Analyzing the Effects of Handling Data Imbalance on Learned Features from Medical Images by Looking Into the Models

Yawei Li [1] Ashkan Khakzar [2] Yang Zhang [2] Mirac Sanisoglu [2] Seong Tae Kim [3] Mina Rezaei [1]
Bernd Bischl [1] Nassir Navab [2]

Abstract

One challenging property lurking in medical datasets is the imbalanced data distribution, where the frequency of the samples between the different classes is not balanced. Training a model on an imbalanced dataset can introduce unique challenges to the learning problem where a model is biased towards the highly frequent class. Many methods are proposed to tackle the distributional differences and the imbalanced problem. However, the impact of these approaches on the learned features is not well studied. In this paper, we look deeper into the internal units of neural networks to observe how handling data imbalance affects the learned features. Moreover, we study several popular cost-sensitive approaches for handling data imbalance and analyze the feature maps of the convolutional neural networks from multiple perspectives such as analyzing the alignment of salient features with pathologies and analyzing the pathology-related concepts encoded by the networks. Our study reveals differences and insights regarding the trained models that are not reflected by quantitative metrics such as AUROC and AP and show up only by looking at the models through a lens.

1. Introduction

Medical imaging datasets often appear in imbalanced distribution where the frequency of the samples between the different classes of the training dataset is not similar or balanced. The low amount of training samples for infrequent classes or tailed distribution makes it challenging to learn optimal classification boundaries in the representation and can lead to a biased model.

Existing methods to tackle class imbalance problem either modify data distribution or learn appropriate costs to reweight class errors. At the data level, the objective is to balance the data distribution through re-sampling techniques which often are prone to over-fitting (Khan et al., 2019; Buda et al., 2018; He & Garcia, 2009). On the other hand, the cost-sensitive approaches modify the learning algorithm to alleviate the bias towards frequent classes or head of distribution (Cui et al., 2019; Rezaei et al., 2018). The efficacy of these approaches is demonstrated by conventional metrics such as precision and recall and their derivatives. However, the effect of these approaches on the learned features is not well studied. The study of learned features not only improves our understanding of what happens within the models, but also is specifically insightful when the conventional evaluation metrics do not reflect the effect of applying such cost-sensitive approaches.

This study explores what happens to the learned features when it is trained by cost-sensitive approaches to handle the data imbalance. To understand the effect on learned features, we analyze the internal units of neural networks. Specifically, we analyze the feature maps (outputs of convolutional layers) in deep layers by class activation maps and network dissection (Bau et al., 2017). First, we optimize models (ResNet (He et al., 2016) and DenseNet (Huang et al., 2017)) with Binary Cross-Entropy (BCE) (Murphy, 2012). Several recent and popular cost-sensitive losses such as Weighted BCE (WBCE) (Paszke et al., 2019), Focal loss (Lin et al., 2017), and Class-Balanced Focal loss (CB-Focal) (Cui et al., 2019) on NIH Chest X-ray (Wang et al., 2017) dataset and report classical metrics: area Under ROC curve (AUROC), average precision (AP) and predicted probabilities. Then, we visually analyze the impact on salient learned features using class activation maps (Zhou et al., 2016; Selvaraju et al., 2017) and provide quantitative evaluations to validate the visual observation. We then proceed to analyze the learned features using network dissection (Bau et al., 2017; Khakzar et al., 2021b) which quantitatively identifies the concepts encoded (learned) by the model.

[1] Department of Statistics, LMU Munich, Munich, Germany [2] Department of Informatics, Technical University of Munich, Munich, Germany [3] Department of Computer Science and Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do, South Korea. Correspondence to: Yawei Li <yawei.li@stat.uni-muenchen.de>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML). Copyright 2022 by the author(s).

Statement of Contribution By placing models trained with different learning strategies under the lens, we observe that while metrics such as AUROC and AP report equivalent performance, the models trained with cost-sensitive losses encode more pathology-related concepts. Moreover, we observe an increased alignment between salient learned features and pathology-related features.

2. Related Works

Handling Data Imbalance: Much of the recent works on imbalanced learning focused on alleviating this problem using novel objective function. Lin Lin et al. (2017) introduce a focal loss for the bounding box classification in object detection where class-specific weights were automatically learned. Cui Cui et al. (2019) re-weight the loss by the inverse effective number of examples to learn balanced representations. Similarly, (Khan et al., 2019) modified the weights according the uncertainty of predictions. Others address this problem by multi-task learning (Kendall et al., 2018; Rezaei et al., 2018) that used selective instances for training on imbalanced sets for each task.

Interpreting Neural Networks: Two principal neural network interpretation approaches are feature attribution (Simonyan et al., 2013; Zhang et al., 2021; Khakzar et al., 2021a; Sundararajan et al., 2017; Lundberg & Lee, 2017; Khakzar et al., 2021c; 2019b) (i.e. saliency methods (Cong et al., 2018)) and analyzing internal units (e.g. feature visualization (Olah et al., 2017) and dissection (Bau et al., 2020; Khakzar et al., 2021b)). Here, we look into the models from both perspectives. We are interested in both the contribution of input features to the output (i.e., feature attribution) and concepts encoded by the network (via analyzing units). The are many attribution methods, however the identified important features are different for different methods (Krishna et al., 2022; Zhang et al., 2021; Khakzar et al., 2022; 2020). This is a caveat for researchers using the attribution toolkit. Within the various feature attribution methods, CAM (Selvaraju et al., 2017) being a classic and intuitive method, it satisfies benchmarks in terms of faithfulness (Hooker et al., 2019; Zhang et al., 2021; Khakzar et al., 2022). Moreover, regardless of its advantages for attribution, the method provides a summary of activation maps in a certain layer by performing a weighted sum of the attributions. We use network dissection (Bau et al., 2017; Khakzar et al., 2021b) to identify concepts associated with individual convolutional feature maps.

3. Method

The question we explore in this paper is what happens to the learned features within the model when we apply cost-sensitive approaches to handle the class imbalance during training. Specifically, we study the intra-class data imbalance, i.e. the imbalance between the number of positive and negative samples for each class. We first introduce the cost-sensitive methods (see Section 3.1) that consider the data imbalance between positive and negative samples. Then, we explain our approaches for analyzing the learned features (see Section 3.2).

3.1. Handling Data Imbalance

Given $\mathcal{D} = \{x^{(k)}, y^{(k)}\}_{k=1}^{K}$, our goal is to learn the optimal boundary θ^* obtained by empirical loss minimization $\mathcal{L}_{\mathcal{D}}(\theta)$ on the training set $\mathcal{D}$ as: $\theta^* = \arg\min_\theta \mathcal{L}_{\mathcal{D}}(\theta)$. The class imbalanced problem exists when the frequency of the samples among different categories are extremely mismatched. Therefore θ learned on $\mathcal{D}$ using conventional empirical loss can be biased towards the low frequent classes and significantly different from the ideal boundary θ^*. In other words, because of the imbalanced proportion between classes, the optimal boundary is more likely to afford a higher empirical error than an alternative hypothesis based on an empirical loss. This is due to the nature of imbalanced class distribution that forces the classifier to shift θ closer to low-frequent classes because it reduces empirical error. A principal strategy to handle data imbalance is through the cost-sensitive loss functions (Lin et al., 2017; Cui et al., 2019; Kendall et al., 2018; Ridnik et al., 2021).

Here, we study data imbalance problem in multi-label and binary classification. Assume N samples $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$ from M classes. we denote the corresponding label set as $\mathcal{Y} = \{y^{(i)}\}_{i=1}^{N}$ with $y^{(i)} \in \{1, 2, ..., M\}$. The probability $p_m^{(i)}$ of class m for sample $x^{(i)}$ given by a neural network is defined with multiple outputs as cross entropy error between predicted value and true label using Binary Cross Entropy (BCE) loss. The cost-sensitive losses applied to class can be summarized into a unified framework:

$$\mathcal{L}_{\mathcal{CE}} = -\sum_{i=1}^{N}[w_+^{(i)} y^{(i)} \log p^{(i)} + w_-^{(i)}(1-y^{(i)}) \log(1-p^{(i)})], \tag{1}$$

where $w_+^{(i)}$ and $w_-^{(i)}$ are the positive and negative class weights for sample $x^{(i)}$, respectively. We omit the class index m for simplicity. The four losses considered in this work can be derived as follows:

- BCE (Murphy, 2012): $w_+^{(i)} = w_-^{(i)} = 1$;
- WBCE (Paszke et al., 2019): $w_+^{(i)} = \frac{N_-}{N_+ + N_-}$ and $w_-^{(i)} = \frac{N_+}{N_+ + N_-}$;
- Focal (Lin et al., 2017): $w_+^{(i)} = \alpha(1 - p^{(i)})^\gamma$ and $w_-^{(i)} = (1 - \alpha){p^{(i)}}^\gamma$;

- CB-Focal (Cui et al., 2019): $w_+^{(i)} = \frac{1-\beta}{1-\beta^{N_+}} \cdot (1-p^{(i)})^\gamma$ and $w_-^{(i)} = \frac{1-\beta}{1-\beta^{N_-}} \cdot {p^{(i)}}^\gamma$;

where N_+, N_- are the number of positive and negative samples of class m, and α, γ and β are hyper-parameters. As we are discussing intra-class imbalance, each loss is considered individually and we apply the cost-sensitive methods on each loss to strike a balance between positive and negative samples for each class.

3.2. Looking into the Models

The objective is to analyze the features learned by the model in different training scenarios. In this work, we analyze the internal units of the model, specifically the activation pattern of convolutional feature maps in deeper layers. The deep convolutional layers are chosen due to the following two reasons: 1) Final layers in principle encode higher-level concepts and we are interested to know if these concepts align with pathological features. 2) We focus on convolutional layers instead of deeper fully connected layers, since convolutional layers keep the spatial correspondence with input features, and therefore, we can leverage image annotations from experts to study if the activations align with pathology related features. We analyze the internal features maps from two different perspectives:

Class Activation Maps (Zhou et al., 2016; Selvaraju et al., 2017) can be deemed as a tool that summarizes the contributions of feature maps of the final convolutional layer in one tensor (for the networks under investigation, GradCAM (Selvaraju et al., 2017), and CAM (Zhou et al., 2016) are equivalent). The method combines the feature maps using their associated weight (or gradient of output with respect to feature map in GradCAM) of their connections to the output, thus summarizing how feature maps contribute to the final output. The method, in essence, shows which areas are contributing to the output for each sample input. Thus using annotations from experts, we can check whether the salient regions are aligned with the pathologies (Wang et al., 2017; Khakzar et al., 2019a).

Analyzing Encoded Concepts (Bau et al., 2017; Khakzar et al., 2021b): We use an approach inspired by network dissection (Bau et al., 2017) which is a tool for individually analyzing the convolutional feature maps of the neural network. In essence, it identifies the concepts associated with each feature map, thus allowing for understanding what concepts are encoded by the network during training. In this work, we use a conceptually similar approach. For each image in the annotated dataset, we identify the connected components in the threshold activations. If the connected component overlaps the bounding box region, we consider the connected component as a pathology-related concept. Using this approach we can report two values that reflect the concepts encoded in the network quantitatively, which are explained in Sec. 4.

4. Experiments

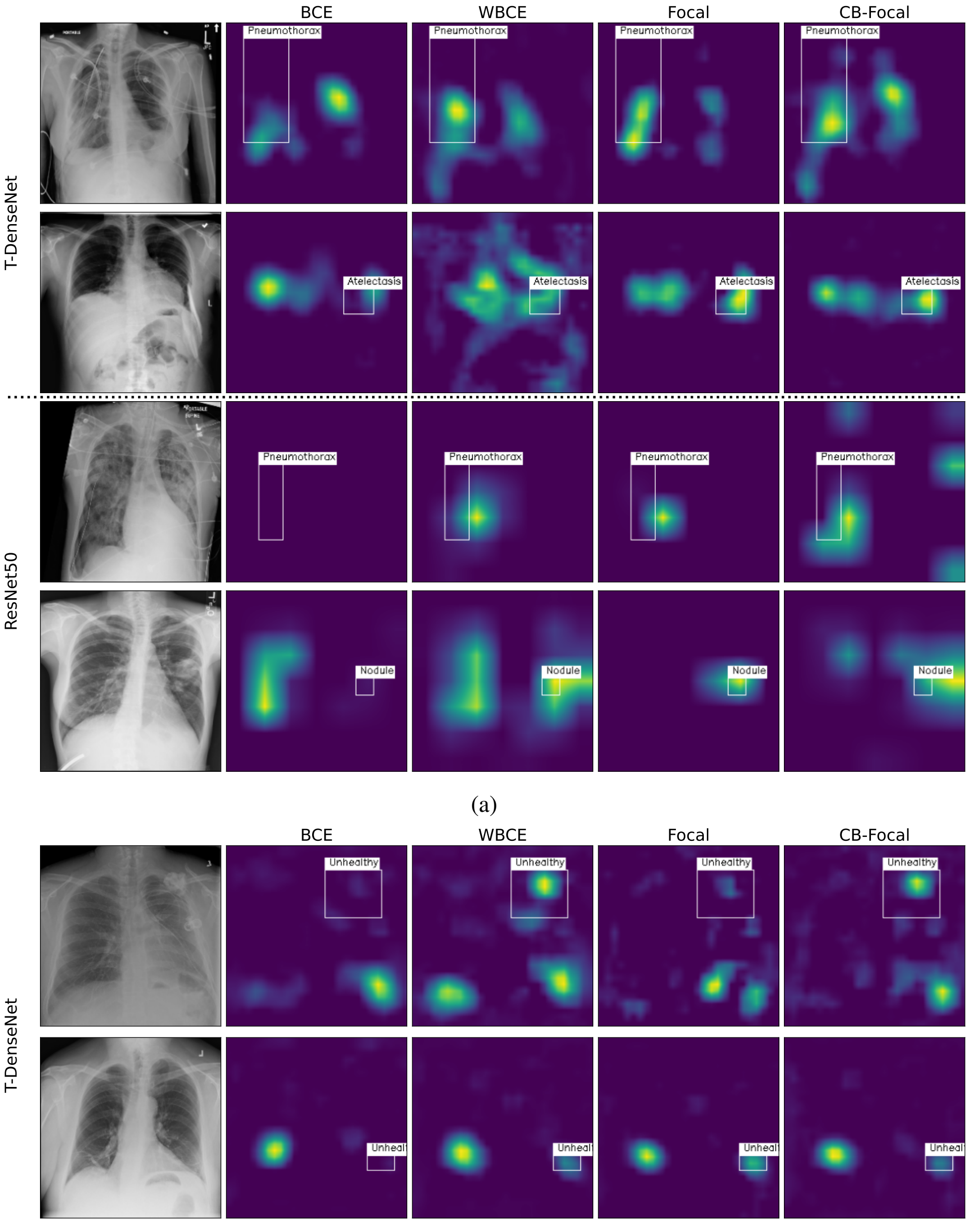
4.1. Experimental Setup

In this paper, we examine and perform two different datasets and learning tasks: multi-label classification and binary classification on the NIH Chest X-ray dataset (Wang et al., 2017). For binary classification. We define samples with the “No Finding” label as “Healthy” samples, and samples with other labels as “Unhealthy” samples. Random crop, horizontal flip, and color jitter are used as the augmentations. In the end, the augmented images are resized to 224×224 before being fed into an image classifier.

We use two ImageNet pre-trained classifiers: ResNet50 (He et al., 2016) and DenseNet (Huang et al., 2017). The last feature maps of ResNet50 have the size 7×7. In order to have a fine-grained analysis on smaller features such as Nodule (for which 7×7 is a coarse resolution) we use a truncated version of DenseNet. We discard the last two dense blocks and the associated transition blocks, so that the last feature maps have size 28×28. We refer to this truncated variant of DenseNet as *T-DenseNet*. We adopt the same training configurations for all models. Specifically, we train the classifiers on 4 GPUs (DGX-A100) for 50 epochs using the Adam (Kingma & Ba, 2015) optimizer with weight decay 10^{-6} and initial learning rate 0.0004. The learning rate is decayed following the cosine-annealing policy. In addition, we set the batch size to 512. For the Focal loss, the γ and α set to 2.0 and 0.25, respectively; for the class-balanced focal loss, we set γ to 2.0 and β to 0.9999.

4.2. Feature Alignment Analysis

The objective is to have a metric for measuring the alignment between the contributing features (from CAMs) for a single prediction and the pathologies. This can be achieved by classical object detection metrics IoBB(Intersection over Bounding Box) and IoR(Intersection of detected Region). IoBB and IoR can be considered as the visual counterparts of recall and precision. Computing these metrics requires thresholding the CAMs. In order to reduce the sensitivity of the results to the chosen threshold, we implement soft IoBB and IoR computation. This is performed by normalizing the CAMs to $[0, 1]$ and applying a summation of values within the regions under consideration (intersection, bounding box, detected region).



(a)

(b)

Figure 1. **Feature Alignment with Pathologies.** (a) multi-label classification task (b) binary classification task – Each row shows CAMs of different models on a single sample (the heatmap values are normalized to $[0, 1]$). Each column is associated with a training strategy (BCE, WBCE, Focal, CB-Focal). We observe that increased activation values due to training with cost-sensitive approaches appear primarily in areas where the pathologies exist. To quantitatively show this effect on the entire dataset, we report IoBB values in Tab. 1.

4.3. Analysis of Concepts

We count the number of concepts within bounding boxes in two ways: **Disjoint:** Number of concepts detected by a feature map (i.e. connected components overlapping the bounding box) for all feature maps in the chosen layer and all images in the annotation subset and normalize the value by the number of images. This approach considers repeated concepts in a bounding box, as some bounding boxes cover huge regions where multiple concepts occur. **Unique:** For each feature map, if there is at least one concept detected, we consider the feature map as a unique concept detector (similar to (Bau et al., 2017)). We count the number of unique concept detectors for all images and normalize by the number of images. The provided metrics reflect how individual feature maps are aligned with pathologies.

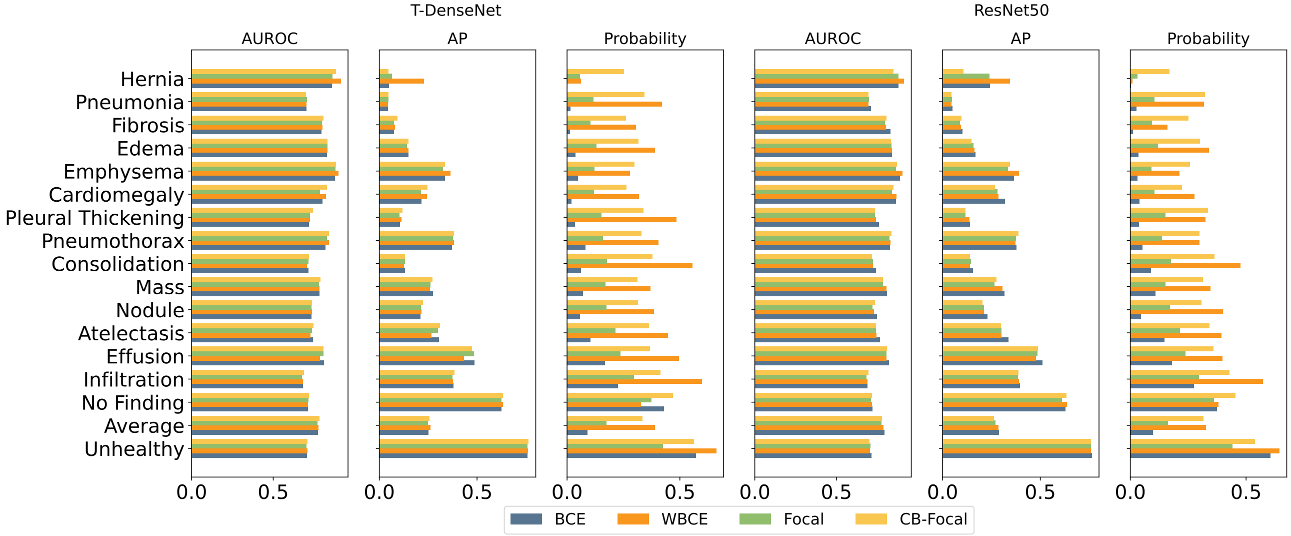


Figure 2. **Performance Analysis.** The sub-figures $1 \sim 3$ (from left to right) and $4 \sim 6$ show the results of the T-DenseNet and ResNet50, respectively. In addition to AUROC and AP, sample-wise averaged predicted probabilities are shown. In each sub-figure, "Average" indicates the class-wise average of the corresponding metric in multi-label classification; "Unhealthy" refers to the positive class in binary classification task. We observe that AUCROC and AP for different learning strategies are similar. Thus these metrics do not tell the entire story.

5. Results and Discussion

5.1. Performance Analysis

We evaluate the classification performance using AUROC and AP (Fig. 2). The performance differences between BCE and weighted losses (WBCE, Focal, and CB-Focal) are marginal. Therefore the effect of handling data imbalance via losses is not visible via these metrics. We also report the predicted probabilities of different losses. The predicted probabilities of the weighted losses are substantially higher than that of the BCE. This is expected as more weight is given to positive samples. Moreover, as more weight is given to the positive weights, the recall is higher. However, it is not clear if the increased recall is due to model's inclination towards identifying samples as positive or due to the model having learned and leveraging more predictive features.

5.2. Feature Alignment Analysis

In this section, we analyze the models' internal feature maps to see if the increased recall and probability values are related to predictive features relevant to the pathologies. This is realized by comparing CAMs and the salient features for individual predictions with annotations from radiologists. In Fig. 1 we observe that in models trained with cost-sensitive losses, the additionally activated regions are aligned with the pathologies (bounding boxes). This shows that the increased recall is not due to the models' propensity towards identifying samples as positive but due to having learned meaningful predictive features. Note that metrics such as AUROC and AP are roughly equivalent for all models (Fig. 2). In order to quantify the observed behavior in Fig. 1 we report IoBB and IoR results in Tab. 1. The increased IoBB for models trained with cost-sensitive losses shows that the features cover more areas of bounding boxes (pathologies). We also observe a decrease in IoR, which is also evident in Fig. 1 as falsely positive active regions have also increased.

5.3. Analysis of Concepts

This section analyzes the effect of handling data imbalance on the concepts encoded by the model. This is performed by counting the number of detected pathology-related concepts. In Tab. 2 we observe that the models trained with cost-sensitive losses consistently have more detectors. Thus it can be concluded that the models encode more pathology-related concepts when the data imbalance is considered in the loss.

6. Conclusion

In this work, we study the effect of handling data imbalance using cost-sensitive losses during training on the learned feature maps. The feature maps are analyzed from two perspectives: class activation maps and the concept encoded by each feature map. We observe that although classical met-

Table 1. **Feature Alignment Analysis.** The average IoBB and IoR results for test sets images that have bounding box annotations. These metrics are provided to quantitatively validate the observation in Fig. 1 for the entire test set.

TASK	MULTI-LABEL				BINARY			
CLASSIFIER	T-DENSENET		RESNET50		T-DENSENET		RESNET50	
METRIC	IOBB	IOR	IOBB	IOR	IOBB	IOR	IOBB	IOR
BCE	0.2105	0.2951	0.2700	**0.2401**	0.2661	**0.2123**	0.3159	**0.2084**
WBCE	**0.2836**	0.2252	**0.3447**	0.2060	**0.2765**	0.1999	**0.3505**	0.2021
FOCAL	0.1855	**0.3223**	0.2915	0.1845	0.2129	0.2064	0.2516	0.2042
CB-FOCAL	0.2753	0.3141	0.2839	0.2375	0.2458	0.1956	0.3409	0.1955

Table 2. **Analysis of Concepts.** The normalized number of detected concepts (Unique and Disjoint) in the chosen convolutional layers of DenseNet and ResNet models for the two tasks. Handling data imbalance consistently increases the number of concepts encoded by the model.

TASK	MULTI-LABEL				BINARY			
CLASSIFIER	T-DENSENET		RESNET50		T-DENSENET		RESNET50	
METRIC	DISJOINT	UNIQUE	DISJOINT	UNIQUE	DISJOINT	UNIQUE	DISJOINT	UNIQUE
BCE	217	141	562	328	280	165	363	278
WBCE	259	157	**641**	441	287	169	542	386
FOCAL	226	143	585	**570**	291	170	**586**	**414**
CB-FOCAL	**263**	**163**	546	390	**292**	**171**	567	400

rics such as AUROC and AP report equivalent performance for the studied models, the learned features are different in these models. When data imbalance is handled, the models encode more pathology-related concepts. Moreover, we observe that overall increased recall and predicted probability of these models are accompanied by an increased alignment between learned features and pathology-related features.

Acknowledgement This work has been funded in part by the German Federal Ministry of Education and Research (BMBF) under Grant No.01IS18036A and 01IS18036B, Munich Center for Machine Learning (MCML). A. K., N. N. were supported with funding from the Bundesministerium fur Bildung und Forschung (BMBF). M. R., and B. B. were supported by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. Y. L. is supported by the German Federal Ministry of Health (2520DAT920). S.T. K. is supported by National Research Foundation of Korea(NRF) grant funded by the Korea Government(MSIT) (No. 2021R1G1A1094990).

References

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Cong, R., Lei, J., Fu, H., Cheng, M.-M., Lin, W., and Huang, Q. Review of visual saliency detection with comprehensive information. *IEEE Transactions on circuits and Systems for Video Technology*, 29(10):2941–2959, 2018.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 9737–9748, 2019.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Khakzar, A., Albarqouni, S., and Navab, N. Learning interpretable features via adversarially robust optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 793–800. Springer, 2019a.

Khakzar, A., Baselizadeh, S., Khanduja, S., Rupprecht, C., Kim, S. T., and Navab, N. Improving feature attribution through input-specific network pruning. *arXiv preprint arXiv:1911.11081*, 2019b.

Khakzar, A., Baselizadeh, S., and Navab, N. Rethinking positive aggregation and propagation of gradients in gradient-based saliency methods. *arXiv preprint arXiv:2012.00362*, 2020.

Khakzar, A., Baselizadeh, S., Khanduja, S., Rupprecht, C., Kim, S. T., and Navab, N. Neural response interpretation through the lens of critical pathways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13528–13538, June 2021a.

Khakzar, A., Musatian, S., Buchberger, J., Valeriano Quiroz, I., Pinger, N., Baselizadeh, S., Kim, S. T., and Navab, N. Towards semantic interpretation of thoracic disease and covid-19 diagnosis models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 499–508. Springer, 2021b.

Khakzar, A., Zhang, Y., Mansour, W., Cai, Y., Li, Y., Zhang, Y., Kim, S. T., and Navab, N. Explaining covid-19 and thoracic pathology model predictions by identifying informative input features. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 391–401. Springer, 2021c.

Khakzar, A., Khorsandi, P., Nobahari, R., and Navab, N. Do explanations explain? model knows best. *arXiv preprint arXiv:2203.02269*, 2022.

Khan, S., Hayat, M., Zamir, S. W., Shen, J., and Shao, L. Striking the right balance with uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 103–112, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., and Lakkaraju, H. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Lundberg, S. M. and Lee, S. I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Rezaei, M., Yang, H., and Meinel, C. Generative adversarial framework for learning multiple clinical tasks. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8. IEEE, 2018.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, 2021.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *34th International Conference on Machine Learning, ICML 2017*, 2017. ISBN 9781510855144.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. 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*, pp. 2097–2106, 2017.

Zhang, Y., Khakzar, A., Li, Y., Farshad, A., Kim, S. T., and Navab, N. Fine-grained neural network explanation by identifying input features with predictive information. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.