NEURAL INFORMATION PROCESSING SYSTEMS

Multi-Granularity Cross-modal Alignment for Generalized Medical Visual Representation Learning

Fuying Wang[1], Yuyin Zhou[2], Shujun Wang[3], Varut Vardhanabhuti[1], Lequan Yu[1]

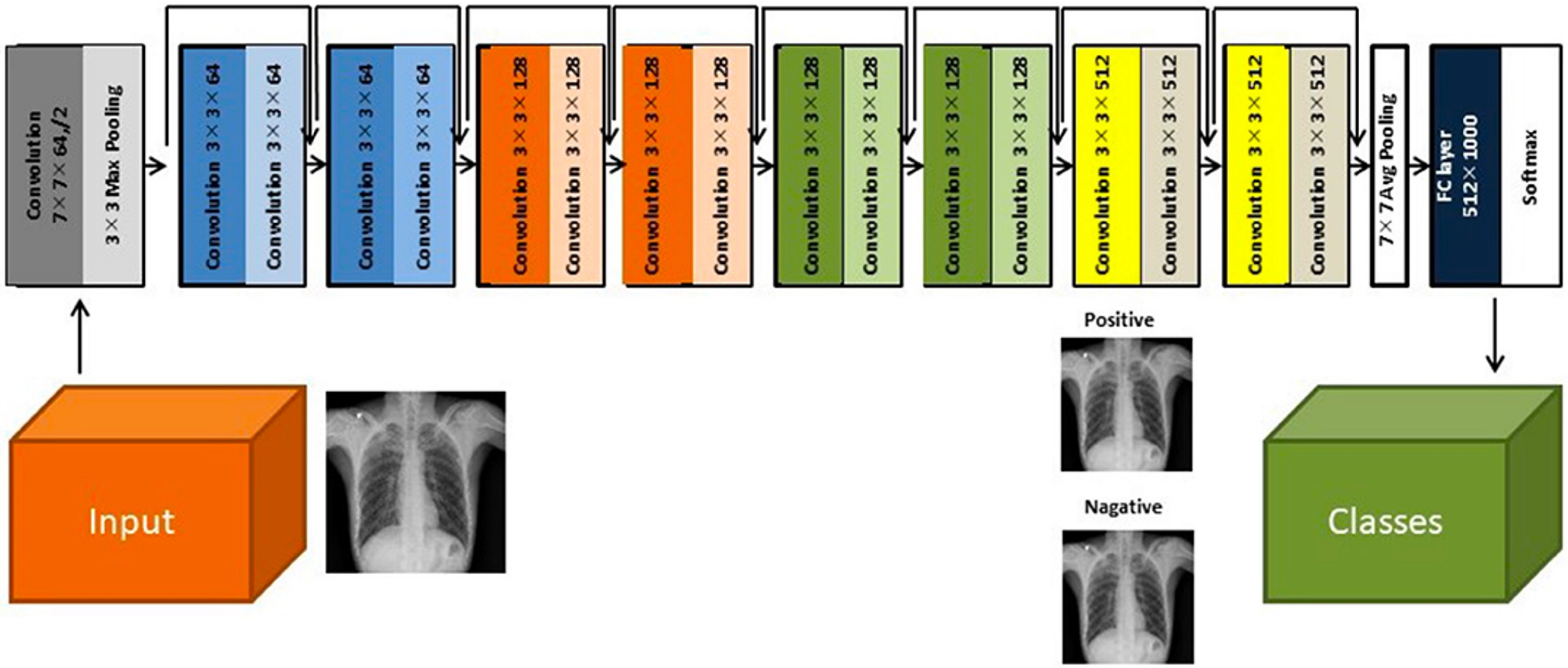
[1]The University of Hong Kong [2]University of California, Santa Cruz

[3]University of Cambridge

香 港 大 學
THE UNIVERSITY OF HONG KONG

UNIVERSITY OF CAMBRIDGE

Bottleneck of supervised chest X-ray classification



Deep learning classifier for COVID-19 diagnosis

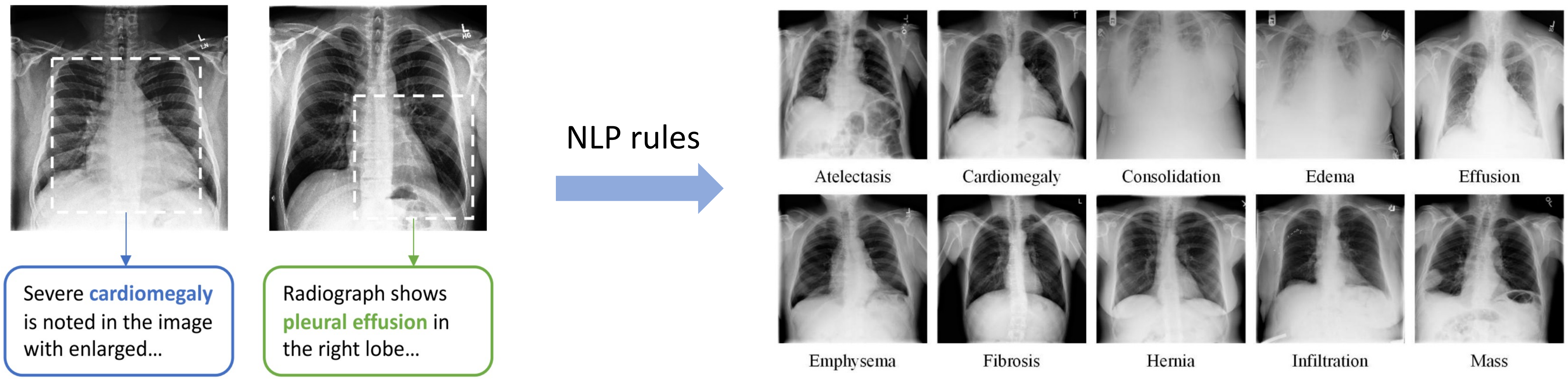
- Diverse and large number of labeled training data are needed
- Collecting such large dataset requires intensive human labor and time

Yoo, S. H., Geng, H., Chiu, T. L., Yu, S. K., Cho, D. C., Heo, J., ... & Lee, H. (2020). Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging. *Frontiers in medicine*, *7*, 427.

Learn medical visual representations from image-text pairs

There are two common approaches to exploit supervision from reports:

- Extract labels from reports via NLP rules



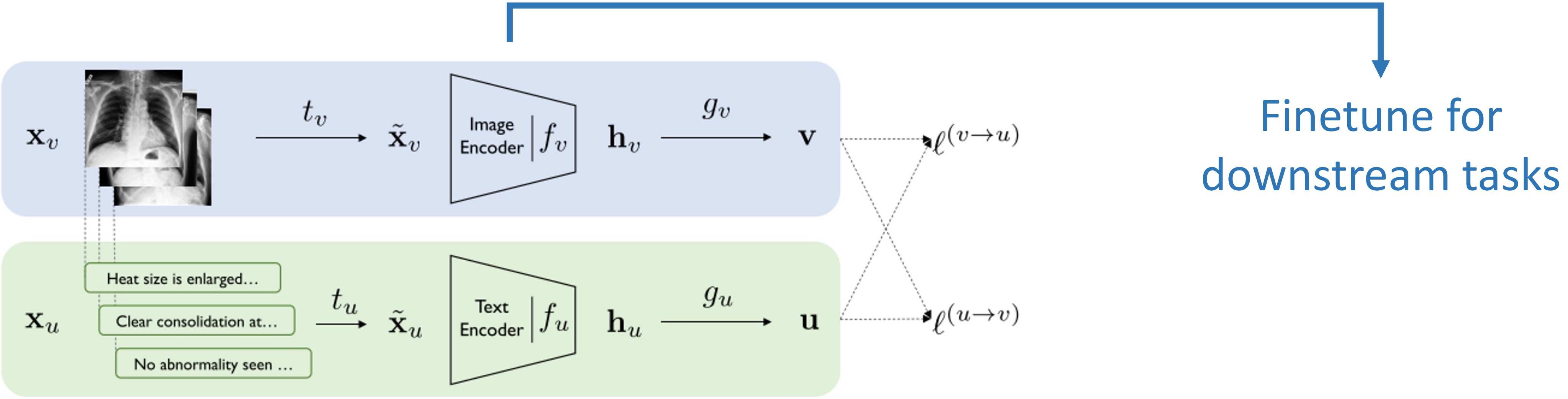
Extract labels from medical reports following NLP rules

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., ... & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, *6*(1), 1-8.

Learn medical visual representations from image-text pairs

There are two common approaches to exploit supervision from reports:

- Extract labels from reports via NLP rules
- Image-text joint representation learning

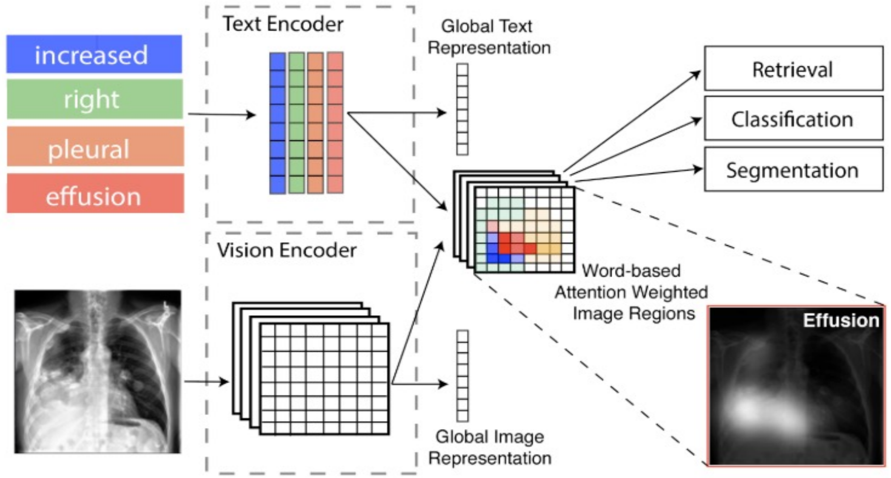


Contrastive learning for image-text joint representation learning

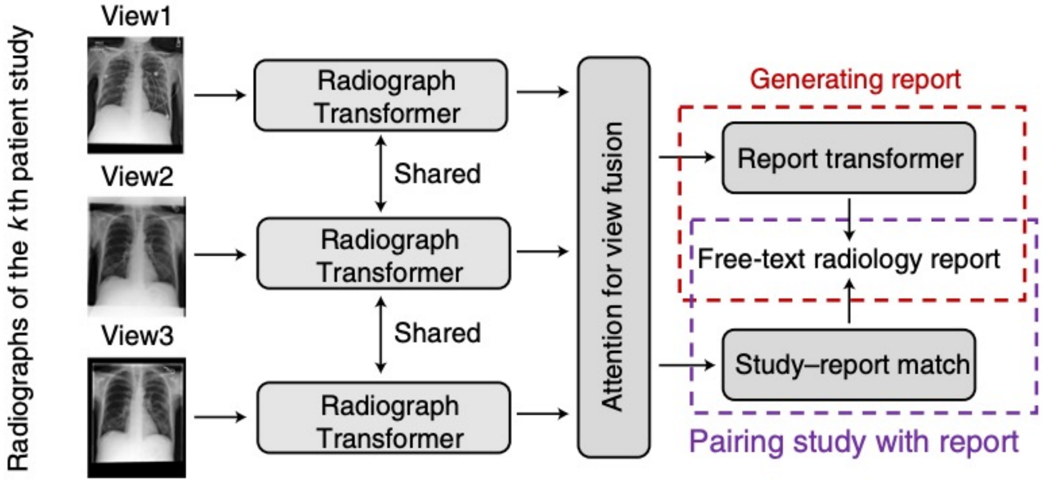
Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., & Langlotz, C. P. (2020). Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747.

Learn medical visual representations from image-text pairs

- Image-text joint representation learning has achieved great success



GLoRIA [1]



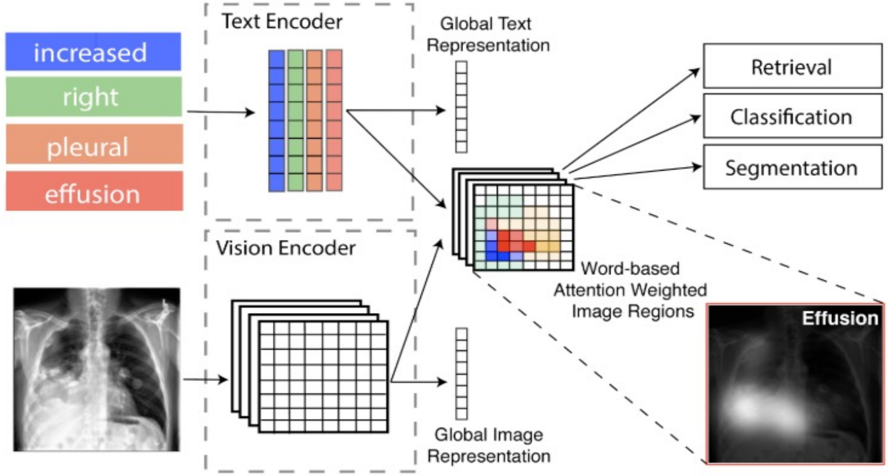
REFERS[2]

[1]. Huang, S. C., Shen, L., Lungren, M. P., & Yeung, S. (2021). 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 (pp. 3942-3951).

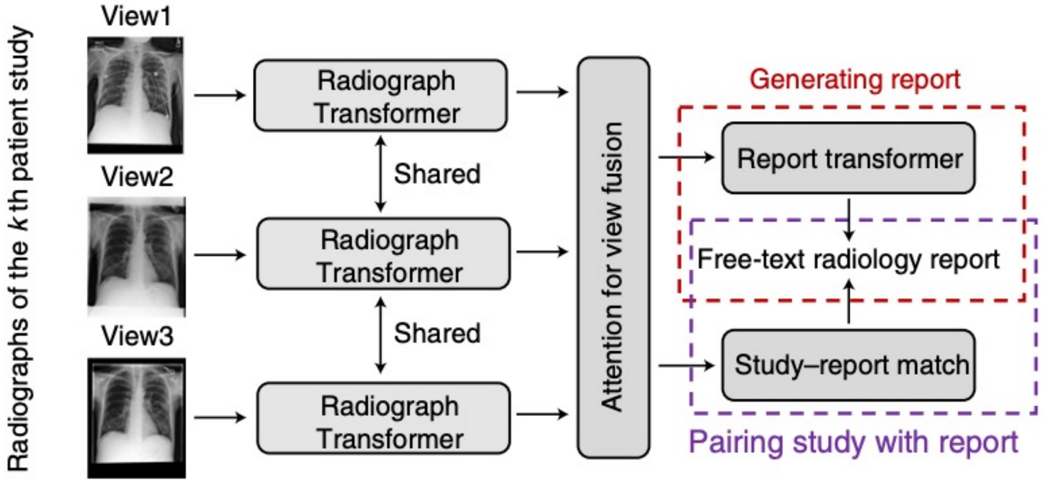
[2]. Zhou, H. Y., Chen, X., Zhang, Y., Luo, R., Wang, L., & Yu, Y. (2022). Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, *4*(1), 32-40.

Learn medical visual representations from image-text pairs

- Image-text joint representation learning has achieved great success



GLoRIA [1]

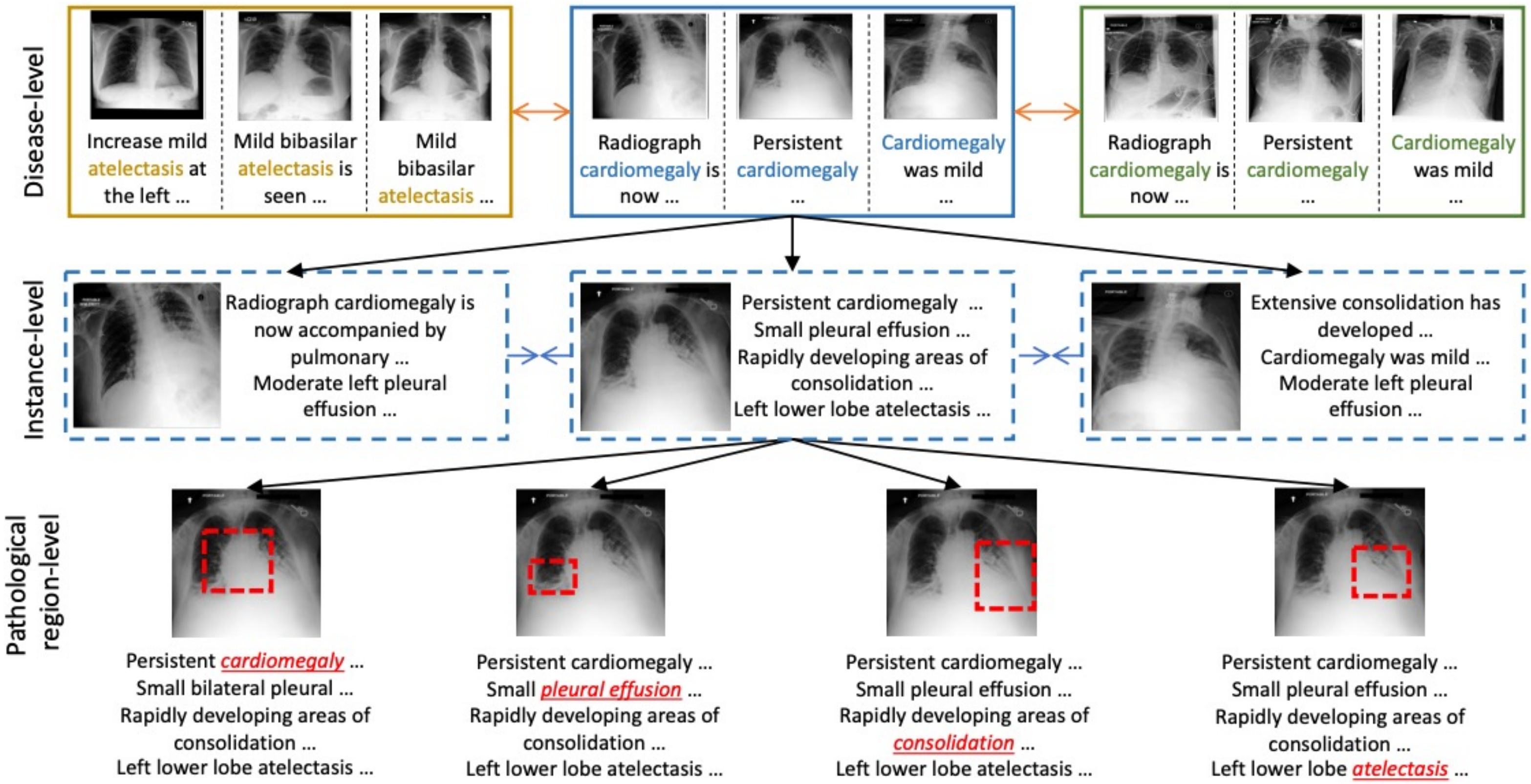


REFERS[2]

- Existing methods utilize insufficient supervision from image-text pairs

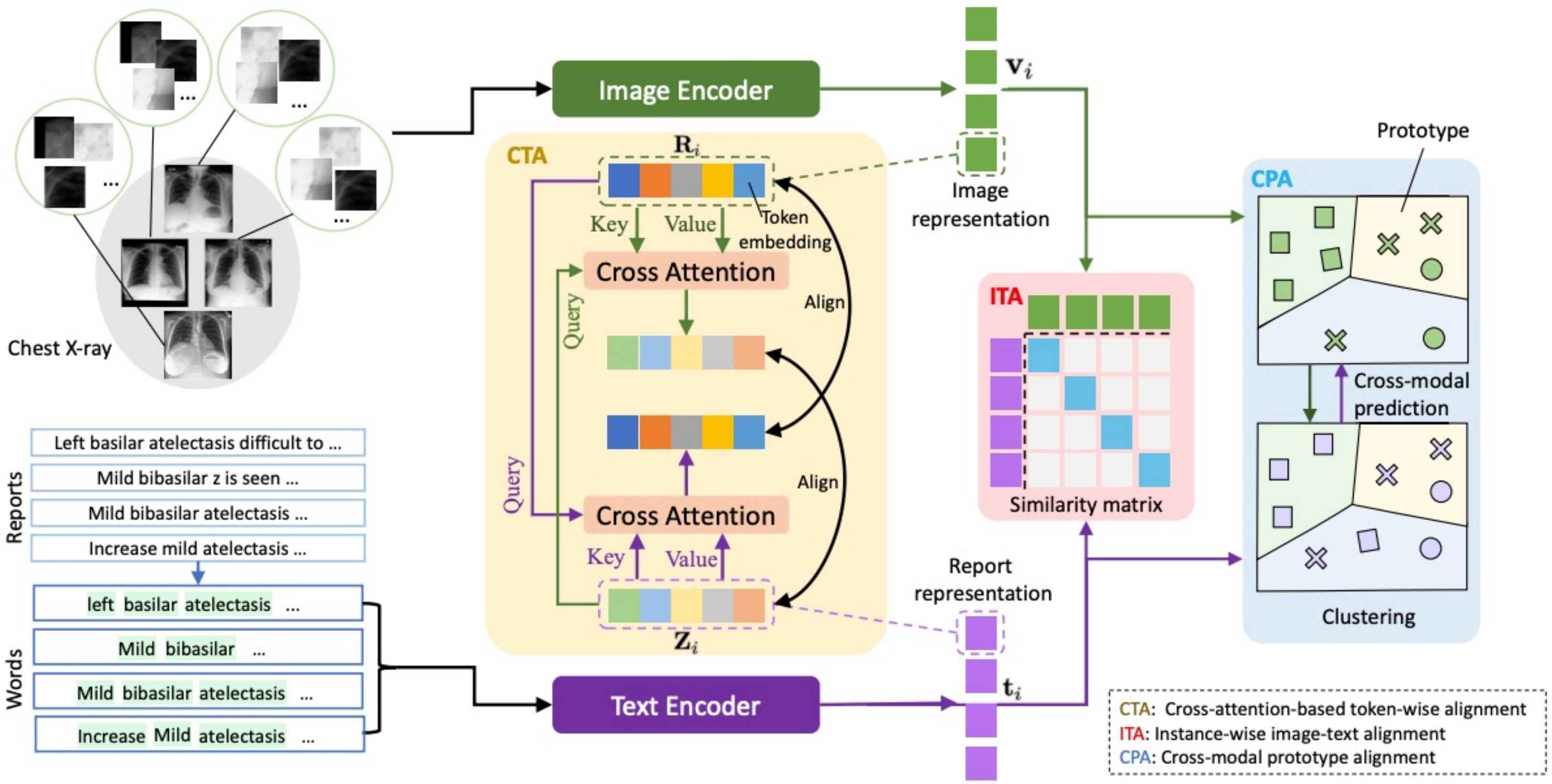
[1]. Huang, S. C., Shen, L., Lungren, M. P., & Yeung, S. (2021). Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3942-3951).
[2]. Zhou, H. Y., Chen, X., Zhang, Y., Luo, R., Wang, L., & Yu, Y. (2022). Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, *4*(1), 32-40.

Our motivation: Multi-granularity correspondence



- **Observation**: image-text pairs naturally exhibit three-level semantic correspondences, i.e., pathological region-level, instance-level and disease-level.

Our framework: Multi-granularity cross-modal alignment (MGCA)



- Exploit multi-granularity image-text alignment (Token-wise, instance-wise and prototype-wise alignment) for visual representation learning

Linear classification performance

Table 1: Linear classification results on CheXpert, RSNA and COVIDx with 1%, 10%, 100% training data. Area under ROC curve (AUROC [%]) are reported for CheXpert and RSNA dataset, and accuracy (ACC [%]) is reported for COVIDx dataset. The best and second-best results are highlighted in red and blue, respectively.

Method	CheXpert (AUC)			RSNA (AUC)			COVIDx (ACC)		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
Random Init	56.1	62.6	65.7	58.9	69.4	74.1	50.5	60.3	70.0
ImageNet Init	74.4	79.7	81.4	74.9	74.5	76.3	64.8	78.8	86.3
pre-trained on CheXpert									
DSVE [16]	50.1	51.0	51.5	49.7	52.1	57.8	-	-	-
VSE++ [14]	50.3	51.2	52.4	49.4	57.2	67.9	-	-	-
GLoRIA [24]	86.6	87.8	88.1	86.1	88.0	88.6	67.3	77.8	89.0
pre-trained on MIMIC-CXR									
Caption-Transformer [6]	77.2	82.6	83.9	-	-	-	-	-	-
Caption-LSTM [54]	85.2	85.3	86.2	-	-	-	-	-	-
Contrastive-Binary [45][44]	84.5	85.6	85.8	-	-	-	-	-	-
ConVIRT [55]	85.9	86.8	87.3	77.4	80.1	81.3	72.5	82.5	92.0
GLoRIA-MIMIC [24]	87.1	88.7	88.0	87.0	89.4	90.2	66.5	80.5	88.8
MGCA(Ours, ResNet-50)	87.6	88.0	88.2	88.6	89.1	89.9	72.0	83.5	90.5
MGCA(Ours, ViT-B/16)	**88.8**	**89.1**	**89.7**	**89.1**	**89.9**	**90.8**	**74.8**	**84.8**	**92.3**

Object detection and semantic segmentation performance

Table 2: Object detection results (mAP [%]) on RSNA and Object CXR. Each dataset is fine-tuned with 1%, 10%, 100% training data. Best results are in boldface. "-" means mAP is smaller than 1%.

Method	RSNA 1%	RSNA 10%	RSNA 100%	Object CXR 1%	Object CXR 10%	Object CXR 100%
Random	1.00	4.00	8.90	-	0.49	4.40
ImageNet	3.60	8.00	15.7	-	2.90	8.30
ConVIRT [55]	8.20	15.6	17.9	-	8.60	15.9
GLoRIA [24]	9.80	14.8	18.8	-	10.6	15.6
GLoRIA-MIMIC [24]	11.6	16.1	24.8	-	8.90	16.6
MGCA (Ours)	**12.9**	**16.8**	**24.9**	-	**12.1**	**19.2**

Table 3: Semantic segmentation results (Dice [%]) on SIIM and RSNA. Each dataset is fine-tuned with 1%, 10%, 100% training data. Best results of each setting are in boldface.

Method	SIIM 1%	SIIM 10%	SIIM 100%	RSNA 1%	RSNA 10%	RSNA 100%
Random	9.00	28.6	54.3	6.90	10.6	18.5
ImageNet	10.2	35.5	63.5	34.8	39.9	64.0
ConVIRT[55]	25.0	43.2	59.9	55.0	67.4	67.5
GLoRIA[24]	35.8	46.9	63.4	59.3	67.5	67.8
GLoRIA-MIMIC [24]	37.4	57.1	64.0	60.3	**68.7**	68.3
MGCA (Ours)	**49.7**	**59.3**	**64.2**	**63.0**	68.3	**69.8**

Ablation study

Table 4: Ablation study of our framework under linear classification setting.

Training tasks			CheXpert (AUC)			RSNA (AUC)		
ITA	CTA	CPA	1%	10%	100%	1%	10%	100%
✓			87.6	88.2	88.5	88.4	89.5	90.5
✓	✓		88.3	88.9	89.1	88.9	89.8	90.7
✓		✓	88.5	88.9	89.0	88.6	89.2	90.4
✓	✓	✓	**88.8**	**89.1**	**89.7**	**89.1**	**89.9**	**90.8**

Table 5: Results of natural VLP pre-trained models on linear classification setting.

	CheXpert (AUC)			RSNA (AUC)		
	1%	10%	100%	1%	10%	100%
BLIP [31]	69.1	74.9	77.7	53.7	82.0	84.1
MGCA (Ours)	**88.8**	**89.1**	**89.7**	**89.1**	**89.9**	**90.8**

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Visualization results

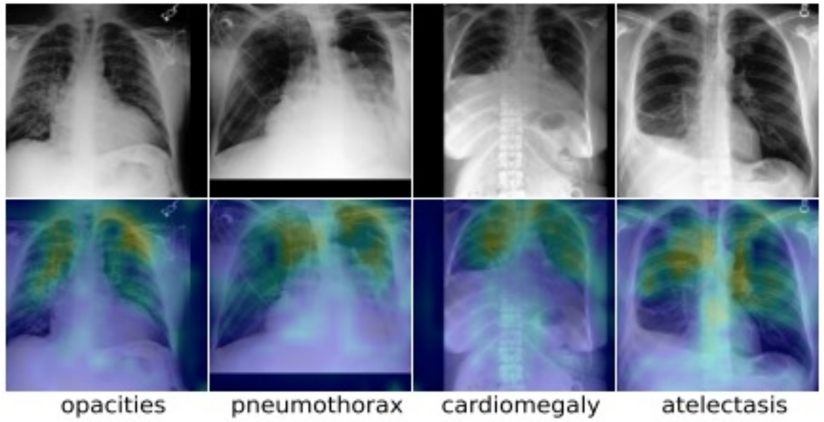


Figure 3: Visualization of learned token correspondence by our MGCA. Highlighted pixels represent higher activation weights by corresponding word.

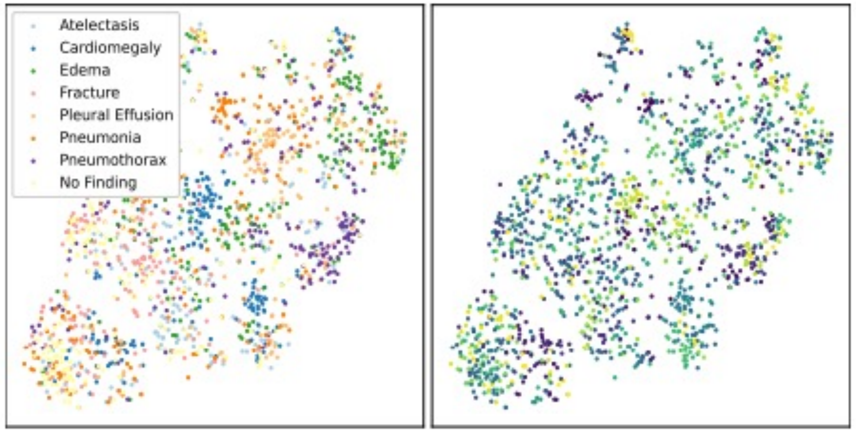


Figure 4: t-SNE visualizations of encoded image representations. Colors indicate the ground truth disease types and cluster assignment in left and right sub-figures.

Conclusion and Future work

- We propose a multi-granularity cross-modal alignment framework for learning better medical visual representation
- (Future work) One potential direction is to explore how to leverage multi-granularity correspondence in a holistic manner
- (Future work) We might also extend our framework as the integration of discrimination-based and generation-based method

Thank you!