Supplementary Materials: Towards Medical Vision-Language Contrastive Pre-training via Study-Oriented Semantic Exploration

Bo Liu, Zexin Lu, Yan Wang
Sichuan University, Chengdu, China

A PSEUDO CODE OF SENSE

To better elaborate the proposed SENSE method, we outline the training process in Algorithm 1.

Algorithm 1 Proposed SENSE Framework.

```
Input: Training set containing $N$ patient studies; preset number of clusters $K$
for number of epochs do
    Generate $N$ momentum study-level features $\{\hat{\boldsymbol{m}}\}$
    Cluster $\{\hat{\boldsymbol{m}}\}$ into $K$ groups through K-means
    for number of mini-batches do
        Sample a single image-report pair $(x_v, x_t)$ randomly for each patient study
        Apply augmentation process on $x_v$ to obtain two views $x_v^1$ and $x_v^2$
        Feed image views to normal and momentum visual encoders to obtain feature $\boldsymbol{v}$ and $\hat{\boldsymbol{v}}$
        Feed report to normal and momentum texture encoders with Max-Max pooling and semantics-level augmentation to generate $\boldsymbol{t}$ and $\hat{\boldsymbol{t}}$
        Compute loss $\mathcal{L}_{S-CMC}$ (Eq. 2), $\mathcal{L}_{S-MMC}$ (Eq. 5), and $\mathcal{L}_{S-UMC}$ (Eq. 7) with corresponding projectors, which compose $\mathcal{L}_{SENSE}$ (Eq. 8) with the loss balancing coefficients $\lambda, \beta, \gamma$
    end for
    Update network parameters through the back-propagation of criterion $\mathcal{L}_{SENSE}$ using AdamW
end for
```

B IMPLEMENTATION DETAILS

Below, we provide implementation details, including data preprocessing, model training, and inference, for downstream tasks.

B.1 Cross-modal Retrieval

We use the test set of MIMIC-CXR with 2, 461 image-report pairs for the evaluation of cross-modal retrieval. The images and reports are pre-processed in the same way as in pre-training.

The pre-trained normal image and report encoders with cross-projectors ($p^v_{cro.}(v)$ and $p^t_{cro.}(t)$) are directly used for cross-modal retrieval without further fine-tuning. Specifically, images and reports are fed to the corresponding encoders and projectors to obtain their embeddings, respectively. Then, for the image-to-report (report-to-image) subtask, top k reports (images) whose features are closest to that of a given image (report) are retrieved based on their cosine similarity [3, 6].

B.2 Data-efficient Image Classification

We conduct classification on two datasets: (1) CheXpert is a multi-label chest X-ray dataset where each image is labeled based on the appearance of 14 kinds of disease symptoms. We follow the same experimental settings in [3, 6] to use the validation set for evaluation because the original test set has not been made publicly available. Moreover, we randomly pick 5, 000 images from the training set for validation; (2) RSNA Pneumonia consists of two types of chest X-ray images, *i.e.*, health and pneumonia. We use the same preprocessing procedure and experimental setting as in [3]. Specifically, about 30,000 front view images are split into training/validation/test sets with a ratio of 70%/15%/15%. For both datasets, we resize the image to 256 × 256 pixels.

We conduct the fine-tuning on an NVIDIA A100 GPU, following the training setting of GLoRIA[3]. For the evaluation protocol, we freeze the weights of the pre-trained normal image encoder and train a randomly initialized linear classifier on top of it. The optimizer is Adam with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-6}. The learning rate scheduler monitors the validation loss and reduces the learning rate by a factor of 0.5 if the validation loss does not decrease for five epochs. The batch size is 64, and the total number of training epochs is 50. Data augmentations are applied during training, including random cropping to 224 × 224 pixels and random horizontal flipping with a probability of 0.5.

B.3 Zero-shot Image Classification

We evaluate the model's zero-shot recognition ability on the test set of RSNA Pneumonia. Each image is resized to 256 × 256 pixels. For class-specific prompts, we set the prompt for the healthy class to "no evidence of pneumonia" and that for the unhealthy class to "findings suggesting pneumonia", following [1].

The pre-trained normal image and report encoders with cross-projectors ($p^v_{cro.}(v)$ and $p^t_{cro.}(t)$) are directly used for zero-shot image classification[5]. Given an input image, the model specifies which class prompt is the best match by calculating the cosine similarity between the image and prompt embeddings which are produced by the corresponding encoders and projectors.

B.4 Medical Image Segmentation

We use the SIIM Pneumothorax dataset for the evaluation of segmentation, which contains 12, 047 chest radiographs and is split into training/validation/test sets in a ratio of 70%/15%/15%. Both images and masks are resized to 512 × 512 pixels.

We conduct the segmentation on an NVIDIA A100 GPU. The Albumentations [2] Python library is used for data augmentations

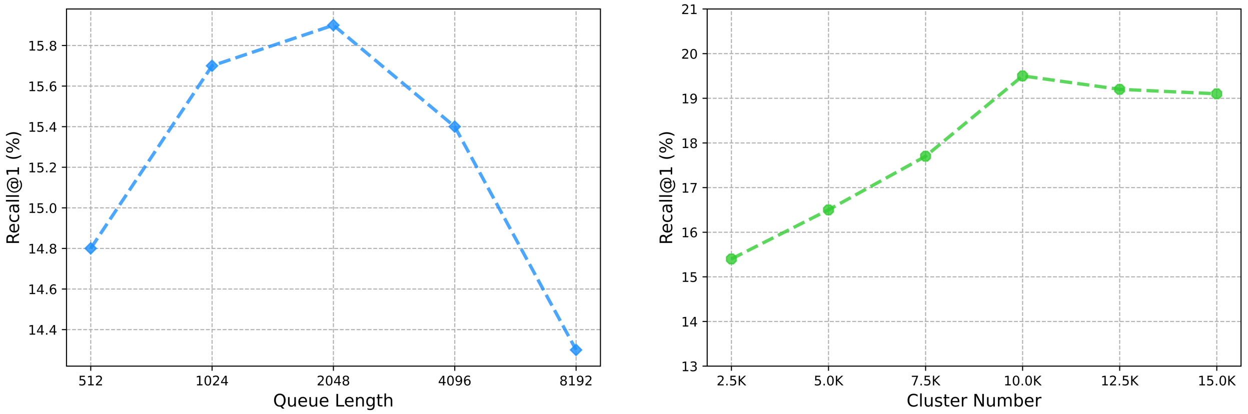


Figure S1: Impact of different queue length and cluster number on image-to-report retrieval task. When ablating the queue length (left), our model here pretrains with only $\mathcal{L}_{S-CMC}$ and Max pooling. Then, we fix the queue length as 2048 and search for the best cluster number using full objective $\mathcal{L}_{SENSE}$ (right).

Table S1: Results of image-to-report retrieval on the test sef of MIMIC-CXR [4] with/without uni-modal projectors. Our model here pre-trains with full objectives, i.e., $\mathcal{L}_{S-CMC}$, $\mathcal{L}_{S-UMC}$, and $\mathcal{L}_{S-MMC}$.

SENSE	Image-to-Report Retrieval		
	R@1	R@5	R@10
w/o uni-projectors	17.1	43.2	53.8
w/ uni-projectors	**19.5**	**45.1**	**57.3**

with a probability of 0.5, including rotation (within ±10°) and resizing (within [0.9, 1.1]). We use the Adam optimizer with an initial learning rate of 5×10^{-4} and a weight decay of 1×10^{-6} for network optimization. The optimization objective is the combination of Focal loss and Dice loss [7]. The learning rate scheduler monitors the validation loss and reduces the learning rate by a factor of 0.5 if the validation loss does not decrease for three consecutive epochs. The batch size is set to 8, and the maximum number of training epochs is 100.

C FURTHER ABLATION STUDY AND ANALYSIS

C.0.1 Queue Length and Cluster Number. As shown in Figure S1 (left), the model's performance first increases and then decreases with the queue length growing. The best results are obtained with a queue length of $2,048$. For results of choosing cluster number in Figure S1 (right), when the number of clusters is within 10K–15K, the performance is relatively stable, with 10K achieving the best performance.

C.0.2 Extra Projectors for Uni-modal Contrast. As stated in Section 3.3.3, we introduce additional projectors for uni-modal contrast to avoid the negative impact of directly using the cross-modal projection heads on cross-modal tasks. Results in Table S1 show that without the additional uni-modal projectors, the performance drops sharply, thus demonstrating the necessity.

C.0.3 Token Activation for Retrieval. We present the token activation weights during extracting the textual features for retrieval in Figure S2. The model is pre-trained with only $\mathcal{L}_{S-CMC}$ for better illustrating the impact of the pooling method. For Max pooling, the whole report is processed together, enclosed by a single pair of special tokens ($<$ cls $>$, $<$ sep $>$); while for the proposed Max-Max pooling, each sentence in the report is enclosed by a pair of ($<$ cls $>$, $<$ sep $>$) tokens for sentence-independent encoding. As we can see, compared with Max pooling, our proposed Max-Max focuses more on the crucial information, such as "chronic scar".

REFERENCES

[1] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. *arXiv preprint arXiv:2204.09817* (2022).

[2] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. 2020. Albumentations: fast and flexible image augmentations. *Information* 11, 2 (2020). https://doi.org/10.3390/info11020125

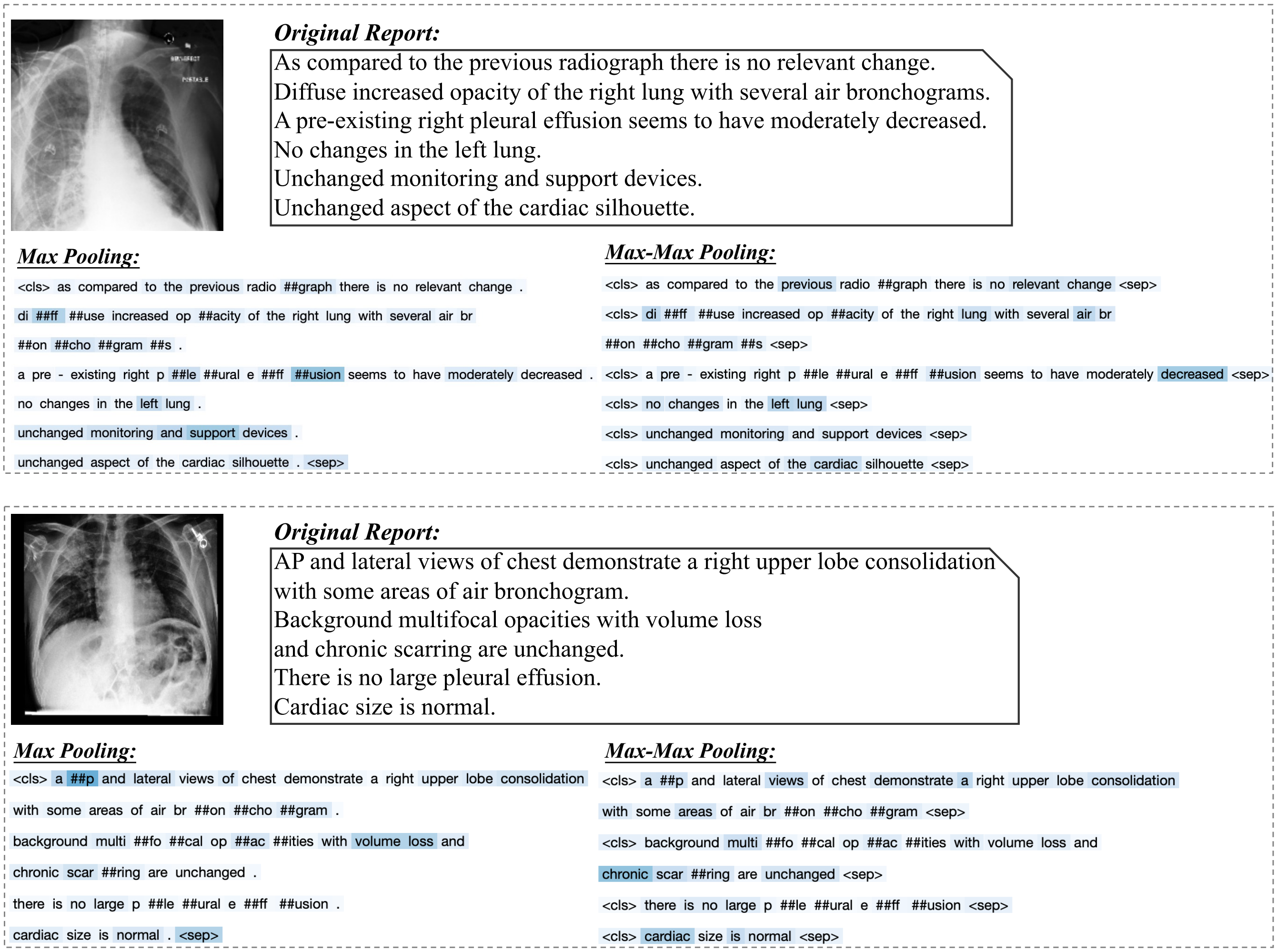
[3] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3942–3951.

[4] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* 6, 1 (2019), 1–8.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[6] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747* (2020).

[7] Wentao Zhu, Yufang Huang, Liang Zeng, Xuming Chen, Yong Liu, Zhen Qian, Nan Du, Wei Fan, and Xiaohui Xie. 2019. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical Physics* 46, 2 (2019), 576–589.



As compared to the previous radiograph there is no relevant change.

Diffuse increased opacity of the right lung with several air bronchograms.

A pre-existing right pleural effusion seems to have moderately decreased.

No changes in the left lung.

Unchanged monitoring and support devices.

Unchanged aspect of the cardiac silhouette.

<cls> as compared to the previous radio ##graph there is no relevant change .

di ##ff ##use increased op ##acity of the right lung with several air br

##on ##cho ##gram ##s .

a pre - existing right p ##le ##ural e ##ff ##usion seems to have moderately decreased .

no changes in the left lung .

unchanged monitoring and support devices .

unchanged aspect of the cardiac silhouette . <sep>

<cls> as compared to the previous radio ##graph there is no relevant change <sep>

<cls> di ##ff ##use increased op ##acity of the right lung with several air br

##on ##cho ##gram ##s <sep>

<cls> a pre - existing right p ##le ##ural e ##ff ##usion seems to have moderately decreased <sep>

<cls> no changes in the left lung <sep>

<cls> unchanged monitoring and support devices <sep>

<cls> unchanged aspect of the cardiac silhouette <sep>

AP and lateral views of chest demonstrate a right upper lobe consolidation with some areas of air bronchogram.

Background multifocal opacities with volume loss and chronic scarring are unchanged.

There is no large pleural effusion.

Cardiac size is normal.

<cls> a ##p and lateral views of chest demonstrate a right upper lobe consolidation

with some areas of air br ##on ##cho ##gram .

background multi ##fo ##cal op ##ac ##ities with volume loss and

chronic scar ##ring are unchanged .

there is no large p ##le ##ural e ##ff ##usion .

cardiac size is normal . <sep>

<cls> a ##p and lateral views of chest demonstrate a right upper lobe consolidation

with some areas of air br ##on ##cho ##gram <sep>

<cls> background multi ##fo ##cal op ##ac ##ities with volume loss and

chronic scar ##ring are unchanged <sep>

<cls> there is no large p ##le ##ural e ##ff ##usion <sep>

<cls> cardiac size is normal <sep>

Figure S2: Visualization of the Max **and** Max-Max **pooling methods for extracting report features. Deeper color indicates larger activation.**