Towards an Interpretable Chest X-ray Classifier through Optimal Transport Regularization

Chih-Chieh Chen[1] JACKFRANK@GMAIL.COM
Chang-Fu Kuo[1,2,3] ZANDIS@GMAIL.COM
[1] *Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan*
[2] *Medical Education Department, Chang Gung Memorial Hospital, Taoyuan, Taiwan*
[3] *Division of Rheumatology, Allergy and Immunology, Chang Gung Memorial Hospital, Taoyuan, Taiwan*

Editors: Under Review for MIDL 2025

Abstract

In addition to model performance, interpretability is essential for integrating artificial intelligence into clinical settings. In this study, we designed a chest X-ray classifier that provides patch-level outputs while training solely on class labels. To align the patch-level outputs with the locations of diseases, we introduced an optimal transport-based regularization into our architecture. We present results and observations to demonstrate the effectiveness of our approach.

Keywords: Chest X-ray Classification, Kernel Optimal Transport, Weakly Supervised localization.

1. Introduction

We plan to focus on weakly supervised localization of chest diseases using only classification annotations. We have designed a classifier with patch-level outputs. Instead of meticulously designing a neural network with the most suitable receptive fields for each specific task, we proposed encoding each patch with a sufficiently large receptive field. We will then use neural optimal transport (Korotin et al., 2022, 2023) to align the patch output distribution with the corresponding label distributions. As a result, rather than relying on CAM-based approaches (Selvaraju et al., 2017; Zhou et al., 2016) to generate heatmaps, we can directly generate heatmaps from the patch outputs, making the process more straightforward.

2. Method

Our method is illustrated in Fig. 1. As mentioned in the previous section, our goal is to construct a neural network that produces patch-level classification outputs. We used a DenseNet-121 backbone pre-trained on ImageNet. To ensure that each patch has a sufficiently large receptive field, we concatenate each patch with a global average pooling feature. This combined input is then passed through a linear layer to ensure that the output channel size for each patch matches the number of classes. The final output is computed as the weighted Log-Sum-Exp (Pinheiro and Collobert, 2015) of all the channels, allowing us to train our model using the classification labels.

To better regularize the distributions of patch outputs, we applied neural optimal transport (Korotin et al., 2022, 2023) to a simulated label distribution Y. For each ground truth

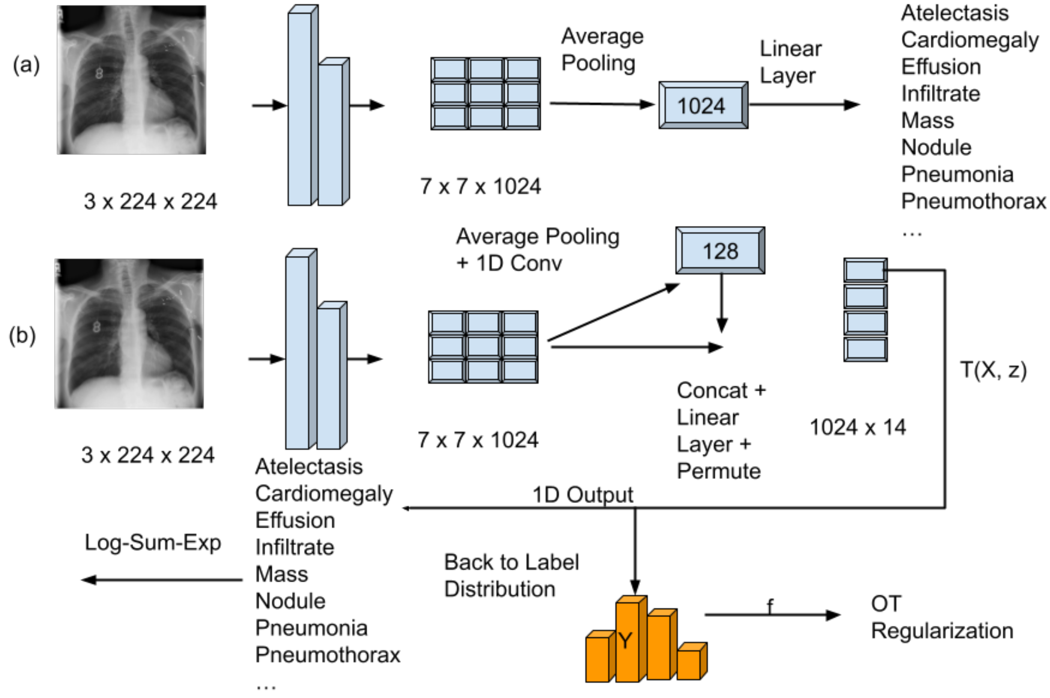


Figure 1: (a) Custom DenseNet-121 architecture. (b) Architecture of our proposed network.

label y, we drop each subtype label with a probability p, reflecting the fact that only a few patches in each X-ray image are abnormal. This process results in a new label y'. Finally, we interpolate y' with a uniform noise w, resulting in $\tilde{y} = \alpha \cdot y' + (1 - \alpha) \cdot w$.

Following (Korotin et al., 2022), we introduce the optimal transport map T and the potential map f. For the network architecture and the loss function, we utilized the 1D counterparts of the architectures described by (Korotin et al., 2022). Each output patch is processed through T before we compute the final weighted Log-Sum-Exp outputs.

For heatmap generation, we filter the values for each subtype by retaining only those that are larger than the tenth largest patch value. All other patch values are set to the minimum value among all patches. After this adjustment, we rescale all patch values to a range of $[0, 255]$.

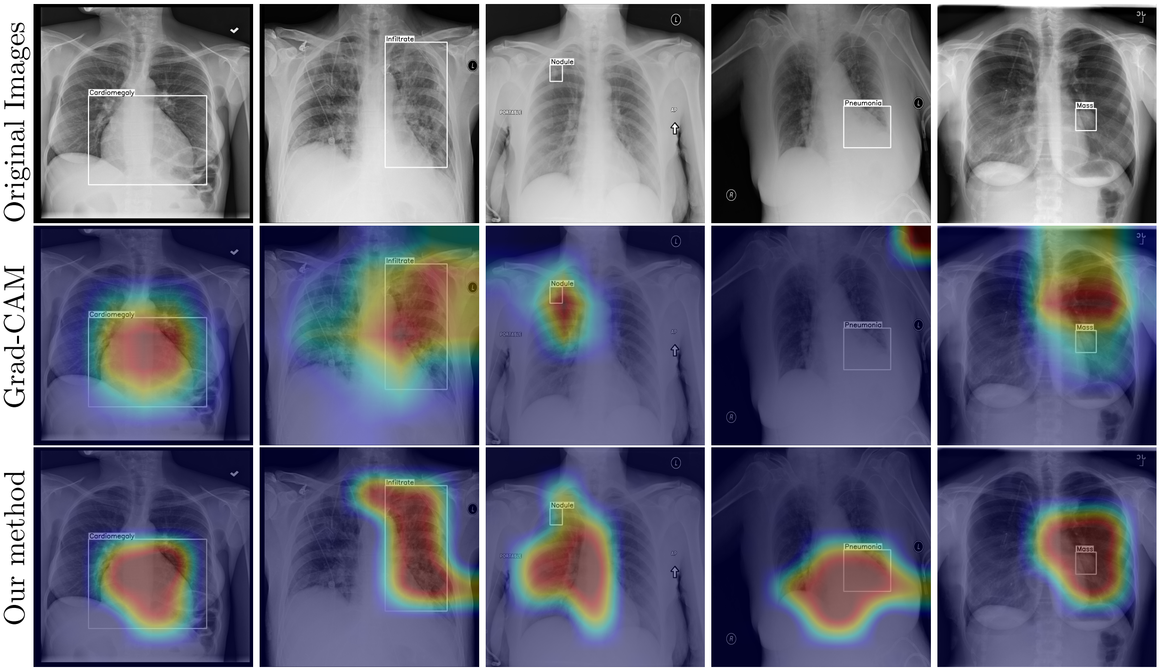
3. Experiments and Results

We tested our method on the ChestX-Ray14 dataset (Wang et al., 2017). All images were resized to 224 × 224 pixels. The batch size was set to 64, with parameters p and α for the simulated distribution Y are set to 0.8 and 0.7, respectively. For comparison, we trained a DenseNet-121 network and applied Grad-CAM on the last layer before the global average pooling layer. All the experiements were conducted with single Nvidia V100 GPU.

The AUCs for our proposed method are presented in 1 while the visualizations of our generated heatmaps and those based on Grad-CAM are shown in Fig. 2. Our empirical observations indicated that both methods performed well for cardiomegaly and infiltrate

Atelectasis	**Cardiomegaly**	**Effusion**	**Infiltrate**	**Mass**
0.8091	0.8757	0.8891	0.7033	0.8120
Nodule	**Pneumonia**	**Pneumothorax**	**Consolidation**	**Edema**
0.7539	0.7697	0.8620	0.7855	0.8820
Emphysema	**Fibrosis**	**Pleural Thickening**	**Hernia**	
0.8975	0.7510	0.7857	0.7902	

Table 1: AUCs of our proposed network on ChestX-ray14 dataset (Wang et al., 2017).



(a)Cardiomegaly (b)Infiltrate (c)Nodule (d)Pneumonia (e)Mass

Figure 2: Heatmap generations of our method and Grad-CAM based method.

cases. we noted instances where Grad-CAM produced highly precise heatmaps, while our method did not perform as effectively, as illustrated in Fig. 2 (c). However, there were times when the heatmaps generated based on Grad-CAM focused on completely irrelevant areas, as seen in Fig. 2 (d). For tiny subtype diseases, they are usually found in the high-intensity areas of our heatmaps; however, our method has not yet been able to accurately pinpoint these locations, as demonstrated in Fig. 2 (e).

4. Conclusion

In this study, we propose a novel approach to predict patch outputs using only classification annotations. Our empirical analysis shows that our method can generate reliable heatmaps for certain subtype diseases, such as cardiomegaly and infiltrates. However, it still struggles to accurately locate small abnormal areas, like masses and nodules. In future work, we plan to enhance our methods and deliver more qualitative results.

Acknowledgments

Both authors acknowledge funding from the Center for Artificial Intelligence in Medicine at Chang Gung Memorial Hospital, via grant agreement no. CLRPG3H0016.

References

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Kernel neural optimal transport. *The Eleventh International Conference on Learning Representations*, 2022.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. *The Eleventh International Conference on Learning Representations*, 2023.

Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1713–1721, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and R Summers. Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR*, volume 7, page 46. sn, 2017.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.