OFF-CLIP: Improving Normal Detection Confidence in Radiology CLIP with Simple Off-Diagonal Term Auto-Adjustment

Junhyun Park[1⋆], Chanyu Moon[1⋆], Donghwan Lee[2], Kyungsu Kim[2⋆⋆], and Minho Hwang[1⋆⋆]

[1] DGIST, Republic of Korea
{sean05071, anscksdb0127, minho}@dgist.ac.kr
[2] Seoul National University, Republic of Korea
{tdr.lee, kyskim}@snu.ac.kr

Abstract. Contrastive Language-Image Pre-Training (CLIP) based models enable zero-shot classification in radiology but often struggle with detecting normal cases due to rigid intra-sample alignment, which leads to poor feature clustering and increased false positive and false negative rates. We propose OFF-CLIP, a simple and effective refinement that introduces an off-diagonal loss term to promote the clustering of normal samples explicitly. In addition, it applies sentence-level filtering to remove typical normal phrases embedded within abnormal reports. OFF-CLIP does not require architectural changes and does not compromise abnormal classification performance. In the VinDr-CXR dataset, normal classification shows a notable 0.61 AUC improvement over the state-of-the-art baseline CARZero. It also improves zero-shot grounding performance by increasing pointing game accuracy and providing more reliable and precise anomaly localization. These results clearly demonstrate that OFF-CLIP serves as an efficient plug-and-play enhancement to existing medical vision-language models. The code and pre-trained models are publicly available at https://github.com/Junhyun-Park01/OFF-CLIP.

Keywords: CLIP · Zero-shot learning · Medical image-text alignment

1 Introduction

Deep learning has significantly advanced medical imaging [3,11], but its reliance on large-scale annotated datasets limits scalability. Zero-shot learning (ZSL) mitigates this by enabling generalization without extensive manual labeling. Contrastive learning, especially vision–language pretraining, has emerged as a powerful approach for aligning large-scale image–text pairs [17], and has been adapted to radiology for zero-shot classification and anomaly detection [9,14,18,20,23,24,19].

GLoRIA [9] introduced text-weighted local attention, and KAD [23] incorporated domain knowledge for entity-aware representations. MedCLIP improved

⋆ Equal contribution as first authors
⋆⋆ Correspondence to Minho Hwang and Kyungsu Kim

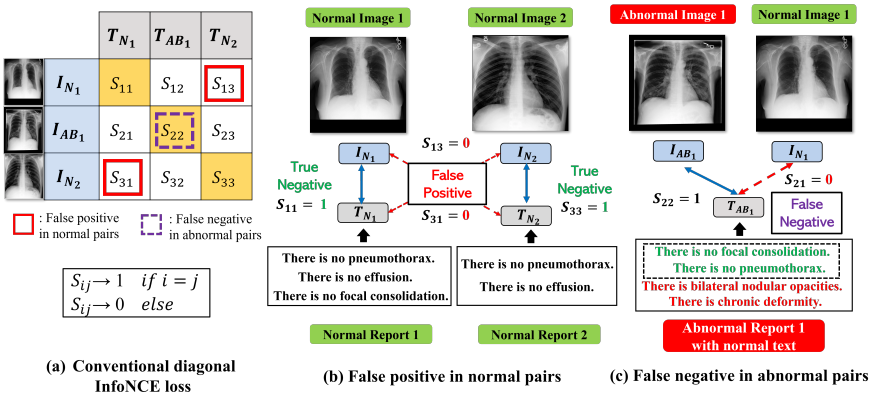

Fig. 1. The figure shows two issues in (a) conventional diagonal InfoNCE loss in Radiology CLIP: (b) high FPs from aligning only matched pairs, separating normal samples, and (c) high FNs caused by normal sentences in abnormal reports, pulling them closer to abnormal images and away from normal ones.

data efficiency by leveraging cross-patient image–text combinations. CARZero [14] enhanced abnormality alignment via cross-attention.

Most methods rely on InfoNCE loss, which aligns only matched image–text pairs in the $B \times B$ similarity matrix (where B is the batch size). This strict diagonal constraint ignores relationships among normal samples, forcing semantically similar cases apart and increasing false positives (FPs) (Fig. 1-(b)). Additionally, normal statements in abnormal reports introduce misalignment, leading to false negatives (FNs) (Fig. 1-(c)). As shown in Table 1, most methods do not consider normal case limitations, resulting in low normal AUCs and poor screening reliability. While MedCLIP acknowledges similar issues, it relies on fixed labels and is incompatible with open-vocabulary CLIP settings.

To address remaining challenges, we propose OFF-CLIP, a novel contrastive learning framework that refines medical image–text alignment. OFF-CLIP introduces an off-diagonal loss to enhance normal clustering and a sentence-level text filter to remove normal statements from abnormal reports, reducing FPs and FNs. These modifications improve both normal detection and anomaly localization, outperforming the CARZero baseline [14] in zero-shot classification.

- We identify two key limitations in medical contrastive learning: (i) poor normal sample clustering leading to high FPs and (ii) misalignment from normal text in abnormal reports increasing FNs.
- We introduce (i) an off-diagonal term loss to enhance normal clustering and (ii) a text filtering strategy to mitigate misalignment, effectively reducing both FPs and FNs.
- We show that OFF-CLIP improves 0.5 on normal AUC while preserving abnormal AUC in zero-shot classification and enhances anomaly localization in the zero-shot grounding task, outperforming the CARZero baseline.

Table 1. Comparison of OFF-CLIP with Existing Radiology CLIP Models

Study	Text Input	Text Encoder	Img Encoder	CLIP Loss	Similarity	Require Annotation?
GLoRIA [9]	Full report	BioClinical BERT [1]	ResNet-50	InfoNCE	Cosine	No
KAD [23]	Extracted entities	PubMed BERT [7]	ResNet-50, ViT-16	InfoNCE, BCE	Cosine	No
CheXZero [18]	Impression sec.	Transformer (CLIP) [17]	ViT-B/32	InfoNCE	Cosine	No
MedCLIP [19]	Full report	BioClinical BERT [1]	Swin Transformer	Multi-label CE	Cosine	Yes
CARZero [14]	Prompting + rand. select	Bio BERT [15]	ViT-B/16	InfoNCE	Cross-Attn	No
OFF-CLIP (Ours)	Prompting + rand. select + **Text filter**	Bio BERT [15]	ViT-B/16	**Off-diag. & abn. InfoNCE**	Agnostic (Cosine & Cross-Attn)	**No** (pseudo-label)

2 Method

2.1 Image and Text Encodings

We employ a text encoder and an image encoder to extract global and local feature representations. OFF-CLIP supports ViT-based [6] and ResNet-based [8] models. In this study, we used ViT-B/16 trained with M3AE [4]. For text encoding, any pre-trained BERT model can be used, and we adopt BioBERT [15] for its domain-specific advantages.

2.2 Off-Diagonal Term Loss and Abnormal InfoNCE Loss

Baseline Loss The conventional InfoNCE loss is formulated as:

$$\mathcal{L}_{\text{baseline}} = -\frac{1}{B}\sum_{i=1}^{B}\left(\log\frac{e^{S^{i,i}}}{\sum_{j=1}^{B} e^{S^{i,j}}} + \log\frac{e^{S^{i,i}}}{\sum_{j=1}^{B} e^{S^{j,i}}}\right). \tag{1}$$

where B is the batch size, and $S \in \mathbb{R}^{B\times B}$ is the similarity matrix between global and local image-text feature vectors. Different methods can be used to construct S, and we adopt a cross-attention-based approach [14] in this study.

Off-Diagonal Term Loss The baseline loss (1) relies on intra-sample image-text associations, misaligning normal representations, and increasing FPs. To mitigate this, we introduce an off-diagonal term loss that clusters normal samples within a batch, formulated as:

$$\begin{aligned}\mathcal{L}_{\text{off}} = -\frac{1}{2B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}&\Big(\hat{Y}_{i,j}\log\sigma(S_{i,j}) + (1-\hat{Y}_{i,j})\log(1-\sigma(S_{i,j}))\Big)\\ &+\Big(\hat{Y}_{j,i}\log\sigma(S_{j,i}) + (1-\hat{Y}_{j,i})\log(1-\sigma(S_{j,i}))\Big).\end{aligned} \tag{2}$$

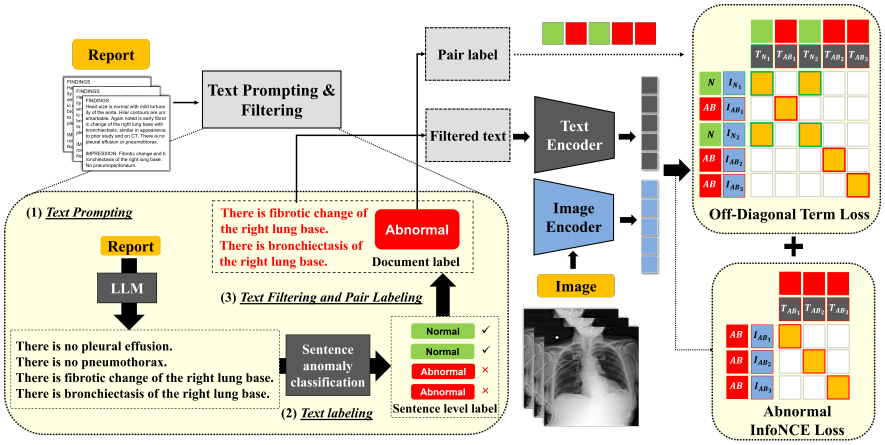

Fig. 2. OFF-CLIP leverages an off-diagonal term loss to effectively cluster normal samples within a batch. Abnormal pairs are further refined using an abnormal-only InfoNCE loss. Reports are processed using an LLM for text prompting, and sentence-level anomaly classification is applied to label each sentence. Normal sentences in abnormal reports are then filtered to reduce misalignment.

where $S \in \mathbb{R}^{B\times B}$ is the similarity matrix between image and text latents, $\sigma(S_{i,j}) = \frac{1}{1+e^{-S_{i,j}}}$ is the sigmoid activation, and $\hat{Y}_{i,j}$ is defined as:

$$\hat{Y}_{ij} = \begin{cases} 1, & \text{if } i = j \text{ (diagonal) or both } i, j \text{ are pseudo-normal pairs } (\hat{n}) \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Pseudo-labels for each pair $(\hat{n}, \hat{a})$ are generated using a pre-trained radiology sentence-level anomaly classifier [13], as detailed in Section 2.3, without human annotation. This formulation extends CLIP loss by reinforcing positive alignment among normal pairs while preserving standard diagonal alignment for all pairs, including abnormal ones.

Abnormal InfoNCE Loss The off-diagonal term loss increases the number of normal labels, which may reduce sensitivity to abnormal cases. To counteract this, we only apply the original InfoNCE loss to abnormal pairs. Let A be the number of abnormal pairs in the batch. The abnormal similarity matrix $S_{ab} \in \mathbb{R}^{A\times A}$ is extracted from the full similarity matrix S, containing only abnormal similarity scores:

$$\mathcal{L}_{\text{ab}} = -\frac{1}{A}\sum_{i=1}^{A}\left(\log\frac{e^{S_{ab}^{i,i}}}{\sum_{j=1}^{A} e^{S_{ab}^{i,j}}} + \log\frac{e^{S_{ab}^{i,i}}}{\sum_{j=1}^{A} e^{S_{ab}^{j,i}}}\right). \tag{4}$$

Total Loss The total loss of OFF-CLIP is defined as:

$$\mathcal{L}_{\text{OFF-CLIP}} = \mathcal{L}_{\text{off}} + \lambda_{\text{ab}}\mathcal{L}_{\text{ab}} \tag{5}$$

Table 2. Zero-shot classification performance evaluated using AUC. OFF-CLIP integrates text filtering, off-diagonal term loss, and abnormal InfoNCE loss. We evaluate it on VinDr-CXR, Open-I, PadChest, and CheXpert, applying OFF-CLIP to CARZero. Note that results may differ from the original baseline paper as we retrained CARZero to ensure consistency.

Dataset	OFF -CLIP	Normal	ATE	CM	EMPH	EFF	NOD	PLT	PNA	TOT
VinDr-CXR	✗	0.25	0.74	0.91	0.93	0.94	0.47	0.75	**0.90**	0.79
	✓	**0.86**	**0.84**	**0.93**	**0.95**	**0.94**	**0.86**	**0.86**	0.83	**0.87**
Open-I	✗	0.32	0.79	0.89	0.83	0.93	0.55	0.73	0.84	0.72
	✓	**0.74**	**0.81**	**0.90**	**0.91**	**0.93**	**0.60**	**0.76**	**0.84**	**0.81**
PadChest	✗	0.24	0.78	0.87	0.78	0.95	0.52	0.66	0.80	0.68
	✓	**0.77**	**0.82**	**0.88**	**0.89**	**0.95**	**0.64**	**0.81**	**0.80**	**0.76**
CheXpert	✗	0.38	0.73	0.86	-	0.92	-	-	0.65	0.73
	✓	**0.81**	**0.87**	**0.88**	-	**0.92**	-	-	**0.79**	**0.86**

* ATE, CM, EMPH, EFF, NOD, PLT, PNA, and TOT represent Atelectasis, Cardiomegaly, Emphysema, Effusion, Nodule, Pleural Thickening, Pneumonia, and Total, respectively.

where λ_{ab} (set to 1) balances the abnormal InfoNCE loss, preserving normal sample clustering while maintaining sensitivity to abnormal cases.

2.3 Text Prompting and Filtering

Text Prompting We employ the GPT-4o model to extract the Findings and Impressions sections from medical reports. The extracted text is reformatted into a structured template: There is disease (e.g., There is no pneumonia, There is pleural effusion at the left lung base), following the prompting format introduced in CARZero [14]. Although no quantitative evaluation was conducted, GPT-4o consistently produced coherent and clinically appropriate prompts, and minor deviations from the template did not cause noticeable performance degradation.

Pseudo-Label Extraction CLIP training operates in a zero-supervision setting, making direct human annotation infeasible. To overcome this, we use a pretrained sentence-level anomaly classification model for radiology [13] to generate pseudo-labels at both the sentence and report levels. The model is based on RadBERT [21] and trained via knowledge distillation from GPT-3.5 without human annotations, achieving an AUC of 0.977 on sentence-level anomaly detection.

Each sentence is labeled as normal, abnormal, or uncertain. Report-level labels are determined by aggregating sentence labels: a report is considered abnormal ($\hat{a}_{report}$) if it contains at least one abnormal sentence; otherwise, it is labeled as normal ($\hat{n}_{report}$). These report-level pseudo-labels are then assigned to the corresponding image–text pairs ($\hat{a}$ or $\hat{n}$), ensuring label consistency for contrastive learning.

Table 3. Ablation Study on CARZero Baseline for Zero-Shot Classification Performance using the Open-I Dataset, and evaluated by AUC. By training the baseline with 6 different options, and we validated on the Open-I dataset.

Text Filter	$\mathcal{L}_{off}$	$\mathcal{L}_{ab}$	Normal	ATE	CALC	CM	EMPH	MASS	NOD	OPAC	PLT	Total
			0.32	0.79	0.59	0.89	0.83	0.68	0.55	**0.79**	0.73	0.72
	✓		0.72	0.29	0.49	0.34	0.33	0.31	0.47	0.24	0.28	0.32
	✓	✓	**0.75**	0.24	0.46	0.30	0.27	0.27	0.44	0.23	0.26	0.29
✓			0.62	0.58	0.50	0.64	0.46	0.27	0.49	0.31	0.41	0.44
✓	✓		0.72	0.77	0.55	0.84	0.85	0.76	**0.61**	0.75	0.74	0.78
✓	✓	✓	0.74	**0.81**	**0.60**	**0.90**	**0.91**	**0.85**	0.60	0.73	**0.76**	**0.81**

* CALC, MASS, OPAC refer to Calcification, Mass and Opacity, respectively. And the another abbreviation are defined on previous table.

Table 4. Ablation study on the CARZero baseline for zero-shot classification performance on the Open-I dataset. We report FP and FN rates, along with their proportions relative to the total sample count. Additionally, we assess the FP-FN balance and demonstrate how text filtering mitigates FN errors. FN reduction is quantified as the relative improvement when text filtering is enabled.

row#	Text Filter	$\mathcal{L}_{off}$	$\mathcal{L}_{ab}$	$\frac{\text{FN}}{\text{Total}}$	$\frac{\text{FP}}{\text{Total}}$	$\frac{\text{FN}}{\text{FP+FN}}$	$\frac{\text{FP}}{\text{FP+FN}}$	$\lvert\frac{\text{FN}}{\text{FP+FN}} - \frac{\text{FP}}{\text{FP+FN}}\rvert$	$\frac{\text{FN}}{\text{Total}}$ reduction
1				0.0003	0.37	0.0007	**0.9993**	**0.9986**	N/A (abn. biased)
2		✓		0.14	0.16	0.47	0.53	0.06	-
3		✓	✓	0.17	0.13	0.57	0.43	0.14	-
4	✓			0.08	0.22	0.27	0.73	0.46	-
5	✓	✓		0.13	0.17	0.43	0.57	0.14	↓**7.1%** (vs. row 2)
6	✓	✓	✓	0.15	0.14	**0.51**	**0.50**	**0.01**	↓**11.8%** (vs. row 3)

Text Filtering for FN Reduction To reduce FNs, we eliminate normal and uncertain sentences from abnormal reports. This prevents contrastive learning from misaligning normal descriptions with abnormal images or pushing normal descriptions away from normal images. By ensuring accurate abnormal image-text associations, this filtering step enhances anomaly detection.

3 Experimental Results

3.1 Datasets

We train OFF-CLIP using the MIMIC-CXR dataset [12], which contains 377,110 chest radiographs with associated reports. Only frontal images are retained; for studies with multiple frontal images, one is randomly selected, and for reports with several prompted sentences, one is chosen per epoch. Training is further restricted to images from the p10-16 folders within the p10-19 range.

For evaluation and ablation, we use four datasets: VinDr-CXR [16] (18,000 X-rays, 28 disease annotations with bounding boxes; 3,000 evaluation scans,

Table 5. Zero-shot grounding performance comparison using the pointing game with top-10% and top-20% attention on the VinDr-CXR dataset, evaluated for the CARZero baseline and CARZero with OFF-CLIP. Attention maps were assessed for each disease-specific prompt.

Attention	Methods	AE	ATE	CALC	CM	CONS	EFF	INF	NOD	OPAC	PLT	PTX
High 10%	Baseline	0.94	0.61	0.30	0.99	0.84	0.69	0.8	0.47	0.66	0.33	0.63
	OFF-CLIP	**1.0**	**0.78**	**0.59**	**1.0**	**0.95**	**0.89**	**0.9**	**0.73**	**0.92**	**0.47**	**1.0**
High 20%	Baseline	0.99	0.78	0.59	1.0	0.97	0.79	0.9	0.64	0.83	0.47	0.63
	OFF-CLIP	**1.0**	**0.89**	**0.82**	**1.0**	**0.97**	**1.0**	**0.95**	**0.84**	**0.96**	**0.67**	**1.0**

* AE, CONS, and INF denote Aortic Enlargement, Consolidation, and Infiltration, respectively. Other abbreviations are defined in the previous table.

68.3% normal), Open-I [5] (7,470 X-rays, 18 disease annotations), CheXpert [10] (224,316 X-rays, 14 disease annotations; evaluated on 500 cases), and PadChest [2] (160,868 X-rays, 192 disease annotations; evaluated on 39,053 manually labeled cases). All datasets are publicly available. MIMIC-CXR, VinDr-CXR, and CheXpert require access approval from PhysioNet.

3.2 Zero-Shot Classification Performance

Table 2 compares the zero-shot classification performance of the state-of-the-art CARZero baseline with and without OFF-CLIP across four test sets, using AUC as the evaluation metric. OFF-CLIP achieves a 177% relative gain in average normal AUC over CARZero, improving it by 0.50. Additionally, AUC scores for multiple chest disease categories (Atelectasis, Cardiomegaly, Emphysema, Effusion, Nodule, Pleural Thickening) show notable gains, leading to a total AUC improvement of 0.095. These results confirm that OFF-CLIP effectively reduces FNs caused by strict loss constraints and FPs from normal sentences in abnormal reports.

3.3 Ablation Study on Zero-shot classification

We conduct an ablation study on Open-I with CARZero to assess the contributions of OFF-CLIP's components: text filtering, the off-diagonal loss ($\mathcal{L}_{off}$), and the abnormal pair InfoNCE loss ($\mathcal{L}_{ab}$).

Table 3 shows that $\mathcal{L}_{off}$ significantly enhances normal detection, maintaining AUC above 0.7 compared to 0.3 in the baseline. Text filtering further mitigates FN errors, particularly when combined with $\mathcal{L}_{off}$ and $\mathcal{L}_{ab}$. Table 4 presents FN and FP ratios and rates. Without OFF-CLIP, the model is heavily biased toward abnormal cases, with FP rates nearing 1. Introducing $\mathcal{L}_{off}$ balances FP and FN misclassifications, reducing FP rates from 0.99 to 0.5. Text filtering further lowers FN ratios ($\mathcal{L}_{off} + \mathcal{L}_{ab}$: 0.17 $\rightarrow$ 0.15) while maintaining a balanced FP-FN distribution. These findings confirm that conventional contrastive learning struggles with normal clustering and misalignment, while OFF-CLIP effectively mitigates these issues, improving zero-shot classification reliability.

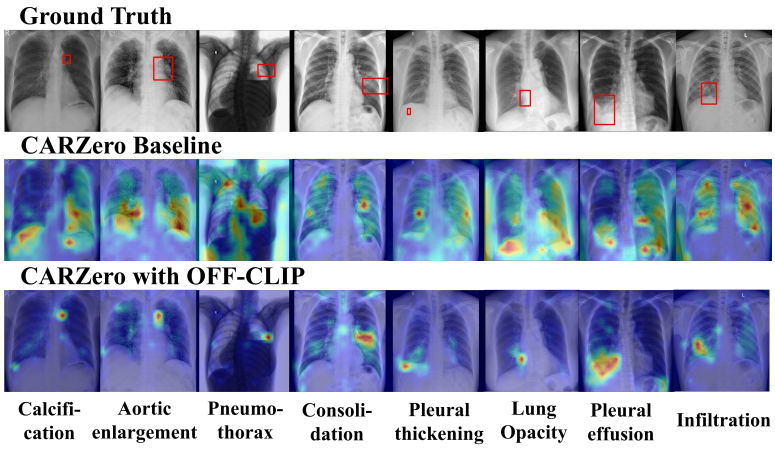

Fig. 3. Visualization of attention maps on VinDr-CXR. Red boxes indicate ground truth bounding boxes for each diseases. Highlighted pixels represent regions with higher activation weights, linking specific words to image areas.

3.4 Zero-shot Grounding Performance

We evaluate anomaly localization using the pointing game [22], selecting the top 10% or 20% high-attention regions instead of only the highest one. A sample is considered successful if any selected region overlaps with the ground truth bounding box, and the average success rate is computed per disease.

Table 5 demonstrates that OFF-CLIP significantly improves attention alignment over the CARZero baseline. With 10% selection, CALC improves from 0.30 to 0.59 (+97%), while with 20%, PTX increases from 0.63 to 1.00 (+58.73%). Figure 3 further reveals that while CARZero's attention is diffuse, OFF-CLIP's is sharply focused on ground truth regions. These results confirm that OFF-CLIP enhances zero-shot grounding and anomaly localization.

4 Conclusion

OFF-CLIP addresses two key limitations in radiology CLIP models: excessive FPs from poor normal sample clustering and high FNs due to misleading normal text in abnormal reports. Through extensive ablation studies, we quantified their impact using FP and FN analyses.

By incorporating an off-diagonal similarity loss ($\mathcal{L}_{off}$) and a text filtering strategy, OFF-CLIP significantly improves normal detection (0.50 AUC, 177% average increase) and enhances anomaly localization (up to 97% improvement in attention alignment). While OFF-CLIP is evaluated on CARZero, its framework-agnostic design suggests potential applicability to other models. Further validation is required to assess its effectiveness across different architectures and its clinical utility in real-world settings.

Acknowledgments. This work was supported by Korea Medical Device Development Fund grant funded by the Korea government (1711196477, RS-2023-00252244), the DGIST R&D Program of the Ministry of Science and ICT (25-IRJoint-09), the Industrial Strategic Technology Development Program (ISTDP) (RS-2024-00443054) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea), and by the collaborative project with ROEN Surgical Inc. This work was supported by a grant from the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korean government (MSIT) (RS2021-II211343), Artificial Intelligence Graduate School Program, Seoul National University.

Junhyun Park and Minho Hwang are affiliated with the Department of Robotics and Mechatronics Engineering at DGIST. Chanyun Moon is affiliated with the Division of Artificial Intelligence, Department of Interdisciplinary Studies at DGIST. Kyungsu Kim is affiliated with the School of Transdisciplinary Innovations, Department of Biomedical Science, Medical Research Center, Interdisciplinary Program in Bioengineering, and Interdisciplinary Program in Artificial Intelligence at Seoul National University. Donghwan Lee is affiliated with the Department of Biomedical Science at Seoul National University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
2. Bustos, A., Pertusa, A., Salinas, J.M., Iglesia-Vaya., M.D.L.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis **66** (2020)
3. Chan, H.P., Hadjiiski, L.M., Samala, R.K.: Computer-aided diagnosis in the era of deep learning. Medical physics **47**(5), e218–e227 (2020)
4. Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., Chang, T.H.: Multi-modal masked autoencoders for medical vision-and-language pretraining. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 679–689. Springer (2022)
5. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
7. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)

9. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: IEEE/CVF International Conference on Computer Vision. pp. 3922–3931 (2021)
10. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI conference on artificial intelligence. pp. 590–597 (2019)
11. Jamshidi, M., Lalbakhsh, A., Talla, J., Peroutka, Z., Hadjilooei, F., Lalbakhsh, P., Jamshidi, M., La Spada, L., Mirmozafari, M., Dehghani, M., et al.: Artificial intelligence and covid-19: deep learning approaches for diagnosis and treatment. IEEE Access **8**, 109581–109595 (2020)
12. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., et al.: Mimic-cxr, a deidentified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(317) (2019)
13. Kim, K., Park, J., Langarica, S., Mahmoud Alkhadrawi, A., Do, S.: Integrating chatgpt into secure hospital networks: A case study on improving radiology report analysis. In: Proceedings of the fifth Conference on Health, Inference, and Learning. pp. 72–87. PMLR (2024)
14. Lai, H., Yao, Q., Jiang, Z., Wang, R., He, Z., Tao, X., Zhou, S.K.: Carzero: Cross-attention alignment for radiology zero-shot classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11137–11146 (2024)
15. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
16. Nguyen, H., Lam, K., Le, L., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(429) (2022)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
18. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Y.Ng, A., Rajpurkar, P.: Expert-level detection of pathologies from unaanotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6** (2022)
19. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. vol. 2022, pp. 3876–3887 (12 2022)
20. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training. In: IEEE/CVF International Conference on Computer Vision. pp. 21372–21383 (2023)
21. Yan, A., McAuley, J., Lu, X., Du, J., Chang, E.Y., Gentili, A., Hsu, C.N.: RadBERT: Adapting transformer-based language models to radiology. Radiology: Artificial Intelligence **4**(4), e210258 (2022)
22. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., , Sclaroff, S.: Top-down neural attention by excitation backprop. International Journal of Computer Vision **126**, 1084–1102 (2018)
23. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14** (2023)
24. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. vol. 182, pp. 2–25. PMLR (2022)