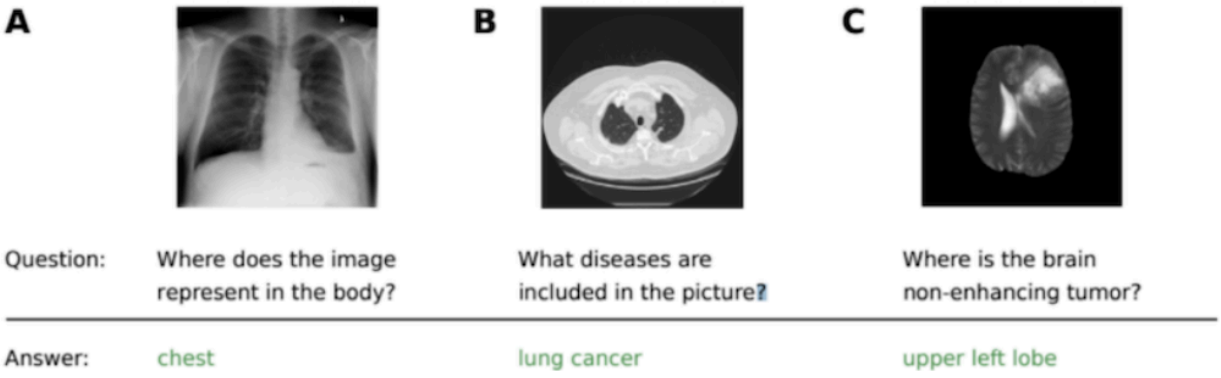
DEEP LEARNING ASPECTS OF

SWIN-VQA

A Caption Pre-Trained Swin Transformer for Visual Question Answering in Radiology

Submitted By: Purusharth Malik (2348542)

Submitted To: Dr. Prabu P.



Swin-VQA tackles the task of Visual Question Answering (VQA) using a combination of deep learning techniques. Here's a breakdown of its key aspects:

Input Encoding

Image: The model leverages CLIP (Contrastive Language-Image Pre-training) to encode the input image into a vector representation. CLIP learns a common embedding space where images and text descriptions are close to each other semantically.

Question: The question is encoded using Llama-2 13b, a pre-trained Transformer model. This model transforms the text question into another vector representation, capturing word meanings and relationships.

Fusion and Decoding

Concatenation: The encoded image and question vectors are then concatenated, creating a single vector that combines visual and textual information.

Custom Decoder Layer: This layer is responsible for processing the concatenated vector and generating the answer. It utilizes two key techniques:

- Rotary Positional Embedding: This technique injects positional information into the vector without relying on learned positional encodings, improving efficiency.
- Window Attention: This attention mechanism allows the model to focus on specific parts of the combined vector representation, potentially attending more to relevant image regions based on the question.

Output and Training

Llama Generator: The processed vector from the decoder layer is fed into a Llama generator, likely a fine-tuned version of the Llama-2 13b model. This generator outputs logits, which are used to predict the most likely next token in the answer sequence.

Loss and Optimization: The predicted answer is compared to the ground truth answer to calculate the loss. This loss function is used to optimize the model parameters during training.

LoRA (Low-Rank Adaptation): This technique is employed for parameter-efficient fine-tuning. It allows the model to adapt to the VQA task without requiring a large number of additional parameters.

Implementation Details

The code for Swin-VQA is built using PyTorch, a popular deep learning framework, and leverages the Transformers library for efficient implementation of the encoder, decoder, and attention mechanisms.

Overall, Swin-VQA combines pre-trained models for image and text encoding with a custom decoder layer that incorporates positional information and focused attention. This architecture allows the model to effectively reason about the relationship between the image and the question to generate accurate answers.