Learning Quality Labels for Robust Image Classification

Xiaosong Wang, Ziyue Xu, Dong Yang, Leo Tam, Holger Roth, Daguang Xu
Nvidia Corporation, CA, US
{xiaosongw, ziyuex, dongy, leot, hroth, daguang}@nvidia.com

Abstract

Supervised learning paradigms largely benefit from the tremendous amount of annotated data. However, the quality of the annotations varies among labelers. Multi-observer studies have been conducted to examine the annotation variances (by labeling the same data multiple times) to see how it affects critical applications like medical image analysis. In this paper, we demonstrate how multiple sets of annotations (either hand-labeled or algorithm-generated) can be utilized together and mutually benefit the learning of classification tasks. A scheme of learning-to-vote is introduced to sample quality label sets for each data entry on-the-fly during the training. Specifically, a label-sampling module is designed to achieve refined labels (weighted sum of attended ones) that benefit the model learning the most through additional back-propagations. We apply the learning-to-vote scheme on the classification task of a synthetic noisy CIFAR-10 to prove the concept and then demonstrate superior results (3-5% increase on average in multiple disease classification AUCs) on the chest x-ray images from a hospital-scale dataset (MIMIC-CXR) and hand-labeled dataset (OpenI) in comparison to regular training paradigms.

1. Introduction

Supervised deep learning methods, although proven to be effective on many tasks, rely heavily on the quality of the data and its corresponding annotations. Some tasks enjoy almost error-free annotation, such as handwritten numbers and simple natural images. However, for other applications, e.g., most medical image analysis tasks, the inherent ambiguity of the task leads to unavoidable noise and fuzziness within the annotations themselves, no matter how experienced the expert labelers are. Meanwhile, under a multi-labeler setting for quality control purposes, the significant intra- and inter-observer variability inject even more uncertainties into the resulting labels. Beyond the above challenges, for the specific task of chest X-ray image classification, due to the fact that most labels of the available

large-scale open datasets are automatically mined by Natural Language Processing (NLP) algorithms, there will be yet another layer of error-prone operation on top of existing variability. Ideally, we would prefer multiple manually and reliably labeled close-to-truth annotations, while in reality, most of the data only have a single annotation from an algorithm with relatively low accuracy.

Learning to learn from a variety of data (labels) falls within the scope of meta-learning, which is popular in many machine learning applications, e.g., domain adaptation/generalization [5,6] and few-shot learning [20,30]. Those previous meta-learner models (as illustrated in Fig. 1(b)) often focus on learning the distribution of data (inputs of tasks) and specifying the update strategy of learner model parameters. Indeed, data from different sets (distributions) will contribute to the final learner model. On the contrary, we do not want the model to learn from erroneous labels (from less-experienced labelers) but learn only from “true” labels. We utilize the learner model parameters (via a meta-training process) to sample “true” labels for training a single learner model (as shown in Fig. 1(c)).

To address these challenges, we propose a learning-to-vote strategy to benefit the training from multiple labels on the same subject. Instead of the resource-demanding process of asking several human annotators to label the same data, we choose to utilize annotations from different algorithm-based labelers, which only add little overhead beyond the single-labeler scenario. A meta-training scheme is adopted and integrated into our proposed label-
sampling module for learning to vote on the labels that benefit the model learning the most through additional back-propagation processes.

Our contributions in this work are three-fold: 1) We proposed a training framework to compute the refined image labels on-the-fly in the classification tasks. The refined labels are voted on and sampled from multiple annotators via additional backpropagation; 2) Gradient flows toward the labels are investigated and implemented. Indeed, the multiple sets of labels are inputs to the training framework. Learnable operations of the labels will require additional updates of label-related model parameters; 3) We not only prove the concept on CIFAR-10 but also perform experiments on two real-world chest X-ray datasets with both image-only and image-text classification tasks. In all datasets, the superior performance of the proposed method is demonstrated in the image classification tasks compared to baseline methods.

2. Related Works

Meta learning: Meta learning aims to learn a generalizable model by situating itself at a higher level than conventional learning. This can be achieved in several ways such as finding weights that can be easily adapted to other models [6] or domains [15] during the training process. Meta-learning results in models that can converge quickly with a few examples [28]. They all share a similar meta-training process while the various goals of meta-training can divide them into different routes as examples shown in Fig. 1. In this work, we target weighting the importance of each label set based on its meta-training feedback and learning to vote for the most effective label for each data entry.

Learning from noisy labels: Learning from noisy labels [16, 23, 42, 43] has been a popular topic in deep learning due to its prevalence in many existing datasets with intra- and inter-observer variability, and the inherent uncertainties of both data and task themselves. For medical imaging applications with a high degree of ambiguity, this issue is even more significant. Recent works attempt to address this challenge via a consistency loss with a teacher model [17], loss weighting with 2nd order derivatives [42], and for medical image specifically, an online uncertainty sample mining strategy [39]. Please note that they all focus on noise labels from a single annotator while we attempt to design a learning-to-vote mechanism to utilize labels from different sources together.

Multi-label classification in chest X-ray: Because of its wide application and easy accessibility, chest X-ray is one of the major research areas in the field of medical image analysis. Among the pioneering works [8, 18, 26, 31, 32, 36, 40] in this area of deep learning, TieNet [37] first introduces an end-to-end trainable CNN-RNN architecture to extract distinctive text representations in addition to image features for improving label quality. More recently, a graph model was incorporated to integrate prior knowledge and enhance learning accuracy [41].

Multi-observer studies: To ensure the annotation quality, especially for medical images where high expertise is required, it is common to have multiple sets of labels on the same set of data [2, 27]. In a sense, each annotation can be regarded as an estimation with uncertainty, good or bad, for the underlying “true label”. Thus, algorithms taking this uncertainty factor into consideration are needed in order to make better use of such multi-observer data. [12] attempt to learn a distribution from a set of diverse but plausible segmentation from multiple graders. A recent work [33] proposed to model annotators by a confusion matrix which is jointly estimated during classification. In our work, instead of human annotators with different skill levels, we employed several “algorithmic labelers” to generate multiple annotations from the same raw data. Compared with their human counterparts, there exists fewer limitations and costs to increase the number of labelers, while the resulting labels can be noisier. Hence, we choose a different strategy of weighting module and meta-learning to benefit the model learning process.

3. Learning From Multiple Noisy Annotations

The accuracy of NLP algorithm-based labelers has been studied [9, 25] and verified by a small set of hand-labeled data (based on associated report texts). The noises in annotations could be traced from many sources, e.g., algorithmic errors, incomplete information in reports, and misjudgment from the clinicians. All of these could elevate the uncertainty and impair the reliability of the published data and associated ground-truth labels, in terms of their utilization in modern machine learning paradigms. “Who to believe?” becomes a fundamental question to answer, which will ultimately have a significant impact on the performance of the trained model.

MIMIC-CXR dataset [10] provides the label sets from two independent algorithm-based annotators with positive, negative, and uncertain cases. The availability of all these different sourced labels enables the observation of the uncertainty inherent to some data samples (with different values in multiple label sets). As shown in Table 1, there are many uncertain cases in each kind of finding, while the un-

<table>
<thead>
<tr>
<th>label sets</th>
<th>atelectasis</th>
<th>consolidated</th>
<th>edema</th>
<th>pneumonia</th>
<th>pneum-x</th>
</tr>
</thead>
<tbody>
<tr>
<td>negbio</td>
<td>10966</td>
<td>3348</td>
<td>13204</td>
<td>19029</td>
<td>1112</td>
</tr>
<tr>
<td>negbio-p</td>
<td>48704</td>
<td>11088</td>
<td>27911</td>
<td>16122</td>
<td>9885</td>
</tr>
<tr>
<td>u/p ratio</td>
<td>0.229</td>
<td>0.301</td>
<td>0.475</td>
<td>1.18</td>
<td>0.112</td>
</tr>
<tr>
<td>chexpert</td>
<td>10662</td>
<td>4446</td>
<td>13817</td>
<td>18915</td>
<td>1177</td>
</tr>
<tr>
<td>chexpert-p</td>
<td>47629</td>
<td>11231</td>
<td>28339</td>
<td>16757</td>
<td>11046</td>
</tr>
<tr>
<td>u/p ratio</td>
<td>0.223</td>
<td>0.395</td>
<td>0.487</td>
<td>1.128</td>
<td>0.106</td>
</tr>
</tbody>
</table>

Table 1. Number of uncertainties (\(\mu\)) and positives (\(\pi\)) of 5 sample disease findings from two labelers (i.e., negbio and chexpert).
certainty / positive ratio may vary in diseases (ranging from 0.106 to 1.18).

Here, we try to tackle this problem by training a multi-label classification model while considering all the available label sets. A novel learning-to-vote training process is introduced. For each set of labels, we perform individual back-propagation as a form of meta-training and then compute the new image/image+text feature using the individual updated model. Based on the new features, we learn to vote the label set with more representative features (with a weight) and sample the weighted summary of labels for the final update of the model in each iteration. Fig. 2 illustrates the overall architecture and learning processes for both image and label model parameters.

3.1. Multi-label Classification Backbone

We represent each image with $x$ and a label of classes with a binary vector $y = [y_1, ..., y_n, ..., y_N], y_n \in \{0, 1\}$. $y_n = 1$ indicates the presence of a corresponding disease pattern or other findings in the image. In the multi-label disease classification setting, the presence of each finding is predicted separately by producing a likelihood after applying the sigmoid on each logit. For the experiments on CIFAR (as multi-class classification), we also use the one-hot binary vectors to represent the labels and predictions. Therefore, the proposed framework could be utilized for both multi-label and multi-class classification tasks.

Our proposed learning-to-vote scheme could be applied to a large variety of pre-trained CNN architectures. Without loss of generality, we take the common ResNet-50 (from Conv1 to Res5c) as our backbone network. A global average pooling (GAP) layer was applied to transform the activation from convolutional layers into a one-dimension image feature $F$. The reason for applying a GAP layer is the necessity to pass concatenated image and text features to a fully-connected layer for the final classification.

We adopt the most common loss functions for the multi-label classification prediction $\tilde{y}$, i.e., binary cross entropy (BCE) loss: $L_C(\tilde{y}, y) = -1/N \sum_{i=1}^{N} y_i \log \tilde{y}_i + (1 - y_i) \log(1 - \tilde{y}_i)$. Other more advanced losses could also be employed, while this type of improvement is out-of-scope in this paper. Here, we would like to demonstrate the feasibility and benefit of applying our proposed meta-training process with learning-to-vote over a vanilla model. Other critical issues, like the unbalanced numbers of pathology compared with “normal” classes, are not considered here either to keep the evaluation simple and effective.

3.2. Attention on Labels

The overall training procedure is illustrated in Algorithm 1. For each training iteration, we input each data entry $(x, Y)$ from the training set and $Y = \{y_1, ..., y_m, ..., y_M\}$ are $M$ sets of image labels. During the meta-training, we compute the BCE loss ($L_C(\hat{y}, y_m, \theta)$) between the prediction $\hat{y}$ of current classification model $\theta$ and label set $y_m$ and then perform the back-propagation to compute a new set of model parameters ($\theta_m$),

$$\hat{\theta}_m = \theta - \alpha \nabla_{\theta} L_C(\hat{y}, y_m, \theta), \quad (1)$$

$\alpha$ is the learning rate for this meta-training process. We then compute a set of new features $\{F_m, m \in \{1, ..., M\}\}$ via the inference of image $x$ using each meta-model $\theta_m$ individually. $\{F_m\}$ could be either image features (i.e., the output of the GAP) or ones concatenated with text embedding (detailed in Section 3.3). Here, we compute a set of new features $\{F_m\}$ as an example to compute the weight $w_m$ for each label set via a softmax-based prediction mechanism,

$$w_m = \text{Softmax}(W_{\text{attn}}(\text{Cat}(\{F_m\}) + b_{\text{attn}}), \quad (2)$$

where $W_{\text{attn}}$ and $b_{\text{attn}}$ are learnable parameters in our learning-to-vote module. Softmax is the activation function. Cat represents the concatenation as a stack of all features. $w_m$ indicates the importance/correctness of label set $y_m$ and is applied to compute the weighted average of all label sets for each data sample. This process is similar to a common softmax-based attention mechanism [1, 38] and many other more complex learning-based attention mechanisms can also be adopted to compute the weights, e.g., self-attention [34].

The values of label vectors after weighted average $\bar{y} = \sum_m w_m y_m \in \{0, 1\}$ are in the range of $[0, 1]$, which is rather ambiguous for the multi-label classification model to learn. Binarization will be useful to cast the value close to either 0 or 1, but it is not differentiable and will disrupt the gradient flow. Therefore, we adopt the differentiable binarization function as first introduced in [19],

$$\tilde{y} = \frac{1}{1 + e^{-k(y - T)}}$$

where $k$ sets the sharpness of a 0 to 1 cliff. $T$ is a threshold to slightly adjust the value range. Finally, we update the model once more with the attended label and a global learning rate $\beta$ for this iteration,

$$\tilde{\theta} \leftarrow \theta - \beta \nabla L_C(\hat{y}, \tilde{y}, \theta). \quad (4)$$

Gradient Flows Towards Labels: Extra gradient flows (highlighted in red in Fig.2) are required for training our learning-to-vote mechanism, specifically the parameters in Eq. 2. Most of the current learning frameworks have gradient flows (highlighted in green in Fig.2) with the images
in the end since the labels are usually fixed or smoothed in advance [22]. However, the gradients in our proposed framework not only flow to the images but also go through towards the labels since the final labels $\tilde{y}$ are computed on-the-fly with learned weights/attentions. To our best knowledge, this concept of gradients towards labels is novel and has not been investigated and implemented before. Indeed, the inputs to the learning-to-vote module are $M$ sets of labels and the computed features $\{F_m\}$, which are detached (without auto-computed gradients) and stacked during the meta-training. Therefore, additional parameter updating is required at the end of each iteration,

$$W_{\text{attn}} \leftarrow W_{\text{attn}} - \beta \nabla L_C(\tilde{y}, \tilde{y}, W_{\text{attn}}), \quad (5)$$

$$b_{\text{attn}} \leftarrow b_{\text{attn}} - \beta \nabla L_C(\tilde{y}, \tilde{y}, b_{\text{attn}}). \quad (6)$$

### 3.3. Image-text Embedding

Clinical textual material, e.g., clinical notes [24] and radiology report [37], contains richer information. We include the text report as input to the classification problem to see if our proposed learning process will still benefit the learning and further improve the classification accuracy. There are a variety of approaches to generate text embedding, e.g., Fisher vectors of word2vec [11], bidirectional LSTMs [35], and the most recently developed BERT model [4]. To keep the simplicity of our baseline model, we embed the text report to a 768 dimension real-valued vector using the uncased version of BioBert features [14], followed by two fully connected layers with 512 neurons each.

### 4. Datasets

**CIFAR-10** [13]: We simulate 5 different types of annotators (with different years of experience) in a similar manner as [33] by injecting label noises into the training set of CIFAR-10, namely 1) hammer-spammer (HS), 2) structured-flips (SF), 3) ordered-confusion (OC), 4) Adversarial (AD), and 5) average (AVG) of previous four. Each set of noisy labels is generated based on the defined confusion matrices for each type (as shown in Fig. 3). Whether each sample would have a noisy label is randomly selected, while the overall noisy distribution should correspond to each confusion matrix individually. Within all the noisy training data, we randomly select 20% as the validation set.

**MIMIC-CXR**: The MIMIC Chest X-ray [10] Database is a large publicly available dataset of chest radiographs with labels mined from image-associated text radiology reports using two different NLP-based annotation tools, i.e., Negbio [25] and CheXpert [9]. The uncertain findings are marked as -1 in the original datasheet. Here, uncertainties...
are set to either 0 or 1 to form 4 different label sets, i.e., neg_u0, neg_u1, che_u0, and che_u1. The dataset contains 377,110 radiographs and labels from the 227,827 free-text radiology reports. Totally, 14 disease findings are listed (N = 14). In our experiments, only the frontal view images with associated reports are adopted. We utilize the official patient split for the training.

**MIMIC-CXR 1K hand-labeled Test**: Following the same labeling protocol proposed by [3], we randomly selected 1000 images and associated textual reports from the official testing split of MIMIC-CXR and one of our staff (trained by a board-certified radiologist) hand-labeled the 1000 images by assigning the 14 labels manually to each image based on the reports, which will be released publicly. We believe it will also benefit the research in chest x-ray disease classification.

**OpenI hand-labeled Test**: OpenI [3] is a public dataset of chest X-rays collected from multiple institutes by Indiana University. In total, we fetch 3,851 unique radiology reports and 7,784 associated frontal/lateral images. To keep the consistency with MIMIC-CXR dataset, we use the same 14 categories of findings as mentioned above. In the experiments, only 3,643 unique front-view images and associated reports are utilized for the evaluation.

### 5. Experiments:

The following methods, in addition to the proposed method (Ours), are included in the comparison:

- **ResNet-50 (Ours)**: We take the classifier based on ResNet-50 as a baseline. It adopts an ImageNet pre-trained ResNet-50 (from Conv1 to Res5c) as the backbone, followed by a GAP layer and a fully-connect layer for the final classification. Optionally, BioBert-embedded text features will be concatenated with the output of GAP before the classification. MV stands for a classifier trained with labels produced by majority voting among multiple annotations with ResNet-50 as the backbone. If the majority is not achieved, we randomly select one from the available sets.

- **CM [33]**: This state-of-the-art method multiplies a confusion matrix with the probability that the model produces for each class, which assumes that the learned confusion matrix can correct the missed labeled data and return the probability for the truth. We carefully implement it according to the code snapshot provided by the authors.

- **NG [41]**: This method utilizes the prior knowledge of the disease relations as a form of the knowledge graph. By injecting such prior knowledge and employing a graph convolutional network, it learns the underlying information for the final classification and report generation task. It is worth noting that the results we report are produced by a model that is both trained and evaluated on the OpenI.

- **TieNet [37]**: It focuses more on how to learn the image and text embedding together using a CNN+RNN framework. Its LSTM-based text embedding is relatively more complicated but also more representative through learning. We directly adopt the text embedding from a pre-trained BioBERT model (without finetuning) for the comparison, which is less customized.

**Evaluation Metric**: Receiver Operating Characteristic (ROC) curve is the standard metric to evaluate the performance of multi-label classifications. Here, Area Under the Curve (AUC) values are computed for all the experiments on MIMIC-CXR and OpenI. We compute the multi-class classification accuracy (using sklearn.metrics.accuracy_score) for all the experiments on CIFAR-10.

**Implementation Details**: For pre-processing, we resize the image to 256×256 (while keeping the size of 32×32 for CIFAR-10) and normalize the image intensities to [0, 1]. No data augmentation is employed in experiments. As mentioned above, we set the learning rate for the meta-training phase as α = 0.2 and the global learning rate as β = 1e−4. The best model for all hyper-parameters is determined via validation. We test k ∈ {10, 20, ..., 100} and T ∈ {0.1, 0.2, ..., 0.9} in the differentiable binarization module and find k = 50 and T = 0.5 provide the best re-
<table>
<thead>
<tr>
<th>CIFAR-10</th>
<th>HS</th>
<th>SF</th>
<th>OC</th>
<th>AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise-Level</td>
<td>30%</td>
<td>40%</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>Accuracy-Label</td>
<td>0.816</td>
<td>0.602</td>
<td>0.600</td>
<td>0.001</td>
</tr>
<tr>
<td>Accuracy-pred</td>
<td>0.808</td>
<td>0.438</td>
<td>0.555</td>
<td>0.025</td>
</tr>
<tr>
<td>AVG</td>
<td>MV</td>
<td>CM</td>
<td>Ours</td>
<td></td>
</tr>
<tr>
<td>Noise-Level</td>
<td>45%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Accuracy-Label</td>
<td>0.510</td>
<td>0.597</td>
<td>-</td>
<td>0.704</td>
</tr>
<tr>
<td>Accuracy-pred</td>
<td>0.521</td>
<td>0.611</td>
<td>0.643</td>
<td>0.705</td>
</tr>
</tbody>
</table>

Table 2. Averaged accuracy of refined labels (in training) and predictions (in testing), evaluated using clean labels on CIFAR-10.

<table>
<thead>
<tr>
<th>CIFAR-10</th>
<th>HS</th>
<th>SF</th>
<th>OC</th>
<th>AD</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaged $w_m$ (noise)</td>
<td>0.189</td>
<td>0.189</td>
<td>0.188</td>
<td>0.191</td>
<td>0.193</td>
</tr>
<tr>
<td>Averaged $w_m$ (clean)</td>
<td>0.209</td>
<td>0.209</td>
<td>0.212</td>
<td>0.0</td>
<td>0.205</td>
</tr>
</tbody>
</table>

Table 3. Averaged weights on noisy and clean labels.

results. A uniform batch size $B = 32$ and Adam optimizer is utilized for training all the compared models, using a single NVIDIA Titan-X Pascal.

5.1. Classification Results on CIFAR-10

To prove the concept, we employ CIFAR-10 with 5 types of added noises in the labels to illustrate that the proposed learning-to-vote scheme can be beneficial for the model training using multiple noise label sets. Table 2 illustrates the multi-class classification accuracy on the CIFAR-10 data set. Noise levels are computed in (1-Accuracy) using corresponding confusion matrices. In general, better-quality labels and data lead to better-trained models. Noise introduced by structured flips confuses the model training more than other types. Here, HS represents a more experienced set of annotators, and models trained with it obtained high accuracy. AVG represents the results of learning from a label set with a simple average of the noise level (defined by the confusion matrices) of all annotators. Our predictions achieve over 17%, 9%, and 6% performance improvements over AVG, MV (majority voting over multiple annotations), and the previous state-of-the-art CM.

Additionally, we record the accuracy of the refined (via weighted average) label set when achieving the best model shown as Accuracy-label. The accuracy of labels (noisy ones and ours, evaluated using clean labels of CIFAR-10 training) is highly correlated with the testing accuracy. In the same run, we also compute the averages of weights on both noisy labels and clean labels (considering the label sets are only partially degraded) for all 5 “labelers”. Table 3 shows the weights on clean labels are overall higher than the ones on noisy labels. This difference results in a more accurate label set than the ones via simple average and majority voting and further leads to a better-trained model.

In addition, we investigate how the label noise level, number of annotators, and annotations with serious errors will affect the model performance.

How Will Different Noise Levels of Labels Affect? Here, we include all 5 sets of stimulated labels, i.e., HS, SF, OC, AD, and AVG. 4 sets (HS, SF, OC, and AVG) are adjusted with the same noise levels (ranging from 10% to 80%). AD will remain the same for all levels since its noise level cannot be tuned. We compare the classification prediction accuracy of 4 models, i.e., R50, that are trained using AVG and labels from majority voting (VM), CM, and our proposed method (Ours). Both CM and Ours are using all 5 label sets for the training. As shown in Figure 4, the classification performance drops along with the increase of noise level in labels, while Ours can constantly achieve better or similar results as others. Particularly, Ours outperforms MV and CM with large margins (over 10%) for noise-level 40% to 60%.

How Will Number of Available Annotation Sets Affect? In this experiment, we try to vary the number of available label sets (2, 3, 4, and 5 different types of annotators) used for training the proposed model (Ours), MV, and CM at different noise level (at 10%, 30%, and 50%). We start with the training model using 2 quite different label sets (#2), i.e., HS and AD. Then we add OC, SF, and AVG one at a time to see how the increasing number of the label sets could affect the final classification performance. HS represents a relatively experienced annotator, and AD can be seen as a ‘bad’ annotator with serious systematic errors. As shown in Figure 5, a surge in the accuracy can be observed after including more than 2 annotators. Considering that the adversarial annotator AD (with noise-level 100%) is among the initial two, all three methods can learn better immediately after a third one (as a confirmation) jumps in. We can also observe that our method performs much better when high noise levels are presented (50% in this case).

How Will Annotations With Serious Errors Can Affect? As shown in previous experiments, AD can influence performance significantly when only a small number of label sets are employed. Here, we want to investigate how it will function if more sets of annotations are included. Figure 6 shows the performance gap when three compared methods are trained either with all available (5 in total) annotation sets or with AD left out (4 sets in total). Both MV and CM will have remarkable accuracy decrease (5% to 10%) for noise level from 40% to 70%. In comparison, our proposed method is much more robust in such scenarios.

5.2. Classification Results on Chest X-Ray Images

Classification Using Image Only: Table 4(a) shows the evaluation results for all the compared methods using only the images as the input on MIMIC-CXR 1K hand-labeled Test and OpenI hand-labeled Test. In the left part of Table 4, we show the accuracies of 4 sets of NLP-mined la-
labels against hand-labeled groundtruth. An average of 15% to 20% noise level are observed. \textit{neg$_u1$}, \textit{neg$_u0$}, \textit{che$_u1$}, and \textit{che$_u0$} are derived from the original NLP-mined labels (via either Negbio or Chexpert labelers) by setting the uncertainty to either 1 or 0. The binarization of uncertainty divides NLP labels to have quite different accuracies (~6% gap) on the MIMIC-CXR 1K hand-labeled Test set. For MIMIC-CXR 1K hand-labeled Test data, we show the AUCs of all the finding categories from R50, R50\_MV, CM and Ours, while additional results from NG and TieNet are presented for OpenI hand-labeled Test set. The AUCs for Ours are constantly higher (~3% on average) than the compared approaches on both datasets. By considering the relatively low noise level presented in the MIMIC-CXR dataset, the reported improvements are consistent with the ones from experiments on CIFAR-10. OpenI dataset has been utilized here for the evaluation purpose only, and our method (without training on OpenI) is able to achieve ~4.5% increase in the averaged AUC, which is also greater than what MV and CM method achieves. Note that the absolute accuracies are overall higher for the one reported on OpenI dataset. It may indicate the domain gap between MIMIC-CXR and OpenI datasets. Although both NG and TieNet partially utilized the report textual information in their image classification framework, Ours still is able to obtain equivalent or better results in most of the detailed disease categories. As mentioned above, those disease categories with a larger amount of uncertainties provide more information and therefore benefit more from the proposed meta-training process, e.g., Atelectasis and Devices.

**Classification Using Both Chest Image and Report:** The text report contains richer information about the disease diagnosis. We observe an increase in all AUCs. In this case, our proposed meta-training with the learning-to-vote scheme also helps to boost the classification performance with a significant margin shown in Table 4. Ours actually achieves higher or similar accuracy in comparison to the NLP annotators while a model trained with NLP mined labels usually can not reach the same level of accuracy as the labels themselves, e.g., R50 is trained with \textit{neg$_u1$}.

### 6. Discussion and Conclusion

The proposed label sampling module converts the hard labels (0 or 1) to soft labels (a value between 0 and 1),
and it also reduces the overfit towards erroneous labels, which is similar to the idea of label smoothing [22]. Unlike hard label smoothing, we assigned the new label on-the-fly (based on the network feedback) instead of instantly replacing 0 and 1 with \(0 + \epsilon \) and \(1 - \epsilon\). Such a setting effectively handles partial label errors in multi-label classification. Indeed, label smoothing is a form of loss-reparameterization which also indicates the upper-bound performance of a model trained using that label set as GT. E-cardio: enlarged-cardiomegaly.

Table 4. Classification results (AUCs) for 14 findings in Chest X-Rays from the models trained on MIMIC-CXR and tested on MIMIC-CXR 1K hand-labeled Test (a) and Open hand-labeled Test (b) data. \(\text{neg}_u, \text{neg}_0, \text{che}_u, \text{che}_0\) are derived from the original NLP mined labels by setting the uncertainty to either 1 or 0. Here shows their accuracy against hand-labeled groundtruth (GT), which actually indicates the up-bound performance of a model trained using that label set as GT. E-cardio: enlarged-cardiomegaly.

In summary, we introduced a novel learning framework (the learning-to-vote module and gradients towards label data) for handling data with multiple noisy label sets. The variability that the multiple labels bring could, in fact, benefit the learning of a more accurate and robust model. The proposed method provides an accurate and robust means for the challenges of learning large-scale data with algorithm-generated labels whenever the annotation remains a burden, e.g., medical image analysis.
References


[18] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8290–8299, 2018.


