D}_L$ - examples with labels and 50% as $\mathcal{D}_U$ - examples without labels. In order to inject error into $\mathcal{D}_L$, we randomly flip labels with an error rate of $\rho_{+1} = \rho_{-1} = 0.15$ for image data sets and $\rho_{+1} = \rho_{-1} = 0.25$ for text data set. The initial learning rate of

the worker is set to $\eta_0 = 0.02$ and it will be gradually decreased as $\eta_t = \frac{20}{20+t}\eta_0$. The memory decay rate β of the workers is not available to the teacher beforehand and we use the image sequence sorting task to estimate each worker's β as $\beta = 1 - \frac{1}{\bar{n}}$ where $\bar{n}$ represents for their mean of maximum number of ordered images they can recover in the image sorting game designed in [47]. Empirically, β is assigned with one of the values of $\{0.75, 0.833, 0.875\}$ where, correspondingly, $\bar{n}$ has been discretized into three memory window ranges $[3, 5], (5, 7], (7, 9]$. Based on our observations, all participants have $3 \leq \bar{n} \leq 9$, then learners' decay rate is grouped into these three values (e.g., $\beta = 1 - \frac{1}{4} = 0.75$, $\beta = 1 - \frac{1}{6} = 0.833$, or $\beta = 1 - \frac{1}{8} = 0.875$). Regarding the implementation of estimating worker's concept, we use the harmonic function [49]. The kernel weight between the i-th and the j-th items denoted by $\boldsymbol{x}_i, \boldsymbol{x}_j$ respectively, is calculated as:

$$\omega_{ij} = \begin{cases} \exp\Big(- \sum_{d=1} \frac{(\boldsymbol{x}_{id} - \boldsymbol{x}_{jd})^2}{\sigma_d^2} \Big), & \Leftrightarrow \text{Image data set} \\ \exp\Big(- \frac{1}{0.03} \Big(1 - \frac{\boldsymbol{x}_i^\top \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|} \Big) \Big), & \Leftrightarrow \text{Text data set} \end{cases}$$

where d is the index for features and σ_d is the sample variance on d-th feature dimension. To deal with the "cold-start" problem, we ad-hocly set the first 20 teaching items as the bottom low influence ones without combining the teaching scores. Therefore, the worker will be taught with easier items as the start and at the same time, the teacher could get stable estimations of the worker's learning progress using these 20 items.

REFERENCES

[1] Gagan Bansal. 2018. Explanatory Dialogs: Towards Actionable, Interactive Explanations. In *AAAI/ACM Conference on AI, Ethics, and Society, AIES*. 356–357.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML*. 41–48.

[3] Nancy Chang, Praveen Paritosh, David Huynh, and Collin Baker. 2015. Scaling Semantic Frame Annotation. In *Proceedings of The 9th Linguistic Annotation Workshop, LAW@NAACL-HLT*. 1–10.

[4] Yuxin Chen, Oisin Mac Aodha, Shihan Su, Pietro Perona, and Yisong Yue. 2018. Near-Optimal Machine Teaching via Explanatory Teaching Sets. In *International Conference on Artificial Intelligence and Statistics, AISTATS*. 1970–1978.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*. 248–255.

[6] Jia Deng, Jonathan Krause, Michael Stark, and Fei-Fei Li. 2016. Leveraging the Wisdom of the Crowd for Fine-Grained Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI* 38, 4 (2016), 666–676.

[7] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv e-prints* (Feb 2017). arXiv:1702.08608

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. 770–778.

[9] Anette Hunziker, Yuxin Chen, Oisin Mac Aodha, Manuel Gomez-Rodriguez, Andreas Krause, Pietro Perona, Yisong Yue, and Adish Singla. 2018. Teaching Multiple Concepts to Forgetful Learners. *CoRR* abs/1805.08322 (2018).

[10] Pengfei Jiang, Weina Wang, Yao Zhou, Jingrui He, and Lei Ying. 2018. A Winners-Take-All Incentive Mechanism for Crowd-Powered Systems. In *Proceedings of the 13th Workshop on Economics of Networks, Systems and Computation, NetEcon@SIGMETRICS*. 3:1–3:6.

[11] Angelos Katharopoulos and François Fleuret. 2018. Not All Samples Are Created Equal: Deep Learning with Importance Sampling. In *Proceedings of the 35th International Conference on Machine Learning, ICML*. 2530–2539.

[12] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML*. 1885–1894.

[13] Todd Kulesza, Margaret M. Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI*. 126–137.

[14] Isaac Lage, Andrew Slavin Ross, Samuel J. Gershman, Been Kim, and Finale Doshi-Velez. 2018. Human-in-the-Loop Interpretability Prior. In *Advances in Neural Information Processing Systems, NeurIPS*. 10180–10189.

[15] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. 2017. Learning from Noisy Labels with Distillation. In *IEEE International Conference on Computer Vision, ICCV*. 1928–1936.

[16] Christopher H. Lin, Mausam, and Daniel S. Weld. 2014. To Re(label), or Not To Re(label). In *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing, HCOMP*.

[17] Christopher H. Lin, Mausam, and Daniel S. Weld. 2016. Re-Active Learning: Active Learning with Relabeling. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI*. 1845–1852.

[18] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network In Network. *CoRR* abs/1312.4400 (2013).

[19] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B. Smith, James M. Rehg, and Le Song. 2017. Iterative Machine Teaching. In *Proceedings of the 34th International Conference on Machine Learning, ICML*. 2149–2158.

[20] Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James M. Rehg, and Le Song. 2018. Towards Black-box Iterative Machine Teaching. In *Proceedings of the 35th International Conference on Machine Learning, ICML*. 3147–3155.

[21] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems, NeurIPS*. 4768–4777.

[22] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 745–754.

[23] Fenglong Ma, Chuishi Meng, Houping Xiao, Qi Li, Jing Gao, Lu Su, and Aidong Zhang. 2017. Unsupervised Discovery of Drug Side-Effects from Heterogeneous Data Sources. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 967–976.

[24] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. 2018. Teaching Categories to Human Learners With Visual Explanations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. 3820–3828.

[25] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. 2013. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems, NeurIPS*. 1196–1204.

[26] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. 2015. Learning Deconvolution Network for Semantic Segmentation. In *IEEE International Conference on Computer Vision, ICCV*. 1520–1528.

[27] F. Pedregosa and et al Varoquaux, G. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, JMLR* 12 (2011), 2825–2830.

[28] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning From Crowds. *Journal of Machine Learning Research, JMLR* 11 (2010), 1297–1322.

[29] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*. 1135–1144.

[30] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision, ICCV*. 618–626.

[31] Victor S. Sheng, Foster J. Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*. 614–622.

[32] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning, ICML*. 3145–3153.

[33] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. 2014. Near-Optimally Teaching the Crowd to Classify. In *Proceedings of the 31th International Conference on Machine Learning, ICML*. 154–162.

[34] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2014. Striving for Simplicity: The All Convolutional Net. *CoRR* abs/1412.6806 (2014).

[35] Sainbayar Sukhbaatar and Rob Fergus. 2014. Learning from Noisy Labels with Deep Neural Networks. *CoRR* abs/1406.2080 (2014).

[36] Stefano Teso and Kristian Kersting. 2019. Explanatory Interactive Machine Learning. In *AAAI/ACM Conference on AI, Ethics, and Society, AIES*. 356–357.

[37] Peter Welinder, Steve Branson, Serge J. Belongie, and Pietro Perona. 2010. The Multidimensional Wisdom of Crowds. In *Advances in Neural Information Processing Systems, NeurIPS*. 2424–2432.

[38] Lijun Wu, Fei Tian, Yingce Xia, Yang Fan, Tao Qin, Jian-Huang Lai, and Tie-Yan Liu. 2018. Learning to Teach with Dynamic Loss Functions. In *Advances in Neural Information Processing Systems NeurIPS*. 6467–6478.

[39] Yan Yan, Rómer Rosales, Glenn Fung, and Jennifer G. Dy. 2011. Active Learning from Crowds. In *Proceedings of the 28th International Conference on Machine*

Learning, ICML. 1161–1168.

[40] Teresa Yeo, Parameswaran Kamalaruban, Adish Singla, Arpit Merchant, Thibault Asselborn, Louis Faucon, Pierre Dillenbourg, and Volkan Cevher. 2018. Iterative Classroom Teaching. *CoRR* abs/1811.03537 (2018).

[41] Ping Zhang and Zoran Obradovic. 2011. Learning from Inconsistent and Unreliable Annotators by a Gaussian Mixture Model and Bayesian Information Criterion. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*. 553–568.

[42] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. 2921–2929.

[43] Dawei Zhou, Zhining Liu, and Jingrui He. 2019. Towards Explainable Representation of Time-Evolving Graphs via Spatial-Temporal Graph Attention Networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM*. 2137–2140.

[44] Dengyong Zhou, John C. Platt, Sumit Basu, and Yi Mao. 2012. Learning from the Wisdom of Crowds by Minimax Entropy. In *Advances in Neural Information Processing, NeurIPS*. 2204–2212.

[45] Yao Zhou and Jingrui He. 2016. Crowdsourcing via Tensor Augmentation and Completion. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI*. 2435–2441.

[46] Yao Zhou and Jingrui He. 2017. A Randomized Approach for Crowdsourcing in the Presence of Multiple Views. In *2017 IEEE International Conference on Data Mining, ICDM*. 685–694.

[47] Yao Zhou, Arun Reddy Nelakurthi, and Jingrui He. 2018. Unlearn What You Have Learned: Adaptive Crowd Teaching with Exponentially Decayed Memory Learners. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD*. 2817–2826.

[48] Yao Zhou, Lei Ying, and Jingrui He. 2019. Multi-task Crowdsourcing via an Optimization Framework. *TKDD* 13, 3 (2019), 27:1–27:26.

[49] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. 2003. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *Proceedings of the 20th International Conference on Machine Learning, ICML*. 912–919.

[50] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. 2018. An Overview of Machine Teaching. *CoRR* abs/1801.05927 (2018).