InVERGe: Intelligent Visual Encoder for Bridging Modalities in Report Generation

[1]Ankan Deria, [1]Komal Kumar, [1]Snehashis Chakraborty, [2]Dwarikanath Mahapatra, [1]Sudipta Roy
[1]Artificial Intelligence & Data Science, Jio Institute, Navi Mumbai, India; [2]Inception Institute of Artificial Intelligence, UAE

CHALLENGES

Challenge 1: In medical imaging capturing pixel-level details such as colour is unnecessary but region, intensity and other details are important.

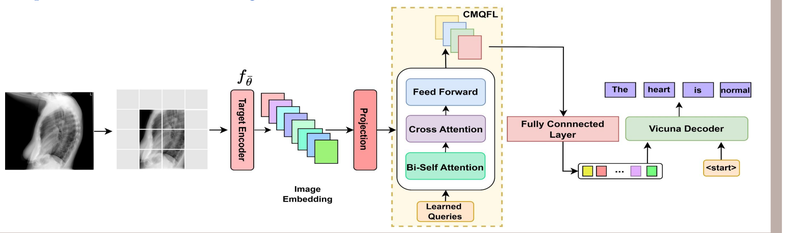
InVERGe: Employ a Self-Supervised Joint-Embedding Predictive Architecture to efficiently train the encoder in a self-supervised manner, obtaining high-level semantic image representations.

Challenge 2: Hard to align two different modalities in a common space and the complex medical visual and textual data biases make this task more challenging.

InVERGe: Incorporates a lightweight transformer known as the CMQFL layer, which utilizes the output from a frozen encoder to identify the most relevant text-grounded image embedding to bridge this modality gap.

Challenge 3: High computational demands due to end-to-end model training of the large scale models.

InVERGe: Apply two-step training approach. In the first stage, we train the CMQFL layer to enhance visual representation. In the second stage, we fine-tune the decoder.



QUANTITIVE RESULT

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
MIMIC-CXR	BLIP-2	0.377	0.221	0.125	0.088	0.152	0.274
	R2GenGPT	0.392	0.229	0.129	**0.101**	0.159	0.283
	InVERGE	**0.425**	**0.240**	**0.132**	0.100	**0.175**	**0.309**
IU X-RAY	BLIP-2	0.476	0.273	0.210	0.168	0.181	0.372
	R2GenGPT	0.481	0.301	0.214	**0.169**	0.189	0.372
	InVERGE	**0.499**	**0.324**	**0.226**	0.168	**0.195**	**0.384**
CDD-CESM	BLIP-2	0.382	0.235	0.139	0.102	0.301	0.342
	R2GenGPT	0.417	0.249	0.165	0.129	0.354	0.377
	InVERGE	**0.453**	**0.267**	**0.185**	**0.134**	**0.391**	**0.430**

Comparison of the proposed InVERGe and other SOTA methods.

➢ Effect of CMQFL Layer :

Baseline refers to one normal ViT encoder and decoder only. CPE (Context Pixel encoder) stands for our trained Model's Target Encoder.

MODEL	BLEU-2	METEOR	ROUGE-L
Baseline	0.161	0.124	0.255
MAE+Decoder	0.178	0.060	0.208
CPE+Decoder	0.183	0.117	0.260
CPE+CMQFL+ Decoder	0.227	0.163	0.290

➢ Effect of Objective Functions :

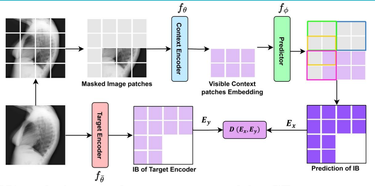
The combined effect of these three objective functions leads to a substantial improvement in the performance of the model.

MODEL'S OBJECTIVE	BLEU-1	BLEU-2	METEOR	ROUGE-L
MLM	0.410	0.227	0.163	0.290
MLM+MMM	0.416	0.231	0.166	0.30
MLM+MMM +MCL	0.425	0.24	0.175	0.309

METHODOLOGY

Semantic Image Representation :

Context encoder (CE) is a ViT which uses the unmasked visible block to predict the originating of the interested blocks by the target encoder (TE). The TE is the same as the CE which is interested in some masked blocks. The predictor is a small ViT that takes the CE's output and predicts the representations of the interested blocks of the TE at a specific location. After the prediction of the predictor, the L_2 loss (i.e., $D(E_x, E_y)$) is computed between the interested block of the Target encoder and the Prediction of those blocks by the Predictor. Simultaneously, the parameters of TE are continuously adjusted, achieved by applying an exponential moving average (EMA) technique to the parameters of the CE.



After completing the training of the entire model, we adopt the target encoder as our primary encoder for the InVERGe.

Text Grounded Image Embedding:

CMQFL layer learns to extract most useful text-grounded image embeddings(z) for the decoder to generate the desired report by jointly optimizing three objective functions.

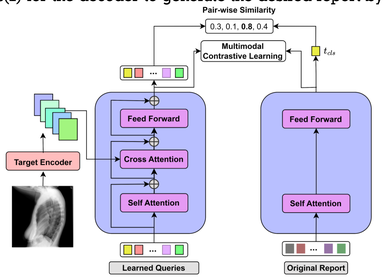
➢ Multimodal Contrastive Learning (MCL) :

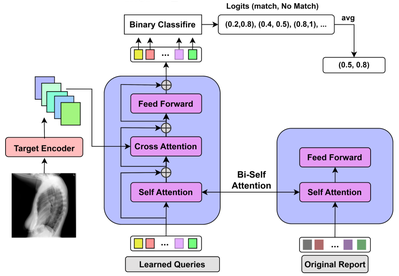
Objective: Align image-text labels by contrasting similarity of positive and negative pairs.

- CMQFL layer with cross-attention mechanisms generates informative text-grounded image embeddings.
- Evaluate similarity between each query output and text [CLS] token.
- Compute similarity using linear transformations (L_z, L_t) of embeddings.
- Image-to-text similarity $\text{sim(Q,T)} = \frac{\max(L_z(z)\cdot L_t(t_{cls}))}{\tau}$
- Text-to-image similarity $\text{sim(T,Q)} = \frac{\max(L_t(t_{cls})\cdot L_z(z))}{\tau}$

Objective Function between Y_f and similarity :

$$\mathcal{L}_{\text{mcl}} = \frac{1}{2}[\text{CE}(Y_f, (\text{sim(Q,T)}) + \text{CE}(Y_f, (\text{sim(T,Q)})]$$





➢ Multi-Modality Matching (MMM) :

Objective: Enhance quality of pairs with hard negative mining.

- Utilize bi-directional self-attention mask for queries and texts.
- Query embedding (Q) captures multimodal information. Cross-Entropy Loss ($\mathcal{L}_{mmm}$) measures dissimilarity between predicted and ground truth class probabilities:

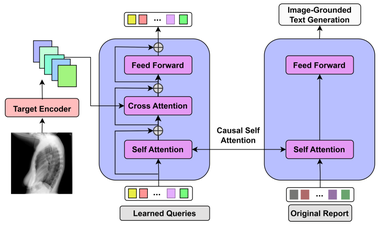
$$\mathcal{L}_{mmm} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} y_{ij}\log(p_{ij})$$

➢ Masked Language Modelling (MLM) :

Objective : ensures that the model maximizes the likelihood of the text tokens, particularly the masked ones, during training, a vital aspect of autoregressive text generation.

Due to the CMQFL layer's architecture, direct interactions between the fixed image encoder and text tokens are not initially feasible. Therefore, the information necessary for text generation is extracted by the queries.

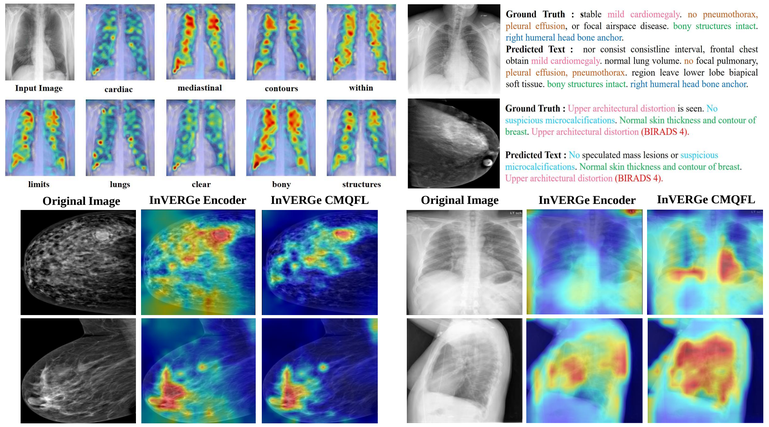
$$\mathcal{L}_{mlm} = -\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}_{mask}(j)\log(p_{ij})$$



The total pre-training objective for training the CMQFL layer of InVERGe consists of a combination of three distinct damage components:

$$\mathcal{L} = \mathcal{L}_{\text{mcl}} + \mathcal{L}_{\text{mmm}} + \mathcal{L}_{\text{mlm}}$$

QUALITATIVE RESULT



Ground Truth : stable mild cardiomegaly. no pneumothorax, pleural effusion, or focal airspace disease. bony structures intact. right humeral head bone anchor.
Predicted Text : nor consist consistline interval, frontal chest obtain mild cardiomegaly. normal lung volume. no focal pulmonary, pleural effusion, pneumothorax. region leave lower lobe biapical soft tissue. bony structures intact. right humeral head bone anchor.

Ground Truth : Upper architectural distortion is seen. No suspicious microcalcifications. Normal skin thickness and contour of breast. Upper architectural distortion (BIRADS 4).
Predicted Text : No speculated mass lesions or suspicious microcalcifications. Normal skin thickness and contour of breast. Upper architectural distortion (BIRADS 4).

Presenting attention map visualizations and results.

CONCLUSIONS

- ❑ we present a novel, high-performing report generation model that significantly advances the alignment of texts with corresponding visual features. By employing a two-stage training procedure, focusing initially on the CMQFL layer and then fine-tuning the decoder, the model demonstrates exceptional performance.
- ❑ Model's ability to generate reports without requiring additional annotations or external task-specific knowledge.
- ❑ The CMQFL layer, with its three objectives, contributes to creating small yet highly informative image embeddings, promoting a more grounded vision and language representation.
- ❑ Experiments performed on multiple dataset having different modalities and data sequences demonstrate the effectiveness and the generalizability of the model. The model consistently outperforms SOTA medical report generation models.
- ❑ Limitation: Our model's decoder, pre-trained on natural language, lacks medical-specific knowledge. Integrating an LLM trained on diverse medical datasets might generate more comprehensive and contextually relevant reports.

REFERENCES

1. Mahmoud Assran et al. "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture" CVPR 2023.
2. Junnan Li et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models" ICML 2023.
3. Junnan Li et al. "Align before fuse: Vision and language representation learning with momentum distillation" NeurIPS 2021.
4. Alistair et al. "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports"
5. Indiana university - chest x-rays (xml reports). https:// openi.nlm.nih.gov/faq.php