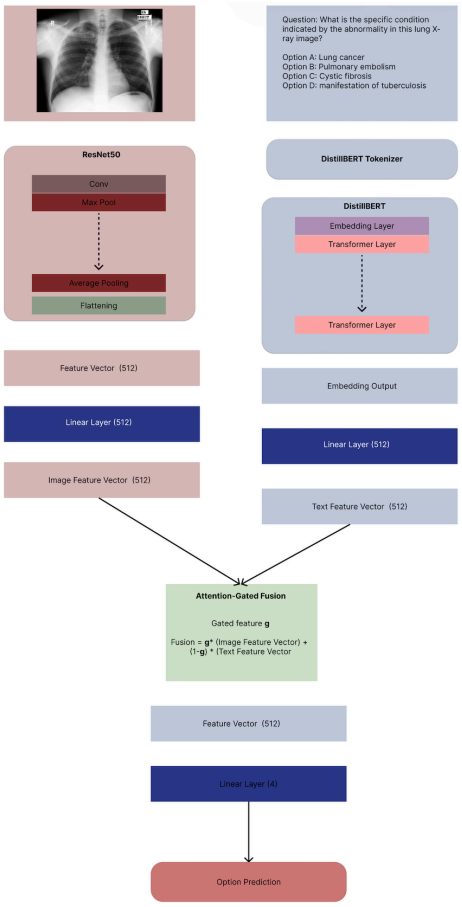
OmniMedVQA Image-Text Fusion Model

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 text input to determine which option is the best to use.

2. Novel Approach

2.1 Model Architecture



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

To train our multimodal model, we used the OmniMedVQA dataset, which includes an image, a question about that image, and four possible diagnosis options. The model's goal is to pick the correct diagnosis based on both the image and the text input.

Training Steps:

1. **Image Preprocessing:**
   Each image was resized to 224×224 and normalized using ImageNet mean and standard deviation.

We used a pretrained ResNet50 model (with the final classification layer removed) to extract image features. The ResNet weights were frozen at first, then fine-tuned later along with the text model.

2. **Text Preprocessing:**
   The question and each disease option were combined into one text input (e.g., *"Question: … Option: …"*). We tokenized this using the DistilBERT tokenizer with a max length of 128 tokens and padded shorter sequences.

3. **Feature Fusion and Classification:**
   Both the image and text embeddings were projected down to 512 dimensions.
   We then fused them using an attention gate that learns how much weight to give each feature.
   The fusion works like this:

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

   where

$$g = \sigma(W[x_{\text{img}}, x_{\text{text}}] + b)$$

   Here, g is a learned gating value between 0 and 1, and σ is the sigmoid activation. The gate parameter g was initialized with an image bias of 0.7, encouraging the model to rely more on visual information early in training while retaining flexibility to adjust during learning. The fused vector is then passed through a final linear layer with four outputs (one per disease option). The model picks the option with the highest softmax probability.

4. **Optimization:**
   We trained the model with the Adam optimizer (learning rate 1×10^{-4}) and cross-entropy loss.
   Early stopping was used to prevent overfitting. Training ran for 50 epochs with a batch size of 32.

5. **Evaluation:**
   After training, we measured accuracy, macro-averaged precision, recall, and F1-score on the test set.

2.3 Design Rationale

The main goal is to leverage image and text information to assist doctors in efficiently identifying potential diagnoses, reducing the time required for manual analysis. In testing we noticed that a lot of the decision comes more so from the image rather than the questions themselves as the images can be much more detailed than a simple question/answer combination.

- **ResNet50 for Image Features:**
  ResNet50 gives a strong visual backbone that can recognize medical patterns like tissue textures or visible abnormalities. It's a reliable model with good generalization from pretrained ImageNet weights.

- **DistilBERT for Text Understanding:**
  DistilBERT is smaller and faster than regular BERT but still captures important language patterns, which helps the model understand questions and disease options efficiently.

- **Attention-Gated Early Fusion:**
  Since many questions in OmniMedVQA were repetitive or ambiguous, most of the discriminative information came from the image rather than the text. To reflect this, we used an attention-gated early fusion mechanism that learns how much to rely on image versus text features for each prediction. The gate adaptively balances the two modalities, allowing the model to emphasize visual information while still incorporating useful textual cues when available.

- **512-Dimensional Shared Space:**
  Keeping both image and text embeddings the same size (512) makes fusion simpler and keeps the model lightweight while maintaining strong feature representation.

Overall, this setup balances simplicity, interpretability, and strong multimodal performance.

---

3. Model Results

After training the model on the OmniMedVQA dataset, we got the following results on the test set:

These results show that the model performs consistently across all classes. It correctly predicts diseases even when visual differences are subtle, which means the attention-gated fusion is helping the model learn which cues matter most for each question.

Because the ResNet50 model has so many parameters it was necessary to use a large computing cluster such as HiPerGator to decrease the amount of time it takes for training.

---

4. Conclusion

In this project, we built a multimodal model that combines image and text information to answer medical visual questions using the OmniMedVQA dataset. The model

Table 1: Test set performance metrics for the multimodal model.

oprule Metric	Value
Accuracy	0.9526
Precision (Macro)	0.9586
Recall (Macro)	0.9498
F1-Score (Macro)	0.9536

used a pretrained ResNet50 network to extract image features and a DistilBERT model to process the questions and disease options. We then fused these two feature types using a learned attention gate, which helped the model decide how much weight to give to the visual versus textual information.

Training was done with the Adam optimizer and cross-entropy loss, and performance was measured using accuracy, precision, recall, and F1-score. The final model performed well across all metrics, showing that combining both image and text features leads to more accurate and meaningful predictions in medical question-answering tasks.