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, daguangx}@nvidia.com

Abstract

Supervised learning paradigms largely benefit from the tremendous amount of annotated data. However, the quality of the annotations often 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.

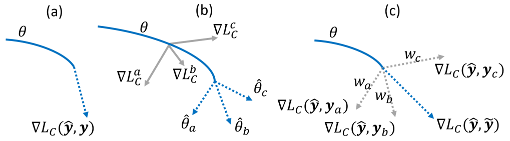

(a)
θ
$\nabla L_C(\hat{\boldsymbol{y}}, \boldsymbol{y})$
(b)
θ
∇L_C^c
∇L_C^a
∇L_C^b
$\hat{\theta}_c$
$\hat{\theta}_a$
$\hat{\theta}_b$
(c)
θ
w_c
$\nabla L_C(\hat{\boldsymbol{y}}, \boldsymbol{y}_c)$
w_a
w_b
$\nabla L_C(\hat{\boldsymbol{y}}, \boldsymbol{y}_a)$
$\nabla L_C(\hat{\boldsymbol{y}}, \boldsymbol{y}_b)$
$\nabla L_C(\hat{\boldsymbol{y}}, \tilde{\boldsymbol{y}})$

Figure 1. The diagram shows the difference in three learning paradigms in how gradients are utilized for the training, i.e., (a) regular gradient-based learning; (b) Meta-learning with multiple learning targets; (c) Proposed: learning the weights of each label ($\mathbf{y}_a, \mathbf{y}_b, \mathbf{y}_c$) in meta-training and then computing the weighted sum ($\tilde{\mathbf{y}}$) of labels for computing the final loss with prediction $\hat{\mathbf{y}}$.

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 backpropagation 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].

label sets	atelectasis	consolid.	edema	pneumonia	pneum-x
negbio_u	10986	3348	13204	19029	1112
negbio_p	47804	11088	27911	16122	9885
u/p ratio	0.229	0.301	0.473	1.18	0.112
chexpert_u	10662	4446	13817	18915	1177
chexpert_p	47629	11231	28339	16757	11046
u/p ratio	0.223	0.395	0.487	1.128	0.106

Table 1. Number of uncertainties (_u) and positives (_p) of 5 sample disease findings from two labelers (i.e., negbio and chexpert).

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 uncertainty / 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 $\mathbf{y} = [y_1, ..., y_n, ..., y_N], y_n \in \{0, 1\}$. $y_n = 1$ 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 $\hat{\mathbf{y}}$, i.e., binary cross entropy (BCE) loss: $\mathcal{L}_C(\hat{\mathbf{y}}, \mathbf{y}) = -1/N \sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)$. 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 $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_m, \ldots, \mathbf{y}_M\}$ are M sets of image labels. During the meta-training, we compute the BCE loss $(\mathcal{L}_C(\hat{\mathbf{y}}, \mathbf{y}_m, \theta))$ between the prediction $\hat{\mathbf{y}}$ of current classification model θ and label set $\mathbf{y}_m$ and then perform the back-propagation to compute a new set of model parameters $(\hat{\theta}_m)$,

$$\hat{\theta}_m = \theta - \alpha \nabla_\theta \mathcal{L}_C(\hat{\mathbf{y}}, \mathbf{y}_m, \theta), \tag{1}$$

α 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 $\hat{\theta}_m$ 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). $\{F_m\}$ represent the feedback of model updates with each label set $\mathbf{y}_m$, i.e., the change that each $\mathbf{y}_m$ has brought to the model θ. Other types of feedback from each noisy label could also be utilized here, e.g., the gradients $\{\nabla_\theta \mathcal{L}_C(\hat{\mathbf{y}}, \mathbf{y}_m, \theta), m \in \{1, ..., M\}\}$. Here, we take $\{F_m\}$ as an example to compute the weight w_m for each label set via a softmax-based prediction mechanism,

$$w_m = \mathrm{Softmax}(\mathbf{W}_{attn}(\mathrm{Cat}(\{F_m\}) + \mathbf{b}_{attn}), \tag{2}$$

where $\mathbf{W}_{attn}$ and $\mathbf{b}_{attn}$ are learnable parameters in our learning-to-vote module. $\mathrm{Softmax}$ is the activation function. Cat represents the concatenation as a stack of all features. w_m indicates the importance/correctness of label set $\mathbf{y}_m$ 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 $\bar{\mathbf{y}} = \sum_m w_m \mathbf{y}_m$ 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],

$$\tilde{\mathbf{y}} = \frac{1}{1 + e^{-k(\bar{\mathbf{y}} - T)}}, \tag{3}$$

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 β for this iteration,

$$\tilde{\theta} \leftarrow \theta - \beta \nabla \mathcal{L}_C(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \theta). \tag{4}$$

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

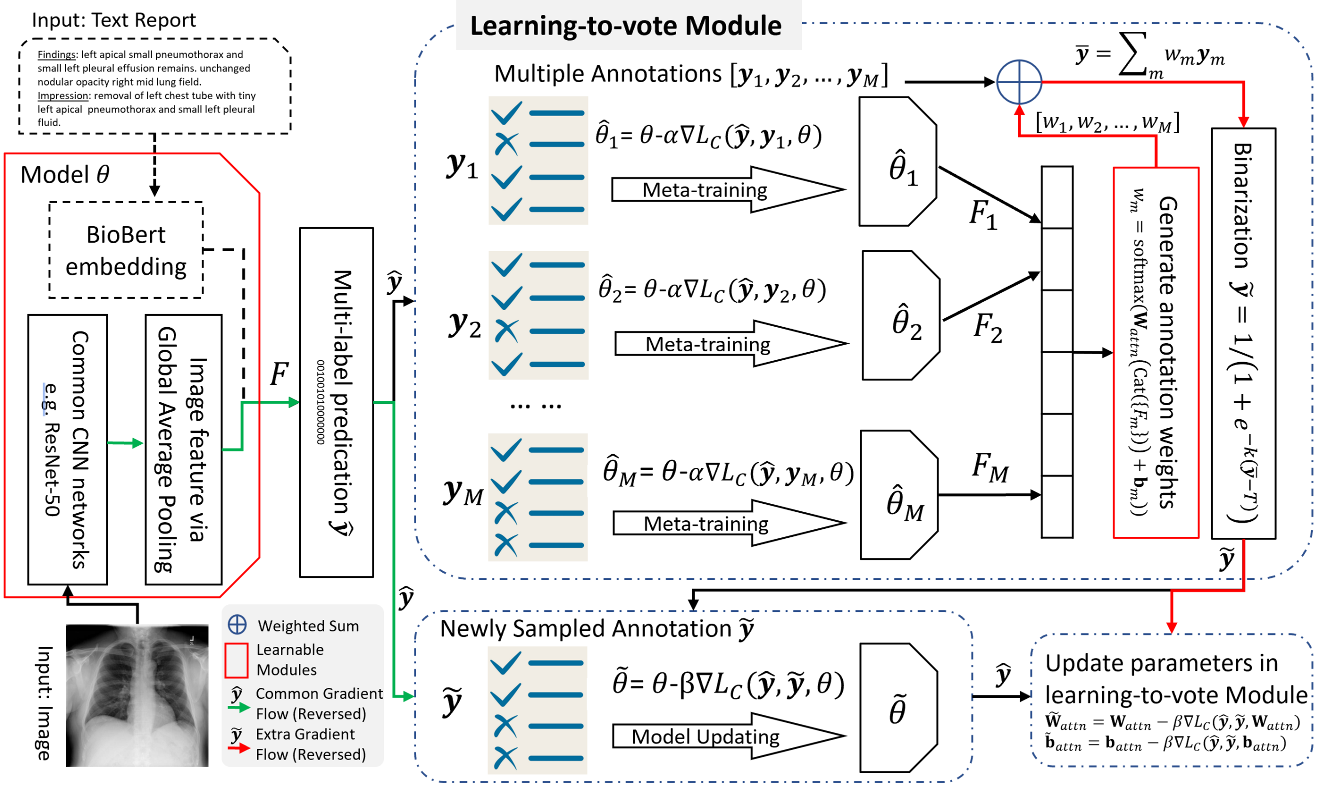

Input: Text Report
Findings: left apical small pneumothorax and small left pleural effusion remains. unchanged nodular opacity right mid lung field.
Impression: removal of left chest tube with tiny left apical pneumothorax and small left pleural fluid.
Model θ
BioBert embedding
Common CNN networks e.g. ResNet-50
Image feature via Global Average Pooling
F
Multi-label predication $\hat{\boldsymbol{y}}$
0010010100000000
$\hat{\boldsymbol{y}}$
Input: Image
Weighted Sum
Learnable Modules
$\hat{\boldsymbol{y}}$ Common Gradient Flow (Reversed)
$\tilde{\boldsymbol{y}}$ Extra Gradient Flow (Reversed)
Learning-to-vote Module
Multiple Annotations $[\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_M]$
$\boldsymbol{y}_1$
$\hat{\theta}_1 = \theta - \alpha \nabla L_C(\hat{\boldsymbol{y}}, \boldsymbol{y}_1, \theta)$
Meta-training
$\hat{\theta}_1$
F_1
$\boldsymbol{y}_2$
$\hat{\theta}_2 = \theta - \alpha \nabla L_C(\hat{\boldsymbol{y}}, \boldsymbol{y}_2, \theta)$
Meta-training
$\hat{\theta}_2$
F_2
… …
$\boldsymbol{y}_M$
$\hat{\theta}_M = \theta - \alpha \nabla L_C(\hat{\boldsymbol{y}}, \boldsymbol{y}_M, \theta)$
Meta-training
$\hat{\theta}_M$
F_M
$\bar{\boldsymbol{y}} = \sum_m w_m \boldsymbol{y}_m$
$[w_1, w_2, \dots, w_M]$
Generate annotation weights
$w_m = \mathrm{softmax}(\mathbf{W}_{attn}(\mathrm{Cat}(\{F_m\})) + \mathbf{b}_m))$
Binarization $\tilde{\boldsymbol{y}} = 1/(1 + e^{-k(\bar{\boldsymbol{y}} - T)})$
$\tilde{\boldsymbol{y}}$
Newly Sampled Annotation $\tilde{\boldsymbol{y}}$
$\tilde{\boldsymbol{y}}$
$\tilde{\theta} = \theta - \beta \nabla L_C(\hat{\boldsymbol{y}}, \tilde{\boldsymbol{y}}, \theta)$
Model Updating
$\tilde{\theta}$
$\hat{\boldsymbol{y}}$
Update parameters in learning-to-vote Module
$\widetilde{\mathbf{W}}_{attn} = \mathbf{W}_{attn} - \beta \nabla L_C(\hat{\boldsymbol{y}}, \tilde{\boldsymbol{y}}, \mathbf{W}_{attn})$
$\tilde{\mathbf{b}}_{attn} = \mathbf{b}_{attn} - \beta \nabla L_C(\hat{\boldsymbol{y}}, \tilde{\boldsymbol{y}}, \mathbf{b}_{attn})$

Figure 2. Overview of the proposed learning pipeline. Learning-to-vote is added for a refined label sampling process.

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 $\tilde{\mathbf{y}}$ 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,

$$\tilde{\mathbf{W}}_{attn} \leftarrow \mathbf{W}_{attn} - \beta \nabla \mathcal{L}_C(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{W}_{attn}), \tag{5}$$

$$\tilde{\mathbf{b}}_{attn} \leftarrow \mathbf{b}_{attn} - \beta \nabla \mathcal{L}_C(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{b}_{attn}). \tag{6}$$

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

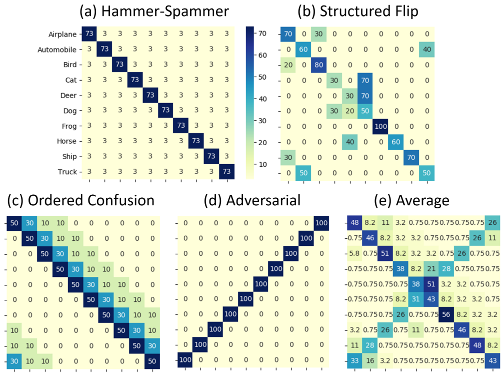


Figure 3. Confusion matrices of noisy labels from 5 different types of simulated annotators.

Algorithm 1 Meta-training with learning-to-vote module

```
Randomly initialize $\theta$
for each data entry $(x, Y)$ do
    Inference with $x, \theta$ to predict $\hat{\mathbf{y}}$
    for $m \in \{1 : M\}$ do
        Update parameters: $\hat{\theta}_m = \theta - \alpha \nabla_\theta \mathcal{L}_c(\hat{\mathbf{y}}, \mathbf{y}_m, \theta)$
        Compute features $F_m$ using the newly updated $\hat{\theta}_m$
    Stack and concatenate the features: $\mathrm{Cat}(\{F_m\})$
    Compute voting weights $w_m$ for each feature $F_m$ (Eq. 2)
    Sample the new label $\bar{\mathbf{y}} = \sum_m w_m \mathbf{y}_m$
    Perform differentiable binarization $\bar{\mathbf{y}} \rightarrow \tilde{\mathbf{y}}$
    Update the final image model $\tilde{\theta} \leftarrow \theta - \beta \nabla \mathcal{L}_C(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \theta)$
    Manually update learning-to-vote module ($\mathbf{W}_{attn}, \mathbf{b}_{attn}$)
```

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 [7] (**R50**, **MV**): 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 $\alpha = 0.2$ and the global learning rate as $\beta = 1e-4$. The best model for all hyper-parameters is determined via validation. We test $k \in \{10, 20, ..., 100\}$ and $T \in \{0.1, 0.2, ..., 0.9\}$ in the differentiable binarization module and find $k = 50$ and $T = 0.5$ provide the best re-

CIFAR-10	HS	SF	OC	AD
Noise-Level	30%	40%	50%	100%
Accuracy-Label	0.816	0.602	0.600	0.001
Accuracy-pred	0.808	0.438	0.555	0.025
	AVG	MV	CM	Ours
Noise-Level	45%	-	-	-
Accuracy-Label	0.510	0.597	-	**0.704**
Accuracy-pred	0.521	0.611	0.643	**0.705**

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

CIFAR-10	HS	SF	OC	AD	AVG
Averaged w_m (noise)	0.189	0.189	0.188	0.191	0.193
Averaged w_m (clean)	0.209	0.209	0.212	0	0.205

Table 3. Averaged weights on noisy and clean labels.

sults. 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 noise 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.

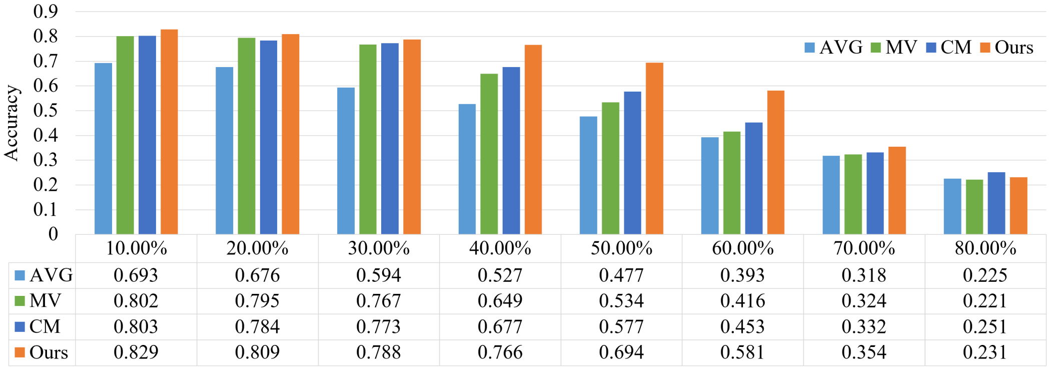
How Will Different Noise Levels of Labels Affect? Here, we include all 5 sets of stimulated labels, i.e., HS, SF, OC, AD, and AVG. 4 sets(HS, SF, OC, and AVG) are adjusted with the same noise levels (ranging from 10% to 80%). AD will remain the same for all levels since its noise level can not be tuned. 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-



	10.00%	20.00%	30.00%	40.00%	50.00%	60.00%	70.00%	80.00%
AVG	0.693	0.676	0.594	0.527	0.477	0.393	0.318	0.225
MV	0.802	0.795	0.767	0.649	0.534	0.416	0.324	0.221
CM	0.803	0.784	0.773	0.677	0.577	0.453	0.332	0.251
Ours	0.829	0.809	0.788	0.766	0.694	0.581	0.354	0.231

Figure 4. Classification accuracy using different label sets and methods over a range of noise levels.

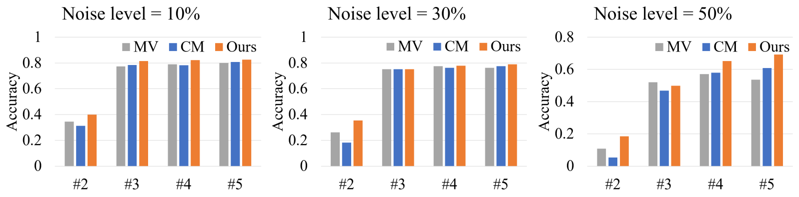


Figure 5. Classification accuracy using a different number of label sets at noise-level 10%, 30%, and 50%.

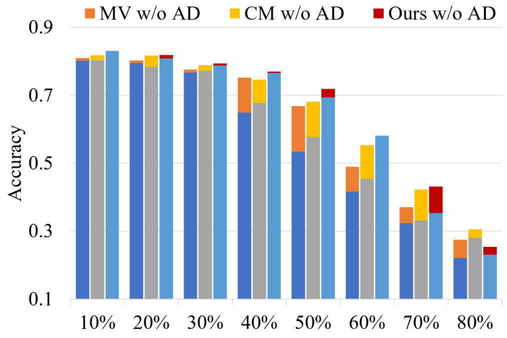


Figure 6. Classification accuracy using all available label sets with and without AD at different noise levels.

bels against hand-labeled groundtruth. An average of 15% to 20% noise level are observed. neg_u1, neg_u0, che_u1, and 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 ($\sim$6% gap) on the MIMIC-CXR 1K hand-labeled Test set. For MIMIC-CXR 1K hand-labeled Test data, we show the AUCs of all the finding categories from R50, R50_MV, CM and Ours, while additional results from NG and TieNet are presented for OpenI hand-labeled Test set. The AUCs for Ours are constantly higher ($\sim$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 $\sim$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 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),

(a) MIMIC-CXR 1K hand-labeled Test

AUC	Noisy NLP labeler (GT for training)				Image-only				Image & Report Text			
Disease	neg_u1	neg_u0	che_u1	che_u0	R50	R50_MV	CM	Ours	R50	R50_MV	CM	Ours
Atelectasis	0.914	0.832	0.908	0.829	0.737	0.743	0.758	0.768	0.962	0.945	0.958	0.963
Cardiomegaly	0.821	0.805	0.822	0.814	0.787	0.769	0.79	0.805	0.865	0.86	0.871	0.862
Consolidation	0.892	0.772	0.875	0.777	0.684	0.657	0.658	0.694	0.802	0.776	0.796	0.869
Edema	0.955	0.901	0.948	0.903	0.807	0.774	0.804	0.804	0.877	0.871	0.88	0.912
E-cardio	0.847	0.759	0.847	0.759	0.592	0.644	0.681	0.629	0.744	0.703	0.745	0.769
Fracture	0.733	0.689	0.747	0.718	0.651	0.743	0.595	0.628	0.671	0.769	0.746	0.769
Lesion	0.777	0.727	0.777	0.719	0.714	0.679	0.69	0.727	0.76	0.71	0.739	0.759
Opacity	0.877	0.861	0.871	0.863	0.662	0.631	0.648	0.685	0.877	0.874	0.89	0.885
No-finding	0.845	0.845	0.815	0.815	0.722	0.72	0.74	0.769	0.909	0.897	0.906	0.909
Effusion	0.941	0.906	0.94	0.913	0.828	0.843	0.856	0.848	0.923	0.91	0.925	0.931
Pleural-other	0.906	0.825	0.906	0.825	0.754	0.782	0.714	0.802	0.803	0.803	0.749	0.844
Pneumonia	0.957	0.685	0.955	0.694	0.731	0.713	0.749	0.753	0.872	0.875	0.878	0.908
Pneumothorax	0.874	0.831	0.917	0.86	0.797	0.795	0.805	0.81	0.852	0.86	0.788	0.89
Devices	0.838	0.837	0.836	0.837	0.793	0.829	0.827	0.843	0.872	0.876	0.86	0.885
Average	**0.868**	0.805	0.868	0.809	0.732	0.737	0.736	**0.755**	0.842	0.837	0.837	**0.868**

(b) OpenI hand-labeled Test

AUC	Image-only						Image & Report Text			
Disease	R50	R50_MV	NG	TieNet	CM	Ours	R50	R50_MV	CM	Ours
Atelectasis	0.781	0.802	0.833	0.774	0.81	0.826	0.901	0.934	0.909	0.925
Cardiomegaly	0.859	0.842	0.913	0.847	0.881	0.879	0.915	0.928	0.928	0.949
Consolidation	0.829	0.872	-	-	0.842	0.906	0.914	0.924	0.891	0.907
Edema	0.895	0.87	0.931	0.879	0.924	0.885	0.903	0.937	0.915	0.939
E-cardio	0.795	0.673	-	-	0.758	0.725	0.581	0.595	0.714	0.598
Fracture	0.513	0.612	0.671	-	0.596	0.632	0.705	0.669	0.683	0.739
Lesion	0.585	0.603	0.643	0.658	0.58	0.643	0.615	0.636	0.607	0.649
Opacity	0.742	0.735	0.803	-	0.738	0.775	0.849	0.858	0.854	0.877
No-finding	0.754	0.743	-	0.747	0.739	0.775	0.79	0.809	0.82	0.867
Effusion	0.912	0.926	0.942	0.899	0.932	0.942	0.944	0.954	0.948	0.943
Pleural-other	0.648	0.678	-	-	0.676	0.705	0.723	0.743	0.778	0.739
Pneumonia	0.781	0.784	0.863	0.731	0.823	0.871	0.812	0.877	0.834	0.889
Pneumothorax	0.793	0.805	0.843	0.709	0.882	0.833	0.879	0.84	0.879	0.853
Devices	0.628	0.662	0.805	-	0.655	0.729	0.796	0.786	0.787	0.821
Average	0.751	0.757	-	-	0.774	**0.795**	0.809	0.820	0.824	**0.835**

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 OpenI hand-labeled Test (b) data. neg_u1, neg_u0, che_u1, and che_u0 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-cardiomediastinum.

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-correction [21], and we study that computed label weights can also be employed to re-weight the losses computed using label sets from different annotators. The computed loss $\mathcal{L}_C(\hat{\mathbf{y}}, \tilde{\mathbf{y}}) = \mathcal{L}_C(\hat{\mathbf{y}}, \sum_{m=1}^{M} w_m \mathbf{y}_m)$ is equivalent to a weighted summation of losses computed with each set of label $\mathbf{y}_m$, i.e., $\mathcal{L}_C(\hat{\mathbf{y}}, \tilde{\mathbf{y}}) = \sum_{m=1}^{M} w_m \mathcal{L}_C(\hat{\mathbf{y}}, \mathbf{y}_m)$, where $\hat{\mathbf{y}}$ is the model prediction, $\tilde{\mathbf{y}} = \sum_{m=1}^{M} w_m \mathbf{y}_m$ is the sampled label, and $\mathcal{L}_C(\hat{\mathbf{y}}, \mathbf{y})$ is the binary cross-entropy loss. Please find the detailed proof in the supplementary material. It shares a similar insight with [29] while Ren et al. [29] re-weighted the losses based on the variation of input images. Similar to the above theorem, we can prove that re-weighting the losses is equivalent to the weighted sample of network predictions (maybe with different data samples as input). Accordingly, the noise in data (on both images and labels) can be measured and corrected by weighting the losses. It also leads to a broader research topic of modeling the distribution of losses, e.g., modeling the losses with Gaussian mixture models [16].

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 wherever the annotation remains a burden, e.g., medical image analysis.

References

[1] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017. 3

[2] Thomas Cherian, E Kim Mulholland, John B Carlin, Harald Ostensen, Ruhul Amin, Margaret de Campo, David Greenberg, Rosanna Lagos, Marilla Lucero, Shabir A Madhi, et al. Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. *Bulletin of the World Health Organization*, 83:353–359, 2005. 2

[3] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015. 5

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[5] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6447–6458, 2019. 1

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1126–1135. JMLR.org, 2017. 1, 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[8] Eui Jin Hwang, Sunggyun Park, Kwang-Nam Jin, Jung Im Kim, So Young Choi, Jong Hyuk Lee, Jin Mo Goo, Jaehong Aum, Jae-Joon Yim, Julien G Cohen, et al. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA network open*, 2(3):e191095–e191095, 2019. 2

[9] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. 2, 4

[10] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a deidentified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. 2, 4

[11] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015. 4

[12] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018. 2

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[14] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 4

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. *arXiv preprint arXiv:1710.03463*, 2017. 2

[16] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 2, 8

[17] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019. 2

[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. 2

[19] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. *arXiv preprint arXiv:1911.08947*, 2019. 3

[20] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10551–10560, 2019. 1

[21] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? *arXiv preprint arXiv:2003.02819*, 2020. 8

[22] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4696–4705, 2019. 4, 8

[23] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013. 2

[24] Obioma Pelka, Felix Nensa, and Christoph M Friedrich. Branding-fusion of meta data and musculoskeletal radiographs for multi-modal diagnostic recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 4

[25] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018. 2, 4

[26] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 2

[27] Coen Rasch, Isabelle Barillot, Peter Remeijer, Adriaan Touw, Marcel van Herk, and Joos V Lebesque. Definition of the prostate in ct and mri: a multi-observer study. *International Journal of Radiation Oncology* Biology* Physics*, 43(1):57–66, 1999. 2

[28] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 2

[29] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018. 8

[30] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 1

[31] Y Tang, Y Peng, K Yan, M Bagheri, BA Redd, CJ Brandon, Z Lu, M Han, J Xiao, and RM Summers. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ Digital Medicine*, 3, 2020. 2

[32] Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pages 249–258. Springer, 2018. 2

[33] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2019. 2, 4, 5

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3

[35] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997, 2016. 4

[36] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 2

[37] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018. 2, 4, 5

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 3

[39] C. Xue, Q. Dou, X. Shi, H. Chen, and P. Heng. Robust learning at noisy labeled medical images: Applied to skin lesion classification. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1280–1283, April 2019. 2

[40] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*, 2017. 2

[41] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of AAAI*, 2020. 2, 5

[42] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9294–9303, 2020. 2

[43] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan T. Dumais. Meta label correction for noisy label learning. In *AAAI Conference on Artificial Intelligence*, 2019. 2