# THE PROBLEM

* vqa là 1 task sẽ yêu cầu trả lời câu hỏi được đặt ra cho 1 bức ảnh, bằng các xem xét cả thông tin hình ảnh và ngôn ngữ
* Tương tự như VQA, medical vqa tập trung vào việc sử dụng để trả lời các câu hỏi liên quan đến hình ảnh y tế
* Tầm Quan Trọng của ứng dụng Medical VQA:
* Hỗ trợ Chẩn đoán và Điều Trị,
* Tiết Kiệm Thời Gian và Nguồn Lực: Việc tự động hóa quá trình phân tích hình ảnh và trả lời các câu hỏi
* Tiếp Cận Chăm Sóc Sức Khỏe
* Thách Thức trong Medical VQA => data medical vqa tương đối ít
* **Độ Phức Tạp Của Dữ Liệu**: Hình ảnh y tế thường có độ phức tạp cao và yêu cầu khả năng phân tích chính xác các chi tiết nhỏ, nên cần sự giám sát của bác sĩ
* Bảo Mật và Quyền Riêng Tư của dữ liệu y tế
* Các phương thức để đánh giá:
* closed accuracy: yes/no
* open accuracy: 1-2 năm trước, câu trả lời dạng multi classification, còn model SOTA gần đây chuyển qua generate

# THE SOLUTION

### MUMC Architecture

Từ phần trình bày của Trung, ta nhận thấy các mô hình được training để giải quyết bài toán med-VQA chỉ có thể được training trên một lượng dữ liệu rất nhỏ, do đó để cải thiện performance. Quá trình train model MUMC sẽ có 2 phase là pre-trained phase và fune-tune phase

Trong đó pre-trained phase sẽ được train trên tập data medical image-captioning (đối với MUMC thì là sử dụng ROCO, MedICaT và the ImageCLEF2022 Image Caption Dataset) là những bộ dataset khá lớn với lần lượt là 80k image-caption pairs cho ROCO, 217k pairs cho MadICat và 90k ảnh cho ImageCLEF2022.

Sau quá trình pre-train, pre-trained model sẽ được sử dụng cho downstream medical VQA task.

Ở bên pre-train phase, model có 12 encoder layers ViT để trích xuất đặc trưng của dữ liệu image và 6 encoder layers Transformers với initial parameters được transfer từ mô hình BERT tiêu chuẩn.

Để việc học được hiệu quả hơn, MUMC sử dụng momentum model, ý tưởng của nó là duy trì một bản sao của Image encoder và Text encoder, các tham số của momentum model sẽ được cập nhật từ tham số của mô hình với một momentum rate, từ đó giúp mô hình ổn định hơn và duy trì tính nhất quán của mô hình từ đó giảm được ảnh hưởng bởi các nhiễu có trong data.

Sau khi đi qua các lớp encoders, chúng ta sẽ tính UCL(image), MCL và UCL(text), công tính tính các loss này sẽ được nói trong phần sau của bài present.

Cuối cùng, ta sẽ đi đến một encoder để encode những thông tin trích xuất được từ các encoder trước đó. Tuy nhiên như mn thấy thì chỉ có features từ image encoder được truyền đến multimodel encoder, điều này là do quá trình pre-trained được thực hiện trên bài toán image-captioning và chủ yếu nhắm vào đào tạo tham số cho hai cái image encoder và text encoder, và các cái loss cũng chủ yếu nhắm vào quá trình đào tạo này.

Ở phase fine tuning, model parameters sẽ được init từ pre-trained model. Ở trong phase này model sẽ có thêm một phần đó là decode câu trả lời từ câu hỏi. The downstream VQA model is fine-tuned via the masked language model (MLM) loss, using ground-truth answers as targets

### 16x16

Đối với các image đầu vào, các bức ảnh sẽ được chia thành các patches 16x16 rồi sau đó sẽ mask 25% trong số chúng lại trước khi cho vào mô hình ViT.

### Unimodal and Multimodal Contrastive Losses

Hai hàm loss được sử dụng trong quá trình pre-training, với hai mục đích khác nhau:

* Hàm Unimodal Contrastive Losses có objective là tối đa hóa khả năng phân biệt trong cùng một modality (đơn phương thức) với mục đích là phân biệt những data point trong cùng một phương thức, cụ thể là nếu hai hình ảnh (hoặc text) đang mô tả chung một cái gì đó thì cái lantern variable sẽ nên có khoảng cách gần nhau và ngược lại nếu hai bức ảnh đang mô hả hai thứ khác nhau hoàn toàn thì điều đó cũng nên được mô tả thông qua độ khác biệt giữa hai lantern vector. Hàm này tập trung vào việc xác định hai bức ảnh (hoặc text) là positive hay negative, nghĩa là ngữ nghĩa được biểu diễn có giống nhau hay không. Và ở trong hàm loss này, để so sánh hai lantern vector thì ta sử dụng cosine similarity. Giá trị nằm ở mẫu số tau là một learnable temperature parameters, có tác dụng điều chỉnh độ nhọn của phân phối xác suất khi sử dụng hàm softmax. Khi tau nhỏ thì sự phân biệt giữa xác suất của các giá trị trong vector sau khi đi qua hàm softmax sẽ lớn hơn, nghĩa là giá trị được chọn sẽ gần 1 hơn trong khi các giá trị không được chọn sẽ gần 0 hơn. Điều ngược lại xảy ra khi giá trị tau nhỏ, khi đó giá trị của vector sau hàm softmax sẽ gần nhau hơn, ví dụ như 0.6 và 0.4. Hàm loss này được tính bằng cách sử dụng cross entropy loss giữa ground truth và giá trị dự đoán của mô hình. Cái Expectation ở đầu ý là lấy giá trị trung bình sau khi tính loss giữa các điểm dữ liệu.
* Hàm multimodal Contrastive Losses, về mặt cơ bản cũng có cách hoạt động tương tự với Unimodal Contrastive Losses, điều duy nhất khác biệt giữa hai hàm này đó là như tên gọi của nó - Unimodal sẽ tính loss giữa images và images, text và text trong khi Multimodal sẽ so sánh giữa image và text, text và image.

### Image Text matching

Sau cái block decoding, ta sẽ có thêm hai hàm loss là ITM và MLM. Đầu tiên ta sẽ nói về ITM hay Image Text Matching. ITM nhắm đến objective là dự đoán cặp hình ảnh/ text có khớp với nhau hay không, tức là có positive hay không. Do chỉ mang giá trị là positive hay negative nên ITM là một một binary classification task được thực hiện bằng cách sử dụng một fully connected layer được gọi là ITM head. Rồi sau đó Loss ITM được tính bằng cách sử dụng công thức bên dưới. Trong đó H là hàm cross entropy, y\_itm là vector one-hot encoder của ground truth label và p\_itm là hàm dự đoán (cái FC layers)

### Masked Language Modeling

MLM nhắm đến việc dự đoán phần bị che trong text. Với mỗi caption thì 15% token sẽ bị che, việc dự đoán từ bị che sẽ dựa trên thông tin thu được thông qua image và text ở trong multimodal encoder. Cũng tương tự như ITM, làm loss MLM được tính bằng hàm cross entropy với y\_mlm đại diện cho ground-truth và p\_mlm là dự đoán của model

# THE RESULT

# THE FUTURE WORK

#### BLIP >< MUMC

BLIP (Bootstrapping Language-Image Pre-training):

* Sử dụng kỹ thuật bootstrap để cải thiện chất lượng dữ liệu huấn luyện
* Có khả năng thực hiện nhiều tác vụ như tạo chú thích hình ảnh, truy vấn hình ảnh bằng ngôn ngữ tự nhiên
* ITC là UCL trong MUMC, không có MCL
* Sử dụng LM thay vì MLM: LM cho phép mô hình với khả năng generalization từ thông tin hình ảnh thành chú thích mạch lạc.

MUMC (Multimodal Understanding via Multimodal Contrastive learning):

* Sử dụng phương pháp học đối nghịch (MCL UCL)
* Tập trung vào việc hiểu mối quan hệ giữa các phương thức khác nhau