This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

BoMD: Bag of Multi-label Descriptors for Noisy Chest X-ray Classification

Yuanhong Chen [1] * Fengbei Liu [1] * Hu Wang [1] Chong Wang [1]
Yuyuan Liu [1] Yu Tian[2] Gustavo Carneiro[3]
[1] Australian Institute for Machine Learning, University of Adelaide
[2] Harvard Medical School, Harvard University
[3] Centre for Vision, Speech and Signal Processing, University of Surrey

Abstract

Deep learning methods have shown outstanding classification accuracy in medical imaging problems, which is largely attributed to the availability of large-scale datasets manually annotated with clean labels. However, given the high cost of such manual annotation, new medical imaging classification problems may need to rely on machine-generated noisy labels extracted from radiology reports. Indeed, many Chest X-Ray (CXR) classifiers have been modelled from datasets with noisy labels, but their training procedure is in general not robust to noisy-label samples, leading to sub-optimal models. Furthermore, CXR datasets are mostly multi-label, so current multi-class noisy-label learning methods cannot be easily adapted. In this paper, we propose a new method designed for noisy multi-label CXR learning, which detects and smoothly re-labels noisy samples from the dataset to be used in the training of common multi-label classifiers. The proposed method optimises a bag of multi-label descriptors (BoMD) to promote their similarity with the semantic descriptors produced by language models from multi-label image annotations. Our experiments on noisy multi-label training sets and clean testing sets show that our model has state-of-the-art accuracy and robustness in many CXR multi-label classification benchmarks, including a new benchmark that we propose to systematically assess noisy multi-label methods. Code is available at https://github.com/cyh-0/BoMD.

1. Introduction

The promising results produced by deep neural networks (DNN) in medical image analysis (MIA) problems [30] is attributed to the availability of large-scale datasets with accurately annotated images. Given the high cost of acquiring such datasets, the field is considering more affordable automatic annotation processes by Natural Language Processing (NLP) approaches that extract multiple labels (each label representing a disease) from radiology reports [20,57]. However, mistakes made by NLP combined with uncertain radiological findings can introduce label noise [44,45], as can be found in NLP-annotated Chest X-ray (CXR) datasets [20, 57] whose noisy multi-labels can mislead supervised training processes. Nevertheless, even without addressing this noisy multi-label problem, current CXR multi-label classifiers [3, 15, 39, 50] show promising results. Although these methods show encouraging multi-label classification accuracy, there is still potential for improvement

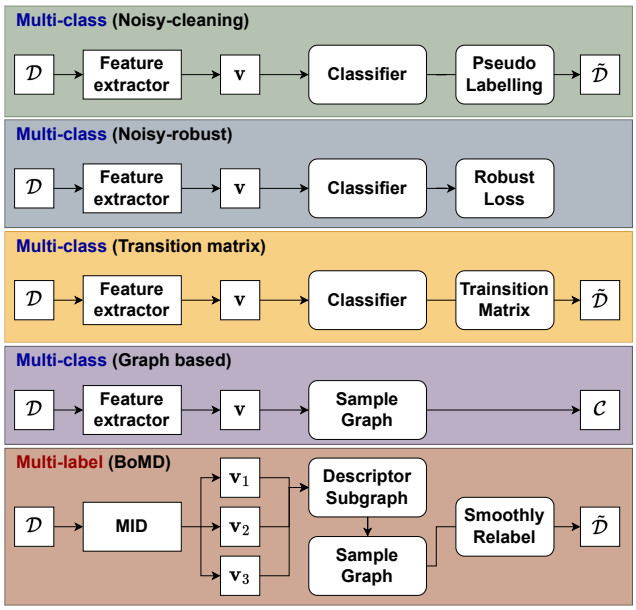


Figure 1: Comparison of multi-class LNL methods [1, 21, 28, 33] and our noisy multi-label approach, BoMD, where the feature extractor returns a single descriptor $\mathbf{v}$ per image, $\mathcal{D}$ is the noisy training set, $\mathcal{C}$ is the clean set, and $\tilde{\mathcal{D}}$ is the re-labelled training set. BoMD has two components: 1) learning of a bag of multi-label image descriptors (MID) $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ to represent the image, and 2) smooth re-labelling of images driven by a graph structure based on the fine-grained relationships between MID descriptors.

*First two authors contributed equally to this work.

that can be realised by properly addressing the noisy multi-label learning problem present in CXR datasets [20, 57].

Current learning with noisy label (LNL) approaches focus on leveraging the hidden clean-label information to assist the training of DNNs (see Fig. 1). This can be achieved with techniques that clean the label noise [25, 28], robustify loss functions [21, 31, 33], estimate label transition matrices [1, 12, 62, 65], smooth training labels [38, 59, 66], and use graphs to explore the latent structure of data. [21, 60, 61]. These methods have been designed for noisy multi-class problems and do not easily extend to *noisy multi-label* learning, which is challenging given the potential multiple label mistakes for each training sample. In addition, a key characteristic of multi-label classification is the inherent positive-negative imbalance [51] issue. Such an issue may cause sample-selection based methods (e.g., DivideMix [28]) to select an extremely imbalanced clean set, where the majority of identified clean samples belong to the 'No Findings' class. Additionally, it could impede the accurate estimation of the posterior probabilities for the noisy or intermediate classes [29]." To the best of our knowledge, the state-of-the-art (SOTA) approach that handles noisy multi-label learning is NVUM [31], which is based on an extension of early learning regularisation (ELR) [33]. NVUM shows promising results, but it is challenged by the different early convergence patterns of multiple labels, which can lead to poor performance for particular label noise conditions, as shown in our experiments. Additionally, NVUM is only evaluated on real-world CXR datasets [20, 57] without any systematic assessment of robustness to varying label noise conditions, preventing a more complete understanding of its functionality.

In this paper, we propose a new solution specifically designed for the noisy multi-label problem by answering the following question: ***can the detection and correction of noisy multi-labelled samples be facilitated by leveraging the semantic information of training labels?*** available from language models [19, 27, 48]? This question is motivated by the successful exploration of language models in computer vision [3, 7, 16, 67], with methods that leverage semantic information to influence the training of visual descriptors; an idea that has not been explored in noisy multi-label classification. To answer this question, we introduce the 2-stage Bag of Multi-label Descriptors (BoMD) method (see Fig. 1) to smoothly re-label noisy multi-label image datasets that can then be used for training common multi-label classifiers. The first stage trains a feature extractor to produce a bag of multi-label image descriptors by promoting their similarity with the semantic embeddings from language models. For the second stage, we introduce a novel graph structure, where each image is represented by a sub-graph built from the multi-label image descriptors, learned in the first stage, to smoothly re-label the noisy multi-label images. Compared with graphs built directly from a single descriptor per image [21], our graph structure with the multi-label image descriptors has the potential to capture more fine-grained image relationships, which is crucial to deal with multi-label annotation. We also propose a new benchmark to systematically assess noisy multi-label methods. In summary, our **contributions** are:

1. A novel 2-stage learning method to smoothly re-label noisy multi-label datasets of CXR images that can then be used for training a common multi-label classifier;
2. A new bag of multi-label image descriptors learning method that leverages the semantic information available from language models to represent multi-label images and to detect noisy samples;
3. A new graph structure to smoothly re-label noisy multi-label images, with each image being represented by a sub-graph of the learned multi-label image descriptors that can capture fine-grained image relationships;
4. The first systematic evaluation of noisy multi-label methods that combine the PadChest [6] and Chest X-ray 14 [57] datasets.

We show the effectiveness of our BoMD on a benchmark that consists of training with two noisy multi-label CXR datasets and testing on three clean multi-label CXR datasets [31]. Results show that our approach has more accurate classifications than previous multi-label classifiers developed for CXR datasets and noisy-label classifiers. Results on our proposed benchmark show that BoMD is generally more accurate and robust than competing methods under our systematic evaluation.

2. Related Works

2.1. CXR multi-label classification

Recently, we have seen many CXR multi-label classifiers being proposed, such as the CXR pneumonia detector [50]. Ma et al. [39] introduce a new cross-attention network to extract meaningful representations. Hermoza et al. [15] propose a weakly-supervised method to diagnose and localise diseases. Although these methods show promising results, there is still potential for improvement that can be realised by addressing the noisy multi-label learning of CXR datasets [20, 57].

2.2. Learning with Noisy Labels

Noise-cleaning methods focus on detecting noisy samples. For instance, Han et al. [13] rely on the small-loss trick (i.e., clean samples have small losses) to co-teach two models. Huang et al. [18] detect noisy samples that have unstable prediction by switching learning rates. Bahri et al. [2]

discard samples whose labels disagree with a KNN classifier prediction. Noise-cleaning methods can be combined with semi-supervised learning [4] to perform both the detection and correction of corrupted data. For example, DivideMix [28] removes the labels of samples classified as noisy and runs a semi-supervised learning method [4, 35]. FINE [25] proposes a robust method to detect noisy samples by verifying the alignment of image features and class-representative eigenvectors. Noise-cleaning methods generally employ two divergent networks to reduce confirmation bias [28, 34], which substantially increases computational complexity. Additionally, it is unclear if these methods can handle noisy multi-label problems since they do not in general capture fine-grained image relationships.

Noise-robust methods rely on robust loss functions to balance the overfitting effects caused by label noise in the training process. Early papers, such as [58], explore the symmetric property of cross-entropy (CE) loss for noise-robust learning. Zhang et al. [68] propose the combination of Mean-Absolute-Error (MAE) and cross-entropy (CE) loss to achieve a good balance between convergence and generalisation. Ma et al. [40] show that any loss function can be robust to label noise by applying a simple normalization term. Recently, [10] proposes a noise-robust Jensen-Shannon divergence (JSD) loss based on a soft transition between MAE and CE losses. Even though these methods can reduce overfitting effects, they also tend to under-fit the training data. This issue has been partially addressed by the early learning regularisation (ELR) [33] that proposes a regularization term which restricts the gradient from samples with corrupt labels. The non-volatile unbiased memory (NVUM) [31] extents ELR to noisy multi-label problems. Although promising, ELR and NVUM are challenged by the different early convergence patterns of multiple labels, which can lead to poor performance for particular label noise conditions, as shown in the experiments.

Transition matrix methods estimate the transition probability between clean and noisy labels. Goldberger et al. [12] propose a noise adaptation layer to estimate label transition. Yao et al. [65] estimate the transition matrix using an intermediate class and a factorised matrix. Xia et al. [62] estimate part-dependent transition matrix for complex noise conditions. Bae et al. [1] proposed a noisy prediction calibration method based on a transition matrix to reduce the gap between noisy prediction and clean label based on a KNN prediction.

Label-smoothing methods rely on modifying [43, 55] the sample-wise label distribution [38]. Zhang et al. [66] propose online label smoothing (OLS) which generates soft labels by considering the relationships among multiple labels. Wei et al. [59] argue that the advantage of label smoothing vanishes under a high label noise regime since the label smoothing tends to over-smooth the estimated label classification, so they propose the generalised label smoothing (GLS) [59] which uses negative smoothing values for higher noise rates. In general, label smoothing methods can underfit the training data since they tend to abandon the optimisation of hard clean-label samples [43].

Graph-based methods leverage the robustness of feature representations to discriminate between clean and noisy samples and regularise the training process. Wu et al. [60] explore the topological property of the data in the feature space to perform noise-cleaning by assuming that clean data are clustered together in this feature space, while the corrupted data are isolated. Wu et al. [61] investigate the geometric structure of the data to model predictive confidence and filter out noisy samples. Iscen et al. [21] introduce a regularisation term that forces samples to have similar predictions to their neighbors. These graph-based methods have been designed for single-label classification, so they cannot be easily adapted to multi-label datasets. Also, building a graph with multi-label data is also an issue for these methods.

Multi-label Noisy label methods have received increasing attention in recent years due to the natural differences with respect to multi-class problems. Instead of the sample-wise noise found in multi-class problems, each label per sample can be corrupted in the multi-label scenario which can be problematic for selecting and correcting the label noise. In addition, the class imbalance [31] and the semantic divergence may also exacerbate the overfitting issue towards the majority classes. Zhao et al. [69] leverage label dependencies to handle noisy labels and use word embeddings to perform context-based regularization to avoid overfitting. Li et al. [29] consider the correlation between labels (i.e., "fish" and "water" have a stronger correlation when comparing with "fish" and "sky") to estimate the transition matrix. Xie et al. [63] mitigate the negative impact of label noise by estimating the confidence for credible labels from the candidate label set. Different from previous methods, we consider using the label semantic information and label smoothing techniques to capture more fine-grained image relationships and prevent the classifier from being overconfident on any of the noisy labels.

2.3. Bag of Words

The Bag of Words (BoW) method [14, 53, 54] is a traditional information retrieval technique, denoted by the representation of documents with a histogram of unordered words. In computer vision, BoW [8, 53] represents images with a histogram of unordered local visual descriptors, learned from the training images in an unsupervised manner. We adopt the BoW concept, but instead of extracting local visual descriptors (e.g., SIFT [37]), we train a DNN to represent each image with a bag of global visual descriptors.

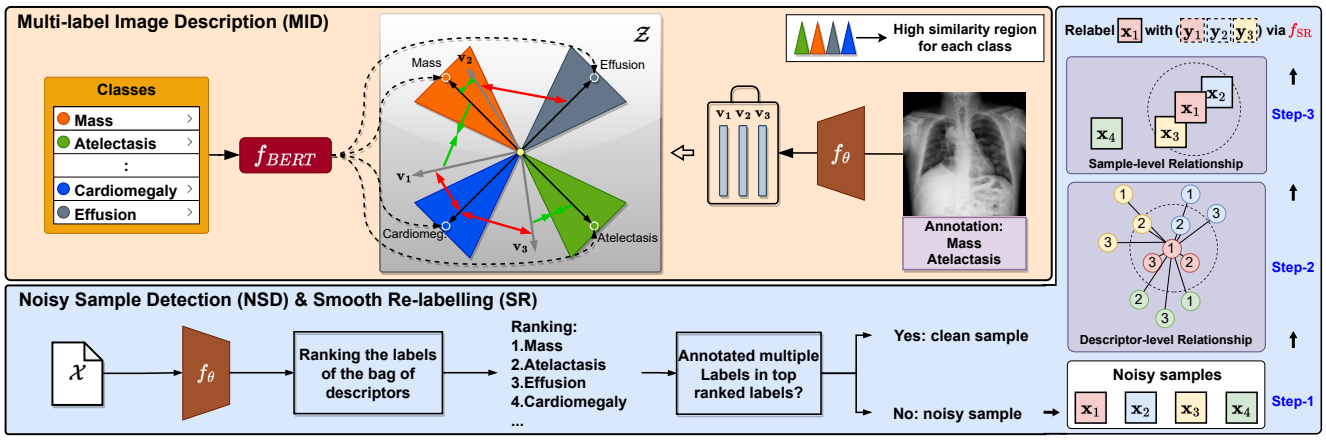


Figure 2: The **multi-label image description (MID)** module estimates a bag of visual descriptors $\{\mathbf{v}_m\}_{m=1}^M$ for each image, which is learned using the semantic information provided by BERT [11,23,48]. BERT represents each class with a descriptor (circle) in the semantic space $\mathcal{Z}$, where the triangular regions indicate high similarity regions of each corresponding class. Please note that the green arrows represent the pulling of image descriptors towards one of the corresponding relevant embeddings in the rank loss of Eq.2. Conversely, the red arrows push the image descriptors away from all irrelevant embeddings and the rectangular regions indicate the high similarity area w.r.t to each class's word embeddings. The **noisy sample detection (NSD)** module leverages the consistency between label ranking prediction from MID and the original annotation to detect the noisy samples, which are then **smoothly re-labelled (SR)** with $f_{\text{SR}}(.)$ in (6).

3. Method

We assume the availability of a noisy multi-label training set denoted by $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{H\times W\times R}$ represents an image of size $H \times W$ and R colour channels[1], and $\mathbf{y}_i \in \mathcal{Y} = \{0,1\}^{|\mathcal{Y}|}$ denotes the multi-label annotation. The testing set is similarly defined.

3.1. Bag of Multi-label Descriptors (BoMD)

Our method, described in Alg. 1, is inspired by the observation that a noisy-labelled sample tends to be an outlier surrounded by clean-labelled samples in the feature space [60]. Hence, the label of each sample should be consistent with the labels of the neighboring samples. This motivated us to develop an approach to re-label a noisy multi-label image with the estimated label distribution from its neighbourhood. The proposed BoMD has two stages (Fig. 1): 1) image description learning that transforms a training image into a bag of visual descriptors that lie in the semantic space $\mathcal{Z} \subset \mathbb{R}^Z$ populated by word embeddings computed from image labels [48] 2) graph construction to smoothly re-label noisy multi-label images, where each image is represented by a sub-graph built from the learned bag of visual descriptors, which can capture fine-grained image relationships. This smoothly re-labelled dataset is then used for training a multi-label classifier.

3.2. Multi-label Image Description (MID)

Motivated by BoW, our MID (Fig. 2) represents an image by associating its multiple labels to a bag of global visual descriptors. MID projects the image into BERT's semantic space using a set of visual descriptors that are optimised to promote their similarity with the semantic descriptors produced by BERT models from the multi-label annotation of the image. The MID of image $\mathbf{x}$ are extracted with $\mathcal{V} = f_\theta(\mathbf{x})$, where $\mathcal{V} = \{\mathbf{v}(m)\}_{m=1}^M$ denotes the M[2] visual descriptors in BERT's semantic space, i.e., $\mathbf{v}(m) \in \mathcal{Z}$. The BERT language models (e.g., BlueBERT [48], medical language model pre-trained on PubMed abstracts [11] and clinical notes [23]) produce semantic descriptors in the form of word embeddings with $\mathbf{w}(c) = f_{BERT}^{\mathcal{Y}}(c)$ for $c \in \{1,...,|\mathcal{Y}|\}$, forming $\mathcal{W} = \{\mathbf{w}(c)\}_{c=1}^{|\mathcal{Y}|}$, where $\mathbf{w}(c) \in \mathcal{Z}$, with $\mathcal{Z}$ being the same space as for $f_\theta(.)$. More specifically, MID is trained with:

$$\theta^* = \arg\min_\theta \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{D}} \omega_i \ell_{mid}(\mathbf{x}_i,\mathbf{y}_i,\theta) + \beta \ell_{reg}(\mathbf{x}_i,\theta) \tag{1}$$

where $\omega_i = \left(\left(\sum_{c=1}^{|\mathcal{C}|} \mathbb{I}(\mathbf{y}_i(c)=1)\right) \div \left(\sum_{c=1}^{|\mathcal{C}|} \mathbb{I}(\mathbf{y}_i(c)=0)\right)\right)$ is a normalisation that controls the ranking weight based on the number of positive and negative labels ($\mathbb{I}(.)$ represents an indicator function) [5,67], and the hyper-parameter

[1] We consider each image to have 3 colour channels (R=3) and dimensions of $H = W = 512$ for NIH, and $H = W = 224$ for CXP.

[2] We empirically set $M = 3$ according the ablation study of the hyper-parameters in Supp. Material.

β weights the regulariser. Also in (1) we have:

$$\ell_{mid}(\mathbf{x}_i, \mathbf{y}_i, \theta) = \sum_{\substack{p=1,\\ \mathbf{y}_i(p)=1}}^{|\mathcal{Y}|} \sum_{\substack{n=1,\\ \mathbf{y}_i(n)=0}}^{|\mathcal{Y}|} \log\left(1 + \exp\left(\ell_{rank}(\mathcal{V}_i, \mathcal{W}, p, n)\right)\right), \text{ where}$$

$$\ell_{rank}(\mathcal{V}_i, \mathcal{W}, p, n) = \left(\max_{\mathbf{v}\in\mathcal{V}_i}(\langle \mathbf{v}, \mathbf{w}(n)\rangle) - \max_{\mathbf{v}\in\mathcal{V}_i}(\langle \mathbf{v}, \mathbf{w}(p)\rangle)\right), \tag{2}$$

with $\langle .,. \rangle$ representing the dot product operator, $p, n \in \{1, ..., |\mathcal{Y}|\}$ denoting the indices to the positive (i.e., $\mathbf{w}(p)$ where $\mathbf{y}_i(p) = 1$) and negative word embeddings (i.e., $\mathbf{w}(n)$ where $\mathbf{y}_i(n) = 0$), respectively, $\mathcal{V}_i = f_\theta(\mathbf{x}_i)$, and

$$\ell_{reg}(\mathbf{x}_i, \theta) = \sum_{\mathbf{v}(m)\in\mathcal{V}_i} \frac{(\mathbf{v}(m) - \bar{\mathbf{v}}(m))^\top (\mathbf{v}(m) - \bar{\mathbf{v}}(m))}{Z - 1} \tag{3}$$

being a regulariser to reduce descriptor variance, where Z is the number of dimensions of $\mathcal{Z}$, and $\bar{\mathbf{v}}(m)$ denoting the MID mean in $\mathcal{V}_i$. The $\ell_{mid}(.)$ in (2), inspired by previous multi-label learning methods [67], forces the dot product $\langle \mathbf{v}, \mathbf{w}(p)\rangle$ to rank higher than the $\langle \mathbf{v}, \mathbf{w}(n)\rangle$, which means that the visual descriptors will be more similar to positive label embeddings than the negative label ones. Intuitively, this loss encourages semantically similar image descriptors to cluster around their related semantic descriptors, which will benefit our graph-based smooth re-labelling.

3.3. Graph Construction and Smooth Re-labelling

Considering that the learned visual descriptors are likely to be closer to clean labels in the semantic space, we formulate the detection of noisy training samples by first ranking (in descending order of similarity) the labels for image $\mathbf{x}_i$ (according to the inner product with the word embeddings from the labels), as follows:

$$\mathbf{r}_i = \arg\text{sort}_{c\in\{1,...,|\mathcal{Y}|\}} \left[\max_{\mathbf{v}\in\mathcal{V}_i}(\langle \mathbf{v}, \mathbf{w}(c)\rangle)\right], \tag{4}$$

where $\mathbf{r}_i(1) \in \{1, ..., |\mathcal{Y}|\}$ is the highest ranked label and $\mathbf{r}_i(|\mathcal{Y}|) \in \{1, ..., |\mathcal{Y}|\}$ is the lowest ranked label. Then, clean samples are the ones where $\mathbf{r}_i(p) < \mathbf{r}_i(n)$ for all positive labels $p \in \mathcal{P}_i$ and negative labels $n \in \mathcal{N}_i$, with $\mathcal{P}_i = \{c|\mathbf{y}_i(c) = 1\}$ and $\mathcal{N}_i = \{c|\mathbf{y}_i(c) = 0\}$; otherwise, the sample is classified as noisy.

The second stage of BoMD re-labels noisy samples using the sample graph built with the MID visual descriptors. The graph is constructed by representing each training image $\mathbf{x}_i$ with M nodes $\{\mathbf{v}_i(m)\}_{m=1}^{M}$ from $\mathcal{V}_i = f_\theta(\mathbf{x}_i)$, where the edge weight between the mth descriptor of the ith image and the nth descriptor of the jth image is defined by $e(\mathbf{v}_i(m), \mathbf{v}_j(n)) = 1/\|\mathbf{v}_i(m) - \mathbf{v}_j(n)\|_2$. This means that the graph has the set of nodes denoted by $\{\mathbf{v}_i(m)\}_{i=1,m=1}^{|\mathcal{D}|,M}$ and edges $\{e(\mathbf{v}_i(m), \mathbf{v}_j(n))\}_{i,j=1,m,n=1}^{|\mathcal{D}|,M}$. The re-labelling is based on finding the K nearest neighbouring training images to image i by using the graph nodes, with:

$$\mathcal{K}(\mathcal{V}_i) = \text{top}K_{j\in\{1,...,|\mathcal{D}|\},m,n\in\{1,...,M\}}(e(\mathbf{v}_i(m), \mathbf{v}_j(n))), \tag{5}$$

where $\mathcal{K}(\mathcal{V}_i)$ contains the unique image indices from the K nodes with the largest edge weights. Next, for all samples identified as noisy, we update their labels with $\tilde{\mathbf{y}}_i = f_{\text{SR}}(\mathbf{y}_i, \bar{\mathbf{y}}_i)$, which is defined as

$$f_{\text{SR}}(\mathbf{y}_i, \bar{\mathbf{y}}_i) = (1 - \lambda) \cdot \mathbf{y}_i + \lambda \cdot \left(\gamma \cdot \mathbf{1}_{|\mathcal{Y}|} + (1 - \gamma) \cdot \bar{\mathbf{y}}_i\right) \odot \mathbf{m}, \tag{6}$$

where $\lambda \in [0, 1], \gamma \in [0, 0.5]$, $\bar{\mathbf{y}}_i = \frac{1}{K}\sum_{j\in\mathcal{K}(\mathcal{V}_i)} \mathbf{y}_j$, $\mathbf{1}_{|\mathcal{Y}|}$ denotes a vector with ones of size $|\mathcal{Y}|$ (uniform distribution) to prevent the re-labelling from being overconfident on any of the labels, $\odot$ is the element-wise vector multiplication, and $\mathbf{m} = \mathbb{I}((\mathbf{y}_i + \bar{\mathbf{y}}_i) > 0)$ is a binary mask to filter out high confident negative labels (with $\mathbb{I}(.)$ being the indicator function) to mitigate the over-smoothing issue.

Algorithm 1 BoMD

require: Training set $\mathcal{D}$, and BERT model embeddings $\mathbf{w}(c) = f_{BERT}^{\mathcal{Y}}(c)$ for $c \in \{1, ..., |\mathcal{Y}|\}$
Build MID
Train $f_\theta(.)$ with (1) using $f_{BERT}^{\mathcal{Y}}(.)$ and $\mathcal{D}$
Create re-labeled set $\tilde{\mathcal{D}} = \emptyset$ and noisy set $\mathcal{B} = \emptyset$
# Detect noisy samples
for $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}$ **do**
  Compute label rank $\mathbf{r}_i$ from (4)
  If $\mathbf{r}_i(p) > \mathbf{r}_i(n)$ for all $p \in \mathcal{P}_i$ and all $n \in \mathcal{N}_i$
  Then $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \bigcup (\mathbf{x}_i, \mathbf{y}_i)$
  Else $\mathcal{B} \leftarrow \mathcal{B} \bigcup (\mathbf{x}_i, \mathbf{y}_i)$
end for
Build graph with nodes $\{\mathbf{v}_i(m)\}_{i=1,m=1}^{|\mathcal{D}|,M}$ and edges $\{e(\mathbf{v}_i(m), \mathbf{v}_j(n))\}_{i,j=1,m,n=1}^{|\mathcal{D}|,M}$ using $f_\theta(.)$ and $\mathcal{D}$
# Re-label noisy samples
for $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{B}$ **do**
  Re-label $\mathbf{x}_i$ with $\tilde{\mathbf{y}}_i$ from (6)
  $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \bigcup (\mathbf{x}_i, \tilde{\mathbf{y}}_i)$
end for
# Train final classifier
Train $f_\phi : \mathcal{X} \rightarrow [0, 1]^{|\mathcal{Y}|}$ with BCE loss using $\tilde{\mathcal{D}}$

3.4. Training and Testing

We build a new training set $\tilde{\mathcal{D}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)|(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}\}_{i=1}^{|\mathcal{D}|}$, where $\tilde{\mathbf{y}}_i = \mathbf{y}_i$ if sample $(\mathbf{x}_i, \mathbf{y}_i)$ is clean from Eq. (4), or computed from Eq. (6) if sample is noisy}. Then, we train a regular classifier $f_\phi : \mathcal{X} \rightarrow [0, 1]^{|\mathcal{Y}|}$ by minimizing a BCE loss on $\tilde{\mathcal{D}}$. Testing is based on applying the trained $f_\phi(.)$ to test images.

4. Experiments

Our experiments are based on the following datasets. **Noisy Training Sets.** The ***NIH Chest X-ray14*** **(NIH)** [57] contains 112,120 frontal-view CXR images from 30,805 patients, where each image has between 0 and 14 annotated pathologies and the training set contains 86,524 images with a maximum of 9 labels per image. The ***CheXpert (CXP)*** [20] has 224,316 frontal-view CXR images from 65,240 patients labelled with 14 common chest radiographic observations, where the training set contains 170,958 images with a maximum of 8 labels per image. The labels of these two datasets are obtained from an NLP algorithm, which forms noisy multi-label annotations [44].

Clean Testing Sets. The ***OpenI*** [9] dataset contains 3,999 radiology reports and 7,470 frontal/lateral-view CXR images from the Indiana Network for Patient Care. We use all frontal-views images for evaluation, resulting in 3,818 images and 19 manually annotated diseases. We also use ***PadChest*** [6], which contains 160,861 images with 27 chest radiographic observations. PadChest has a mixture of machine and manually labelled images, but we only use the manually labelled frontal-view images (about 15.25%[3] of the images). Additionally, we follow previous works [31, 64] to evaluate the model on a re-organised subset of the official NIH test set [57], referred to as ***NIH-GOOGLE***, with 1,962 CXR images that focuses on two findings (pneumothorax and nodule/mass). Each image is manually re-labelled by at least three certified radiologists [41] with final label pooled from an adjudication process.

Systematic Noisy-label Assessment. We introduce the first systematic noisy multi-label assessment benchmark by combining the NIH [57] dataset and the clean test samples from Padchest [6] as a new noisy-label training set with the OpenI [9] as the clean testing set. We then apply symmetric label noise [1, 28] to the PadChest training subset **only**, where we flip the labels from present to absent, and vice-versa, based on two control variables, namely: 1) the proportion of noisy samples, and 2) the probability of switching a label. This dataset is referred to as ***NIHxPDC***.

4.1. Implementation Details

We resize the NIH [57] images to 512×512, and CXP [20] images to 224×224, where images are normalised using ImageNet [52] mean and standard deviation. We use random resize crop and random horizontal flipping as data augmentation. The BlueBERT word embeddings in $\mathcal{W}$ are L2 normalised. For the MID model $f_\theta(.)$, we use the ImageNet [52] pre-trained DenseNet121 [17], which is trained with Adam optimiser [26] using a learning rate of 0.0001 with cosine annealing decay [36], batch size of 16 and 20 epochs. We set the number of descriptors $M = 3$ and weight of regulariser $\beta = 0.3$. The descriptor graph is implemented with Faiss [24] for efficient search (the search process in Eq. (5) takes 5 seconds for all training samples from NIH). The classifier $f_\phi(.)$ uses another ImageNet pre-trained DenseNet121 [17], and then it is trained with Adam optimiser [26] using a learning rate of 0.05, batch size of 16 and 30 epochs. We empirically set the mixup coefficient λ and γ in Eq. (6) to 0.6 and 0.25, and use $K = 10$ in Eq. (5). The classification results are assessed with the mean of the class-wise area under the receiver operating characteristic curve (AUC) for all disease classes [15, 20, 32]. All experiments are implemented with Pytorch [46] and run on an NVIDIA RTX3090 GPU (24GB). Training takes 23h for NIH and 15h for CheXpert, and testing for a single image takes 13.41ms for NIH and 12.24ms for CheXpert.

[3]The 27% of PadChest's annotated images include the lateral-view CXRs.

Methods	Models	OpenI	PadChest
General	Hermoza et al [15]	85.54 ± 0.42	83.90 ± 0.57
	CAN [39]	84.26 ± 0.35	83.10 ± 0.25
Noise-cleaning	DivideMix [28]	72.76 ± 1.09	75.49 ± 0.21
	FINE [25]	63.67 ± 1.78	70.91 ± 0.20
Noise-robust	ELR [33]	86.62 ± 0.87	85.24 ± 0.11
	NVUM [31]	**88.17 ± 0.48**	**85.49 ± 0.06**
Transition matrix	NPC [1]	86.21 ± 0.07	83.88 ± 0.05
Graph-based	NCR [21]	85.06 ± 0.96	83.79 ± 0.48
Label smoothing	LS [38]	83.72 ± 1.29	80.93 ± 0.82
	OLS [66]	85.08 ± 0.31	83.51 ± 0.61
	GLS [59]	83.80 ± 0.34	81.56 ± 0.24
	BoMD	**89.57 ± 0.22**	**86.45 ± 0.08**

Table 1: Mean ± standard deviation AUC results for the testing sets from OpenI and PadChest, using models **trained on NIH [57]**. Best and the second best results are in **red**/**blue**.

4.2. Classification Results on Real-world Datasets

We first compare the performance of state-of-the-art (SOTA) methods with our BoMD in Tab. 1 and Tab. 2. We run each experiment three times and show mean and standard deviation of AUC results. Tab. 1 shows the testing AUC results of the training with NIH [57] and testing on OpenI [9] and PadChest [6]. Our model surpasses the second best method [31] by 1.4% and 0.96% on the two test sets with p-values 0.0018 and 10^{-14}, respectively (one-sided z-test). We also report testing performance on OpenI [9] and PadChest [6] using training on CXP [20] in Tab. 2. Our model surpasses the second best result [20] by 2.82% and 1.13% on the two test sets with p-values 0.0002 and 10^{-14}, respectively (one-sided z-test). We also notice that there is a large gap between all models' performance for certain classes. For example, in our model, the AUC results for *Pneumonia* classification when training on NIH are much better than when training on CXP, with a gap of 23.20% and 15.48% on OpenI [9] and PadCh-

Methods	Models	OpenI	PadChest
General	Hermoza et al [15]	74.94 ± 0.50	77.24 ± 0.04
	CAN [39]	76.34 ± 0.49	78.92 ± 0.58
Noise-cleaning	DivideMix [28]	73.23 ± 0.60	74.21 ± 0.55
	FINE [25]	71.68 ± 0.54	73.83 ± 0.52
Noise-robust	ELR [33]	77.16 ± 0.79	79.97 ± 0.71
	NVUM [31]	**77.21 ± 0.81**	**80.62 ± 0.10**
Transition matrix	NPC [1]	75.32 ± 0.40	77.30 ± 0.03
Graph-based	NCR [21]	76.93 ± 0.38	79.36 ± 0.84
Label Smoothing	LS [38]	72.86 ± 0.23	75.34 ± 0.50
	OLS [66]	76.52 ± 0.83	77.72 ± 0.70
	GLS [59]	76.50 ± 0.26	78.80 ± 0.70
	BoMD	**80.03 ± 0.73**	**81.76 ± 0.40**

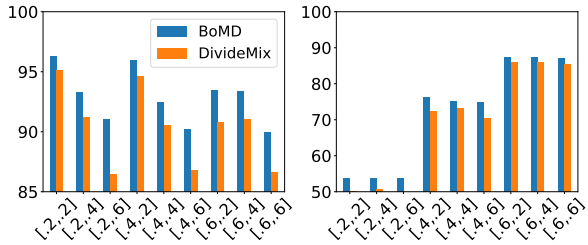
Table 2: Mean ± standard deviation AUC results for the testing sets from OpenI and PadChest, using models **trained on CXP [20]**. Best and the second best results are in **red**/**blue**.

est [6], respectively, which may be due to CXP's smaller image size. The NIH-GOOGLE [31, 64] evaluation on classes pneumothorax and nodule/mass (obtained from the average classification scores for Mass and Nodule) is displayed in Tab. 3, which shows that our method outperforms the SOTA methods on both Pneumothorax (+0.6% compared to [64]) and Mass/Nodule (+2.4% compared [31]) classifications. The per-finding results are reported in the *Supplementary Material*.

4.3. Systematic Noisy-label Benchmark

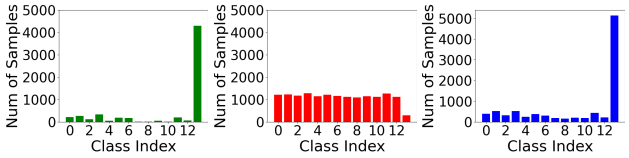
In our systematic noisy-label benchmark NIHxPDC, we compare our BoMD with the SOTA method NVUM [31] and a baseline model trained with BCE loss. Recall that this benchmark relies on two control variables: a) percentage of noisy samples p_s, and b) probability of switching a label p_l, where $p_s, p_l \in \{0\%, 20\%, 40\%, 60\%\}$. We show the mean AUC classification over the 14 classes on clean OpenI in Tab. 4. Notice that our BoMD has better results than NVUM and BCE for the majority of the cases, with a 3% to 5% improvement compared with NVUM and BCE when $p_s = 20\%$, 1% to 4% improvement for $p_s = 40\%$, and 1% to 5% improvement for $p_s = 60\%, p_l = 20\%, 40\%$, but for $p_s, p_l = 60\%$, NVUM improves 2.8% over BoMD. Hence, BoMD provides a consistently better classification result across many noise rates, but for large noise rates, our re-labelling method may be injecting too much noise into the training set. Another interesting point to note is that with 0% controlled noise, BoMD shows 89.76% AUC, which means that the AUC drops an average of 3.8% for each addition of 20% for p_s with a fixed $p_l = 20\%$.

We now study the effectiveness of the detection and re-labelling of noisy samples by BoMD. In Fig. 3, we compare the precision and recall of our detection of noisy-label samples compared with the traditional small-loss approach used by DivideMix [28]. Notice that our noisy-label sample



(a) Recall (b) Precision

Figure 3: Noisy-label sample detection performance on NIHxPDC. We compare our proposed rank-based detection approach with DivideMix's small loss method [28] for a) recall, and b) precision. The horizontal axes show the values for $[p_s, p_l]$.



(a) Clean (b) Noise (c) Re-labelled

Figure 4: Visualisation of the changes in the histogram of label distributions after applying our re-labelling.**Left**: label distribution for the clean set OpenI. **Middle**: label distribution after injecting symmetric noise $p_s = 0.4, p_l = 0.2$. **Right**: label distribution after the re-labelling by BoMD.

detection consistently outperforms DivideMix's small-loss method on both measures. An interesting note is that while recall worsens, precision improves with increasing noise rates. This happens because of the natural imbalance found in the distribution of classes in CXR datasets, where the class "No Findings" is dominant. A larger synthetic noise rate implies that this class is more affected than the others, but at the same time easier to detect given that the NIH dataset has relatively smaller noise rates.

We also visualise in Fig. 4 the label distribution change before and after applying our re-labelling. Note that our method successfully corrects the noisy label distribution to be closer to the original clean label distribution. The mean AUC over the labels in the re-labelled dataset before and after our re-labelling process is presented in Fig. 5, where results show that our re-labelling process significantly improves the label cleanliness of the training set for all benchmark noise rates. Recall that in multi-class problems, such re-labelling is facilitated by the fact that each sample can only have a single label. However, such constraint is dropped for multi-label problems, making the re-labelling more complicated because the feature space will be populated with multiple clusters containing different combinations of multi-labels. In the supplementary material, we evaluate the amount of consistency the KNN neighboring samples need to have for a clean re-labelling. This eval-

	BCE	F-correction [47]	MentorNet [22]	Decoupling [42]	Co-teaching [13]	ELR [33]	Xue et al. [64]	NVUM [31]	BoMD
Pneu	87.0	80.8	86.6	80.1	87.3	87.1	**89.1**	88.9	**89.7**
M/N	84.3	84.8	83.7	84.3	82.0	83.2	84.6	**85.5**	**87.9**

Table 3: Pneumothorax and Mass/Nodule AUC of NIH-Google [41] for models trained on NIH [57]. Best and the second best results for OpenI and PadChest are in **red**/**blue**.

p_s	0%	20%			40%			60%		
p_l	0%	20%	40%	60%	20%	40%	60%	20%	40%	60%
BCE	85.65	83.99	81.42	79.63	80.99	78.51	75.79	77.14	75.35	72.39
NVUM	87.89	85.34	82.83	81.35	82.52	80.64	78.66	78.49	77.19	**76.91**
BoMD	**89.76**	**88.00**	**86.26**	**84.55**	**84.47**	**81.86**	**78.68**	**82.23**	**78.15**	74.11

Table 4: Mean testing AUC results for the 13 OpenI classes with models trained on NIHxPDC. Best results in **red**.

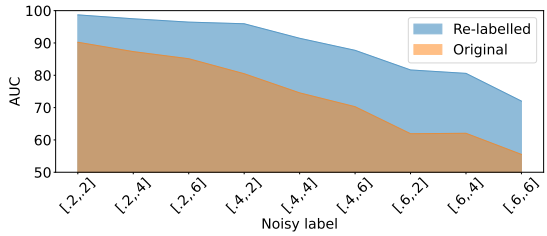


Figure 5: Mean AUC over labels before (orange) and after (blue) our re-labelling w.r.t PadChest's clean labels. The horizontal axes show values for $[p_s, p_l]$.

uation is based on the label-wise precision and recall results of our graph-based re-labelling method as a function of a threshold on the minimum number of nearest neighbors containing the same label. Results suggest that precision and recall increase until this threshold is between 4 and 6 nearest neighbours, and plateaus afterwards. Hence, when 4 to 6 neighbours (depending on the noisy rate) share a particular label, it is probable that the noisy training sample has this clean label.

4.4. Ablation Study

Language models. Our ablation study starts with an investigation of the language models for our BoMD, where we consider three types of models: 1) a randomly initialised model (without relying on language models); 2) a computer vision language model (e.g., Common Crawl data [4]); and 3) a medical language model (e.g., BioBERT[5], ClinicalBERT[6], BlueBERT[7]). In Tab. 5, the BERTs box shows that BERT models enable better performance than other models, with gains from 2.50% to 1.50% on OpenI and PadChest. We argue that this is because the word embeddings from BERT models contain relevant clinical semantic meaning (e.g., Atelectasis and Pneumonia are both correlated to lung opacity, but uncorrelated with Enlarged Cardiomediastinal [20]) that facilitates the multi-label descriptor learning of our method. Among the models trained on

[4] https://commoncrawl.org
[5] Biomedical language model, pretrained on PubMed [27].
[6] MIMIC corpus (FT on BioBERT) [19].
[7] Pretrained on PubMed abstract + MIMIC-III (clinical notes) [48].

Ablation Study	M	Language Models	Open-i	PadChest
BERTs	3	Random Init.	87.02	84.99
	3	Glove [49]	87.62	85.08
	3	ClinicalBERT [19]	88.27	85.72
	3	BioBERT [27]	89.11	86.27
	3	BlueBERT [48]	**89.52**	**86.50**
Stage-one training	1	Self-supervised [70]	84.50	83.21
	1	NVUM [31]	86.69	84.66
	1	MID	88.34	86.02
	3	MID	**89.52**	**87.50**

Table 5: Ablation study that compares the mean testing AUC results of our BoMD with the use of different language models (BERTs) and descriptor training (Stage-one training), M shows the number of descriptors per image.

$\mathbf{1}$	$\bar{\mathbf{y}}$	$\mathbf{m}$	OpenI	PadChest
✓			83.72	80.93
✓	✓		87.92	85.48
	✓	✓	89.11	86.27
✓	✓	✓	**89.52**	**86.50**

Table 6: Ablation study of the testing AUC results of the components of our re-labelling in (6). $\bar{\mathbf{y}}$ indicates the KNN propagated label, $\mathbf{1}$ is the uniform distribution, and $\mathbf{m}$ is the binary mask.

BERT models, we observe small variations, which can be related to: 1) the size of the training set, and 2) the relatedness of the medical dataset to our CXR classification problem (BlueBERT is arguably more related than BioBERT).

Evaluation of MID. In Tab. 5 the Stage-one training box shows a study of the effectiveness of MID for graph construction and downstream classification task, by comparing it against the use of the descriptors from NVUM [31]. Given that NVUM produces one descriptor per image, we set MID's number of descriptors per image at $M = 1$ for fairness. Results from the table show that MID descriptors allow an improvement of between 2% and 3% compared to NVUM's descriptors. In addition, $M = 1$ represents a graph with one (instead of multiple) descriptor per image, where we can observe a 1% to 2% drop, which indicates that our aggregating sub-graph is a more suitable strategy for multi-label images.

Smoothly re-labelling. We show an ablation study of the mixup terms in Eq. (6) in terms of the testing AUC results in Tab. 6. First, we mix up $\mathbf{y}$ and the uniform distribution $\mathbf{1}$ (i.e., label smoothing) with a fixed mixup coefficient of $\lambda = 0.6$ (first row of Tab. 6), then we introduce $\bar{\mathbf{y}}$ with another fixed mixup coefficient of $\gamma = 0.25$ (second row of Tab. 6) and observe improvements of over 4%. Next, we

remove $\mathbf{1}$ and add the binary mask $\mathbf{m}$ (third row of Tab. 6) to filter out the confident negative labels, which increases the performance by around 1%. The integration of all re-labelling components further increases performance from 0.23% to 0.41%. These results suggest that the mask $\mathbf{m}$ combined with the KNN average label $\bar{\mathbf{y}}$ mitigate the over-smoothing promoted by the uniform distribution $\mathbf{1}$.

Pre-training with vision-language model [70]. CoOp [70] is an effective visual-textual pre-training that can be considered for improving any of the general methods in Tab. 1 and 2. Hence, we pre-trained with CoOp the method in [15] and noticed a 1.04% and 0.69% performance drop in Tab. 1 and 2. This drop can be explained by the fact that CoOP requires backpropagation for the context tokens of the labels during training, where noisy labels may have caused the learned token to be inaccurate.

Additional results. In the supplementary material, we include a visualisation of different label smoothing techniques using a t-SNE [56] graph for a toy problem, additional evaluation for descriptors extracted by the MID module, and detailed sensitivity testing of hyper-parameters.

5. Discussion and Conclusion

In this work, we proposed BoMD, a new method to learn from noisy multi-label CXR datasets. BoMD explores the clinical semantic information, represented by word embeddings from BlueBERT [48], to optimise the multi-label image descriptors which are used to find noisy multi-label training samples. We then use the learned image descriptors to build a graph for smoothly re-labelling the training data. BoMD outperforms current SOTA methods on three real-world CXR benchmarks that consist of training on two large-scale noisy multi-label CXR datasets and testing on three clean multi-label CXR datasets. We additionally evaluate BoMD on our proposed systematic benchmark to further show the effectiveness and robustness of our method.

Limitations and future work. We identify three limitations of BoMD. The first issue is the longer training time (+8h compared with NVUM [31]) since it requires multiple training stages. We plan to tackle this problem by better integrating the training stages. The second issue is that BoMD decreases its performance under extremely noisy label setup (i.e., [0.6, 0.6] in Tab. 4), which is due to mistakes in the smooth re-labelling. However, such a high noise rate may not be applicable in real-world scenarios since the usual F1 score for text mining performance is between 80% to 94% [9,44,57], which suggests noise rates much smaller than 60%. Another drawback of BoMD is that it does not address imbalanced learning, which is an important point when training with CXR datasets. BoMD demonstrates the possibility of leveraging semantic information and sample graphs to estimate the label distribution for training a better CXR classifier. However, noisy-cleaning methods are still dominant in the LNL under multi-class scenarios. One potential future direction is to study a unified framework for addressing both multi-class and multi-label tasks with minimum adaptation.

6. Acknowledgement

This work was supported by funding from the Australian Research Council through grant FT190100525.

References

[1] HeeSun Bae, Seungjae Shin, Byeonghu Na, JoonHo Jang, Kyungwoo Song, and Il-Chul Moon. From noisy prediction to true label: Noisy prediction calibration via generative model. In *International Conference on Machine Learning*, pages 1277–1297. PMLR, 2022. 1, 2, 3, 6, 7

[2] Dara Bahri, Heinrich Jiang, and Maya Gupta. Deep k-nn for noisy labels. In *International Conference on Machine Learning*, pages 540–550. PMLR, 2020. 2

[3] Avi Ben-Cohen, Nadav Zamir, Emanuel Ben-Baruch, Itamar Friedman, and Lihi Zelnik-Manor. Semantic diversity learning for zero-shot multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 640–650, 2021. 1, 2

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[5] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005. 4

[6] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020. 2, 6, 7

[7] Guillaume Couairon, Matthijs Douze, Matthieu Cord, and Holger Schwenk. Embedding arithmetic of multimodal queries for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4950–4958, 2022. 2

[8] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. *Workshop on statistical learning in computer vision, ECCV*, 1(1-22):1–2, 2004. 3

[9] 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, 2016. 6, 9

[10] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021. 3

[11] Nicolas Fiorini, Robert Leaman, David J Lipman, and Zhiyong Lu. How user intelligence is improving pubmed. *Nature biotechnology*, 36(10):937–945, 2018. 4

[12] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. *4th International Conference on Learning Representations (ICLR)*, 2016. 2, 3

[13] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 2, 8

[14] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954. 3

[15] Renato Hermoza, Gabriel Maicas, Jacinto C Nascimento, and Gustavo Carneiro. Region proposals for saliency map refinement for weakly-supervised disease localisation and classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 539–549. Springer, 2020. 1, 2, 6, 7, 9

[16] Hexiang Hu, Wei-Lun Chao, and Fei Sha. Learning answer embeddings for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5428–5436, 2018. 2

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6

[18] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3326–3334, 2019. 2

[19] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019. 2, 8

[20] Jeremy Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, volume 33, pages 590–597, 2019. 1, 2, 6, 7, 8

[21] Ahmet Iscen, Jack Valmadre, Anurag Arnab, and Cordelia Schmid. Learning with neighbor consistency for noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4672–4681, 2022. 1, 2, 3, 6, 7

[22] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018. 8

[23] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. 4

[24] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 6

[25] Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24137–24149, 2021. 2, 3, 6, 7

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] 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. 2, 8

[28] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 1, 2, 3, 6, 7

[29] Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. Estimating noise transition matrix with label correlations for noisy multi-label learning. *Advances in Neural Information Processing Systems*, 35:24184–24198, 2022. 2, 3

[30] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[31] Fengbei Liu, Yuanhong Chen, Yu Tian, Yuyuan Liu, Chong Wang, Vasileios Belagiannis, and Gustavo Carneiro. Nvum: Non-volatile unbiased memory for robust medical image classification. *arXiv e-prints*, pages arXiv–2103, 2021. 2, 3, 6, 7, 8, 9

[32] Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. *arXiv preprint arXiv:2111.12918*, 2021. 6

[33] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020. 1, 2, 3, 6, 7, 8

[34] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4267, 2022. 3

[35] Yuyuan Liu, Yu Tian, Chong Wang, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Translation consistent semi-supervised segmentation for 3d medical images. *arXiv preprint arXiv:2203.14523*, 2022. 3

[36] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[37] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 3

[38] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise?

In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020. 2, 3, 6, 7

[39] Congbo Ma, Hu Wang, and Steven CH Hoi. Multi-label thoracic disease image classification with cross-attention networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 730–738. Springer, 2019. 1, 2, 6, 7

[40] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pages 6543–6553. PMLR, 2020. 3

[41] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020. 6, 8

[42] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". *Advances in Neural Information Processing Systems*, 30, 2017. 8

[43] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019. 3

[44] Luke Oakden-Rayner. Exploring the chestxray14 dataset: problems. *Wordpress: Luke Oakden Rayner*, 2017. 1, 6, 9

[45] Luke Oakden-Rayner. Exploring large-scale public medical image datasets. *Academic radiology*, 27(1):106–112, 2020. 1

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[47] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017. 8

[48] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019. 2, 4, 8, 9

[49] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 8

[50] 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. 1, 2

[51] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021. 2

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[53] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003. 3

[54] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):591–606, 2008. 3

[55] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[56] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 9

[57] 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. 1, 2, 6, 8, 9

[58] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019. 3

[59] Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. To smooth or not? when label smoothing meets noisy labels. *Learning*, 1(1):e1, 2021. 2, 3, 6, 7

[60] Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. A topological filter for learning with label noise. *Advances in neural information processing systems*, 33:21382–21393, 2020. 2, 3, 4

[61] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: a unified framework for learning with open-world noisy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 62–71, 2021. 2, 3

[62] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33:7597–7610, 2020. 2, 3

[63] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3676–3687, 2021. 3

[64] Cheng Xue, Lequan Yu, Pengfei Chen, Qi Dou, and Pheng-Ann Heng. Robust medical image classification from noisy labeled data with global and local representation guided co-training. *IEEE Transactions on Medical Imaging*, 41(6):1371–1382, 2022. 6, 7, 8

[65] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33:7260–7271, 2020. 2, 3

[66] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30:5984–5996, 2021. 2, 3, 6, 7

[67] Yang Zhang, Boqing Gong, and Mubarak Shah. Fast zero-shot image tagging. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5985–5994. IEEE, 2016. 2, 4, 5

[68] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 3

[69] Wenting Zhao and Carla Gomes. Evaluating multi-label classifiers with noisy labels. *arXiv preprint arXiv:2102.08427*, 2021. 3

[70] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 8, 9