MI2RL

Multi modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training

Jong Hak Moon, Hyung yung Lee*,*

Woncheol Shin,

Young-Hak Kim, and Edward Choi

발표자 : 임지섭

울산대학교 UNIVERSITY OF ULSAN

서울아산병원 Asan Medical Center

Abstract

- 최근 많은 연구가 Multi modal 사전 훈련 목표로 BERT 아키텍처를 확장하여 image captioning 및 visual question answering과 같은 다양한 vision-language multi-modal task에서 인상적인 성능을 보여주었다.
- 우리는 radiology이미지 와 판독문을 사용해 medical domain에서 multi-modal representation learning task 들을 탐구한다.
- 우리는 Vision Language Understanding task들(진단 분류, 의료 이미지 보고서 검색, 시각적 질문 답변)과 Vision Language Generation task들(판독문 생성) 모두에 대한 일반화 성능을 극대화하기 위해 새로운 multi-modal attention masking 체계와 결합된 BERT 기반 아키텍처를 채택하는 Medical Vision Language Learner(MedViLL)를 제안한다
- 세 개의 radiographic image-report datasets(MIMIC-CXR, Open-I, VQA-RAD)를 사용하여 네 개의 downstream 작업에 대해 제안된 모델을 통계적이고 엄격하게 평가함으로써 작업별 아키텍처를 포함한 다양한 기준선에 대해 MedViLL의 우수한 다운스트림 작업 성능을 경험적으로 입증한다.

Introduction

- radiographic images 및 관련 free-text description(예: Chest X-rays 와radiology report)을 사용하는 VL(Vision-Language) multi-modal연구는 의료 정보학에서 가장 중요하고 흥미로운 작업 중 하나이다.
- 각 VL(Vision-Language) 양식은 연구자에게 서로 다른 표현을 제공하지만 이미지와 판독문에는 상호 도움이 되는 의미 정보가 포함되어 있다. 따라서 VL multi-modal 연구의 발전은 diagnosis classification, report generation과 같은 다양한 작업에 대한 자동화된 지원을 제공함으로써 임상 치료의 품질을 향상시키는 데 도움이 될 수 있다.
- 그러나 높은 차원성, 이질성 및 체계적 편향으로 인해 공동 표현을 학습하기 위해 이미지와 임상 판독문를 모두 처리하는 것은 상당한 기술적 과제를 제기한다

Introduction

- VL 멀티모달 학습의 개발은 최근 딥 러닝 영역에서 BERT 기반 아키텍처를 확장함으로써 엄청난 발전을 이루었다.
- 그러나 사전 훈련된 모델을 사용하는 광범위한 다운스트림 작업에 대해 상당한 개선이 보고되었음에도 불구하고, 대부분의 이전 연구는 VLU(Vision Language Understanding) 작업 또는 VLG (Vision Language Generation) 작업 하나 중점을 두었으며, 이는 VLU와 VLG 모두에 대해 의미 있는 표현을 동시에 학습하는 어려운 특성을 시사했다.
- VL multi-modal pre-training은 최근 몇 년 동안 의심할 여지 없이 상당한 진전을 보였지만, 주로 일반 도메인(예: MS-COCO 사용)의 맥락에서 개발되었다.
- VL multi-modal pre-training은 진단 정확도 향상, 보고서 자동 생성 또는 의사의 질문에 답변하는 등 의료 분야에서 널리 사용될 수 있는 큰 잠재력을 가지고 있다.

Introduction

- 본 논문에서는 의학 영역에서 joint representations of vision and language를 학습할 수 있는 모델을 개발하는 것을 목표로 한다.
  이 논문의 주요 contributions은 다음과 같이 요약될 수 있다

1. 우리는 새로운 self-attention scheme를 가진 의료 이미지와 판독문의 multi-modal pre-training model인 MedViLL(MedVision Language Learner)을 제안한다

1. 진단 분류(diagnosis classification), 의료 이미지 판독문 검색, 의료 Visual Question Answering(VQA) 및 판독문 생성을 포함하여 광범위한 VLU 및 VLG 다운스트림 작업에 대한 자세한 추가 연구를 통해 접근 방식의 효과를 입증한다.

1. 두 개의 다른 Chest X-ray datasets(MIMIC-CXR, Open-I)를 사용하여 transfer learning 에서 일반화 능력을 입증한다. 여기서 우리는 한 데이터 세트(MIMIC-CXR)에서 모델을 사전 훈련하고 다른 데이터 세트(Open-I)에서 다양한 다운스트림 작업을 수행한다

- 우리가 아는 한, 이것은 의료 영역에서 통합된 VL pre-train model로 VLU와 VLG 작업을 모두 수행하는 첫 번째이다

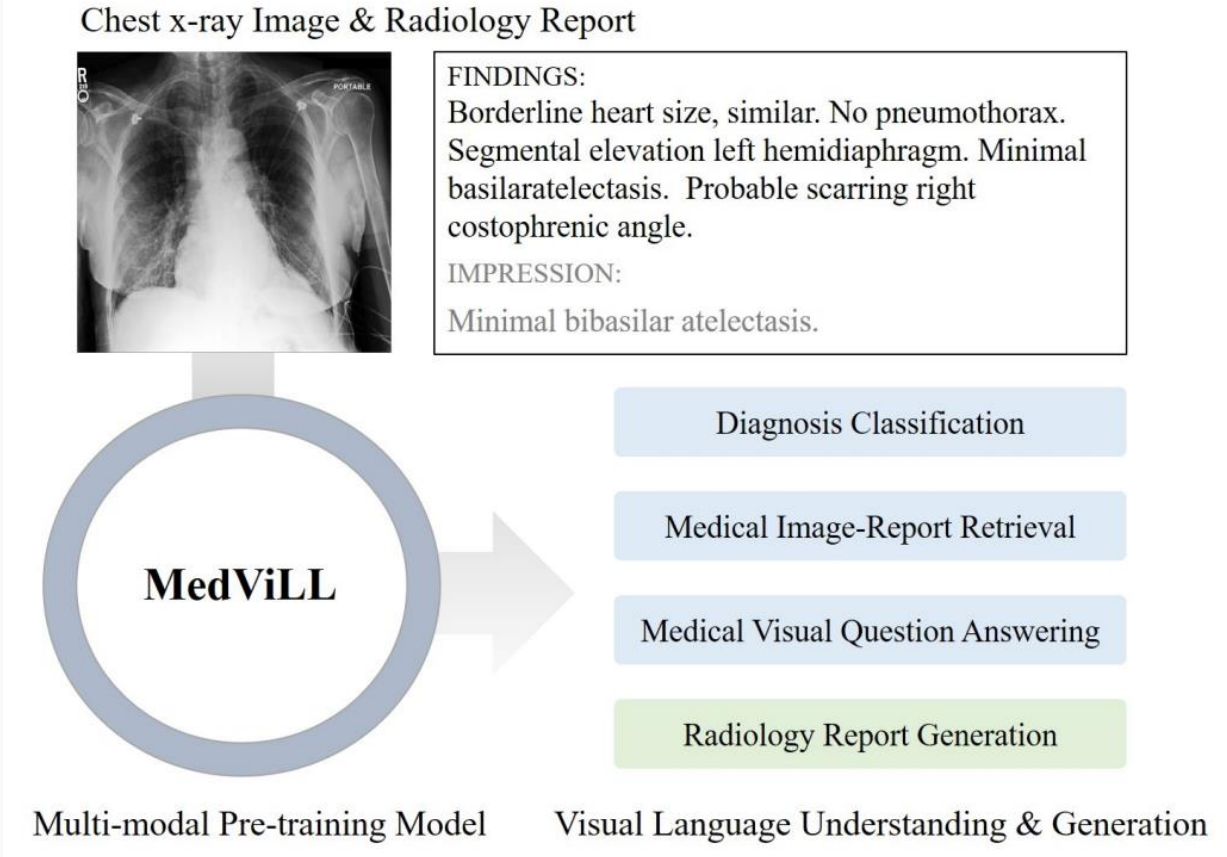


Fig. 1: **Overview of MedViLL.** During the pre-training, MedViLL learns joint representation, then fine-tuned for VLU and VLG tasks.

Related Work

A. Radiology Practices

의사는 방사선 이미지와 환자 이력을 기반으로 다양한 소견을 식별한 다음 이러한 소견과 전체적인 인상을 임상 보고서에 **요약**한다.

진단은 임상 소견에 대해 긍정적, 부정적 또는 불확실한 것으로 설명되며, 여기에는 소견의 세부 위치와 심각도가 포함된다. 이러한 임상 보고서는 현재 임상 환경에서 의사소통하기 위한 표준 방법으로 사용되고 있다.

vision and language data의 조합은 image annotation과 판독문 생성 모두에서 모델 성능을 더욱 향상시키는 데 도움이 된다

Related Work

B. VL Multimodal Researches in the Medical Domain

다양한 모델이 점진적으로 개발되었지만 의료 영역에서는 CNN-RNN 기반 모델이 VL multi-modal task에서 지배적이며, 이러한 모델은 주로 VLU 또는 VLG 작업의 작업별 방법을 위해 설계되었다.

TieNet은 ChestXray14 데이터 세트를 사용하여 VLU(예: 진단 분류) 및 VLG(예: 판독문 생성) 작업에 대한 이미지 보고서 주attention mechanism 을 갖춘 선구적인 CNN-RNN 모델이다.

가장 최근의 연구도 VLU 또는 VLG 작업에만 초점을 맞추고 있다.

Related Work

C. VL Multimodal Researches in the General Domain

VL multimodality에 대한 더 나은 이해를 위해, 최근 general domain 에서 많은 연구가 제안되었다.
우리의 접근 방식과 가장 관련이 있는 세 가지 구성 요소, 즉 입력 임베딩 스트림, 시각적 기능 임베딩 및 다운스트림 작업에 중점을 둔다.

1. Input Embedding Stream

기존 모델은 다운스트림 작업 성능에서 근소한 차이가 있는 single 또는 two-stream 아키텍처를 기반으로 두 그룹으로 나눌 수 있다.

two-stream아키텍처는 더 많은 parameters를 가지고 있는 반면, single stream 아키텍처는 parameters를 공유하고 스택을 처리함으로써 두 양식 간의 초기 상호작용을 허용한다. 아키텍처의 단순성과 시간/공간 효율성을 위해 single stream아키텍처로 모델을 설계한다.

Related Work

C. VL Multimodal Researches in the General Domain

2. Visual Feature Embedding

최근 연구의 대부분은 영역 기반 시각적 입력을 추출하기 위해 pre-trained object detectors를 활용하는 것에서 영감을 받지만 이 접근 방식은 language understanding에 대한 정보 격차가 발생한다.

픽셀 BERT는 시각적 특징 학습의 견고성을 향상시키고 과적합을 방지하기 위해 random pixel sampling이 있는 CNN 기반 시각적 인코더를 제안한다. 의료 영역에서 영역 기반 기능을 추출하는 데 적용 가능한 기성 객체 탐지기 모델이 없으므로 CNN 기반 시각적 기능 임베딩을 채택한다

Related Work

C. VL Multimodal Researches in the General Domain

3. Downstream tasks

VLU 및 VLG 작업은 비전과 언어를 결합하는 더 복잡한 작업을 해결하기 위한 VL 사전 훈련 모델의 일반적인 다운스트림 작업이다. 이와 관련하여, 많은 이전 연구는 VLU 작업을 수행하기 위해 BERT 기반 vision-language joint encoder 를 사용한다.

반면, VLG 작업에는 텍스트를 생성하는 디코더가 필요하다. Unified VLP는 사전 훈련 동안 bidirectional 과 sequence to sequence mask에 고정된 비율로 마스크 유형을 반복적으로 교대하여 단일 BERT 기반 아키텍처로 이러한 두 가지 작업(VLU 및 VLG)을 수행한다.

이 통합 사전 훈련 접근법에서 영감을 받아, 우리는 다양한 유형의 마스크와 다양한 VLU 및 VLG 다운스트림 작업에 미치는 영향을 탐구한다

MATERIALS AND METHODS

- **Dataset**

  우리는 공개적으로 사용 가능한 MIMIC-CXR 및 Open-I 데이터 세트를 사용
  MIMIC-CXR 에는 377,110개의 흉부 X선 영상과 해당하는 판독문이,
  Open-I 데이터 세트에는 3,851개의 판독문과 7,466개의 흉부 X선 이미지가 포함되어 있다.

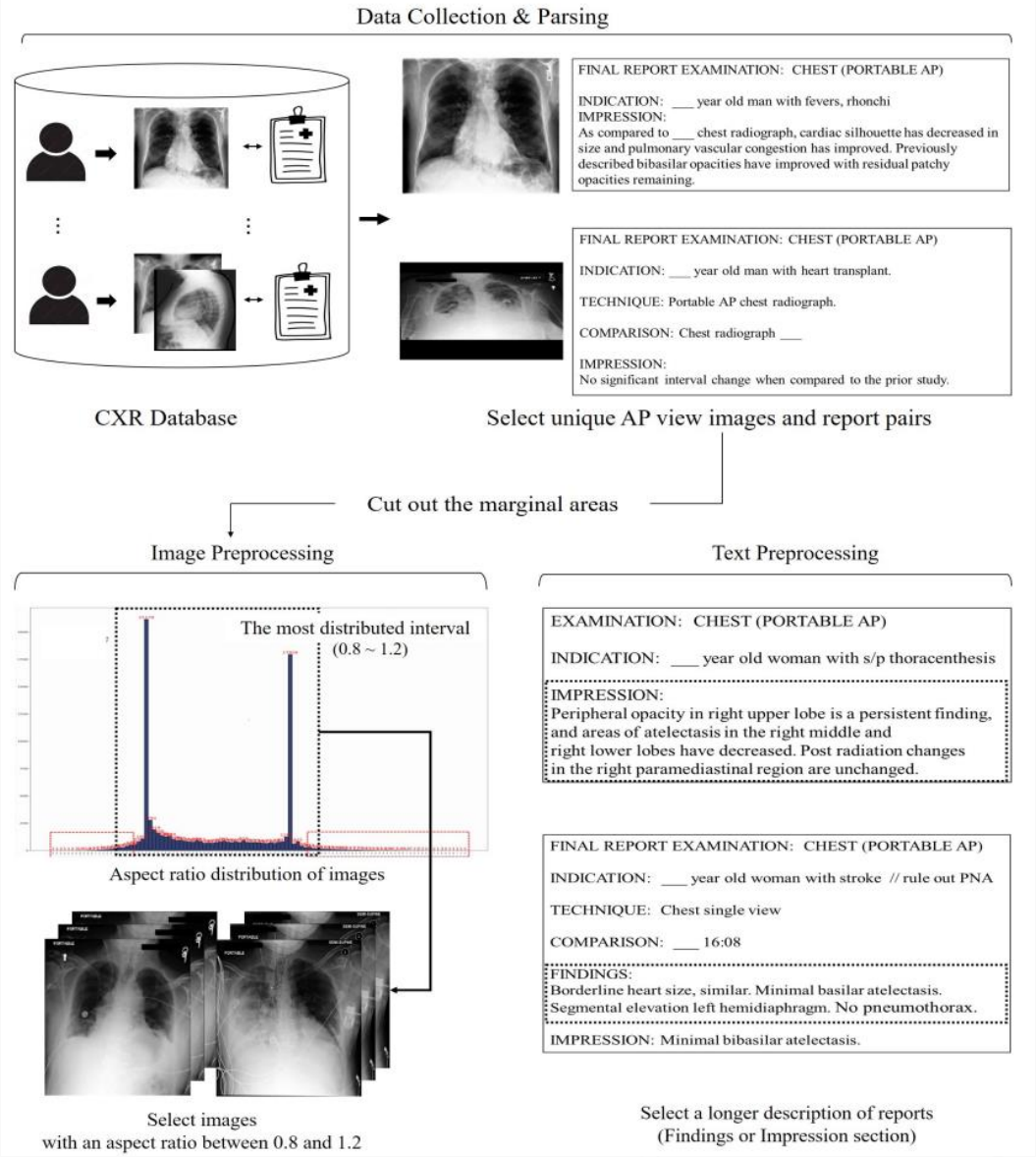
  데이터 세트에는 정면 및 측면 보기 영상이 포함되어 있으므로 영상과 리포트 쌍 간의 불일치 소견을 방지하기 위해 보기 위치를 구별해야 합니다.
  따라서 ICU(Intensive Care Units: 중환자실) 설정에서 전후방(AP) 정면 보기의 우세(예: 하나 이상의 AP 보기 이미지를 포함하는 모든 연구의 38.89%)를 고려하여 MIMICXR(train: 89,395, valid: 759, test: 1,531)과의 공식 분할에 따라 고유한 91,685개의 AP 이미지 및 관련 판독문 쌍에 대한 모든 실험을 수행한다

  공식 Open-I 데이터 세트의 3,547개 이미지, 판독문 쌍을 사용하여 모델의 일반화 능력을 테스트한다.

MATERIALS AND METHODS

- **Dataset**

  먼저 X선 이미지의 경우 원본 이미지의 주변 공간을 잘라내고 모든 이미지의 크기를 512 x 512로 조정하여 가로 세로 비율을 유지한 다음 판독문의 경우 X선 영상과 관련된 자세한 정보를 포함할 수 있는 더 긴 설명(소견 또는 인상 섹션)을 선택한다.



MATERIALS AND METHODS

- **VL Pre-training Model**

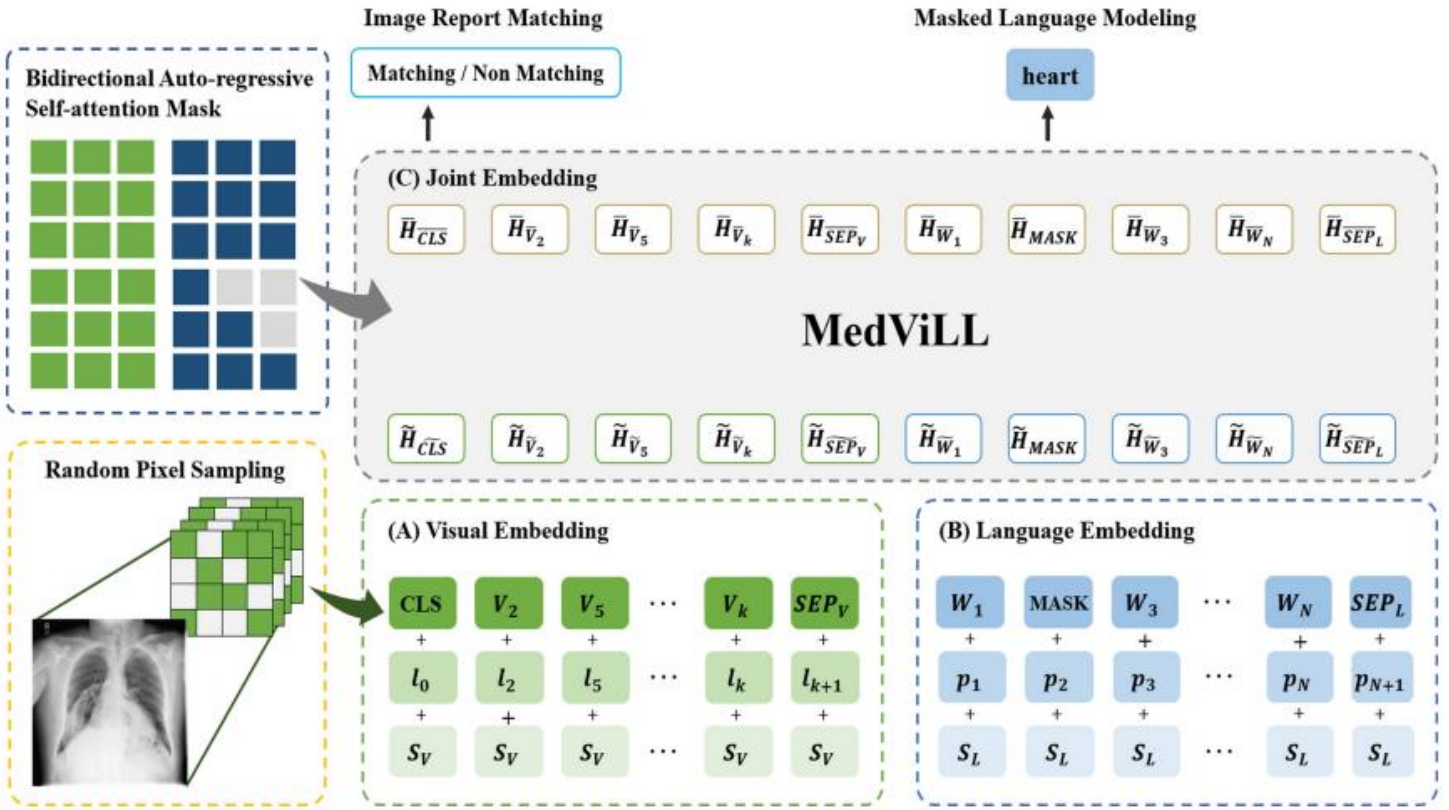


Fig. 3: **Architecture of the MedViLL.** MedViLL is a single stream BERT model for the cross-modal embedding. Chest X-ray images are randomly sampled from the last feature map of the CNN model as visual inputs. Also, each report is parsed with the BERT tokenizer to get language input. MedViLL is pre-trained with masked language modeling and image report matching tasks, and flexibly applied to VLU and VLG downstream tasks.

MATERIALS AND METHODS

- **VL Pre-training Model**
  1) Visual Feature Embedding

  의료 이미지에서 visual features을 CNN의 마지막 컨볼루션 레이어에서 얻은 다음 flatten 한다.
  X선 이미지에 동일한 추가 정보로 visual input의 position information을 인코딩한다

  V = visual features
  I = location feature
  K = visual features의 수(height × width)
  c = hidden dimension size(채널 크기)

$$\mathbf{v} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K\}, \quad \mathbf{v}_i \in \mathbb{R}^c \tag{1}$$

$$\mathbf{l} = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_K\}, \quad \mathbf{l}_i \in \mathbb{R}^c \tag{2}$$

  최종 시각적 특징 임베딩 다음과 같이 계산된다

$$\tilde{\mathbf{v}}_i = \mathbf{v}_i + \mathbf{l}_i + \mathbf{s}_V \tag{3}$$

  sV는 language 임베딩과 차별화하기 위해 모든 visual features이 공유하는 semantic 임베딩 벡터.

  사전 훈련 동안, visual input의 semantic 지식 학습을 강화하고 overfitting을 방지하기 위해 random sample한다.

  k = 샘플링된 visual features의 수
  K = 모든 visual features의 수

  Segment Embedding : 단어가 첫번째 문장에 속하는지 두번째 문장에 속하는지 알려준다.

MATERIALS AND METHODS

- **VL Pre-training Model**
  2) Language Feature Embedding

  language feature임베딩의 경우 BERT로 인코딩.
  주어진 판독문 w는 WordPiece tokenizer를 사용하여 먼저 N개의 토큰 시퀀스(즉, 하위 단어) {w1, ..., wN}로 분할된다.
  그 다음 토큰은 lookup table 을 통해 vector representation인 w = {w1, w2, ., wN}, wi∈R^d 로 변환됩니다.
  여기서 d는 임베딩 차원 크기.
  position embedding은 p = {p1, p2, ., pN}, pi∈Rd로 나타낸다.

  w = {w1, w2, ., wN}, wi∈R^d
  p = {p1, p2, ., pN}, pi∈R^d

  (vision task의 식 1,2 와 같다)

  최종 language feature 임베딩 은 다음과 같이 얻는다 $$\tilde{\mathbf{w}}_i = \mathbf{w}_i + \mathbf{p}_i + \mathbf{s}_L \quad (4)$$

  sL는 visual 임베딩과 차별화하기 위해 모든 language features이 공유하는 semantic 임베딩 벡터.

  Segment Embedding : 단어가 첫번째 문장에 속하는지 두번째 문장에 속하는지 알려준다.

MATERIALS AND METHODS

- **VL Pre-training Model**
  3) Joint Embedding

  v∈Rd 와 w∈Rd를 얻은 후 연결하여 joint 임베딩에 대한 입력 시퀀스를 구성

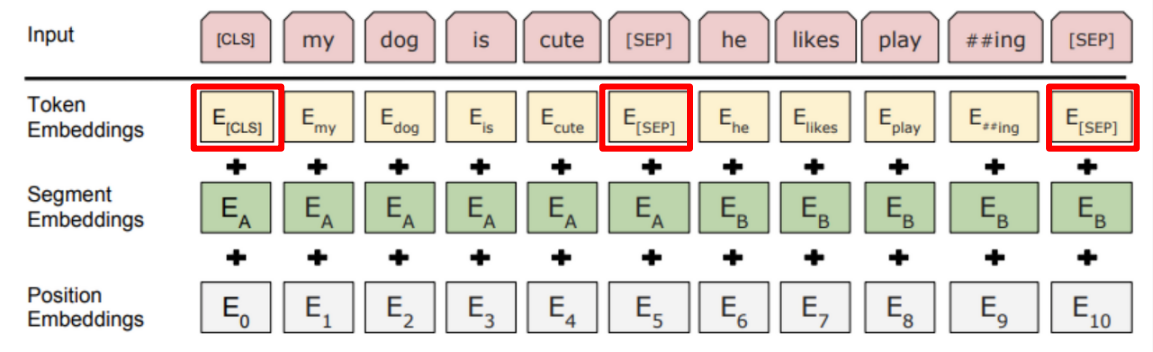
  특수 토큰 CLS와 SEP를 사용하여, 우리는 공동 임베딩 블록에 대한 입력을 H˜∈R^d×S로 정의.

  CLS, SEP_V, SEP_L은 특수 토큰을 임베딩하여 얻는다.

  공동 임베딩 블록에 의해 생성된 상황별 임베딩은 H¯로 표시된다.

  S = N(워드) + K(비전) + 3

  Special Classification token(CLS)
  Special Separator token(SEP)
  Segment Embedding : 단어가 첫번째 문장에 속하는지 두번째 문장에 속하는지 알려준다.



MATERIALS AND METHODS

- **VL Pre-training Model**
  3) Joint Embedding

  v∈Rd 와 w∈Rd를 얻은 후 연결하여 joint 임베딩에 대한 입력 시퀀스를 구성

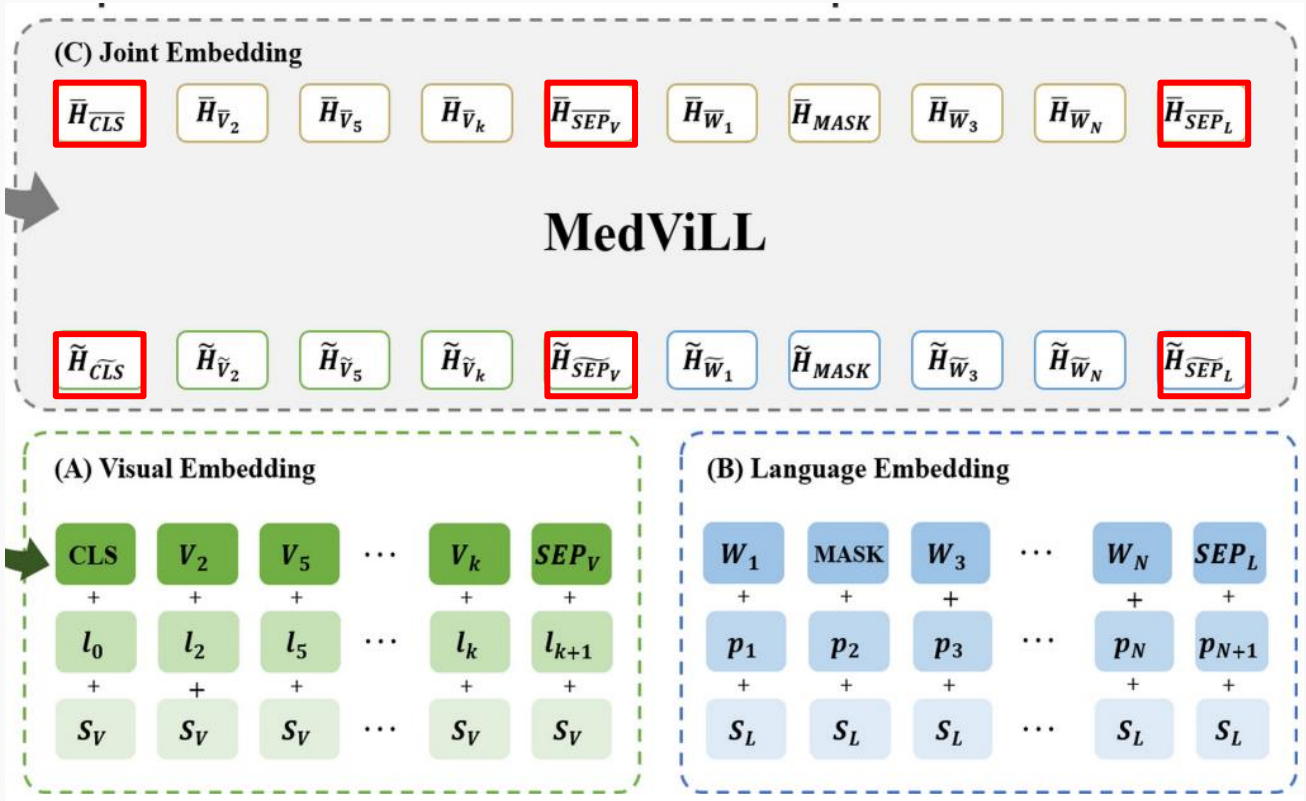
  특수 토큰 CLS와 SEP를 사용하여, 우리는 공동 임베딩 블록에 대한 입력을 H˜∈R^d×S로 정의.

  CLS, SEP_V, SEP_L은 특수 토큰을 임베딩하여 얻는다.

  공동 임베딩 블록에 의해 생성된 상황별 임베딩은 H˜로 표시된다.

  S = N(워드) + K(비전) + 3

  Special Classification token(CLS)
  Special Separator token(SEP)
  Segment Embedding : 단어가 첫번째 문장에 속하는지 두번째 문장에 속하는지 알려준다.



MATERIALS AND METHODS

- **VL Pre-training Model**
  4) Pre-training Objectives

  Image Report Matching (IRM) 작업은 주어진 이미지와 판독문(v, w) 쌍이 일치하는지 여부를 예측하도록 모델을 훈련한다.
  훈련 동안, 데이터 세트에서 일치하는 이미지-판독문 쌍과 일치하지 않는 쌍을 1:1 비율로 무작위로 샘플링한다.

  일치하는 쌍을 선택하는 것은 간단하지만 (X-ray영상에 해당 판독문이 함께 제공) 일치하지 않는 쌍은 “No findings.”, “Nothing noticeable.” 등 있기 때문에 샘플링이 어렵기 때문에 **일치하지 않는 이미지 판독문 쌍에 대해 샘플링할 때 진단 레이블을 사용**한다.

  IRM 작업에서, 불일치 판독문은 일치 판독문과 다른 positive 진단 레이블을 추출한 판독문으로 정의된다.

  joint contextualized embedding CLS는 입력 이미지와 판독문이 일치하는 쌍인지 여부를 분류하는 데 사용된다.
  다음과 같은 손실 함수를 사용한다,

$$L_{IRM}(\theta) = -\mathbb{E}_{(v,w)\sim D}\left[y \log P_\theta(v,w)\right] - \mathbb{E}_{(v,w')\sim D}\left[(1-y)\log(1-P_\theta(v,w'))\right] \quad (6)$$

  (v, w) = 일치하는 이미지 판독문 쌍
  (v, w’) = 일치하지 않는 쌍
  y = 레이블(일치하는 경우 1, 일치하지 않는 경우 0)
  E(v,w)~D는 훈련 세트 D에 대한 평균이고, Pθ(v,w)는 (v, w)가 쌍을 이루는 확률을 나타낸다.

MATERIALS AND METHODS

- **Self-Attention Mask Schemes**

  a) Bidirectional attention mask

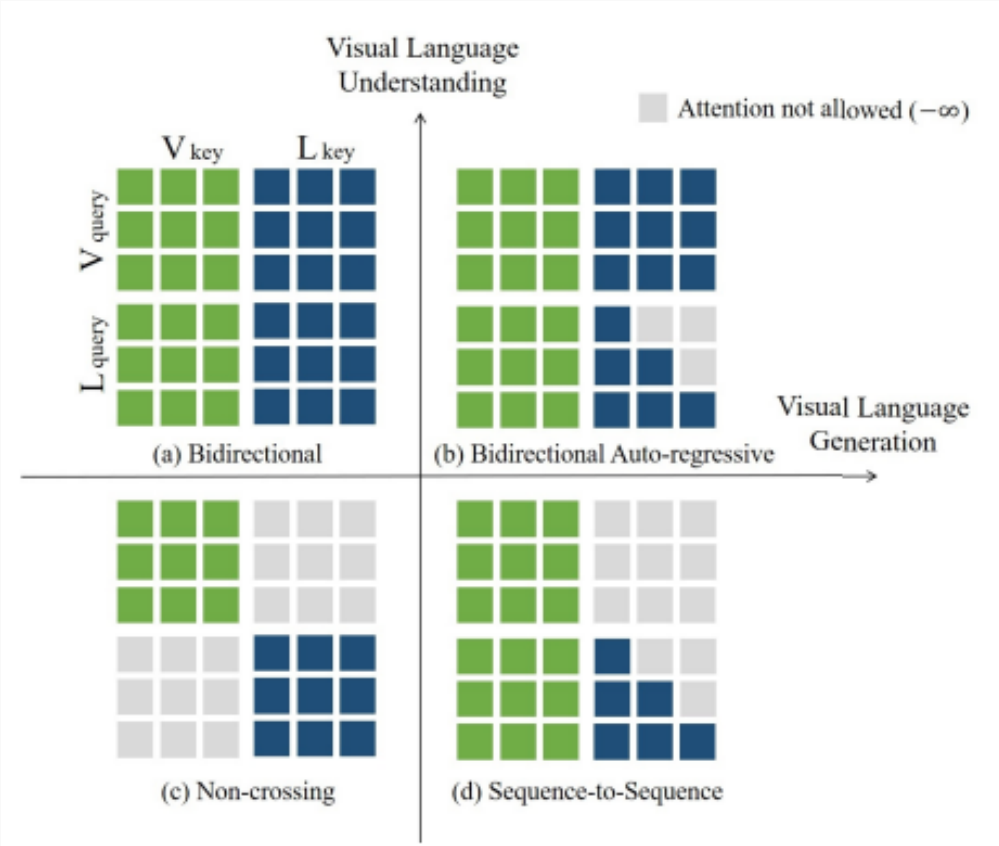
  Bidirectional attention mask (a)는 visual language modality 간의 unconstrained context learning(제약 없는 문맥 학습)을 위해 모든 입력이 자유롭게 상호 작용할 수 있도록 한다.

  d) Sequence-to-Sequence

  S2S(Sequence-to-Sequence) causal attention mask (d)는 restricted context learning(제한된 문맥 학습)을 허용한다. language features는 이전 단어에만 attention 할 수 있고 visual features은 미래의 정보 유출을 방지하기 위해 어떤 language features에도 attention 할 수 없다.

  a + d) Bi & S2S

  Bi & S2S는 pre-train 중에 Bi와 S2S 마스크를 번갈아 사용(모든 미니 배치에서 75% 확률로 S2S를 사용하고 25% 확률로 Bi를 사용)
  VLU와 VLG 다운스트림 작업을 모두 수행한다.



MATERIALS AND METHODS

- **Self-Attention Mask Schemes**

  a) Bidirectional attention mask
  Bidirectional attention mask (a)는 visual language modality 간의 unconstrained context learning(제약 없는 문맥 학습)을 위해 모든 입력이 자유롭게 상호 작용할 수 있도록 한다.

  d) Sequence-to-Sequence
  S2S(Sequence-to-Sequence) causal attention mask (d)는 restricted context learning(제한된 문맥 학습)을 허용한다. language features는 이전 단어에만 attention 할 수 있고 visual features은 미래의 정보 유출을 방지하기 위해 어떤 language features에도 attention 할 수 없다.

  a + d) Bi & S2S
  Bi & S2S는 pre-train 중에 Bi와 S2S 마스크를 번갈아 사용(모든 미니 배치에서 75% 확률로 S2S를 사용하고 25% 확률로 Bi를 사용)
  VLU와 VLG 다운스트림 작업을 모두 수행한다.

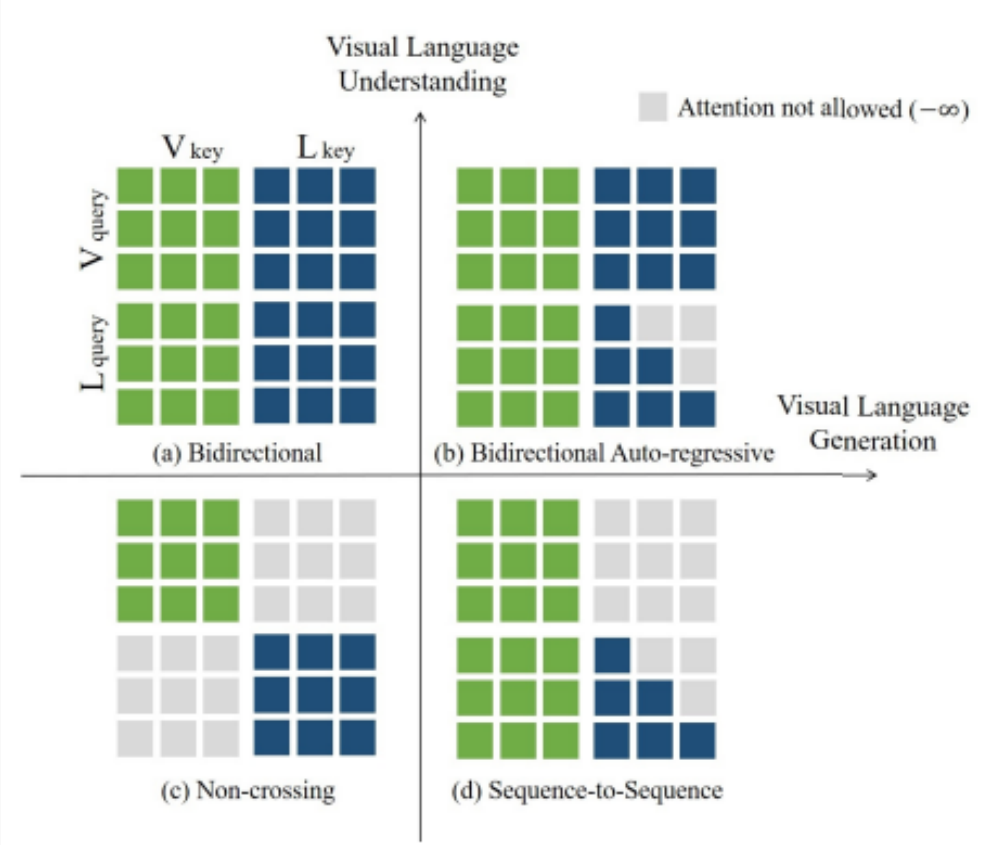
MATERIALS AND METHODS

- **Self-Attention Mask Schemes**

  b) Bidirectional Auto Regressive(BAR)
  Bi와 S2S 사이의 격차를 좁히고 두 가지 이점을 모두 활용하기 위해 Bidirectional Auto Regressive(BAR) (b)를 제안한다. BAR는 auto regressive language generation의 causal한 특성을 보존하면서 pre-train 중에 image feature을 language features와 혼합할 수 있도록 한다.

  self-attention mask $M \in RS \times S$, $S = N + K + 3$은 아래와 같이 0과 음의 무한대로 구성된다

  S = N(워드) + K(비전) + 3



MATERIALS AND METHODS

- **Self-Attention Mask Schemes**

  self-attention mask는 0과 음의 무한대로 구성된다
  S = N(워드) + K(비전) + 3

  self-attention 행렬은 식(9)에 따라 vision-language modality의 쿼리 및 핵심 벡터에서 계산되므로, 계산된 self-attention 행렬은 modality type별로 쿼리의 4개 하위 부분과 키 조합으로 나눌 수 있다.

  식 (10) = vision에서 q와 k의 attention
  식 (11) = vision에서 q와 language의 k 와의 attention mask
  식 (12) = language에서 q와 vision의 k의 attention mask
  식 (13) = language에서 q와 k의 주의 attention mask

  계산된 attention 값에 음수 무한대를 추가하면 소프트맥스 연산에서 0이 되기 때문에 SA 하위 부분에 대한 주의 마스크 행렬 M을 결합한다.
  식 14

$$M_{jk} = \begin{cases} 0, & \text{(attention allowed)} \\ -\infty, & \text{(attention not allowed)} \end{cases} \qquad j,k = 1,\ldots,S. \tag{7}$$

$$Attention = \text{softmax}\left(SA + M\right)V, \quad SA = \frac{QK^T}{\sqrt{d_k}} \tag{8}$$

$$SA = \begin{bmatrix} CLS_q \cdot CLS_k & \cdots & CLS_q \cdot W_{1k} & \cdots & CLS_q \cdot SEP_{Lk} \\ V_{1q} \cdot CLS_k & \cdots & V_{1q} \cdot W_{1k} & \cdots & V_{1q} \cdot SEP_{Lk} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ SEP_{Vq} \cdot CLS_k & \cdots & SEP_{Vq} \cdot W_{1k} & \cdots & SEP_{Vq} \cdot SEP_{Lk} \\ W_{1q} \cdot CLS_k & \cdots & W_{1q} \cdot W_{1k} & \cdots & W_{1q} \cdot SEP_{Lk} \\ W_{2q} \cdot CLS_k & \cdots & W_{2q} \cdot W_{1k} & \cdots & W_{2q} \cdot SEP_{Lk} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ W_{Kq} \cdot CLS_k & \cdots & W_{Kq} \cdot W_{Lk} & \cdots & W_{Kq} \cdot SEP_{Lk} \\ SEP_{Lq} \cdot CLS_k & \cdots & SEP_{Lq} \cdot W_{Lk} & \cdots & SEP_{Lq} \cdot SEP_{Lk} \end{bmatrix} \tag{9}$$

$$SA_{q,k} = SA_{CLS_q:SEP_{Vq},CLS_k:SEP_{Vk}} \tag{10}$$

$$+ SA_{CLS_q:SEP_{Vq},W_{1k}:SEP_{Lk}} \tag{11}$$

$$+ SA_{W_{1q}:SEP_{Lq},CLS_k:SEP_{Vk}} \tag{12}$$

$$+ SA_{W_{1q}:SEP_{Lq},W_{1k}:SEP_{Lk}} \tag{13}$$

MATERIALS AND METHODS

- **Self-Attention Mask Schemes**

  self-attention mask는 0과 음의 무한대로 구성된다
  S = N(워드) + K(비전) + 3

  self-attention 행렬은 식(9)에 따라 vision-language modality의 쿼리 및 핵심 벡터에서 계산되므로, 계산된 self-attention 행렬은 modality type별로 쿼리의 4개 하위 부분과 키 조합으로 나눌 수 있다.

  식 (10) = vision에서 q와 k의 attention
  식 (11) = vision에서 q와 language의 k 와의 attention mask
  식 (12) = language에서 q와 vision의 k의 attention mask
  식 (13) = language에서 q와 k의 주의 attention mask

  계산된 attention 값에 음수 무한대를 추가하면 소프트맥스 연산에서 0이 되기 때문에 SA 하위 부분에 대한 주의 마스크 행렬 M을 결합한다.
  식 14

$$BAR_M = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & -\infty & \cdots & -\infty \\ 0 & \cdots & 0 & \cdots & -\infty \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & -\infty \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{S \times S} \qquad (14)$$

MATERIALS AND METHODS

- **Self-Attention Mask Schemes**

  BAR attention mask(14)는 식 (13)을 제외한 나머지 가능한 조합의 attention 계산을 가능하게 한다.

  attention mask 체계는 autoregressive attention mask를 language modality에 적용하여 vision과 language modality 간의 joint 임베딩을 향상시키고 VLU, VLG 모두에서 잘 수행한다.

  noncrossing attention mask는 두 modality 간의 상호작용을 제한하기 때문에 language feature의 시작 부분에 CLS 토큰을 하나 더 추가하여 CLSV와 CLSL을 모두 IRM 사전 훈련 작업에 사용할 수 있다..

$$M_{jk} = \begin{cases} 0, & \text{(attention allowed)} \\ -\infty, & \text{(attention not allowed)} \end{cases} \qquad j,k = 1,\ldots,S. \tag{7}$$

$$Attention = \text{softmax}\left(SA + M\right)V, \quad SA = \frac{QK^T}{\sqrt{d_k}} \tag{8}$$

$$BAR_M = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & -\infty & \cdots & -\infty \\ 0 & \cdots & 0 & \cdots & -\infty \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & -\infty \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{S\times S} \tag{14}$$

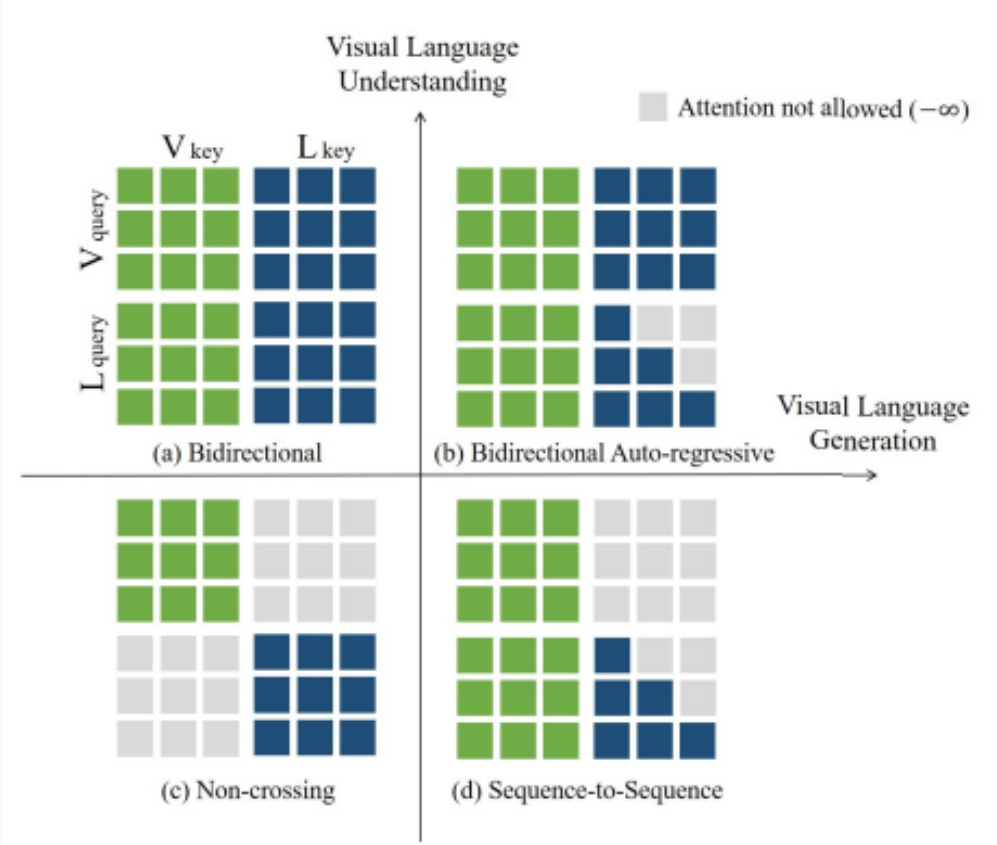
MATERIALS AND METHODS

- **Self-Attention Mask Schemes**

  BAR attention mask(14)는 식 (13)을 제외한 나머지 가능한 조합의 attention 계산을 가능하게 한다.

  attention mask 체계는 autoregressive attention mask를 language modality에 적용하여 vision과 language modality 간의 joint 임베딩을 향상시키고 VLU, VLG 모두에서 잘 수행한다.

  noncrossing attention mask는 두 modality 간의 상호작용을 제한하기 때문에 language feature의 시작 부분에 CLS 토큰을 하나 더 추가하여 CLSV와 CLSL을 모두 IRM 사전 훈련 작업에 사용할 수 있다.



MATERIALS AND METHODS

- **Self-Attention Mask Schemes**

  참고:

  Causality Masking
  https://velog.io/@crosstar1228/NLPTransformer-Attention-is-all-you-need-%EC%83%85%EC%83%85%EC%9D%B4-%ED%8C%8C%ED%97%A4%EC%B9%98%EA%B8%B0

  셀프어텐션
  https://ratsgo.github.io/nlpbook/docs/language_model/tr_self_attention/

RESULTS AND DISCUSSION

A. Dataset Analysis

MIMIC-CXR과 Open-I는 별도의 기관에서 수집되었기 때문에 서로 다른 특성을 가질 수 있다.

특히 진단 정보는 서로 다르게 분포될 수 있습니다.따라서 두 데이터 세트 간 진단 레이블 분포의 차이를 분석하기 위해 Chexpert 레이블러 결과에서 획득한 양의 레이블을 비교했다.

데이터 전반적으로 약간의 불균형이 관찰되었다 MIMIC-CXR에서 클래스 비율은 13.39%(support devices)에서 1.2%(pneumonia와 pleur, alother))까지 다양했다.

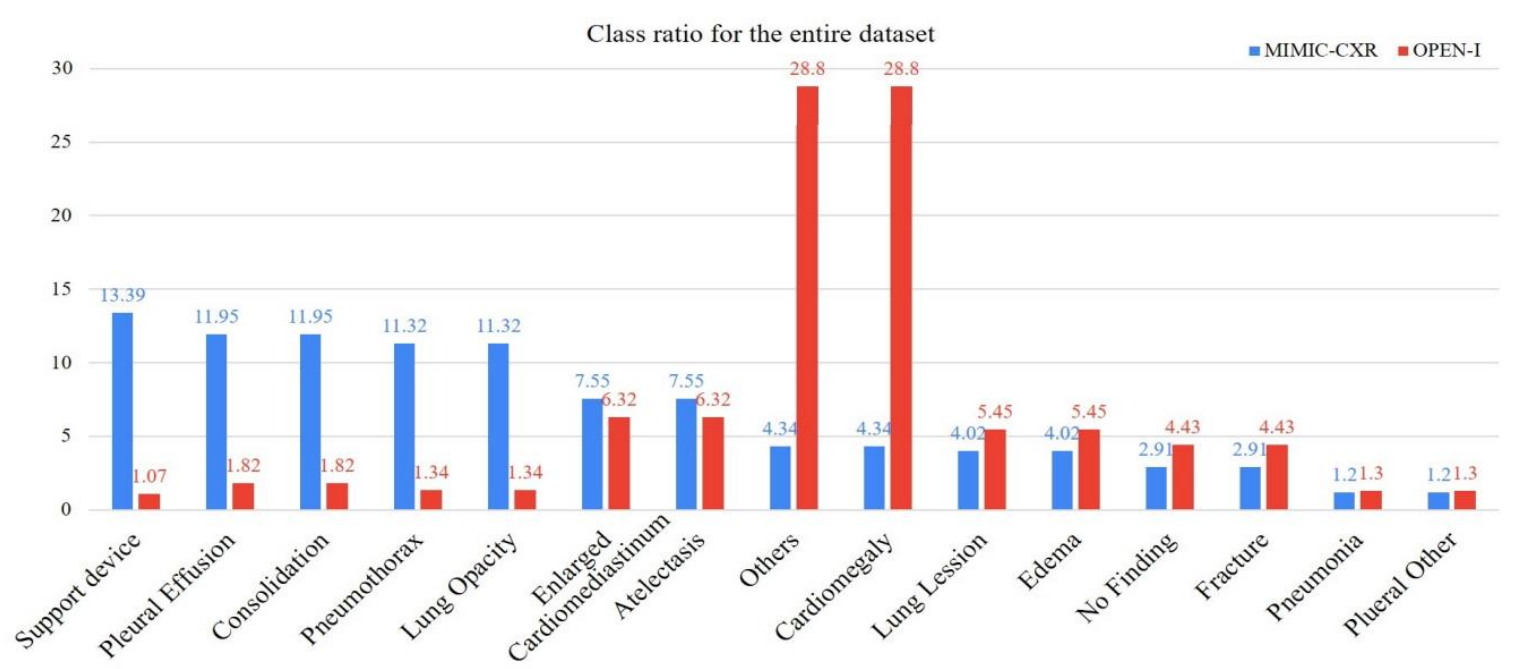
MIMIC-CXR에 비해 Open-I에서 최대 등급 비율이 28.8%(기타 및 심혈관계), 최소 1.07%(지원 장치)로 심각한 불균형 Open-I가 데이터 볼륨뿐만 아니라 임상 특성 측면에서도 MIMIC-CXR과 다르다는 것을 보여준다.

따라서,Open-I에서 MIMIC-CXR-사전 훈련된 모델을 평가하는 것이 모델의 일반화 능력을 테스트하기 위한 적절한 설정이라고 믿는다

RESULTS AND DISCUSSION

A. Dataset Analysis

전체 데이터 세트에 대한 진단 레이블의 분포를 비교. 두 데이터 세트의 척도가 다르기 때문에 각 레이블은 전체 데이터 세트에 대한 백분율로 표시



RESULTS AND DISCUSSION

B. Implementation details

ImageNet에서 pre-train된 ResNet-50을 visual feature extractor로 사용한다.

input 이미지 크기는 (512x512x3)이며 ResNet-50의 마지막 feature 맵(16x16x2048)은 spatial 차원에서 flatten하고
pre-train때는 180개의 visual feature(180x2048)을 랜덤 샘플링
반면 모든 다운스트림 작업에 모든 feature(256x2048)을 사용한다.

text token을 임베딩 하기위해 판독문의 각 문장의 최대 임베딩 크기를 고려하여 길이가 253 token으로 자르거나 패딩한다.

token 임베딩을 위해 12개의 트랜스포머 계층으로 구성된 BERT 기반 아키텍처를 채택한다.

각 레이어에는 12개의 attention heads, 768 임베딩 hidden size 및 0.1의 drop out
visual backbone 및 Transformer에 대한 learning rate 1e-5 설정이 있는 AdamW 최적화 도구를 채택한다.

모든 모델은 batch size 128 및 50 epochs 의 RTX-3090 GPU 8개에서 훈련되었다

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

1) Diagnosis Classification

이미지 판독문 쌍에 대해, Chexpert 라벨러에 의해 판독문에서 추출된 positive 라벨을 Diagnosis 라벨로 사용한다.

단일 쌍은 최대 14개의 진단 레이블을 가질 수 있으므로(즉, multi-label classification) CLS 위에 14개의 linear heads 를 사용하고 binary cross entropy loss을 사용하여 모델을 fine-tune한다.

micro average AUROC 및 F1 점수로 평가

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

1) Diagnosis Classification

TABLE II: Model AUROC and F1 scores for the diagnosis classification task on MIMIC-CXR and Open-I. Inference time(ms) on MIMIC-CXR: MedViLL(12.5), Bi&S2S (13), Bi (13), S2S (13), Non-crossing (12.5), Fine-tuning Only (15.5), CNN & Transformer (10.5).

Dataset	Metrics	MedViLL	Bi&S2S	Bi	S2S	Non-crossing	Fine-tuning Only	CNN & Transformer
MIMIC-CXR	avg AUROC	0.980 (0.00)	0.979 (0.00)	**0.984 (0.00)**	0.982 (0.00)	0.980 (0.00)	0.969 (0.00)	0.831 (0.00)
	avg F1	0.839 (0.00)	0.846 (0.00)	**0.852 (0.00)**	0.846 (0.00)	0.824 (0.00)	0.807 (0.00)	0.491 (0.00)
	p-value (avg AUROC)	-	0.005	1.97E-15	0.003	0.254	1.70E-36	3.41E-102
	p-value (avg F1)	-	9.59E-28	7.85E-42	1.62E-26	2.02E-43	4.90E-63	2.70E-122
Open-I	avg AUROC	**0.892 (0.00)**	0.827 (0.00)	0.758 (0.00)	0.720 (0.00)	0.589 (0.00)	0.723 (0.00)	0.709 (0.00)
	avg F1	**0.407 (0.01)**	0.301 (0.01)	0.295 (0.01)	0.256 (0.01)	0.185 (0.00)	0.300(0.00)	0.245 (0.01)
	p-value (avg AUROC)	-	6.94E-83	4.00E-98	1.23E-101	4.66E-122	1.41E-103	1.35E-109
	p-value (avg F1)	-	1.05E-93	7.04E-95	7.17E-101	2.08E-110	6.49E-94	2.69E-104

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

2) Medical Image-Report Retrieval

image-to-report (I2R) 검색은 모델이 이미지가 주어진 대규모 판독문에서 가장 관련성이 높은 보고서를 검색
Report-to-Image(R2I) 검색의 경우 그 반대이다.

이미지 제공되면 원래 일치하는 판독문과 동일한 Cexpert 진단 레이블이 포함된 리포트는 이미지 판독문 쌍으로 간주

최종 multi-modal representation CLS는 주어진 쌍을 분류하기 위해 이진 분류기의 입력으로 사용된다.

inference 시, 각 시행에서 모델은 100개의 이미지 판독문 쌍이 주어지며, 가능한 한 positive 쌍의 순위를 매기기 위해 예측된 점수를 사용해야 한다. 평가 지표는 Hit@K, Recall@K, Precision@K(K = 5) 및 mean reciprocal rank(MRR)입니다

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

2) Medical Image-Report Retrieval

TABLE III: Medical Image-Report Retrieval performance on MIMIC-CXR and Open-I. Inference time(ms) of Report-to-Image and Image-to-Report on MIMIC-CXR: MedViLL(7.6, 7.8), Bi&S2S (8.2, 7.8), Bi (7.6, 7.6), S2S (7.6, 7.7), Non-crossing (7.7, 7.8), Fine-tuning Only (7.8, 7.6), CNN & Transformer (5.3, 5.3).

Task	Models	MIMIC-CXR					OpenI				
		MRR	H@5	R@5	P@5	p-value	MRR	H@5	R@5	P@5	p-value
Report-to-Image	MedViLL	56.5(0.01)	77.0(0.01)	47.4(0.01)	19.9(0.00)	-	51.3(0.01)	73.0(0.01)	12.9(0.00)	31.7(0.00)	-
	Bi&S2S	55.5(0.01)	76.7(0.01)	46.7(0.01)	19.7(0.00)	1.20E-05	46.4(0.01)	68.1(0.01)	10.5(0.00)	28.8(0.01)	3.71E-27
	Bi	58.0(0.01)	78.2(0.01)	48.2(0.01)	20.2(0.00)	1.60E-10	**51.4(0.01)**	**74.8(0.01)**	**13.3(0.00)**	32.0(0.01)	0.843
	S2S	**58.8(0.01)**	**79.1(0.01)**	**48.9(0.01)**	**20.3(0.00)**	1.89E-18	48.6(0.01)	67.2(0.01)	10.3(0.01)	**32.9(0.01)**	2.28E-14
	Non-crossing	54.7(0.01)	77.0(0.01)	47.2(0.01)	19.5(0.00)	4.07E-12	48.6(0.01)	68.4(0.01)	11.2(0.00)	31.1(0.01)	3.88E-18
	Fine-tuning Only	41.8(0.01)	61.6(0.01)	35.8(0.01)	15.8(0.00)	3.14E-53	36.9(0.01)	54.4(0.01)	5.4(0.00)	20.7(0.01)	1.72E-53
	CNN & Transformer	11.4(0.01)	15.2(0.02)	5.1(0.00)	3.6(0.01)	9.43E-71	36.2(0.04)	56.6(0.04)	5.0(0.00)	21.4(0.04)	4.94E-19
Image-to-Report	MedViLL	55.8 (0.01)	75.5(0.01)	47.1(0.01)	19.7(0.00)	-	**50.4(0.01)**	63.8(0.01)	12.9(0.00)	35.5(0.01)	-
	Bi&S2S	54.5(0.01)	75.5(0.01)	47.8(0.01)	19.9(0.00)	6.32E-08	45.8(0.01)	54.0(0.01)	10.1(0.00)	35.8(0.00)	8.55E-29
	Bi	56.7(0.01)	76.3(0.01)	47.6(0.01)	20.2(0.00)	0.0002	48.5(0.01)	**65.8(0.01)**	**13.7(0.00)**	32.3(0.01)	3.17E-12
	S2S	**57.9(0.01)**	**78.5(0.01)**	**49.7(0.01)**	**20.7(0.00)**	2.72E-13	45.4(0.01)	53.6(0.01)	8.9(0.00)	**36.9(0.00)**	6.84E-31
	Non-crossing	54.6(0.01)	75.7(0.01)	47.6(0.01)	20.0(0.00)	3.84E-07	42.6(0.01)	61.2(0.01)	11.0(0.00)	28.0(0.01)	1.15E-40
	Fine-tuning Only	41.4(0.01)	60.8(0.01)	36.3(0.01)	15.7(0.00)	5.56E-56	45.2(0.00)	49.7(0.01)	5.1(0.00)	35.0(0.00)	2.64E-29
	CNN & Transformer	12.0(0.02)	15.3(0.02)	5.1(0.00)	4.0(0.01)	1.09E-52	37.9(0.06)	54.0(0.06)	5.0(0.00)	23.0(0.06)	1.16E-12

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

3) Medical Visual Question Answering

VQA-RAD(Visual Question Answering in Radiology) 데이터 세트에서 VQA를 수행

315개의 이미지(1104개의 두부 CT 또는 MRI, 107개 흉부 X선 및 104개 복부 CT))에 대한 3,515개의 질문-답변 쌍이 포함되어 있다

모델은 흉부 X선 이미지에 대해 사전 훈련을 받았기 때문에, VQA-RAD는 사전 훈련된 모델이 단일 이미지 도메인을 넘어 일반화될 수 있는지 여부를 연구할 수 기회를 제공한다.

이미지와 free-text question이 주어지면, final representation CLS을 이용 one-hot encoding 답변을 예측한다 (가능한 모든 응답은 단일 토큰으로 처리됨).

성능은 VQA-RAD 논문에 따라 closed questions (예/아니오 등의 짧은 형식의 답변)과 open-ended questions(즉, 긴 형식의 답변)에 대해 별도로 평가되었다

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

3) Medical Visual Question Answering

TABLE IV: Model accuracy on the VQA-RAD dataset. O.E. stands for Open-ended question and C.E. stands for close-ended question. For MEVF [27], we used the reported results from the original paper. Inference time(ms) on MIMIC-CXR: MedViLL(19.46), Bi&S2S(19.52), Bi(19.43), S2S(19.51), Non-crossing(19.58), Fine-tuning Only(19.61), CNN & Transformer(17.42).

Models	ALL				CHEST			
	O.E.	C.E.	p-value of O.E.	p-value of C.E.	O.E.	C.E.	p-value of O.E.	p-value of C.E.
MedViLL	**0.595(0.032)**	0.777(0.071)	-	-	**0.587(0.033)**	**0.782(0.123)**	-	-
Bi&S2S	0.541(0.038)	0.76(0.027)	2.93E-07	0.224	0.566(0.074)	0.766(0.035)	0.164	0.519
Bi	0.58(0.038)	**0.784(0.03)**	0.124	0.643	0.562(0.04)	0.767(0.035)	0.013	0.549
S2S	0.505(0.042)	0.73(0.025)	1.81E-12	0.002	0.517(0.07)	0.723(0.048)	1.57E-05	0.021
Non-crossing	0.531(0.015)	0.734(0.017)	5.58E-12	0.003	0.474(0.083)	0.732(0.03)	3.94E-08	0.043
Fine-tuning Only	0.232(0.019)	0.649(0.026)	2.98E-43	5.38E-11	0.124(0.014)	0.606(0.035)	1.08E-42	1.50E-08
CNN & Transformer	0.24(0.029)	0.667(0.015)	7.66E-46	2.70E-9	0.124(0.067)	0.523(0.033)	7.60E-32	1.62E-12
MEVF [27]	0.407	0.741	-	-	-	-	-	-

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

4) Radiology Report Generation

fine-tuning process는 모든 모델에 대해 self-attention mask를 S2S로 수정한다는 점을 제외하고 Masked Language Modeling(MLM) 사전 훈련 작업과 동일.

inference 시, MASK 토큰을 순차적으로 복구하여 판독문을 생성할 수 있다.

visual features 다음에 단일 MASK 토큰이 오면 모델은 첫 번째 language token 을 예측할 수 있다. 그런 다음 첫 번째 MASK를 샘플링된 토큰으로 대체하고 새 MASK 토큰이 추가됩니다. 이 과정을 SEP 토큰을 stop sign으로 예측할 때까지 반복

성능은 Clinical efficacy(임상 효과), Perplexity(복잡성) 및 BLEU(Bilingual Evaluation Understudy) 점수의 세 가지 메트릭으로 측정된다.

임상 효과는 원래 보고서와 생성된 보고서 모두에 Chexpert 라벨을 적용하여 얻는다. 추출된 레이블을 기반으로 accuracy, precision, recall, 그리고 F1을 계산

복잡성은 모델의 언어적 유창성을 평가하고 임상 효과는 모델이 주어진 이미지의 의미를 포착할 수 있는지 평가하는 데 사용된다.

또한 생성된 판독문이 reference 보고서와 얼마나 유사한지 평가하기 위해 4그램 BLEU 점수를 계산한다.

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

4) Radiology Report Generation

TABLE V: Report generation performance in terms of Perplexity and Label Accuracy, Precision Recall and F1 and BLEU4. Inference time(ms) on MIMIC-CXR: MedViLL(32.81), Bi&S2S(33.55), Bi(32.54), S2S(33.04), Non-crossing(32.71), Fine-tuning Only(32.92).

Dataset	Models	Perplexity (↓)	Accuracy (↑)	Precision (↑)	Recall (↑)	F1 Score (↑)	BLEU4 (↑)	p-value
MIMIC	MedViLL	4.185(0.022)	**0.841(0.003)**	**0.698(0.002)**	**0.559(0.004)**	**0.621(0.002)**	0.066(0.001)	-
	Bi&S2S	6.515(0.12)	0.786(0.007)	0.619(0.003)	0.435(0.009)	0.511(0.006)	0.066(0.001)	4.17E-43
	Bi	849.67(5.225)	0.637(0.004)	0.283(0.007)	0.07(0.024)	0.11(0.032)	0.015(0.004)	1.11E-36
	S2S	4.258(0.069)	0.797(0.007)	0.662(0.004)	0.448(0.01)	0.534(0.007)	0.043(0.001)	2.32E-38
	Non-crossing	718.122(9.484)	0.634(0.005)	0.277(0.013)	0.076(0.004)	0.12(0.005)	0.007(0.001)	2.30E-75
	Fine-tuning Only	224.343(0.204)	0.664(0.003)	0.417(0.012)	0.305(0.006)	0.352(0.005)	0.009(0.004)	4.14E-65
	TieNet	**4.132(0.033)**	0.687(0.003)	0.487(0.003)	0.380(0.006)	0.426(0.006)	**0.123(0.002)**	7.17E-54
Open-I	MedViLL	5.637(0.259)	0.734(0.001)	0.512(0.002)	0.594(0.001)	0.55(0.001)	0.049(0.001)	-
	Bi&S2S	15.97(1.071)	0.712(0.003)	0.497(0.003)	0.369(0.006)	0.423(0.004)	0.024(0.01)	4.02E-52
	Bi	787.66(55.492)	0.686(0.004)	0.356(0.025)	0.103(0.006)	0.16(0.008)	0.015(0.004)	2.14E-52
	S2S	**4.732(0.537)**	**0.736(0.003)**	**0.517(0.002)**	0.538(0.004)	0.527(0.002)	0.043(0.002)	1.76E-44
	Non-crossing	217.27(12.139)	0.693(0.003)	0.337(0.025)	0.085(0.005)	0.135(0.007)	0.002(0.001)	1.46E-57
	Fine-tuning Only	292.60(19.858)	0.684(0.003)	0.291(0.023)	0.073(0.035)	0.112(0.047)	0.006(0.002)	7.02E-30
	TieNet	7.901(0.483)	0.732(0.007)	**0.517(0.013)**	**0.610(0.017)**	**0.553(0.013)**	**0.189(0.005)**	0.2181

RESULTS AND DISCUSSION

D. Downtream Task Result

MedViLL은 서로 다른 attention masks 로 훈련된 네 가지 사전 훈련 모델과 비교.

두 가지 기준을 더 포함

1) MedViLL과 동일한 모델 아키텍처를 pre-train 없이 각 다운스트림 작업에서 Fine-tuning Only.

2) CNN & Transformer, 이미지 인코딩에만 CNN 사용, 판독문 인코딩에는 Transformer 모듈(MedViLL과 동일한 크기)을 사용, 각 출력은 다운스트림 작업에 사용, CNN & Transformer도 pre-train하지 않는다.

모든 작업에 대해 30개의 무작위 multiple-bootstrap 실험을 수행하고 평균 성능과 표준 편차를 보고한다.

델에 대해 얻은 다양한 메트릭의 평균 값을 기반으로 중요도 값 0.05로 독립적인 t-테스트를 수행

요약하면, MedViLL은 다음과 같이 최고 또는 차선의 성능을 달성했습니다

VLU 및 VLG 작업의 다양한 기준 모델로 통계적 유의성을 분석한다.또한 MedViLL은 도메인 외부 평가에서 대부분의 모델을 능가하여 우수한 일반화 능력을 보여준다

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

1) Diagnosis Classification

TABLE II: Model AUROC and F1 scores for the diagnosis classification task on MIMIC-CXR and Open-I. Inference time(ms) on MIMIC-CXR: MedViLL(12.5), Bi&S2S (13), Bi (13), S2S (13), Non-crossing (12.5), Fine-tuning Only (15.5), CNN & Transformer (10.5).

Dataset	Metrics	MedViLL	Bi&S2S	Bi	S2S	Non-crossing	Fine-tuning Only	CNN & Transformer
MIMIC-CXR	avg AUROC	0.980 (0.00)	0.979 (0.00)	**0.984 (0.00)**	0.982 (0.00)	0.980 (0.00)	0.969 (0.00)	0.831 (0.00)
	avg F1	0.839 (0.00)	0.846 (0.00)	**0.852 (0.00)**	0.846 (0.00)	0.824 (0.00)	0.807 (0.00)	0.491 (0.00)
	p-value (avg AUROC)	-	0.005	1.97E-15	0.003	0.254	1.70E-36	3.41E-102
	p-value (avg F1)	-	9.59E-28	7.85E-42	1.62E-26	2.02E-43	4.90E-63	2.70E-122
Open-I	avg AUROC	**0.892 (0.00)**	0.827 (0.00)	0.758 (0.00)	0.720 (0.00)	0.589 (0.00)	0.723 (0.00)	0.709 (0.00)
	avg F1	**0.407 (0.01)**	0.301 (0.01)	0.295 (0.01)	0.256 (0.01)	0.185 (0.00)	0.300(0.00)	0.245 (0.01)
	p-value (avg AUROC)	-	6.94E-83	4.00E-98	1.23E-101	4.66E-122	1.41E-103	1.35E-109
	p-value (avg F1)	-	1.05E-93	7.04E-95	7.17E-101	2.08E-110	6.49E-94	2.69E-104

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

2) Medical Image-Report Retrieval

TABLE III: Medical Image-Report Retrieval performance on MIMIC-CXR and Open-I. Inference time(ms) of Report-to-Image and Image-to-Report on MIMIC-CXR: MedViLL(7.6, 7.8), Bi&S2S (8.2, 7.8), Bi (7.6, 7.6), S2S (7.6, 7.7), Non-crossing (7.7, 7.8), Fine-tuning Only (7.8, 7.6), CNN & Transformer (5.3, 5.3).

Task	Models	MIMIC-CXR					OpenI				
		MRR	H@5	R@5	P@5	p-value	MRR	H@5	R@5	P@5	p-value
Report-to-Image	MedViLL	56.5(0.01)	77.0(0.01)	47.4(0.01)	19.9(0.00)	-	51.3(0.01)	73.0(0.01)	12.9(0.00)	31.7(0.00)	-
	Bi&S2S	55.5(0.01)	76.7(0.01)	46.7(0.01)	19.7(0.00)	1.20E-05	46.4(0.01)	68.1(0.01)	10.5(0.00)	28.8(0.01)	3.71E-27
	Bi	58.0(0.01)	78.2(0.01)	48.2(0.01)	20.2(0.00)	1.60E-10	**51.4(0.01)**	**74.8(0.01)**	**13.3(0.00)**	32.0(0.01)	0.843
	S2S	**58.8(0.01)**	**79.1(0.01)**	**48.9(0.01)**	**20.3(0.00)**	1.89E-18	48.6(0.01)	67.2(0.01)	10.3(0.01)	**32.9(0.01)**	2.28E-14
	Non-crossing	54.7(0.01)	77.0(0.01)	47.2(0.01)	19.5(0.00)	4.07E-12	48.6(0.01)	68.4(0.01)	11.2(0.00)	31.1(0.01)	3.88E-18
	Fine-tuning Only	41.8(0.01)	61.6(0.01)	35.8(0.01)	15.8(0.00)	3.14E-53	36.9(0.01)	54.4(0.01)	5.4(0.00)	20.7(0.01)	1.72E-53
	CNN & Transformer	11.4(0.01)	15.2(0.02)	5.1(0.00)	3.6(0.01)	9.43E-71	36.2(0.04)	56.6(0.04)	5.0(0.00)	21.4(0.04)	4.94E-19
Image-to-Report	MedViLL	55.8 (0.01)	75.5(0.01)	47.1(0.01)	19.7(0.00)	-	**50.4(0.01)**	63.8(0.01)	12.9(0.00)	35.5(0.01)	-
	Bi&S2S	54.5(0.01)	75.5(0.01)	47.8(0.01)	19.9(0.00)	6.32E-08	45.8(0.01)	54.0(0.01)	10.1(0.00)	35.8(0.00)	8.55E-29
	Bi	56.7(0.01)	76.3(0.01)	47.6(0.01)	20.2(0.00)	0.0002	48.5(0.01)	**65.8(0.01)**	**13.7(0.00)**	32.3(0.01)	3.17E-12
	S2S	**57.9(0.01)**	**78.5(0.01)**	**49.7(0.01)**	**20.7(0.00)**	2.72E-13	45.4(0.01)	53.6(0.01)	8.9(0.00)	**36.9(0.00)**	6.84E-31
	Non-crossing	54.6(0.01)	75.7(0.01)	47.6(0.01)	20.0(0.00)	3.84E-07	42.6(0.01)	61.2(0.01)	11.0(0.00)	28.0(0.01)	1.15E-40
	Fine-tuning Only	41.4(0.01)	60.8(0.01)	36.3(0.01)	15.7(0.00)	5.56E-56	45.2(0.00)	49.7(0.01)	5.1(0.00)	35.0(0.00)	2.64E-29
	CNN & Transformer	12.0(0.02)	15.3(0.02)	5.1(0.00)	4.0(0.01)	1.09E-52	37.9(0.06)	54.0(0.06)	5.0(0.00)	23.0(0.06)	1.16E-12

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

3) Medical Visual Question Answering

TABLE IV: Model accuracy on the VQA-RAD dataset. O.E. stands for Open-ended question and C.E. stands for close-ended question. For MEVF [27], we used the reported results from the original paper. Inference time(ms) on MIMIC-CXR: MedViLL(19.46), Bi&S2S(19.52), Bi(19.43), S2S(19.51), Non-crossing(19.58), Fine-tuning Only(19.61), CNN & Transformer(17.42).

Models	ALL				CHEST			
	O.E.	C.E.	p-value of O.E.	p-value of C.E.	O.E.	C.E.	p-value of O.E.	p-value of C.E.
MedViLL	**0.595(0.032)**	0.777(0.071)	-	-	**0.587(0.033)**	**0.782(0.123)**	-	-
Bi&S2S	0.541(0.038)	0.76(0.027)	2.93E-07	0.224	0.566(0.074)	0.766(0.035)	0.164	0.519
Bi	0.58(0.038)	**0.784(0.03)**	0.124	0.643	0.562(0.04)	0.767(0.035)	0.013	0.549
S2S	0.505(0.042)	0.73(0.025)	1.81E-12	0.002	0.517(0.07)	0.723(0.048)	1.57E-05	0.021
Non-crossing	0.531(0.015)	0.734(0.017)	5.58E-12	0.003	0.474(0.083)	0.732(0.03)	3.94E-08	0.043
Fine-tuning Only	0.232(0.019)	0.649(0.026)	2.98E-43	5.38E-11	0.124(0.014)	0.606(0.035)	1.08E-42	1.50E-08
CNN & Transformer	0.24(0.029)	0.667(0.015)	7.66E-46	2.70E-9	0.124(0.067)	0.523(0.033)	7.60E-32	1.62E-12
MEVF [27]	0.407	0.741	-	-	-	-	-	-

RESULTS AND DISCUSSION

C. Task-specific Downstream Model Strategy

4) Radiology Report Generation

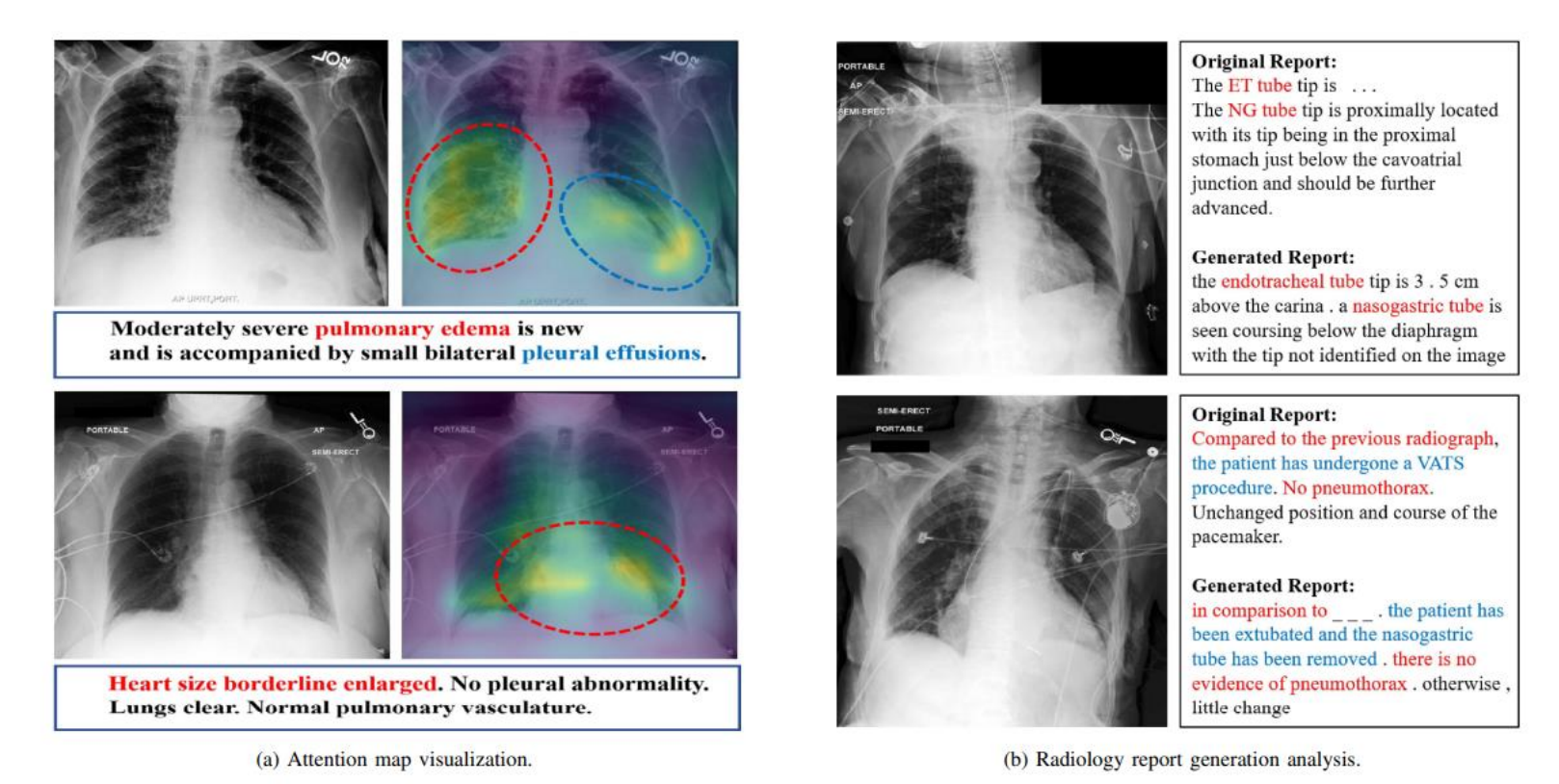
TABLE V: Report generation performance in terms of Perplexity and Label Accuracy, Precision Recall and F1 and BLEU4. Inference time(ms) on MIMIC-CXR: MedViLL(32.81), Bi&S2S(33.55), Bi(32.54), S2S(33.04), Non-crossing(32.71), Fine-tuning Only(32.92).

Dataset	Models	Perplexity (↓)	Accuracy (↑)	Precision (↑)	Recall (↑)	F1 Score (↑)	BLEU4 (↑)	p-value
MIMIC	MedViLL	4.185(0.022)	**0.841(0.003)**	**0.698(0.002)**	**0.559(0.004)**	**0.621(0.002)**	0.066(0.001)	-
	Bi&S2S	6.515(0.12)	0.786(0.007)	0.619(0.003)	0.435(0.009)	0.511(0.006)	0.066(0.001)	4.17E-43
	Bi	849.67(5.225)	0.637(0.004)	0.283(0.007)	0.07(0.024)	0.11(0.032)	0.015(0.004)	1.11E-36
	S2S	4.258(0.069)	0.797(0.007)	0.662(0.004)	0.448(0.01)	0.534(0.007)	0.043(0.001)	2.32E-38
	Non-crossing	718.122(9.484)	0.634(0.005)	0.277(0.013)	0.076(0.004)	0.12(0.005)	0.007(0.001)	2.30E-75
	Fine-tuning Only	224.343(0.204)	0.664(0.003)	0.417(0.012)	0.305(0.006)	0.352(0.005)	0.009(0.004)	4.14E-65
	TieNet	**4.132(0.033)**	0.687(0.003)	0.487(0.003)	0.380(0.006)	0.426(0.006)	**0.123(0.002)**	7.17E-54
Open-I	MedViLL	5.637(0.259)	0.734(0.001)	0.512(0.002)	0.594(0.001)	0.55(0.001)	0.049(0.001)	-
	Bi&S2S	15.97(1.071)	0.712(0.003)	0.497(0.003)	0.369(0.006)	0.423(0.004)	0.024(0.01)	4.02E-52
	Bi	787.66(55.492)	0.686(0.004)	0.356(0.025)	0.103(0.006)	0.16(0.008)	0.015(0.004)	2.14E-52
	S2S	**4.732(0.537)**	**0.736(0.003)**	**0.517(0.002)**	0.538(0.004)	0.527(0.002)	0.043(0.002)	1.76E-44
	Non-crossing	217.27(12.139)	0.693(0.003)	0.337(0.025)	0.085(0.005)	0.135(0.007)	0.002(0.001)	1.46E-57
	Fine-tuning Only	292.60(19.858)	0.684(0.003)	0.291(0.023)	0.073(0.035)	0.112(0.047)	0.006(0.002)	7.02E-30
	TieNet	7.901(0.483)	0.732(0.007)	**0.517(0.013)**	**0.610(0.017)**	**0.553(0.013)**	**0.189(0.005)**	0.2181

RESULTS AND DISCUSSION

E. Qualitative results and analysis

우리는 MedViLL(a)에서 추출한 주의 영역을 시각화한다. 또한 생성된 리포트를 동일한 흉부 X선 영상에 대한 원본 리포트와 비교합니다(b)



(a) Attention map visualization.

(b) Radiology report generation analysis.

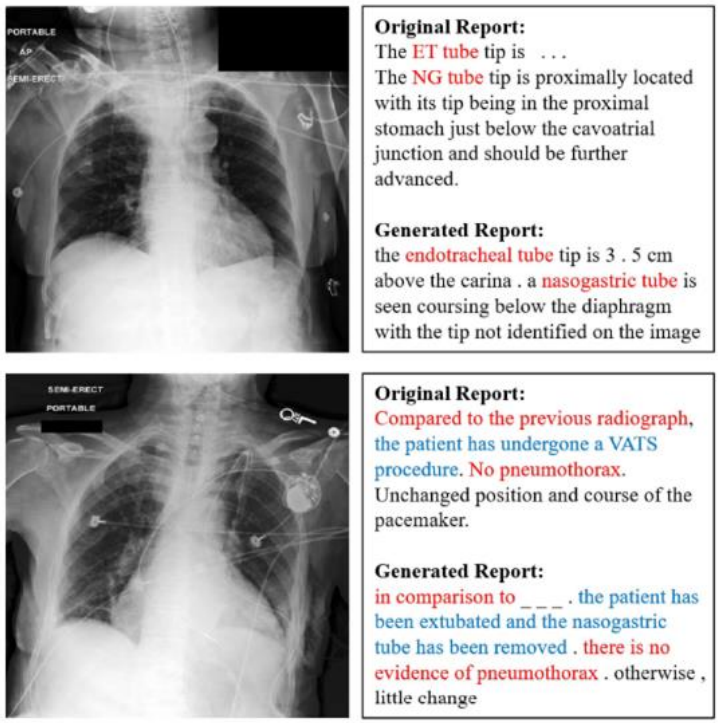
RESULTS AND DISCUSSION

E. Qualitative results and analysis

생성된 보고서는 VATS의 일부인 절제술과 비위관 제거를 설명한다.
이는 BLEU 점수가 특히 의료 영역에서보고서 생성을 평가하는 데 적합한 척도가 아님을 나타낸다.

(i.e. “ET tube”, “NG tube”)

(i.e. “endotracheal tube”,“nasogastric tube”)

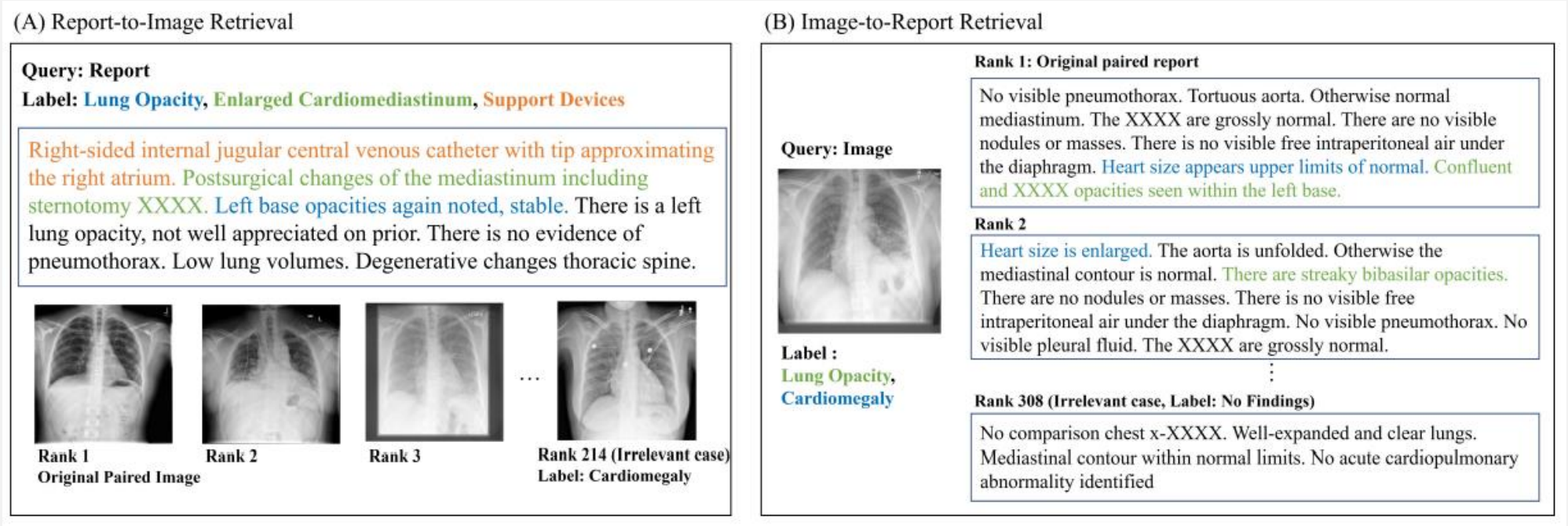


(b) Radiology report generation analysis.

RESULTS AND DISCUSSION

E. Qualitative results and analysis

판독문 또는 이미지를 쿼리로 지정하면 검색된 이미지 또는 보고서를 (a)와 (b)에 각각 표시한다.(a)에서, 주어진 보고서는 Cexpert 라벨러에 의해 진단 라벨 Lung Opacity(파란색), Envidated Cardiomediastinum(녹색) 및 Support Devices(오렌지색)로 주석이 달렸다. 상위 3개 이미지는 또한 지정된 쿼리와 동일한 레이블을 포함하는 반면, 마지막 이미지(순위 214)는 지정된 쿼리와 관련이 없다. (B)에서 상위 3개 리포트에는 Cexpert 레이블, Lung Opacity(녹색) 및 Cardiomegaly(파란색)에서 인식하는 동일한 레이블도 포함됩니다. 마지막 보고서(순위 308)에는 레이블이 없으며 지정된 쿼리 이미지와 관련이 없습니다



CONCLUSION

- VLU VLG의 여러 다운스트림 작업에 유연하게 적응하기 위해 새로운 self-attention 을 사용하는 multi-modal pre-training model인 MedViLL을 제안한다.

- 3개의 radiographic 이미지 판독문 데이터 세트를 사용하여 4개의 다운스트림 작업 모두에 대해 MedViLL을 통계적이고 엄격하게 평가함으로써 작업별 아키텍처를 포함한 다양한 기준선에 대해 MedViLL의 우수한 성능을 경험적으로 입증했다.

- MedViLL의 인상적인 성능에도 불구하고, 이것은 medical 영역에서 vision-language representation learning의 시작일 뿐이며, 우리는 이러한 접근 방식을 시간이 지남에 따라 multi-view Chest X-ray 연구 또는 일련의 연구와 같은 더 다양한 설정으로 확장할 계획이다

APPENDIX

- https://arxiv.org/pdf/2105.11333.pdf