Eye-gaze Guided Multi-modal Alignment for Medical Representation Learning

Chong Ma[1], Hanqi Jiang[2], Wenting Chen[3], Yiwei Li[2], Zihao Wu[2], Xiaowei Yu[4], Zhengliang Liu[2], Lei Guo[1], Dajiang Zhu[4], Tuo Zhang[1], Dinggang Shen[5], Tianming Liu[2], Xiang Li[6]*

[1]Northwest Polytechnical University [2]University of Georgia [3]City University of Hong Kong

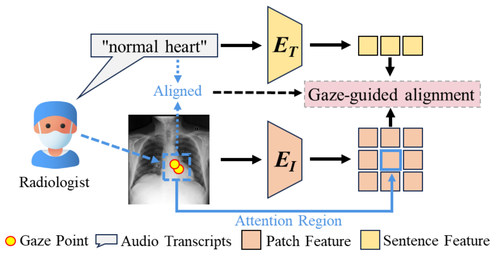
[4]University of Texas at Arlington [5]ShanghaiTech University & Shanghai United Imaging Intelligence Co.

[6]Massachusetts General Hospital & Harvard Medical School

NEURAL INFORMATION PROCESSING SYSTEMS

EGMA Paper EGMA Code

Introduction



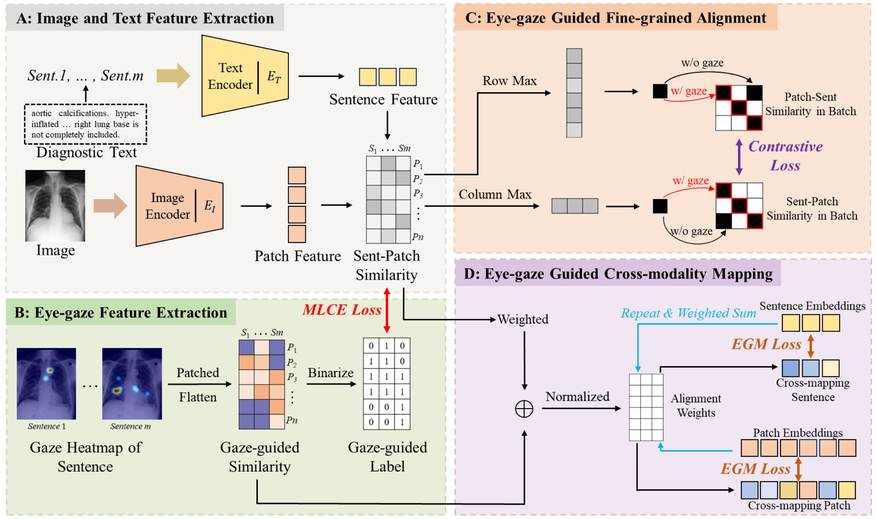
Motivation:

- Overcoming alignment complexity in medical multi-modal learning.
- Leveraging radiologists' eye-gaze data for efficient multi-modal alignment.

Contribution:

- A novel framework EGMA for medical multi-modal alignment, which makes the first attempt to integrate eye-gaze data into vision-language pre-training.
- EGMA demonstrates that even a small amount of eye-gaze data can effectively assist in multi-modal pre-training and improve the feature representation ability of the model.

Framework



Overview: Our EGMA framework processes images and text through an encoder, extracting patch features and sentence feature representations to generate a fine-grained similarity matrix for instances. Subsequently, two types of eye-gaze-based auxiliary information are utilized to achieve fine-grained alignment and cross-mapping alignment at different stages.

Cross-modality Mapping

- We generate image-to-text and text-to-image alignment weight matrices W^{I2T} and W^{T2I} using matrices GS_k, x_k^{P2S} and x_k^{S2P}, as shown below:

$$W^{I2T} = norm(\omega(x_k^{P2S}) + GS_k), W^{T2I} = norm(\omega(x_k^{S2P}) + (GS_k)^T$$

- Here, $norm$ denotes normalization, and ω applies sparse and binarization operations. Using the weight matrices, we map text features S_k^m to image features $Cross\_P_k^m$ and vice versa:

$$Cross\_P_k^i = \sum_{j=1}^{m} S_k^j \cdot W_{ij}^{I2T}, Cross\_S_k^j = \sum_{i=1}^{n} P_k^i \cdot W_{ji}^{T2I}$$

- Finally, these mapped features and their corresponding target features are used to compute the alignment contrastive loss:

$$L_{EGM} = \frac{1}{2}\sum_{k=1}^{b}(mL_k^I + mL_k^T)$$

Eye-gaze Guided Fine-grained Alignment

- We utilize eye-gaze-guided fine-grained alignment in our model to enhance the interaction between patch and sentence features. Based on local patch and sentence features, we compute the similarities between sentences and patches in both directions. Using these similarity matrices, we generate a gaze-guided label matrix GL_k and optimize it through multi-label cross-entropy (MLCE) loss. The fine-grained features are calculated as:

$$\hat{z}_k^I = \frac{1}{n}\sum_{i=1}^{n} max_j[(x_k^{P2S})_{ij}], \hat{z}_k^T = \frac{1}{m}\sum_{j=1}^{m} max_i[(x_k^{S2P})_{ji}]$$

- These fine-grained features are used to enhance the alignment between image and text, ensuring that relevant information at the local level is properly captured. Finally, the total Eye-Gaze Fine-grained (EGF) alignment loss is calculated as:

$$L_{EGF} = \frac{1}{2b}\sum_{k=1}^{b}(fL_k^{S2P} + fL_k^{P2S}) + \frac{1}{2}\sum_{k=1}^{b}(\hat{L}_k^{T2I} + \hat{L}_k^{I2T})$$

Experimental Results

Method	CheXpert			RSNA			SIIM-ACR		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
ConVIRT	85.90	86.80	87.30	77.40	80.10	88.60	-	-	-
BioViL	81.95	85.37	88.62	81.76	85.68	88.64	80.26	82.79	90.51
MedKLIP	-	-	-	87.31	87.99	89.31	85.27	90.71	91.88
MGCA	85.80	87.66	89.30	85.22	87.54	89.24	86.12	89.66	92.16
GLoRIA	86.60	87.80	88.10	86.10	88.00	88.60	-	-	-
PRIOR	86.16	87.08	89.08	86.72	88.07	89.19	88.35	89.72	92.49
MedCLIP	85.74	87.49	88.02	87.61	88.19	89.10	88.84	91.13	92.18
EGMA(Ours)	**87.71**	**88.92**	**89.50**	**88.41**	**89.40**	**90.10**	**90.78**	**92.17**	**93.29**

Supervised Classification: Comparison results of supervised classification task with other SOTA models on CheXpert, RSNA, and SIIM-ACR datasets. Area under ROC curve (AUROC) is reported with different portions of training data: 1%, 10%, 100%.

Method	CheXpert 5x200		RSNA		SIIM-ACR	
	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
CLIP	20.10	9.12	25.03	22.07	49.39	47.98
GLoRIA	53.30	48.99	29.15	28.54	22.57	22.57
PRIOR	34.90	30.56	76.77	51.80	50.00	33.33
MGCA	43.60	41.37	60.83	**57.77**	30.03	25.45
MedCLIP	57.50	55.97	43.09	31.01	58.40	57.85
EGMA(Ours)	**61.30**	**60.38**	**76.97**	43.49	**63.62**	**61.46**

Zero-shot Classification: Comparison results of zero-shot classification tasks with other SOTA models on CheXpert 5x200, RSNA, and SIIM-ACR datasets. The Accuracy (Acc.) and F1-score (F1) metrics are reported.

Method	Image-to-text			Text-to-image		
	P@1↑	P@5↑	P@10↑	P@1↑	P@5↑	P@10↑
CLIP	12.75	12.48	10.03	5.00	12.50	12.50
MedCLIP	14.50	15.98	15.86	12.50	12.50	15.00
MGCA	35.00	27.80	23.33	45.00	47.50	44.00
GLoRIA	38.75	31.62	24.51	52.50	49.00	50.25
ConVIRT	-	-	-	60.25	60.00	57.50
EGMA(Ours)	**42.65**	**37.50**	**28.84**	**80.00**	**74.50**	**69.50**

Image Retrieval: Comparison results of zero-shot retrieval task with other SOTA models on CheXpert 8x200 dataset. The Precision at Top-1, Top-5, and Top-10 are reported.

Visualization Results