Build an interpretable model to provide answers to clinically relevant questions based on radiology images

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2021
Technology, Social Impact

MMBERT: Multimodal BERT Pretraining for Improved Medical VQA

ABSTRACT

Images in the medical domain are fundamentally different from the general domain images. Consequently, it is infeasible to directly employ general domain Visual Question Answering (VQA) models for the medical domain. Additionally, medical image annotation is a costly and time-consuming process. To overcome these limitations, we propose a solution inspired by self-supervised pre-training of Transformer-style architectures for NLP, Vision, and Language tasks. Our method involves learning richer medical image and text semantic representations using Masked Vision-Language Modeling as the pretext task on a large medical image+caption dataset. The proposed solution achieves new state-of-the-art performance on two VQA datasets for radiology images – VQA-Med 2019 and VQA-RAD, outperforming even the ensemble models of previous best solutions. Moreover, our solution provides attention maps which help in model interpretability.

OBJECTIVE

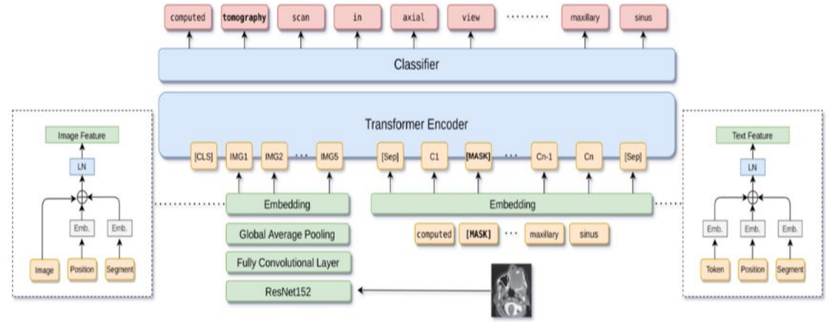
Build an interpretable model to provide answers to clinically relevant questions based on radiology images

METHOD

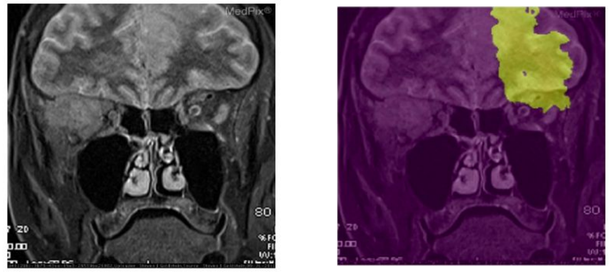
Self-supervised pretraining of BERT-like architectures has proven quite effective in Natural Language Processing (NLP), Vision, and Language space. Our Multimodal Medical BERT (MMBERT) model is inspired by these approaches. We first pretrain our MMBERT model on a set of medical images and their corresponding captions with the Masked Vision-Language Modeling task. Later this model is finetuned for the VQA task.

In masked vision-language modeling, the task is to predict the original token in place of a [MASK] token with the usage of text and image features. To ensure that the model learns to predict medical words from the context, we mask only medical keywords (provided with the dataset) from the captions and leave the common words untouched.

MODEL



RESULTS



Question: What imaging modality was used?
Answer: MR-T2 Weighted

Authors:Yash Khare, Viraj Bagal, Minesh Mathew et. al.

Research Center Name: HAI