Deep Learning Towards Robustness in Medical Images

EE 797 : MTP Phase - 1

submitted in partial fulfillment of the requirements
for the degree of

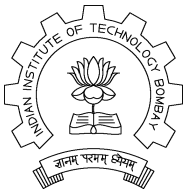
Master of Technology

by

Varsha S
Roll No: 193079005

under the guidance of

Prof. Amit Sethi



Department of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai - 400076.

2021

Abstract

Deep learning models underperform when the quality of supervisory labels degrades. Traditional methods focused on addressing the performance drop caused by closed-set label noise, neglecting the presence of a more general open-set label noise. The natural occurrence of before-mentioned label noise samples in medical imaging modalities necessitates the development of more novel training strategies that remain robust to training set contamination. We develop a unified framework utilizing a SimCLR framework and feature aggregating memory banks to re-emphasize clean training samples. We conduct extensive experiments on two publicly available datasets - BACH and TCGA-BRCA. Our method demonstrates superior results on experiments performed to validate the robustness in a general label noise scheme.

Second part of this document showcases work done in removing speckle noise in Ultra sound images. Ultrasound imaging is vital tool for diagnosis, it provides in non-invasive manner the internal structure of the body to detect diseases or abnormalities tissues. The speckle noise in this imaging process will be mixed with effective information, thus reducing the image quality and affect the diagnosis. Therefore, it is of great significance to study the denoising method of medical ultrasound images. Our experiments include use of super-resolution and wavelet transforms to de-speckle satellite images which are equally affected by speckle noise.

Acknowledgments

I express my sincere gratitude towards my guide **Prof. Amit Sethi** for his constant help, encouragement and inspiration throughout the project work. Without his invaluable guidance, this work would never have been a successful one.

I would like to thank the research scholars **Nikhil Cherian Kurian** and **Sahar Almahfouz Nasser** and for their advice and guidance in techniques and implementations used in this work and for their valuable suggestions and helpful discussions.

I would like to thank **Akshay Bajpai** who worked with me on a number of experiments of this work.

Last, but not the least, I would like to thank the whole MeDAL family for always being helpful in the times of need.

Varsha S
IIT Bombay
Oct 17, 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction : Robust Deep Learning Framework to address General Label Noise in Medical Imaging

Deep learning (DL) has been able to empower supervised medical image analysis largely due to the availability of high quality labeled data [1]. However, preparation of accurately labeled medical image datasets is challenging and a certain level of label noise is to be expected because the labels given by experts have a certain level of subjectivity. Moreover, several disease classes are based on arbitrary discretization of underlying disease states lying in continuum [2]. Furthermore, the affected region in a large medical image might be very small. Additionally, the low image quality in such modalities due to tissue preparation or preservation artifacts can also degrade the supervisory signal to a DL algorithm [3].

Such practical challenges can lead to both closed-set label noise (mislabeling among a known set of classes) and open-set label noise (mislabeling a sample from an unknown class as one from the known set of classes) in the training data (Figure 2.1). DL models underperform when the quality of supervisory labels degrades. Hence, developing DL based pipelines that are robust to both open-set and closed-set label noise problems are becoming increasingly important.

Training DL models in the face of closed-set label noise has been a subject of growing interest [4, 5]. However, only few works have addressed open-set label noise and have given a unified treatment to both types of label noise [6, 7], We propose a unified framework to train DL models that is robust to both open and closed-set label noise.

Our pipeline consists of a warm-up phase that incorporates contrastive learning on each mini batch, after a weight update based on cross-entropy to aggregate the probable clean features into a class-specific memory bank. At the end of the warm-up phase, we reweight the training samples based on their similarity to the features stored in the memory bank to emphasize clean data points.

We conducted experiments on BACH dataset [8] with synthetically introduced label noise, and on TCGA-BRCA dataset where the label noise occurs as a natural consequence of intra-tumor heterogeneity and other artifacts in whole slide images. Our experiments affirm the benefits of using robust training procedures, especially in histology datasets where the supervision is intrinsically error-prone.

1.1 Related Work

Deep learning pipelines addressing open-set label noise are receiving widespread attention considering the practical implications of the problem in a real world setting. Conventionally, the techniques addressing closed-set label noise most often utilizes the small loss trick of a cross-entropy loss function in-order to segregate noisy labeled samples. Raghav et al has re-purposed such schemes in a general label noise additionally utilizing the principles of semi-supervision and model uncertainties. Wang et al [6] used an iterative technique to detect noisy labels using a probabilistic and cumulative Local Outlier Factor (pcLOF) score while learning deep discriminative features. However, these methods doesn't perform well on medical images owing to high intra-class heterogeneity present in these images. In metric learning framework,a method addressing label noise by learning robust class prototypes was reported in [9]. They utilised a first in first out (FIFO) memory bank to aggregate clean samples based on the similarity score with the features extracted from the previous iterations.

Recently, self-supervised contrastive learning frameworks are widely adopted as robust initalization that can significantly improve state of art results using fewer labels. The robustness of such models in the label noise settings are also reported in literature. SimCLR models, tailor made for histology images have been developed in [10].

Very few works have addressed the issue of open-set noise in histology images. A sample-specific generalized cross entropy has been proposed in [3] as an alternative to conventional cross entropy. Though this loss effectively mitigates the degradation due to closed-set label noise, it does not explicitly address the issues of open-set label noise.

Chapter 2

Data & Dataset Preparation

2.1 Datasets

The datasets - BACH, Kather, TCGA are considered for our experiments.

2.1.1 BACH Dataset

The BACH challenge microscopy dataset is composed of microscopy images annotated image-wise by two expert pathologists from the Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP) and from the Institute for Research and Innovation in Health. The dataset is composed of 400 training and 100 test images, with the four classes equally represented. The provided images are on RGB .tiff format and have a size of 2048×1536 pixels and a pixel scale of 0.42 μm×0.42 μm [11]. The dataset has H&E stained breast histology microscopy images in four classes:

1. Normal
2. Benign
3. InSitu Carcinoma
4. Invasive Carcinoma

2.1.2 TCGA Dataset - Basal vs Luminal A Classification

For our second set of experiments, we chose the basal vs Luminal A PAM50 subtype classification from H&E stained WSI. A total of 130 WSI, 65 from each class, were sampled from the TCGA-BRCA dataset. We followed a patch-based classification approach in which potential tumour regions from the WSI were identified and annotated by a pathologist. As it is done conventionally, all the patches inherited the same slide level labels. The presence of intra-tumour heterogeneity naturally cast this classification problem into a noisy labeled classification. In addition to the dominant subtype, other vestigial subtypes of breast cancer could also be present in the tissue images. The occurrence of such intra-tumour heterogeneity along with other image degradation that weakens the correlation with the supervisory label, reframes this problem to the domain of open-set noises. Further, though the Luminal A dataset majority consisted of ductal morphology, some samples were listed as morphologically as Lobular further contaminating the dataset as OOD samples. For the experiments patches of size 512×512 were selected at 40x magnification. A batch size of 64 was used.

2.1.3 Kather Dataset

The dataset is collected from the Institute of Histological Images of Pathology of Human Colorectal Cancer taken from the pathology archive by Kather, et al.[12]. The dataset consists of 5000 non-duplicated histological images of human colorectal cancer (CRC) using hematoxylin and eosin (HE) and healthy normal tissue images. This dataset created images with 150 $\times$ 150 pixels (74 $\times$ 74 μm) each for every RGB color, and contains eight different tissue texture features and original tissue images with a size of 5000 pixel. The eight classes are categorized as following:

1. TUMOR: tumor is an abnormal new growth of cells.
2. STROMA: stroma is the part of a tissue or organ with a structural or connective role.
3. COMPLEX: complex stroma contain a single or a few tissue cells.
4. LYMPHO: lymphoma is a group of blood malignancies that develop from lymphocytes.
5. DEBRIS: debris or HE stain is one of the principal tissue stains used in histology.
6. MUCOSA: mucus is produced by many tissues in the body, and acts as a protective force.
7. ADIPOSE: adipose tissue is mainly composed of adipocytes.
8. EMPTY: histological image background.

2.2 Dataset Preparation

Following is a brief explanation of various data augmentations used in the training experiments discussed in this work:

- Random Horizontal Flips: Random horizontal/left-to-right flip of images with 50% probability.
- Color Jitter: Color distortion is composed by color jittering. It involves randomly changing the brightness, contrast, saturation and hue of an image.

2.2.1 Noise addition - BACH

The dataset is contaminated with open-set noise from the class 'Invasive'. The dataset is contaminated with open-set noise from two classes - 'DEBRIS', 'MUCOSA'. the open-set noise levels are fixed to 15 amd 20 per class. The label noise levels are fixed to 10, 14 and 18 per class.

2.2.2 Noise addition - Kather

The dataset is contaminated with open-set noise from two classes - 'DEBRIS', 'MUCOSA'. the open-set noise levels are fixed to 30% and 35% per class. The label noise levels are fixed to 30%, 40% and 50% per class.

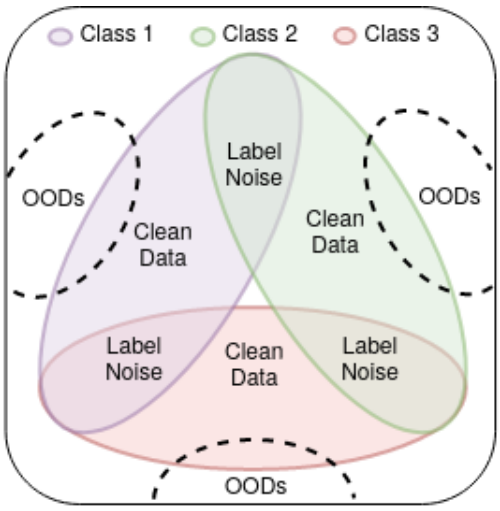


Figure 2.1: Noisy Dataset visualisation for 3 clean classes, consisting open-set (label) noise and closed-set (OOD) noise

Chapter 3

Method

Our framework primarily relies on the robust nature of SimCLR based pretraining on data that includes noisy labels (closed-set) and out-of-distribution (OOD) samples (open-set) [13]. We use a histology-specific ResNet-18 SimCLR model [14] as a backbone network for the classification task. Further, we adopt class specific FIFO memory banks as a feature aggregation module in the early stages of training to mine samples that have maximum cosine similarities within a mini-batch. Our training proceeds via an initial warm-up stage that consist of two phases – cross-entropy based weight update followed by a supervised contrastive loss based training. With this interleaved training procedure, we encourage cross-entropy to build class-specific features, while a supervised contrastive framework identifies class-wise outliers or noisy labeled samples. At the end of warm-up, the FIFO memory bank is expected to collect clean samples and their augmented replicas that help us in temporal ensembling of the data. We re-weight each training sample class-wise based on its cosine similarity to the sample present in the memory bank. Further, for more robustness we extend the weight calculation step using a K-medoids algorithm that essentially computes class specific prototypes. Each aspect of our training framework is further elucidated in the following sections, and the overview is given in Figure 3.3 and 3.4.

3.1 SimCLR Framework

SimCLR learning framework involves training using a contrastive loss minimization objective. The goal of the loss would be to minimize the distance between features extracted from different transformed versions of one image while increasing the distance between features from different images. SimCLR [15] first learns generic representations of images on an unlabeled dataset, and then it can be fine-tuned with a small amount of labeled images to achieve good performance for a given classification task. Figure 3.1 showcases the concept behind SimCLR, x_i and x_k are augmented versions of an image, the distance between them is minimised in the embedding space by SimCLR.

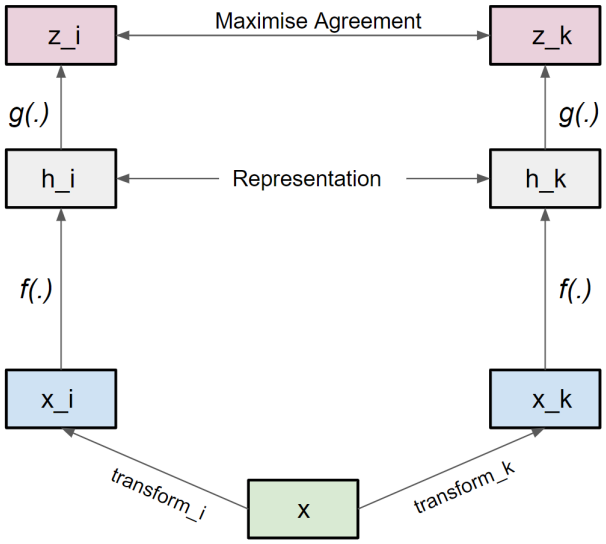


Figure 3.1: SimCLR concept

3.2 K-Medoids

The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. K-medoids clustering is a variant of K-means that is more robust to noises and outliers. Instead of using the mean point as the center of a cluster, K-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points. Figure 1 shows the difference between mean and medoid in a 2-D example. The group of points in the right form a cluster, while the rightmost point is an outlier. Mean is greatly influenced by the outlier and thus cannot represent the correct cluster center, while medoid is robust to the outlier and correctly represents the cluster center [16]. Figure 3.2 shows difference between K-Means and K-Medoid clustering

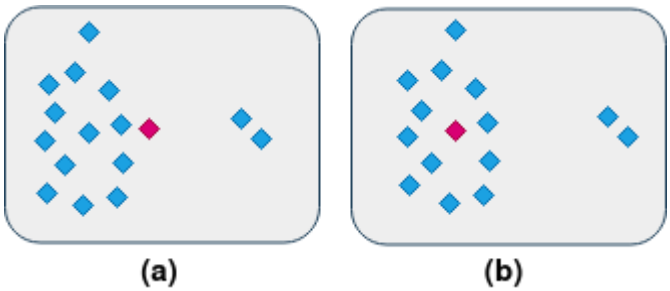


Figure 3.2: (a) shows mean in K-Means clustering, which is influenced by the outlier. (b) shows the mediod in K-medoid clustering which remains unaffected by the outliers.

Partitioning around medoids is a representative K-medoids clustering method. The basic idea is as follows: Select K representative points to form initial clusters, and then repeatedly moves to better cluster representatives. All possible combinations of representative and non representative points are analyzed, and the quality of the resulting clustering is calculated for each pair. An original representative point is replaced with the new point which causes the greatest reduction in distortion function. At each iteration, the set of best points for each cluster form the new respective medoids.

3.3 Loss Functions

We train our model sequentially on two loss functions - Cross Entropy loss and Contrastive loss.

3.3.1 Cross Entropy Loss

Cross-entropy is a measure of the difference between two probability distributions for a given random variable or set of events. Cross entropy is given by

$$L_{CE} = -\sum_{c=1}^{M} y_{o,c} log(p_{o,c}) \tag{3.1}$$

where,
y - binary indicator (0 or 1) if class label c is the correct classification for observation o.
p - predicted probability observation o is of class c.

3.3.2 Contrastive Loss

A minibatch of N examples is sampled and a contrastive prediction task is defined on pairs of augmented examples derived from the minibatch, resulting in $2N$ data points. Negative examples are not sampled explicitly. Instead, given a positive pair, the other $2(N-1)$ augmented examples within a minibatch are treated as negative examples. The dot product between l_2 normalized u and v, i.e., cosine similarity is denoted by

$$sim(u, v) = u^T v / ||u|| ||v|| \tag{3.2}$$

$$\mathcal{L}_{Con} = [d_p - m_{pos}]_+ + [m_{neg} - d_n]_+ \tag{3.3}$$

where $[x]_+$ is $max(0, x)$. The values of m_{pos} and m_{neg} are set to 1 and 0 respectively in our experiments. We use $\mathcal{L}_{Con}$ loss in conjunction with maximum loss miner to obtain features of clean samples to aggregate to the memory bank.

3.3.3 Algorithm

Thise section outlines the algorithm proposed in this work. The Warm-up phase and the weight calculation phase is shown.

Algorithm 1: Warm-up & Weight Calculation

Input: $\mathcal{B} = \{(x_0, y_0), (x_1, y_1)....(x_n, y_n)\}$: minibatch of size n in dataset $\mathcal{S}$;
$f(.)$: The deep learning framework;
Parameter: $\mathcal{M}$: Fixed size memory bank;
Warm-up Phase
foreach $\mathcal{B} \in \mathcal{S}$ **do**
  $\phi(x_i), \theta(x_i) \leftarrow f((x_i, y_i))$;
  $hardpairs \leftarrow miner(\phi_{\mathcal{B}}, L_{con})$;
  if $\phi(x_i) \notin hardpairs$ **then**
    $\mathcal{M}_{Y_i} \leftarrow \phi(x_i)$;
  end
  Calculate $L_{CE}(\theta(x_i), y_i)$ and update $f(.)$;
  Calculate $L_{con}(\phi(x_i))$ and update $f(.)$;
end
Calculate Weights
foreach $(x_i, y_i) \in \mathcal{S}$ **do**
  $\phi(x_i) \leftarrow f((x_i, y_i))$;
  $w_i \leftarrow cosinesimilarity(\mathcal{M}_{Y_i}, \phi(x_i))$
end

3.4 Warm-up Phase

We train a ResNet-18 neural architecture using SimCLR self-supervised training for histopathology images without labels. In the first stage, we use an interleaved training based on cross entropy loss ($\mathcal{L}_{CE}$) and contrastive loss ($\mathcal{L}_{Con}$), where the latter attempts to make the distance between positive pairs (d_p) smaller than some margin (m_{pos}) and the distance between negative pairs (d_n) larger than some threshold (m_{neg}). We use $\mathcal{L}_{Con}$ loss in conjunction with maximum loss miner to obtain features of clean samples to aggregate to the memory bank.

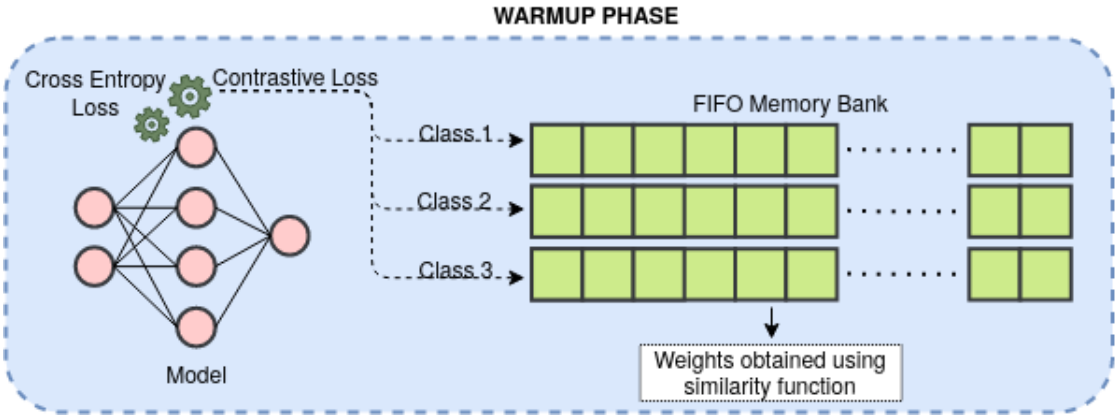


Figure 3.3: Warm-up phase representation

We can define the algorithm as shown in Algorithm 1. The algorithm inputs a mini-batch (B) of size n from the dataset. The warm-up phase uses a 128-dimensional feature vector ($\phi(x_i)$) and a n-class softmax output ($\theta(x_i)$) from the model for each sample in the minibatch, for updating the memory bank and training the model. $\phi(x_i)$ for all the samples in the batch and the labels are fed into a maximum loss miner function which computes the 'hard-pairs' using contrastive loss (L_{con}) and cosine similarity as the distance metric. The maximum loss miner creates random subsets of the given minibatch, calculates the contrastive loss for the subsets and gives the indices of the samples in the subset with maximum loss value as the output. The samples in the miner output are the hard pairs and would contain the noisy labels within the given minibatch. The contrastive loss takes the feature vectors for a positive sample and calculates its cosine similarity with the samples of the same class and contrasts that with the distance to the negative samples. The loss function forms positive and negative pairs based on the given labels. All the feature vectors ($\phi(x_i)$) except the hard-pairs are updated class-wise into a memory bank (M_{Y_i}). The memory bank used in the experiments is limited to a fixed size and the features are appended in a FIFO manner. The softmax output ($\theta(x_i)$) is used to calculate simple cross entropy loss which is backpropagated. In the same epoch, contrastive loss is calculated for the minibatch and backpropagated to train the model. Cross entropy encourages the model to improve the classification accuracy. Since we are using self-supervised contrastive pre-trained model backpropagating the contrastive loss interleaved with cross entropy further improves the classification performance of the model.

3.5 Weight calculation Phase

Post warm-up phase, i.e., second step, we calculate the weights for all the samples in the train dataset using the memory bank. The algorithm iterates over all the samples in the train dataset, calculates $\phi(x_i)$ and computes the cosine similarity with the prototypes in the memory bank of the same class. We also find K-Medoids for the features in the memory bank and cosine similarity is calculated for the samples with the medoids in the memory bank of the same class. The similarity scores are stored and used in the next step as the weights for cross entropy loss for the final phase of training in Section 3.6.

$$S_{u,v} = \frac{u^T v}{||u||.||v||} \tag{3.4}$$

$$W_{u,v} = \frac{S_{u,v} - (S_{u,v})_{min}}{(S_{u,v})_{max} - (S_{u,v})_{min}} \tag{3.5}$$

u - feature vector of input
v - feature prototype of class corresponding to input
$S_{u,v}$ - cosine similarity between l_2 normalized u and v
$W_{u,v}$ - weights for the cross entropy loss in the range [0,1]

3.6 Final Training Phase

In the final phase of the algorithm, we train the model using a weighted cross entropy loss function. The weights for each sample of the dataset are the cosine similarity scores which are calculated in the step described in the section 3.5.

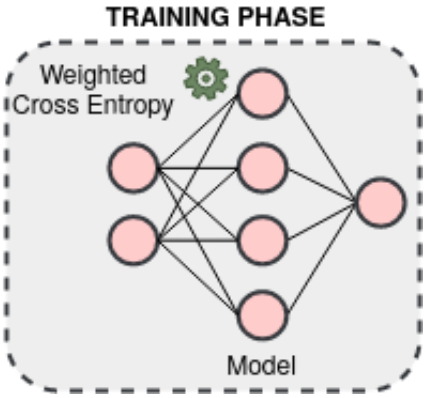


Figure 3.4: Training phase representation

Chapter 4

Experiments and Results

4.1 Overview

For training, the images were reshaped to 224 × 224. Color jitter, random horizontal and vertical flips was used to augment the training data. A batch size of 8 and a stochastic gradient descent optimizer with a momentum of 0.9 and a learning rate of 0.01 was used.

4.2 Experiments-I: BACH

The ICIAR 2018 Grand Challenge dataset called Breast Cancer Histology (BACH) dataset [8] was used for the first set of experiments. The dataset consists of 400 breast cancer microscopy image patches of size 2048 × 1536, labeled into four categories: normal, benign, ductal carcinoma in-situ (DCIS), and invasive ductal carcinoma (IDC) based on the predominant cancer type present in these images. All images were used with the original 33x magnification. We used four-fold cross validation to perform our experiments where 75 random images from the first three categories were used for training and the remaining 25 images for validation. The DCIS class was used to generate the open-set noisy samples. In the training dataset, labels were changed symmetrically and randomly in a controlled manner to simulate the closed-set label noise.

For training, the images were reshaped to 224 × 224. Color jitter and random horizontal and vertical flips were used to augment the training data. A batch size of 8 and a stochastic gradient descent optimizer with a momentum of 0.9 and a learning rate of 0.01 were used. For the warm-up phase, ten epochs of training were used. The batch size of the maximum loss threshold miner was set to 7. The class-specific memory banks of size 300 were used in our warm-up phase. For the K-medoids experiment, K value was set to 3 and one prototype was computed per class.

It was observed that the mean weight scores for the clean samples were much higher than the mean scores for the noisy samples as seen in the Figure 4.2. The Table 4.1 shows the results we obtained with memory bank and K-medoids approach contrasted with the baselines on BACH dataset.

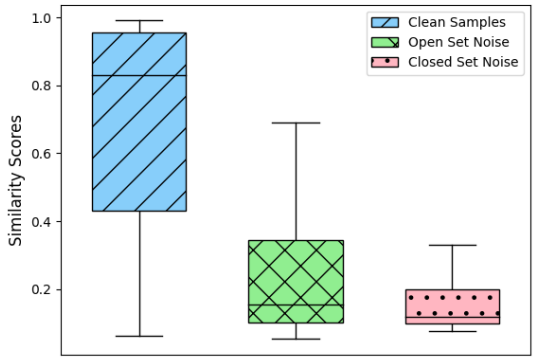


Figure 4.1: Distribution of similarity scores of samples with the memory bank obtained after warm-up phase, corresponding to clean data and noisy data.

OOD Noise	15			20		
Label Noise	10	14	18	10	14	18
CE Imagenet	82.33±1.12	81.00±1.61	77.67±1.26	80.66±1.99	81.66±0.90	79.67±1.06
CE Conloss	84.00±2.11	84.38±1.94	82.33±0.84	85.66±1.99	83.32±2.47	82.32±2.57
SSGCE Imagenet	86.00±0.47	84.66±0.47	83.33±0.47	85.35±0.94	81.33±1.88	79.33±0.47
Mem-Bank (ours)	91.67±1.14	**90.33**±1.00	**89.03**±0.35	90.33±0.84	89.33±0.54	**88.66**±1.38
Mem-Bank+K-Medoid (ours)	**92.33**±0.33	90.00±0.86	88.08±2.05	**91.00**±1.26	**90.66**±0.94	88.00±0.54

Table 4.1: Four-fold cross-validation classification accuracies on BACH dataset with different levels of label noise and OOD noise.

4.3 Experiments-II: TCGA

For our second set of experiments, we chose the basal versus Luminal A PAM50 subtype classification from H&E stained WSI. A total of 130 WSI, 65 from each class, were sampled from the TCGA-BRCA dataset. We followed a class-wise balanced data split in which 90 WSI were taken for training, 10 WSI for validation and the rest of 30 for testing. We followed a patch-based classification approach in which potential tumor regions from the WSI were identified and annotated by a pathologist. All the patches from the same slide were given the same label. The presence of intra-tumor heterogeneity naturally introduces label noise in this dataset. In addition to the dominant subtype, other subtypes of breast cancer could also be present in the tissue images. The occurrence of such intra-tumor heterogeneity along with image degradation introduces open-set label noise. Furthermore, although the Luminal A dataset consisted mostly of ductal morphology, some samples were also had lobular morphology, which further contaminated the dataset as OOD samples.

For our experiments, we used patches of size 512×512 at 40x magnification. Approximately, 20,000 patches were extracted from the training cases. As before, we used a SimCLR pretrained ResNet-18 architecture, a batch size of 128, miner size of 32 and SGD optimizer with a learning rate of 0.01 for training this model. The warmup phase was run for three epochs. Images were augmented using color jitter, random horizontal and vertical flips for training. Further, class-specific memory banks of size 10,000 were used in our warm-up phase. The number of prototypes per class was increased to six for

K-medoids considering the intra-class heterogeneity in the dataset. The computations were performed on an Nvidia DGX station with 8 V100 cards.

The final evaluation was carried out at patient level by averaging the results of the constituent patches. Results on 30 held out patients, 15 from each class, is shown in Table 4.2. We observed that using K-mediods in the memory bank improved results compared to using all samples, most likely because the mediods are more robust prototypes for large datasets that are likely to have outlier samples.

	Slide-level Accuracy
CE-Imagenet	73.33
CE-SimCLR	80.00
SSGCE-Loss	83.33
Mem-Bank (ours)	80.00
Mem-Bank+K-Medoids (ours)	**86.67**

Table 4.2: Basal vs Luminal A Classification accuracy percentages on 30 held out WSI.

4.4 Experiments-III: Kather

The results discussed in this section were obtained by using a ResNet-18 as the main architecture on Kather dataset. The OOD noise levels were set to 30% and 35% per class, the label noise levels were fixed to 30%, 40% and 50% per class. The table 4.3 shows the comparison of performance with different noise levels. For training, the images were reshaped to 224×224. Color jitter, random horizontal and vertical flips was used to augment the training data. A batch size of 8 and a stochastic gradient descent optimizer with a momentum of 0.9 and a learning rate of 0.01 was used. For the warm-up phase, 10 epochs of training was used. The miner value of maximum loss threshold miner was set to 7. The class-specific memory banks of size 300 were used in our warm-up phase. For the K-Medoid experiment, K value was set to 3 and one prototype was computed per class.

OOD Noise	30%			35%		
Label Noise	30%	40%	50%	30%	40%	50%
CE-Imagenet pretrained	89.12±1.07	88.91±0.64	87.05±0.05	89.28±0.69	89.19±0.50	85.53±0.93
CE-SimCLR pretrained	92.6±0.2	92.25±0.45	91.05±0.45	92.75±0.85	91.93±1.14	91.09±0.59
Mem-Bank (ours)	**94.08±0.26**	**92.91±0.59**	**91.68±0.32**	**92.57±0.81**	**93.71±0.32**	**92.33±0.11**
Mem-Bank+K-Medoids (ours)	92.33±0.33	90.00±0.86	88.08±2.05	91.00±1.26	90.66±0.94	88.00±0.54

Table 4.3: Two-fold cross-validation classification accuracies on Kather dataset with different levels of label noise and OOD noise.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

Through our experiments‘, we establish the performance in model degradation in presence of both open-set and closed-set label noise. We proposed a simple and effective sample weighting method that refocuses model optimization to samples with cleaner labels. Our experiments, on much harder TCGA dataset, also establish the inherent ubiquitous nature of open-set and closed-set noise in a medical imaging dataset.

5.2 Future Work

Following is a list of future work that can be looked into.

- Method of aggregating features in the memory bank - In this work, we were adding features to the memory bank in FIFO manner, but this does not explicitly ensure the features are similar. We can compute contrastive similarity between the existing clean features and the new features and add higher similarity samples into memory bank.
- Contrastive methods tend to work better with more number of negative examples, since presumably larger number of negative examples may cover the underlying distribution more effectively and thus give a better training signal. In this work, the batch size was fixed to 8 for both BACH and Kather dataset, but further experiments can be run to better the performance with various batch sizes. Also, techniques including momentum based contrastive learning can be used to improve the self-supervised embedding further.
- Re-weighting strategy - Use of non-linear functions on the similarity scores before using them with the weighted cross entropy loss can be experimented with.
- Study the effect of varying contrastive loss threshold for small loss trick on the algorithm performance.
- The memory bank size was fixed to 300 in our experiments for BACH and Kather datasets. This can be changed to see the effect on the model's classification performance.
- The in K-Medoids experiments, we had fixed the number of clusters to be 3 which can be varied and better performance can be expected on Kather dataset.

Chapter 6

De-speckling Ultrasound Images

6.1 Introduction

Ultrasound imaging is a popular non-invasive and low cost technique to observe the dynamical behavior of organs. This technique uses ultrasonic waves which are produced from the transducer and travel through body tissues. The return sound wave vibrates the transducer which turns into electrical pulses that travel to the ultrasonic scanner where they are processed and transformed into a digital image.

The speckle noise in ultrasound images tends to obscure diagnostically important features and degrades the image quality significantly, making the recognition and analysis of the image details difficult [17]. Ultrasound pulses randomly interfere with objects of comparable size to the sound wavelength and the superposition of acoustical echoes produces an intricate interference pattern [18]. It degrades the fine details and edges definition and limits the contrast resolution by making it difficult to detect small and low contrast lesions in body.

6.2 Dataset

The performance assessment of speckle suppression efficiency can be measures using synthetic speckled images. To achieve this, generally a clean image is used as a reference and speckle noise is added to it to generate clean and noisy data pairs.

Satellite images are equally affected by speckle noise. Since the ground truth and noisy image pairs of ultrasound images were not available at the time of experimentation, we use UC Merced land use dataset [19], whose images are manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the country, and which of the pixel resolution is 1 foot.

In this set of experiments 21,000 images were used for training the model and 10,500 images were used for validating the model. The input high resolution images were of size 256×256 while low resolution noisy images were of size 128×128.

6.3 Experiments & Results

The model is based on the U-Net architecture. The U-Net architecture contains two paths. First path is the contraction path (also called as the encoder) which is used to capture the context in the image. The encoder is a stack of convolutional and max pooling layers. The second path is the symmetric expanding path (also called as the

decoder) which is used to enable precise localization using transposed convolutions. Thus it is an end-to-end fully convolutional network (FCN), i.e. it only contains Convolutional layers and does not contain any Dense layer because of which it can accept image of any size [20].

To use this architecture for super-resolution, after the last layer an upsampling layer was added.

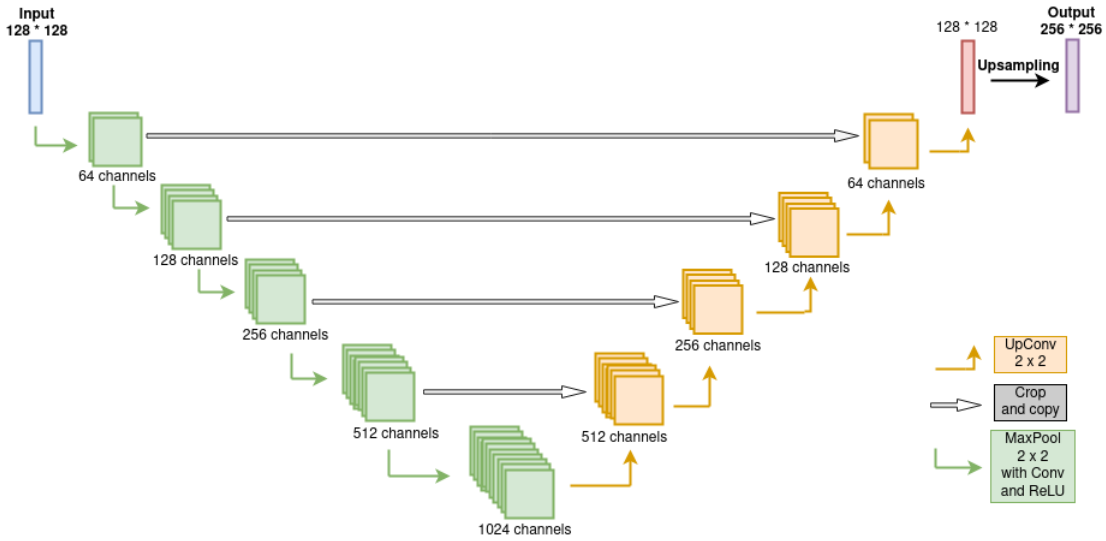


Figure 6.1: Modified U-Net Model

6.3.1 Metrics

The experiments use one or more than one of the following as either a loss function or a performance metric.

1. Mean Absolute Error / L1 Loss : Given two images I_1 and I_2 with N pixels, L1 is defined as
$$L1 = \frac{1}{N}\|I_1 - I_2\|_1^1 \tag{6.1}$$

2. Mean Squared Error (MSE) / L2 Loss : It denotes the average difference of the pixels all over the image. Given two images I_1 and I_2 with N pixels, MSE between them is defined as
$$MSE = \frac{1}{N}\|I_1 - I_2\|_2^2 \tag{6.2}$$

3. Structural Similarity Index Measure (SSIM) : It is a full reference metric that measures perceptual loss between two images I_1 and I_2.
$$SSIM(I_1, I_2) = \frac{2\mu_{I_1}\mu_{I_2} + k_1}{\mu_{I_1}^2 + \mu_{I_2}^2 + k_1} \times \frac{\sigma_{I_1 I_2} + k_2}{\sigma_{I_1}^2 + \sigma_{I_2}^2 + k_2} \tag{6.3}$$
where μ_x and σ_x are the mean and standard deviation of an image x.

4. Peak Signal to Noise Ratio (PSNR) : It is the ratio between the maximum possible power of an image and the power of corrupting noise that affects the quality of its representation. Assuming $L = 255$,

$$PSNR = 10log_{10}\frac{L^2}{MSE} \tag{6.4}$$

6.3.2 Results

Experiments include use of super-resolution to de-speckle satellite images which are equally affected by speckle noise.

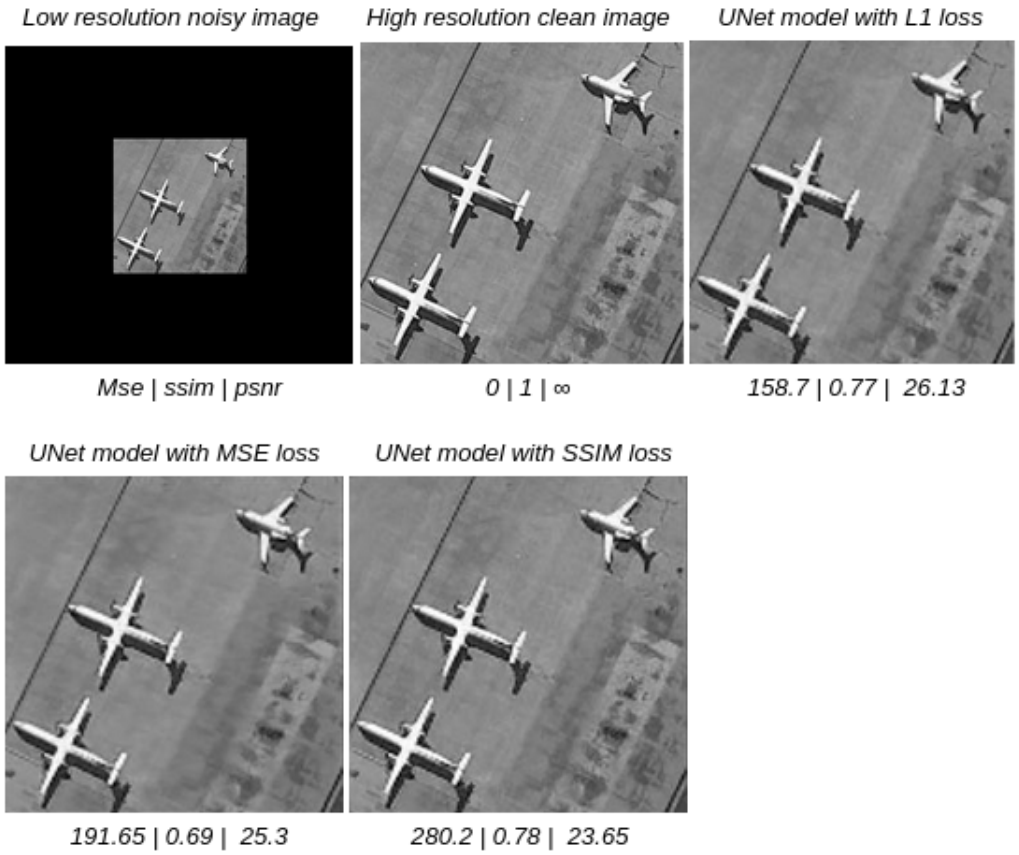


Figure 6.2: Comparison of results UNet model trained with different loss functions.

Loss function	MSE Loss	SSIM Value	PSNR
L1	158.7	0.77	26.13
MSE	191.65	0.69	25.3
SSIM	280.2	0.78	23.65

Table 6.1: UNet experiments on Satellite dataset, values calculated on a test image.

Chapter 7

Future Work

1. Use of wavelet transforms to denoise the speckle data - Because wavelets localize features in the data to different scales, we can preserve important image features while removing noise. At each scale, some operations, such as thresholding can be performed to suppress noise. Denoising is accomplished by transforming back the processed wavelet coefficients into spatial domain.

2. Implementing the performed set of experiments on synthetically generated ultrasound image noisy and clean pairs.

Bibliography

[1] Nikhil Cherian Kurian, Amit Sethi, Anil Reddy Konduru, Abhishek Mahajan, and Swapnil Ulhas Rane, "A 2021 update on cancer image analytics with deep learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 4, pp. e1410, 2021.

[2] JA Morris, "Information and observer disagreement in histopathology," *Histopathology*, vol. 25, no. 2, pp. 123–128, 1994.

[3] Nikhil Cherian Kurian, Pragati Shuddhodhan Meshram, Abhijeet Patil, Sunil Patel, and Amit Sethi, "Sample specific generalized cross entropy for robust histology image classification," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1934–1938.

[4] Aritra Ghosh, Himanshu Kumar, and PS Sastry, "Robust loss functions under label noise for deep neural networks," *arXiv preprint arXiv:1712.09482*, 2017.

[5] 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*, 2017, pp. 1944–1952.

[6] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia, "Iterative learning with open-set noisy labels," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8688–8696.

[7] Hongxin Wei, Lue Tao, Renchunzi Xie, and Bo An, "Open-set label noise can improve robustness against inherent label noise," *arXiv preprint arXiv:2106.10891*, 2021.

[8] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al., "Bach: Grand challenge on breast cancer histology images," *Medical image analysis*, vol. 56, pp. 122–139, 2019.

[9] Chang Liu, Han Yu, Boyang Li, Zhiqi Shen, Zhanning Gao, Peiran Ren, Xuansong Xie, Lizhen Cui, and Chunyan Miao, "Noise-resistant deep metric learning with ranking-based instance selection," in *CVPR*, 2021.

[10] Shallu and Rajesh Mehra, "Breast cancer histology images classification: Training from scratch or transfer learning?," *ICT Express*, vol. 4, no. 4, pp. 247–254, 2018.

[11] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan,

Michael Donovan, Gerardo Fernandez, Jack Zeineh, Matthias Kohl, Christoph Walz, Florian Ludwig, Stefan Braunewell, Maximilian Baust, Quoc Dang Vu, Minh Nguyen Nhat To, Eal Kim, Jin Tae Kwak, Sameh Galal, Veronica Sanchez-Freire, Nadia Brancati, Maria Frucci, Daniel Riccio, Yaqi Wang, Lingling Sun, Kaiqiang Ma, Jiannan Fang, Ismael Kone, Lahsen Boulmane, Aurélio Campilho, Catarina Eloy, António Polónia, and Paulo Aguiar, "Bach: Grand challenge on breast cancer histology images," 2019, vol. 56, pp. 122–139.

[12] Min-Jen Tsai and Yu-Han Tao, "Deep learning techniques for colorectal cancer tissue classification," in *2020 14th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2020, pp. 1–8.

[13] Aritra Ghosh and Andrew Lan, "Contrastive learning improves model robustness under label noise," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2703–2708.

[14] Ozan Ciga, Anne L. Martel, and Tony Xu, "Self supervised contrastive learning for digital histopathology," *ArXiv*, vol. abs/2011.13971, 2020.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, Hal Daumé III and Aarti Singh, Eds. 13–18 Jul 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, PMLR.

[16] Leonard Kaufman and Peter J Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, vol. 344, John Wiley & Sons, 2009.

[17] Jinhua Yu, Jinglu Tan, and Yuanyuan Wang, "Ultrasound speckle reduction by a susan-controlled anisotropic diffusion method," *Pattern Recognition*, vol. 43, no. 9, pp. 3083–3092, 2010.

[18] Christoph B. Burckhardt, "Speckle in ultrasound b-mode scans," *IEEE Transactions on Sonics and Ultrasonics*, vol. 25, no. 1, pp. 1–6, 1978.

[19] Yi Yang and Shawn Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," 01 2010, pp. 270–279.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.