2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Deep-ReAP: Deep Representations And Partial label learning for Multi-pathology Classification

Sohini Roychowdhury
Director ML Curriculum, FourthBrain,
Adjunct Faculty, Santa Clara University,
Email: sohini@fourthbrain.ai

***Abstract*—Automated detection of pathology in images with multiple pathologies is one of the most challenging problems in medical diagnostics. The primary hurdles for automated systems include data imbalance across pathology categories and structural variations in pathological manifestations across patients. In this work, we present a novel method to detect a minimal dataset to train deep learning models that classify and explain multiple pathologies through the *deep representations*. We implement partial label learning with 1% false labels to identify the under-fit pathological categories that need further training followed by fine-tuning the deep representations. The proposed method identifies 54% of available training images as optimal for explainable classification of upto 7 pathological categories that can co-exist in 36 various combinations in retinal images, with overall precision/recall/F_β scores of 57%/87%/80%. Thus, the proposed method can lead to explainable inferencing for multi-label medical image data sets.**

***Index Terms*—Deep Learning, Semi-supervised Learning, Label Propagation, Partial label learning, multi-label classification.**

I. Introduction

Computer-aided diagnostics for retinal pathology has been a well studied topic for over a decade now [1]. With the surge in transfer learning applications to transfer weights from pre-trained deep learning (DL) layers to new data sets, the need for "new labelled data" has significantly reduced. However, the medical domain continues to suffer from the "*small data challenge*", wherein, there is a lack of quality annotated data. Additionally medical data often suffers from inter-observer variability, wherein multiple annotators may arrive at different conclusions regarding the pathological state for the same patient owing to poor quality images or the presence of multiple pathologies [1]. In this work, we present a novel method to maximize the learning from the DL feature extraction layers and to detect data dependencies while being robust to observer variabilities, or false labels.

Prior works in [2] [3] refer to DL networks as a combination of a *trainable feature extractor* followed by a classification layer. Motivated by these existing works, we analyze the impact of training data on these two DL components separately. The three-step method presented in this work aims to optimally train the DL feature extractor for multi-label classification. Additionally, we identify the under-fit pathological categories that require additional DL training and assess the fine-tuning capability of single pathology training samples for classification of multiple pathology images.

This paper makes two key contributions. First, we identify a minimal training image set to classify and explain multiple pathologies in retinal images. We apply partial label learning (PLL) [4] [5] to identify the under-fit pathology categories (UPC) followed by retraining the DL feature extractor on additional 100 images from UPC. Second, we present a framework for multi-label image classification method using optimally trained *deep representations* ($\mathcal{D}$) such that the proposed system achieves categorical recall and F_β scores in the range of 68-98% and 56-88%, respectively. Fig. 1 shows the proposed three-step method, namely extraction of $\mathcal{D}$ per image from the dense layer, followed by semi-supervised label propagation (LP) with limited false labels to identify UPC images. Finally, a combination of propagated labels and training data is used to retrain the a DL network to classify and visually explain regions of interest (ROIs) corresponding to pathological sites.

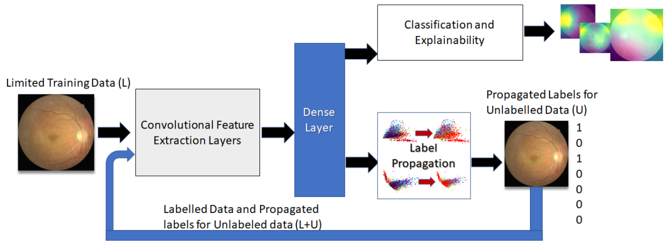

Classification and Explainability
Limited Training Data (L)
Convolutional Feature Extraction Layers
Dense Layer
Label Propagation
Propagated Labels for Unlabelled Data (U)
1
0
1
0
0
0
0
Labelled Data and Propagated labels for Unlabeled data (L+U)

Fig. 1. The proposed method to train deep representations from the dense layer for explainable multi-label classification.

II. Methods and Materials

So far, incremental learning methods for multi-class classification problems have been analyzed in [2] [3] with the intention to process data in incremental small batches. In [2], a knowledge distillation component in the overall loss function enables weighted learning for the deep representations and subsequent classification. However, in multi-label scenarios, where several combination classes can occur in incremental batches, and when individual classes can manifest in a wide variety of ways, the weighted training approach for deep representations in batches requires further modification. For instance, the retinal pathology of diabetic retinopathy (DR) can be detected by the presence of many small red dots (micro-aneurysms) or combination with bright lesions (hard exudates) and red haemorrhages in different patients. Thus,

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

balancing the number of images with DR with respect to other individual pathology samples may not extract all representative features for DR. The proposed approach of fine-tuning deep representations enables weighted learning for multi-label classification instances. [1] Explanations of the data set, mathematical framework and methods are presented below.

A. Data: Viet AI Retinal Challenge

The publicly available VietAI data set [6] contains 3435 annotated retinal images and 350 un-annotated images of size $[512 \times 512 \times 3]$ each, such that the labels are representative of multiple pathology categories namely: Opacity, DR, Glaucoma, Macular Edema (ME), Macular Degeneration (MD), Retinal Vein Occlusion (RVO), or Normal. Exploratory data analysis shows that the top 5 highly occurring images belong to categories of Opacity, Normal, Glaucoma, MD and DR. Also, 36 unique label combinations from the 7 pathological categories and normal category exist. Here, we split the annotated data set by 80/20, such that representations ($\mathcal{D}$) from 2748 images correspond to the training data $(\mathcal{X}, Y)$, and the remaining 687 images from the annotated set are test data $(\mathcal{X}', Y')$. Finally, $\mathcal{D}$ from the 350 unlabelled images for which labels will be generated by semi-supervised learning are in $\mathcal{X}^*$. Here, $\mathcal{D}$ represent DL output from an additional dense layer applied after global average pooling from the final convolutional layers.

B. Mathematical Framework

For our analysis, we begin with a minimum subset of the training data that can provide similar $\mathcal{D}$ using a DL model as the complete training data set. This minimum subset or fraction-f of all training images with l number of samples is referred to as the labelled data (L) with samples $\mathcal{X}_L^f$ and corresponding labels Y_L^f. The remaining images from train data become a part of an unlabelled set (U), where U contains u number of samples from the annotated ($\mathcal{X}_U^{1-f}$,Y_U^{1-f}) and un-annotated sets $\mathcal{X}^*$ combined. As a first step, we analyze the DL classification performance using L. Next, for UPC identification, we apply *rbf kernel*-based label propagation (k-LP) [4], where, the training labels are propagated in the $\mathcal{D}$ feature space to generate labels for unlabelled samples. Here, we prefer k-LP over nearest neighbor methods since prior work in [7] has shown that k-LP minimizes adaptive representation and classification errors in kernel space.

k-LP initiates by the estimation of a weight affinity matrix (W) for all unlabelled samples (x_i) with respect to the labelled samples (x_j) as follows.

$$W_{ij} = e^{-\gamma||x_i - x_j||^2} \geq 0, \forall x_i \in U, x_j \in L, \qquad (1)$$

where, γ is an input kernel parameter. Next, a normalized Laplacian $(\mathcal{L})$ is computed from the diagonal matrix (D) that uses W as follows.

$$D_i = \sum_j W_{ij}, \quad \mathcal{L} = D^{-(1/2)} W D^{-(1/2)}. \qquad (2)$$

Finally, label spreading is performed based on [7], wherein the initial state includes $l + u$ number of labels corresponding to the labelled samples $(y_1, y_2...)$ and the unlabelled samples (assigned to placeholder label -1) referred to as $\hat{\mathcal{Y}}^{(0)} = \{y_1, y_2 .... y_l, -1, -1 ....\}$. Next, the steps in (3-4) are repeated for t iterations or until convergence.

$$\text{Iterate }, \hat{\mathcal{Y}}^{(t+1)} = \alpha \mathcal{L} \hat{\mathcal{Y}}^{(t)} + (1 - \alpha)\hat{\mathcal{Y}}^{(0)}, \qquad (3)$$

$$\forall t = [0, \infty], x_i \longrightarrow \hat{y}_i^{\infty}.$$

$$\hat{Y}_U^{1-f} = \hat{\mathcal{Y}}^{\infty}_{[l+1, l+u]}. \qquad (4)$$

In (3), α is parameterized in range [0,1], such that a large value indicates iterative label modifications while a small value indicates initial label retention. The k-LP process is shown in Fig. 2, wherein the unlabelled samples (black color in middle column) are assigned a propagated label at the end of (3-4).

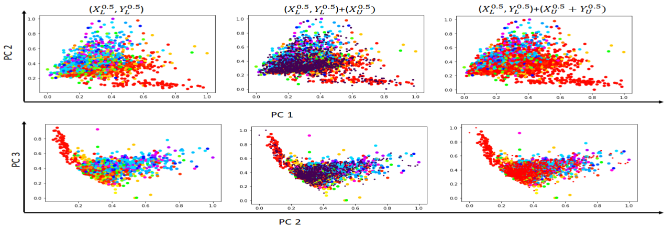


Fig. 2. Example of k-LP to generate labels for the samples U. PC1, PC2, PC3 represent the top 3 principal components for $\mathcal{D}$. Column 1, 2, 3 represent samples from L, $L + U$, L and propagated U labels, respectively.

Next, we compare the propagated labels for samples $\mathcal{X}_U^{1-f}$ that are referred to as $\hat{Y}_U^{1-f}$, with their actual labels Y_U^{1-f}. The pathology categories with the most error are thereby identified as UPC. Next, we extract a limited data subset for DL model fine-tuning from the remaining training data for samples corresponding to UPC and some propagated samples for set $\mathcal{X}^*$. The DL model is thus fine-tuned and analyzed for classification performance and explainability.

The multi-label classification metrics evaluated on the test data with n samples are in terms of correctly classified pathology or true positives (tp), falsely classified pathology or false positives (fp) and missed pathology or false negatives (fn) as defined in (5-6). Here, test data labels $y_i \in Y'$ and $\hat{y}_i$ is the classification outcome vector with $C = 7$ dimensions. The performance of trained representations $\mathcal{D}$ for classifying multiple pathologies is analyzed in terms of average precision (Pr), recall (Re), F_1 and F_β scores defined in (7-8). Since fn are more detrimental than fp in medical diagnostics, $\beta = 2$ for (8) to ensure higher weightage to Re than Pr.

$$tp = \sum_i^n (y_i^T \hat{y}_i), fp = \sum_i^n \sum_{c=1}^C \{(\hat{y}_i(c) - y_i(c)) \geq 0\}, \qquad (5)$$

$$fn = \sum_i^n \sum_{c=1}^C \{(y_i(c) - \hat{y}_i(c)) \geq 0\}. \qquad (6)$$

$$Pr = \frac{tp}{tp + fp}, Re = \frac{tp}{tp + fn}, \qquad (7)$$

$$F_1 = \frac{2Pr.Re}{Pr + Re}, F_\beta = \frac{(1 + \beta^2)Pr.Re}{(\beta^2 Pr) + Re}. \qquad (8)$$

[1] Github: https://github.com/sohiniroych/Deep-ReAP-Multi-label

The three steps in the proposed system are described below.

C. Step 1: Initial Training of Deep Representations

The first step is training a DL model to extract network weights and biases that can be combined with the individual input images to extract $\mathcal{D}$ per sample. We implemented several DL pre-trained models such as ResnetV2, Resnet50 and InceptionV3, and observed that using the complete training data set with 2748 images and hyper-parameterization with 80/20 cross-validation split for learning rate using Adam optimizer in range $[10^{-5} - 10^{-3}]$ on all these models resulted in similar F_β scores in the range [0.70-0.723]. This observation further aligns with the prior submission in [6]. Thus, we select the InceptionV3 model with the highest F_β to train and analyze the classification capability of $\mathcal{D}$. For DL training, the input images are resized to [224 × 224] and zoom, width and height shift augmentations are applied to the images. Training proceeds with batch sizes of 20, for 40 epochs, with binary cross entropy loss function and monitoring the F_β score.

To identify a smaller training data subset with similar classification performance as the complete training data, we apply random stratified sampling to select a fraction f of each multi-label combination from the training data. We vary $f = [0.2, 0.3, 0.4, 0.5, 0.6]$ and retrain the InceptionV3 model with an additional dense layer with 512 neurons and a 7 neuron classifier layer. We observe that for $f = 0.5$, $F_\beta = 0.722$, which is the closest in classification performance to the complete data set. Thus, we select these 1374 images (at 50% representation) as the *initial training dataset* ($\mathcal{X}_L^f$,Y_L^f).

D. Step 2: UPC Detection by kernel-PLL

To identify the pathology categories that need further training (or UPC) we apply the PLL framework wherein we randomly drop sample labels and apply k-LP, as shown in [4], such that 1% of the labels from the set L are false-labels. To achieve this, we randomly select 28 images (1% of training data) from $\mathcal{X}_L^f$ and randomly add to or remove from a pathology category in the label. For instance, a label [1,0,0,1,1,0,0] can become [0,0,0,1,1,0,0] or [1,0,1,1,1,0,0], etc., in different sample runs. Next, labels from L are randomly removed and substituted for a vector with values -1 to depict dropped label samples. We vary the fraction of randomly dropped labels from L as $p = [0.1, 0.2, ..0.8]$, and for each sample run, we apply k-LP to obtain propagated labels for all the dropped samples based on the work in [5]. The averaged Pr, Re and F_β metrics per pathology category are then analyzed to detect UPC as categories that either have the highest variation across changing values of p or that have consistently low F_β scores.

The PLL method with k-LP and a few false labels analyzes the realistic scenario when inter-observer variability between manual annotators can propagate through an automated learning system. This analysis isolates the contributions of the learned $\mathcal{D}$ from the classification process for robust identification of UPC.

E. Step 3: Fine-tuned Deep Representation Analysis

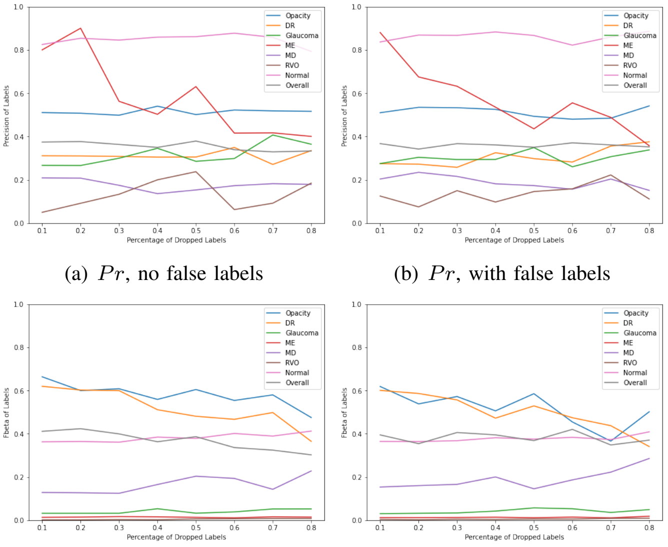
Images from the under-fit categories are isolated from the ($\mathcal{X}_U^{1-f}$, Y_U^{1-f}) data set used for fine-tuning the DL model from Section II-C. The goal is to improve classification performances for all pathologies. Representations $\mathcal{D}$ for test data images obtained after fine-tuning are then analyzed quantitatively and qualitatively to explain the pathological ROIs using the Gradcam and tf_explain libraries in Python.

III. Experiments and Results

In this work, we perform three major experiments. First, we identify a subset UPC samples using k-PLL with false labels. Second, we analyze the classification performances with InceptionV3 model using the initial training dataset and after fine-tuning with the UPC images. Third, we qualitatively assess the ROIs for explainability of multiple patholgies in retinal images.

A. PLL for Under-fit Data Detection

In this experiment, we implement PLL as shown in Section II-D with and without the introduction of false labels. The optimal $\gamma = 5$ for rbf-kernel upon cross-validation. The variations in Pr, Re and F_β per-category are analyzed for varying proportions (p) of dropped labels in Fig. 3. Here, we observe that the PLL curves for $p \leq 0.5$ have relatively similar trends with and without false labels. This demonstrates that per-pathology sample clusters are well-defined and robust against inter-observer variability.



(a) Pr, no false labels

(b) Pr, with false labels

(c) F_β, no false labels

(d) F_β, with false labels

Fig. 3. PLL curves for Pr and F_β scores across variations in proportions of dropped labels. Re curves have similar trends as F_β score.

Also we observe that Normal and Opacity categories are most stable across variations in p for all the metrics. This implies that sample $\mathcal{D}$s are well trained to identify normal retinal images and images with opacity that may have a small disc-like appearance around the macula and blurry image quality. Also, we observe that categories of Glaucoma, ME and RVO have the least Re and F_β metrics and can be referred to as the UPC.

B. Multi-label Classification Performance

As a next step we identify an additional fine-tuning training set to contain $\{50,25,25\}$ additional images with $\{$Glaucoma, ME, RVO$\}$, respectively. Additionally, for randomization, we add 20 random images to this set as the propagated labels from the k-LP process for the image set $\mathcal{X}^*$. The averaged classification performance for the initial training set and the combination of initial training and fine-tuning sets across 20 random runs are shown in Table I. The number of training images corresponding to each individual pathology category are also shown here. Thus, if a training image contains Opacity and Glaucoma, then the same image will be counted twice, once for the category Opacity and once for Glaucoma, respectively. From Table I, we observe a significant improvement in overall classification performance metrics with about 7% and 9% improvements in F_β and Re, respectively. We observe a categorical increase in Re for all but DR, and the F_β increment for UPCs, namely, Glaucoma, ME and RVO to be 7%, 6%, 28%, respectively. Thus, PLL-based identification of underfit images aids detection of a minimal training image set for multi-pathology classification.

TABLE I
AVERAGE CLASSIFICATION PERFORMANCES BY VARYING TRAINING DATA.

Category	# images	Precision	Recall	F_1	F_β
Data: Initial Training Set					
Overall	1374	0.5399	0.7887	0.641	0.7222
Opacity	631	0.6495	0.9205	0.7616	0.8496
DR	283	0.5963	0.7471	0.6633	0.7112
Glaucoma	244	0.3973	0.8018	0.5313	0.6662
ME	215	0.4678	0.7407	0.5735	0.6633
MD	225	0.3913	0.9153	0.5482	0.7219
RVO	182	0.4468	0.2593	0.3281	0.283
Normal	218	0.8864	0.78	0.8298	0.80
Data: Fine-tuned Training Set					
Overall	1494	0.567	0.8773	0.6888	0.8007
Opacity	646	0.6245	0.9801	0.7629	0.8799
DR	285	0.5426	0.8046	0.6481	0.7338
Glaucoma	300	0.5	0.8288	0.6237	0.7325
ME	231	0.4185	0.8796	0.5672	0.7208
MD	229	0.5389	0.822	0.651	0.7439
RVO	211	0.3293	0.679	0.4435	0.5601
Normal	222	0.7681	0.94	0.8393	0.9069

C. Qualitative Assessment: Pathology Explainability

Finally, the fine-tuned weights from the InceptionV3 models are used to visualize the ROIs for test images in Fig. 4. In these visualizations, deeper yellow color represents concentration of features or ROIs. Thus, multiple pathology images can be classified and explained by fine-tuning on individual pathology images for sensitive categories.

IV. Conclusions

In this work, we identify an optimal training dataset with minimal number of images to train a DL model for classification and explain-ability of retinal images with multiple pathologies. The novel framework includes the use of deep representations learned from DL models and the semi-supervised approach of label propagation to identify categories of retinal images that need further training. The proposed framework is capable of increasing overall classification Pr/Re by 2.5%/9% by using about 54% of all training samples when compared to the complete training data set itself. Visualizations of single and multiple pathologies are shown in Fig. 5. Here, we observe that for single and multiple pathologies, the ROIs help identify the most prominent pathology first (opacity in this case) followed by the next prominent ones. Future works can be directed towards applying correlated label propagation and variants of iterative classification methods [2] to extend the proposed framework to other medical images.

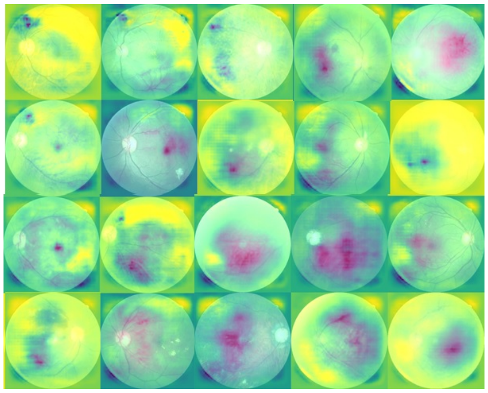
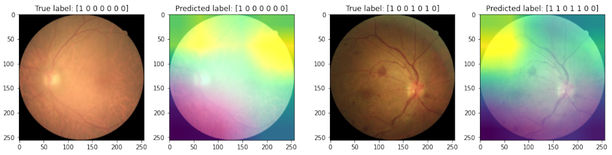


Fig. 4. Examples of visualizations for test images with multiple pathologies.


True label: [1 0 0 0 0 0 0]
Predicted label: [1 0 0 0 0 0 0]
True label: [1 0 0 1 0 1 0]
Predicted label: [1 1 0 1 1 0 0]
0
50
100
150
200
250

(a) Image with Opacity only.
(b) Image with Opacity and DR.

Fig. 5. Examples of visualized pathologies to gauge detection preferences.

References

[1] D. S. Wanderley, T. Araújo, C. B. Carvalho, C. Maia, S. Penas, Â. Carneiro, A. M. Mendonça, and A. Campilho, "Analysis of the performance of specialists and an automatic algorithm in retinal image quality assessment," in *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*. IEEE, 2019, pp. 1–4.

[2] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[3] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 233–248.

[4] K. Sun, Z. Min, and J. Wang, "Pp-pll: Probability propagation for partial label learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 123–137.

[5] S. Roychowdhury, E. Alareqi, and W. Li, "Opam: Online purchasing-behavior analysis using machine learning," *IEEE, IJCNN*, 2021.

[6] C. T. E. Hospital, "Vietai advance course - retinal disease detection," 2020. [Online]. Available: https://www.kaggle.com/c/vietai-advance-retinal-disease-detection-2020/overview

[7] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in neural information processing systems*, 2004, pp. 321–328.