**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN HỆ THỐNG THÔNG TIN**

****

**NGUYỄN ĐÌNH LỘC : 15110243**

**PHẠM QUỐC BẢO : 15110161**

Đề tài:

**TÌM HIỂU BÀI TOÁN MÔ TẢ NỘI DUNG ẢNH**

**TIỂU LUẬN CHUYÊN NGÀNH HỆ THỐNG THÔNG TIN**

**GIÁO VIÊN HƯỚNG DẪN**

**TS. NGUYỄN THIÊN BẢO**

**KHÓA 2015-2019**

|  |  |
| --- | --- |
| **ĐH SƯ PHẠM KỸ THUẬT TP.HCM**  **KHOA CNTT**  **\*\*\*\*\*\*\*** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh Phúc**  **\*\*\*\*\*\*\*** |

# PHIẾU NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

Họ và tên Sinh viên 1 :.................................................................................. MSSV 1: .................................

Họ và tên Sinh viên 2 :.................................................................................. MSSV 2: .................................

Ngành: Công nghệ Thông tin

Tên đề tài :........................................................................................................................................................

Họ và tên Giáo viên hướng dẫn :......................................................................................................................

**NHẬN XÉT**

1. Về nội dung đề tài & khối lượng thực hiện :

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

2. Ưu điểm :

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

3. Khuyết điểm :

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

4. Đề nghị cho bảo vệ hay không ?

5. Đánh giá loại :

6. Điểm :

*Tp. Hồ Chí Minh, ngày tháng năm 2018*

Giáo viên hướng dẫn

*(Ký & ghi rõ họ tên)*

|  |  |
| --- | --- |
| **ĐH SƯ PHẠM KỸ THUẬT TP.HCM**  **KHOA CNTT**  **\*\*\*\*\*\*\*** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh Phúc**  **\*\*\*\*\*\*\*** |

# PHIẾU NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

Họ và tên Sinh viên 1 :.................................................................................. MSSV 1: .................................

Họ và tên Sinh viên 2 :.................................................................................. MSSV 2: .................................

Ngành: Công nghệ Thông tin

Tên đề tài :........................................................................................................................................................

Họ và tên Giáo viên phản biện :......................................................................................................................

**NHẬN XÉT**

1. Về nội dung đề tài & khối lượng thực hiện :

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

2. Ưu điểm :

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

3. Khuyết điểm :

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

...........................................................................................................................................................................

4. Đề nghị cho bảo vệ hay không ?

5. Đánh giá loại :

6. Điểm :

*Tp. Hồ Chí Minh, ngày tháng năm 2018*

Giáo viên phản biện

*(Ký & ghi rõ họ tên)*

# LỜI CẢM ƠN

Tiểu luận được hoàn thành tại Trường Đại học Sư phạm Kỹ thuật Tp. Hồ Chí Minh. Trong quá trình làm bài tiểu luận chuyên ngành chúng em đã nhận được rất nhiều sự giúp đỡ để hoàn tất đề tài.

Trước tiên em xin gửi lời cảm ơn chân thành đến thầy Nguyễn Thiên Bảo đã tận tình hướng dẫn, truyền đạt kiến thức, kinh nghiệm cho chúng em trong suốt quá trình thực hiện đề tài.

Xin gửi lời cảm ơn đến quý thầy cô Khoa Công nghệ Thông tin, Trường Đại học Sư phạm Kỹ thuật Tp. Hồ Chí Minh, những người đã truyền đạt kiến thức quý báu cho chúng em suốt trong thời gian học tập vừa qua.

Sau cùng xin gửi lời cảm ơn đến gia đình, bạn bè và các bạn sinh viên lớp 151102A đã luôn động viên, giúp đỡ chúng em trong quá trình nghiên cứu.

Một lần nữa, xin chân thành cảm ơn!

|  |  |
| --- | --- |
| **ĐH SƯ PHẠM KỸ THUẬT TP.HCM**  **KHOA CNTT**  **\*\*\*\*\*\*\*** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh Phúc**  **\*\*\*\*\*\*\*** |

# ĐỀ CƯƠNG TIỂU LUẬN CHUYÊN NGÀNH

Họ và tên Sinh viên 1 :................................................................ MSSV 1:...........................

Họ và tên Sinh viên 2 :................................................................ MSSV 2:...........................

Thời gian làm luận văn : từ :......................................... Đến :................................................

Chuyên ngành : ......................................................................................................................

Tên luận văn : .......................................................................................................................

................................................................................................................................................

GV hướng dẫn :......................................................................................................................

**Nhiệm Vụ Của Luận Văn :**

1. …

2. …

3. …

4. …

Đề cương viết luận văn :

**MỤC LỤC**

**1. Phần MỞ ĐẦU**

1.1. Tính cấp thiết của đề tài

1.2. Mục đích của đề tài

1.3. Cách tiếp cận và phương pháp nghiên cứu

- Đối tượng nghiên cứu

- Phạm vi nghiên cứu

1.4. Phân tích những công trình có liên quan

1.5. Kết quả dự kiến đạt được

**2. Phần NỘI DUNG**

1. Chương 1 : Giới thiệu về bài toán mô tả nội dung ảnh và giải pháp

1.1. Giới thiệu về bài toán mô tả nội dung ảnh

1.2. Giải pháp cho bài toán mô tả nội dung ảnh

1.3. Kiến trúc cơ bản của mô hình mô tả nội dung ảnh

- Ý tưởng

- Bộ mã hóa CNN

- Bộ giải mã RNN

2. Chương 2 : Lựa chọn thuật toán và xây dựng mô hình

2.1. …

2.2. …

3. Chương 3 : Triển khai huấn luyện và đánh giá hiệu năng của mô hình

3.1. …

3.2. …

**5. Phần KẾT LUẬN**

6. Tài liệu tham khảo

[1]…

[2]…

**KẾ HOẠCH THỰC HIỆN**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thời gian** | **Công việc** | **Ghi chú** |
| **1** |  |  |  |
| **2** |  |  |  |
| **3** |  |  |  |
| **4** |  |  |  |
| **5...** |  |  |  |

Ngày tháng năm 2018

**Người viết đề cương**

**Ý kiến của giáo viên hướng dẫn**

***(ký và ghi rõ họ tên)***

# MỤC LỤC

[**TÓM TẮT NỘI DUNG ĐỀ TÀI 1**](#_Toc533023764)

[**PHẦN MỞ ĐẦU 2**](#_Toc533023765)

[1.1. Tính cấp thiết của đề tài 2](#_Toc533023766)

[1.2. Mục đích của đề tài 2](#_Toc533023767)

[1.3. Cách tiếp cận và phương pháp nghiên cứu 2](#_Toc533023768)

[1.4. Phân tích những công trình có liên quan 2](#_Toc533023769)

[**PHẦN NỘI DUNG 3**](#_Toc533023770)

[**CHƯƠNG 1: 3**](#_Toc533023771)

[**PHẦN KẾT LUẬN 5**](#_Toc533023772)

[**TÀI LIỆU THAM KHẢO 6**](#_Toc533023773)

# DANH MỤC HÌNH

[**Hình 1. Câu mô tả được tạo ra từ hình ảnh.** 3](file:///C:\Users\hider\Downloads\baocao-tlcn.docx#_Toc533367680)

[**Hình 2. Mô hình mô tả ảnh với kiến trúc bộ mã hóa-giải mã [9].** 6](file:///C:\Users\hider\Downloads\baocao-tlcn.docx#_Toc533367681)

[**Hình 3. Kết nối trong lớp convolutional.** 7](file:///C:\Users\hider\Downloads\baocao-tlcn.docx#_Toc533367682)

[**Hình 4. Phép toán convolution.** 8](file:///C:\Users\hider\Downloads\baocao-tlcn.docx#_Toc533367683)

[**Hình 5. Max pooling** [21]**.** 10](#_Toc533367684)

# DANH MỤC BẢNG

# KÝ HIỆU & CHỮ VIẾT TẮT

ANN: Artifical Neural Network

CNN: Convolutional Neural Network

MLP: Multi-layer Perceptron

RNN: Recurrent Neural Network

# TÓM TẮT NỘI DUNG ĐỀ TÀI

Đề tài Tìm hiểu bài toán mô tả nội dung ảnh được chia làm ba chương với các nội dung chính được nêu ra sau đây:

* **CHƯƠNG 1:** Giới thiệu và tìm hiểu sơ lược về bài toán mô tả nội dung ảnh cũng như các kiến thức liên quan. Nội dung tập trung vào các nghiên cứu, tiến bộ gần đây, các cách tiếp cận từ đó đưa ra một kiến trúc cơ bản cho bài toán. Chương này cung cấp các kiến thức cơ bản để có thể xây dựng một mô hình hoàn chỉnh ở Chương 2.
* **CHƯƠNG 2:** Lựa chọn thuật toán, xây dựng và tối ưu hoàng chỉnh một mô hình mô tả nội dung ảnh. Chương này đi sâu vào cấu trúc của mô hình, cách thức áp dụng và xây dựng một mô hình hoàn chỉnh từ Chương 1 cùng một số phương pháp tối ưu.
* **CHƯƠNG 3:** Triển khai huấn luyện, thu thập kết quả, đánh giá và hiệu chỉnh mô hình. Chương này thực hành huấn luyện, đánh giá mô hình đã xây dựng và cung cấp các kiến thức để hiệu chỉnh, sửa lỗi, tối ưu quá trình huấn luyện nhằm đạt được kết quả cao.

# PHẦN MỞ ĐẦU

* 1. **Tính cấp thiết của đề tài**

Mô tả ảnh là một bài toán mới nổi trong thời gian gần đây. Không chỉ trên thế giới mà Việt Nam cũng đang nghiên cứu vấn đề này. Làm sao để máy tính có thể hiểu được ngôn ngữ tự nhiên của con người? Đòi hỏi phải phải có thuật toán phù hợp cùng với một dữ liệu khổng lồ và sạch sẽ, đặc biệt là các nhãn – câu miêu tả bức ảnh, phải phù hợp về nội dung bức ảnh cũng như về ngữ pháp có trong câu. Bên cạnh đó, mô tả ảnh ảnh hưởng rất quan trọng trong nền công nghiệp hiện nay như Facebook đã áp dụng thành công vào ứng dụng của mình. Nhận thấy tầm quan trọng của bài toán mô tả ảnh, nhóm em xin chọn đề tài “Mô tả ảnh” làm đề tài tiểu luận chuyên ngành.

* 1. **Mục đích của đề tài**

Học hỏi và nghiên cứu các mạng thần kinh nhân tạo như CNN, RNN và các chức năng ưu việt của nó

Áp dụng cả hai mạng CNN và RNN để ứng dụng vào bài toán mô tả ảnh và đánh giá mô hình đó dựa trên tập esty.

* 1. **Cách tiếp cận và phương pháp nghiên cứu**

**Cách tiếp cận:** Sử dụng mô hình Attention và mạng LSTM

**Phương pháp nghiên cứu lý thuyết:** Nghiên cứu các tài liệu, các bài báo có liên quan đến mô tả ảnh.

**Phương pháp nghiên cứu thực nghiệm**: Sau khi nghiên cứu lý thuyết, chúng ta sẽ đưa ra đánh giá và mô phỏng thử nghiệm chương trình demo. Đánh giá kết quả đạt được.

* 1. **Phân tích những công trình có liên quan**

Các bài toán liên quan đến mô tả ảnh như:

* Video captioning: Mô tả video trong thời gian thực
* Question answering: trả lời câu hỏi trong bức hình
* Image Generaiton: Sinh hình ảnh từ một đoạn chữ cho trước
* Và các bài toán khác: Visual Analogy, Surface Normal Estimation, Crowd Counting,…
  1. **Kết quả dự kiến đạt được**

Thông qua việc đọc các bài báo cũng như code mà tác giả đã cung cấp, kết quả nhóm em sẽ gần giống với tác giả. Bên cạnh đó, nhóm em còn train trên tập dữ liệu esty - dữ liệu này tác giả không đề cập, và dự đoán rằng mô hình cũng sẽ phù hợp với dữ liệu và qua đó tăng độ chính xác hơn so với các mô hình khác.

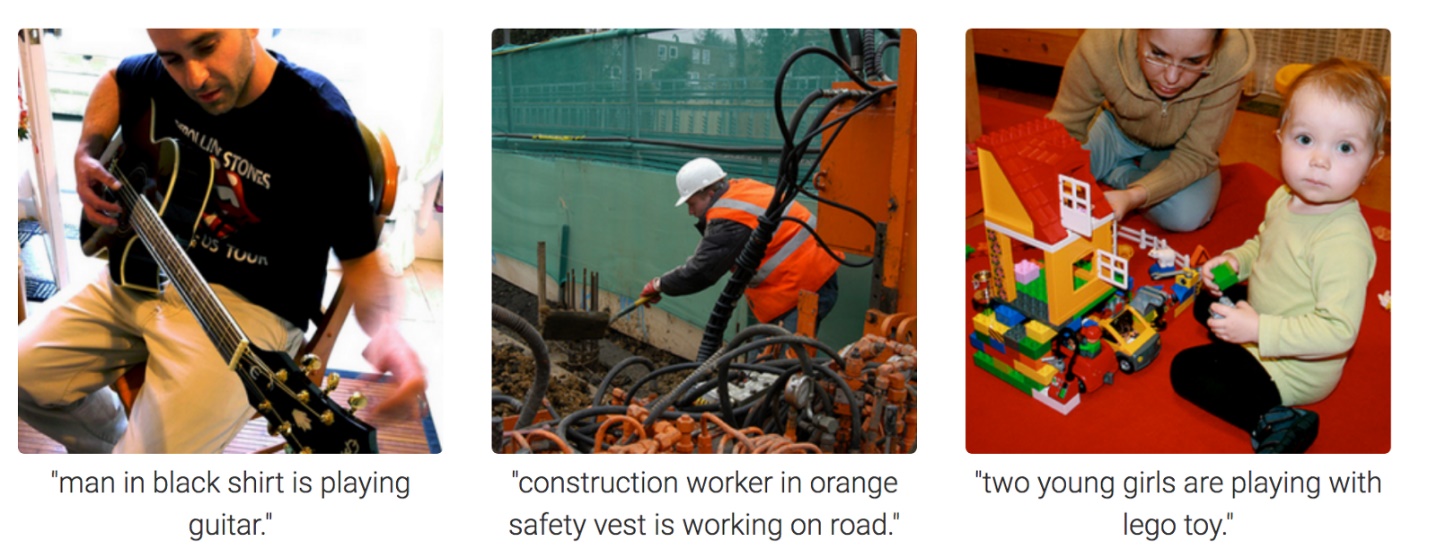
# PHẦN NỘI DUNG

## CHƯƠNG 1 : GIỚI THIỆU VỀ BÀI TOÁN MÔ TẢ NỘI DUNG ẢNH VÀ GIẢI PHÁP

1. **GIỚI THIỆU VỀ BÀI TOÁN MÔ TẢ ẢNH :**

Bài toán mô tả nội dung ảnh là một lĩnh vực nghiên cứu trong ngành Trí tuệ nhân tạo đòi hỏi một mô hình thông minh phải hiểu được nội dung của hình ảnh và diễn tả nội dung bằng ngôn ngữ tự nhiên. Để hiểu được nội dung, mô hình phải phát hiện và nhận dạng được các đối tượng xuất hiện trong hình ảnh. Mô hình cũng cần hiểu được khung cảnh hoặc vị trí, các thuộc tính của đối tượng và tương tác giữa chúng. Không những thế, tạo ra các câu mô tả nội dung tốt còn đòi hỏi cả sự hiểu biết về cú pháp và ngữ nghĩa của ngôn ngữ [1].

**Hình 1. Câu mô tả được tạo ra từ hình ảnh.**



Hiểu được nội dung của hình ảnh phụ thuộc phần lớn vào kỹ thuật trính xuất thông tin từ hình thành ở định dạng thô sang định dạng mà mô hình có thể sử dụng được, các thông tin này được gọi là *features*, chúng thường được biểu diễn dưới dạng véc-tơ. Các kỹ thuật được sử dụng cho mục đích này có thể được chia thành hai loại: (1) Kỹ thuật dựa trên *machine learning* truyền thống và (2) Kỹ thuật dựa trên *deep learing*.

Trong *machine learning* truyền thống, *features* được trính xuất từ hình ảnh bằng các kỹ thuật thủ công như Mô hình nhị phân cục bộ (Local Binary Patterns) [2], Scale-Invariant Feature Transform [3], Biểu đồ của các lớp định hướng (Histogram of Oriented Gradients) [4], và sử dụng kết hợp các kỹ thuật này được sử dụng rộng rãi. Trong các kỹ thuật này, *features* được trích xuất từ dữ liệu đầu vào. Sau đó, chúng được chuyển đến một mô hình phân loại như Support Vector Machines [5] để phân loại các đối tượng. Vì các kỹ thuật thủ công này chỉ phù hợp với các nhiệm vụ có tính đặc thù riêng, nên trích xuất *features* từ một tập hợp dữ liệu lớn và đa dạng là không khả thi. Hơn nữa, dữ liệu trong thế giới thực như hình ảnh, video rất phức tạp và còn có nhiều cách hiểu ngữ nghĩa khác nhau.

Mặt khác, trong các kỹ thuật dựa trên *deep learing*, *features* được học tự động từ dữ liệu huấn luyện và chúng có thể xử lý một tập hợp lớn các hình ảnh và video. Ví dụ: Mạng Convolutional Neural Network (CNN) [6] được sử dụng rộng rãi để học *features* và một mô hình phân loại như Softmax được sử dụng để phân loại. CNN thường được theo sau bởi Recurrent Neural Network (RNN) để tạo ra câu mô tả.

Trong 5 năm gần đây, một số lượng lớn các bài báo viết về chủ đề mô tả nội dung ảnh với *deep learning* được xuất bản và được ứng dụng phổ biến. Các thuật toán *deep learning* có thể xử lý các vấn đề phức tạp và đầy thử thách trong bài toán mô tả nội dung ảnh khá tốt. Hơn nữa, sự sẵn có của các tập dữ liệu lớn và mới đã làm cho bài toán mô tả nội dung hình ảnh dựa trên *deep learning* trở thành một lĩnh vực nghiên cứu thú vị. Trong phạm vi đề tài, chúng em trình bày chủ yếu tập trung vào bài toán mô tả nội dung hình ảnh dựa trên *deep learning*.

1. **GIẢI PHÁP CHO BÀI TOÁN MÔ TẢ NỘI DUNG ẢNH :**

Các phương pháp để giải quyết bài toán mô tả nội dung ảnh có thể được chia làm ba nhóm chính (1) Template-based, (2) Retrieval-based, và (3) Translation-based [7].

**Template-based:** Phương pháp tiếp cận dựa trên mẫu gồm có các ô trống với số lượng cố định tương ứng với một từ trong câu mô tả. Trong các cách tiếp cận này, các đối tượng, thuộc tính, hành động khác nhau được mô hình xác định và sau đó các ô trống trong mẫu được lấp đầy. Các phương pháp dựa trên mẫu có thể tạo ra câu mô chính xác về mặt ngữ pháp. Tuy nhiên, các mẫu được xác định trước và không thể tạo câu mô tả có độ dài thay đổi.

**Retrieval-based:** Các câu mô tả có thể được truy hồi từ không gian ảnh và không gian đa phương thức. Trong các cách tiếp cận dựa trên phương pháp truy hồi, câu mô tả được lấy từ tập hợp các câu mô tả hiện có. Các phương pháp này trước tiên tìm các hình ảnh tương tự nhau về mặt hình ảnh cùng với câu mô tả của chúng từ tập dữ liệu huấn luyện. Những câu mô này được gọi là các câu mô tả ứng viên. Khi truy vấn câu mô tả nội dung cho hình ảnh, câu mô tả phù hợp nhất trong nhóm ứng viên sẽ được chọn [8]. Những phương pháp này tạo ra câu mô tả khái quát và cú pháp chính xác. Tuy nhiên, chúng không thể tạo câu mô tả cho từng hình ảnh cụ thể và chính xác về mặt ngữ nghĩa.

**Translation-based:** Phương pháp phiên dịch có thể tạo ra câu mô tả từ cả không gian ảnh và không gian đa phương thức. Cách tiếp cận chung của phương pháp này là phân tích nội dung của hình ảnh trước và sau đó tạo câu mô tả từ nội dụng phân tích được bằng mô hình ngôn ngữ [9]. Các phương pháp này có thể tạo câu mô tả mới cho từng hình ảnh chính xác hơn về mặt ngữ nghĩa so với các phương pháp trước đây. Hầu hết các phương pháp dựa trên *deep learning*đều áp dụng Translation-based.

Ngoài ra còn có thể phân loại các phương pháp mô tả nội dung ảnh dựa trên *deep learning* theo các kỹ thuật học khác nhau: học có giám sát (Supervised learning), học tăng cường (Reinforcement learning), và học không giám sát (Unsupervised learning). Thông thường câu mô tả được tạo cho toàn bộ khung cảnh trong hình ảnh. Tuy nhiên, câu mô tả cũng có thể được tạo cho các vùng khác nhau trong hình ảnh (Dense captioning) [10]. Các mô hình mô tả nội dung ảnh có thể sử dụng kiến trúc bộ mã hóa-giải mã đơn giản (Encoder-Decoder) [11] hoặc kến trúc thành phần (Compositional) [12]. Còn có các phương pháp sử dụng cơ chế chú ý (Attention mechanism) [13], dựa trên các khái niệm về ngữ nghĩa (Semantic concept) [14], và các phương pháp đi theo nhiều hướng khác nhau trong bài toán mô tả hình ảnh. Một số phương pháp cũng có thể tạo ra câu mô tả cho cả đối tượng không xuất hiện trong tập huấn luyện [15].

1. **KIẾN TRÚC CƠ BẢN :**

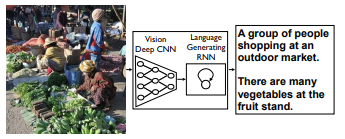
Để hiểu rõ hơn các phương pháp tiếp cận bài toán mô tả nội dung ảnh, một mô hình mẫu với kiến trúc cơ bản được trình bày trong chương này. Mô hình áp dụng các phương pháp được sử dụng phổ biến, cụ thể mô hình *deep learning* học có giám sát với kiến trúc bộ mã hóa-giải mã được sử dụng.

* 1. **Ý tưởng :**

Một mô hình duy nhất nhận một hình ảnh làm dữ liệu đầu vào và được huấn luyện để tối đa hóa xác suất tạo ra câu mô tả mô tả hình ảnh một cách đầy đủ, với mỗi là một từ cho trước trong từ điển.

Nguồn ý tưởng chính của kiến trúc bộ mã hóa-giải mã là từ những tiến bộ trong lĩnh vực dịch máy (Machine translation), nhiệm vụ trong lĩnh vực này là chuyển đổi một câu được viết bằng ngôn ngữ nguồn sang bản dịch được viết bằng ngôn ngữ đích, bằng cách tối đa hóa xác suất . Trong nhiều năm, dịch máy được thực hiện bằng cách kết hợp một loạt các nhiệm vụ riêng biệt (dịch các từ riêng lẻ, sắp xếp các từ, sắp xếp lại, v.v.), nhưng những nghiên cứu tiến bộ sau đó đã chỉ ra rằng dịch thuật có thể được thực hiện theo cách đơn giản hơn nhiều bằng cách sử dụng mạng RNN [11] mà vẫn đạt được hiệu suất cao. Một bộ mã hóa RRN đọc câu nguồn và trích xuất thông tin thành một *features* véc-tơ, véc-tơ này sau đó được “tiêm” vào một bộ giải mã RRN để tạo ra câu đích.

**Hình 2. Mô hình mô tả ảnh với kiến trúc bộ mã hóa-giải mã [9].**



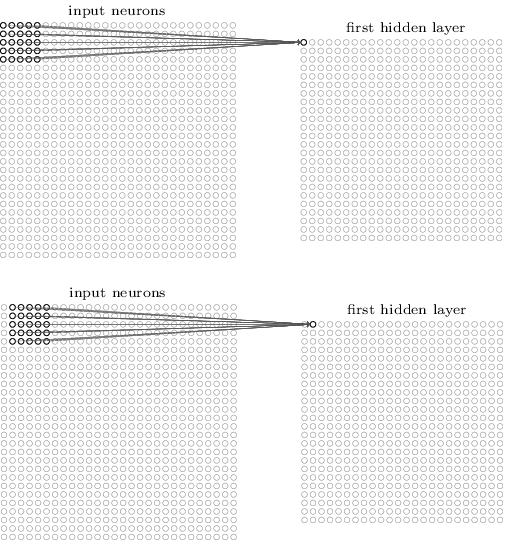
Ý tưởng này được áp dụng vào bài toán mô tả ảnh bằng cách thay thế mạng RNN của bộ mã hóa bằng một mạng CNN. Trong vài năm qua, các nhà nghiên cứu đã chứng minh một cách thuyết phục rằng các mạng CNN có thể trích xuất các thông tin của hình ảnh đầu vào và nén chúng lại thành một véc-tơ có độ dài cố định hay thường được gọi là các *features* véc-tơ, sao cho các véc-tơ này có thể sử dụng trong các tác vụ về Computer vision [16]. Do đó, việc sử dụng mạng CNN làm bộ mã hóa hình ảnh là điều tự nhiên, bằng cách huấn luyện mạng CNN cho một tác vụ phân loại hình ảnh trước và sau đó sử dụng lớp nơ-ron cuối làm dữ liệu đầu vào cho bộ giải mã RNN để tạo câu mô tả [9].

Mô hình này là các mạng lưới nơ-ron thần kinh nhân tạo (Artificial neural network) [17] hoàn toàn có thể huấn luyện bằng các thuật toán tối ưu hàm như Gradient descent [18]. Mô hình có thể dễ dàng đạt được hiệu suất cao bằng cách kết hợp các mạng nơ-ron hiện đại nhất trong lĩnh vực Computer vision và ngôn ngữ tự nhiên. Bộ giải mã còn có thể được huấn luyện trước trên các tập tài liệu ngôn ngữ khổng lồ và do đó có thể nắm được các hiểu biết về ngôn ngữ nằm ngoài tầm của tập dữ liệu huấn luyện.

* 1. **Bộ mã hóa CNN :**

Convolutional neural network (CNN) là một mạng nơ-ron thần kinh thần kinh nhân tạo được sử dụng chính trong các tác vụ về hình ảnh như phân loại, nhận diện, khoang vùng đối tượng,... CNN được thiết kế để có thể nắm bắt các thông tin về cấu trúc không gian của dữ liệu như là dữ liệu 2D của hình ảnh hay 3D của vật thể, điều mà một mạng Multi-layer perceptron (MLP) [17] khó làm được. Ngoài ra kiến trúc của CNN còn có các tham số được chia sẽ giữa kết nối của các nơ-ron làm cho số lượng các tham số trong mạng giảm mạnh, điều này làm cho mạng dễ dàng huấn luyện hơn, tăng số lượng các lớp lên, từ đó cải thiệt hiệu năng của mạng. Khả năng mạnh mẽ và hiệu quả của CNN đã thúc đẩy một phần không nhỏ sự phát triển của *deep learning* trong lĩnh vực Computer vision.

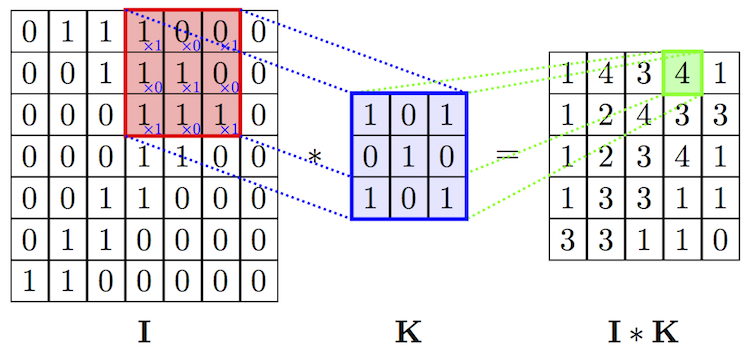
Trong CNN đưa ra các khái niệm chính (1) Receptive field (2) Shared-weight (3) Convoutional (4) Pooling. Một mạng CNN bao gồm các lớp *convolutional*, lớp *pooling* và các lớp với kết nối đầy đủ, có thể hiểu là các lớp điển hình trong mạng MLP.



**Hình 3. Kết nối trong lớp convolutional.**

**Receptive field:** Trong một lớp MLP với kết nối đầy đủ, mỗi nơ-ron nhận đữ liệu đầu vào từ tất cả các nơ-ron của lớp trước đó. Trong một lớp *convolutional*, mỗi nơ-ron chỉ nhận dữ liệu đầu vào từ một vùng nhất định của lớp trước đó. Thông thường các vùng như vậy có dạng hình vuông (ví dụ: kích thước 5 x 5). Các vùng này được gọi là các *receptive field*. Như vậy, trong một lớp MLP với kết nối đầy đủ, *receptive field* là toàn bộ lớp trước đó. Trong một lớp *convolutional*, *receptive field* là một phần nhỏ của lớp trước đó.

**Shared-weight:** Mỗi nơ-ron trong mạng tính toán một giá trị đầu ra bằng cách áp dụng một số hàm cho các giá trị đầu vào đến từ một vùng *receptive field* trong lớp trước. Hàm được áp dụng cho các giá trị đầu vào được chỉ định bởi một véc-tơ chứa các tham số được huấn luyện. Quá trình học của mạng chính là việc thực hiện các điều chỉnh tăng hoặc giảm các tham số này sao cho độ chính xác trong kết quả dự đoán của mạng được gia tăng. Véc-tơ chứa các tham số này được gọi là *filter*, các véc-tơ học được từ quá trình huấn luyện có khả năng phát hiện được một số đặc tính của dữ liệu đầu vào (ví dụ: nhận diện được các góc, cạnh hay cao cấp hơn là các đối tượng cụ thể trong ảnh). Một đặc điểm khác biệt của CNN là nhiều nơ-ron chia sẻ cùng một *filter*. Điều này làm giảm số lượng các tham số cần huấn luyện vì với mỗi *filter* chỉ cần một véc-tơ tham số được sử dụng trên tất cả các vùng *receptive field*, thay vì mỗi vùng *receptive field* cần một véc-tơ tham số riêng.



**Hình 4. Phép toán convolution.**

**Convolutinal:** Là các lớp áp dụng phép toán *convolution* lên dữ liệu đầu vào và chuyển kết quả sang lớp tiếp theo, các dữ liệu này là các dữ liệu 2D hoặc 3D được biểu diễn bằng các ma trận đại số 2 hoặc 3 chiều [19].

Để hiểu rõ hơn, ta xét ví dụ Hình 4. Trong đó ma trận là ma trận chứa dữ liệu đầu vào với số chiều 7 x 7, ma trận là ma trận các tham số cần huấn luyện của mạng với số chiều 3 x 3 và ma trận chứa kết quả của phép toán *convolution* giữa ma trận và với số chiều thu được là 5 x 5. Ta có công thức tính như sau:

Ma trận thu được có số chiều được tính như sau:

Trong đó là chỉ số trượt hay còn gọi là *strike*, ví dụ này chọn nên ta thu được ma trận 5 x 5 như Hình 4. Ở đây có một số khái niệm cần lưu ý: ma trận chính là *filter*, ma trận kết quả được gọi là *features map*.

Kết quả đầu ra của lớp *convolutional* được áp dụng một hàm phi tuyết tính để mang lại tính chất phi tuyết tính cho mạng, các hàm này được gọi chung là *activation function*. Thông thường hàm ReLU được sử dụng trong các lớp *convolutional*:

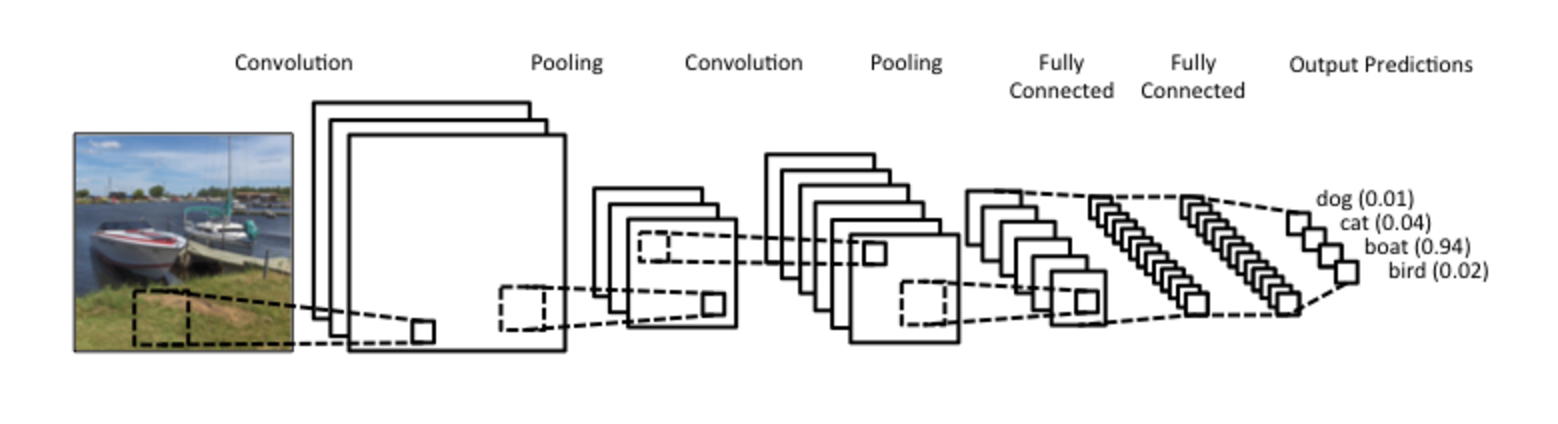
Mỗi tế bào nơ-ron trong lớp *convolutional* chỉ xử lý dữ liệu trong từng vùng *recptive field* của nó. Mỗi *filter* có nhiệm vụ nhận diện một đặc tích riêng của dữ liệu đầu vào nên trong một lớp *convolutional* sẽ bao gồm nhiều *filter* khác nhau để có thể nắm bắt được càng nhiều đặc tính của dữ liệu, các kết quả đầu ra sẽ được xếp chồng lên nhau theo thứ tự. Mặc dù có thể sử dụng các mạng MLP với kết nối đầy đủ giữa các nơ-ron để học các *features* cũng như phân loại dữ liệu, nhưng việc áp dụng kiến ​​trúc này vào dữ liệu hình ảnh là không thực tế. Các mạng này yêu cầu một số lượng lớn nơ-ron, ngay cả đối với các mạng đơn giản để xử lý dữ liệu, do kích thước đầu vào của hình ảnh thường rất lớn, trong đó mỗi pixel đều là một tham số có liên quan. Ví dụ như một lớp MLP xử lý dữ liệu cho một hình ảnh (nhỏ) có kích thước 100 x 100 thì cần tới 10000 tham số cho mỗi nơ-ron trong lớp tiếp theo. Phép toán *convolution* mang lại một giải pháp cho vấn đề này vì nó làm giảm số lượng các tham số cần thiết, cho phép mạng có thể có nhiều lớp hơn với ít tham số hơn [20]. Ví dụ cụ thể như, với bất kể hình ảnh có kích thước nào, các vùng *receptive field* có kích thước 5 x 5 đều có cùng tham số được chia sẻ nên chỉ yêu cầu 25 tham số.

**Pooling:** Các mạng CNN có thể có thêm các lớp *pooling*, các lớp này kết hợp dữ liệu đầu vào theo từng cụm, biến mỗi cụm này thành một nơ-ron duy nhất ở lớp tiếp theo [22] [23].



**Hình 5. Max pooling** [21]**.**

Có thể hiểu lớp *pooling* tương tự như lớp *convolutional* với các vùng trượt *receptive field* nhưng thay vì áp dụng các *filter* và *activation function*, lớp *pooling* chỉ áp dụng một hàm đơn giản như *max pooling* hoặc *average pooling* lên các vùng *receptive field*. Cụ thể, *max pooling* chỉ lấy giá trị lớn nhất trong mỗi vùng làm kết quả [23], *average pooling* thì tính kết quả bằng cách lấy trung bình cộng các giá trị trong mỗi vùng. Các lớp *pooling* có nhiệm vụ làm giảm số chiều của dữ liệu lưu thông trong mạng từ đó làm giảm số lượng các tham số cần huấn luyện và một chức năng quan trọng, lớp *pooling* giúp các đặc tính phát hiện được trong dữ liệu nổi bật hơn, bất biến với các thay đổi về không gian của dữ liệu (ví dụ như phóng to, thu nhỏ hoặc xoay) [21].



**Hình 7. Kiến trúc một mạng CNN cơ bản.**

Một mạng CNN cơ bản bao gồm các cụm được gọi là *conv-pool* nối tiếp nhau, mỗi cụm này gồm một hay nhiều lớp *convolutional* theo sau là một hoặc nhiều lớp *pooling*, số chiều của dữ liệu giảm đi sau mỗi cụm *conv-pool* thể hiện mực độ cao cấp của đặc tính mà mạng nhận diện được. Ví dụ như cụm *conv-pool* đầu tiên sẽ học hỏi để nhận diện các góc, cạnh, cụm tiếp theo sẽ nhận diện các hình khối cơ bản, … cho đến khi mạng nhận diện được các thuộc tính cấp cao như con người, đồ vật, … trong các bài toán nhận diện, phân loại hình ảnh. Cụm *conv-pool* cuối cùng được nối với các lớp nơ-ron kết nối đầy đủ để mạng CNN xử lý, đưa ra kết quả dự đoán mong muốn.



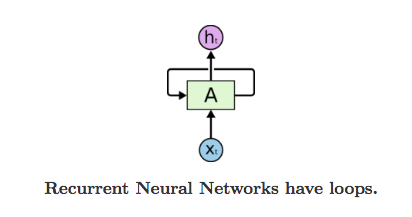
**Hình 6. Kiến trúc mạng VGG16 [25].**

Để sử dụng CNN làm bộ mã hóa. Trước tiên mạng sẽ được huấn luyện cho một tác vụ về hình ảnh như phân loại, nhận diện hình ảnh,… Tập dữ liệu sử dụng để huấn luyện thường lớn hơn và tương đồng về mặt chủ đề, đối tượng trong hình ảnh so với tập dự liệu mô tả ảnh nhằm đảm bảo CNN trích xuất được các thông tin hữu ích cho quá trình tạo câu mô tả. Sau đó lớp cho ra kết quả dự đoán của mạng (ví dụ như lớp Softmax trong bài toán phân loại) [26] bị loại bỏ vì các thông tin này thuộc về tác vụ huấn luyện trước đó, kết quả đầu ra của lớp nơ-ron với kết nối đầy đủ cuối cùng (cũng có thể chọn lớp trước nó) được chọn làm *features* véc-tơ cho quá trình giải mã.

* 1. **Bộ giải mã RNN :**

Bộ giải mã đóng vai trò là một mô hình ngôn ngữ bằng cách nhận dữ liệu đầu vào và tạo ra câu mô tả dữ liệu đó bằng ngôn ngữ tự nhiên. Vì độ dài của câu được tạo ra là không xác định nên việc tạo ra cả câu mô tả cùng một lúc thường không khả thi, thay vào đó các từ trong câu sẽ được sinh ra theo thứ tự. Trong ngôn ngữ tự nhiên, một từ trong câu phụ thuộc rất lớn vào các từ đừng trước nó hay thậm chí là các từ đứng sau nó, điều này đòi hỏi một mạng nơ-ron thần kinh nhân tạo phải nắm bắt được các thông tin đầu ra trước đó để có thể tạo ra được một câu mô tả mạch lạc, có nghĩa. Các mạng nơ-ron chuyển tiếp (Feedforward neural network – các mạng nơ-ron mà trong đó các kết nối giữa các nơ-ron không tạo thành một chu kỳ ví dụ như MLP, CNN) [27] không có khả năng làm được việc này vì dữ liệu đầu vào và đầu ra của chúng là độc lập với nhau.

Ngược lại, Recurrent neural network (RNN) là các mạng nơ-ron thần kinh nhân tạo mà kết nối giữa các nơ-ron tạo thành một chuỗi liên tục theo thời gian. Hay nói một cách khác các thông tin về trạng thái trước đó của nơ-ron có thể được lưu giữ lại để xử lý cho dòng dữ liệu đầu vào tiếp theo. Trong mỗi nơ-ron tồn tại một trạng thái ẩn là các thông tin về dòng dữ liệu hiện tại và một bộ nhớ cho phép nơ-ron lưu giữ các thông tin cần thiết về các trạng thái ẩn trước đó của nó. Điều này cho phép RNN được áp dụng vào các tác vụ đòi hỏi phải xử lý một chuỗi các dữ liệu đầu hoặc cho ra một chuỗi kết quả đầu ra liên tục (ví dụ: nhận diện giọng nói, dịch máy,…).

****

**Hình 9. Các kết nối trong RNN tạo thày một vòng lặp.**

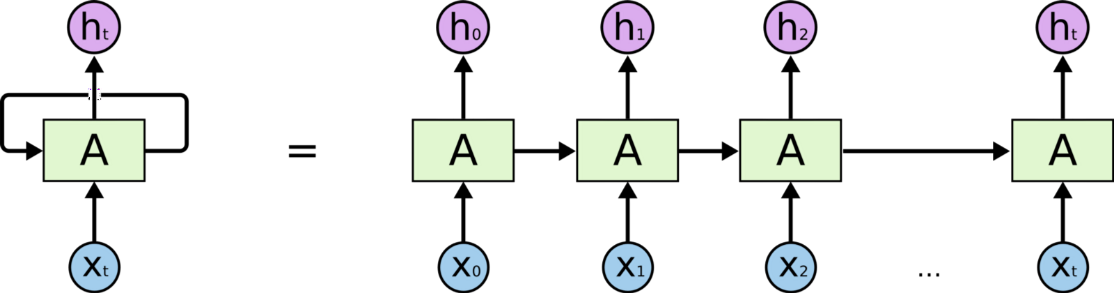
Để hiểu rõ hơn các định nghĩa về RNN, chúng ta có thể xây dựng một mô hình RNN đơn giản từ các công thức của mạng MLP như sau:

Công thức tính kết quả đầu ra của nơ-ron trong mạng MLP [17]:

Trong đó là kết quả đầu ra của nơ-ron, là dữ liệu đầu vào, và là các tham số huấn luyện, là hàm *activation*.

Để có được tính chất lưu giữ các thông tin trước đó, chúng ta sẽ bổ sung thêm một dòng dữ liệu nữa đi vào nơ-ron, đó là kết quả đầu ra trước đó của nó. Công thức được chỉnh sửa lại như sau:

Trong đó là kết quả đầu ra của nơ-ron tại thời điểm , là kết quả đầu ra tại thời điểm , , là các tham số huấn luyện.



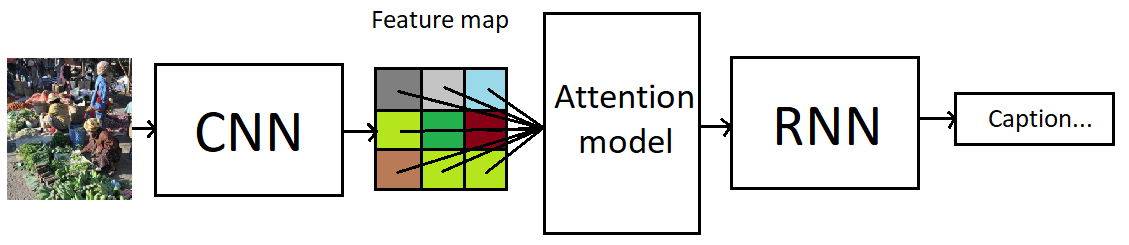
**Hình 8. Mạng RNN được thể hiện qua từng chu kỳ.**

Mạng với cấu trúc như vậy được gọi là Simple recurrent network. Tuy nhiên mạng Long short-term memory (LSTM) thường được sử dụng. Bộ giải mã RNN sẽ được huấn luyện thành một mô hình ngôn ngữ trước, nghĩa là mạng sẽ được huấn luyện trên một tập tài liệu ngôn ngữ tự nhiên để mạng có thể học được cách trình bày ngôn ngữ theo đúng ngữ pháp.

## CHƯƠNG 2 : LỰA CHỌN THUẬT TOÁN VÀ XÂY DỰNG MÔ HÌNH

1. **KIẾN TRÚC**

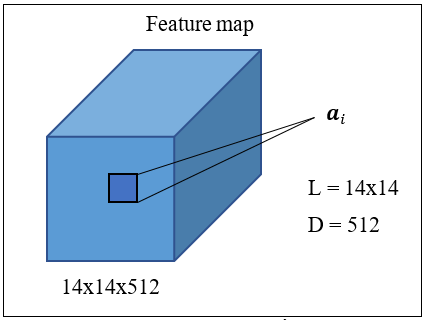
Mô hình cho bài toán mô tả nội dung ảnh sẽ nhận vào một hình ảnh ở định dạng thô và tạo ra câu mô tả cho hình ảnh đó. Dựa trên ý tưởng của mô hình sequence-to-sequence với cơ chế attention trong bài toán máy dịch – tập trung vào các từ có ngữ cảnh cần thiết ở mỗi lần sinh từ tiếp theo, một mô hình image-to-sequence với cơ chế attention được sử dụng – tập trung vào các mảng hình cần thiết trong bức ảnh ở mỗi lần sinh từ mô tả tiếp theo. Với cấu trúc mã hoá – giải mã, một mạng CNN được sử dụng làm bộ mã hoá – mã hoá hình ảnh thành các thông tin cần thiết, một mạng RNN làm bộ giải mã – giải mã các thông tin về hình ảnh thành câu mô tả và để thực hiện cơ chế attention, một mô hình attention sẽ đảm nhận giao tiếp trung gian giữa CNN và RNN.



**Hình 10. Mô hình mô tả nội dung ảnh với cơ chế attention.**

### **XÂY DỰNG BỘ MÃ HÓA**

Mạng CNN đảm nhận công việc trích xuất các thông tin về hình ảnh thông qua các lớp của nó, thay vì lấy các thông đã được nén lại trong một véc-tơ ở lớp fully connected (các lớp với kết nối dày đặc của mạng ANN) khi mà các thông tin riêng lẻ của hình ảnh đã được nén lại khó cho việc thực hiện cơ chế attention, một lớp convolution được sử dụng. Các thông tin từ lớp convolution được trính xuất thành một bộ véc-tơ tương ứng với từng phần của hình ảnh.



**Hình 11. Bộ véc-tơ được trích xuất từ lớp convolution.**

Cụ thể bao gồm L véc-tơ được trính xuất, mỗi véc-tơ có số chiều là D.

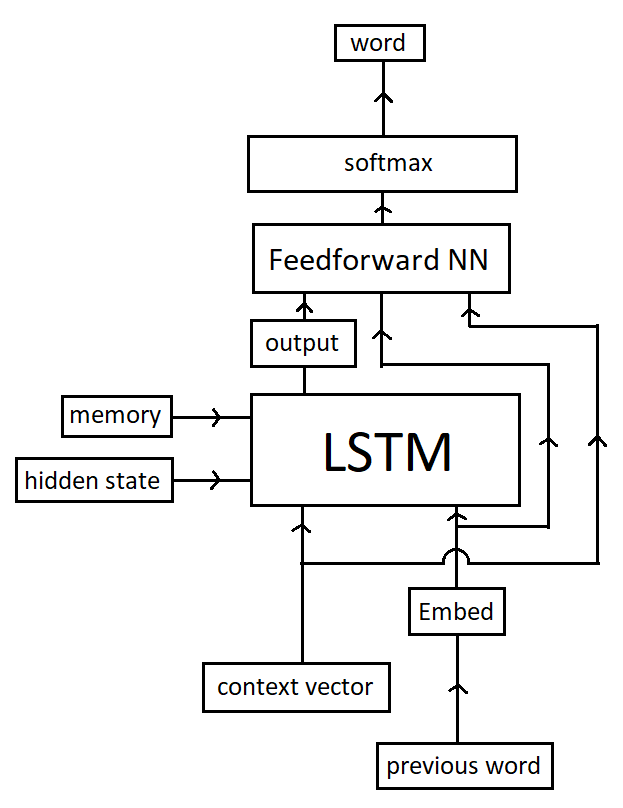
Trích xuất các thông tin từ lớp convolution cho phép bộ giải mã có thể lựa chọn tập trung vào từng phần cụ thể của hình ảnh.

### **XÂY DỰNG BỘ GIẢI MÃ**

Mạng RNN với các đơn vị LSTM [28] được sử dụng để tạo ra câu mô tả, mỗi bước mô hình sinh ra từ tiếp theo trong câu dựa vào véc-tơ ngữ cảnh (context vector), trạng thái trước đó của nơ-ron (previous hidden state) và từ được sinh ra trước đó.

Ở đây, , , , , lần lượt là cổng forget, cổng input, cổng output, memory và hidden state của đơn vị LSTM. Véc-tơ là véc-tơ ngữ cảnh, chưa các thông tin về hình ảnh ứng với từng mảng của hình, được tính toán dựa vào mô hình attention. là ma trận word embeding với là one-hot véc-tơ biểu thị từ được sinh ra trước đó. và là các tham số cần học của mạng. là hàm sigmoid activation.

Các giá trị khởi tạo memory và hidden state của LSTM được dự đoán thông qua hai mạng ANN (, ) bằng giá trị trung bình cộng của bộ véc-tơ trích xuất từ lớp convolution:



**Hình 12. Kiến trúc bộ giải mã.**

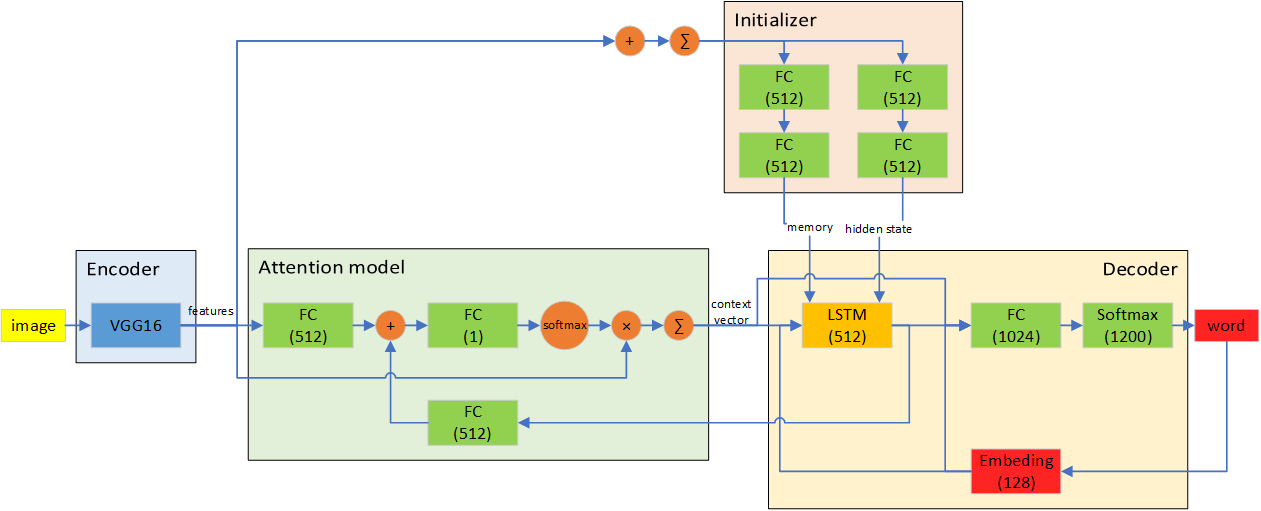
Trong mô hình này, mạng RNN sử dụng một “deep output layer [29] để dự đoán từ tiếp theo trong câu sử dụng hidden state, véc-tơ ngữ cảnh và từ dự đoán trước đó.

### **MÔ HÌNH ATTENTION**

Véc-tơ ngữ cảnh chứa các thông tin về từng phần của hình ảnh ở thời điểm t trong quá trình tạo câu mô tả. được tính đưa theo bộ véc-tơ trích xuất được từ lớp convolution tương ứng với từng vị trí khác nhau trên hình ảnh. Đối với từng vị trí , mô hình tính toán ra một trọng số dương tương ứng với tầm quan trọng, ảnh hưởng mà vị trí được “chú ý” trong quá trình tạo ra từ tiếp theo trong câu. Trọng số của từng véc-tơ được tính toán bằng mô hình attention dựa trên một mạng ANN với dữ liệu đầu vào là véc-tơ và giá trị hidden state trước đó .

Khi các trọng số (có tổng bằng 1) được tính, véc-tơ ngữ cảnh được xác định bằng kỳ vọng của bộ véc-tơ (Bahdanau, et al., September 2014):

Cuối cùng, mô hình được tổng hợp lại từ ba mô-dun chính: bộ mã hóa, mô hình attention, bộ giải mã.



**Hình 13. Kiến trúc chi tiết của mô hình.**

## CHƯƠNG 3 : TRIỂN KHAI HUẤN LUYỆN VÀ ĐÁNH GIÁ HIỆU NĂNG CỦA MÔ HÌNH

1. **TẬP DỮ LIỆU ETSY**

Tập dữ liệu Etsy được sử dụng trong đề tài để đánh giá hiệu năng của mạng. Tập dữ liệu Etsy là tập dự liệu thời trang bao gồm 5000 hình ảnh có chủ đề về quần áo đi kèm với câu mô tả về chất liệu, phong cách, màu sắc của quần áo trong hình.



**Hình 14. Tập dữ liệu Etsy**

Tập dữ liệu được chia làm 4000 hình cho mục đích huấn luyện, 1000 hình cho quá trình đánh giá.

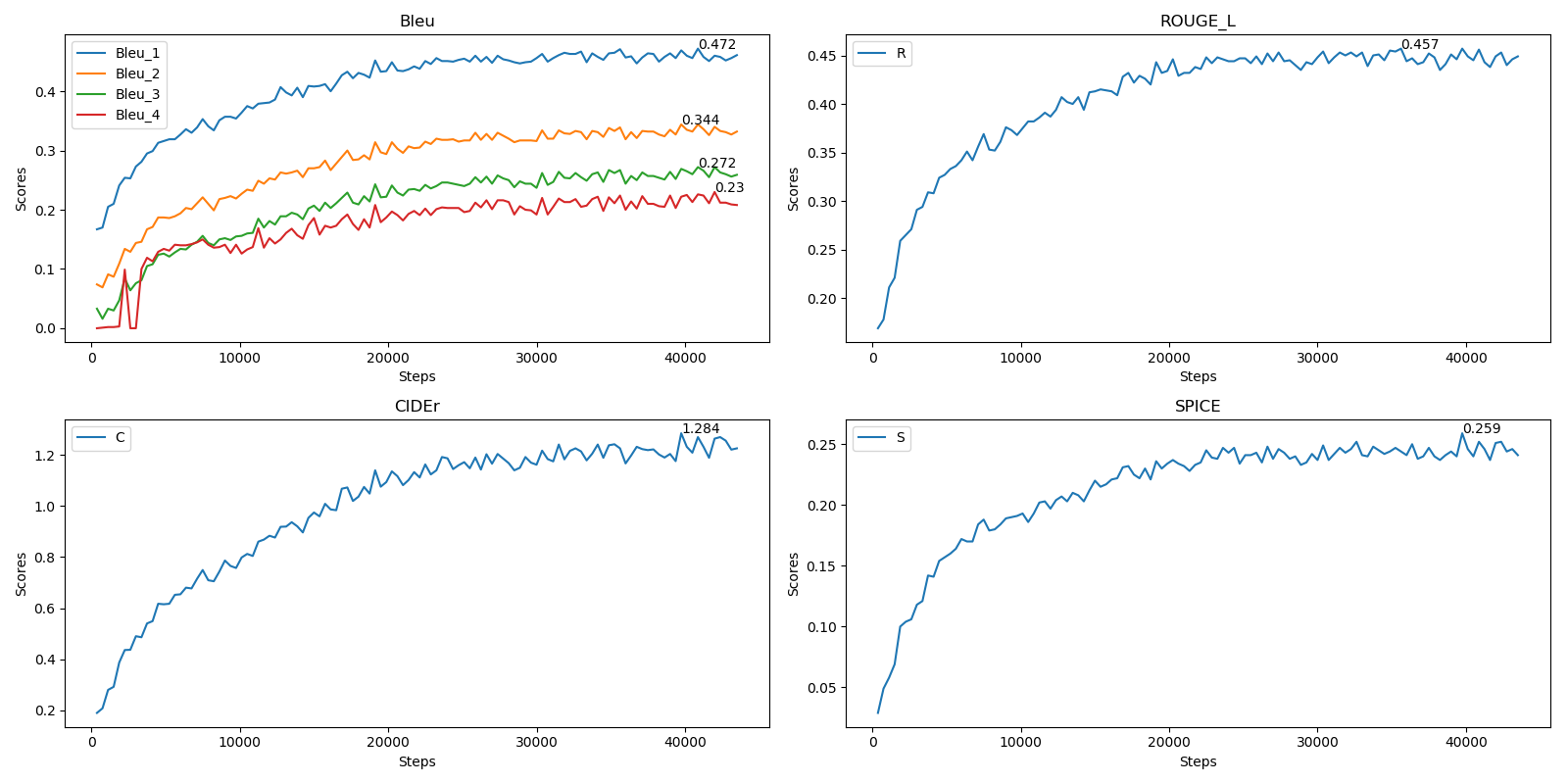
1. **LỰA CHỌN THUẬT TOÁN VÀ CÁC THÔNG SỐ HUẤN LUYỆN**

Mô hình được train với:

* Cross entropy loss
* Adam optimizer
* Learning rate decay
* Batch size 32
* Input Normalization
* Image Distortion
* Dropout Regularization
* L2 Regularization
* Clip gradient

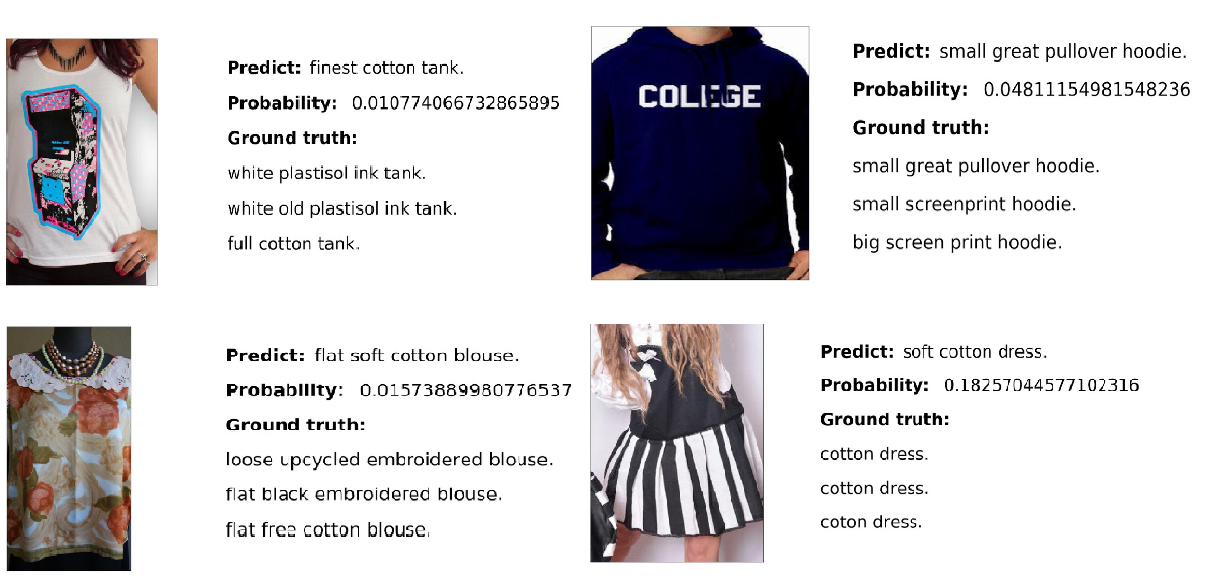
1. Đánh giá và thu thập kết quả

Đánh giá mô hình với tập test

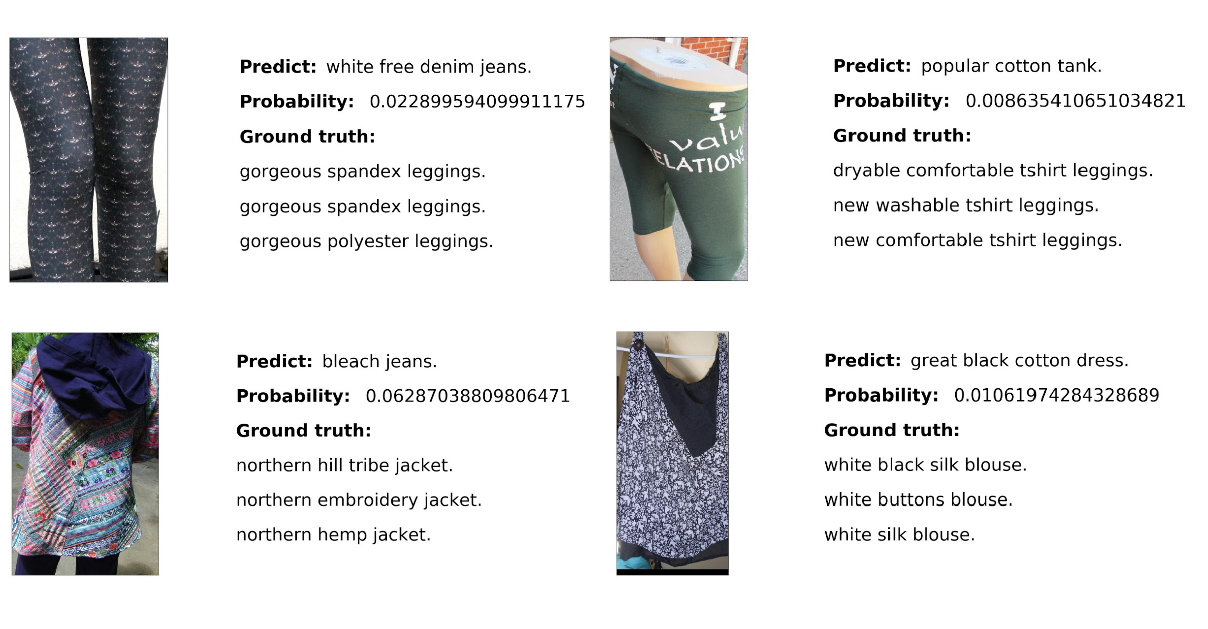


Các độ đo như BLEU, ROUGE\_L, CIDEr, SPICE đánh giá khả năng hiểu ngôn ngữ của mô hình so với con người.

Mô hình cho ra kết quả tốt:



Mô hình cho ra kết quả xấu:



# PHẦN KẾT LUẬN

# Mô tả ảnh là một bài toán rất quan trọng trong thị giác máy tính. Làm sao để máy tính có thể mô tả những bức ảnh được cho một cách chính xác là một vấn đề lớn trong bài toán. Vì ngôn ngữ rất đa dạng, một bức hình có nhiều câu mô tả và mỗi câu mô tả có thể có nhiều ý khác nhau. Bên cạnh đó không có công thức hay đơn vị đo nào có thể đáp ứng thỏa mãn sự tương quan giữa ngôn ngữ con người và ngôn ngữ máy, làm sao máy tính có thể hiểu được ngôn ngữ của con người? Đó cũng là vấn đề nan giải của các nhà nghiên cứu cũng đang cố gắng có thể đưa ra các thuật toán cũng như các độ đo mới để giúp cho máy tính ngày càng giống với con người hơn. Qua đó, giúp ích cho con người cũng như nền công nghiệp 4.0 phát triển như hiện nay.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | O. Vinyals, A. Toshev, S. Bengio và D. Erhan, “Show and tell: Lessons learned from the 2015 mscoco image captioning challenge,” *IEEE transactions on pattern analysis and machine intelligence,* pp. 652-663, 2017. |
| [2] | T. Ojala, M. PietikÃďinen và T. MÃďenpÃďÃď, “Gray scale and rotation invariant texture classification,” trong *European Conference on Computer Vision*, Dublin, 2000. |
| [3] | D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer,* pp. 91-110, 2004. |
| [4] | D. Navneet và T. Bill, “Histograms of oriented gradients for human detection,” trong *Computer Vision and Pattern Recognition*, San Diego, 2005. |
| [5] | B. E. Boser, I. M. Guyon và V. N. Vapnik, “A training algorithm for optimal margin classifiers,” trong *Proceedings of the fifth annual workshop on Computational learning theory*, Pittsburgh, 1992. |
| [6] | L. B. Y. B. a. P. H. Yann LeCun, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE,* pp. 2278-2234, 1998. |
| [7] | M. Z. Hossain, F. Sohel, M. F. Shiratuddin và H. Laga, “A Comprehensive Survey of Deep Learning for Image Captioning,” *arXiv.org,* pp. 2-3, 13 May 2018. |
| [8] | M. Hodosh, P. Young và J. Hockenmaier, “Framing image description as a ranking task: Data, models,” *Journal of Artificial Intelligence Research 47,* pp. 853-899, 2013. |
| [9] | O. Vinyals, A. Toshev, S. Bengio và D. Erhan, “Show and Tell: A Neural Image Caption Generator,” trong *The IEEE conference on computer vision and pattern recognition*, Boston, 2015. |
| [10] | J. Johnson, A. Karpathy và L. Fei-Fei, “ Densecap: Fully convolutional localization networks for dense captioning,” trong *The IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016. |
| [11] | Cho, Kyunghyun, v. Merrienboer, Bart, Gulcehre, Caglar, Bougares, Fethi, Schwenk, Holger, Bengio và Yoshua, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” trong *EMNLP*, Doha, 2014. |
| [12] | R. Socher, A. Karpathy, Q. V. Le, C. D. Manning và A. Y. Ng, “Grounded compositional semantics for finding and describing images with sentences,” *Transactions of the Association for Computational Linguistics,* pp. 207-218, 2014. |
| [13] | K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel và Yoshua, “Show, attend and tell: Neural image caption generation with visual attention,” trong *International Conference on Machine Learning*, Lille, 2015. |
| [14] | A. Karpathy, A. Joulin và F. F. F. Li, “Deep fragment embeddings for bidirectional image sentence mapping,” trong *Advances in neural information processing systems*, Montreal, 2014. |
| [15] | S. V. M. R. R. M. K. S. T. D. J. M. J. H. A. T. O. C. Lisa Anne Hendricks, “Deep compositional captioning: Describing novel object categories without paired training data,” trong *The IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016. |