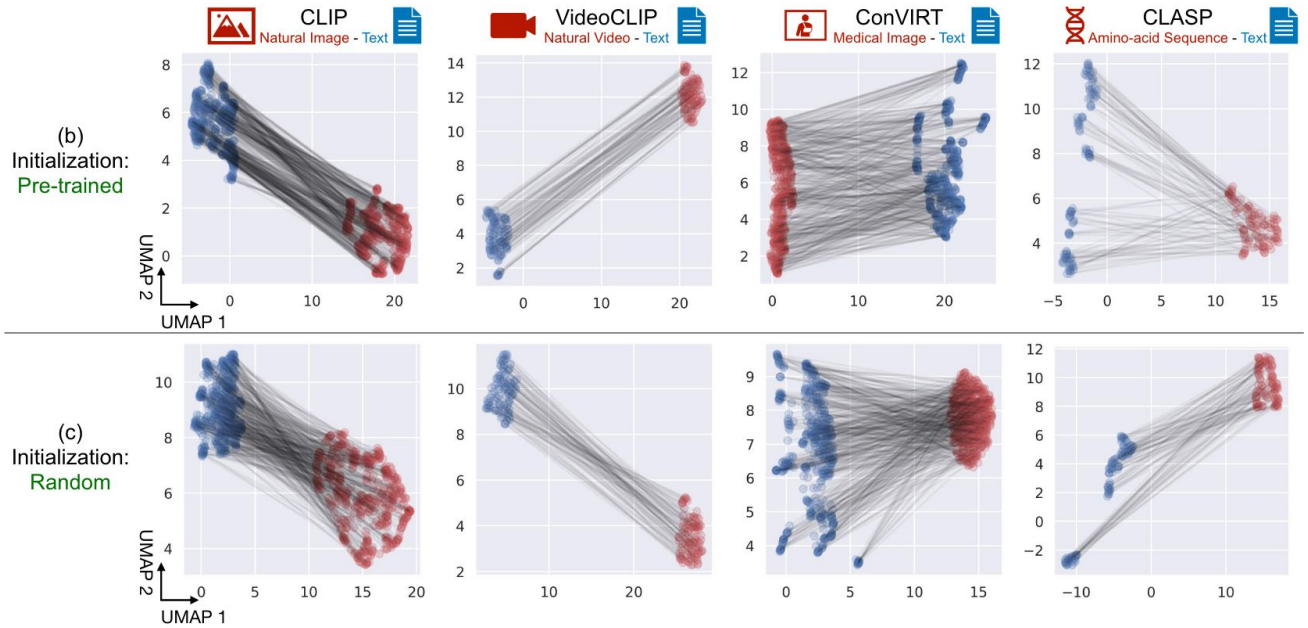
# Improving Medical Multi-modal Contrastive Learning with Expert Annotations

Yogesh Kumar and Pekka Marttinen

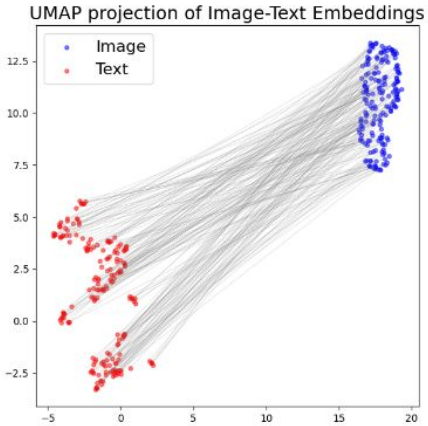
Department of Computer Science, Aalto University, Finland

# Modality Gap in CLIP Embeddings

UMAP projection reveals that CLIP embeddings form distinct clusters that pertain to their respective modalities (e.g., images and text)



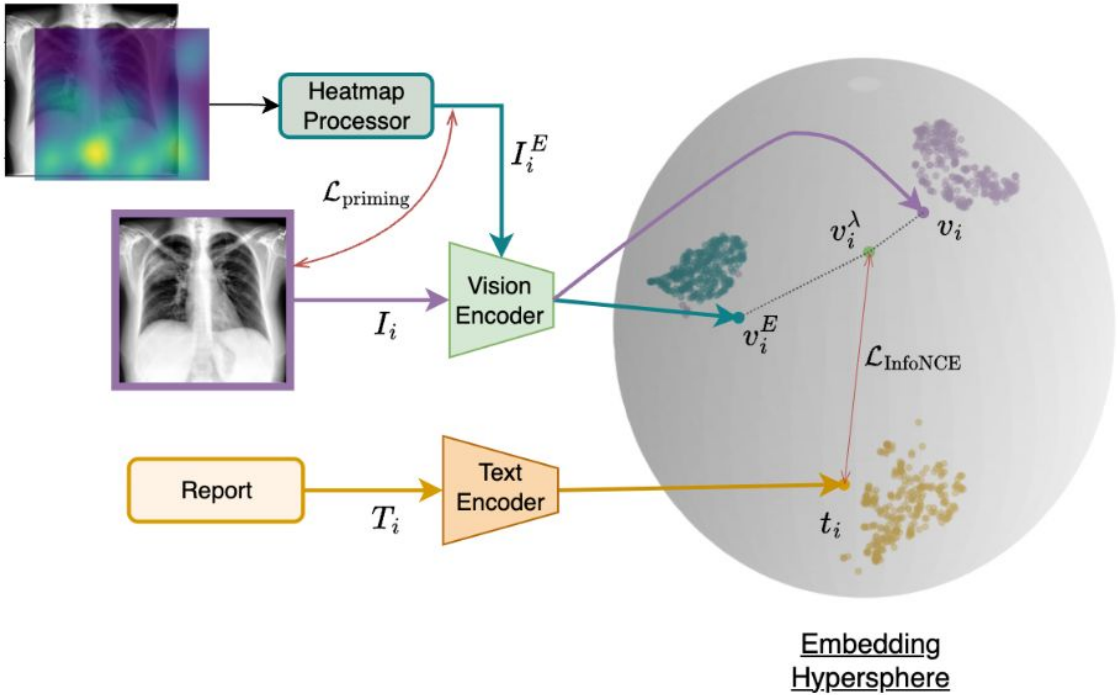

CLIP on OpenI Dataset
Chest X-ray - Report pairs




Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning, Liang et al. 2022

# Expert Annotated CLIP - eCLIP

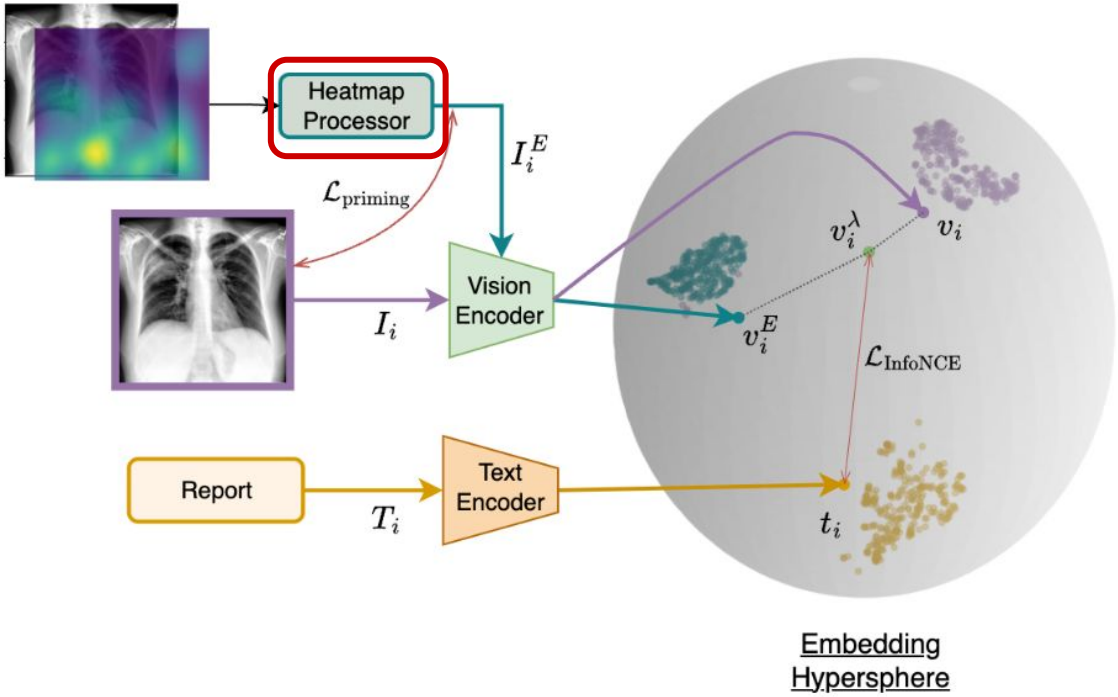
- For medical data, pairs of chest X-rays and their corresponding radiologist report form the image-text pairs used for CLIP pretraining
- Additionally, we can leverage the heatmaps of radiologist's gaze patterns obtained through eye-tracking

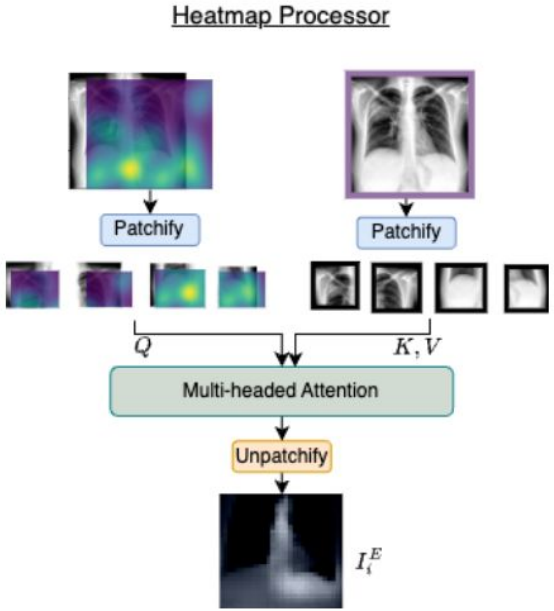



eCLIP creates an additional modality – expert image $I_i^E$

# Heatmap Processor

Heatmap processor is a single layer of multi-headed attention used to fuse the eye-gaze heatmaps with X-ray image, forming the "expert image" $I_i^E$

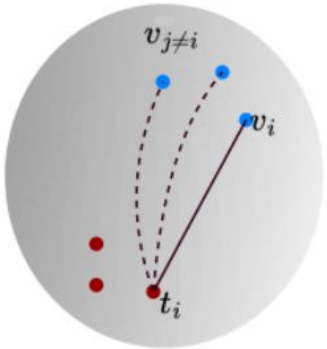


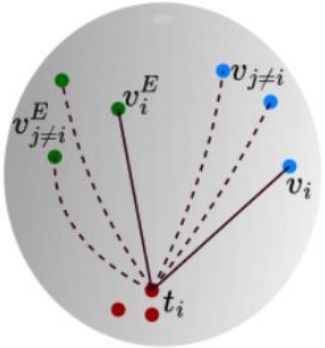
# Additional Image-Text pairs for Contrastive Loss

- eCLIP exploits the modality gap to create additional pairs for contrastive loss using the expert image embeddings $v_i^E$
- Heatmap processor is crucial to ensure that image embeddings $v_i$ and the expert image embeddings $v_i^E$ are not too similar

CLIP




eCLIP




$$\mathcal{L}_{\text{text}} = \mathop{\mathbb{E}}_{(t_i, v_i)\sim pos}\left[-\log \frac{\exp(sim(t_i, v_i)/\tau)}{\exp(sim(t_i, v_i)/\tau) + \sum_{j\neq i}\exp(sim(t_i, v_j)/\tau)}\right]$$

# Experiments

Zero Shot Image Classification (F1-Score)

| Model | Dataset | | | |
|---|---|---|---|---|
| | Chexpert 5x200 | MIMIC 5x200 | RSNA | CXR 14x100 |
| $\text{CLIP}_{\text{ViT Base}}$ | $0.540_{\pm.017}$ | $0.465_{\pm.004}$ | $0.805_{\pm.001}$ | $0.183_{\pm.011}$ |
| +naive | $0.506_{\pm.011}$ | $0.426_{\pm.006}$ | $0.805_{\pm.004}$ | $0.151_{\pm.009}$ |
| +DACL | $0.474_{\pm.007}$ | $0.400_{\pm.002}$ | $0.759_{\pm.001}$ | $0.106_{\pm.003}$ |
| +$m^3$-mix | $0.542_{\pm.021}$ | $0.465_{\pm.013}$ | $0.798_{\pm.004}$ | $0.183_{\pm.020}$ |
| +expert (ours) | $\mathbf{0.563_{\pm.021}}$ | $\mathbf{0.477_{\pm.004}}$ | $0.814_{\pm.003}$ | $\mathbf{0.193_{\pm.017}}$ |

Pre-training Sample Efficiency