Combining Graph Transformers Based Multi-Label Active Learning and Informative Data Augmentation for Chest Xray Classification

Dwarikanath Mahapatra[1], Behzad Bozorgtabar[2], Zongyuan Ge[3], Mauricio Reyes[4], Jean-Philippe Thiran[2]

[1]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
[2]EPFL, Switzerland
[3]Monash University, Australia
[4]University of Bern, Switzerland.

Abstract

Informative sample selection in active learning (AL) helps a machine learning system attain optimum performance with minimum labeled samples, thus improving human-in-the-loop computer-aided diagnosis systems with limited labeled data. Data augmentation is highly effective for enlarging datasets with less labeled data. Combining informative sample selection and data augmentation should leverage their respective advantages and improve performance of AL systems. We propose a novel approach to combine informative sample selection and data augmentation for multi-label active learning. Conventional informative sample selection approaches have mostly focused on the single-label case which do not perform optimally in the multi-label setting. We improve upon state-of-the-art multi-label active learning techniques by representing disease labels as graph nodes, use graph attention transformers (GAT) to learn more effective inter-label relationships and identify most informative samples. We generate transformations of these informative samples which are also informative. Experiments on public chest xray datasets show improved results over state-of-the-art multi-label AL techniques in terms of classification performance, learning rates, and robustness. We also perform qualitative analysis to determine the realism of generated images.

Introduction

Annotating medical images is necessary for state-of-the-art (SOTA) supervised learning methods, but poses challenges due to the high expertise and costs involved. Active Learning (AL) allows an expert to label informative samples which enable a model to have high performance with minimal labeled samples (i.e., high learning rates). Data augmentation is also effective with few labeled samples and synthetic data is used to improve model generalization (Perez and Wang 2017). Despite increasing dataset size, conventional data augmentation (flipping, rotating, etc.) does not add new informative samples to the training set. Recent works for data augmentation use Generative Adversarial Networks (GAN) (Goodfellow et al. 2014), Spatial Transform Networks (STN) (Jaderberg et al. 2015), Variational Autoencoder (VAE) (Kingma and Welling 2013), and Mixup (Zhang et al. 2017) for synthetic feature generation.

Combining data augmentation and active learning can leverage both approaches' advantages. (Tran et al. 2019) select informative samples by an acquisition function and generate augmented samples from them. However, the acquisition function does not evaluate the potential information gain from augmented samples and the generated data does not guarantee informativeness. Kim et al. in (Kim et al. 2021) propose Look Ahead Data Augmentation (LADA), that evaluates the informativeness of potential augmentations and generates informative samples. Thus, augmented samples provide qualitatively different information from the base samples. However, LADA is not equally effective for the multi-label setting. Chest xrays (CXRs) have multiple disease labels making informative sample selection a challenge since one needs to consider the mutual influence and similarity of all potential class labels,. Additionally, augmenting such informative samples should ensure appropriate informativeness of the new samples. In this paper we propose a novel method for multi-label active learning combined with informative sample data augmentation.

Prior Work

Active Sample Selection: Different informative sample selection approaches for deep learning based medical image analysis include sample entropy (Zhu and Bento 2017), model uncertainty (Gal, Islam, and Ghahramani 2017), Fisher information (Sourati et al. 2019), and clustering-based sample selection (Zheng et al. 2019). Sample entropy quantifies it's difficulty in classification, with higher entropy characterizing higher sample informativeness. (Wang et al. 2017a) use sample entropy, a least-confidence component, and margin sampling to select informative samples. (Zhou et al. 2016) use GANs to synthesize samples close to the decision boundary, which are then annotated by human experts. (Mayer and Timofte 2018) generate high entropy samples, which are used as a proxy to find the most similar samples from a pool of real sample candidates to be annotated by experts. The state-of-the-art in active learning is mostly dominated by methods relying on uncertainty estimations. Uncertainty-based methods identify the most informative samples for which a model is most uncertain. (Yang et al. 2017) propose a two-step sample selection approach based on uncertainty estimation, followed by a second selection step based on a maximum set coverage similarity met-

ric. Test-time Monte-Carlo dropout (Gal, Islam, and Ghahramani 2017) has been used to estimate sample uncertainty, and consequently select the most informative ones for label annotation (Bozorgtabar et al. 2019). (Mahapatra et al. 2021) propose an interpretability-guided sample selection approach featuring state-of-the-art performance for classification and segmentation tasks. However these approaches are not designed for multi-label classification problems.

A comprehensive survey on multi-label deep active learning can be found in (Wu et al. 2020). Some specific approaches include application to remote sensing images (Mollenbrok, Sumbul, and Demir 2023) and sub-example querying. However there is very limited work on ML active learning for medical image analysis, with (Reyes, Morell, and Ventura 2018) using a measure of inconsistency of a predicted label set to select the most informative samples.

Active Learning with Data Augmentation: Prior work on leveraging data augmentation for active learning includes Bayesian Generative Active Deep Learning (BGADL), which combines acquisition and augmentation steps in a pipelined approach (Tran et al. 2019). However, a large number of labeled instances are needed to train the generative model, and BGADL does not measure the potential information gain from data augmentation. Consistency-based Active Learning (CAL) algorithms consider data augmentation by replacing uncertainty with an augmentation-based inconsistency term. (Kim et al. 2021) propose look-ahead-data-augmentation (LADA) to select informative samples and also evaluate informativeness of generated samples, but does not perform accurately in a multi-label setting.

Contributions: Different from previous works, we combine data augmentation and informative sample selection in a multi-label setting with the following novelties: **1:** Using graph attention transformers to incorporate the importance of different nodes and quantify a graph's informativeness. Simple aggregation such as sum and mean of the weights give equal importance to all nodes and do not emphasize nodes with greater information that could have greater contribution to the task in hand. **2:** We propose a novel multi-label informativeness score, derived from graph attention transformers, that quantifies the importance of each sample based on multi-label interactions. **3:** A novel data augmentation approach to reduce redundancy in sample selection that takes base informative images (identified from the multi-label informativeness score) and generates novel transformations such that new images are informative compared to the base image.Our method is dubbed as DAMLAL (**D**ata **A**ugmentation with **M**ulti **L**abel **A**ctive **L**earning), and outperforms previous active learning approaches for chest X-ray classification.

Methods

Outline Of Proposed Method: Figure 1 depicts the different stages of our workflow. We identify multi-label informative samples using a multi-label sample informativeness measure by jointly considering the mutual influence of all potential class labels. New informative samples are synthesized from identified base informative samples to provide new information to the training set. The classifier model is then updated and these set of steps repeated till no new informative samples are found.

Figure 1: Workflow of proposed DAMLAL method. Given unlabeled samples a graph-based transformer is used with a novel metric to rank informative samples. Selected samples are used to synthesize more informative and non-redundant samples which are added to the training dataset for the next active learning cycle.

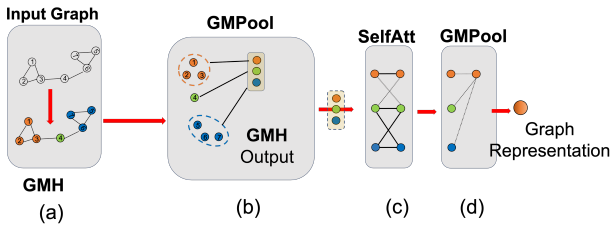

Figure 2: Graph Multiset Transformer: A graph with n nodes depicting multi-label information of a sample is passed through several message-passing layers (a) and an attention-based pooling block (GMPool$_k$) (b) to get $k(< n)$ nodes. A self-attention block (SelfAtt) (c) encodes the relationship between k nodes, and passes through GMPool$_1$ (d), to obtain a single node value. Different node colors indicate different classes and the edge length denotes node similarity.

Multi-Label Sample Informativeness

To identify most informative samples in multi-label settings, we incorporate class-label interactions using a graph-based ranking metric scheme that leverages graph transformers, and a graph pooling approach to learn better global relationships among graph nodes. We model each sample as follows. Assuming K nodes in each graph (corresponding to K classes), we define edge weights between nodes i, j, as $w_{ij} = cosine\_similarity(z_{S_{I,i}}, z_{S_{I,j}})$. $z_{S_{I,i}}$ and $z_{S_{I,j}}$ are latent feature vectors derived from saliency maps S of sample image I for class labels i and j, which have shown to perform better than using feature maps in classification due to their intrinsic focused attention mechanism (Mahapatra, Poellinger, and Reyes 2022b). Cosine similarity is used since its range is bounded and suitable as a loss term.

Graph Transformers: Graph pooling is important to obtain a lower dimensional informative representation of graphs. While Node drop methods (Zhang et al. 2018; Lee, Lee, and Kang 2019) and Node clustering methods (Ying

et al. 2018; Bianchi, Grattarola, and Alippi 2019) are popular, they tend to drop informative nodes at every pooling step, and have high computational complexity respectively. To overcome these limitations, we use a Graph Multiset Transformer (GMT) (Baek, Kang, and Hwang 2021) attention unit. GMT pooling condenses a graph into a set of representative nodes, and then encodes their relationships to enhance the representation power of the graph (Figure 2). We refer the readers to (Baek, Kang, and Hwang 2021) for full details and provide a short summary below. A simple sum pooling over a graph does not consider each node's relevance to the task and treats them equally. This limitation is addressed using an attention mechanism on the multiset pooling function to better capture structural dependencies among nodes.

Graph Multi-head Attention: Assuming n node vectors, input query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$ the output of the attention stage is: $Att(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \phi(\mathbf{Q}\mathbf{K}^T)\mathbf{V}$. Multi-head attention (Vaswani et al. 2017) can be used by linearly projecting $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, h times respectively to yield h different representation subspaces. To enable better representation learning, we instead use Graph Multi-Head attention (GMH) to generate key and value pairs using Graph Neural Networks (GNNs) as: $GMH(\mathbf{Q}, \mathbf{H}, \mathbf{A}) = [O_1, \cdots, O_h]\mathbf{W}^O$; where $O_i = Att(QW_i^Q, GNN_i^K(\mathbf{H}, \mathbf{A}), GNN_i(\mathbf{H}, \mathbf{A}))$. The output of GNN_i contains neighboring information of the graph, and improves over previously used linear node embeddings for key and value matrices. Given node features $\mathbf{H}$ from GNNs, a Graph Multiset Pooling (GMPool$_k$) compresses the n nodes into $k(< n)$ nodes. To quantify the interactions among graph nodes, a Self-Attention function (SelfAtt), inspired by the Transformer architecture (Vaswani et al. 2017; Lee et al. 2019) is used. SelfAtt captures inter-relationships among n nodes by using node embeddings H on both query and key locations of GMH.

Overall Architecture: For a graph G with node features $\mathbf{X}$ and an adjacency matrix $\mathbf{A}$, the Encoder is denoted as:

$$Encoder(X, A) = GNN_2(GNN_1(\mathbf{X}, \mathbf{A}), \mathbf{A}), \quad (1)$$

where we stack 2 GNNs to construct the deep structures, although more GNNs may be stacked depending on the application. After obtaining a set of node features $\mathbf{H}$ from an encoder, the pooling layer aggregates the features into a single vector. Finally, we obtain the entire graph representation by using GMPool with $k = 1$ as follows:

$$\begin{aligned} MIS &= Pooling(H, A) \\ &= GMPool_1(SelfAtt(GMPool_k(H, A)), A'), \end{aligned} \quad (2)$$

where $\mathbf{A}'$ is the identity or coarsened adjacency matrix, adjusted after compressing the nodes from n to k with $GMPool_k$. The above step condenses the entire graph and represents it as a single value: the multi-label informativeness score MIS with higher values indicating higher sample informativeness. To verify whether the value of MIS truly reflects sample informativeness, we select 500 images at random and calculated their MIS and uncertainty values (most commonly used sample informativeness metric), and generated a scatter plot of the two metrics (Figure 3). We found a strong correlation coefficient of 0.92 showing that the proposed MIS also measures sample informativeness.

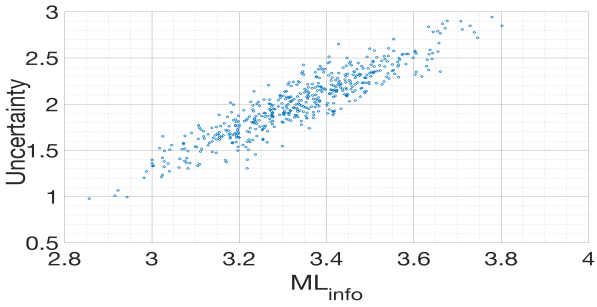

Figure 3: Scatter plot between ML_{info} and uncertainty values of 500 samples. The points are concentrated around a line with a high correlation coefficient (0.92), suggesting that ML_{info} is a good measure of sample informativeness.

Variational Auto Encoder Training

The VAE encoder outputs two vectors of same size: mean vector, $\vec{\mu}$, and a standard deviation vectors, $\vec{\sigma}$. Given an encoding of an image, one can generate variations of it by sampling from the parameters $\vec{\mu}$ and $\vec{\sigma}$. For smooth interpolation amongst encodings and construction of new samples, a cost function defined as a combination of the reconstruction error term and the Kullback–Leibler (KL) divergence is commonly used:

$$\mathcal{L}_{VAE} = \sum \|x - \hat{x}\|^2 + KL\left[\mathcal{N}(\mu_x, \sigma_x), \mathcal{N}(0, 1)\right] \quad (3)$$

The KL loss is equivalent to the sum of all KL divergences between the component $X \sim N(\mu_x, \sigma_x^2)$ and the standard normal, and is minimized when $\mu_x = 0, \sigma_x = 1$.

Informative Synthetic Image Generation Given image I, we pass it through the encoder to obtain latent distribution parameters μ_I, σ_I. By sampling from this distribution, we generate different transformations of the original image I, which we denote as I^n. We first identify informative samplesfrom a given pool using the MIS score (Eqn. 2). We aim that I^n is similar to I and at the same time has novel information such that by adding to the training set we add qualitatively novel data. Hence, I^n should fulfill the following novelty criteria:

1. *Class label preservation*: Image I and all generated I^n should have the same class labels.
2. *Redundancy avoidance*: I^n's semantic content should be sufficiently different from I to ensure that I^n are identified as informative.

We propose a novel scoring function that quantifies the degree to which the generated images meet the above criteria. The final score for each generated image is used to rank the images and select the most informative ones.

Preserving Label Similarity: To preserve the labels between I and I^n we enforce that class probability values of I and I^n be close. Let us denote the probability of class k for image I as p_I^k, and the corresponding probability for

generated image I^n is $p^k_{I^n}$. Class probabilities for each label is determined using the current classification model If $p^k_I > 0.5 (< 0.5)$ we expect $p^k_{I^n} > 0.5 (< 0.5)$. The change in probability values should not be too high to avoid introducing spurious information. Consequently, we define the label score as a function of the relative difference between class probabilities computed for I and I^n, as follows:

$$\text{Score}_{label} = \begin{cases} \frac{|p^k_{I^n} - p^k_I|}{p^k_I}, & \text{if } \frac{|p^k_{I^n} - p^k_I|}{p^k_I} \leq \eta_1 \\ -\gamma_1, & \text{otherwise} \end{cases} \quad (4)$$

We set $\eta_1 = 0.3$ and $\gamma_1 = 0.1$, and we put a condition that $p^k_I - 0.5$ and $p^k_{I^n} - 0.5$ always have the same sign. If the signs are different, then the labels are different and the generated sample is not considered for informativeness evaluation. As per the above formulation if the difference of probability values is less than equal to 30%, the score is same as this value. If the difference is greater than 30% it indicates a significant change of the probability distributions. Such a high change may bias the probability to too high or low values which indicate that the generated image is not very informative since the classifier is very confident in predicting its label. Such an approach ensures that generated images with high distorted content are not included in the training set.

Redundancy avoidance: To ensure generated images have new information compared to the original image, we use the difference in MIS values (Eqn.2). We put a condition that the relative difference between their scores is above a threshold and define the redundancy avoidance score as:

$$\text{Score}_{red} = \begin{cases} \frac{|MIS^I - MIS^{I^n}|}{MIS^I}, & \eta_2 \leq \frac{|MIS^I - MIS^{I^n}|}{MIS^I} \leq \eta_3 \\ -\gamma_2, & otherwise \end{cases} \quad (5)$$

We set $\eta_2 = 0.05, \eta_3 = 0.25$, and $\gamma_2 = 0.15$ to define the range of thresholds within which the MIS score of the new sample can vary. The total informativeness of the generated sample is then calculated as,

$$\text{Score}_{sample} = \lambda_1 \text{Score}_{label} + \lambda_2 \text{Score}_{red}. \quad (6)$$

Higher Score_{sample} indicates higher overall informativeness. We rank the generated samples based on Score_{sample} and select the top-n informative samples to add to the training set. λ_1, λ_2 determine relative contribution of each term.

Baseline Methods For Comparison: We compare our method's performance with the following baselines: 1) **Fully supervised Learning (FSL):** using a DenseNet-121 classifier (Huang et al. 2016) on the designated training sets. 2) **Random Sample Selection** without informativeness criteria. 3) **Uncertainty** based informative sample selection.

Implementation Details: Our method is implemented in TensorFlow. Initially we choose 5% training samples and train a DenseNet-121 (Huang et al. 2016) on the NIH ChestXray14 dataset (Wang et al. 2017b). We select a batch of 128 images from the unlabeled dataset. For each image we set up the graph and calculate the MIS (multi-label informativeness) score. Top 3 samples of each class

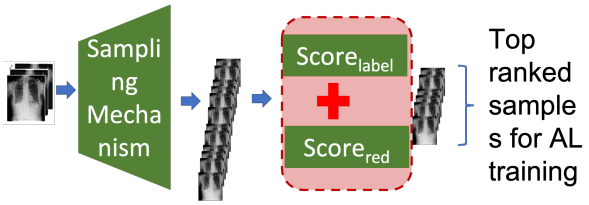

Figure 4: Augmenting and choosing informative samples from a base informative sample. Given base informative images we generate additional images by sampling from the variational autoencoder. To add only informative generated images to the training set, we calculate sum of label score Score_{label} and redundancy avoidance score Score_{red}.

are the base informative samples. We generate more informative samples, rank them and add the top samples alongwith the base image to the training set. The classifier is updated, and the cycle continues till there are no more informative samples.We used Adam (Kingma and Ba 2014) with $\beta_1 = 0.93$, $\beta_2 = 0.999$, batch normalization, binary cross-entropy loss, learning rate $1e-4$, 10^5 update iterations and early stopping based on the validation accuracy. The architecture and trained parameters were kept constant across compared approaches. Training and testing were performed on an NVIDIA Titan X GPU having 12 GB RAM. The input image size is 320×320 pixels. To obtain interpretability saliency maps, we used default parameters of the iNNvestigate implementation of Deep Taylor (Alber et al. 2019), and GradCAM (Selvaraju et al. 2017). For uncertainty estimation, we used a total of $T = 20$ dropout samples with dropout distributed across all layers (Kendall and Gal 2017).

Comparison Methods: Our proposed method 'DAMLAL' is compared with other AL methods: 1) 'GAL'- Graph-Based Active Learning (GAL) approach of (Long et al. 2008); 2) 'LEMAL': the "example-label based sampling strategy (LEMAL)" approach by (Wu et al. 2014); 3) 'CVIRS'- the Uncertainty sampling based on "Category Vector Inconsistency and Ranking of Scores (CVIRS)" approach of (Reyes, Morell, and Ventura 2018); 4) 'AlphaMix' - the "Active Learning by Feature Mixing (Alpha-Mix)" method of (Parvaneh et al. 2022). We also compare with 5) LADA (Kim et al. 2021), and 6) A graph-based multi-label active learning method called 'GESTALT' (Mahapatra, Poellinger, and Reyes 2022a).

Results and Discussion

Dataset Description: We use the CheXpert dataset (Irvin, Rajpurkar, and et al. 2019) consisting of $224,316$ chest radiographs of $65,240$ patients labeled for the presence of common chest conditions. The training set has $223,414$ images, while validation and test sets have 200 and 500 images. The validation ground truth was obtained using majority voting from annotations of 3 board-certified radiologists. The test set evaluation protocol is based on 5 disease labels: *Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, and *Pleural Effusion*.We also use the NIH ChestXray14 dataset (Wang et al. 2017b) having $112,120$ expert-annotated frontal-view

X-rays from $30,805$ unique patients. For each task, the dataset was split into training (70%), validation (10%), and test (20%) at the patient level such that all images from one patient are in a single fold.

Comparative Results for CheXpert Dataset: Table 1 shows the Area Under the Curve (AUC) for every 10% increment of training data using different methods. Except for random sample selection, all AL-based methods outperform the fully-supervised learning model (FSL) with fewer samples, confirming the benefits of selecting samples based on their informativeness. This finding aligns with other works (Mayer and Timofte 2018; Yang et al. 2017; Sourati et al. 2019) indicating the capability of AL methods to boost the learning rate of trained models. GESTALT does better than most methods while LADA is slightly worse than GESTALT. We show results for two versions of DAMLAL - DAMLAL$_{DT}$, when using saliency maps obtained with the Deep Taylor method, and DAMLAL$_{GC}$, when using Grad-CAM saliency maps. Of the two, DAMLAL$_{DT}$ shows better results and we refer to it as DAMLAL. DAMLAL significantly improves over GESTALT and LADA by integrating data augmentation with multi-label AL. GESTALT outperforms FSL at 45% of labeled data, whereas DAMLAL outperforms FSL at 37% of training data. DAMLAL requires significantly less labeled data to attain better performance. Although LADA combines data augmentation with active learning, it does not consider the multi-label scenario.

Ablation Studies: We attribute the improved performance of DAMLAL to: 1) integration of data augmentation with multi-label active learning; 2) use of graph attention transformers that improve the representation learning. We conduct ablation studies on DAMLAL$_{DT}$ to quantify their individual contributions and summarize results in Table 1.

First we exclude data augmentation from the pipeline and use only the initially identified informative samples and use very basic augmentation strategies like rotation, translation, and scaling instead of our proposed informative augmentation. We refer to this method as DAMLAL$_{con-DA}$ (i.e., DAMLAL$_{DT}$ using conventional data augmentation). DAMLAL$_{con-DA}$ does better than GESTALT and shows the merit of using GATs and data augmentation for multi-label sample selection. A second variant of DAMLAL$_{DT}$ uses only the sum of graph weights which we term as DAMLAL$_{Sum}$. It does better than GESTALT due to added informative augmentation but fares worse than DAMLAL. This is due to graph transformer networks doing a better job of learning the global dependencies, and being more accurate in identifying informative samples.

Importance of Graph Multi Set pooling: We replace the multi-set pooling with conventional pooling and report the results in Table 1 as DAMLAL$_{pooling}$. We observe a significant drop in performance compared to DAMLAL$_{DT}$ showing the multiset pooling has a significant role in the improved performance. We also completely remove the graph attention, and self attention, and show the results in Table 1 as DAMLAL$_{no-GAT}$, and DAMLAL$_{no-SA}$. Individual results clearly show the importance of each component of the graph pooling stage.

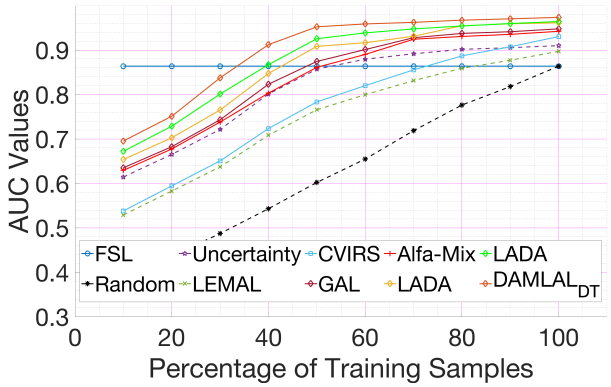

Figure 5: Additional Dataset - NIH DataSet: AUC measures at different percentage levels of training percentage for baselines (indicated with dotted lines) and the proposed DAMLAL approach.

We also investigate the role of the parameters k and n. n is the number of input nodes which corresponds to the number of labels for the input image. In our experiments we use $n = 5$ since the test set of CheXpert dataset evaluates on 5 diseased labels. Decreasing n leads to reduced performance since fewer nodes result in inferior learning of features. Increasing n improves performance since more nodes (disease labels) improve the feature learning ability of the network. k denotes the dimensions of the Graph Multiset Pooling stage. In our experiments we set $k = 3$ (k_{int}) for the intermediate layers while $k = 1$ for the final layer (k_{fin}).

Performance on Additional Datasets

In Figure 5 we show result obtained for the NIH dataset across different methods. The trend is similar to the results shown in Table 1 wherein different active learning methods outperform the fully supervised learning method at a lower training data percentage, and the proposed DAMLAL approach outperform other methods.

Robustness and Generalization

To test the robustness of the proposed approach, we added simulated Gaussian noise of $\mu = 0$ and different $\sigma \in \{0.005, 0.01, 0.015, 0.05, 0.1\}$. Figure 6 shows the AUC values for the baseline performance of DAMLAL and different σ. The results are close to DAMLAL for $\sigma = 0.005, 0.01$, but start to degrade significantly for noise levels above $\sigma = 0.01$, which we term as noise threshold. With added noise, the performance of all methods degrades. However, our proposed DAMLAL performs better than others and is more robust.

Hyperparameter Settings

We adopt the following steps to set the value of η_1, η_2, η_3. For η_1 we varied the values from $[0, 1]$ in steps of 0.05, keeping $\eta_2 = 0.05, \eta_3 = 0.35$. The best results were obtained for $\eta_1 = 0.3$, which was our final value. Following similar

Train Data %	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	p-
FSL - reference	90.23	90.23	90.23	90.23	90.23	90.23	90.23	90.23	90.23	90.23	< 0.001
Random	41.69	47.5	52.6	58.1	64.07	69.34	75.72	81.49	85.64	90.23	< 0.001
Uncertainty	62.15	66.56	72.41	80.16	85.80	88.12	90.34	90.72	90.84	91.03	< 0.001
LEMAL(Wu et al. 2014)	64.70	69.17	76.08	81.25	88.32	89.03	91.09	91.40	91.74	92.08	< 0.001
CVIRS(Reyes, Morell, and Ventura 2018)	66.05	70.88	76.05	84.12	89.09	91.02	91.48	91.52	92.03	92.54	0.001
AlphaMix(Parvaneh et al. 2022)	68.30	72.57	79.11	87.22	91.43	93.12	93.47	93.87	94.09	95.01	0.005
GAL(Long et al. 2008)	68.91	73.01	79.61	87.99	91.74	92.82	93.30	93.91	94.24	94.56	0.007
LADA(Kim et al. 2021)	69.15	73.98	80.82.	89.08	92.93	93.64	94.04	94.72	95.13	95.73	0.001
GESTALT	70.82	74.64	80.82	89.27	93.28	94.13	94.92	95.24	95.88	96.71	0.02
DAMLAL$_{DT}$	**72.24**	**76.18**	**82.47**	**91.02**	**94.90**	**95.88**	**96.42**	**96.92**	**97.43**	**98.21**	-
DAMLAL$_{GC}$	71.63	75.84	81.16	90.35	94.53	95.11	95.79	96.07	96.7	97.57	0.07
Ablation Studies											
DAMLAL$_{con-DA}$	71.23	75.32	81.83	89.93	94.08	94.45	95.62	95.89	96.24	97.01	0.034
DAMLAL$_{Sum}$	71.01	75.1	81.43	89.51	93.71	94.21	95.31	95.61	96.01	96.81	0.032
DAMLAL$_{red}$	68.22	71.83	78.23	84.68	87.12	90.34	91.97	92.47	93.47	94.01	0.03
DAMLAL$_{label}$	65.61	68.25	74.37	79.51	82.02	85.48	87.15	88.02	89.42	90.14	0.03
DAMLAL$_{pooling}$	64.13	67.24	71.34	76.23	79.21	83.02	86.25	87.12	87.64	88.52	0.032
DAMLAL$_{no-GAT}$	62.21	65.52	69.71	73.95	77.12	80.94	83.84	85.36	86.02	86.78	0.023
DAMLAL$_{no-SA}$	61.35	64.93	68.62	73.02	76.21	79.41	82.29	84.62	85.11	85.77	0.011

Table 1: AUC values on CheXpert dataset for different baselines, proposed DAMLAL approach, and ablation studies. The $p-$values are with respect to DAMLAL. DT: DeepTaylor; GC: GradCAM; con-DA: Conventional Data Augmentation. FSL: Fully supervised model trained with all 100% data shown as reference. Bold indicates best performance.

Training And Test Time								
DenseNet-121	Random	Unc	GESTALT	GAL	LEMAL	CVIRS	AlfaMix	DAMLAL
18h(0.67T)	18.5h(0.69T)	19.5h(0.72T)	25h(0.93T)	23.5h(0.87T)	20h(0.74T)	21.5h(0.8T)	24h(0.89T)	27h(T)
0.18s	0.19s	0.2s	0.32s	0.28s	0.22s	0.24s	0.3s	0.4s

Table 2: Training(h)/Inference(s) time in hours and seconds for different methods.

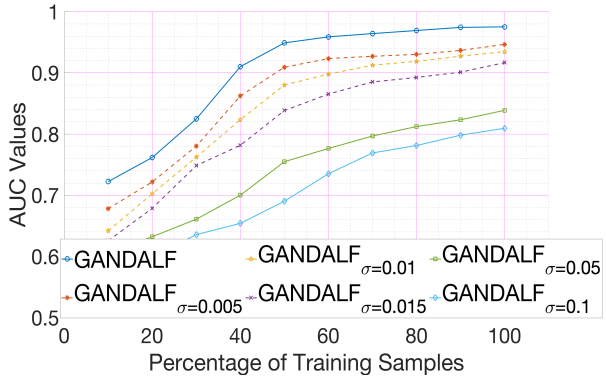

Figure 6: AUC measures for different features for added Gaussian noise of $\mu = 0$ and different σ.

steps, we set $\eta_2 = 0.1$, while keeping $\eta_1 = 0.3, \eta_3 = 0.35$. Thereafter we fix $\eta_1 = 0.3, \eta_2 = 0.1$, and vary the values for η_3 and get the best results for $\eta_3 = 0.3$. While obtaining the optimal values of η_1, η_2, η_3 we keep fixed $\gamma_1 = 0.2$ and $\gamma_2 = 0.2$. After setting the values of η_1, η_2, η_3 we vary γ_1, γ_2 and obtain the best values for $\gamma_1 = 0.1$ and $\gamma_2 = 0.1$. The sensitivity of the different parameters is shown in Table 3.

The threshold η_1 is used to ensure that the probability values of the generated image do not change significantly as to make it uninformative. For example, if $p^k_{I^n}$ is close to 0.9 then the generated image is not very informative as the classifier is very confident about the prediction. In such a case the score function $Score_{label}$ is assigned a negative value of 0.1 (γ_1). This ensures that this particular sample's informativeness is reduced and is given less importance in selecting informative synthesized samples. We observe that too high negative values for γ_1 will give disproportionate importance to the probability score and will unfairly reduce the score of the sample despite a high value for $Score_{red}$.

The threshold parameters η_2, η_3 control the degree of redundancy that may be allowed for the generated images. We want that the generated images should have a minimum degree of novelty which is controlled by $\eta_2 = 0.05$. Quantitatively, this may be interpreted that the multilabel informativeness score changes by atleast 5%. On the other hand too much change of the informativeness score indicates major distortions to the image and could also be due to an 'outlier' image. Hence the upper threshold $\eta_3 = 0.25$ indicates that we allow upto 25% change of the image's informativeness score. This ensures that the transformed images are not too different from the base image. The penalty γ_2's optimal value is 0.1 since too high values give disproportionate importance to the redundancy score.

Vals→	0.0	0.1	0.2	0.3	0.4	0.5
η_1	93.6	94.1	94.7	**95.3**	94.8	94.0
η_2	95.5	**96.4**	95.9	95.2	94.8	94.3
η_3	94.3	95.2	96.1	**97.5**	96.7	96.0
γ_1	96.7	**98.1**	96.8	96.0	95.6	94.9
γ_2	96.8	**98.2**	96.9	96.2	95.7	94.8
Vals→		0.6	0.7	0.8	0.9	1.0
η_1		93.5	92.8	92.1	91.7	91.2
η_2		93.5	92.9	92.3	91.7	91.1
η_3		95.4	94.8	94.1	93.6	92.9
γ_1		94.2	92.7	92.1	91.8	91.2
γ_2		94.1	92.5	91.8	91.3	90.6

Table 3: AUC values for DAMLAL for different values of the parameters η_1, η_2, η_3, γ_1, γ_2.

Realism of Synthetic Features

Data augmentation is an important part of our pipeline, where informative synthetic samples are added to the training set. Hence it is imperative that the generated images be realistic else it will adversely affect the performance of the trained classifier. We perform a qualitative assessment of our synthetic images to determine their degree of realism. We select 1000 synthetic images almost equally distributed between the 14 classes and ask two trained radiologists, having 12 and 14 years experience in examining chest xray images for abnormalities, to identify whether the images are realistic or not. Each radiologist was blinded to the other's answers.

Results in Table 4 show one radiologist (RAD 1) identified $912/1000$ (91.2%) images as realistic while RAD 2 identified 919 (91.9%) generated images as realistic. Both of them had a high agreement with 890 common images (89.0% -"$Both\ Experts$" in Table 4) identified as realistic. Considering both RAD 1 and RAD 2 feedback, a total of 941 (94.1%) unique images were identified as realistic ("$Atleast\ 1\ Expert$"). Subsequently, $59/1000$ (5.9%) of the images were not identified as realistic by any of the experts ("$No\ Expert$").

We also generate images using GANs and repeat the qualitative assessment with the two radiologists. The agreement statistics are summarized in Table 4. GANs show a higher degree of agreement since it is a well established fact that GANs generate more realistic images than VAEs, but are more time consuming and difficult to train. However the difference in the percentage of identified realistic images is not very high. We believe this is to be an important assessment to ensure there are no abnormal artefacts introduced in the entire active learning setup.

Importance of Score Values

We investigate the importance of each of the scoring terms in Eqn. 6. Table 1 show the performance measures when using only Score_{red} (DAMLAL_{red}) and only Score_{label} (DAMLAL_{label}. The results clearly show that discarding either of the terms degrades the performance. Excluding Score_{red} leads to worse performance than excluding Score_{label}. This may be explained by the fact that the redundancy score uses the multilabel informativeness score MIS to determine informative samples.

	RAD 1	RAD 1	Both Experts	Atleast 1 Expert	No Expert
VAE	912	919	**89.0** (890)	**94.1** (941)	**5.9** (59)
GANs	927	931	**90.5** (905)	**95.8** (958)	**4.2** (42)

Table 4: Agreement statistics for different image generation methods amongst 2 radiologists. Numbers in bold indicate agreement percentage while numbers within brackets indicate actual numbers out of 1000 samples.

Computation Time

For a training dataset of $100,000$ images of size 320×320, the training time (in hours) for different methods on an NVIDIA Titan X GPU having 12 GB RAM is summarized in Table 2. Compared to GESTALT, our proposed DAMLAL method has an 8% higher training time. This is due to the extra computations involved in the informative augmentation and graph transformer attention which is an integral part of the process. However, the resulting performance improvement justifies the added complexity of our method. The inference time for a single image (in seconds) is also summarized for different methods.

Conclusions

In this paper, we present a novel approach that combines multi-label active learning with data augmentation and the key motivation is to leverage their mutually complementary strengths. By using graph attention transformers with graph neural networks we learn more discriminative graph aggregations. Complementing the improved graph aggregation strategy is the informative augmentation step that takes a base informative image, generates augmented versions, and calculates a score based on label preservation and informativeness of the augmented images. The overall informativeness of the augmented samples is the sum of the two scores, and the most informative samples are added to the training set for further training. Our proposed method yields better results than competing methods and ablation studies highlight the importance of the graph attention transformers and the informative augmentation step in the overall performance of DAMLAL. We also engage two experienced radiologists to perform qualitative assessment of images generated by our method and GANs, which confirms the high degree of realism of our synthetic images.

References

Alber, M.; Lapuschkin, S.; Seegerer, P.; Hagele, M.; Schutt, K. T.; Montavon, G.; Samek, W.; Muller, K.-R.; Dahne, S.; and Kindermans, P.-J. 2019. iNNvestigate neural networks. *Journal of Machine Learning Research*, 20(93): 1–8.

Baek, J.; Kang, M.; and Hwang, S. J. 2021. Accurate Learning of Graph Representations with Graph Multiset Pooling. In *International Conference on Learning Representations*.

Bianchi, F. M.; Grattarola, D.; and Alippi, C. 2019. Spectral clustering with graph neural networks for graph pooling. In *arXiv preprint arXiv:1907.00481*.

Bozorgtabar, B.; Mahapatra, D.; von Teng, H.; Pollinger, A.; Ebner, L.; Thiran, J.-P.; and Reyes, M. 2019. Informative sample generation using class aware generative adversarial networks for classification of chest Xrays. *Computer Vision and Image Understanding*, 184: 57–65.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep Bayesian Active Learning with Image Data. In *Proc. International Conference on Machine Learning*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Proc. NIPS*, 2672–2680.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. 2016. Densely Connected Convolutional Networks. In *https://arxiv.org/abs/1608.06993,*.

Irvin, J.; Rajpurkar, P.; and et al. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *arXiv preprint arXiv:1901.07031*.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial Transformer Networks. In *NIPS*, –.

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems.*

Kim, Y.-Y.; Song, K.; Jang, J.; and Moon, I.-c. 2021. LADA: Look-Ahead Data Acquisition via Augmentation for Deep Active Learning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 22919–22930. Curran Associates, Inc.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. In *arXiv preprint arXiv:1412.6980,*.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. In *arXiv preprint arXiv:1312.6114*.

Lee, J.; Lee, I.; and Kang, J. 2019. Self-attention graph pooling. In *ICML*, 3734–3743.

Lee, J.; Lee, Y.; Kim, J.; Kosiorek, A. R.; Choi, S.; and Teh., Y. W. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 3744–3753.

Long, J.; Yin, J.; Zhao, W.; and Zhu, E. 2008. Graph-Based Active Learning Based on Label Propagation. In Torra, V.; and Narukawa, Y., eds., *Modeling Decisions for Artificial Intelligence*, 179–190. Berlin, Heidelberg: Springer Berlin Heidelberg.

Mahapatra, D.; Poellinger, A.; and Reyes, M. 2022a. Graph Node Based Interpretability Guided Sample Selection for Active Learning. *IEEE Transactions on Medical Imaging*, 1–1.

Mahapatra, D.; Poellinger, A.; and Reyes, M. 2022b. Interpretability-guided inductive bias for deep learning based medical image. *Medical Image Analysis*, 81: 102551.

Mahapatra, D.; Poellinger, A.; Shao, L.; and Reyes, M. 2021. Interpretability-Driven Sample Selection Using Self Supervised Learning For Disease Classification And Segmentation. *IEEE TMI*, 40(10): 2548–2562.

Mayer, C.; and Timofte, R. 2018. Adversarial sampling for active learning. In *arXiv preprint arXiv:1808.06671*.

Mollenbrok, L.; Sumbul, G.; and Demir, B. 2023. Deep Active Learning for Multi-Label Classification of Remote Sensing Images. arXiv:2212.01165.

Parvaneh, A.; Abbasnejad, E.; Teney, D.; Haffari, G. R.; van den Hengel, A.; and Shi, J. Q. 2022. Active Learning by Feature Mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12237–12246.

Perez, L.; and Wang, J. 2017. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. In *arXiv preprint arXiv:1712.04621*.

Reyes, O.; Morell, C.; and Ventura, S. 2018. Effective active learning strategy for multi-label learning. *Neurocomputing*, 273: 494–508.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proc. ICCV*, 618–626.

Sourati, J.; Gholipour, A.; Dy, J. G.; Tomas-Fernandez, X.; Kurugol, S.; and Warfield, S. K. 2019. Intelligent labeling based on fisher information for medical image segmentation using deep learning. *IEEE transactions on medical imaging*, 38(11): 2642–2653.

Tran, T.; Do, T.-T.; Reid, I.; and Carneiro, G. 2019. Bayesian Generative Active Deep Learning. In *arXiv preprint arXiv:1904.11643*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; and Lin., L. 2017a. Cost-Effective Active Learning for Deep Image Classification. *IEEE Trans. CSVT.*, 27(12): 2591–2600.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. 2017b. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *In Proc. CVPR*.

Wu, J.; Sheng, V. S.; Zhang, J.; Li, H.; Dadakova, T.; Swisher, C. L.; Cui, Z.; and Zhao, P. 2020. Multi-Label Active Learning Algorithms for Image Classification: Overview and Future Promise. *ACM Comput. Surv.*, 53(2).

Wu, J.; Sheng, V. S.; Zhang, J.; Zhao, P.; and Cui, Z. 2014. Multi-label active learning for image classification. In *2014 IEEE International Conference on Image Processing (ICIP)*, 5227–5231.

Yang, L.; Zhang, Y.; Chen, J.; Zhang, S.; and Chen, D. 2017. Suggestive Annotation: A Deep Active Learning Framework

for Biomedical Image Segmentation. In *Proc. MICCAI*, 399–407.

Ying, Z.; You, J.; Morris, C.; Ren, X.; Hamilton, W. L.; and Leskovec, J. 2018. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, 4805–4815.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond Empirical Risk Minimization. In *arXiv preprint arXiv:1710.09412*.

Zhang, M.; Cui, Z.; Neumann, M.; and Chen, Y. 2018. An end-to-end deep learning architecture for graph classification. In *AAAI*, 4438–4445.

Zheng, H.; Yang, L.; Chen, J.; Han, J.; Zhang, Y.; Liang, P.; Zhao, Z.; Wang, C.; and Chen, D. Z. 2019. Biomedical image segmentation via representative annotation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5901–5908.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proc. CVPR*, 2921–2929.

Zhu, J.-J.; and Bento, J. 2017. Generative adversarial active learning. *arXiv preprint arXiv:1702.07956*.