Medical Visual Question Answering

Code Link:

https://drive.google.com/drive/folders/1y7ZAlukwNs_RP0sz5TsJckwSSpYpN-z1?usp=sharing

Approaches

Since the dataset itself is not very large, we cannot expect a model to learn to perform sophisticated tasks like Medical Visual Question Answering all by itself. So, we thought of fine-tuning the existing pre-trained medical VQA models on the given dataset. But still we did try training a model from scratch. So our approaches were:-

- BiomedCLIP
- Cross-Modal Attention
- SAN-stacked attention network

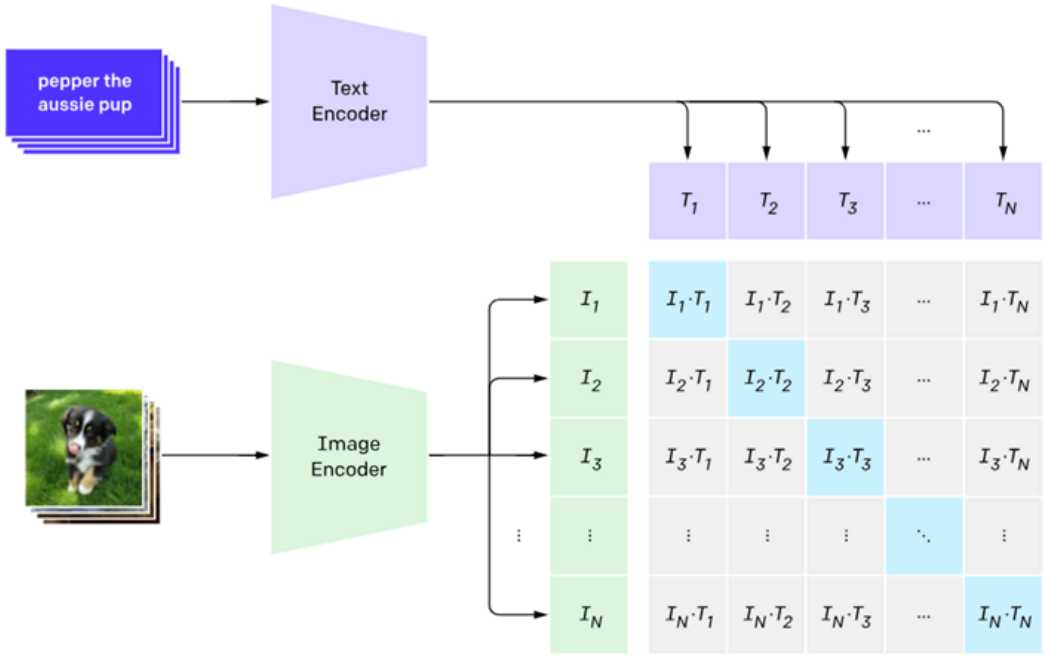
The main aim of fine-tuning it on different pre-trained models was to try out different approaches and methods and select the one that gives the best accuracy. All the models that we have listed here are different from each other and have a unique architecture of themselves thus providing different approaches to solve the same task. Let's talk in detail about each of the approaches that we have tried out.

1)BiomedCLIP

Since the given dataset consists of mostly one-word answers and only around 2.5k data points, we cannot expect to train a model from scratch and expect it to generalize well. So, we try to leverage pre-trained models and transfer their knowledge in the medical domain to train the model on the given VQA-RAD.

CLIP – Contrastive Language-Image Pre-Training

Given a batch of N (image, text) pairs, CLIP learns a multimodal embedding space by jointly training an image encoder and a text encoder to maximize the cosine similarity between the image and text embeddings of the N pairs in the batch while minimizing the cosine similarity of the embeddings of the other N2-N non-pairs.



Fine-tuning CLIP for bio-medical tasks

On the text side, the researchers replaced the blank-slate GPT-2 with PubMedBERT - a pre-trained language model more suited for biomedicine, which shows substantial gains from domain-specific pretraining. On the image side, Vision Transformers were used for pre-training. The pre-training was performed on the PMC-15M dataset.

BiomedCLIP learns to encode both images and text in a joint embedding space, enabling it to perform tasks such as cross-modal retrieval and image-text understanding in the biomedical domain.

Fine Tuning BiomedCLIP on VQA-RAD

Now, we used the pre-trained image and text embeddings from BiomedCLIP and added some dense layers on top of it for the classification task at hand. The given method showed 60-70% accuracy on the output labels.

2) Cross-Modal Attention

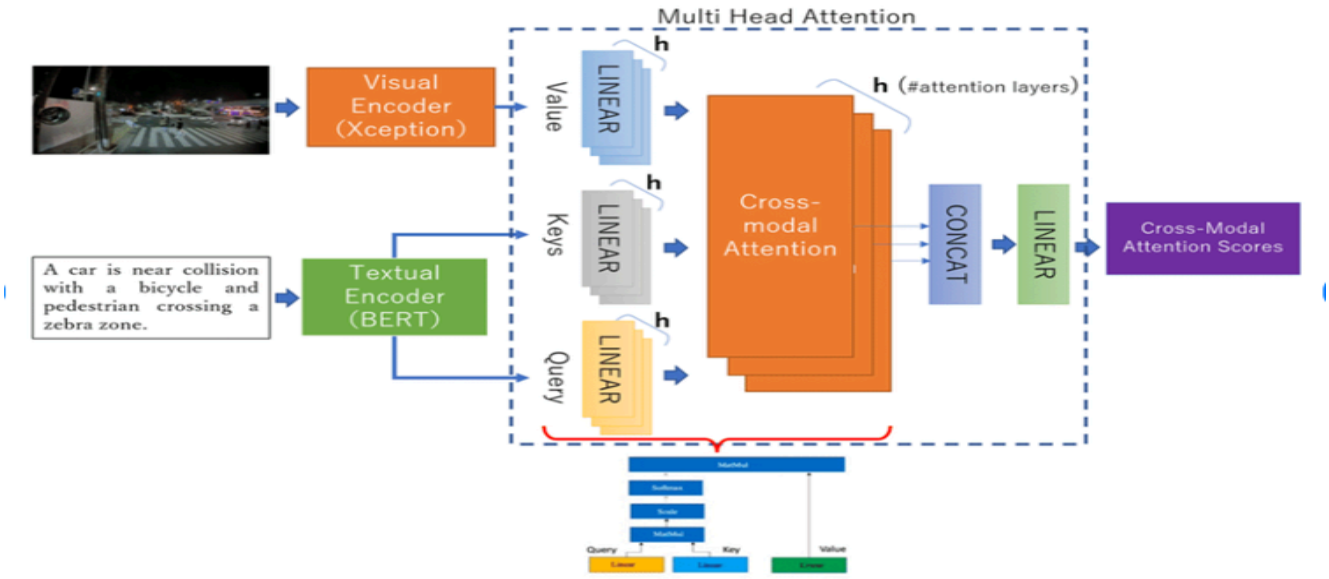
For text embeddings, **"PubMedBERT ",** a specialized variant of the BERT (Bidirectional Encoder Representations from Transformers) model designed for biomedical and healthcare-related natural language processing tasks, particularly for the text from PubMed articles and clinical

texts. It has been pre-trained on a vast amount of biomedical text data to capture domain-specific knowledge and context.

For image embeddings two approaches were used - VGG-16 and **Vision Transformers**.

Vision Transformers (ViTs) are a class of deep learning models that adapt the Transformer architecture, originally designed for natural language processing, to computer vision tasks. They process images by breaking them into smaller patches and use self-attention mechanisms to capture both local and global relationships.

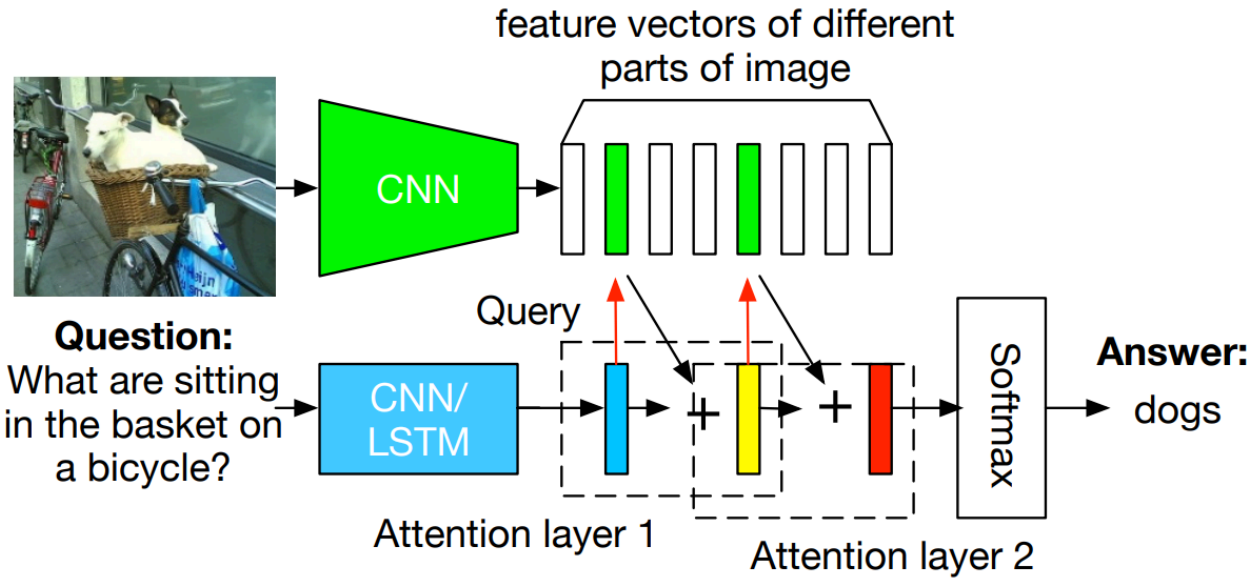
Now, we apply cross attention on the image and text embeddings so that the model can learn what part of the image to focus on given the question at hand. So, we add a multi-head cross-attention layer on top of image and text embeddings.



Now the combined embeddings are passed through a classification layer to finally obtain the classification logits.

The proposed model was able to achieve an accuracy of 72-75% on the classification labels (the top 10 most frequent answers in the dataset)

3)SAN-Stacked attention network



Firstly I loaded all the dependencies and libraries required. After that, the dataset was preprocessed. Then, the questions and answers were tokenized, and images were preprocessed. After that dataset of the processed images and questions is prepared. Both The feature vectors are shaped accordingly for the input. For the language part, I used an LSTM layer. For the image part, I used the VGG16 model. After that, two approaches were followed. The first one did not include any attention layer; the image and text features are simply concatenated and passed through a linear layer and finally a dense layer with softmax activation fn. The accuracy for that model on the validation set is around 53-54%. The second approach includes adding an attention layer between the images and questions in the network to answer the questions more appropriately by focusing on the required and specific parts of the image. The accuracy increased to 67-68%. The model could only be trained on 700 samples for both the approaches in this model due to computational inefficiency(collab notebook crashing when trained on more no. of samples and PC running out of RAM & storage.)

Reference paper:https://arxiv.org/pdf/1511.02274.pdf

Other literature and papers around VQA-RAD that we used:

1)https://arxiv.org/pdf/2305.10415v5.pdf-PMC-VQA
The research report presents a comprehensive analysis of the PMC-VQA dataset, a large-scale and diverse MedVQA dataset constructed using a scalable pipeline. The dataset comprises 227k image-question pairs, covering various modalities and

diseases. The proposed method, PMC-VQA, utilizes a generative learning approach by aligning a pre-trained vision encoder with a large language model through visual instruction tuning. The architecture of PMC-VQA consists of a visual encoder, a language encoder, and a multimodal decoder. The model achieves state-of-the-art performance on existing MedVQA datasets and outperforms other models. The report highlights the importance of multimodal understanding and the challenges faced by general visual-language models in the medical domain.

References

1. Vision–Language Model for Visual Question Answering in Medical Imagery
2. A Question-Centric Model for Visual Question Answering in Medical Imaging
3. Knowledge-Enhanced Medical Visual Question Answering
4. Attention Based VQA Methods
5. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering
6. Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing