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

Yuanhong Chen¹ * Fengbei Liu¹ * Hu Wang¹ Chong Wang¹
Yuyuan Liu¹ Yu Tian² Gustavo Carneiro³
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

1. Dataset Statistics

Table 1 shows the statistics of our training noisy training set (NIH [8] and ChestXpert (CXP) [4]) and clean testing sets (OpenI [3] and PadChest [1]). Due to inconsistencies in the number of labels for each dataset, we trim the original datasets and only keep the samples that contain labels present in all datasets based on [2, 5]. After our data pre-processing, there are 83,672 frontal-view images with 14 common chest radiographic observations for NIH [8] dataset where the corresponding testing sets for OpenI [3] and PadChest [1] contain 2,917 and 14,714 frontal-view images respectively. For CXP, we have 170,958 frontal-view images with 8 chest radiographic observations where the corresponding testing set for OpenI [3] and PadChest [1] contain 2,823 and 12,885 frontal-view images, respectively.

2. Further Ablation Studies

We evaluate the number of KNN neighboring samples that are required for a clean re-labelling. We measure the precision and recall for the detection of noisy-labels of our graph-based relabelling method in Fig. 1 as a function of the threshold of the minimum number of nearest neighbors for NIH [8] dataset where the corresponding testing sets for OpenI [3] and PadChest [1] contain 2,917 and 14,714 frontal-view images respectively. For CXP, we have 170,958 frontal-view images with 8 chest radiographic observations where the corresponding testing set for OpenI [3] and PadChest [1] contain 2,823 and 12,885 frontal-view images, respectively.

3. Visualisation of Smoothing Techniques

To visualise the performance of different label smoothing techniques, we plot the t-SNE [7] for a toy problem. More specifically, we first generate two isotropic Gaussian clusters as the clean set (Fig. 2a) and randomly inject 20% of symmetric noise (Fig. 2b) to form a noisy set. We show that our BoMD demonstrates a better tradeoff when correcting the labels since it re-labels the noisy samples without being overconfident in the detection (like shown by GLS [9]) and without over-smoothing the labels (like displayed by LS [6]). Note that we set the smoothing parameter $r$ to 0.6 and -0.4 respectively for LS [6] and GLS [9].

4. Additional Results

4.1. Per-finding results

We show per-finding results over all available findings for NIH [8] in Tables 3 and 4 and for CheXpert [4] in Tables 5 and 6.

4.2. Hyper-parameter sensitivity

Tab. 2 studies the four hyper-parameters ($\lambda$, $\gamma$, $M$ and $K$) of BoMD. In general, for $\lambda$, we note that relying too much on the pseudo-labels from the graph ($\lambda = 0.2$) or the original noisy labels ($\lambda = 1.0$) worsens the performance, with the best result achieved with a balanced $\lambda = 0.6$. We noticed that the method is robust to $\gamma$ and $M$ with little variation in results. As for $K$, values larger than 10 over-smooth the decision boundary of our classifier, causing under-fitting. The values $\lambda = 0.6$ and $\gamma = 0.25$, $M = 3$, and $K = 10$ reach the best results.

4.3. Evaluation for Descriptors from MID

Visualisation of distance distribution. To verify the separation of positive descriptors (labelled as 1) and negative descriptors (labelled as 0) based on their edge weight, we
Table 1: Statistics for all datasets after data pre-processing, where the digit on the left is the total number of samples and the digit inside brackets is the number of classes.

| Datasets   | Train on NIH | CXP [4] | OpenI [3] | PadChest [1] |
|------------|--------------|---------|-----------|--------------|
| Train on NIH | 83,672 (14)  | -       | 2,971 (14) | 14,714 (14)  |
| Train on CXP | -            | 170,958 (8) | 2,823 (8) | 12,885 (8)   |

Table 2: Ablation study of the hyper-parameters using mean AUC. Models are trained on NIH [8] and tested on OpenI [3] and PadChest [1]. Note that for each hyper-parameter, we fix the others to their best values (i.e., $\lambda = 0.6$, $\gamma = 0.25$, $M = 3$ and $K = 10$).

| Settings | $\lambda$ | $\gamma$ | $M$ | $K$ |
|----------|-----------|----------|-----|-----|
| AUC      | 0.2       | 0.05     | 1   | 5   |
|          | 0.4       | 0.15     | 3   | 89.52 86.50 10 89.52 86.50 |
|          | 0.6       | 0.25     | 5   | 88.92 86.39 20 88.23 85.79 |
|          | 0.8       | 0.35     | 7   | 89.03 86.43 50 87.59 85.49 |
|          | 1.0       | 0.45     | 9   | 88.45 86.29 100 87.36 85.48 |

References

[1] 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. 1, 2

[2] Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, et al. Torchxrayvision: A library of chest x-ray datasets and models. arXiv preprint arXiv:2111.00595, 2021. 1

[3] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2):304–310, 2016. 1, 2

[4] 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

[5] 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. 1

[6] 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. 1, 3

[7] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(11), 2008. 1, 2

[8] 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

[9] 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. 1, 3
KNN Threshold
0.0
0.2
0.4
0.6
0.8
1.0
2 4 6 8 10
[.2,.2] 

Figure 1: Label-wise precision and recall of our KNN propagated label under $\hat{y}$ w.r.t the clean annotation from PadChest. The horizontal axis shows a threshold of the minimum number of nearest neighbors containing each class.

Figure 2: Visualisation of different label smoothing techniques. The color of each data point indicates the confidence score. We start with two isotropic Gaussian clusters in (a) as the clean set where red points indicate class 1 and blue points represent class 2. We randomly inject 20% of symmetric noise to form the noisy set in (b). We compare our method (in (d)) with two baseline methods, namely: label smoothing (LS) [6] (in (c)) and generalised label smoothing (GLS) [9] (in (e)). We show that our method alleviates the noisy label problem by modifying the confidence score based on the nearest neighbors, while LS pushes the labels toward the uniform distribution and GLS pushes the labels toward the sharp binary distribution. Note that GLS has a different scale for confidence scale which is from -0.2 to +1.2, while the others have a range from 0 to 1.

Table 3: Disease-level testing AUC results for models trained on NIH.

| Datasets      | Models       | Atelectasis | Cardiomegaly | Effusion | Infiltration | Mass | Nodule | Pneumonia | Pneumothorax | Edema | Emphysema | Fibrosis | Pleural Thicken | Hernia | Mean AUC |
|---------------|--------------|-------------|--------------|----------|--------------|------|--------|-----------|--------------|-------|-----------|---------|----------------|--------|----------|
| OpenI PadChest | OpenI PadChest | 86.85       | 91.25        | 94.05    | 77.48        | 95.72 | 81.68  | 85.72     | 81.50         | 84.31 | 83.26     | 85.85   | 77.99          | 92.90  | 85.54    |
| OpenI PadChest | OpenI PadChest | 83.59       | 91.72        | 96.29    | 73.78        | 86.93 | 75.99  | 75.73     | 75.34         | 79.78 | 79.81     | 96.46   | 71.85          | 89.90  | 83.00    |
| OpenI PadChest | OpenI PadChest | 84.83       | 94.74        | 84.49    | 84.03        | 87.47 | 80.71  | 80.71     | 74.78         | 78.06 | 81.26     | 93.17   | 77.59          | 88.79  | 85.98    |
| OpenI PadChest | OpenI PadChest | 79.88       | 81.63        | 97.75    | 81.61        | 85.81 | 74.14  | 74.14     | 73.22         | 59.91 | 59.91     | 63.89   | 72.32          | 62.98  | 73.00    |
| OpenI PadChest | OpenI PadChest | 70.98       | 77.93        | 74.39    | 73.41        | 71.31 | 57.35  | 57.35     | 56.22         | 58.18 | 43.31     | 58.17   | 57.35          | 83.45  | 74.28    |
| OpenI PadChest | OpenI PadChest | 73.48       | 84.54        | 86.76    | 67.28        | 71.41 | 57.66  | 57.66     | 59.89         | 85.46 | 31.33     | 95.20   | 75.59          | 85.99  | 74.94    |
| OpenI PadChest | OpenI PadChest | 67.70       | 90.79        | 94.74    | 78.92        | 69.54 | 81.90  | 81.90     | 88.99         | 78.65 | 35.91     | 93.06   | 75.59          | 87.79  | 75.56    |
| OpenI PadChest | OpenI PadChest | 86.21       | 90.28        | 94.74    | 78.92        | 81.90 | 81.90  | 81.90     | 88.99         | 85.78 | 43.31     | 93.06   | 75.59          | 87.79  | 75.56    |
| OpenI PadChest | OpenI PadChest | 85.69       | 92.81        | 96.67    | 73.40        | 84.51 | 85.99  | 85.99     | 80.28         | 85.78 | 35.91     | 93.06   | 75.59          | 87.79  | 75.56    |
| OpenI PadChest | OpenI PadChest | 88.16       | 92.91        | 96.56    | 74.30        | 84.51 | 85.99  | 85.99     | 80.28         | 85.78 | 43.31     | 93.06   | 75.59          | 87.79  | 75.56    |
| OpenI PadChest | OpenI PadChest | 85.66       | 92.94        | 96.56    | 74.30        | 84.51 | 85.99  | 85.99     | 80.28         | 85.78 | 43.31     | 93.06   | 75.59          | 87.79  | 75.56    |

Mean AUC
| Models       | NPC | NCR | LS  | OLS  | GLS  | BoMD  |
|--------------|-----|-----|-----|------|------|-------|
| Datasets     | Open| PadChest | Open| PadChest | Open| PadChest | Open| PadChest | Open| PadChest |
| Atelectasis  | 86.04| 85.23 | 83.80| 85.46 | 85.34| 84.74 | 87.27| 85.18 | 88.23| 83.00 | 87.91| 86.19 |
| Cardiomegaly | 91.42| 92.12 | 89.42| 91.45 | 88.08| 89.17 | 84.59| 89.83 | 89.12| 91.40 | 91.37| 92.17 |
| Effusion     | 95.58| 96.19 | 93.96| 95.89 | 94.54| 95.63 | 94.28| 96.75 | 93.67| 96.36 | 95.28| 96.71 |
| Infiltration | 68.76| 64.08 | 60.48| 67.98 | 72.26| 74.20 | 76.10| 76.19 | 82.08| 71.27 | 81.65| 76.64 |
| Mass         | 80.20| 86.04 | 85.00| 85.98 | 88.08| 80.56 | 82.79| 84.80 | 75.12| 80.67 | 92.31| 88.48 |
| Nodule       | 87.60| 75.68 | 85.12| 75.60 | 86.44| 74.82 | 83.42| 75.27 | 82.10| 74.34 | 84.05| 75.28 |
| Pneumonia    | 91.01| 76.87 | 88.87| 76.40 | 83.50| 76.17 | 87.18| 78.20 | 85.65| 74.83 | 89.99| 78.71 |
| Pneumothorax | 84.28| 79.22 | 83.07| 76.98 | 74.07| 76.10 | 75.89| 80.02 | 73.93| 76.45 | 88.89| 85.82 |
| Edema        | 82.27| 92.40 | 85.66| 93.57 | 83.38| 88.23 | 87.31| 89.55 | 85.92| 93.01 | 87.60| 98.68 |
| Emphysema    | 82.05| 80.87 | 82.36| 75.80 | 76.94| 73.10 | 80.94| 78.15 | 75.16| 74.21 | 85.28| 81.94 |
| Fibrosis     | 87.53| 91.50 | 90.67| 94.57 | 92.09| 96.43 | 90.19| 95.35 | 91.06| 95.29 | 94.56| 97.44 |
| Pleural Thicken | 87.37| 76.06 | 82.66| 76.62 | 82.83| 72.82 | 84.12| 70.55 | 80.10| 68.14 | 86.94| 71.53 |
| Hernia       | 96.60| 94.17 | 94.69| 92.74 | 80.85| 70.11 | 91.95| 85.84 | 87.29| 81.38 | 98.57| 94.22 |
| Mean AUC     | 86.21| 83.88 | 85.06| 83.79 | 83.72| 80.93 | 85.08| 83.51 | 83.80| 81.56 | 89.57| 86.45 |

**Table 5:** Disease-level testing AUC results for models that trained on CheXpert.

| Models       | Hermoza et al | CAN | DivideMix | FINE | ELR | NVUM |
|--------------|---------------|-----|-----------|------|-----|------|
| Datasets     | Open| PadChest | Open| PadChest | Open| PadChest | Open| PadChest | Open| PadChest |
| Cardiomegaly | 86.12| 87.20 | 82.83| 85.89 | 79.53| 85.42 | 83.62| 83.99 | 90.48| 87.46 | 85.15| 88.48 |
| Edema        | 87.92| 94.35 | 86.46| 97.47 | 81.24| 83.41 | 86.43| 87.07 | 90.88| 96.12 | 87.35| 97.21 |
| Pneumonia    | 65.56| 57.15 | 61.88| 63.38 | 55.98| 51.20 | 55.58| 55.58 | 61.59| 64.13 | 64.42| 67.89 |
| Atelectasis  | 78.40| 75.65 | 80.13| 72.87 | 72.74| 68.34 | 72.87| 72.87 | 79.63| 73.68 | 80.81| 75.03 |
| Pneumothorax | 62.09| 78.65 | 74.69| 79.50 | 75.49| 79.98 | 65.34| 68.85 | 74.12| 83.95 | 82.18| 83.32 |
| Effusion     | 87.00| 93.94 | 88.43| 92.92 | 83.75| 88.91 | 85.92| 85.92 | 86.65| 92.42 | 83.54| 89.74 |
| Fracture     | 57.47| 53.77 | 59.96| 60.44 | 63.87| 62.23 | 51.97| 62.50 | 56.75| 62.00 | 57.02| 62.67 |
| Mean AUC     | 74.94| 77.24 | 76.34| 78.92 | 73.23| 74.21 | 71.68| 73.83 | 77.16| 79.97 | 77.21| 80.62 |

**Table 6:** Disease-level testing AUC results for models that trained on CheXpert.
Figure 3: L2 distance between positive/negative descriptors and semantic descriptor
Figure 4: Visualisation of descriptor distribution in latent space.