MI2L

# MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain

Dhruv Sharma, Sanjay Purushotham & Chandan K. Reddy, Scientific Reports (2022)


**Jeeyoung Kim**

University of Ulsan College of Medicine,

Asan Medical Center

77imjee@gmail.com


2023.01.27 Fri


울산대학교 UNIVERSITY OF ULSAN

서울아산병원 Asan Medical Center

# Introduction

- Medical images are difficult to comprehend for a person without expertise
- The scarcity of medical practitioners across the globe often face the issue of physical and mental fatigue due to the high number of cases, inducing human errors during the diagnosis.
- In such scenarios, having an additional opinion can be helpful in boosting the confidence of the decision maker. Thus, it becomes crucial to have **a reliable visual question answering (VQA) system** to provide a **‘second opinion’** on medical cases.

# Introduction

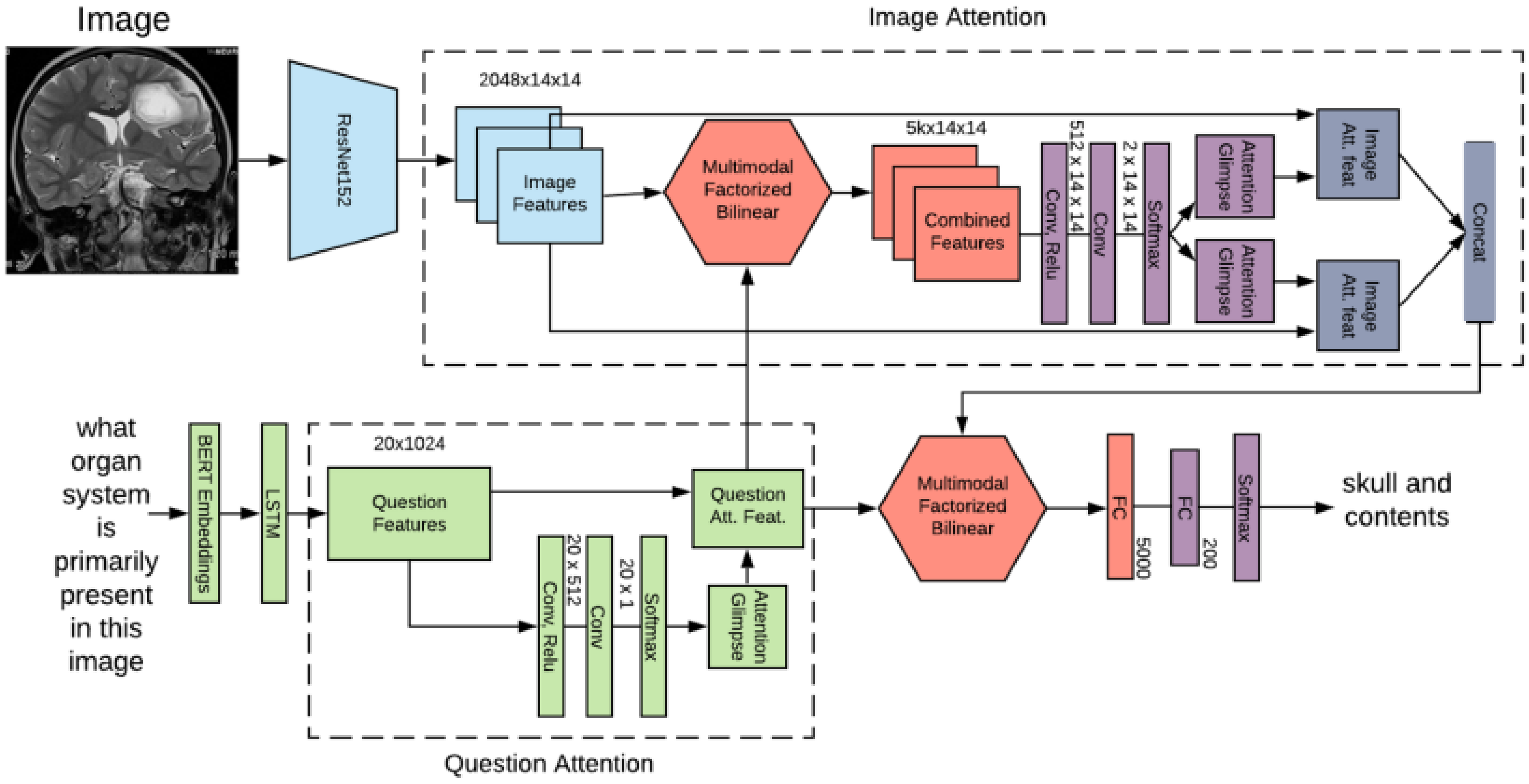
- However, most of the VQA systems that work today cater to real-world problems and **are not specifically tailored for handling medical images**.
  - the main challenge is the limited availability of labeled medical data
  - the number of VQA data samples in medical domain are quite less compared to the VQA datasets for the other real-world domains.

# Introduction

- We propose **MedFuseNet, an attention based multimodal deep learning model for answer categorization and answer generation tasks in medical domain VQA**. We show that a LSTM-based generative decoder along with heuristics can improve our model performance for the answer generation task.
- We **demonstrate state-of-the-art results on two real-world medical VQA datasets**. In addition, we conducted an exhaustive ablation study to investigate the importance of each component in our proposed model.
- We study the **interpretability** of our MedFuseNet by **visualizing various attention mechanisms** used in the model. Tis provides a deeper insight into understanding the VQA capability of our model.

# MedFuseNet

- Image feature extraction
- Question feature extraction
- Feature fusion techniques
- Attention mechanisms



# MedFuseNet

- Image feature extraction
  - ResNet-152
  - Since the medical images are complex compared to the standard real-world images, models like DenseNet-121 and ResNet-152 which have skip connections, provide more robust feature representations through deeper convolutional layers.
- Question feature extraction
  - positional semantics of each word and the word-level semantics
  - BERT + XLNet

# MedFuseNet

- Feature fusion techniques
  - MFB : simplicity of the algorithm, ease of implementation, high convergence rate

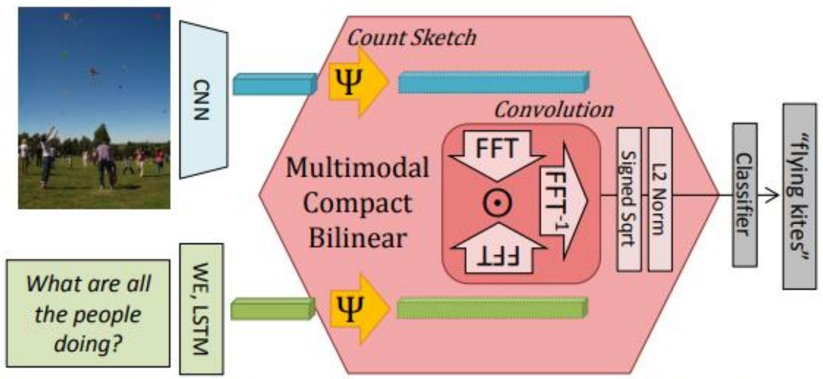
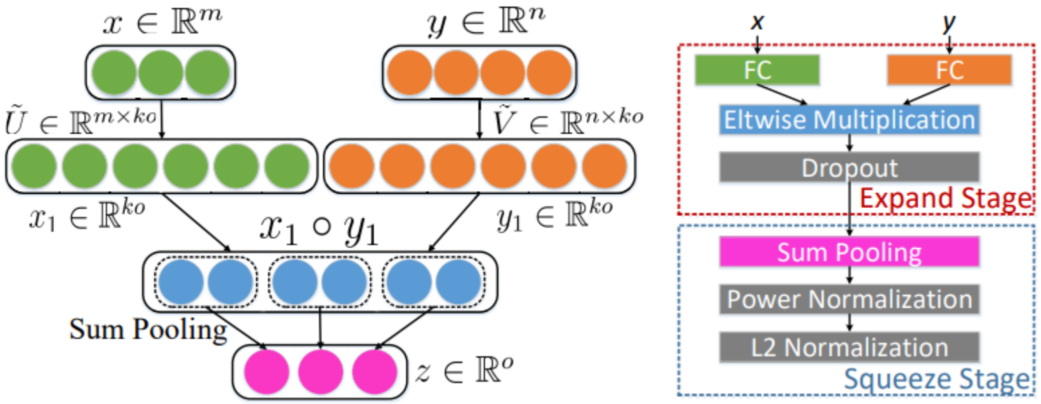



Figure 1: Multimodal Compact Bilinear Pooling for visual question answering.




(a) Multi-modal Factorized Bilinear Pooling (b) MFB module

# MedFuseNet

- Attention mechanisms
  - Image attention : The image attention mechanism aims at spanning the attention of the MedFuseNet model to **the most relevant part of the image based on the input question**
  - Image-Question Co-Attention : use the attended vector as an input to the image attention mechanism

**Algorithm 1:** *MedFuseNet* Training Algorithm

**Input:** Image $v$, Question $q$, Answer $a$, Batch size $N_b$
**Output:** Trained model parameters $\Theta$

Extract the image features ($\hat{v}$), from image ($v$)
Extract the question features ($\hat{q}$) from question ($q$)
**for** *a few iterations* **do**
  **for** *batch of size $N_b$ in* $\{\hat{v}, \hat{q}, a\}$ **do**
    Perform Question Attention $\mathscr{E}_q(q)$ on $\hat{q}$ to get attended question features ($\hat{q}_e$)
    Perform Image Attention $\mathscr{E}_v(\hat{v}, \hat{q}_e, MFB, 2)$ on $\hat{v}$ to get attended image features ($\hat{v}_e$)
    Combine $\hat{q}_e$ and $\hat{v}_e$ using MFB($\hat{q}_e$, $\hat{v}_e$, 5000, 3) to get intermediate vector ($z$)
    Find the predicted answer ($\tilde{a}$) depending on the task as defined in Eq. (1) and Eq. (2)
    Calculate the loss $\mathscr{L}$ for $a$ and $\hat{a}$ using Eq. (3)
    Update the model parameters $\Theta$ with the loss $\mathscr{L}$
  **end**
**end**
**return** trained model parameters $\Theta$
**Procedure** MFB ($\hat{v}, \hat{q}, d_o, k$)
  $v' = Fully-Connected(\hat{v}, m, d_o)$
  $q' = Fully-Connected(\hat{q}, n, d_o)$
  Compute and store inner product ($\circ$) of vector $v'$ and vector $q'$ in vector z
  Perform SumPooling of vector $z$ with a window size of $k$
  Normalize vector $z$ using L2-normalization
  **return** $z$
**Procedure** Image Attention ($\hat{v}, \hat{q}, \mathscr{F}, g$)
  Combine $\hat{v}$ and $\hat{q}$ using $\mathscr{F}(\hat{q}_e, \hat{v}_e)$ to get intermediate vector $f$
  $f_{conv}$ = ReLU(Conv2d($f, d_o$, 512))
  $f_{AttMaps}$ = Softmax(Conv2d($f_{conv}$, 512, $g$))
  Initialize $v_e$ as an empty list to store the attention glimpses
  **for** $i \leftarrow 1$ **to** $g$ **do**
    Find the attended image feature $e_i$ for $i_{th}$ glimpse as follows:
    $e_i = f_{AttMaps}[i] \circ \hat{v}$
    Add $e_i$ to the list $v_e$
  **end**
  Sum over all the attention glimpses in $v_e$ to get attended image feature vector ($\hat{v}_e$)
  **return** $\hat{v}_e$

# MedFuseNet

- ResNet and BERT models are pretrained on very large datasets, and they provide a much better generalization for the features by the virtue of transfer learning.
- Due to the simplistic implementation of MFB, it reduces the complexity of calculating the outer product to a large extent, while conserving the information from the fusion of the two modalities. Tis reduces the computation of model parameters and works well for the limited MED-VQA datasets.
- The attention and co-attention mechanisms help in reducing the attention span of the model to the significant parts of the input, thus, reducing the search space for the model.

# Datasets

- MED-VQA, PathVQA

(a)
- ***what kind of image is this?***
- cta - ct angiography
- ***which plane is this image taken?***
- axial
- ***which organ is captured by this ct scan?***
- lung, mediastinum, pleura
- ***what is abnormal in the ct scan?***
- cryptococcal pneumonia in an immunocompetent host

(b)
- ***is this a t1 weighted, t2 weighted, or flair image?***
- T2
- ***what imaging plane is depicted here?***
- Coronal
- ***what organ system is shown in the image?***
- skull and contents
- ***what is abnormal in the mri?***
- colloid (neuroepithelial) cyst of the third ventricle

(c)
- ***what modality was used to take this image?***
- xr - plain film
- ***what plane is this?***
- Ap
- ***what organ system is shown in this x-ray?***
- Musculoskeletal
- ***what is the primary abnormality in this image?***
- psoriatic arthritis

(d)
- ***is coronary artery anomalous origin left from pulmonary artery present?***
- no
- ***what does this image show?***
- x-ray three horizontal slices of ventricles showing quite well the penetrating arteries
- ***where is this from?***
- heart

**Figure 1.** Sample radiology scans and the corresponding question-answer pairs from the MED-VQA and PathVQA dataset. The first three (**a**–**c**) belong to the MED-VQA dataset and the last one (**d**) belongs to the PathVQA dataset.

# Datasets

- MED-VQA
  - Modality – 3825 triplet(image-question-answer), 35 classes
  - Plane – 3825 triplet, 16 classes
  - Organ – 3825 triplet, 10 unique organ systems
  - maximum question length for the three questions combined is 13 words and the average question length is around 8 words

| Split | Modality | Plane | Organ |
|---|---|---|---|
| Train | 3200 | 3200 | 3200 |
| Validation | 500 | 500 | 500 |
| Test | 125 | 125 | 125 |

# Datasets

- PathVQA
  - only use the yes-no type question
  - the average question length is about 6 words

| Split | Medical Images | 'Yes' type QA Pairs | 'No' type QA Pairs |
|---|---|---|---|
| Train | 4271 | 9305 | 9163 |
| Validation | 1176 | 2359 | 2335 |
| Test | 942 | 1874 | 1853 |

# Dataset preprocessing

- Image
  - resize to the same dimension of 224 x 224 x 3
- Question
  - tokenized using the NLTK library in python
  - questions were padded to make them all of the same lengths

# Implementation details

- Image feature extractor : pre-trained models available in Keras
- Question feature Extractor : Embedding –as-a-Service (BERT, XLNet)
- questions : 20 tokens
- combined feature vector : 5000
- optimizer : ADAM
- batch size : 32
- epochs : 100

# Results

| Methods | Accuracy | | | AUC-ROC | | | AUC-PRC | | |
|---|---|---|---|---|---|---|---|---|---|
| | Modality | Plane | Organ | Modality | Plane | Organ | Modality | Plane | Organ |
| VIS + LSTM[50] | 0.704(0.012) | 0.701(0.017) | 0.652(0.020) | 0.899(0.012) | 0.851(0.011) | 0.775(0.015) | 0.478(0.024) | 0.453(0.022) | 0.456(0.025) |
| d-LSTM + n-CNN[52] | 0.723(0.014) | 0.719(0.018) | 0.672(0.022) | 0.909(0.010) | 0.862(0.014) | 0.777(0.017) | 0.474(0.025) | 0.459(0.023) | 0.450(0.027) |
| SAN[18] | 0.669(0.013) | 0.729(0.015) | 0.669(0.023) | 0.926(0.011) | 0.870(0.011) | 0.783(0.015) | 0.459(0.025) | 0.415(0.023) | 0.406(0.026) |
| HiCAt[19] | 0.760(0.010) | 0.740(0.015) | 0.668(0.018) | 0.929(0.011) | 0.869(0.010) | 0.797(0.014) | 0.468(0.023) | 0.431(0.025) | 0.430(0.028) |
| BAN[21] | 0.820(0.011) | 0.766(0.016) | **0.750(0.014)** | **0.961(0.010)** | **0.929(0.009)** | 0.800(0.016) | 0.600(0.024) | 0.521(0.022) | 0.456(0.025) |
| *MedFuse-Net* | **0.840(0.010)** | **0.780(0.017)** | 0.746(0.015) | 0.942(0.010) | 0.901(0.010) | **0.800(0.013)** | **0.618(0.023)** | **0.526(0.024)** | **0.510(0.023)** |

**Table 4.** Comparison of *MedFuseNet* with the baseline models on MED-VQA answer classification dataset.

| Methods | Accuracy |
|---|---|
| VIS + LSTM[50] | 0.603(0.025) |
| d-LSTM + n-CNN[52] | 0.607(0.021) |
| SAN[18] | 0.627(0.023) |
| HiCAt[19] | 0.629(0.018) |
| BAN[21] | 0.604(0.021) |
| *MedFuseNet* | **0.636(0.020)** |

**Table 5.** Comparison of *MedFuseNet* with the baseline models on PathVQA yes-no answer type dataset.

# Results

| Question Category | Image Feature | MCB | | MUTAN | | MFB | |
|---|---|---|---|---|---|---|---|
| | | BERT | XLNet | BERT | XLNet | BERT | XLNet |
| *Accuracy* | | | | | | | |
| Category 1 Modality | VGG16 | 0.718(0.019) | 0.697(0.018) | 0.751(0.016) | 0.686(0.019) | 0.805(0.012) | 0.680(0.019) |
| | DenseNet121 | 0.704(0.015) | 0.675(0.019) | 0.768(0.014) | 0.688(0.021) | 0.813(0.014) | 0.675(0.020) |
| | ResNet152 | 0.731(0.014) | 0.663(0.017) | 0.783(0.018) | 0.716(0.017) | **0.840(0.011)** | 0.701(0.018) |
| Category 2 Plane | VGG16 | 0.706(0.018) | 0.697(0.016) | 0.750(0.017) | 0.605(0.022) | 0.749(0.014) | 0.629(0.019) |
| | DenseNet121 | 0.719(0.016) | 0.643(0.018) | 0.754(0.016) | 0.643(0.017) | 0.757(0.011) | 0.655(0.021) |
| | ResNet152 | 0.712(0.015) | 0.659(0.019) | 0.763(0.015) | 0.693(0.019) | **0.780(0.010)** | 0.735(0.016) |
| Category 3 Organ System | VGG16 | 0.718(0.018) | 0.625(0.015) | 0.785(0.012) | 0.683(0.016) | **0.798(0.011)** | 0.692(0.019) |
| | DenseNet121 | 0.753(0.013) | 0.630(0.018) | 0.774(0.015) | 0.696(0.018) | 0.774(0.012) | 0.720(0.016) |
| | ResNet152 | 0.669(0.016) | 0.672(0.013) | 0.705(0.016) | 0.649(0.019) | 0.746(0.010) | 0.682(0.015) |
| *AUC-ROC* | | | | | | | |
| Category 1 Modality | VGG16 | 0.845(0.011) | 0.697(0.016) | 0.896(0.010) | 0.710(0.015) | **0.954(0.011)** | 0.738(0.015) |
| | DenseNet121 | 0.854(0.013) | 0.675(0.018) | 0.898(0.010) | 0.659(0.014) | 0.934(0.010) | 0.703(0.016) |
| | ResNet152 | 0.861(0.012) | 0.703(0.018) | 0.906(0.011) | 0.740(0.017) | 0.942(0.013) | 0.700(0.014) |
| Category 2 Plane | VGG16 | 0.833(0.012) | 0.697(0.018) | 0.866(0.011) | 0.718(0.017) | 0.899(0.013) | 0.729(0.014) |
| | DenseNet121 | 0.832(0.013) | 0.743(0.017) | 0.867(0.012) | 0.801(0.013) | 0.894(0.012) | 0.839(0.015) |
| | ResNet152 | 0.840(0.010) | 0.685(0.017) | 0.881(0.010) | 0.849(0.014) | **0.921(0.012)** | 0.891(0.013) |
| Category 3 Organ System | VGG16 | 0.655(0.015) | 0.619(0.019) | 0.689(0.014) | 0.622(0.017) | 0.691(0.014) | 0.730(0.016) |
| | DenseNet121 | 0.667(0.013) | 0.700(0.016) | 0.691(0.013) | 0.626(0.018) | 0.690(0.013) | 0.650(0.014) |
| | ResNet152 | 0.803(0.010) | 0.674(0.018) | **0.854(0.012)** | 0.795(0.014) | 0.800(0.010) | 0.790(0.015) |
| *AUC-PRC* | | | | | | | |
| Category 1 Modality | VGG16 | 0.322(0.019) | 0.312(0.017) | 0.379(0.017) | 0.373(0.020) | 0.590(0.016) | 0.352(0.019) |
| | DenseNet121 | 0.287(0.021) | 0.310(0.019) | 0.407(0.016) | 0.390(0.019) | 0.572(0.018) | 0.219(0.021) |
| | ResNet152 | 0.361(0.021) | 0.208(0.018) | 0.469(0.017) | 0.343(0.019) | **0.618(0.016)** | 0.224(0.018) |
| Category 2 Plane | VGG16 | 0.252(0.018) | 0.368(0.018) | 0.331(0.019) | 0.370(0.021) | 0.439(0.017) | 0.288(0.020) |
| | DenseNet121 | 0.269(0.017) | 0.279(0.021) | 0.347(0.018) | 0.335(0.021) | 0.437(0.019) | 0.351(0.019) |
| | ResNet152 | 0.248(0.020) | 0.293(0.021) | 0.365(0.017) | 0.321(0.020) | **0.526(0.016)** | 0.435(0.017) |
| Category 3 Organ System | VGG16 | 0.341(0.016) | 0.348(0.020) | 0.393(0.018) | 0.289(0.019) | 0.443(0.019) | 0.351(0.016) |
| | DenseNet121 | 0.364(0.018) | 0.420(0.018) | 0.377(0.016) | 0.289(0.021) | 0.433(0.021) | 0.330(0.018) |
| | ResNet152 | 0.428(0.017) | 0.322(0.017) | 0.473(0.019) | 0.396(0.018) | **0.510(0.016)** | 0.352(0.018) |

**Table 7.** Performance metric scores for the ablation study experiments on MED-VQA dataset.

# Experiments

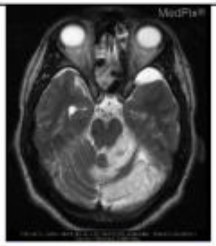
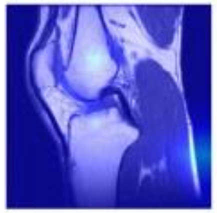
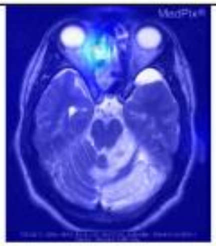
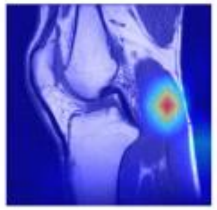
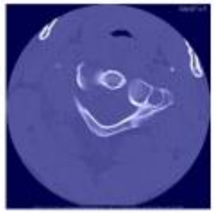
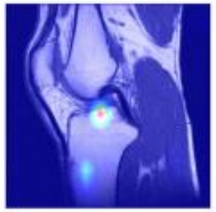
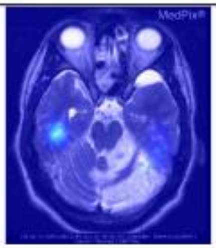
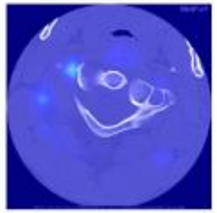
| Method | musculoskeletal - ankle | knee | skull and contents | spine and contents |
|---|---|---|---|---|
| Original |  |  |  |  |
| SAN[18] |  |  |  |  |
| HiCAt[19] |  |  |  |  |
| *MedFuseNet* |  |  |  |  |

**Table 9.** Image Attention visualization for SAN, Hie. Co-Att, and *MedFuseNet*.


(a) Modality : xr - plain film

(b) Plane : lateral

(c) Organ : musculoskeletal

**Figure 5.** Co-Attention Maps for a sample case to display the attention span of *MedFuseNet* with the input image and the corresponding question attention. **(a)** Displays the image attention map and the corresponding question attention map for category 1—modality, **(b)** for category 2—plane, and **(c)** for category 3—organ.

# Conclusions

- Visual questions answering systems for medical images can be extremely helpful in providing the doctors with a second-opinion
- We presented MedFuseNet, an attention-based multimodal deep learning model for VQA on medical images
- Ablation study was conducted to investigate the role of image features, question features, and fusion techniques on the model performance for the two VQA tasks