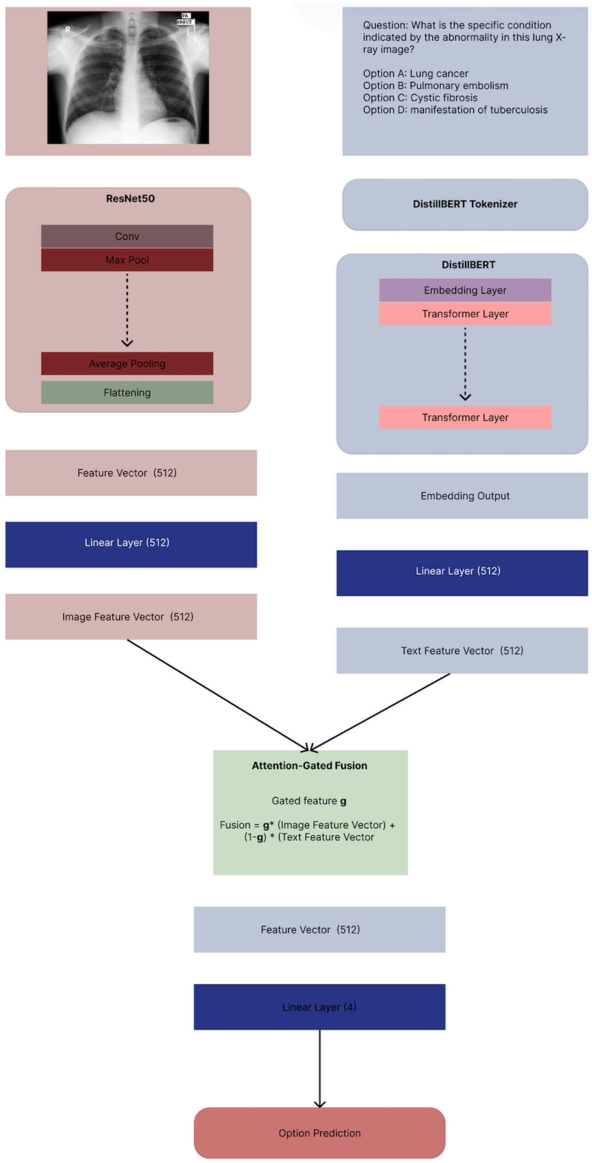
OmniMedVQA Image Test Fusion Model

1. Introduction

The dataset OmniMedVQA consists of medical vision question answering tasks. The tasks consists of predicting the disease given a image of the specific region, a question, and up to four possible disease options. To complete this task, we must design a multimodal model that can use image input and test input to determine which option is the best to use.

2. Novel Model

2.1 Model Architecture


Question: What is the specific condition indicated by the abnormality in this lung X-ray image?

Option A: Lung cancer
Option B: Pulmonary embolism
Option C: Cystic fibrosis
Option D: manifestation of tuberculosis

To create a multimodal model, we decided to use deep learning image and text models, extract features from each modality, and fuse them to create a unified representation.

To obtain the image features, we use a ResNet50 model where multiple layers of convolutions and pooling are performed to produce an output of (512,1,1) for each image. We remove the final fully connected layer in the ResNet50 model to obtain an embedding representation, which is then passed to a linear layer to produce a feature vector of size 512, our chosen fusion dimension.

Once we obtain the feature representation of the image, we obtain the text feature representation. We tokenize the text of the question and the four options. Then, using DistillBert, we produce an embedding of the text which we apply a linear layer of 512 nodes to get a text feature representation of size 512.

Once we have two separate feature vectors of the same size, we can combine them by using attention-gated fusion, where a gate parameter, restricted to be between 0 and 1, weighs the importance of the image feature and the text feature before combining them. The gate parameter g is learned during model training as part of the network, typically by passing the concatenated image and text features through a sigmoid-activated linear layer, allowing the model to adaptively balance the contributions of each modality. The formula below shows how we combine the two features to obtain y, our combined feature representation of image and text.

This combined feature is then passed through a linear layer of four nodes to obtain the probability for each option. We select the option with the highest probability.

$$y = g \cdot x_{\text{img}} + (1 - g) \cdot x_{\text{text}}$$

2.2 Training Design

The image is trained by

2.3 Design Rationale

3. Model Results

4. Conclusion