Natural Language Processing
Medical Visual Question Answering

Team EnEoPy

21280115 - Tran Duc Trung

21280035- Nguyen Phuc Gia Nghi

July 6, 2024

Content I

1. The problem

2. Solution
   - 2.1 Model Architecture
   - 2.2 Unimodal and Multimodal Contrastive Losses
   - 2.3 Image Text Matching
   - 2.4 Masked Language Modeling
   - 2.5 Masked Image Strategy

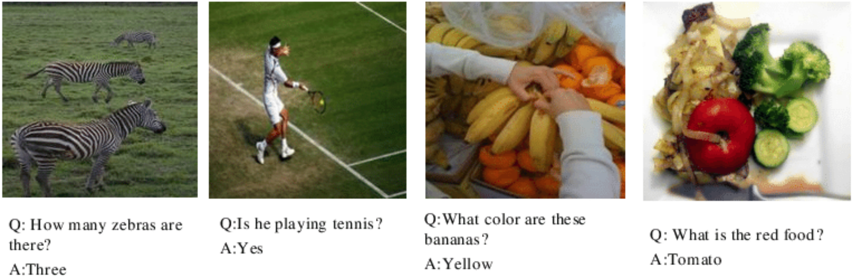
3. Result

4. Other work

5. Future work

1. VQA Problem

VQA Prolems



Q: How many zebras are there?
A:Three

Q:Is he playing tennis?
A:Yes

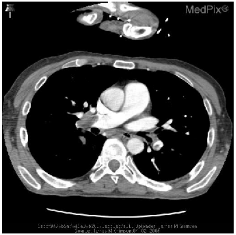
Q:What color are these bananas?
A:Yellow

Q: What is the red food?
A:Tomato

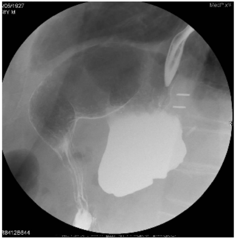
Figure 1: An example of VQA

- Visual question answering (VQA) is a challenging task that requires answering questions of a given image, by taking consider of both visual and language information.

Medical VQA Problems



(g) **Q**: which organ system is shown in the ct scan? **A**: lung, mediastinum, pleura



(h) **Q**: what is abnormal in the gastrointestinal image? **A**: gastric volvulus (organoaxial)

- Medical Visual Question Answering (VQA) is a combination of medical artificial intelligence and popular VQA challenges.

Dataset Description

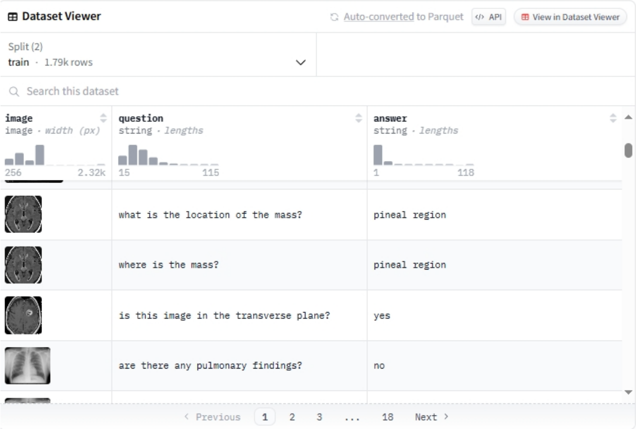


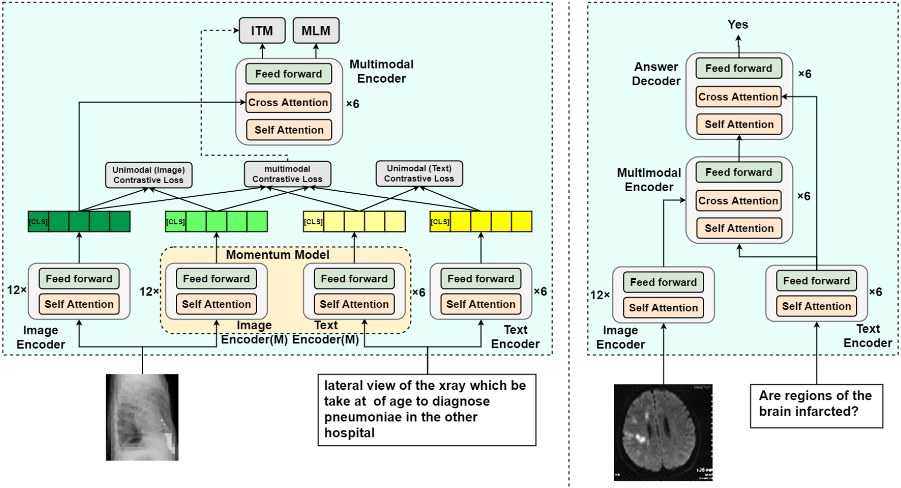
image	question	answer
image · *width (px)*	string · *lengths*	string · *lengths*
256 2.32k	15 115	1 118
	what is the location of the mass?	pineal region
	where is the mass?	pineal region
	is this image in the transverse plane?	yes
	are there any pulmonary findings?	no

Figure 2: Dataset VQA-RAD

- The dataset contains 2,248 question-answer pairs and 315 images, with 1,793 question-answer pairs in training set and 451 in testing set.

2. Solution

MUMC Architecture



lateral view of the xray which be take at of age to diagnose pneumoniae in the other hospital

Are regions of the brain infarcted?

(a) Pre-training

(b) Fine-tuning

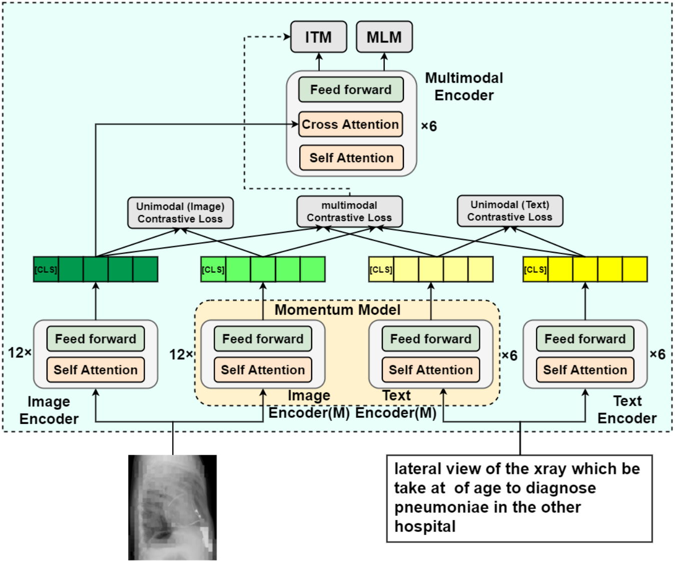
MUMC architecture

Introduction

- Due to the small scale of training data, pre-training fine-tuning paradigms have been used solution to improve model performance
- We proposed a self-supervised vision language pre-training (VLP) approach that applied Masked image and text modeling with Unimodal and Multimodal Contrastive losses (MUMC) in the pre-training phase for solving downstream VQA tasks.
- The model was pretrained on image caption datasets for aligning visual and text information, and transferred to downstream VQA datasets. The unimodal and multimodal contrastive losses in our work are applied to align image and text features; learn unimodal image encoders via momentum contrasts of different views of the same image; learn unimodal text encoder via momentum contrasts
- Used a masked image strategy by randomly masking the patches of the image with a probability of 25%

2.1 Model Architecture

MUMC pre-training phase



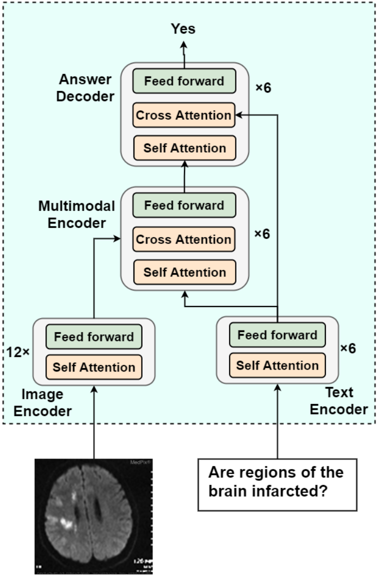
lateral view of the xray which be take at of age to diagnose pneumoniae in the other hospital

(a) Pre-training

- Pre-training phase comprises an image encoder, a text encoder, a multimodal encoder, which are all based on the transformer
- Image encoder leverages a 12-layer Vision Transformer (ViT)
- Text encoder leverages the first 6 layers of pre-trained BERT
- The last 6 layers of BERT are utilized as the multimodal encoder and incorporated cross-attention at each layer which fuses the visual and linguistic features to facilitate learning of multimodal interactions

- The model is trained on image-caption pairs. An image is partitioned into patches of size 16 × 16, and 25% of the patches are randomly masked. The remaining unmasked image patches are converted into a sequence of embeddings by an image encoder.
- Text is tokenized into a sequence of tokens using a WordPiece tokenizer and fed into the BERT-based text encoder
- [CLS] are appended to the beginning of both the image and text sequence

MUMC fine-tuning phase



(b) Fine-tuning

- To transfer the models trained to the downstream VQA tasks, we utilize the weights from the pre-training stage to initialize the image encoder, text encoder and multimodal encoder
- To generate answers, add an answering decoder with a 6-layer transformer-based decoder, which receives the multimodal embeddings and output text tokens.
- [CLS] token serves as the initial input token for the decoder, and a [SEP] token is appended at the end of the generated sequence
- The downstream VQA model is fine-tuned via the masked language model (MLM) loss, using ground-truth answers as targets

2.2 Unimodal and Multimodal Contrastive Losses

$$L_{ucl} = \frac{1}{2}\mathbb{E}_{(V,T)D}\left[H\left(y_{i2i}(V), \frac{\exp\left(s\left(V,V_i\right)/\tau\right)}{\sum_{n=1}^{N}\exp\left(s\left(V,V_i\right)/\tau\right)}\right) + H\left(y_{t2t}(T), \frac{\exp\left(s\left(T,T_i\right)/\tau\right)}{\sum_{n=1}^{N}\exp\left(s\left(T,T_i\right)/\tau\right)}\right)\right]$$

$$L_{mcl} = \frac{1}{2}\mathbb{E}_{(V,T)D}\left[H\left(y_{i2t}(V), \frac{\exp\left(s\left(V,T_i\right)/\tau\right)}{\sum_{n=1}^{N}\exp\left(s\left(V,T_i\right)/\tau\right)}\right) + H\left(y_{t2i}(T), \frac{\exp\left(s\left(T,V_i\right)/\tau\right)}{\sum_{n=1}^{N}\exp\left(s\left(T,V_i\right)/\tau\right)}\right)\right]$$

Note: - The image and caption embeddings from the unimodal image encoder and text encoder as v_{cls} and t_{cls}

- The transformations are g_v and g_t, to normalize and map the image and text embeddings to be lower-dimensional representations.

- The ground-truth one-hot similarity by $y_{i2i}(V), y_{t2t}(T), y_{i2t}(V)$, and $y_{t2i}(T)$, where the probability of negative pairs is 0 and positive pair is 1 .

- s denotes cosine similarity function,

$s\left(\mathrm{~V}, V_i\right) = g_v\left(v_{cls}\right)^T g_v\left(v_{cls}\right)_i$,

s($\mathrm{T}, T_i) = g_t\left(t_{cls}\right)^T g_t\left(t_{cls}\right)_i$,

s($\mathrm{V}, T_i) = g_v\left(v_{cls}\right)^T g_t\left(t_{cls}\right)_i$,

s($\mathrm{T}, V_i) = g_t\left(t_{cls}\right)^T g_v\left(v_{cls}\right)_i$ and τ is a learnable temperature parameter.

- The proposed self-supervised objective attempts to capture the semantic discrepancy between positive and negative samples across both unimodal and multimodal domains at the same time
- The unimodal contrastive loss (UCL) aims to differentiate between examples of one modality (image-image or text-text)
- The multimodal contrastive loss (MCL) learns the alignments between both modalities (image-text)

2.3 Image Text Matching

- Image-Text Matching Loss (ITM) predicts whether a pair of image and text is positive (matched) or negative(not matched). It aims to learn image-text multimodal representation that captures the fine-grained alignment between vision and language. ITM is a binary classification task, where the model uses an ITM head (a linear layer) to predict whether an image-text pair is positive (matched) or negative (unmatched) given their multimodal feature. The ITM task is optimized using the cross-entropy loss:

$$L_{itm} = \mathbb{E}_{(V,T)\sim D} H\left(y_{itm}, p_{itm}(V,T)\right)$$

Note: $H\left(\ ,\ \right)$ represents a cross-entropy computation, where y_{itm} is a 2-dimensional one-hot vector representing the ground-truth label, and $p_{\text{itm}}\left(V,T\right)$ is a function for predicting the class.

2.4 Masked Language Modeling

- Masked Language Modeling (MLM) is another pre-trained objective in our approach, that predicts masked tokens in text based on both the visual and unmasked contextual information. For each caption text, 15% of tokens are randomly masked and replaced with the special token, [MASK]. Predictions of the masked tokens are conditioned on both unmasked text and image features. We minimize the cross-entropy loss for MLM:

$$\mathcal{L}_{mlm} = \mathbb{E}_{(V,\hat{T})D} H\left(y_{mlm}, p_{mlm}(V,\widehat{T})\right) \tag{4}$$

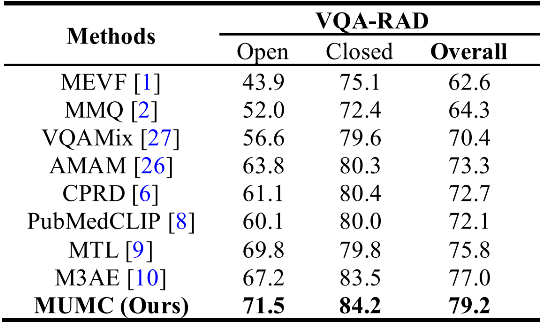
Note: $H(,)$ is a cross-entropy calculation, $\hat{T}$ denotes the masked text token, y_{mlm} represents the ground-truth of the masked text token and $p_{mlm}(V,\widehat{T})$ is the predicted probability of a masked token.

2.5 Masked Image Strategy

- Using a masked image strategy as a data augmentation technique. Input images are partitioned into patches which are randomly masked with a probability of 25%, and only the unmasked patches are passed through the network

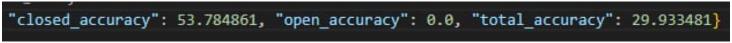
3. Result

Result



Methods	VQA-RAD		
	Open	Closed	**Overall**
MEVF [1]	43.9	75.1	62.6
MMQ [2]	52.0	72.4	64.3
VQAMix [27]	56.6	79.6	70.4
AMAM [26]	63.8	80.3	73.3
CPRD [6]	61.1	80.4	72.7
PubMedCLIP [8]	60.1	80.0	72.1
MTL [9]	69.8	79.8	75.8
M3AE [10]	67.2	83.5	77.0
MUMC (Ours)	**71.5**	**84.2**	**79.2**

Figure 3: Results of the base MUMC and SOTA methods on the dataset VQA-RAD



```
"closed_accuracy": 53.784861, "open_accuracy": 0.0, "total_accuracy": 29.933481}
```

Figure 4: Our MUMC's result on the dataset VQA-RAD

4. Other work

BLIP model

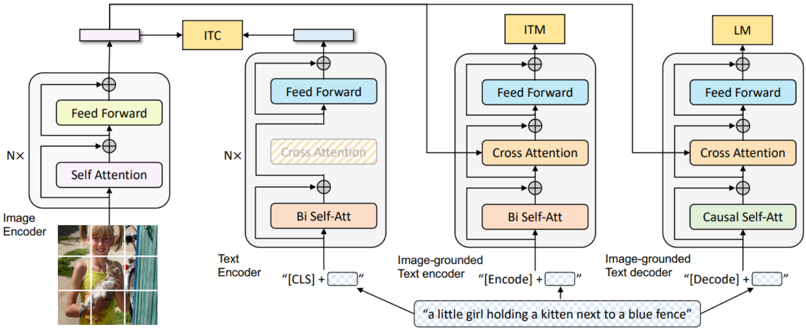
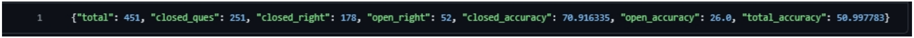


Figure 5: BLIP model



```
{"total": 451, "closed_ques": 251, "closed_right": 178, "open_right": 52, "closed_accuracy": 70.916335, "open_accuracy": 26.0, "total_accuracy": 50.997783}
```

Figure 6: BLIP result

5. Future work

Apply compact transformers to MUMC

- With the rise of Transformers as the standard for language processing, along with their benefits, is the unprecedented size and amount of training data $->$ leads to great concerns: limited availability of data in certain scientific domains and the exclusion of those with limited resource from research in the field
$=>$Compact transformers dispel the myth that transformers are “data hungry”

Team EnEoPy's source code I

[1] MUMC: `https://github.com/nprm1243/Compact-MUMC`

[2] BLIP: `https://github.com/nprm1243/Compact-BLIP`

References I

[1] Li, P. (2022). MUMC. GitHub repository. Retrieved from https://github.com/pengfeiliHEU/MUMC/tree/main

[2] Masked Vision and Language Pre-training with Unimodal and Multimodal Contrastive Losses for Medical Visual Question Answering. https://arxiv.org/pdf/2307.05314v1

[3] Align before Fuse: Vision and Language Representation Learning with Momentum Distillation https://arxiv.org/pdf/2107.07651

[4] Dataset: https://huggingface.co/datasets/flaviagiammarino/vqa-rad

[5] BLIP. GitHub repository. Retrieved from https://github.com/salesforce/BLIP

End