TRƯỜNG ĐẠI HỌC XÂY DỰNG HÀ NỘI
KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN KHOA HỌC MÁY TÍNH

BÁO CÁO BÀI TẬP LỚN
XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Đề tài: Ứng dụng mô hình CLIP và LSTM trong bài toán tạo chú thích ảnh

Giảng viên hướng dẫn: ThS. Nguyễn Đình Quý

Sinh viên thực hiện: Nguyễn Hải Cường - 0174067 - 67CS1
Lã Minh Khánh - 4004267 - 67CS1
Trịnh Quỳnh Anh - 0279367 - 67CS1
Phạm Hồng Thái - 0127067 - 67CS1

Hà Nội, ngày 14 tháng 05 năm 2025

Mục lục

Lời nói đầu

Chúng em xin chân thành cảm ơn thầy Nguyễn Đình Quý đã trang bị những kiến thức quý báu cho chúng em trong suốt quá trình học tập môn học phần Xử lý ngôn ngữ tự nhiên - Natural Language Processing. Chính nhờ công giảng dạy, chỉ bảo tận tình của thầy mà chúng em mới có những kiến thức cơ bản để bước vào chuyên ngành và đặt những bước chân tiếp theo trên con đường học tập, vận dụng và sáng tạo ra những sản phẩm hữu ích góp phần phục vụ các lĩnh vực khác nhau của đời sốn

Chúng em xin chân thành cảm ơn thầy đã tận tình hướng dẫn, chỉ bảo cho chúng em trong suốt quá trình thực hiện đề tài này.

Chúng em xin gửi lời cảm ơn đến gia đình, bạn bè, các anh chị tiền bối đã động viên và cung cấp các tài liệu hữu ích cho nhóm để chúng em hoàn thành đề tài một cách chỉnh chu nhất

Mặc dù đã cố gắng nỗ lực thực hiện đề tài với quyết tâm cao nhưng bài làm chắc chắn không thể tránh khỏi thiếu sót. Chúng em rất mong nhận được những ý kiến đóng góp của các thầy để bài tập và kiến thức của chúng em ngày càng hoàn thiện hơn và rút kinh nghiệm trong những lần làm đề tài sau ạ.

Em xin chân thành cảm ơn!

Ngày 14 tháng 05 năm 2025

Từ các thành viên của nhóm

Lã Minh Khánh

Nguyễn Hải Cường

Phạm Hồng Thái

Trịnh Quỳnh Anh

Chương I: Giới thiệu

1. Giới thiệu đề tài

Đề tài ”Tạo chú thích hình ảnh tự động (Image Captioning)” tập trung vào việc kết hợp mô hình học sâu để phân tích hình ảnh và tạo mô tả ngôn ngữ tự nhiên phù hợp. Phương pháp này ứng dụng mô hình CLIP (Contrastive Language-Image Pretraining) của OpenAI để trích xuất đặc trưng hình ảnh hiệu quả, đồng thời sử dụng mạng LSTM (Long Short-Term Memory) để sinh chuỗi chữ mô tả nội dung hình ảnh dựa trên các đặc trưng đó.

Trong dự án này, chúng ta sẽ khám phá cách mô hình CLIP mã hóa hình ảnh thành các embedding đa chiều, sau đó đưa vào mạng LSTM kết hợp với xử lý ngôn ngữ tự nhiên (NLP) thông qua thư viện NLTK để tối ưu hóa quá trình tạo chú thích. Mô hình sẽ học cách ánh xạ giữa đặc trưng hình ảnh và từ ngữ, đồng thời tạo ra câu mô tả chính xác và tự nhiên.

Ứng dụng của Image Captioning rất đa dạng, từ hỗ trợ người khiếm thị, tự động gắn thẻ hình ảnh, đến nâng cao trải nghiệm tìm kiếm hình ảnh. Dự án này không chỉ cung cấp cái nhìn sâu sắc về cách kết hợp thị giác máy tính (Computer Vision) và xử lý ngôn ngữ tự nhiên (NLP), mà còn mở ra hướng phát triển các hệ thống AI đa phương thức (Multimodal AI) trong tương lai.

2. Mục tiêu, đối tượng và phạm vi

Mục tiêu

- Xây dựng mô hình Image Captioning tự động, kết hợp CLIP (ViT-B/32) để trích xuất đặc trưng hình ảnh và LSTM với cơ chế Attention để sinh chú thích.
- Tận dụng bộ dữ liệu Flickr8k (8,000 ảnh + chú thích) để huấn luyện và đánh giá mô hình.
- Ứng dụng NLTK để tiền xử lý văn bản (tokenization, stopwords removal) và đánh giá độ chính xác của caption (BLEU).

Đối tượng

- Bộ dữ liệu: Flickr8k (ảnh + chú thích tiếng Anh).
- Mô hình chính:
  - Encoder: CLIP (ViT-B/32) → trích xuất đặc trưng hình ảnh (global hoặc patch tokens).
  - Decoder: LSTM + Attention Mechanism → sinh caption dựa trên features từ CLIP.
- Công cụ: Python, PyTorch/TensorFlow, NLTK (xử lý ngôn ngữ).

Phạm vi

- Dữ liệu: Giới hạn ở Flickr8k (không mở rộng sang COCO hoặc Conceptual Captions).
- Mô hình:
  - Chỉ sử dụng CLIP ViT-B/32 (không so sánh với ResNet, Faster R-CNN).
  - Decoder dùng LSTM + Attention.
- Đánh giá: Tập trung vào độ chính xác (BLEU) và khả năng mô tả tự nhiên (qualitative analysis).

Chương II: Cơ sở lý thuyết

1. Image Captioning là gì?

Image Captioning (hay thuật toán sinh văn bản dựa theo ảnh) là một thuật toán giúp máy tính sinh ra một đoạn văn bản mô tả một ảnh. Đây là một bài toán trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing, NLP) và được ứng dụng rộng rãi trong các ứng dụng như dịch thuật, tìm kiếm hình ảnh, đánh giá hình ảnh, dịch thuật, v.v.

Thuật toán này thực hiện nhiệm vụ dự đoán chú thích cho một hình ảnh nhất định dựa theo các đặc điểm thuộc tính có trong bức ảnh. Các ứng dụng phổ biến trong thế giới thực của nó bao gồm hỗ trợ người khiếm thị có thể giúp họ điều hướng qua các tình huống khác nhau. Do đó, chú thích hình ảnh giúp cải thiện khả năng tiếp cận nội dung cho mọi người bằng cách mô tả hình ảnh cho họ.

Theo đó, thuật toán này gồm input-output như sau:

- Input: Một hình ảnh;
- Output: Một chuỗi các từ mô tả hình ảnh dựa theo các đặc điểm.

2. Ứng dụng

Image Captioning (tạo chú thích hình ảnh tự động) có nhiều ứng dụng thiết thực trong đời sống, công nghệ và nghiên cứu, bao gồm:

1. **Hỗ trợ người khiếm thị:** Mô tả hình ảnh tự động giúp người khiếm thị hiểu nội dung ảnh thông qua giọng nói (screen reader).

   Ví dụ: Ứng dụng như Seeing AI (Microsoft) sử dụng AI để mô tả cảnh vật, chữ viết, cảm xúc.

2. **Tìm kiếm hình ảnh thông minh:** Cải thiện kết quả tìm kiếm bằng cách hiểu nội dung ảnh thay vì chỉ dựa vào metadata hoặc tags.

   Ví dụ: Google Images sử dụng AI để phân tích và gán nhãn ảnh.

3. **Mạng xã hội & Nền tảng chia sẻ ảnh:** Tự động gợi ý chú thích (caption) cho ảnh đăng tải trên Facebook, Instagram.

   Ví dụ: Phát hiện nội dung không phù hợp (ảnh bạo lực, ảnh hưởng xấu).

4. **Y tế & Chẩn đoán hình ảnh:** Mô tả tự động kết quả X-quang, MRI, CT scan để hỗ trợ bác sĩ.

   Ví dụ: AI mô tả tổn thương trong ảnh y tế.

5. **Giáo dục & Học tập:** Tạo mô tả cho hình ảnh trong sách giáo khoa, tài liệu học tập.

   Ví dụ: Hỗ trợ học ngoại ngữ (ghép ảnh với từ vựng).

6. **An ninh & Giám sát:** Tự động mô tả sự kiện trong camera an ninh.

   Ví dụ: Phát hiện hành vi đáng ngờ qua phân tích hình ảnh.

7. **Thương mại điện tử:** Tự động gán nhãn sản phẩm từ ảnh (”Áo thun màu xanh, chất cotton”).

   Ví dụ: Pinterest sử dụng AI để đề xuất sản phẩm tương tự.

8. **Robot & Xe tự lái:** Giúp robot/xe tự lái ”hiểu” môi trường xung quanh qua mô tả bằng ngôn ngữ.

   Ví dụ: Tesla sử dụng AI để nhận diện vật thể và cảnh báo nguy hiểm.

3. Các mô hình phổ biến

Trong thế giới hiện tại, có nhiều mô hình Image Captioning được phát triển và áp dụng, trong đó có những mô hình được sử dụng phổ biến như:

1. Mô hình dựa trên CNN và LSTM: đây là các mô hình dạng cổ điển và được sử dụng phổ biến nhất. Mô hình này sử dụng CNN để trích xuất đặc trưng từ ảnh và LSTM để sinh ra chú thích. Trong đó:
   - Show and Tell (2015)
     - Kiến trúc: CNN + LSTM
     - Ý tưởng: dùng CNN làm encoder, LSTM làm decoder.
     - Ưu điểm: đơn giản, hiệu quả với dữ liệu nhỏ.
   - Show, Attend and Tell (2015)
     - Bổ sung cơ chế Attention để tập trung vào vùng ảnh quan trọng khi sinh từng từ.
     - Ưu điểm: Cải thiện độ chính xác cho ảnh phức tạp.
   - NIC (Neural Image Captioning) (2015)
     - Kiến trúc: CNN + LSTM, ngoại trừ việc nó sử dụng ResNet (hoặc VGG16) làm encoder thay vì CNN thông thường.
     - Ứng dụng tốt cho Flickr8k, Flickr30k.
2. Mô hình dựa trên Transformer: đây là các mô hình mới và được phát triển gần đây. Mô hình này sử dụng kiến trúc Transformer để sinh ra chú thích. Trong đó:
   - Vision Transformer + Transformer Decoder (2020)
     - Kiến trúc: CNN + Transformer
     - Encoder: ViT chia ảnh thành các patch và mã hóa thành các token.
     - Decoder: Transformer sinh caption dựa trên token ảnh.
     - Ưu điểm: Hiệu suất cao với dữ liệu lớn.
   - OFA (One For All) (2021)
     - Đa nhiệm: Unified model cho Image Captioning, VQA, Text-to-Image.
     - Kiến trúc: Transformer, các patch được huân luyện trước có thể sử dụng trên đa dạng tác vụ.
   - BLIP (Bootstrap Language-Image Pre-training) (2022)
     - Kết hợp CLIP và GPT: Vừa hiểu ảnh, vừa sinh văn bản mạch lạc.
     - Ứng dụng: Tạo caption chất lượng cao, chỉnh sửa caption.
   - GIT (GenerativeImage2Text) (2022)
     - Tận dụng Vision Transformer (ViT) và Transformer Decoder.
     - Đặc điểm: Huấn luyện trên lượng dữ liệu khổng lồ (hàng tỷ ảnh)

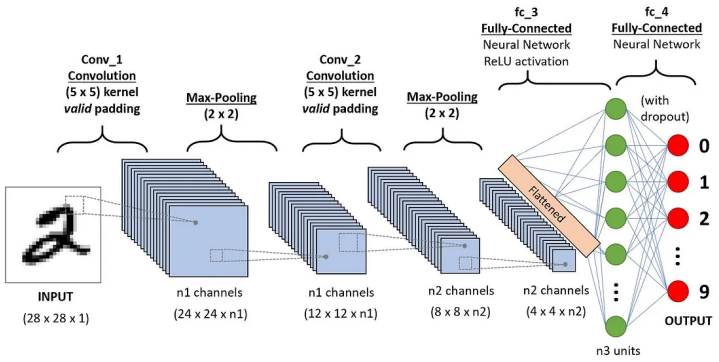
3. Ngoài ra còn có các mô hình khác như VinVL (Sử dụng CNN và Transformer), hay CoCa (Contrastive Captioners).

4. Các kiến thức liên quan

4.1 CNN

CNN là một kiến trúc mạng nơ-ron được thiết kế đặc biệt để xử lý dữ liệu có cấu trúc lưới, như hình ảnh. Mạng CNN mô phỏng cơ chế hoạt động của vỏ não thị giác trong con người, có khả năng học được các đặc trưng không gian và cục bộ trong dữ liệu đầu vào thông qua các lớp tích chập (convolutional layers).

Một mô hình CNN điển hình bao gồm các thành phần chính như lớp tích chập (convolution), lớp kích hoạt (activation function), lớp gộp (pooling), và lớp kết nối đầy đủ (fully connected). Các lớp tích chập giúp mô hình tự động trích xuất đặc trưng như đường viền, cạnh, hình dạng trong ảnh đầu vào mà không cần thiết kế thủ công.



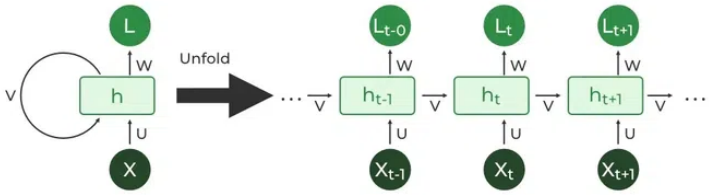
Hình 1: CNN

Nhờ khả năng học đặc trưng mạnh mẽ và hiệu quả tính toán cao, CNN được ứng dụng rộng rãi trong các bài toán nhận dạng ảnh, phân loại ảnh, phát hiện đối tượng, và các hệ thống thị giác máy tính.

4.2 RNN

RNN là một kiến trúc mạng nơ-ron được thiết kế để xử lý dữ liệu tuần tự như văn bản, âm thanh, hoặc chuỗi thời gian. Khác với CNN, RNN có khả năng ghi nhớ thông tin từ các bước thời gian trước thông qua cơ chế lan truyền trạng thái ẩn (hidden state).

Tại mỗi bước thời gian, RNN nhận đầu vào mới cùng với trạng thái ẩn từ bước trước để cập nhật trạng thái hiện tại, cho phép mô hình xử lý các chuỗi dữ liệu có độ dài linh hoạt. Tuy nhiên, RNN truyền thống gặp khó khăn với hiện tượng tiêu biến hoặc bùng nổ gradient, dẫn đến việc ghi nhớ kém các thông tin dài hạn.

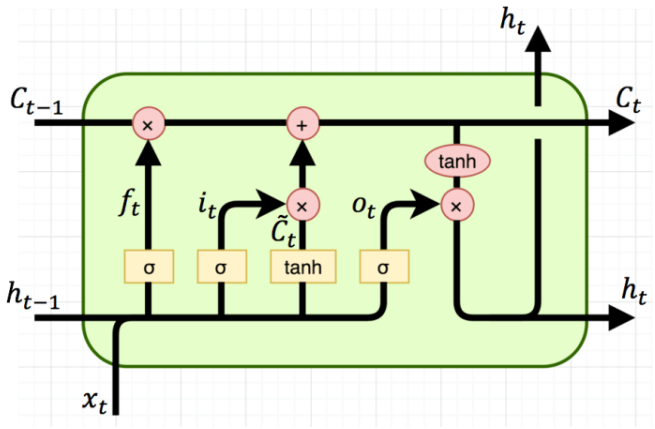


Hình 2: RNN

Do đó, trong thực tế, RNN thường được thay thế hoặc cải tiến bởi các kiến trúc mạnh hơn như LSTM hoặc GRU.

4.3 LSTM

LSTM là một biến thể của RNN được thiết kế để giải quyết vấn đề quên thông tin dài hạn bằng cách sử dụng cơ chế "bộ nhớ" (memory cell). Mỗi đơn vị LSTM bao gồm ba cổng điều khiển: cổng vào (input gate), cổng quên (forget gate), và cổng đầu ra (output gate), cho phép mạng quyết định nên ghi nhớ, quên, hay xuất ra thông tin nào tại mỗi bước thời gian.



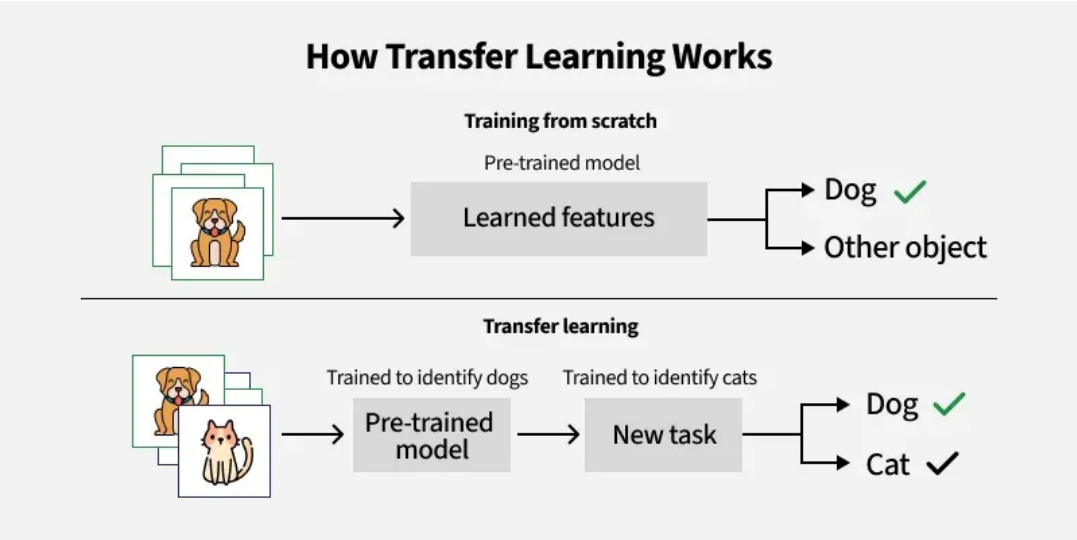
Hình 3: RNN

Điểm khác biệt giữa LSTM và RNN là LSTM có thể lưu trữ thông tin dài hạn hơn so với RNN thông qua các cổng ghi và đọc. Điều này giúp LSTM có thể xử lý chuỗi dữ liệu có độ dài lớn hơn so với RNN. Cơ chế này giúp LSTM duy trì được thông tin quan trọng trong thời gian dài, đồng thời tránh hiện tượng tiêu biến gradient khi huấn luyện. Nhờ đó, LSTM đặc biệt phù hợp với các bài toán liên quan đến ngôn ngữ tự nhiên, phân tích chuỗi thời gian, và dịch máy.

4.4 Transfer Learning

Transfer Learning (học chuyển giao) là một kỹ thuật trong học máy, trong đó mô hình đã được huấn luyện trên một nhiệm vụ sẽ được tận dụng lại để áp dụng cho một nhiệm vụ mới có liên quan. Thay vì huấn luyện mô hình từ đầu với dữ liệu lớn, Transfer Learning cho phép tái sử dụng kiến thức đã học từ một mô hình gốc, đặc biệt hữu ích trong các bài toán có dữ liệu hạn chế.

Ý tưởng của học chuyển giao là khai thác các đặc trưng phổ quát đã được học trong một mô hình huấn luyện trước đó – chẳng hạn như các đặc trưng cạnh, hình dạng trong ảnh – và sau đó tinh chỉnh (fine-tune) mô hình này trên tập dữ liệu mục tiêu để đạt hiệu quả tốt mà không cần nhiều dữ liệu huấn luyện.



Hình 4: Transfer Learning

Transfer Learning thường được thực hiện theo hai cách chính:

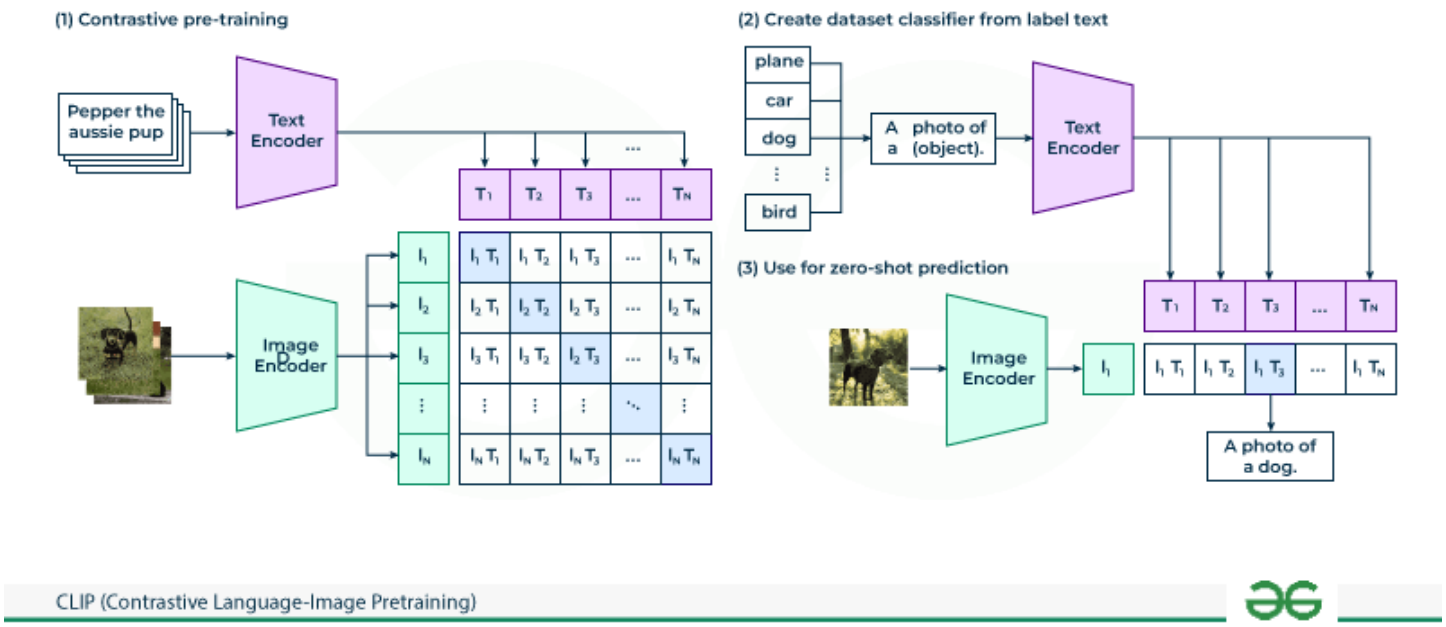
- Feature Extraction: sử dụng mô hình đã huấn luyện để trích xuất đặc trưng, sau đó đưa vào một mô hình phân loại đơn giản như SVM hoặc một lớp fully connected.

- Fine-tuning: điều chỉnh một phần hoặc toàn bộ tham số của mô hình đã huấn luyện để phù hợp với dữ liệu mới.

Transfer Learning đã chứng minh hiệu quả vượt trội trong nhiều lĩnh vực như thị giác máy tính, xử lý ngôn ngữ tự nhiên và y sinh, đặc biệt khi dữ liệu mục tiêu khó thu thập hoặc có quy mô nhỏ. Nhờ khả năng tiết kiệm chi phí huấn luyện và cải thiện độ chính xác, Transfer Learning đã trở thành một chiến lược quan trọng trong thực tiễn ứng dụng học sâu hiện nay.

4.5 Mô hình CLIP

CLIP (Contrastive Language-Image Pre-training) là một mô hình học sâu do OpenAI phát triển, được thiết kế để học cách biểu diễn hình ảnh và văn bản bằng cách so sánh chúng với các cặp ảnh–đoạn văn bản. Mô hình này được huấn luyện trên một tập dữ liệu lớn và được sử dụng để học cách biểu diễn hình ảnh và văn bản trong không gian vector.

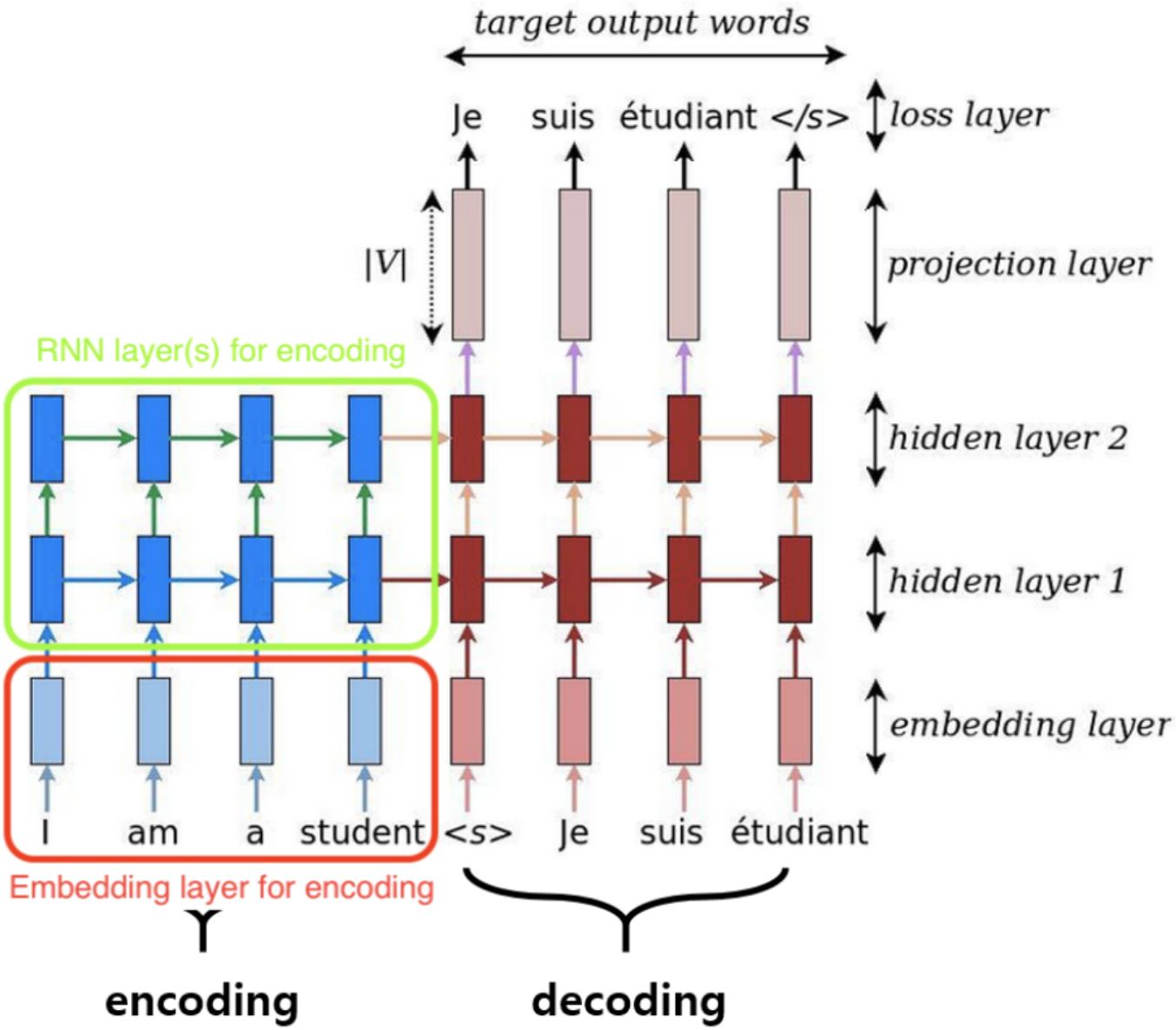


Hình 5: CLIP

CLIP sử dụng hai mạng neural riêng biệt: một mạng mã hóa hình ảnh (ví dụ ResNet hoặc Vision Transformer), và một mạng mã hóa văn bản (Transformer). Mỗi đầu vào được ánh xạ vào cùng một không gian nhúng, và mô hình học cách đưa các cặp tương ứng lại gần nhau, trong khi đẩy xa các cặp không tương ứng. Nhờ vào đó, CLIP có thể áp dụng cho nhiều tác vụ mà không cần huấn luyện lại, như tìm kiếm ảnh theo mô tả văn bản hoặc phân loại ảnh theo nhãn chưa từng thấy.

4.6 Cơ chế Attention

Cơ chế Attention là một kỹ thuật trong học sâu cho phép mô hình tập trung vào những phần thông tin quan trọng trong đầu vào, thay vì xử lý toàn bộ thông tin một cách đồng đều. Cơ chế này được áp dụng rộng rãi trong các mô hình xử lý ngôn ngữ tự nhiên và thị giác máy tính.

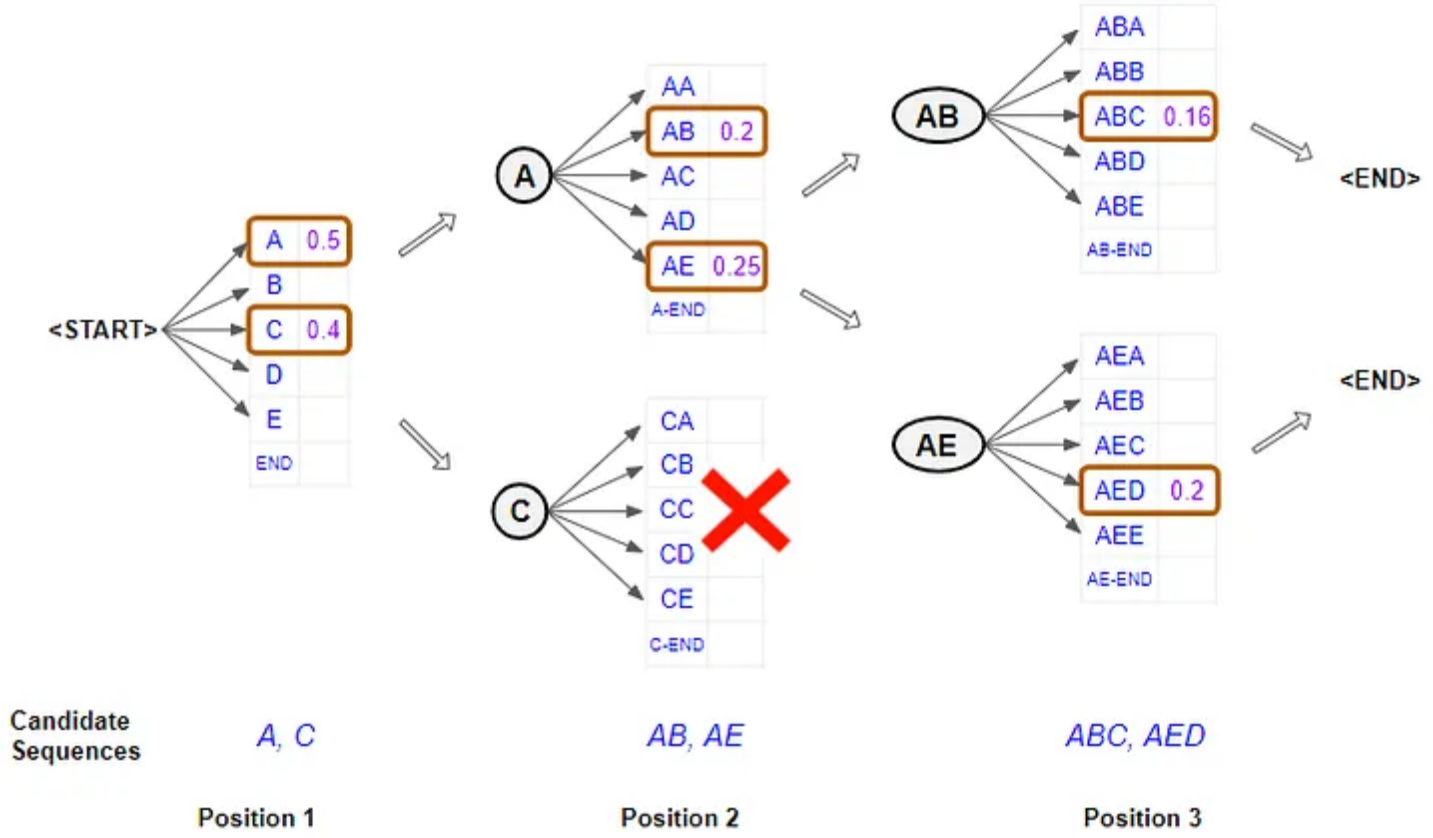


Hình 6: Cơ chế Attention

Attention hoạt động bằng cách gán trọng số khác nhau cho từng phần tử đầu vào khi tính toán đầu ra. Điều này cho phép mô hình học được mối quan hệ giữa các từ trong câu hoặc các vùng trong ảnh, bất kể khoảng cách vị trí. Attention là thành phần cốt lõi trong kiến trúc Transformer, mở ra bước tiến lớn trong các ứng dụng như dịch máy, sinh văn bản và tổng hợp ngôn ngữ.

4.7 Beam Search

Thuật toán này giúp cải thiện chất lượng sinh văn bản bằng cách tránh rơi vào lựa chọn tối ưu cục bộ. Beam Search thường được áp dụng trong các bài toán như dịch máy, sinh mô tả ảnh, và tổng hợp văn bản tự động.



Hình 7: Beam Search

Công thức đánh giá xác suất chuỗi trong Beam Search:

$$\log P(y|x) = \sum_{t=1}^{T} \log P(y_t|y_1, ..., y_{t-1}, x)$$

Trong đó, mô hình sẽ chọn chuỗi đầu ra có tổng log xác suất cao nhất trong số các chuỗi được duy trì tại mỗi bước thời gian.

4.8 Chỉ số BLEU

BLEU (Bilingual Evaluation Understudy) là một phương pháp đánh giá tự động thường được sử dụng để đo lường chất lượng của các câu do mô hình dịch máy tạo ra bằng cách so sánh chúng với một hoặc nhiều câu tham chiếu được viết bởi con người. BLEU hoạt động dựa trên việc so khớp các n-gram giữa câu sinh ra và câu tham chiếu, từ đó tính toán điểm số thể hiện mức độ tương đồng.

Chỉ số BLEU được tính bằng trung bình hình học của độ chính xác n-gram (thường từ 1-gram đến 4-gram), kết hợp với một hệ số phạt (brevity penalty) nhằm điều chỉnh cho các câu sinh quá ngắn. Giá trị BLEU nằm trong khoảng từ 0 đến 1, trong đó giá trị càng cao thì chất lượng câu sinh ra càng gần với câu tham chiếu. Đây là thước đo phổ biến và hiệu quả trong việc đánh giá các mô hình sinh ngôn ngữ tự động.

Công thức tổng quát của BLEU như sau:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Trong đó:

- p_n là độ chính xác n-gram (precision) với n từ 1 đến N (thường là 4)
- w_n là trọng số (thường đặt bằng nhau, $w_n = \frac{1}{N}$)
- BP (brevity penalty) là hệ số phạt độ dài, được tính như sau

$$\text{BP} = \begin{cases} 1 & \text{nếu } c > r \\ e^{(1-\frac{r}{c})} & \text{nếu } c \leq r \end{cases}$$

với c là độ dài câu sinh, r là độ dài tham chiếu gần nhất.

Chương III: Xây dựng hệ thống

1. Đặt vấn đề bài toán

Phát biểu bài toán: Cho một hình ảnh X bất kỳ, nhiệm vụ là tạo ra một câu mô tả (chú thích) tự động mô tả nội dung của hình ảnh đó một cách chính xác và tự nhiên.

Input:

- Một hình ảnh X

Output:

- Một câu văn bản Y mô tả nội dung của hình ảnh X

2. Ý tưởng chung

Khi đối diện với một bức ảnh bất kỳ, con người thường dựa vào kinh nghiệm, hiểu biết về thế giới và khả năng nhận diện các đối tượng, hành động, bối cảnh để mô tả nội dung ảnh bằng ngôn ngữ tự nhiên. Ý tưởng của bài toán này là mô phỏng quá trình đó bằng cách kết hợp sức mạnh của mô hình học sâu đa phương thức (CLIP) để trích xuất đặc trưng hình ảnh và mô hình sinh ngôn ngữ (LSTM) để tạo ra câu chú thích phù hợp.

Cụ thể, hệ thống sẽ sử dụng CLIP để chuyển đổi ảnh thành vector đặc trưng mang thông tin nội dung, sau đó LSTM sẽ tiếp nhận đặc trưng này và sinh ra câu mô tả tự động. Nhờ đó, mô hình có thể học cách liên kết giữa hình ảnh và ngôn ngữ, từ đó tạo ra các chú thích sát nghĩa, tự nhiên và giàu thông tin cho ảnh đầu vào mà không cần sự can thiệp thủ công của con người.

3. Mô tả bộ dữ liệu sử dụng và tiền xử lí tạo dữ liệu huấn luyện

3.1 Bộ dữ liệu sử dụng

- Sử dụng bộ dữ liệu Flickr8k
- Mỗi ảnh đi kèm nhiều chú thích (caption) do con người viết, mô tả nội dung ảnh.
- Dữ liệu gồm hai phần chính:
  - Thư mục ảnh: chứa các file ảnh (dataset/Images/)
  - File chú thích: file văn bản (dataset/captions.txt) với mỗi dòng gồm tên ảnh và chú thích, phân tách bởi dấu phẩy.

3.2 Tiền xử lý dữ liệu

Việc xử lí dữ liệu gồm các bước sau:

- **Bước 1:** Đọc và gom nhóm chú thích: Đọc file chú thích, gom tất cả caption tương ứng với từng ảnh vào một dictionary
- **Bước 2:** Xây dựng từ điển (Vocabulary)
  - Duyệt qua toàn bộ caption, tách từ (tokenize) bằng NLTK.
  - Đếm tần suất xuất hiện của từng từ.

  - Chỉ thêm các từ xuất hiện đủ số lần tối thiểu (ví dụ: ≥5) vào từ điển để giảm nhiễu.
  - Gán chỉ số cho các token đặc biệt: <PAD>, <SOS>, <EOS>, <UNK>.
- **Bước 3:** Mã hóa caption
  - Mỗi caption được chuyển thành một chuỗi số nguyên dựa trên từ điển (numericalize).
  - Thêm token <SOS> ở đầu và <EOS> ở cuối mỗi caption.
- **Bước 4:** Tiền xử lý ảnh
  - Ảnh được đọc bằng PIL, chuyển sang RGB.
  - Ảnh được resize và chuẩn hóa bằng hàm preprocess của CLIP để phù hợp với encoder.
- **Bước 5:** Tạo Dataset và DataLoader
  - Mỗi phần tử của dataset gồm:
    * Ảnh đã tiền xử lý (tensor)
    * Caption đã mã hóa (chuỗi số nguyên)
  - Sử dụng DataLoader để tạo batch, tự động padding caption về cùng độ dài trong batch.
- **Bước 6:** Chia tập dữ liệu: Ngẫu nhiên chia dữ liệu thành 3 phần:
  - Tập huấn luyện (train): 80% dữ liệu
  - Tập kiểm tra (val): 10% dữ liệu
  - Tập kiểm thử (test): 10% dữ liệu

4. Lựa chọn mô hình và hàm chi phí

4.1 Mô hình

Trong bài toán tạo chú thích ảnh, nhóm em lựa chọn kiến trúc kết hợp giữa mô hình CLIP và LSTM có tích hợp cơ chế Attention. Cụ thể, đặc trưng hình ảnh được trích xuất từ encoder CLIP (ViT-B/32), sau đó được đưa vào mạng LSTM nhiều lớp để sinh ra chuỗi từ mô tả. Cơ chế Attention giúp mô hình tập trung vào các đặc trưng quan trọng của ảnh tại mỗi bước sinh từ, từ đó nâng cao chất lượng chú thích.

4.2 Hàm chi phí

Để huấn luyện mô hình, nhóm em đã sử dụng hàm mất mát Cross-Entropy Loss, vốn là lựa chọn phổ biến cho các bài toán sinh chuỗi. Hàm này đo lường sự khác biệt giữa phân phối xác suất dự đoán của mô hình và nhãn thực tế (chuỗi từ đúng). Ngoài ra, trong quá trình sinh chú thích, nhóm em áp dụng thêm các kỹ thuật như length penalty và repetition penalty để tối ưu hóa chất lượng câu sinh ra, hạn chế lặp từ và khuyến khích độ dài hợp lý.

Công thức hàm mất mát tổng quát như sau:

$$\mathscr{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\log P(y_t^{(i)} \mid y_{1:t-1}^{(i)}, X^{(i)})$$

Trong đó:

- $y_t^{(i)}$ là từ thứ t trong câu mô tả (caption) của ảnh thứ i

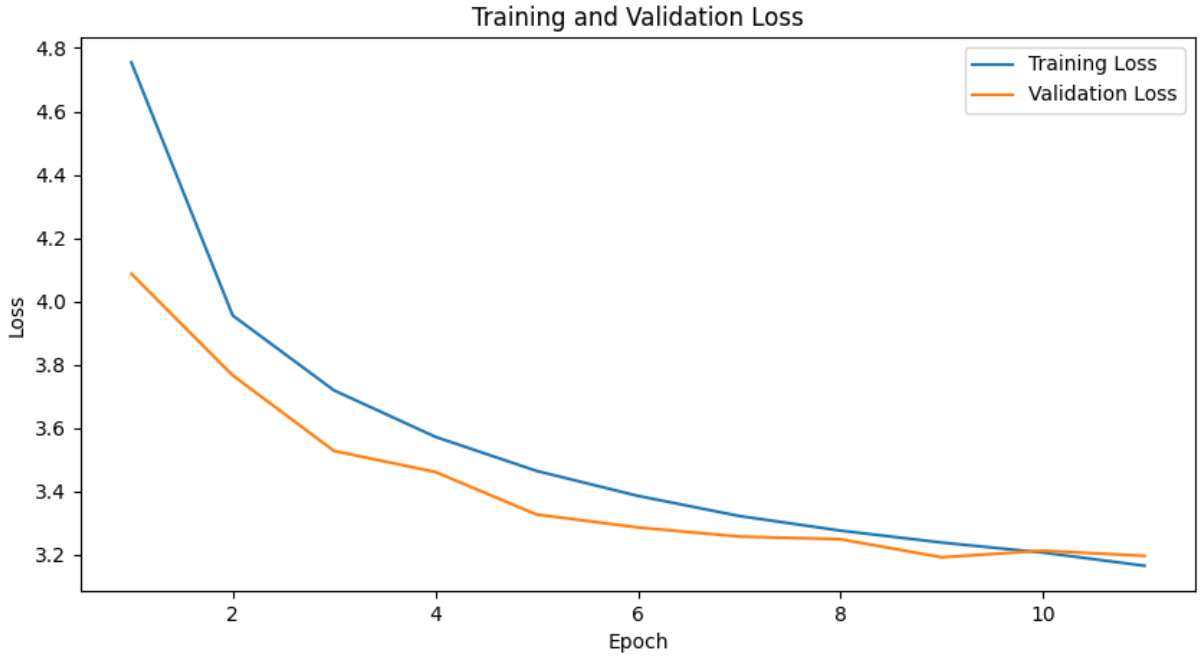
- $y_{1:t-1}^{(i)}$ là dãy từ trước đó
- $X^{(i)}$ là đặc trưng hình ảnh đầu vào
- N là số lượng mẫu trong một batch huấn luyện.

4.3 Tham số được sử dụng

- embed_size (Kích thước vector embedding cho mỗi từ trong caption) = 256
- hidden_size (Số chiều của trạng thái ẩn (hidden state) trong LSTM) = 512
- batch_size (Số lượng mẫu (ảnh và caption) được xử lý trong mỗi lần lặp (iteration) của quá trình huấn luyện) = 32
- learning_rate (Tốc độ học của thuật toán tối ưu (Adam)) = 3e-4
- num_epochs (Số vòng huấn luyện) = 15
- use_clip_cache (Các đặc trưng ảnh từ CLIP sẽ được tính trước và lưu lại để tăng tốc quá trình huấn luyện) = True
- early_stopping_patience (Số epoch tối đa cho phép mô hình không cải thiện trên tập validation trước khi dừng sớm quá trình huấn luyện) = 2
- attention (Bật/tắt cơ chế Attention trong mô hình LSTM) = True
- beam_size (Kích thước beam search khi sinh caption) = 5

4.4 Kết quả huấn luyện

Quá trình huấn luyện mô hình được thực hiện trong 15 epoch với batch size là 32. Đồ thị dưới đây thể hiện sự thay đổi của giá trị loss trên tập huấn luyện và tập validation qua từng epoch:



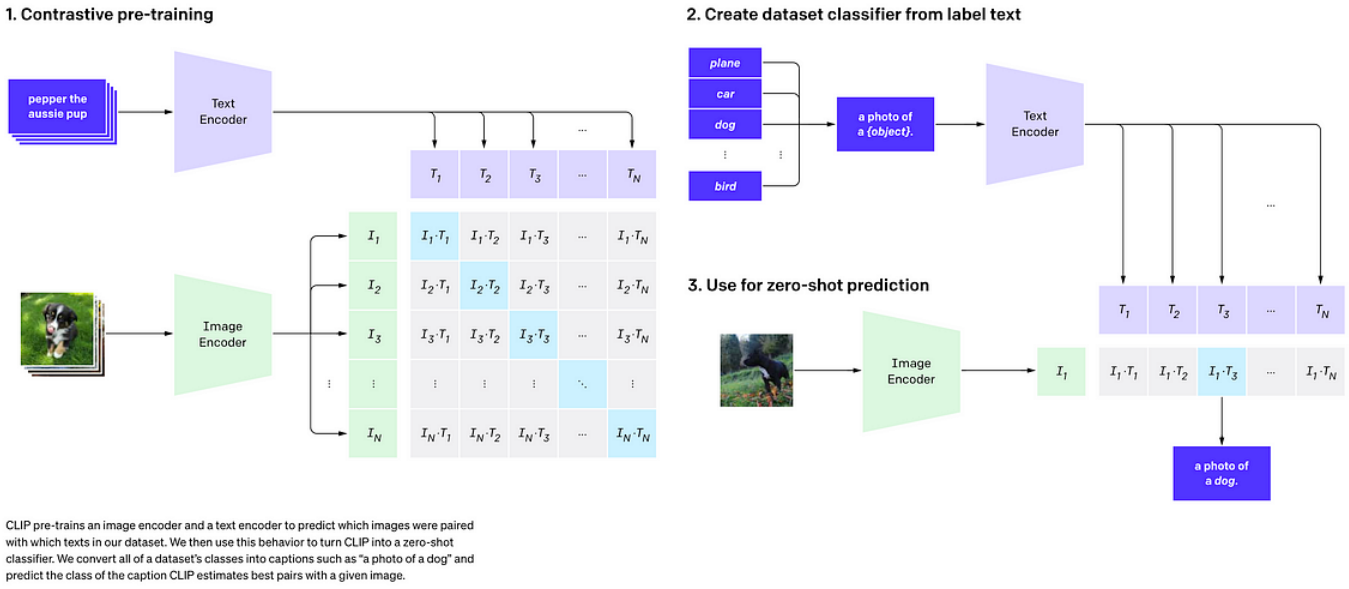
Hình 8: Sự biến thiên của giá trị loss trên tập huấn luyện và validation qua các epoch

Nhìn vào đồ thị, có thể thấy giá trị loss giảm đều qua các epoch, cho thấy mô hình học tốt và không bị overfitting. Đến cuối quá trình huấn luyện, loss trên tập validation tiệm cận loss trên tập huấn luyện, chứng tỏ mô hình tổng quát hóa tốt trên dữ liệu đáng kinh ngạc.

5. Quy trình xử lí

Quy trình tạo chú thích ảnh tự động bao gồm các bước chính sau đây:

- **Bước 1:** Tiền xử lý dữ liệu: Ảnh và chú thích được chuẩn hóa trước khi đưa vào mô hình. Ảnh được resize, chuyển sang định dạng phù hợp và chuẩn hóa bằng hàm tiền xử lý của CLIP. Các chú thích được tách từ, chuyển về chữ thường, loại bỏ ký tự đặc biệt và mã hóa thành chuỗi số nguyên dựa trên từ điển.
- **Bước 2:** Trích xuất đặc trưng ảnh: Ảnh sau khi tiền xử lý được đưa vào mô hình CLIP để trích xuất vector đặc trưng, đại diện cho nội dung hình ảnh trong không gian đa chiều.
- **Bước 3:** Sinh chú thích: Đặc trưng ảnh từ CLIP được sử dụng làm đầu vào cho mạng LSTM nhiều lớp có tích hợp Attention. Mạng LSTM sẽ lần lượt sinh ra từng từ trong câu chú thích, với sự hỗ trợ của Attention để tập trung vào các đặc trưng quan trọng của ảnh tại mỗi bước sinh từ.
- **Bước 4:** Tối ưu hóa và đánh giá: Quá trình huấn luyện sử dụng hàm mất mát Cross-Entropy để tối ưu hóa mô hình. Sau khi huấn luyện, mô hình được đánh giá bằng các chỉ số BLEU trên tập kiểm tra để đo lường chất lượng chú thích sinh ra so với chú thích tham chiếu.
- **Bước 5:** Sinh chú thích cho ảnh mới: Khi có ảnh mới, hệ thống sẽ thực hiện các bước tiền xử lý, trích xuất đặc trưng, và sử dụng mô hình đã huấn luyện để sinh ra câu chú thích tự động cho ảnh đó.



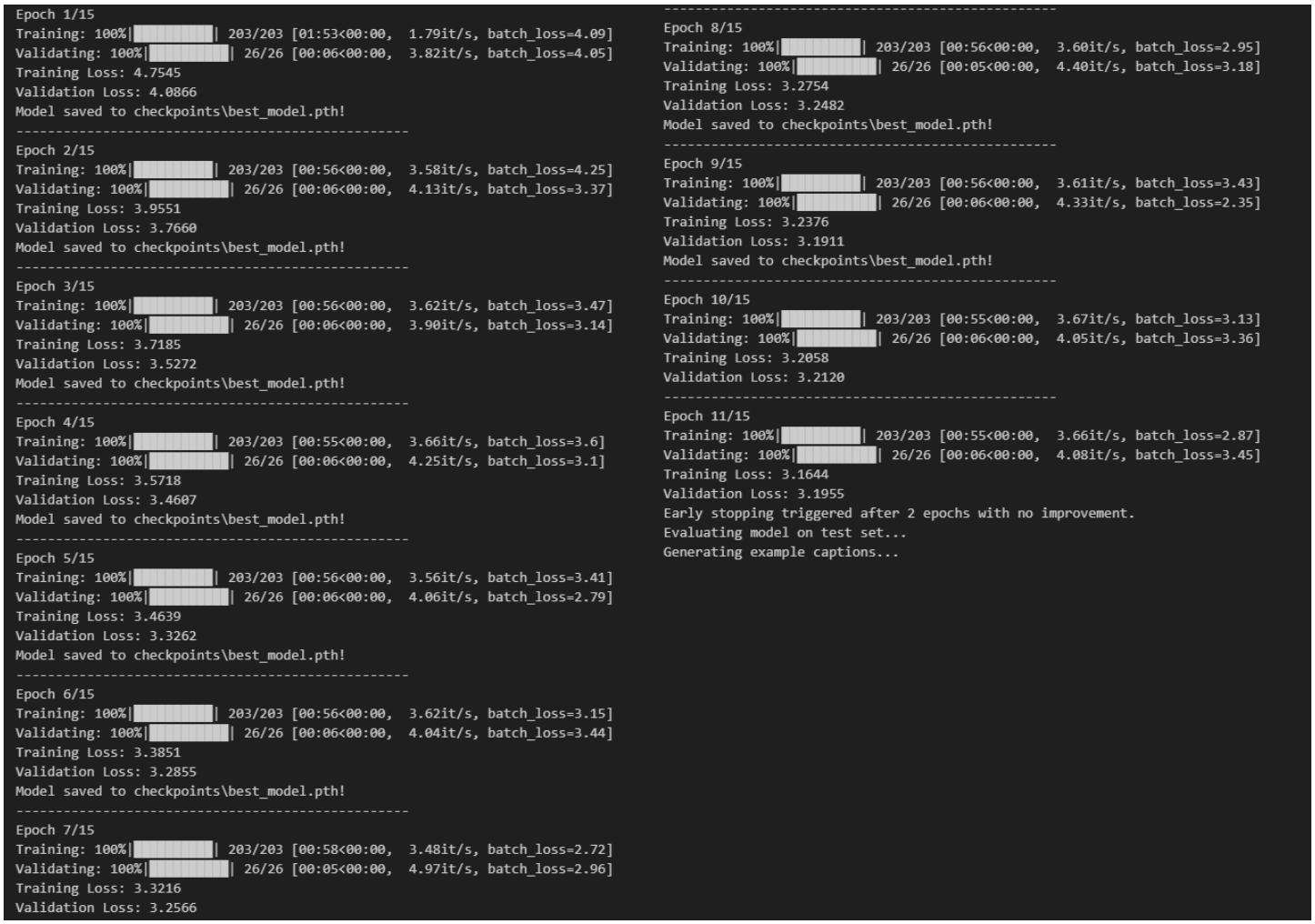
Hình 9: Quy trình sinh chú thích ảnh ảnh

6. Kết quả

Trong phần này, nhóm em trình bày kết quả thực nghiệm của mô hình kết hợp CLIP và LSTM trong bài toán tạo chú thích ảnh. Mô hình được đánh giá dựa trên quá trình huấn luyện, các chú thích ảnh được tạo ra, cũng như chỉ số BLEU thể hiện độ chính xác giữa mô tả sinh ra và mô tả gốc.

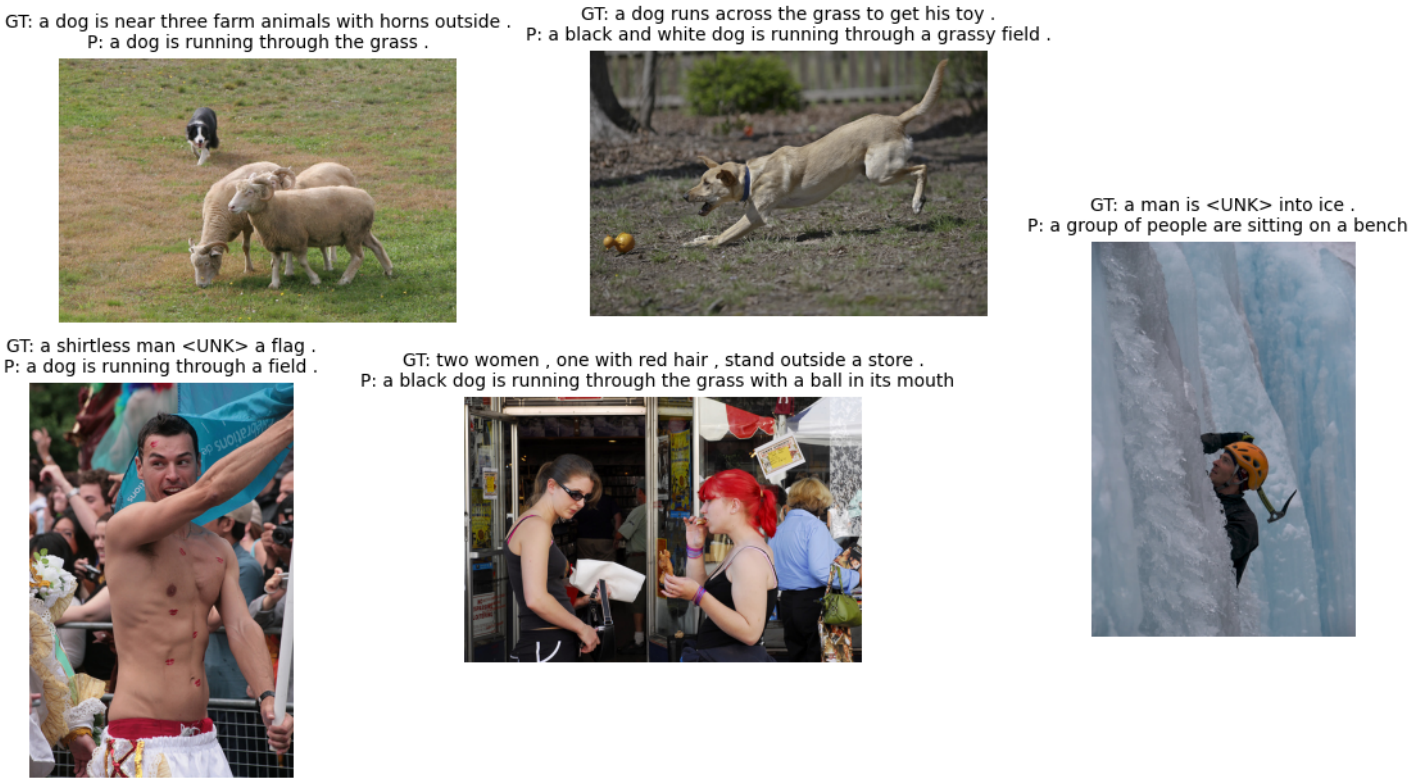
6.1 Quá trình huấn luyện

Quá trình huấn luyện diễn ra trong 15 epoch, tuy nhiên mô hình đã kích hoạt cơ chế dừng sớm (early stopping) tại epoch thứ 11 do không có cải thiện về độ lỗi trên tập validation trong 2 epoch liên tiếp. Trong suốt quá trình huấn luyện, ta nhận thấy rằng loss của cả tập huấn luyện và tập validation đều giảm đều đặn, cho thấy mô hình học tốt và tránh được hiện tượng quá khớp (overfitting). Mô hình tốt nhất đạt được khi validation loss là 3.1911 tại epoch thứ 9.



Hình 10: Kết quả huấn luyện mô hình

6.2 Kết quả tạo chú thích ảnh



Hình 11: Kết quả tạo chú thích ảnh

Sau khi huấn luyện, mô hình được đánh giá trên tập kiểm tra bằng cách so sánh câu mô tả sinh ra (P - Prediction) với câu mô tả gốc (GT - Ground Truth). Tuy nhiên, kết quả cho thấy mô hình vẫn gặp nhiều khó khăn khi tạo ra mô tả chính xác cho ảnh. Nhiều ảnh được mô hình gán chú thích lặp lại hoặc không phù hợp với nội dung thật của ảnh. Ví dụ:

- Ảnh một người đàn ông cầm cờ được mô tả sai thành "a dog is running through a field".
- Ảnh hai người phụ nữ đứng trước cửa hàng bị mô hình mô tả thành "a black dog is running through the grass with a ball in its mouth".

Điều này cho thấy mô hình có xu hướng bị hội tụ về các mô tả phổ biến trong dữ liệu huấn luyện, thay vì sinh ra câu mô tả cụ thể theo nội dung ảnh.

6.3 Đánh giá mô hình

Chỉ số BLEU được sử dụng để đánh giá định lượng chất lượng mô tả ảnh. Kết quả cho thấy các chỉ số BLEU đều ở mức thấp, đặc biệt là BLEU-4 chỉ đạt 0.0201, cho thấy khoảng cách đáng kể giữa câu mô tả mô hình tạo ra và câu mô tả gốc. Cụ thể:

Chỉ số	Giá trị
BLEU-1	0.2706
BLEU-2	0.0850
BLEU-3	0.0382
BLEU-4	0.0201

Bảng 11: Kết quả đánh giá mô hình theo các chỉ số BLEU

6.4 Nhận xét

Mặc dù mô hình kết hợp CLIP và LSTM có khả năng học và giảm lỗi trong quá trình huấn luyện, nhưng chất lượng mô tả đầu ra chưa cao. Chúng ta có thể lý giải lý do vì sao:

- Mô hình LSTM đơn giản có thể chưa đủ mạnh để học được các đặc trưng phức tạp của ảnh.
- Dữ liệu huấn luyện chưa đủ phong phú, dẫn đến hiện tượng mô hình tạo ra mô tả thiên về các câu phổ biến.
- Việc mã hóa thông tin hình ảnh bằng CLIP cần được khai thác hiệu quả hơn hoặc kết hợp thêm kiến trúc sinh mạnh mẽ hơn (ví dụ như Transformer).

7. Hạn chế và hướng phát triển

Tuy mô hình đã hoạt động hiệu quả trong một số trường hợp, đặc biệt là với các ảnh có ngữ cảnh rõ ràng, tuy nhiên cũng còn tồn tại những điểm cần cải thiện để nâng cao độ chính xác và sự phù hợp của mô tả ảnh. Một số vấn đề tiêu biểu có thể kể đến như sau:

- **Thứ nhất:** Mô hình có xu hướng tạo ra các mô tả chung chung, lặp lại hoặc không đúng với nội dung ảnh cụ thể. Ví dụ, nhiều ảnh khác nhau được gán cùng một mô tả như "a dog is running through the grass", mặc dù trong ảnh không hề có chó hoặc hoạt động tương ứng. Điều này cho thấy mô hình bị ảnh hưởng bởi sự thiên lệch dữ liệu và học theo các mô tả phổ biến thay vì chú thích đúng nội dung.
- **Thứ hai:** Mô hình gặp khó khăn trong việc sinh mô tả cho các ảnh có bố cục phức tạp, nhiều đối tượng, hoặc các chi tiết không phổ biến trong tập huấn luyện. Trong các trường hợp này, câu mô tả thường không đề cập đúng tới hành động, vật thể chính, hoặc thậm chí mô tả sai hoàn toàn.
- **Thứ ba:** Chỉ số BLEU ở các cấp độ (từ BLEU-1 đến BLEU-4) đều đạt giá trị thấp, đặc biệt BLEU-4 chỉ ở mức 0.0201, cho thấy mô hình chưa học được tốt mối quan hệ giữa đặc trưng ảnh và ngôn ngữ mô tả. Điều này phần lớn do giới hạn của mô hình LSTM trong việc học chuỗi dài và khả năng sinh ngữ cảnh chính xác.
- **Thứ tư:** Beam search với kích thước cố định không phải lúc nào cũng sinh ra chú thích tối ưu. Đôi khi các chú thích có điểm số cao nhất lại thiếu tự nhiên hoặc quá đơn giản so với nội dung phức tạp của ảnh. Hiện tượng này xuất hiện khi mô hình ưu tiên các chuỗi từ an toàn nhưng ít thông tin.

Nhóm em đã nghiên cứu và đưa ra kết luận rằng các vấn đề trên xảy ra do hai nguyên nhân chính: hạn chế về kích thước và đa dạng của tập dữ liệu huấn luyện, cùng với cấu trúc đơn giản của mô hình so với độ phức tạp của bài toán tạo chú thích ảnh.

Về hướng phát triển trong tương lai, nhóm em dự định sẽ:

- **Hướng thứ nhất:** Tích hợp cơ chế Attention tinh vi hơn để mô hình có thể tập trung vào các vùng quan trọng khác nhau của ảnh khi sinh từng từ trong caption. Cụ thể, có thể áp dụng Self-Attention hoặc Multi-Head Attention để nắm bắt tốt hơn mối quan hệ không gian giữa các đối tượng trong ảnh.
- **Hướng thứ hai:** Sử dụng bộ dữ liệu lớn và đa dạng hơn như COCO hoặc Conceptual Captions để mở rộng khả năng nhận biết đối tượng và ngôn ngữ của mô hình. Việc kết hợp nhiều bộ dữ liệu khác nhau cũng có thể giúp mô hình khái quát hóa tốt hơn.
- **Hướng thứ ba:** Thay thế LSTM bằng các kiến trúc hiện đại hơn như Transformer, kết hợp với pre-trained language models như GPT để cải thiện chất lượng ngôn ngữ của các caption được tạo ra. Điều này sẽ giúp tạo ra các chú thích tự nhiên và đa dạng hơn về mặt ngôn ngữ.

Do hướng tài nguyên phần cứng và thời gian có hạn, ba hướng nghiên cứu trên nhóm em chưa thể triển khai đầy đủ ở thời điểm hiện tại. Tuy nhiên, các thử nghiệm ban đầu với việc tăng cường cơ chế Attention đã cho thấy kết quả khả quan, hứa hẹn cải thiện đáng kể chất lượng chú thích trong tương lai.

Chương IV: Kết luận

Bài tập lớn về ứng dụng mô hình CLIP và LSTM trong bài toán tạo chú thích ảnh đã cơ bản hoàn thành mục tiêu đề ra, bước đầu xây dựng được hệ thống có khả năng sinh chú thích dựa trên nội dung hình ảnh đầu vào. Việc kết hợp mô hình CLIP để trích xuất đặc trưng ngữ nghĩa từ hình ảnh và LSTM để sinh văn bản đầu ra đã cho thấy hiệu quả nhất định trong nhiều trường hợp thử nghiệm.

Tuy nhiên, trong quá trình thực hiện, nhóm cũng gặp một số khó khăn như chất lượng dữ liệu huấn luyện chưa đồng đều, khả năng sinh câu còn đơn giản, và điểm số BLEU thu được còn khá thấp. Điều này phản ánh việc mô hình chưa thực sự hiểu sâu nội dung ảnh hoặc chưa học được ngữ cảnh ngôn ngữ phức tạp.

Dù vậy, dự án vẫn là một bước khởi đầu có ý nghĩa cho việc tìm hiểu và áp dụng mô hình đa phương thức. Nhóm xin ghi nhận mọi đóng góp, nhận xét để có thể hoàn thiện sản phẩm hơn trong các nghiên cứu tiếp theo.

Nhóm hi vọng bài tập lớn này sẽ là nền tảng cho các bước phát triển sâu hơn về sau trong lĩnh vực học sâu đa phương thức và xử lý ảnh – ngôn ngữ.