Medical Contrastive Learning of Positive and Negative Mentions

WeiLong Wu[1], Jingzhi Yang[4], Xun Zhu[1], Xiao Zhang[1], ZiYu Liu[1], Miao Li[1]†, and Ji Wu[1,2,3]†

[1] Department of Electronic Engineering, Tsinghua University, Beijing 10084, China
[2] College of AI, Tsinghua University, Beijing 100084, China
[3] Beijing National Research Center for Information Science and Technology, Beijing 10084, China
[4] School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
{wu-wl23, zhu-x24, xzhang19, ziyu-liu22, miao-li, wuji_ee}@mails.tsinghua.edu.cn
yangjingzhi@bupt.edu.cn

Abstract. Contrastive learning techniques have achieved significant success and have been widely applied in both general and medical domains. However, there is a data difference between the general domain and the medical domain about negative mentions, which almost never appear in general domain but almost always in medical domain. We find that most existing medical contrastive learning methods do not effectively utilize or even overlook the numerous negative mentions present in the data during training, resulting in deficient multimodal feature alignment capabilities. To address this issue, we propose the **V**isual **E**ntailment Based **C**ontrastive **L**earning (VECL) method. By introducing a ternary visual entailment contrast relationship of entailment, neutral, and contradiction, our method effectively utilizes both positive and negative mentions for modeling fine-grained sample relationships, enhancing the model's multimodal feature alignment capabilities. The experiment results show that we achieves SOTA performance on classification, grounding and report generation tasks. Resources are maintained at https://github.com/WVeLong/VECL.

Keywords: Contrastive Learning · Multimodality · Medical Pretraining

1 Introduction

Attributed to the rapid development of contrastive learning, many medical vision-language model are proposed, such as MedCLIP [1], GLoRIA [2], CheXzero [3], MedKLIP [4], KAD [5], CARZero [6]. However, existing methods typically do not effectively utilize or even overlook the data difference between the general domain and the medical domain: data in the general domain usually contains only

† Corresponding author

positive mentions (confirming the presence of target or disease), whereas data in the medical domain includes both positive and negative mentions (confirming the absence of target or disease). For example, we conduct a rough statistical analysis of the MIMIC-CXR [7] dataset, and find that samples containing negative mentions account for approximately 49%. Because existing methods typically do not pay attention to negative mentions, they fail to learn better multimodal representations, which consequently limits their performance on downstream tasks.

For the zero-shot classification evaluation of visual language pre-trained Models, in general domain, images are only computed for similarities with texts that contain positive mentions, and we refer to this as the Positive-Only Similarity (POS) evaluation method. But in medical domain, it is necessary to compute the similarities between images and texts that contain positive or negative mentions, and we refer to this as the Positive-Negative Similarity (PNC) evaluation method, which has been proposed in CheXzero [3] but is not widely adopted in the community. In our experiments, most previous methods show a certain degree of drop in metrics when evaluated using the PNC evaluation method compared to the POS evaluation method, which supports the hypothesis above.

To address this issue, we propose the **V**isual **E**ntailment Based **C**ontrastive **L**earning (VECL) method. We firstly extract positive and negative mention labels from radiology reports and then introduce ternary visual entailment contrast relationships, utilizing both labels to model fine-grained sample relationships. Meanwhile, we modify the InfoNCE loss function to train the model more effectively. The experimental results show that our model learns better multimodal feature alignment capabilities and achieves SOTA performance on various downstream tasks, such as classification, grounding and report generation.

2 Method

In this section, we will illustrate the details of the **V**isual **E**ntailment Based **C**ontrastive **L**earning (VECL) method. Its framework is shown in Figure 1.

2.1 Extract Positive and Negative Mention Labels from Report

We can choose an LLM or other specialized label extraction tool to extract labels from report sentences. During label extraction, we need to choose p categories of diseases to form the set of label categories C, along with the label 0 representing "other diseases / without abnormalities". Although "other diseases" and "without abnormalities" have different meanings, they serve the same effect in subsequent processing, so they are both represented by the label 0.

$$C = \{0, 1^+, 1^-, 2^+, 2^-, \ldots, p^+, p^-\} \tag{1}$$

Assuming that the sentence i is related to the disease category r, $1 \leq r \leq p$, the extracted label is denoted as C_i, $C_i \in C$. If the sentence i contains positive / negative mention of the disease category r, then $C_i = r^+/r^-$; If the sentence

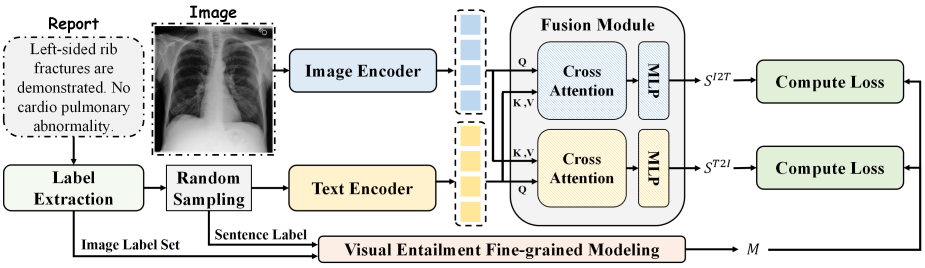

Fig. 1. The framework of VECL method. The image enters the image encoder directly, while the report is first segmented into sentences and extracted labels, and then one of the sentences sampled from the report enters the text encoder. After encoding, the image and text features are fused to obtain cross-modal similarity S^{I2T} and S^{T2I} in both way. Meanwhile, we introduce visual entailment and utilize the positive and negative mentions in the labels to finely model the sample relationships, obtaining the training labels M. Finally, S^{I2T}, S^{T2I} and M are used to compute loss.

contains other diseases or without abnormalities, then $C_i = 0$. If a report sentence contains multiple diseases, a list of labels will be generated. Since one image corresponds to a complete report, the set of labels for all sentences in a report is regarded as the image label set.

2.2 Encode and Fuse Feature

Assuming that x_i represents the i-th image in training set and y_j represents the j-th report in training set, Φ_I, Φ_T and Φ_F represent the image encoder, the text encoder and the fusion module, respectively. So the intermediate results and the image-text similarities are as follows:

$$x_i^I = \Phi_I(x_i),\ y_j^T = \Phi_T(y_j) \tag{2}$$

$$S_{ij}^{I2T} = \Phi_F(x_i^I, y_j^T) = MLP(CrossAtten(x_i^I, y_j^T, y_j^T)) \tag{3}$$

$$S_{ji}^{T2I} = \Phi_F(y_j^T, x_i^I) = MLP(CrossAtten(y_j^T, x_i^I, x_i^I)) \tag{4}$$

x_i^I and y_j^T represent the image features and the text features, respectively, while S_{ij}^{I2T} and S_{ji}^{T2I} represent the image-text similarities in both way, with each serving as the query in the cross attention process, respectively. For any image x_i and any report y_j within a batch, the image-text similarities among them ultimately form two similarity matrices S^{I2T} and S^{T2I}, where $S^{I2T}, S^{T2I} \in \mathbb{R}^{N\times N\times 3}$. Here N is the batch size.

2.3 Modeling Fine-grained Contrast Relationships

As shown in Figure 2, to finely model the contrast relationships among samples using positive and negative mentions labels from report, we introduce ternary

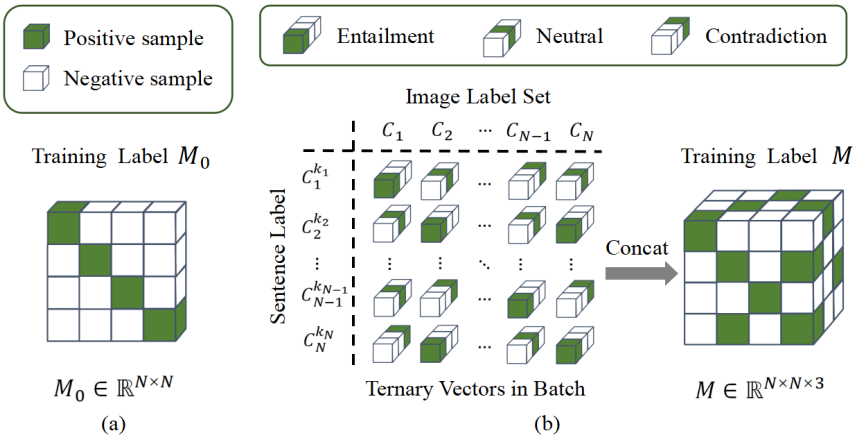

Fig. 2. Comparison of contrast relationship modeling. (a) Traditional binary contrast relationships for the InfoNCE [8] loss, which include positive and negative samples represented by number 1 and 0. (b) Visual entailment contrast relationships, which include entailment, neutral and contradiction, are represented by three basis vectors.

visual entailment contrast relationships: entailment, neutral and contradiction, which are represented by three basis vectors:$[1, 0, 0]$, $[0, 1, 0]$, and $[0, 0, 1]$, respectively. Specifically, we assess the relationship between the image label set and the sentence label within a batch:

- If the image label set and sentence label contain the same disease category r and are the same / opposite symbol, we consider the visual entailment contrast relationship to be entailment / contradiction.
- If the image label set and sentence label contains different disease category r, we consider the visual entailment contrast relationship to be neutral.
- Specifically, when the sentence label is 0, we consider the relationship is entailment with itself and neutral with others.

After obtaining the visual entailment contrast relationships within a batch, we get a group vectors whose number is $N \times N$, then we concatenate them and finally obtain the training label M, where $M \in \mathbb{R}^{N\times N\times 3}$.

2.4 Compute Loss

The entire loss function consists of two parts: the original InfoNCE [8] loss and the proposed 3D-InfoNCE loss. For the InfoNCE loss, it's label M_0 remains with the diagonal being 1 and the off-diagonal being 0.

The complete predictions and labels of the entire loss function are S^{I2T}, S^{T2I}, and M, M_0. Let d represent the position along the third dimension of S^{I2T}, S^{T2I} and M, $d \in \{0, 1, 2\}$. The InfoNCE loss computes the loss between $S^{I2T}(d = 0)$, $S^{T2I}(d = 0)$ and M_0:

$$\mathcal{L}_{Info} = \text{InfoNCE}(S^{I2T}(d=0), M_0) + \text{InfoNCE}(S^{T2I}(d=0), M_0) \tag{5}$$

And the 3D-InfoNCE loss computes the loss between S^{I2T}, S^{T2I} and M:

$$\mathcal{L}^{I2T}(d) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\text{Norm}(M_{ij}(d)) \cdot \log(\text{Softmax}(e^{S_{ij}^{I2T}(d)})) \tag{6}$$

$$\mathcal{L}_{3DInfo} = \sum_{d=0}^{2}(\mathcal{L}^{I2T}(d) + \mathcal{L}^{T2I}(d)) \tag{7}$$

The total loss is the sum of both:

$$\mathcal{L} = \mathcal{L}_{Info} + \mathcal{L}_{3DInfo} \tag{8}$$

Here, Norm refers to the normalization:

$$\text{Norm}(M_{ij}(d)) = \begin{cases} \frac{M_{ij}(d)}{\sum_{i=1}^{N} M_{ij}(d)} & \text{if } \sum_{i=1}^{N} M_{ij}(d) \neq 0, \\ 0 & \text{if } \sum_{i=1}^{N} M_{ij}(d) = 0 \end{cases} \tag{9}$$

3 Experiment and Result

3.1 Dataset, Evaluation Metric and Implementation Details

For training, we only use MIMIC-CXR [7]. For downstream task evaluation, we use Open-I [9], CheXpert [10], ChestXray14 [11], ChestXDet10 [12], PadChest [13] for zero-shot classification, ChestXray14 for fine-tuning classification, ChestXDet10 for zero-shot grounding, and MIMIC-CXR for retrieval based report generation.

For classification task, we adopt AUC, F1, MCC, and mAP. For grounding task, we adopt Pointing Game [14]. For generation task, we adopt NLG metrics: ROUGE [15], BLUE [16], CIDEr [17] and CE metrics: Precision, Recall and F1.

We use Meta-Llama-3-8B-Instruct [18] for label extraction, and the number of diseases categories p is 24, which exactly covers all the categories in MIMIC-CXR, Open-I, CheXpert, ChestXray14, and ChestXDet10, but does not including any categories in PadChest. we choose ViT-B/16 [19] as the image encoder which utilizes M3AE [20] for pretraining on the MIMIC-CXR, and choose BioBERT [21] as the text encoder which is fine-tuned on MIMIC-CXR and PadChset. The Adam [22] optimizer is utilized with a learning rate of 5e-5, and the batch size is 256. All experiments are conducted with an 80G A800 GPU.

3.2 Comparison with Different Methods

Zero-Shot Classification As shown in Table 1, in both the POS and PNC evaluation methods, our model achieves SOTA across all metrics on all datasets, even in PadChest, where the labels is not covered in the set of label categories C. It is worth noting that CARZero [6] who did not care about negative mentions during training, has a significant drop in the PNC. And models such as MedCLIP [1] and MedKLIP [4], although they indirectly consider negative mentions in different ways during training, still have different degrees of drop in the PNC, due to failing to effectively model the complex contrast relationships among images-sentence pairs.

Method	Eval Method	Open-I		CheXpert		ChestXray14		ChestXDet10		PadChest	
		AUC↑	F1↑	AUC↑	F1↑	AUC↑	F1↑	AUC↑	F1↑	AUC↑	F1↑
MedCLIP	POS	0.500	0.134	0.528	0.389	0.510	0.146	0.517	0.322	0.477	0.033
	PNC	0.756	0.184	0.819	0.531	0.704	0.180	0.647	0.347	0.700	0.051
GLoRIA	POS	0.588	0.086	0.585	0.331	0.620	0.131	0.602	0.308	0.563	0.027
	PNC	0.524	0.074	0.540	0.307	0.539	0.115	0.560	0.297	0.532	0.022
MedKLIP	POS	0.565	0.111	0.717	0.416	0.623	0.135	0.571	0.297	0.612	0.032
	PNC	0.470	0.074	0.496	0.349	0.498	0.105	0.529	0.303	0.500	0.024
KAD	POS	0.818	0.283	0.849	0.549	0.796	0.289	0.749	0.449	0.748	0.087
	PNC	0.695	0.169	0.786	0.514	0.695	0.168	0.675	0.383	0.568	0.043
CARZero	POS	0.839	0.242	0.909	0.370	0.803	0.242	0.795	0.308	0.804	0.114
	PNC	0.352	0.109	0.144	0.244	0.347	0.133	0.422	0.333	0.419	0.040
VECL(Ours)	POS	**0.839**	**0.341**	**0.915**	**0.664**	**0.816**	**0.309**	**0.811**	**0.498**	**0.822**	**0.147**
	PNC	**0.813**	**0.333**	**0.922**	**0.689**	**0.792**	**0.291**	**0.780**	**0.470**	**0.775**	**0.116**

Table 1. Comparison of different methods on Open-I, CheXpert, ChestXray14, ChestXDet10, PadChest for zero-shot classification.

Fine-Tuning Classification As shown in Table 2, on 1%, 5%, 10% ChestXray14 fine-tuning data, our model continues to achieve SOTA on all metrics.

[Fine-Tuning Classification]

Method	1%				5%				10%			
	AUC↑	F1↑	MCC↑	mAP↑	AUC↑	F1↑	MCC↑	mAP↑	AUC↑	F1↑	MCC↑	mAP↑
KAD	0.750	0.117	0.208	0.163	0.712	0.096	0.169	0.129	0.666	0.096	0.121	0.096
CARZero	0.813	0.153	0.263	0.215	0.835	0.189	0.303	0.264	0.839	0.196	0.314	0.277
VECL (Ours)	**0.826**	**0.338**	**0.310**	**0.264**	**0.842**	**0.352**	**0.324**	**0.286**	**0.845**	**0.361**	**0.334**	**0.295**

[Zero-Shot Grounding and Retrieval Based Report Generation]

Method	ChestXDet10	MIMIC-CXR						
	Point Game↑	RG-L↑	BL-1↑	BL-2↑	CIDEr↑	Pr↑	Re↑	F1↑
KAD	0.391	0.115	0.189	0.087	0.019	0.504	0.137	0.197
CARZero	0.543	0.128	0.223	0.105	0.028	0.496	0.160	0.218
VECL (Ours)	**0.683**	**0.128**	**0.223**	**0.105**	**0.029**	**0.574**	**0.204**	**0.273**

Table 2. Combined Tables: Comparison of different methods on 1%, 5%, 10% ChestXray14 data for fine-tuning classification, on ChestXDet10 for zero-shot grounding, and on MIMIC-CXR for retrieval based report generation.

Zero-Shot Grounding As shown in Table 2, on ChestXDet10, our model not only achieved SOTA but also showed a very significant improvement over the baseline.

Retrieval Based Report Generation As shown in Table 2, on MIMIC-CXR, our model achieves SOTA on all metrics.

In the three downstream tasks in the Table 2, the outstanding performance of our method still demonstrates its ability to effectively align the multimodal representations, even without using the PNC evaluation method. By considering both positive and negative mentions during training, our model is capable of accurately matching the normal/lesion areas in the images with the prompt.

[Ablation Study of Visual Entailment and Loss Function]

Visual Entail		Loss Function				Eval Method	CheXpert				PadChest			
False	True	BCE	CE	InfoNCE	Ours		AUC↑	F1↑	MCC↑	mAP↑	AUC↑	F1↑	MCC↑	mAP↑
✓		×	×	✓	×	POS	0.899	0.331	0.574	0.639	0.798	0.078	0.126	0.067
						PNC	0.463	0.219	0.208	0.255	0.445	0.028	0.050	0.018
	✓	✓		×		POS	0.896	0.638	0.568	0.639	0.739	0.072	0.089	0.065
						PNC	0.906	0.652	0.586	0.660	0.745	0.067	0.086	0.037
	✓		✓	×		POS	0.882	0.614	0.544	0.614	0.739	0.070	0.088	0.040
						PNC	0.893	0.623	0.552	0.639	0.749	0.069	0.088	0.038
	✓			×	✓	POS	**0.915**	**0.664**	**0.615**	**0.677**	**0.822**	**0.147**	**0.165**	**0.097**
						PNC	**0.922**	**0.689**	**0.631**	**0.690**	**0.775**	**0.116**	**0.129**	**0.074**

[Analysis of the Model's Robustness to Extracted Labels]

Method	Eval Method	Open-I		CheXpert		ChestXray14		ChestXDet10		PadChest	
		AUC↑	F1↑	AUC↑	F1↑	AUC↑	F1↑	AUC↑	F1↑	AUC↑	F1↑
VECL(labeler)	POS	0.838	0.336	0.915	**0.665**	0.811	0.305	0.798	0.485	0.813	**0.150**
	PNC	**0.833**	0.307	0.919	0.687	**0.811**	0.288	0.763	0.453	**0.777**	0.082
VECL(LLM)	POS	**0.839**	**0.341**	**0.915**	0.664	**0.816**	**0.309**	**0.811**	**0.498**	**0.822**	0.147
	PNC	0.813	**0.333**	**0.922**	**0.689**	0.792	**0.291**	**0.780**	**0.470**	0.775	**0.116**

Table 3. Combined Tables: Ablation study of visual entailment and loss function, and comparison on zero-shot classification of VECL models trained from labels extracted by Chexpert-Labeler and LLM. Visual entailment being set to False indicates that only the first dimension of training label M is used, while being set to True indicates that the entire label M is used. Note that × indicates combinations that cannot be achieved due to dimension mismatches.

3.3 Ablation Study

Ablation Study of Visual Entailment As shown in Table 3, we compare with baseline model for the zero-shot classification performance on CheXpert and PadChest. In baseline, we only use the first dimension of the training label M for optimization. The baseline showed varying degrees of drop in both POS and PNC, with a particularly significant drop in PNC. In fact, the baseline model can still model the entailment and neutral from positive and negative samples, but it fails to model the contradictory from negative mentions, which indicates that modeling the contradictory of negative mentions is quite important and visual entailment method is proved to be an effective way to model them.

Ablation Study of Loss Function As shown in Table 3, we compare with different baseline models on the same tasks. In baseline, there use Binary Cross-Entropy Loss (BCE) loss and Cross-Entropy (CE) loss. Compared with them, our model has better performance, which indicates that our loss function is more suitable for optimizing features that contain complex visual entailment contrast relationships.

Analysis of the Model's Robustness to Extracted Labels We use label extraction tool Chexpert-Labeler [10] to extract new labels, which contain 13 types of disease categories and "No Finding" as 14th category, and retrained the model. Meanwhile, we use new labels as the reference to evaluate the accuracy of the labels extracted by LLM. The average precision, recall, and F1 score across these 14 categories are 0.552, 0.608, and 0.585, respectively. As shown in Table 3, although there is some noise in the labels of LLMs, the VECL models trained on both types of labels show comparable performance, both outperforming other baseline models and achieving SOTA. This indicates that our method has robustness to resist label noise and can effectively learn meaningful visual entailment contrast relationships among them.

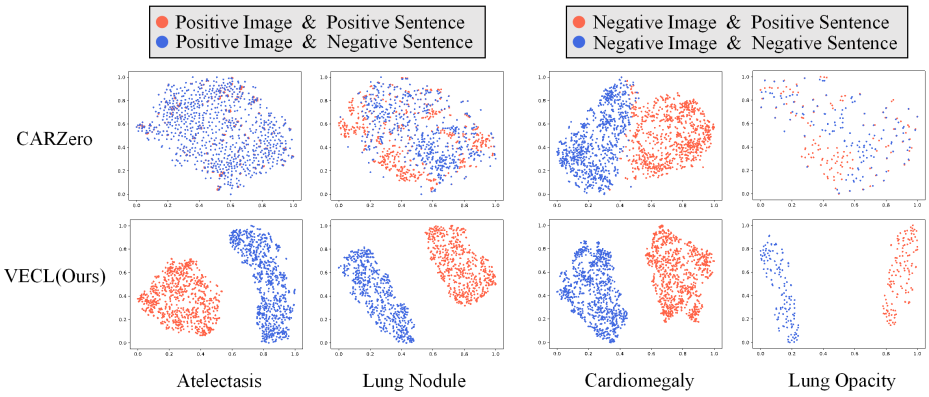

Fig. 3. t-SNE visualization on the similarities of different image-sentence pairs between CARZero and VECL. Visualization results demonstrate that our method can learn better multimodal aligned representations. Here, "positive" indicates the presence of the specific disease, while "negative" indicates the absence of the specific disease.

3.4 Visualization

We perform t-SNE [23] visualization on the similarities of different image-sentence pairs, selecting CARZero as comparison. As shown in Figure 3, in each specific disease category, our method can clearly distinguish the distributions of opposite features, while CARZero is much poorer. In the feature space, our model can bring positive images closer to positive sentences and farther from negative

sentences, while simultaneously bringing negative images closer to negative sentences and farther from positive sentences, which directly demonstrates that our method has learned better multimodal aligned representations.

4 Conclusion

In this paper, we find that existing methods typically overlook the data difference between the general domain and the medical domain and do not effectively utilize the negative mentions, consequently limits multimodal feature alignment capability and downstream tasks performance. To address this issue, we propose the **V**isual **E**ntailment Based **C**ontrastive **L**earning (VECL) method. The experiment results show that our model has better multimodal feature alignment capability and achieves SOTA performance on classification, grounding and report generation tasks.

Additionally, since the label extraction process relies only on reports, our method exhibits high generality and transferability. In the future, we will validate our method across various types of medical data, such as CT, MRI, and so on.

Acknowledgments. Supported by Beijing Natural Science Foundation NO. 4252046

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

References

1. Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163, 2022.
2. Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3942–3951, 2021.
3. Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering, 6(12): 1399–1406, 2022.
4. Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 21372–21383, 2023.
5. Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications, 14(1):4542, 2023.
6. Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, and S Kevin Zhou. Carzero: Cross-attention alignment for radiology zero-shot classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11137–11146, 2024.

7. 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 StevenHorng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042, 2019.
8. Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.
9. Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2):304–310, 2016.
10. Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence, pages 590–597, 2019.
11. Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestxray8: 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, pages 2097–2106, 2017.
12. Jingyu Liu, Jie Lian, and Yizhou Yu. Chestx-det10: chest x-ray dataset on detection of thoracic abnormalities. arXiv preprint arXiv:2006.10550, 2020.
13. Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis, 66:101797, 2020.
14. Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. International Journal of Computer Vision, 126(10):1084–1102, 2018.
15. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81, 2004.
16. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318, 2002.
17. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4566–4575, 2015.
18. AI@Meta. Llama 3 model card. 2024.
19. Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
20. Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pretraining. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 679–689. Springer, 2022.
21. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240, 2020.
22. Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
23. Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11).