DAUG: DIFFUSION-BASED CHANNEL AUGMENTATION FOR RADIOLOGY IMAGE RETRIEVAL AND CLASSIFICATION

Ying Jin[†‡] *Zhuoran Zhou*[‡] *Jenq-Neng Hwang*[‡]

[†]Microsoft
[‡]University of Washington

ABSTRACT

Medical image understanding requires meticulous examination of fine visual details, with particular regions requiring additional attention. While radiologists build such expertise over years of experience, it is challenging for AI models to learn where to look with limited amounts of training data. This limitation results in unsatisfying accuracy in medical image understanding. To address this issue, we propose DAug, a portable method that improves model performance by augmenting the image feature. Specifically, we extend the single-channel radiology image to multiple channels, with the additional channels being the heatmaps of abnormalities which demand extra attention. Although ground truths of such abnormality heatmaps are unavailable, we generate them with a diffusion-based image-to-image translation model, trained with disease class labels. Our method is motivated by the fact that generative models learn the distribution of normal and abnormal images, and such knowledge is complementary to perception tasks. In addition, we propose a novel Image-Text-Class Hybrid Contrastive learning criterion to utilize both text and class labels. With two novel approaches combined, our method surpasses baseline models without changing the model architecture, and achieves state-of-the-art performance on both medical image retrieval and classification tasks.

Index Terms— Chest X-ray, Diffusion Models, Image Retrieval, Image Classification

1. INTRODUCTION

Shortages and burnout of radiologists are significant problems worldwide and leave risks to patient care [1, 2]. Training a radiologist takes thirteen to fifteen years, making AI models assisting in diagnostics a scalable solution. Specifically, the classification and retrieval of Chest X-ray images are fundamental tasks, as classification can cross-check with doctors, and retrieval allows comparison with historical cases for more accurate diagnoses.

Vision models pretrained on large datasets provide a strong baseline for understanding common, scenic images. However, when applied to medical images, their pretrained capabilities are underexplored for two reasons. First, medical images require meticulous examination of fine details, which differs from scenic images where salient objects are large and have clear boundaries. Second, radiology images like Chest X-rays are monochrome, preventing full utilization of the pretrained model's capability to utilize all three color channels. Consequently, vision models trained on limited radiology images struggle to focus on the correct regions for accurate understanding. In this study, we tackle this issue by augmenting the image feature with abnormality heatmaps as additional channels alongside the original monochrome medical image. The augmented feature provides additional information to the model on areas requiring extra attention, mimicking the expertise of experienced radiologists. Our method, dubbed DAug, is portable onto a wide range of model architectures and leverages their native compatibility with multi-channel (RGB) images.

Specifically, we repurpose a diffusion-based image generation model to generate attention heatmaps for predefined disease classes. Each heatmap emphasizes the region where the disease could potentially occur. The heatmap, as an additional feature channel, explicitly directs the model's attention to clinical significant areas, a skill difficult to acquire through conventional training. The augmentation module is implemented with a classifier-guided diffusion model [3]. We add Gaussian noise into a Chest X-ray image, and the diffusion model, guided by the disease classifier, removes the noise toward a direction where a disease is either mitigated or exacerbated. The difference between the input and output images yields an attention heatmap that highlights the potential disease area. Through this generative learning process, the model acquires knowledge beyond what is typically learned by classification or retrieval models, and enhance the performance of these tasks. Utilizing generative learning as the mechanism, our method produces heatmaps in a self-supervised manner, eliminating the need for human-annotated heatmaps. Our diffusion-based heatmap generation is inspired by [4], and we extensively improved the method to handle the co-occurrence of multiple diseases and to generate cleaner and more accurate outputs.

In addition, different from existing work which trains retrieval and classification models separately, we are motivated by the synergy of learning two tasks together. Therefore, we design an Image-Text-Class Hybrid Contrastive learning criterion: during training, the loss function contrasts both image-text and image-class pairs to leverage both text and class labels. It results in a single model that supports both classification and retrieval tasks, making real-world deployment at ease. To validate the performance gain, we combine both methods and evaluate on the largest Chest X-ray dataset, MIMIC-CXR [5]. Our model outperforms existing state of the arts on both retrieval and classification tasks.

To sum up, the contribution of this paper is three-fold:

- We propose DAug, a portable feature augmentation method which improves medical image understanding performance by adding abnormality heatmaps as additional image channels.
- We introduce Image-Text-Class Hybrid Contrastive learning which leverage both image-text and image-class labels to improve performance on both retrieval and classification tasks.
- We deliver a single model which is capable of both retrieval and classification tasks with state-of-the-art performance. The proposed methods can be applied on standard pre-trained models (such as CLIP ViT [6]), making real-world deployment as ease.

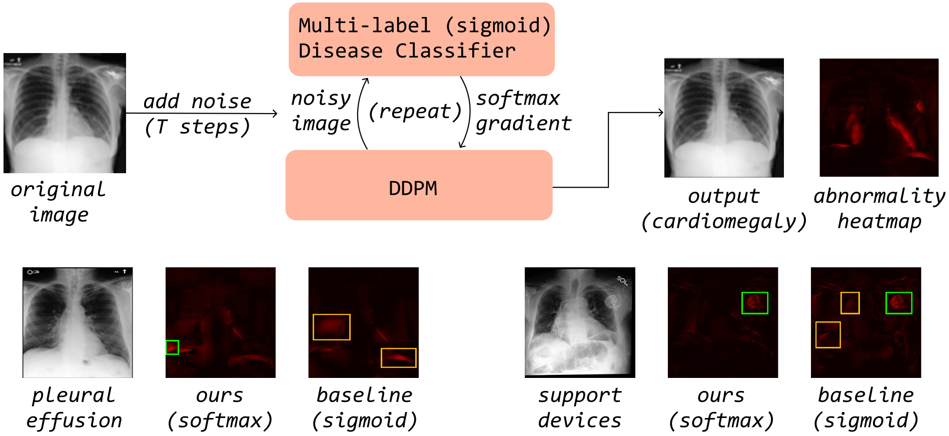

Fig. 1: Diffusion-based channel augmentation (DAug) pipeline. The original image is translated into diseased versions (or vise versa) with a classifier-guided diffusion model. The difference between the input and output images produce a heatmap highlighting the potential area of the corresponding disease. The heatmaps are added to the original monochrome radiology image as additional image channels, resulting in augmented features which improve the performance on downstream tasks. The DAug features supports multiple disease categories (two examples in the second row), and our softmax-based approach generates more accurate heatmaps than the existing baseline. Green and orange bounding boxes indicate correctly and wrongly highlighted regions, respectively.

2. RELATED WORK

2.1. Generation for Perception

Perception and generation models have long been regarded as two distinct paradigms in machine learning. In computer vision, perception tasks such as classification and segmentation require training data of {image, ground truth} pairs labeled in a defined format, often leading to a limitation in the amount of training data. Generative models, on the other hand, reconstruct the original image during training without requiring additional annotation, enabling the use of larger amounts of training data. Image generation models learn the distribution of image features, and this knowledge could aid in the understanding of the images as well. This insight has led us to explore the use of a diffusion-based generation model in aiding medical image understanding.

Common image generation methods include Generative Adversarial Networks (GANs) [7], Variational Autoencoders (VAEs) [8] and Diffusion Models [9]. Diffusion models became the main stream due to the ease of training and high image quality. During training, the diffusion model learns to remove noise from a noisy input and gradually turns a Gaussian noise into images. The de-noising steps can be guided by a classifier, whose gradients are used to determine the direction of the denoising process, encouraging the output to maximize the probability of a certain class based on the classifier. The result will be an image of the chosen class.

Existing works in image understanding have been benefiting from generative modeling. One stream of work explores using generative models to augment or synthesize training data [10, 11, 12]. Another stream of work uses generative modeling as a pre-training task. For example, Masked AutoEncoders (MAE) [13] reconstructs the original image from a partial input during pretraining, and finetunes on downstream classification tasks. In terms of medical images, [4] proposed medical anomaly detection with classifier-guided diffusion. When applied on Chest X-rays, their work produces reasonable anomaly heatmaps for easy cases like pleural effusion, but cannot handle the co-occurrence of multiple diseases reliably. Our work is based on [4], and we make significant algorithmic improves for turning the output a useful feature which aids downstream tasks. Our improvements are detailed in section 3.2.

2.2. Image Retrieval and Classification

In the era of transformer models, pretraining on web-scale data boosts the performance on downstream tasks like retrieval and classification. CLIP [6] proposes image-text contrastive learning, which aligns the feature spaces of an image encoder and a text encoder. Retrieval methods base on the cosine similarity of CLIP features has become dominant since then. Due to the significant domain gap between medical images and the common scenic images, CLIP shows limited zero-shot performance on the medical domain. A stream of work [14, 15] aims to fine-tune CLIP on medical image-text data.

In this study, we address the challenges of medical image retrieval and classification from two perspectives. Specifically, DAug is from the feature augmentation perspective and the Image-Text-Class Hybrid Contrastive learning is from the loss function perspective. On the feature augmentation perspective, maintaining feature banks for prototypes built on the whole dataset is often effective for tasks with limited training data, and this applies to medical images [16, 17]. Recently, X-TRA [18] proposes improving radiology multi-modal retrieval and classification by a retrieval-based feature augmentation. They use a CLIP model to select the top-K similar samples from the dataset and construct an augmented feature for each input sample. Differently, DAug, our method, leverage generative models and outperforms existing methods.

Related to our work, UniCL [19] combines image-text and image-class datasets for contrastive pre-training. However, each training sample is assumed to have either text label or class label, but not both. The scenario is different from the medical domain, where datasets such as MIMIC-CXR and CheXpert [5, 20, 21] often contain both text (medical report) and class (multiple disease classes) labels at the same time. Our Image-Text-Class Hybrid Contrastive loss function utilizes this feature by combining both image-text contrastive loss and image-class contrastive loss into a single loss term per sample. Intuitively, the method treats image classification as a special image-to-class retrieval task. It not only enables the synergy of learning

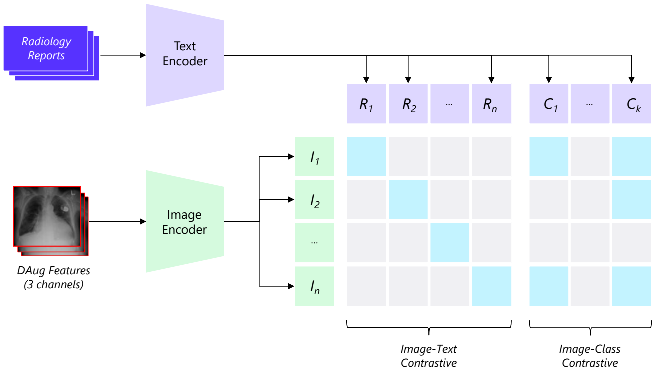

Fig. 2: Model architecture and Image-Text-Class Hybrid Contrastive Loss. The inputs are pairs of radiology reports and DAug features (3-channel images including both medical image and abnormality heatmap channels). The image and text encoders are pretrained by CLIP. R and C are text embeddings for the reports and class prompts, respectively. The hybrid contrastive loss includes both the image-text CLIP loss and image-class contrastive loss. Blue cells are positive pairs, in which the image-class matchings are derived from ground-truth class labels.

both retrieval and classification tasks together, but also delivers a single model that can perform both two tasks. Therefore, both model performance and deployment ease are improved.

3. METHOD

3.1. Overview

We propose a unified model for multi-modal retrieval and image classification in the radiology domain. This section is divided into three parts. Section 3.2 focuses on feature augmentation, detailing the use of a diffusion-based generative model to create heatmaps and the methods for integrating these additional features. Section 3.3 describes the model architecture and training process, with a special emphasis on unified training for both retrieval and classification tasks. Lastly, section 3.4 discusses the testing pipeline for the retrieval and classification tasks.

3.2. DAug: Diffusion-based Channel Augmentation

We aim to use generative models to enhance medical image perception models by providing additional information. Building on existing work [4], validated on pleural effusion—a relatively straightforward disease—we extend its application to multiple diseases and optimize its suitability for our scenario.

3.2.1. Dataset and Label Generation

For experiments on both retrieval and classification tasks, we use MIMIC-CXR [5], the largest Chest X-ray image-text dataset. As class labels are unavailable, we generate pseudo-labels by converting the text to 14 disease classes with CheXpert labeler [22], a text classification model.

3.2.2. Training of Classifier-Guided Diffusion Model

We train a Denoising Diffusion Probabilistic Model (DDPM) model [23] on Chest X-ray images. Given an input image I, we generate a series of noisy images $\{x_0, x_1, \ldots, x_T\}$ by gradually adding Gaussian noise for T steps. x_0 is the original image with no noise and $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A U-Net model is trained to reverse the process by estimating the noise added in any timesteps. During evaluation, the model can recover the original image x_0 by removing the estimated noise for T steps from Gaussian noise X_T. We set $T = 1,000$.

To translate a radiology image into its healthy or diseased variant, the generated image must retain the same anatomic structure as the input image. For this purpose, following [4], we add noise by only 500 steps to the original input, resulting in x_{500} where the anatomic structures are still visible. The diffusion model is used to recover the original input from x_{500}. To produce a diseased or healthy version of the input, the denoising process is guided by an image classifier that assigns an X-ray to defined disease classes. The gradients for a selected class are used to condition the denoising process, making the output image maximize or minimize the probability of the class in the classifier.

For Chest X-ray images, each image may contain multiple diseases, making the binary classifier from the original work unsuitable. From our evaluation, particularly challenging cases involve diseases that often co-occur (e.g., pleural effusion and lung opacity). In such instances, the generated heatmap highlights areas for both diseases without accurately distinguishing them. This problem arises because the image classifier cannot differentiate diseases that frequently co-exist, where the presence of one disease serves as a strong bias for the presence of the other. To address these challenges, we propose the following improvements:

First, we replace disease classes with disease super-classes. Instead of training the classifier on 14 disease classes, we utilize medical knowledge to group these into 7 super-classes. For example, Consolidation, Edema, and Pneumonia are grouped into one class, as edema and pneumonia are child nodes of consolidation. The rationale for using super-classes is twofold: it eliminates ambiguity in distinguishing between a parent class and its child classes, and it acknowledges that some diseases share similar visual features but differ in their causes. This approach enables the classifier to focus on distinguishing visual features rather than the underlying causes.

Second, we employ sigmoid activation for training but softmax for testing. As multiple diseases can co-exist on a medical image, we train the classifier with sigmoid activation to use multiple disease class labels. During inference, however, we observed that the generated outputs using sigmoid gradients as de-noising guidance tend to highlight false positive regions. This issue can be intuitively explained with an example. Suppose that we are generating a diseased version of the input image where the disease class is cardiomegaly. We want the generated output to contain only cardiomegaly so that the difference heatmap will accurately highlight this particular disease. During the denoising process, we update the noisy image toward a direction that increases the sigmoid probability z_i for cardiomegaly. As the sigmoid equation eq. (1) considers only one classes, the generated output is not guaranteed to be free of other diseases (i.e. minimized $z_j, j \neq i$). In fact, as some diseases tend to co-exist (e.g. cardiomegaly and pleural effusion), the generated image tends to have both diseases, which introduces false positives. Using softmax as the activation during inference solves this issue because mathematically eq. (2) enforces maximized z_i while minimized probability of other classes ($z_j, j \neq i$). Intuitively, the task transforms from "generating an image where the chance of cardiomegaly is maximized" to "generating an image where the chance of cardiomegaly, compared to other diseases, is maximized". We validate the effectiveness with qualitative comparisions, as shown in Figures Figure 1 (bottom row) and fig. 3.

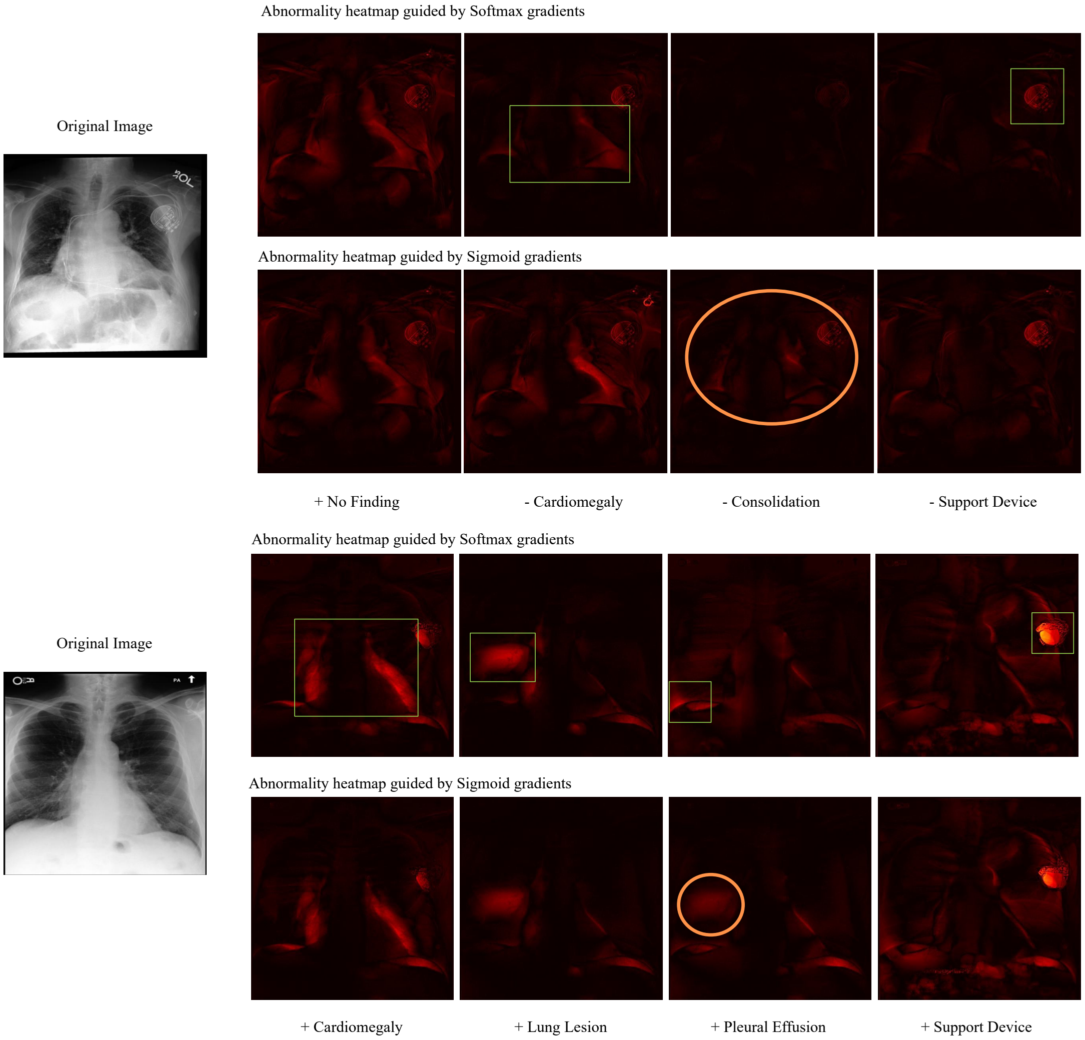
Abnormality heatmap guided by Softmax gradients
Original Image
Abnormality heatmap guided by Sigmoid gradients
+ No Finding
- Cardiomegaly
- Consolidation
- Support Device
Abnormality heatmap guided by Softmax gradients
Original Image
Abnormality heatmap guided by Sigmoid gradients
+ Cardiomegaly
+ Lung Lesion
+ Pleural Effusion
+ Support Device

Fig. 3: Example output of abnormality heatmaps generated by DAug. We use two chest X-rays as examples, each one covers four types of abnormalities (cardiomegaly, consolidation, etc. in each column). The **plus** $(+)$ and **minus** $(-)$ signs indicate the direction of the classifier gradient, meaning amplifying the disease and reducing the disease, respectively. **"+ No Findings"** reduces the probability of all potential diseases. For each input, the **first row** shows the output heatmap guided by gradients of **softmax** probabilities, and the **second row** shows the results guided by gradients of the **sigmoid** probabilities. The **green bounding boxes** shows that our method correctly highlights the region of the disease, which can help the mode to establish better image-text correspondence. Also, using softmax gradient is better than sigmoid gradient as guidance, as softmax successfully removes false positives (see orange circles). **Orange circle** in the second row highlights false positive areas of consolidation, and the orange circle in the last row highlights a wrong activation of lung lesion when it is supposed to detect pleural effusion. The corresponding softmax version (green box) makes correct detections

$$\text{sigmoid}(z_i) = \frac{1}{1+e^{-z_i}} \tag{1}$$

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} = \frac{1}{\sum_{j=1}^{K} e^{z_j - z_i}} \tag{2}$$

3.2.3. Channel-wise Feature Augmentation

The generated heatmaps can be integrated into the downstream perception model in various ways. A simple method involves processing them through a vision backbone and then concatenating or adding the heatmap features to the image features for downstream tasks. This straightforward approach necessitates modifications to the model architecture. We propose an alternative, channel-wise feature augmentation method, which incorporates the heatmaps as additional image channels alongside the medical image.

This channel-wise feature augmentation offers two main advantages. First, it leverages the capabilities of powerful pretrained vision models designed for RGB three-channel input. As radiology images are typically monochrome, it results in waste of model compute. Second, channel-wise augmentation does not require any changes to the model architecture, making it easier to utilize a wide range of pretrained transformers. This approach is complementary to other methods, enhancing performance on downstream tasks without the

		No Finding	Enl. Cardiomed.	Cardiomegaly	Lung Opacity	Lung Lesion	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural other	Fracture	Support Devices	wAvg	Avg
Zhang et al. [24]		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.485
X-TRA *CLIP ($\mathcal{L}_{CLIP}$)*		.62	.52	.93	.88	.50	.60	.29	.44	.75	.54	.85	.50	.36	.71	.606	.723
X-TRA *CNN+BERT*	$r \rightarrow x$	**.77**	.54	.73	.91	.52	**.83**	.39	**.87**	.77	.63	.74	.23	.61	.73	.662	.735
X-TRA *PubmedCLIP*		.75	.65	.92	**.99**	.23	.79	.21	.51	.59	**.72**	.81	.56	.43	.67	.645	.720
X-TRA *CLIP*		.63	.62	**.96**	.94	.62	.69	.47	.61	**.85**	.69	**.91**	.57	.46	**.82**	.703	**.779**
DAug *CLIP (ours)*		**.77**	**.86**	.83	.89	**.64**	.79	**.80**	.71	.81	.64	.79	**.66**	**.76**	.79	**.799**	.767
Yu et al. [25]		-	.65	.75	.72	.43	.80	.73	.60	.76	.76	.85	.43	.16	.86	-	.680
X-TRA[18] *CLIP ($\mathcal{L}_{CLIP}$)*		.71	.52	.74	.78	.39	.79	.39	.40	.76	.42	.67	.44	.43	.64	.578	.761
X-TRA *CNN+BERT*	$x \rightarrow x$	.87	.63	.88	**.90**	.49	.90	.57	.60	.85	**.85**	.83	.29	.47	.82	.678	.769
X-TRA *PubmedCLIP*		**.90**	.63	.82	.83	.39	.86	.45	.63	.87	.53	**.90**	.48	.51	.79	.685	.795
X-TRA *PubmedCLIP*		**.90**	.63	.82	.83	.39	.86	.45	.63	.87	.53	**.90**	.48	.51	.79	.685	.795
X-TRA *CLIP*		.84	.62	**.89**	.89	**.56**	**.91**	.55	.59	**.89**	.60	.86	.49	.57	**.84**	.713	**.840**
DAug *CLIP (ours)*		.72	**.86**	.83	.88	.53	.78	**.79**	**.65**	.81	.47	.77	**.59**	**.69**	.74	**.771**	.721

Table 1: **Retrieval performance** for both report-to-image (r-x) and image-to-image(x-x) scenarios, measured with retrieval mAP@K, where K=5. wAvg and Avg are weighted average of classes and average of classes, respectively. Our model, DAug-CLIP surpasses existing methods by a clear margin in the prevalent report-to-image scenario. In terms of image-to-image retrieval, DAug-CLIP achieves state-of-the-art performance on the most common classes and in terms of wAvg. This could be attributed to the reliability of the DAug feature in these classes, as the classifier guiding the diffusion model is affected by class imbalance.

	Aug	No Finding	Enl. Cardiomed.	Cardiomegaly	Lung Opacity	Lung Lesion	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural other	Fracture	Support Devices	wAvg	Avg
CNN+BERT	-	.81	.63	.73	.67	.62	.83	.69	.59	.68	.75	.83	.70	.58	.84	.71	.79
	X-TRA[18]	.81	.74	.75	.69	.63	.81	.72	.63	.75	.75	.83	.69	.63	.85	.73	.82
PubmedCLIP	-	.78	.65	.72	.66	.61	.82	.70	.61	.73	.76	.81	.62	.54	.84	.70	.78
	X-TRA[18]	.84	.76	.78	.69	.64	.83	.73	.64	.76	.75	.82	.75	.67	.85	.75	.83
	-	.77	.65	.71	.67	.62	.85	.73	.61	.72	.75	.80	.59	.51	.83	.70	.80
CLIP	X-TRA[18]	.82	.78	.74	.70	.71	.82	.75	.63	.79	.78	.86	.74	.72	.91	.77	.85
	DAug, ours	**.93**	**.90**	**.90**	**.91**	**.85**	**.89**	**.89**	**.85**	**.89**	**.90**	**.90**	**.87**	**.84**	**.89**	**.89**	**.89**

Table 2: **Image classification** performance on MIMIC-CXR dataset. Results are in AUC-ROC. wAvg is the weighted average by number of samples per class. Avg is the average. Augmentation method X-TRA improves baseline performance, with CLIP achieving the best result. DAug-CLIP, our method, surpasses the best setting of X-TRA by a clear margin in all classes.

need for architectural modifications.

3.3. Image-Text-Class Hybrid Contrastive Learning

Image-to-text retrieval and image classification are deeply interconnected tasks. Essentially, image classification can be seen as a retrieval problem focusing on a more defined set of targets. When the classification head is a linear layer without bias, the class logits become unscaled cosine similarities between the image feature and the weights in the linear layer. In image-to-text retrieval, the weights in the linear layer is replaced by text embeddings dynamically generated for each target text. Motivated by the potential benefits of jointly training both retrieval and classification tasks, we integrate image-text and image-class labels into our training loss. Unlike existing methods [26, 19] that expand contrastive learning to class labels, our approach uniquely addresses scenarios where a single sample is associated with both text and class labels, aiming to train a unified model for retrieval and classification. To distinguish with existing work, we name our method image-text-class hybrid contrastive learning.

As illustrated in fig. 2, we first transform each class into a fixed set of texts by converting each class into prompts. For instance, the class "cardiomegaly" is rephrased as "A photo of a Chest X-ray image with cardiomegaly". During training, as depicted in fig. 2, class prompts and radiology reports are transformed into text embeddings C and R, respectively. We then calculate two sets of losses and adjust their balance using a weight hyper-parameter:

$$\mathcal{L} = \mathcal{L}_{CLIP} + w * \mathcal{L}_{i2c}, \tag{3}$$

where $\mathcal{L}_{CLIP}$ is the CLIP loss including both image-to-text and text-to-image cross-entropy loss, and $\mathcal{L}_{i2c}$ is the image-to-class binary cross-entropy loss. Specifically,

$$\mathcal{L}_{i2c} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\mathrm{sim}(i,j)) + (1-y_i)\log(1-\mathrm{sim}(i,j))\right], \tag{4}$$

where $\mathrm{sim}(i,j)$ is the cosine similarity between the image embedding I_i and text embedding for class prompts C_i. During training, embedding C_i is regenerated with each update to the text encoder. Several works discussed the connection between contrastive learning and cross-entropy loss [26, 19]. In our scenario, $\mathcal{L}_{i2c}$ is essentially

Method	Feature Aug	Criterion	No Finding	Enl. Cardiomed.	Cardiomegaly	Lung Opacity	Lung Lesion	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural other	Fracture	Support Devices	wAvg	Avg
Image Retrieval: $r \rightarrow x$																		
X-TRA (Baseline)	X-TRA	CLIP	.62	.52	.93	.88	.50	.60	.29	.44	.75	.54	.85	.50	.36	.71	.606	.723
DAug (Ablation)	DAug	CLIP	**.77**	.85	.80	.86	.60	.76	.78	.67	.79	.61	.76	.65	.75	**.79**	.778	.745
DAug (Full)	DAug	Hybrid Contrastive	**.77**	**.86**	**.83**	**.89**	**.64**	**.79**	**.80**	**.71**	**.81**	**.64**	**.79**	**.66**	**.76**	**.79**	**.799**	**.767**
Image Retrieval: $x \rightarrow x$																		
X-TRA (Baseline)	X-TRA	CLIP	.71	.52	.74	.78	.39	.79	.39	.40	.76	.42	.67	.44	.43	.64	.578	.761
DAug (Ablation)	DAug	CLIP	.71	**.86**	.81	**.88**	.49	.76	.78	.62	**.81**	.44	.76	.58	**.69**	**.74**	.765	.711
DAug (Full)	DAug	Hybrid Contrastive	**.72**	**.86**	**.83**	**.88**	**.53**	**.78**	**.79**	**.65**	**.81**	**.47**	**.77**	**.59**	**.69**	**.74**	**.771**	**.721**
Image Classification																		
CLIP (Baseline)	-	CLIP	.77	.65	.71	.67	.62	.85	.73	.61	.72	.75	.80	.59	.51	.83	.70	.80
X-TRA (Baseline)	X-TRA	CLIP	.82	.78	.74	.70	.71	.82	.75	.63	.79	.78	.86	.74	.72	.91	.77	.85
DAug (Ablation)	DAug	CLIP	.77	.86	.83	.89	.64	.79	.80	.71	.81	.64	.79	.66	.76	.79	.80	.77
DAug (Full)	DAug	Hybrid Contrastive	**.93**	**.90**	**.90**	**.91**	**.85**	**.89**	**.89**	**.85**	**.89**	**.90**	**.90**	**.87**	**.84**	**.89**	**.89**	**.89**

Table 3: **Ablation studies** on image retrieval and classification tasks. Compared to baselines, our feature augmentation method DAug gains performance over no augmentation (CLIP Baseline) and the X-TRA augmentation (X-TRA Baseline). Besides, using the proposed Image-Text-Class Hybrid Contrastive loss outperforms the baselines using the original CLIP loss. Baseline results from [18].

contrasting image I with the class prompts, where multiple positive pairs could exist determined by the ground truth class labels. The inherent cross-entropy nature of both loss terms facilitates training both tasks simultaneously without causing the embedding space to diverge for each task.

3.4. Multi-modal Retrieval and Classification

After the model is trained, we leverage the model in different ways for the classification and retrieval tasks. For classification, we use only the image encoder and connect it with a linear classifier. The linear classifier consists of a single layer where the weight vectors per class are populated with the text embeddings for class prompts generated by the final text encoder. The bias is set to zeros. Essentially, the output logit per class is equivalent to the unscaled cosine similarity between the image and classes $\text{logit}_i = sigmoid(I \cdot C_i)$.

For retrieval tasks, both the image and text encoders are used to convert each sample into embeddings. We use the cosine similarity of the embeddings to rank the association for retrieval.

4. EXPERIMENTS

4.1. Implementation details

To benchmark our ideas, we need an image-text-class dataset on medical images. We select **MIMIC-CXR**, the largest Chest X-ray (CXR) medical report dataset. It contains 227,835 image-text pairs, where the texts are radiology reports which lists the normal and abnormal findings. As class labels are unavailable, and generate pseudo-class labels with the CheXpert labeler [22]. It is a text classification model which converts a radiology report into binary labels on 14 disease classes. One of them is "No Findings", indicating a healthy case.

For DAug feature augmentation, we construct a three-channel image with the first two channels containing the medical image and the third channel filled with the diffusion-generated heatmap. In our experiments, we evaluate on all 14 disease classes and compare with the original CLIP as a baseline. Therefore, we selected the heatmap for "No Finding" which combines all diseases. This requires no change on the model architecture for a fair comparison with the vanilla CLIP. In real-world applications, heatmaps for individual disease groups (e.g., cardiomegaly) can be selected for optimal performance gain according to the scenario.

Following existing work [18], we resize images to 256 × 256 and use a CLIP pretrained ViT-Base/32 model for fair comparison. We fine-tune the model with a cosine learning rate scheduler with a base learning rate of $2e^{-5}$ for 10 epochs. We use a batch size of 256 over eight $V-100$ GPUs. With all images resized in advance, training and evaluation takes around 2 hours. Abnormality heatmaps are pre-generated.

4.2. Results

We compare our method, DAug CLIP with existing methods on both multi-modal retrieval and image classification on the MIMIC dataset. Table 1 demonstrates that our method outperforms existing state-of-the-art approaches in retrieving radiology images with medical reports, a critical clinical scenario where radiologists refer to previous cases to confirm diagnoses. Table 2 shows performance of the same model on image classification, which outperforms existing methods by a clear margin.

Specifically, our model clearly outperforms existing work on the classification task, thanks to the DAug feature augmentation, which guides the model on where to look for diseases, and the Hybrid Contrastive Loss, which enables the model to learn from both text and class labels. In the retrieval tasks, although we outperform X-TRA on the weighted average mAP, our method does not perform as well on the non-weighted average mAP. This is reasonable because X-TRA augments features with similar samples from the dataset, which particularly benefits the performance of tail classes.

We conduct ablation studies in Table 3. The results show that performance on both tasks improved incrementally with the addition of the DAug feature augmentation and the Hybrid Contrastive criterion. This validates that both methods aid in medical image understanding tasks.

5. CONCLUSION

We propose DAug, a single model that achieves state-of-the-art performance in both medical image retrieval and classification tasks. DAug consists of two novel methods: a diffusion-based feature augmentation method and an Image-Text-Class Hybrid Contrastive learning criterion. Experiments show that both methods improve model performance in medical image understanding tasks.

6. REFERENCES

[1] Daniel J Cao, Casey Hurrell, and Michael N Patlas, "Current status of burnout in canadian radiology," *Canadian Association of Radiologists Journal*, vol. 74, no. 1, pp. 37–43, 2023.

[2] Abi Rimmer, "Radiologist shortage leaves patient care at risk, warns royal college," *BMJ: British Medical Journal (Online)*, vol. 359, 2017.

[3] Prafulla Dhariwal and Alexander Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[4] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin, "Diffusion models for medical anomaly detection," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*. Springer, 2022, pp. 35–45.

[5] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, pp. 317, 2019.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[8] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[9] Yang Song and Stefano Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[10] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim, "Data synthesis based on generative adversarial networks," *arXiv preprint arXiv:1806.03384*, 2018.

[11] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*. Springer, 2018, pp. 1–11.

[12] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 289–293.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16000–16009.

[14] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun, "Medclip: Contrastive learning from unpaired medical images and text," *arXiv preprint arXiv:2210.10163*, 2022.

[15] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel, "Does clip benefit visual question answering in the medical domain as much as it does in the general domain?," *arXiv preprint arXiv:2112.13906*, 2021.

[16] Jun Wang, Abhir Bhalerao, and Yulan He, "Cross-modal prototype driven network for radiology report generation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*. Springer, 2022, pp. 563–579.

[17] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan, "Cross-modal memory networks for radiology report generation," *arXiv preprint arXiv:2204.13258*, 2022.

[18] Tom van Sonsbeek and Marcel Worring, "X-tra: Improving chest x-ray tasks with cross-modal retrieval augmentation," in *International Conference on Information Processing in Medical Imaging*. Springer, 2023, pp. 471–482.

[19] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao, "Unified contrastive learning in image-text-label space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19163–19173.

[20] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[21] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren, "Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert," *arXiv preprint arXiv:2004.09167*, 2020.

[22] Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar, "Visualchexbert: addressing the discrepancy between radiology report labels and image labels," in *Proceedings of the Conference on Health, Inference, and Learning*, 2021, pp. 105–115.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[24] Yong Zhang, Weihua Ou, Jiacheng Zhang, and Jiaxin Deng, "Category supervised cross-modal hashing retrieval for chest x-ray and radiology reports," *Computers & Electrical Engineering*, vol. 98, pp. 107673, 2022.

[25] Yang Yu, Peng Hu, Jie Lin, and Pavitra Krishnaswamy, "Multimodal multitask deep learning for x-ray image retrieval," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer, 2021, pp. 603–613.

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.