TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

----------



**BÁO CÁO BÀI TẬP LỚN HỌC MÁY**

***Đề tài****:* **Ứng dụng học sâu**

**cho bài toán mô tả hình ảnh**

***Giảng viên hướng dẫn*: TS Thân Quang Khoát**

**Giảng viên hướng dẫn: TS Thân Quang Khoát**

**Sinh viên thực hiện: Đàm Minh Tiến 20156599**

**Lương Thành Long 20155970**

**Phan Xuân Phúc 20156248**

**Nguyễn Bình Minh 20156063**

**Học kỳ: 20181**

**MỤC LỤC**

[MỞ ĐẦU 4](#_Toc532309413)

[CHƯƠNG I: CƠ SỞ LÝ THUYẾT HỌC SÂU 5](#_Toc532309414)

[1. Học sâu là gì? 5](#_Toc532309415)

[2. Mạng nơ ron tích chập (Convolutional neural network) 7](#_Toc532309416)

[5.1 Convolution 8](#_Toc532309417)

[5.2 Pooling 10](#_Toc532309418)

[5.3 Fully Connected 11](#_Toc532309419)

[5.4 CNN Structure 11](#_Toc532309420)

[3. Recurrent Neural Network 12](#_Toc532309421)

[4. Long Short Term Memory 12](#_Toc532309422)

[CHƯƠNG II: BÀI TOÁN TỰ ĐỘNG MÔ TẢ HÌNH ẢNH 13](#_Toc532309423)

[1. Tổng quan 13](#_Toc532309424)

[1. Ý tưởng chính 13](#_Toc532309425)

[2. Dữ liệu và tiền xử lí 13](#_Toc532309426)

[3. Xây dựng và thử nghiệm mô hình 13](#_Toc532309427)

[4. Đánh giá mô hình và kết quả 13](#_Toc532309428)

[CHƯƠNG III: KẾT LUẬN 14](#_Toc532309429)

[1. Kết quả đạt được 14](#_Toc532309430)

[2. Hướng phát triển 14](#_Toc532309431)

[3. Lời kết 14](#_Toc532309432)

[Tài liệu tham khảo 16](#_Toc532309433)

[Phụ lục 17](#_Toc532309434)

# 

# MỞ ĐẦU

Trong những năm gần đây, trí tuệ nhân tạo nói chung, học máy nói riêng đang trở thành một một xu hướng phát triển của công nghệ. Những bài toán lập trình có thể giải bằng cách cho máy tính học tập với dữ liệu có trước, từ đó giải quyết các lớp bài toán giống nhau mà không phải đi vào lập trình cụ thể.

Những năm gần đây, khi mà khả năng tính toán của các máy tính được nâng lên một tầm cao mới và lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn, học máy đã tiến thêm một bước dài và một lĩnh vực mới được ra đời gọi là học sâu (deep learning). Học sâu đã giúp máy tính thực thi những việc tưởng chừng như không thể vào 10 năm trước: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói và chữ viết của con người, giao tiếp với con người, hay thậm chí cả sáng tác văn hay âm nhạc.

Trong đó, tự tạo chú thích cho hình ảnh là một đề tài khó với việc yêu cầu kết hợp cả xử lí hình ảnh và xử lí ngôn ngữ ngôn ngữ tự nhiên. Tuy nhiên, đây cũng là một đề tài hay và có thể ứng dụng được nhiều vào thực tiễn như camera dẫn đường cho người mù hay hệ thống tự gán alt của facebook.

Chính vì những lý do trên nên chúng em chọn đề tài **phương pháp học sâu tự động gán chú thích cho hình ảnh** cho bài tập lớn lần này.

***Hà Nội, ngày 01 tháng 12 năm 2018***

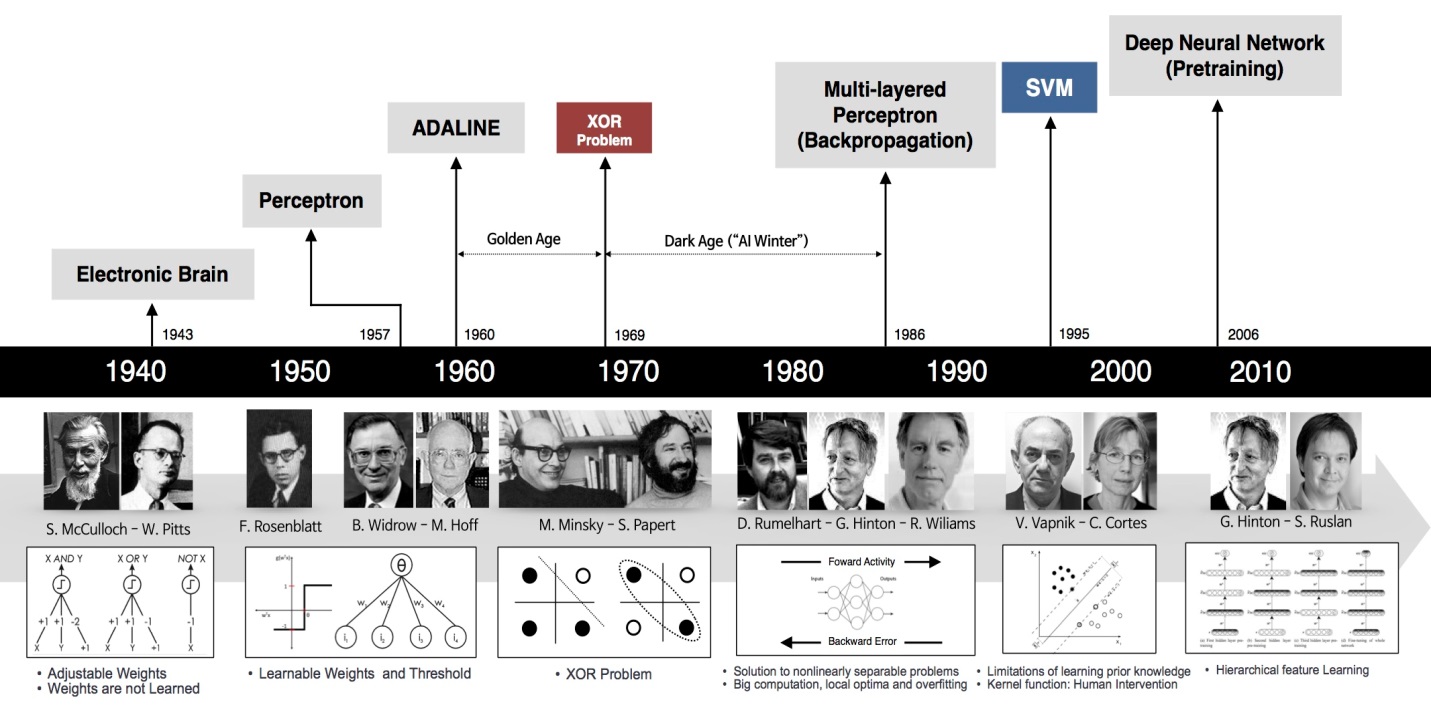
# CHƯƠNG I: CƠ SỞ LÝ THUYẾT HỌC SÂU

## 1. Học sâu là gì?

Học sâu (hay deep learning) là một nhánh của học máy. Những năm gần đây, khi mà khả năng tính toán của các máy tính được nâng lên và lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn, học sâu nổi lên như một làn sóng công nghệ mới của thế giới. Học sâu với cấu trúc mạng nhiều lớp phi tuyến tính để trích xuất các đặc trưng quan trọng dựa trên việc huấn luyện một bộ dữ liệu mẫu đã phát triển rất nhanh và từng bước được nâng cao và gần tiệm cận với khả năng của con người trong một số lĩnh vực.

Nói đến học sâu và mạng noron không thể không nhắc tới các mô hình mạng tích chập (convolutional neural network). Chính nhờ sự thành công của mạng noron tích chập trong lĩnh vực phân loại ảnh đã giúp cho học sâu nổi lên và được biết đến nhiều hơn. Các mạng noron cơ bản đã được sử dụng vào những năm 1980. Tuy nhiên công nghệ này chỉ phát triển trong khoảng một chục năm trở lại đây với các thuật toán và cấu trúc mạng tối ưu hơn, sự hỗ trợ tính toán của các hệ thống máy tính cấu hình mạnh và hỗ trợ tính toán song song (GPU) cũng như số lượng dữ liệu được xử lý gán nhãn lớn (big data). Công nghệ học sâu này thực sự bùng nổ mạnh mẽ từ năm 2012 sau chiến thắng của Alex Krizhesky với mạng AlexNet nâng cấp từ mạng CNN cơ bản trong cuộc thi ILSVRC (ImageNet Large-Scale Visual Recognition Challenge – là một cuộc thi hằng năm trong lĩnh vực phân loại ảnh). Kết quả của AlexNet đã làm giảm tỉ lệ phân loại ảnh gán nhãn sai xuống từ 26% còn 16%, một kì tích tại thời điểm đó, tốt hơn nhiều so với các phương pháp học máy thời bấy giờ chủ yếu dựa trên SVM với trích chọn đặc trưng HoG, SIFT… Sau AlexNet, tất cả các mô hình giành giải cao trong các năm tiếp theo đều là các deep networks (ZFNet 2013, GoogLeNet 2014, VGG 2014, ResNet 2015). Xu thế chung có thể thấy là các mô hình càng ngày càng sâu.

Những công ty công nghệ lớn cũng để ý tới việc phát triển các phòng nghiên cứu học sâu trong thời gian này. Rất nhiều các ứng dụng công nghệ đột phá đã được áp dụng vào cuộc sống hàng ngày. Google cho phép tìm kiếm hình ảnh, google dịch… Facebook sử dụng cho thuật toán tự động gán tag, chú thích… Amazon với hệ thống khuyến nghị sản phẩm. Cũng kể từ năm 2012, số lượng các bài báo khoa học về học sâu tăng lên theo hàm số mũ. Các blog về học sâu cũng tăng lên từng ngày.



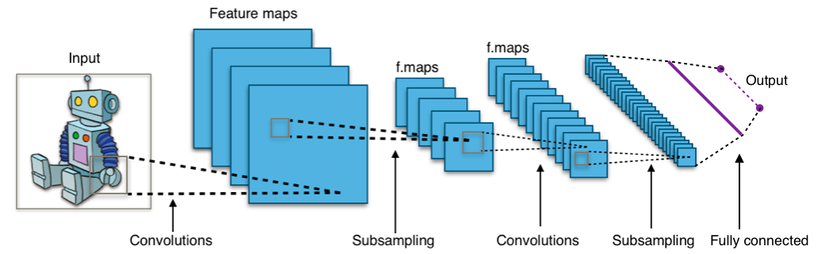
*Hình 1 Lược sử phát triển của học sâu*

Rất nhiều những ý tưởng cơ bản của deep learning được đặt nền móng từ những năm 80-90 của thế kỷ trước, tuy nhiên deep learning chỉ đột phá trong khoảng 5-6 năm nay. Có nhiều nhân tố dẫn đến sự bùng nổ này:

* Sự ra đời ngày càng nhiều của các bộ dữ liệu lớn được gán nhãn và public.
* Khả năng tính toán song song tốc độ cao của GPU, TPU. Clould computing ngày càng phổ biến.
* Sự ra đời của các hàm kích hoạt liên quan làm hạn chế vấn đề vanishing gradient (ReLU…)
* Sự cải tiến của các kiến trúc: GoogLeNet, VGG, ResNet, … và các kỹ thuật transfer learning, fine tuning, enssemble.
* Nhiều kỹ thuật regularization mới: dropout, batch normalization, data augmentation.
* Nhiều thư viện mới hỗ trợ việc huấn luyện deep network với GPU: theano, caffe, mxnet, tensorflow, pytorch, keras, …
* Nhiều kỹ thuật tối ưu mới: Adagrad, RMSProp, Adam, …
* Cộng đồng chia sẻ, các cuộc thi và sự đầu tư của chính phủ cũng như các tập đoàn lớn.

## 2. Mạng nơ ron tích chập (Convolutional neural network)

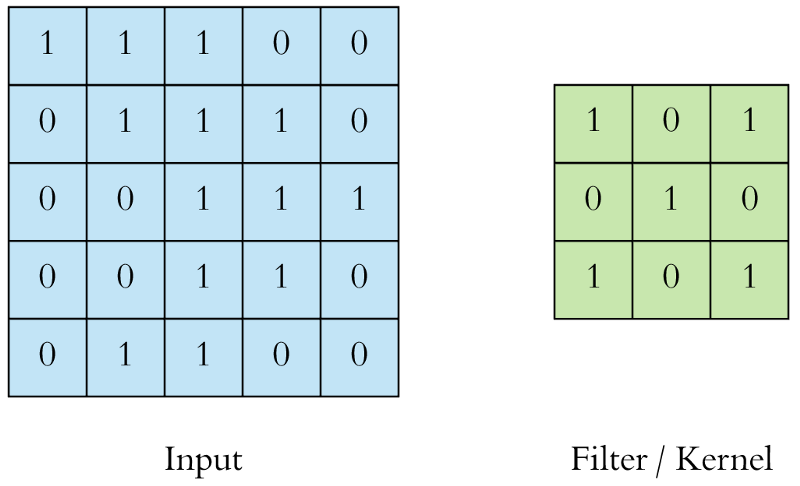
Mạng nơ ron tích chập hay convolutional neural network (CNN) là mạng dạng tiếp thuận (feedforward) trong đó thông tin chỉ đi theo một chiều từ đầu vào đến đầu ra. CNN là một deep neural network (DNN). Hiểu đơn giản, nó cũng chính là một dạng artificial neural network (ANN), một multi-layer perceptron (MLP) nhưng mang thêm 1 vài cải tiến, đó là convolution và pooling.

*Hình 2: Trong suốt quá trình huấn luyện, CNNs sẽ tự động học được các thông số cho các filter. Ví dụ trong tác vụ phân lớp ảnh, CNNs sẽ cố gắng tìm ra thông số tối ƣu cho các filter tƣơng ứng theo thứ tự raw pixel > edges > shapes > facial > high-level features. Layer cuối cùng đƣợc dùng để phân lớp ảnh.*

CNNs có tính bất biến và tính kết hợp cục bộ (Location Invariance and Compositionality). Với cùng một đối tượng, nếu đối tượng này được chiếu theo các gốc độ khác nhau (translation, rotation, scaling) thì độ chính xác của thuật toán sẽ bị ảnh hưởng đáng kể. Lớp tổng hợp (pooling layer) sẽ cho tính bất biến đối với phép dịch chuyển (translation), phép quay (rotation) và phép co giãn (scaling). Tính kết hợp cục bộ cho ta các cấp độ biểu diễn thông tin từ mức độ thấp đến mức độ cao và trừu tượng hơn thông qua convolution từ các filter. Đó là lý do tại sao CNNs cho ra mô hình với độ chính xác rất cao. Cũng giống như cách con ngƣời nhận biết các vật thể trong tự nhiên. Ta phân biệt được một con chó với một con mèo nhờ vào các đặc trƣng từ mức độ thấp (có 4 chân, có đuôi) đến mức độ cao (dáng đi, hình thể, màu lông).

### 2.1. Lớp tích chập (Convolutional layers)

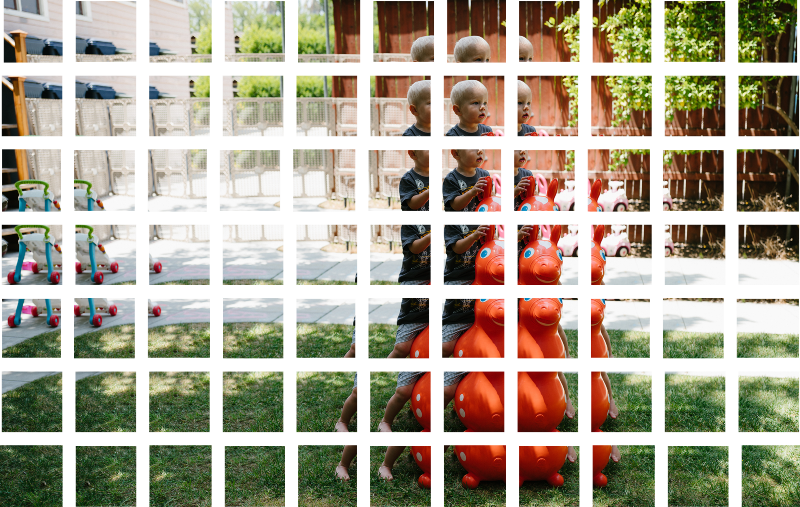
**Lớp tích chập được dùng để phát hiện và trích xuất đặc trưng - chi tiết của ảnh.** Lớp đầu tiên trong một mạng nơ ron tích chập hẳn là một lớp tích chập. Đây chính là cải tiến đáng kể của mạng LeNet (mạng CNN đầu tiên) so với các mạng MLP truyền thống.

Thực chất mình không biết phải giải thích khái niệm mới này trong CNN thế nào cho chính xác nhất. Theo ý hiểu của mình, convolution gồm 2 khái niệm khác là Convolution Filter và Convolutional Layer. Trong mạng neural network thông thường, từ input, ta cho qua các hidden layer rồi ra được output. Với CNN, Convolutional Layer cũng chính là hidden layer, khác ở chỗ, Convolutional Layer là một tập các feature map và mỗi feature map này là một bản scan của input ban đầu, nhưng được trích xuất ra các feature/đặc tính cụ thể. Scan như thế nào thì lại dựa vào Convolution Filter hay kernel. Đây là một ma trận sẽ quét qua ma trận dữ liệu đầu vào, từ trái qua phải, trên xuống dưới, và nhân tương ứng từng giá trị của ma trận đầu vào mà ma trận kernel rồi cộng tổng lại, đưa qua activation funciton (sigmoid, relu, elu, ... ), kết quả sẽ là một con số cụ thể, tập hợp các con số này lại là 1 ma trận nữa, chính là feature map.

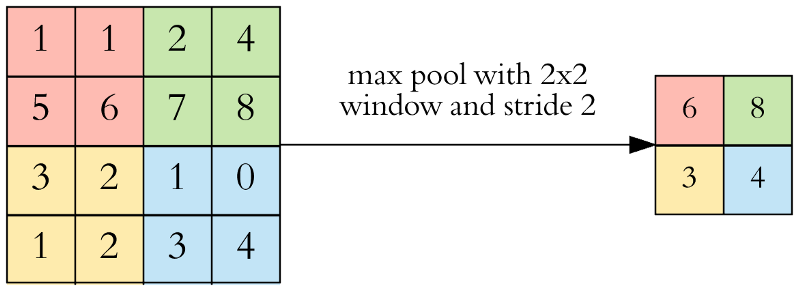
***Stride and Padding***

Stride là khoảng cách giữa 2 kernel khi quét. Với stride = 1, kernel sẽ quét 2 ô ngay cạnh nhau, nhưng với stride = 2, kernel sẽ quét ô số 1 và ô số 3. Bỏ qua ô ở giữa. Điều này nhằm tránh việc lặp lại giá trị ở các ô bị quét.

Chúng ta chọn stride và size của kernel càng lớn thì size của feature map càng nhỏ, một phần lý do đó là bởi kernel phải nằm hoàn toàn trong input. Có một cách để giữ nguyên kích cỡ của feature map so với ban đầu. Đấy là Padding. Khi ta điều chỉnh padding = 1, tức là ta đã thêm 1 ô bọc xung quanh các cạnh của input, muốn phần bọc này càng dày thì ta cần phải tăng padding lên.

Với stride=1 và padding=0, từ bức ảnh input ban đầu, ta sẽ quét kernel qua và tạo thành các ô như sau để map thành feature map

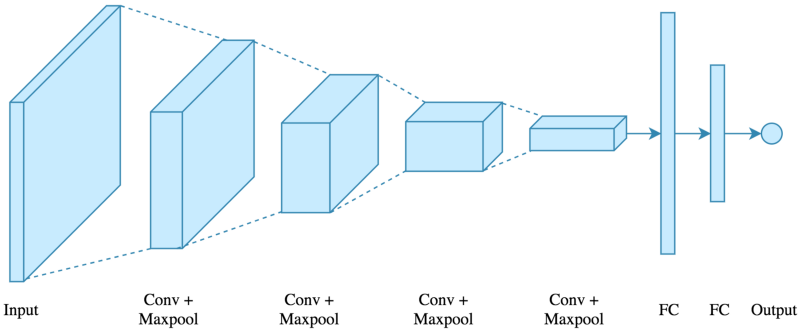
### 2.2. Lớp tổng hợp (Pooling layers)

Mục đích của pooling rất đơn giản, nó làm giảm số hyperparameter mà ta cần phải tính toán, từ đó giảm thời gian tính toán, tránh overfitting. Loại pooling ta thường gặp nhất là max pooling, lấy giá trị lớn nhất trong một pooling window. Pooling hoạt động gần giống với convolution, nó cũng có 1 cửa sổ trượt gọi là pooling window, cửa sổ này trượt qua từng giá trị của ma trận dữ liệu đầu vào (thường là các feature map trong convolutional layer), chọn ra mộtgiá trị từ các gía trị nằm trong cửa sổ trượt (với max pooling ta sẽ lấy giá trị lớn nhất). Hãy cùng nhìn vào ví dụ sau, tôi chọn pooling window có kích thước là 2 \* 2, stride = 2 để đảm bảo không trùng nhau, và áp dụng max pooling:

### 2.3. Lớp kết nối đầy đủ (Fully connected layers)

Layer này cũng chính là 1 fully connected ANN. Thường thì sau các lớp Conv+Pooling thì sẽ là 2 lớp Fully connected, 1 layer để tập hợp các feature layer mà ta đã tìm ra, chuyển đổi dữ liệu từ 3D, hoặc 2D thành 1D, tức chỉ còn là 1 vector. Còn 1 layer nữa là output, số neuron của layer này phụ thuộc vào số output mà ta muốn tìm ra. Giả sử với tập dữ liêu MNIST chẳng hạn, ta có tập các số viết tay từ 0 -> 9. Vậy output sẽ có số neuron là 10.

### 2.4. Huấn luyện mạng noron tích chập



Hình vẽ dưới đây đã biểu diễn rất rõ ràng kiến trúc của 1 mạng CNN  Ta có ảnh input đầu vào. Qua hàng loạt các Convolutional Layer cùng Max Pool Layer (thường pooling sẽ theo ngay sau 1 convolutional layer), cuối cùng là 2 fully connected.

## 3. Mạng hồi quy (Recurrent neural network)

### 3.1. Vấn đề phụ thuộc xa

### 3.2. Mạng hồi quy RNN ra đời

## 4. Mạng bộ nhớ dài hạn – ngắn hạn (Long short term memory)

### 4.1. Vấn đề của mạng hồi quy RNN

### 4.2. Những cải tiến mới của LSTM

# CHƯƠNG II: BÀI TOÁN TỰ ĐỘNG MÔ TẢ HÌNH ẢNH

## 1. Tổng quan

## 1. Ý tưởng chính

## 2. Dữ liệu và tiền xử lí

## 3. Xây dựng và thử nghiệm mô hình

## 4. Đánh giá mô hình và kết quả

# CHƯƠNG III: KẾT LUẬN

## 1. Kết quả đạt được

* Nhận thêm kiến thức về cơ bản về học máy, các bài toán phân loại, mạng noron nhân tạo
* Có thêm kiến thức về ngôn ngữ Python, Java
* Chứng chỉ Python for Machine Learning và Data Analys của Cognitive class
* Tìm hiểu kiến thức về các công nghệ như Docker, Android NDK, SDK
* Mô hình Inception\_v3, mobile\_net, tensorflow
* Giải quyết các lớp bài toán nhận dạng hình ảnh đồ vật
* Đưa các ứng dụng về học máy thành sản phẩm
* Source code dự án: https://github.com/damminhtien/tensorflow-android-classifier

## 2. Hướng phát triển

* Tiếp tục tìm hiểu và nâng cao kiến thức về học máy
* Hoàn thành các khóa học tiếp theo
* Tăng tốc và phát triển ứng dụng với lượng dữ liệu lớn hơn

## 3. Lời kết

Ứng dụng nhận diện các loại xe ô tô được xây dựng trên nền framework Deep Learning Tensorflow của Google, mạng Inception v3. Thông qua quá trình đào tạo trên tập dữ liệu ở máy tính và cài đặt lên thiết bị di động với Tensorflow SDK. Kết quả cho tỉ lệ nhận diện chính xác ở mức khá cao.

Chúng em xin gửi lời cảm ơn trân thành tới thầy **Nguyễn Văn Linh** đã hướng dẫn em những kiến thức để có thể hoàn thành việc xây dựng ứng dụng này. Em sẽ cố gắng tiếp tục học tập và phát triển ứng dụng để đạt được những kết quả lớn hơn.

## Tài liệu tham khảo

[[1](https://arxiv.org/pdf/1502.03044.pdf)] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv preprint arXiv:1502.03044, 2016.

[[2](https://cs.stanford.edu/people/karpathy/deepimagesent/)] Andrej Karpathy, Li Fei-Fei Deep Visual-Semantic Alignments for Generating Image Descriptions. arXiv preprint arXiv:1412.2306, 2015.

[[3](https://arxiv.org/pdf/1411.4555.pdf)] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555, 2014.

[[4](https://vision.cornell.edu/se3/wp-content/uploads/2018/03/1501.pdf)] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, Serge Belongie. Learning to Evaluate Image Captioning.

[[5](https://arxiv.org/pdf/1410.1090.pdf)] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Alan L. Yuille. Explain Images with Multimodal Recurrent Neural Networks. arXiv preprint arXiv:1410.1090, 2014.

[[6](https://arxiv.org/pdf/1411.4389.pdf)] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. arXiv preprint arXiv:1411.4389, 2016.

[[7](https://arxiv.org/pdf/1411.4389.pdf)] Xinlei Chen, C. Lawrence Zitnick. Learning a Recurrent Visual Representation for Image Caption Generation. arXiv preprint arXiv:1411.5654, 2016.

## Phụ lục