ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN CUỐI KỲ

MÔN CS406 – XỬ LÝ ẢNH VÀ ỨNG DỤNG

Đề tài: Image captioning

Sinh viên thực hiện: Trần Ngọc Thiện

MSSV: 21521465

Giảng viên hướng dẫn: Cáp Phạm Đình Thăng

Mục lục

1. Giới thiệu

Bài toán image captioning (tạo chú thích cho hình ảnh) là một lĩnh vực trong thị giác máy tính và xử lý ngôn ngữ tự nhiên. Đây là một bài toán phức tạp, đòi hỏi máy tính không chỉ hiểu được nội dung của hình ảnh mà còn có khả năng diễn đạt chúng bằng ngôn ngữ tự nhiên.

1.1. Mô tả bài toán

- Đầu vào:
    - Tập huấn luyện: { (x, y)n } gồm n mẫu dữ liệu, mỗi mẫu gồm 1 ảnh đầu vào và 1 câu mô tả.
        - x: hình ảnh
        - y: câu mô tả nội dung ảnh tương ứng
        - n: số mẫu dữ liệu trong tập huấn luyện
    - X: ảnh đầu vào
- Đầu ra: Y: câu mô tả nội dung hình ảnh

1.2. Thách thức của bài toán

- Nhiều cách diễn đạt
- Chi tiết phức tạp
- Ngữ cảnh
- Hình ảnh có nhiều đối tượng, các mối quan hệ phức tạp giữa các đối tượng
- Hình ảnh có đối tượng nhỏ hoặc bị che khuất

1.3. Ứng dụng của bài toán

- Tìm kiếm hình ảnh dựa trên văn bản
- Hỗ trợ người khuyết tật
- Truyền thông xã hội

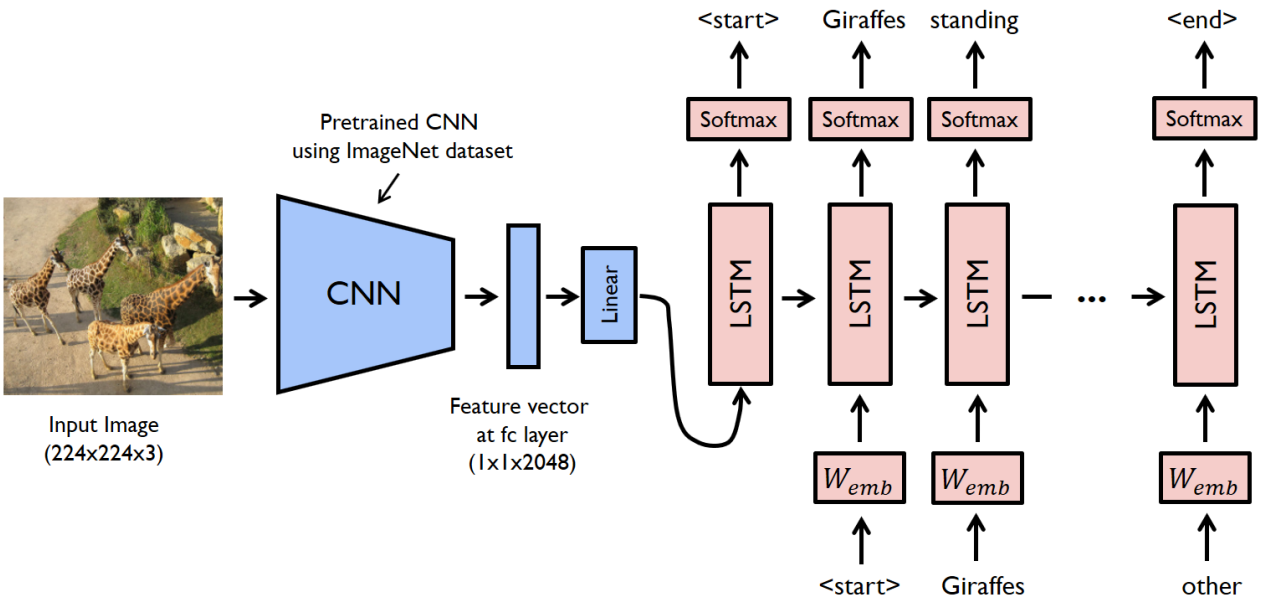
- Phân tích hình ảnh y tế

2. Các công trình liên quan

- UIT-ViIC: A Dataset for the First Evaluation on Vietnamese Image Captioning
- Show and Tell: A Neural Image Caption Generator
- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

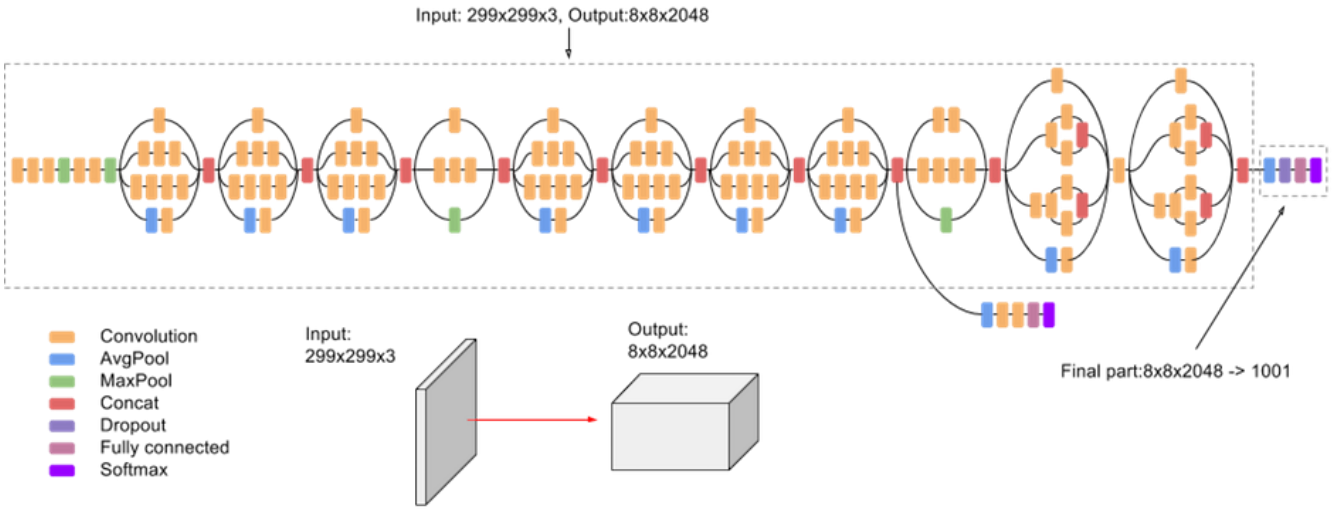
3. Phương pháp đề xuất

3.1. Phương pháp CNN + LSTM



Phương pháp này sử dụng mạng InceptionNetv3 được huấn luyện trước trên bộ dữ liệu ImageNet để trích xuất đặc trưng hình ảnh và một mạng neurol dựa trên LSTM để tạo mô tả cho ảnh.

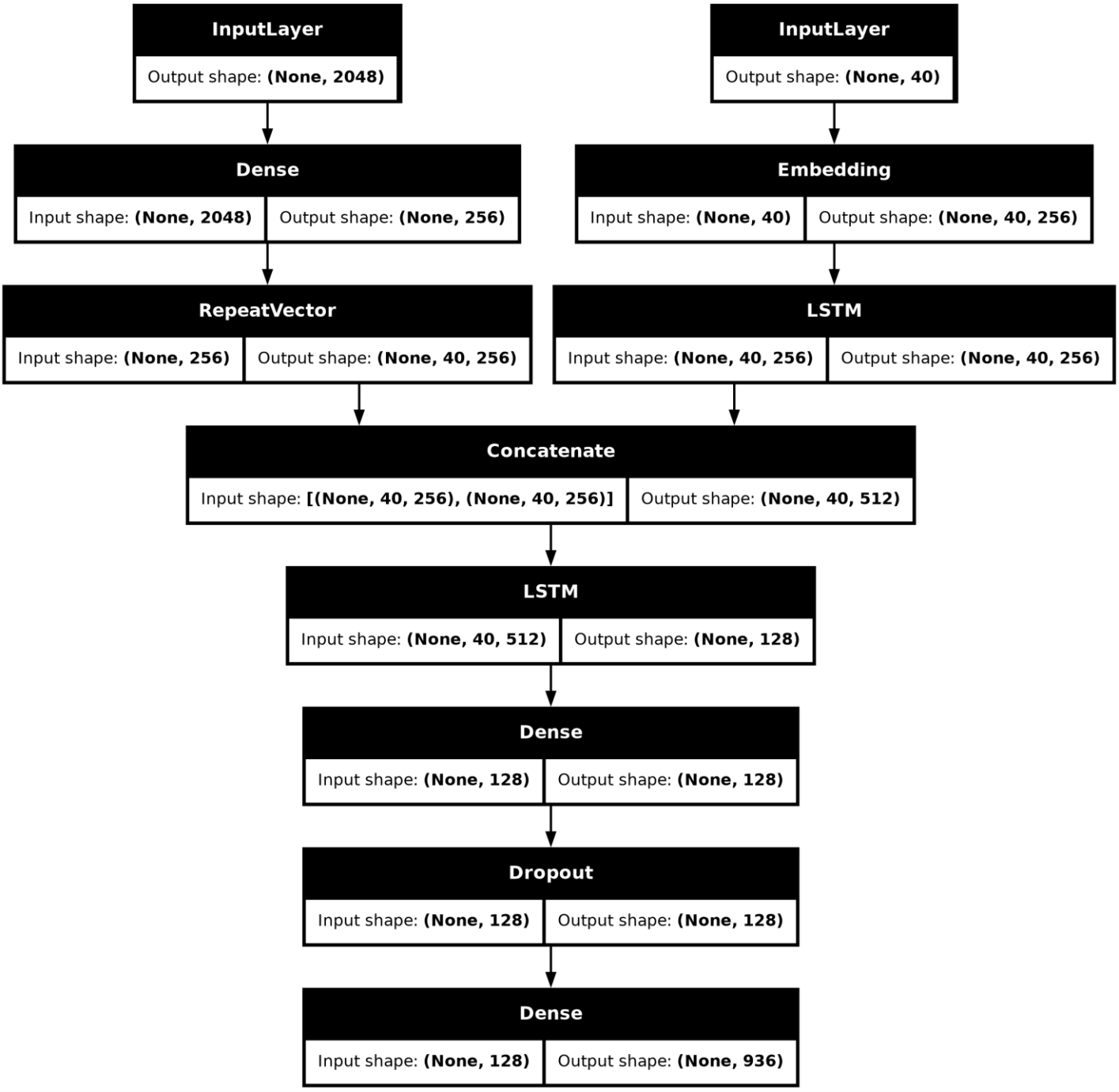
3.1.1. Mạng InceptionNetv3



Mạng InceptionV3 là một kiến trúc mạng nơ-ron sâu được thiết kế để nhận diện hình ảnh, nổi bật với khả năng xử lý và phân tích đặc trưng hình ảnh hiệu quả. Cấu trúc của InceptionV3 bao gồm nhiều khối Inception, mỗi khối tích hợp các phép biến đổi khác nhau (như convolution với các kích thước kernel khác nhau) để thu thập thông tin từ nhiều mức độ chi tiết khác nhau trong hình ảnh. Điều này giúp mạng có khả năng học được các đặc trưng phong phú và đa dạng.

Trong mô hình chung, mạng InceptionV3 được sử dụng như một **encoder** để chuyển đổi hình ảnh thành các vector đặc trưng. Các vector này sau đó được đưa vào mạng **decoder** để tạo ra các câu mô tả cho hình ảnh đó.

3.1.2. Mạng decoder



Decoder là một mạng neurol tự thiết kế với cấu trúc như sau:

- Đầu vào văn bản (bên phải):

- InputLayer: Lớp này nhận vào một chuỗi các từ (token) biểu diễn cho câu mô tả hình ảnh hiện có.
- Embedding: Lớp này chuyển đổi các từ (token) thành các vector đặc trưng có chiều cao hơn, giúp mạng học được mối quan hệ ngữ nghĩa giữa các từ.
- LSTM: Lớp LSTM (Long Short-Term Memory) được sử dụng để xử lý tuần tự thông tin trong chuỗi từ.

- Đầu vào hình ảnh (bên trái):
  - InputLayer: Lớp này nhận vào một vector đặc trưng được trích xuất bởi mạng encoder InceptionNetV3.
  - RepeatVector: Lớp này lặp lại vector đặc trưng của hình ảnh để có cùng chiều với đầu ra của lớp LSTM, tạo điều kiện cho việc kết hợp thông tin từ hình ảnh và văn bản.
- Concatenate: Lớp này kết hợp vector đặc trưng của hình ảnh và đầu ra của LSTM, tạo ra một vector đặc trưng mới chứa cả thông tin thị giác và ngữ nghĩa.
- LSTM (2): Lớp LSTM thứ hai tiếp tục xử lý thông tin kết hợp, giúp mạng học được mối quan hệ phức tạp giữa hình ảnh và văn bản.
- Dense: Các lớp Dense (fully-connected) được sử dụng để dự đoán từ tiếp theo trong câu mô tả. Lớp Dense cuối cùng có số lượng node bằng với số lượng từ vựng, mỗi node đại diện cho xác suất của một từ cụ thể.
- Dropout: Lớp này được sử dụng để ngăn chặn hiện tượng overfitting bằng cách ngẫu nhiên bỏ đi một số kết nối giữa các neurol.

3.1.3. Cách hoạt động của mạng

Encoder (InceptionNetV3): Hình ảnh đầu vào được đưa vào mạng InceptionNetV3 để trích xuất các đặc trưng thị giác.

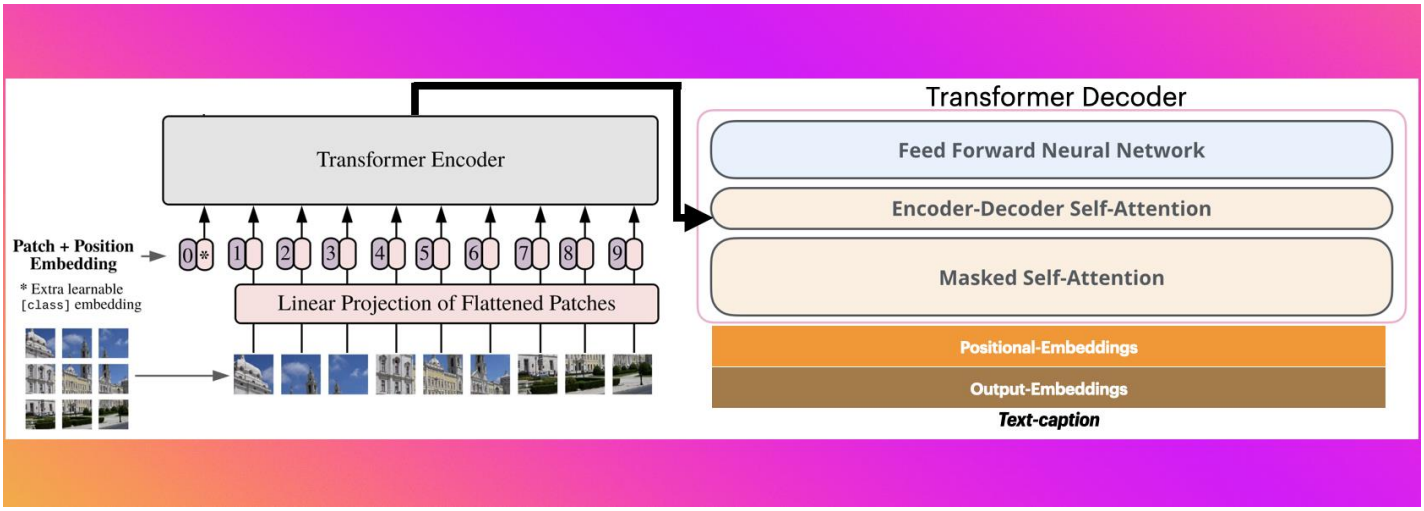
Decoder:

- **Khởi tạo:** Quá trình tạo câu mô tả bắt đầu bằng một token đặc biệt (ví dụ: <start>).
- **Lặp:**
    - **Embedding:** Token hiện tại được chuyển thành vector embedding.
    - **LSTM:** Vector embedding được đưa vào lớp LSTM cùng với trạng thái ẩn của lớp LSTM ở thời điểm trước.
    - **Kết hợp:** Đầu ra của LSTM được kết hợp với vector đặc trưng của hình ảnh.
    - **Dự đoán:** Vector kết hợp được đưa vào các lớp Dense để dự đoán xác suất của từng từ trong từ vựng.
    - **Chọn từ:** Từ có xác suất cao nhất được chọn làm từ tiếp theo trong câu mô tả.
    - **Cập nhật trạng thái:** Trạng thái ẩn của LSTM được cập nhật cho thời điểm tiếp theo.
- **Kết thúc:** Quá trình lặp lại cho đến khi gặp token kết thúc (<end>) hoặc đạt đến độ dài tối đa của câu.

3.1.4. Ưu nhược điểm

- Ưu điểm:
    - Kiến trúc dễ hiểu và triển khai.
    - Đã được chứng minh hiệu quả trên nhiều bài toán khác nhau.
- Nhược điểm:
    - Khó xử lý các chuỗi dài do vấn đề gradient vanishing/exploding.
    - Không tận dụng tốt thông tin toàn cục: RNN chỉ xử lý thông tin theo trình tự tuần tự, khó nắm bắt các mối quan hệ giữa các từ cách xa nhau trong câu.

3.2. Phương pháp transformer



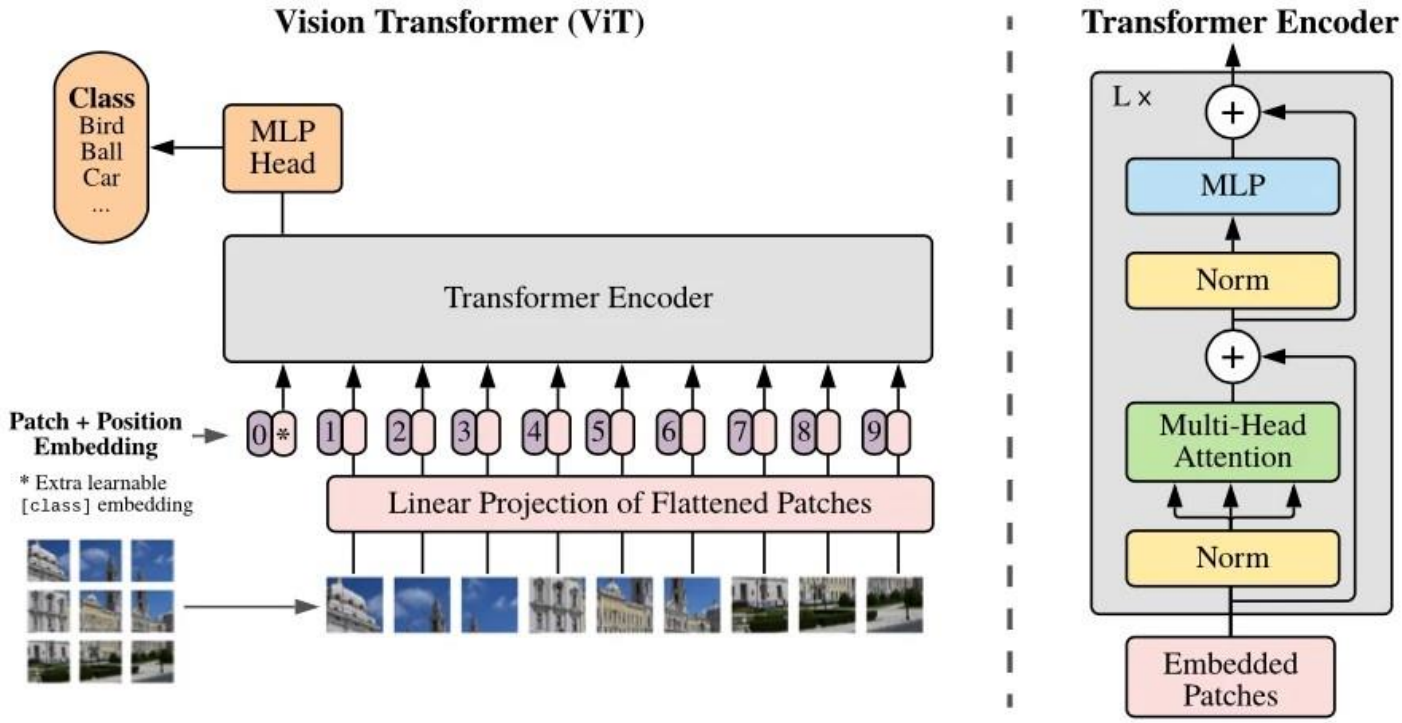
Transformer là một mạng neurol cải tiến sử dụng cơ chế attention giúp nắm bắt tốt các mối quan hệ phức tạp của dữ liệu. Trong phương pháp này, mạng vision transformer được dùng làm encoder và gpt-2-vietnamese làm decoder.

3.2.1. Vision transformer

Mạng Vision Transformer (ViT) là một kiến trúc mạng nơ-ron sâu được thiết kế để xử lý hình ảnh bằng cách sử dụng các khối transformer, nổi bật với khả năng học các đặc trưng không gian mà không cần các phép biến đổi convolution truyền thống. Cấu trúc của ViT chia hình ảnh thành các đoạn (patches) nhỏ và chuyển đổi chúng thành các vector nhúng, sau đó áp dụng các cơ chế attention để xác định mối quan hệ giữa các đoạn này. Điều này giúp ViT nắm bắt được các đặc trưng hình ảnh một cách hiệu quả và linh hoạt.

Trong mô hình chung, mạng ViT được sử dụng như một **encoder** để chuyển đổi hình ảnh thành các vector đặc trưng. Các vector này sau đó được đưa vào mạng **decoder** để tạo ra các câu mô tả cho hình ảnh đó.

3.2.2. GPT-2-Vietnamese



Mạng GPT-2 là một mô hình ngôn ngữ dựa trên kiến trúc transformer với khả năng sinh văn bản tự nhiên một cách mạch lạc và phong phú. GPT-2 sử dụng cơ chế attention để xử lý các đầu vào tuần tự, cho phép nó ghi nhớ và liên kết thông tin từ các bước trước đó trong quá trình sinh câu. GPT-2-Vietnamese là phiên bản mô hình GPT-2 được huấn luyện bổ sung (finetune) trên dữ liệu tiếng Việt.

Vai trò chính của GPT-2 là tạo ra văn bản mô tả từ các đặc trưng hình ảnh mà ViT đã trích xuất. Nhờ vào khả năng hiểu ngữ nghĩa và cấu trúc ngôn ngữ, GPT-2 có thể tạo ra những câu mô tả không chỉ chính xác mà còn tự nhiên và phong phú.

3.2.3. Ưu nhược điểm

- Ưu điểm:
    - Xử lý song song

    - Cơ chế attention giúp Transformer nắm bắt các mối quan hệ giữa các từ cách xa nhau trong câu.
    - Hiệu suất cao
- Nhược điểm:
    - Phức tạp
    - Tốn tài nguyên tính toán

4. Thực nghiệm

4.1. Phương pháp đánh giá BLEU-4

Độ đo BLEU-4 (Bilingual Evaluation Understudy) là một chỉ số phổ biến trong đánh giá chất lượng của các hệ thống sinh ngôn ngữ tự nhiên. BLEU-4 so sánh các câu sinh ra với các câu tham chiếu (ground truth) bằng cách tính toán tỷ lệ giữa số lượng n-gram (chuỗi liên tiếp của n từ) trùng khớp giữa câu sinh và câu tham chiếu.

Cách hoạt động:

- N-gram Matching: BLEU-4 tính toán số lượng n-gram trùng khớp cho n = 1, 2, 3, và 4. Mỗi loại n-gram được tính điểm riêng.
- Precision Calculation: Tỷ lệ chính xác được tính bằng cách chia số lượng n-gram trùng khớp cho tổng số n-gram trong câu sinh.
- Geometric Average Precision: Đuọc tính bằng tích của từng n-gram precision lũy thừa với trọng số.

$$
\begin{aligned}
Geometric\ Average\ Precision\ (N) &= exp(\sum_{n=1}^{N} w_n\ log\ p_n) \\
&= \prod_{n=1}^{N} {p_n}^{w_n} \\
&= (p_1)^{\frac{1}{4}} \cdot (p_2)^{\frac{1}{4}} \cdot (p_3)^{\frac{1}{4}} \cdot (p_4)^{\frac{1}{4}}
\end{aligned}
$$

- Brevity Penalty: Để khuyến khích độ dài câu sinh gần với câu tham chiếu, một hình phạt độ dài (brevity penalty) được áp dụng nếu câu sinh ngắn hơn câu tham chiếu.

$$
Brevity\ Penalty = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c <= r \end{cases}
$$

- BLEU-4: Chỉ số BLEU-4 là tích của geometric average precision của các n-gram precision 1-4 và brevity penalty.

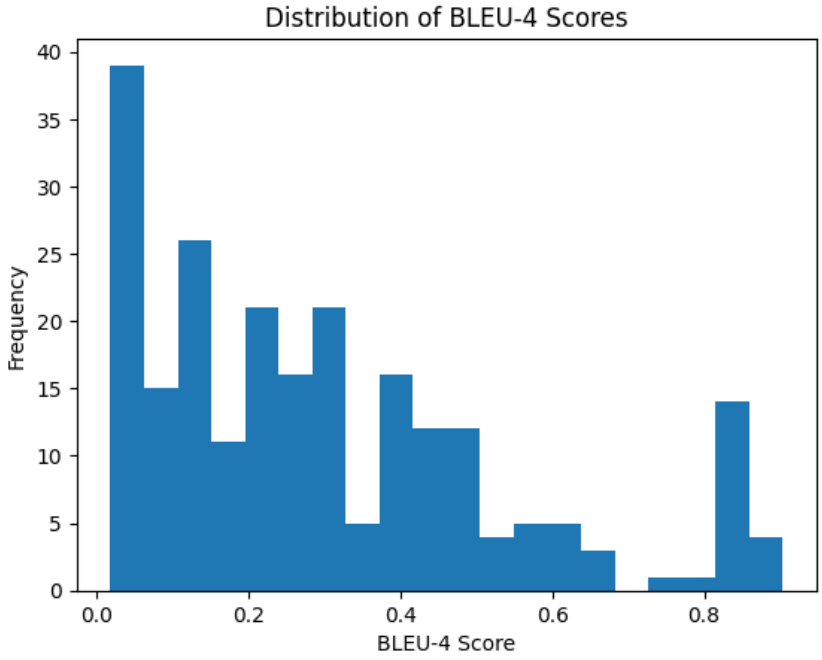
$$
Bleu\ (N) = Brevity\ Penalty \cdot Geometric\ Average\ Precision\ Scores\ (N)
$$

4.2. Bộ dữ liệu UIT-ViIC

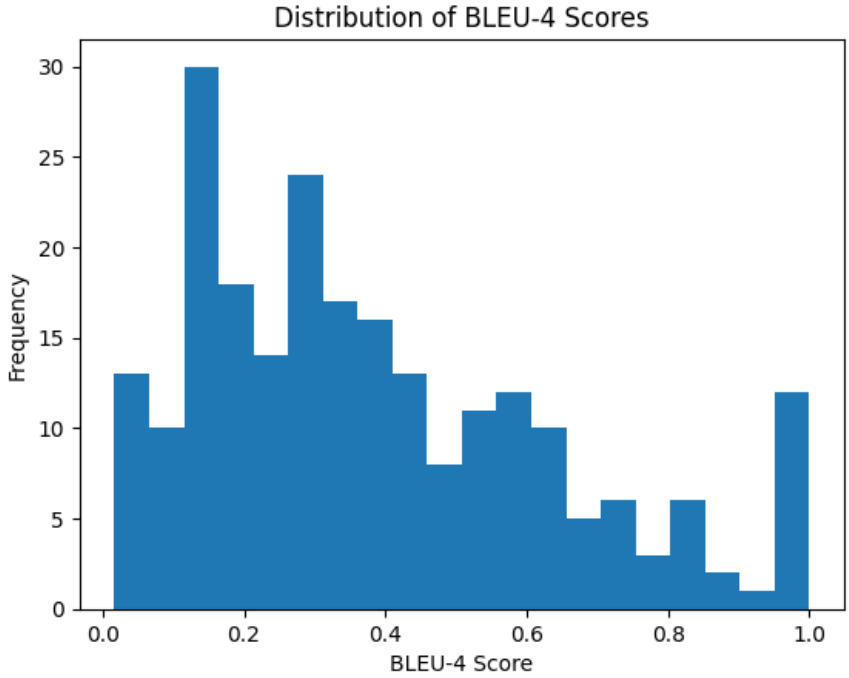
UIT-ViIC là bộ dữ liệu chứa các chú thích được viết thủ công cho các hình ảnh từ tập dữ liệu Microsoft COCO liên quan đến các môn thể thao với bóng. UIT-ViIC bao gồm 19.250 chú thích tiếng Việt cho 3.850 hình ảnh.

4.3. Kết quả

Mô hình CNN-LSTM đạt điểm BLEU-4 trên tập kiểm thử là **0.2978** với phân bố như sau:



Mô hình ViT-GPT-2 đạt điểm BLEU-4 trên tập kiểm thử là **0.3875** với phân bố như sau:



5. Kết luận

- Mô hình ViT-GPT-2 cho kết quả tốt hơn nhưng có thời gian chạy lâu hơn mô hình CNN-LSTM.
- Bộ dữ liệu có đa phần hình ảnh thuộc các môn bóng đá, bóng chày và tennis nên cả 2 mô hình đều không hoạt động tốt với các hình ảnh không thuộc các môn trên. => Cả 2 phương pháp đều phụ thuộc vào dữ liệu huấn luyện.

6. Tài liệu tham khảo

- LAM, Quan Hoang, et al. Uit-viic: A dataset for the first evaluation on vietnamese image captioning. In: *International Conference on Computational Collective Intelligence*. Cham: Springer International Publishing, 2020. p. 730-742.
- VINYALS, Oriol, et al. Show and tell: A neural image caption generator. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015. p. 3156-3164.
- DOSOVITSKIY, Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.