Uni-Mlip: Unified Self-supervision for Medical Vision Language Pre-training

Technology Innovation Institute

Ameera Bawazir, Kebin Wu, Wenbin Li

{ameera.bawazir, Kebin.wu, wenbin.li}@tii.ae

BMVC 2024

PAPER-ID: 329

Introduction

• **Vision-and-Language Pre-training (VLP)** models align image and text representations, improving multimodal understanding.

• In the **medical field**, multimodal data is common, but privacy concerns and complex annotations hinder the acquisition of large datasets.

• **Medical VLP** models use self-supervised learning but struggle to fully integrate domain-specific knowledge.

• **Uni-MLIP** addresses these challenges with a unified self-supervision framework for medical vision-language pre-training, enhancing tasks like image-text retrieval, classification, and VQA

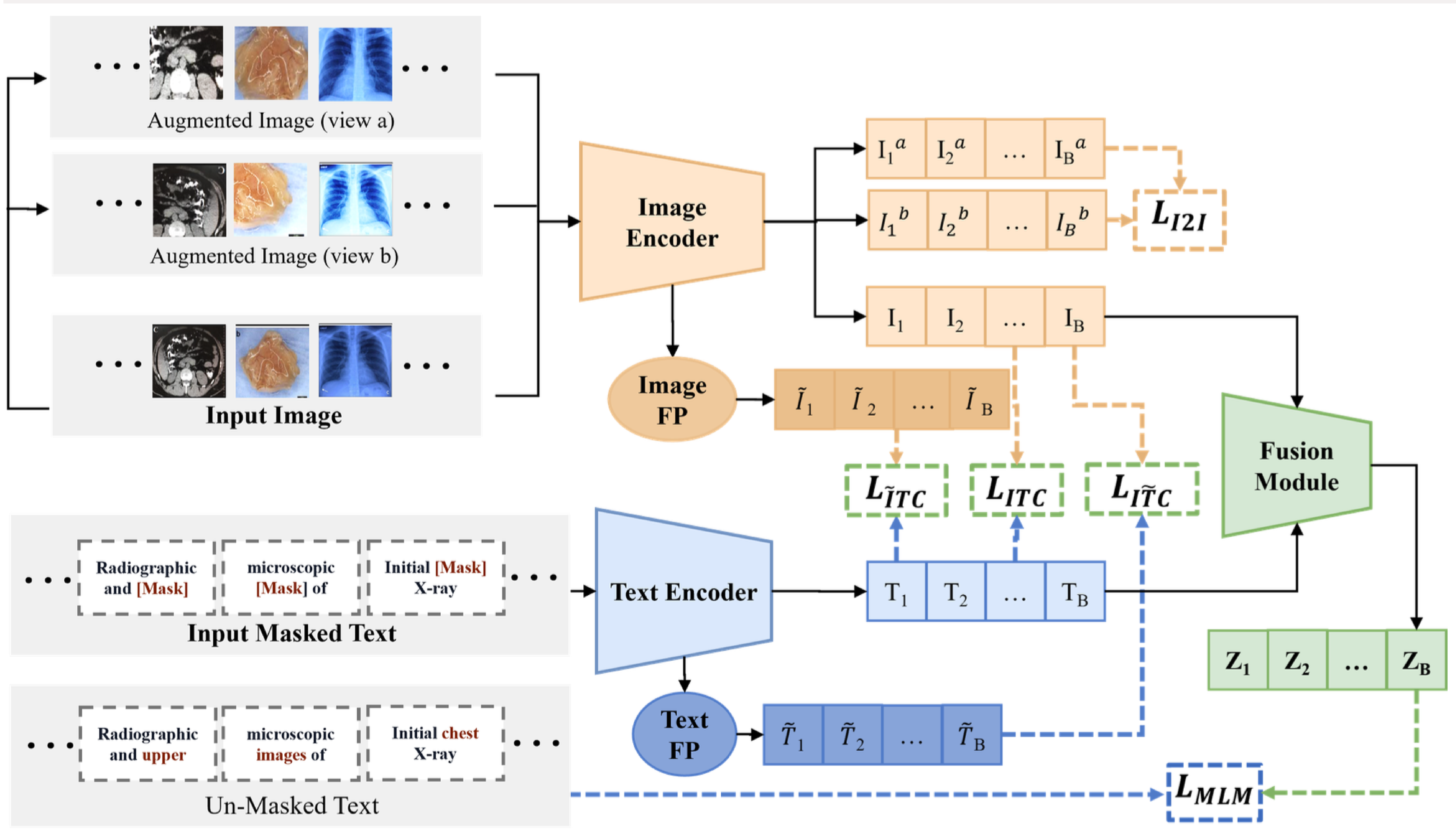
Contribution

1) **Unified Self-Supervision Framework**: Uni-Mlip integrates feature-level and data-level self-supervision across uni-modal and multimodal contexts, aligning image and text modalities.

2) **Specialized for Medical Images**: Uni-Mlip adapts self-supervision for medical images, overcoming intensity challenges and improving representation learning.

3) **State-of-the-Art Performance**: Experiments show that Uni-Mlip outperforms existing methods in tasks like image-text retrieval, classification, and VQA.

Uni-Mlip



Cross-modal input level self-supervision:

- Image-Text Contrastive (ITC) loss aligns input image and input text embeddings.

$$\mathcal{L}_{ITC} = -\frac{1}{2B}\sum_{i=1}^{B} log \frac{e^{sim(I_i,T_i)/\tau}}{\sum_{j=1}^{B} e^{sim(I_i,T_j)/\tau}} - \frac{1}{2B}\sum_{i=1}^{B} log \frac{e^{sim(I_i,T_i)/\tau}}{\sum_{j=1}^{B} e^{sim(I_j,T_i)/\tau}}$$

Cross-modal feature-level self-supervision:

- Image-Text Contrastive (ITC) loss aligns perturbed image and text embeddings.

Uni-modal self-supervision – image:

- Image-Image Contrastive loss (I2I) tailored for medical images.

$$\mathcal{L}_{I2I} = -\frac{1}{2B}\sum_{i=1}^{B} log \frac{e^{sim(I_i^a,I_i^b)/\tau}}{\sum_{j=1}^{B} e^{sim(I_i^a,I_j^b)/\tau}} - \frac{1}{2B}\sum_{i=1}^{B} log \frac{e^{sim(I_i^a,I_i^b)/\tau}}{\sum_{j=1}^{B} e^{sim(I_j^a,I_i^b)/\tau}}$$

Fused-modal self-supervision:

- Improves text representation using masked language modeling (MLM).

$$\mathcal{L}_{MLM} = \mathbb{E}_{(V,C)\sim D}[CE(y^{mask}, p^{mask}(V,C))]$$

Total training loss:

$$\mathcal{L}_{Total} = \lambda_{cm}\cdot(\mathcal{L}_{ITC} + \mathcal{L}_{\tilde{I}TC} + \mathcal{L}_{I\tilde{T}C}) + \lambda_{um}\cdot(\mathcal{L}_{MLM} + \mathcal{L}_{I2I})$$

Ablation

• Performance Gain from Pre-training Objectives

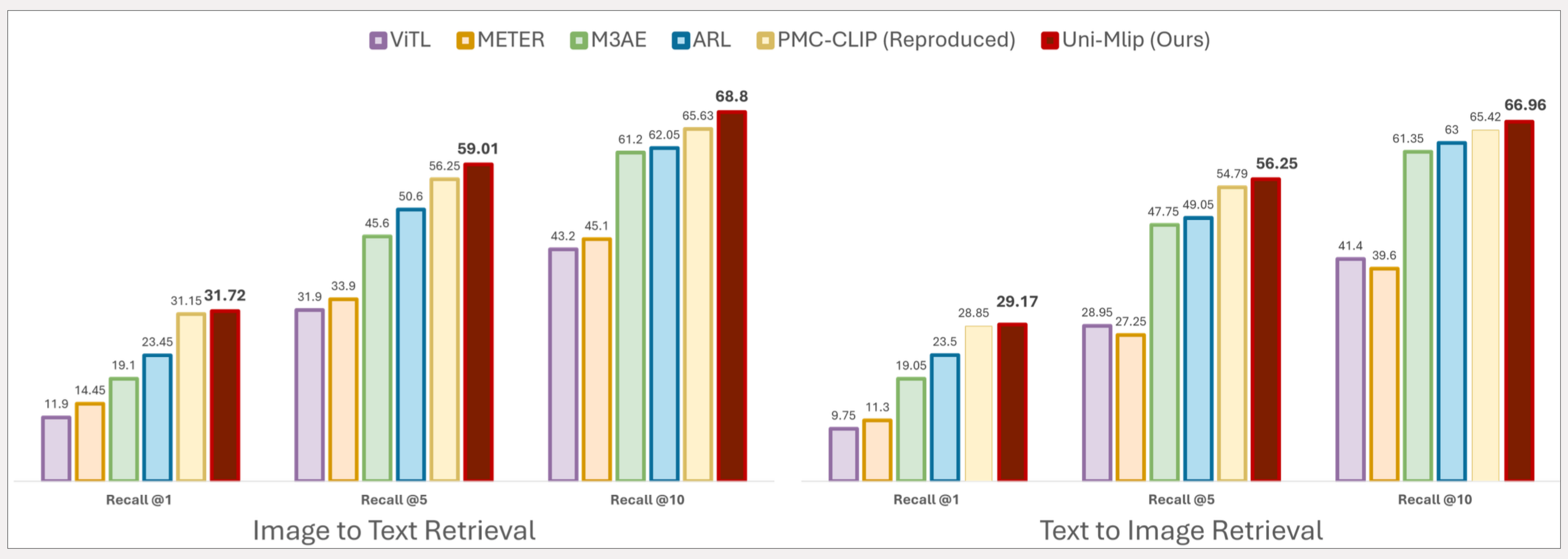
$\mathcal{L}_{ITC}$	$\mathcal{L}_{MLM}$	$\mathcal{L}_{I\tilde{T}C}$	$\mathcal{L}_{\tilde{I}TC}$	$\mathcal{L}_{I2I}$	I2T	T2I	Performance Gain (I2T/T2I)
✓	✓	✗	✗	✗	22.2	21.7	0.0 / 0.0
✓	✓	✓	✗	✗	23.2	23.3	+1.0 / +1.6
✓	✓	✗	✓	✗	23.8	23.4	+1.6 / +1.7
✓	✓	✗	✗	✓	25.7	24.0	+3.5 / +2.3
✓	✓	✓	✓	✓	**25.9**	**24.5**	+3.7 / + 2.8

• Effect of Freezing Batch Normalization on Image SSL Objective

Vision Encoder Input	Training Objectives	Freeze BN	I2T	T2I
V	$\mathcal{L}_{ITC}, \mathcal{L}_{MLM}$	✗	22.2	21.7
V, V^a, V^b	$\mathcal{L}_{ITC}, \mathcal{L}_{MLM}, \mathcal{L}_{I2I}$	✗	5.0	1.0
V, V^a, V^b	$\mathcal{L}_{ITC}, \mathcal{L}_{MLM}$	✗	4.0	6.9
V, V^a, V^b	$\mathcal{L}_{ITC}, \mathcal{L}_{MLM}, \mathcal{L}_{I2I}$	✓	25.7	24.0

Results

• Medical Image-Text Retrieval



• Medical Image Classification

Method	MIMIC			CXP			NIH		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
Random	53.6	66.5	78.2	62.6	69.0	76.9	56.4	67.1	76.9
ImageNet	67.8	70.5	79.3	63.7	70.7	77.7	59.7	68.9	78.1
ConVIRT [45]	67.8	73.4	80.1	63.2	71.3	77.7	60.0	69.0	76.6
GLoRIA [17]	67.5	72.6	80.1	62.9	69.0	77.8	60.1	71.2	77.7
MGCA [40]	68.4	74.4	80.2	63.4	72.1	78.1	61.1	67.8	77.3
M-FLAG [30]	69.5	74.8	80.2	64.4	71.4	78.1	62.2	71.6	78.7
PMC-CLIP [27] (reproduced)	73.1	77.4	81.8	69.1	74.9	79.1	64.9	76.3	82.3
Uni-Mlip (ours)	**73.2**	**79.1**	**82.0**	**69.1**	**75.3**	**79.8**	**65.6**	**76.4**	**82.9**

• Medical Visual Question Answering

Methods	VQA-RAD			Slake		
	Open	Closed	Overall	Open	Closed	Overall
MEVF-BAN [33]	49.20	77.20	66.10	77.80	79.80	78.60
CPRD-BAN [28]	52.50	77.90	67.80	79.50	83.40	81.10
PubMedCLIP [8]	60.10	80.00	72.10	78.40	82.50	80.10
PMC-CLIP [27] (reproduced)	**61.45**	80.14	72.73	**80.16**	84.38	81.81
Uni-Mlip (ours)	60.43	**81.62**	**73.17**	79.38	**85.82**	**81.90**