

Peer Review

Review of: "DAug: Diffusion-based Channel Augmentation for Radiology Image Retrieval and Classification"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper proposes helping a model learn where to look on a chest X-ray by adding abnormality heatmaps as extra image channels next to the monochrome radiograph (Abstract, Section 1). The method is called Diffusion-based Feature Augmentation (DAug). To make the heatmaps, a DDPM trained on chest X-rays receives an input image, noise is added for 500 steps, and denoising is then guided by a disease classifier toward a diseased or healthy version (Section 3.2). The difference between the input and the output is the heatmap. The classifier is trained with sigmoid and used with softmax gradients at inference to reduce false positives (Section 3.2.1, Figure 2). It is trained on seven disease super-classes instead of the 14 CheXbert classes (Section 3.1.1, Appendix B, Table 4). The heatmaps enter as image channels, so a pretrained RGB model needs no architecture change (Section 3.2.2).

The second contribution is Image-Text-Class Hybrid Contrastive learning (Section 4, Figure 3). Each class becomes a text prompt, and the loss is a weighted average of the CLIP image-text loss and an image-class binary cross-entropy loss. One model then serves both retrieval and classification (Section 4.1). The setup uses MIMIC-CXR with 227,835 image-text pairs, the official splits, and pseudo class labels from the CheXbert labeler, with CLIP ViT-Base/32 as the backbone (Section 5.1). Channels one and two hold the image, and channel three holds the heatmap guided by "No Findings". The authors claim state-of-the-art performance on medical image retrieval (Table 1, mAP@5) and on classification (Table 2, AUC-ROC), and they report an ablation in Table 3 (Section 5.2, Section 6).

2. Whether the conclusion holds on the data

The idea of giving the generative model's knowledge of normal and abnormal images to the network as an extra input channel is simple, and I like that it keeps the pretrained architecture unchanged (Section 3.2.2). The softmax-at-inference trick is also shown clearly in Figure 2.

My first remark is about Table 3. In the image classification block, the row "DAug (Ablation), DAug, CLIP" reads .77 .86 .83 .89 .64 .79 .80 .71 .81 .64 .79 .66 .76 .79, with Avg .77 and wAvg .80. These are the same per-class numbers as the retrieval row "DAug CLIP (ours)" in the report-to-image block of Table 1, where Avg is .767 and wAvg is .799. That Table 1 row is mAP@5, not AUC-ROC, so it looks like a copied row. If so, the gain attributed to the hybrid loss in classification (wAvg .80 to .89, Table 3) cannot be read from the table. I would ask the authors to check this row and replace it.

My second remark is about the image-to-image block of Table 1. The caption says DAug-CLIP "achieves state-of-the-art performance on the most common classes". Against X-TRA CLIP, DAug CLIP is lower on No Finding (.72 vs .84), Atelectasis (.81 vs .89), Pleural Effusion (.77 vs .86), and Support Devices (.74 vs .84). It is higher on Enlarged Cardiomedastinum (.86 vs .62), Consolidation (.79 vs .55), and Fracture (.69 vs .57). Section 5.2 says X-TRA benefits more in the tail classes, but the numbers in this block suggest the opposite pattern. It would help to look again at both the caption and the text so that they match the table.

On the main metric in Table 1, Avg and wAvg point in different directions, and the caption calls wAvg the primary metric. For report-to-image retrieval, DAug CLIP has Avg .767 against .779 for X-TRA CLIP, but wAvg .799 against .703. For image-to-image retrieval, DAug CLIP has Avg .721 against .840 for X-TRA CLIP, but wAvg .771 against .713. A weighted mean that rises while the plain mean falls by .119 depends entirely on the class weights, so I would print the number of test samples per class next to the table. The baseline in Table 3 also needs clarification. The retrieval rows "X-TRA (Baseline)" (wAvg .606 and .578) match the first, unlabelled rows of Table 1, not the stronger X-TRA CLIP rows (wAvg .703 and .713). It would help to state which X-TRA configuration the ablation compares against.

Only one heatmap is used: channel three is the heatmap guided by "No Findings" (Section 5.1). In Table 4, the classifier for super-class 1 (No Finding) has an AP of 0.631, among the lowest of the seven, and Appendix B says heatmap quality is highly correlated with classifier performance of the selected class. I would add a control with the same three-channel input where channel three is zeros, or a copy of the image. It would show how much of the gain comes from the heatmap itself. The heatmaps are also judged only through downstream tasks (Section 5.2), since the authors note there is no ground truth. A small

check against radiologist-drawn boxes on a public chest X-ray set would support the claim that the model learns where to look.

3. Reproducibility

Much of what a reader needs is already here. Section 5.1 gives the MIMIC-CXR official splits, CLIP ViT-Base/32, a cosine learning rate schedule, a loss weight of 0.7 chosen on validation, and a batch size of 256 on eight V-100 GPUs, with about 2 hours of training and evaluation on pre-generated heatmaps. The 500 noising steps are in Section 3.2 and Appendix A, and Appendix D records the PhysioNet approval for data access.

I would add a code and model release statement, since the paper has none. It would help to report the number of runs, seeds, and variance for each table, because several per-class gaps in Table 1 are only a few hundredths. The cost of generating the heatmaps is also missing: Appendix D calls the diffusion overhead significant but gives no time per image. In Table 4, the super-classes cover eleven of the 14 classes, and Lung Opacity, Pneumothorax, and Fracture are in none of them; a sentence on why would help. Finally, the test labels are CheXbert pseudo-labels (Section 5.1), yet Appendix C names the labeler "CheXpert" and cites a 96.9% F1, while Sections 3.1 and 5.1 say CheXbert. One name throughout would avoid confusion.

4. What specifically to fix

1. Table 3, image classification block. The row "DAug (Ablation), DAug, CLIP" carries the same numbers as the report-to-image row "DAug CLIP (ours)" of Table 1 (mAP@5). Could you replace it with the AUC-ROC values?
2. Table 1 needs its caption and Section 5.2 brought in line with the image-to-image numbers. In that block, DAug CLIP sits below X-TRA CLIP on No Finding (.72 vs .84), Pleural Effusion (.77 vs .86), and Support Devices (.74 vs .84). I would also add the number of test samples per class that was used for wAvg.
3. A formatting matter in the Qeios HTML version of Table 1. Some rows have no method name, since the first row of each block starts with ")" or "-". In the image-to-image block, the row "X-TRA PubmedCLIP" appears twice with identical numbers. It would help to name every row and remove the duplicate.

4. Table 3 should say which X-TRA configuration is the baseline. Its rows (wAvg .606 and .578) are the weaker unlabelled rows of Table 1, and they are not X-TRA CLIP (.703 and .713).
5. For the channel three idea, I would add a control that uses zeros or a copy of the image in place of the "No Findings" heatmap. I would also report seeds or variance for the main tables.
6. Last, a few items for the paper as a whole. Please state code availability and the heatmap generation time per image. Please also explain why Lung Opacity, Pneumothorax, and Fracture are outside the seven super-classes (Table 4). And I would use one name for the labeler, either CheXbert or CheXpert, across Sections 3.1, 5.1, and Appendix C.

5. Recommendation

Major revisions. The idea of feeding diffusion heatmaps to a pretrained model as an extra channel is simple and portable, the softmax guidance is shown clearly (Figure 2), and a single model for retrieval and classification is practical. Before acceptance, I would correct the classification row in Table 3, which repeats retrieval numbers from Table 1, and align the Table 1 caption with the image-to-image numbers. It would also help to show per-class sample counts for wAvg and to add a channel-three control. I would be glad to read a revised version.

Declarations

Potential competing interests: No potential competing interests to declare.