Leveraging Foundation Models for Content-Based Medical Image Retrieval in Radiology

Stefan Denner[1,2], David Zimmerer[1], Dimitrios Bounias[1,2], Markus Bujotzek[1,2], Shuhan Xiao[1,3], Rafael Stock[1,3], Lisa Kausch[1], Philipp Schader[1,3], Tobias Penzkofer[4], Paul F. Jäger[5,6], and Klaus Maier-Hein[1,2,3,7]

[1] German Cancer Research Center (DKFZ), Division of Medical Image Computing, Heidelberg, Germany
[2] Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany
[3] Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany
[4] Department of Radiology, Charité - Universitätsmedizin Berlin, Berlin, Germany
[5] German Cancer Research Center (DKFZ), Interactive Machine Learning Group (IML), Heidelberg, Germany
[6] German Cancer Research Center (DKFZ), Helmholtz Imaging, Heidelberg, Germany.
[7] Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
stefan.denner@dkfz-heidelberg.de

Introduction. In recent years, a key advancement in Content-based image retrieval (CBIR) systems has been the adoption of deep learning models to encode the visual features of images [2]. However, medical CBIR systems often lack generalizability, being limited to the few pathologies they were explicitly trained on, which does not meet the diverse needs of radiology for a versatile retrieval system [1,5]. Developing a feature extractor that accurately retrieves a wide range of pathologies is challenging due to data and compute requirements. To address this, we explore using vision foundation models as robust and versatile feature extractors for CBIR.

Materials and Methods. We evaluate foundation models' performance in medical image retrieval using four public datasets—RadImageNet [6], NIH14 [7], MIMIC [4], and CheXpert [3]—comprising over 1.6 million 2D images across four modalities, 12 anatomical regions, and 185 classes. Our evaluation includes models from self-supervised, weakly-supervised, and fully-supervised learning schemes, such as ResNet, Vision Transformer (ViT), SAM, MedSAM, CLIP, MedCLIP, BiomedCLIP, MAE, and DINOv2, using their respective embeddings. The models are used off-the-shelf without fine-tuning. To this end, we (1) resize the images to the required input size for each model, (2) extract the features, (3) perform $L2$ normalization on the models' embeddings, and (4) store these normalized embeddings in a vector database. During retrieval (5), we measure the cosine similarity between the query image encoding and all other image encodings in the index. We retrieve the top N images with the highest similarity, sorted by descending order. We measure the retrieval performance between the query image and the top N retrieved images, quantified by Precision at N ($P@N$).

Table 1. Retrieval performance of the evaluated models on the combined dataset.

	Sample-wise (micro)				Class-wise (macro)			
	P@1	**P@3**	**P@5**	**P@10**	**P@1**	**P@3**	**P@5**	**P@10**
ResNet	0.544	0.538	0.535	0.529	0.204	0.193	0.189	0.182
ViT	0.560	0.554	0.550	0.543	0.217	0.204	0.199	0.190
SAM	0.520	0.517	0.515	0.511	0.197	0.187	0.182	0.175
MedSAM	0.489	0.483	0.478	0.470	0.154	0.146	0.142	0.134
CLIP	0.571	0.567	0.565	0.559	0.222	0.213	0.209	0.202
MedCLIP	0.566	0.561	0.559	0.554	0.223	0.211	0.205	0.198
BiomedCLIP	**0.594**	**0.590**	**0.588**	**0.583**	**0.240**	**0.230**	**0.224**	**0.217**
MAE	0.513	0.507	0.503	0.497	0.183	0.177	0.172	0.166
DINOv2	0.553	0.548	0.545	0.540	0.219	0.204	0.199	0.192

Results and Discussion. Evaluating the retrieval performance of the foundation models on the combined dataset, we observe that BiomedCLIP achieves the highest score across all metrics (Table 1). In general, weakly-supervised approaches outperform other training schemes. Segmentation models (SAM and MedSAM) show lower performance, suggesting their embeddings are not well-suited for medical image retrieval. Furthermore, medical adaptations (MedSAM and MedCLIP) fail to surpass their natural image counterparts (SAM and CLIP), indicating that their embeddings don't generalize. The superior performance of CLIP-based approaches can probably be attributed to its training objective of semantically meaningful aligning the representations of images and texts. BiomedCLIP in particular performs well, probably due to its exposure to a wide variety of biomedical image-text pairs, leading to a nuanced understanding of pathologies, which is reflected in the embedding space. However, sample-wise scores are significantly higher than class-wise scores across all models, 0.594 and 0.240 for BiomedCLIP P@1, respectively.

Conclusion. To conclude, our analysis highlights the promising off-the-shelf capabilities of weakly-supervised models, especially BiomedCLIP, strongly supporting the pursuit of further exploration into even more advanced vision foundation models, that can accurately understand and represent pathological similarities.

References

1. Choe, J., et al.: Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest ct. Radiology **302**(1), 187–197 (2022)
2. Gordo, A., et al.: Deep image retrieval: Learning global representations for image search. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 241–257. Springer (2016)
3. Irvin, J., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)

4. Johnson, A., et al.: Mimic-iii, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)

5. Langlotz, C.P.: The future of ai and informatics in radiology: 10 predictions (2023)

6. Mei, X., et al.: Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)

7. Wang, X., et al.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)