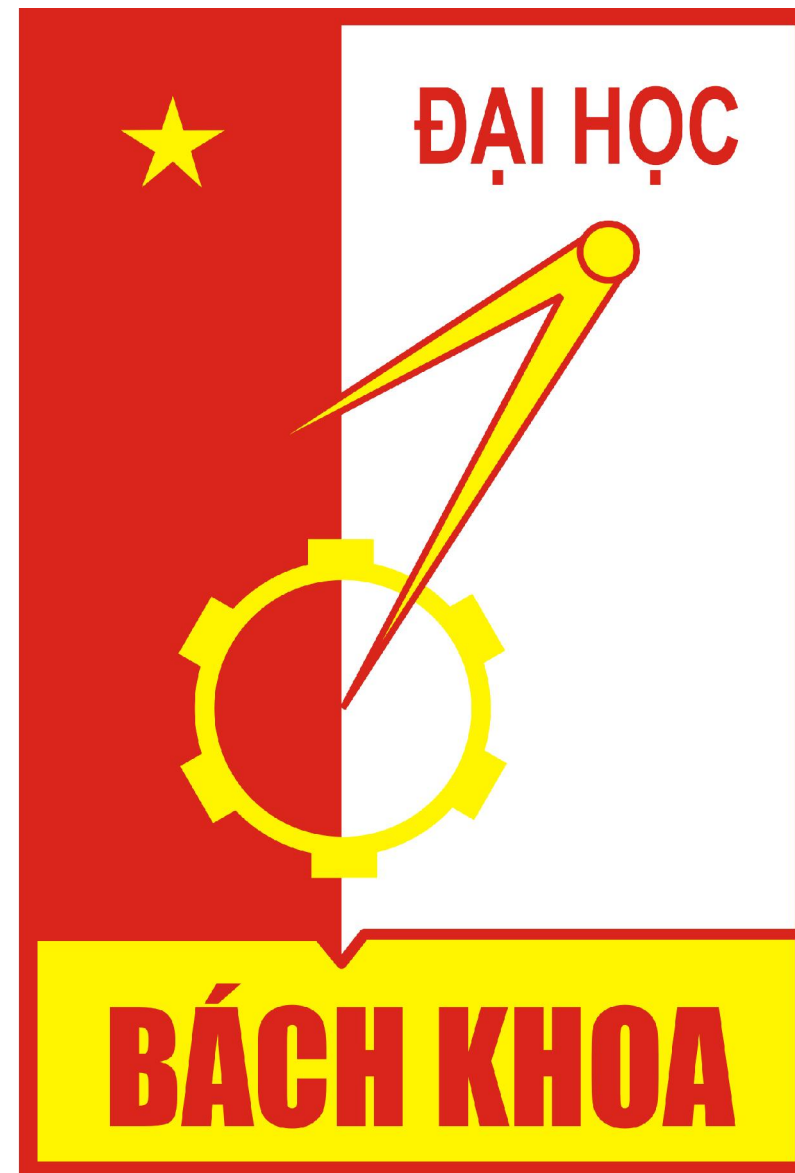



TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

BÁO CÁO  
BÀI TẬP LỚN  
MÔN HỌC MÁY

Nhóm 17





Đề tài: Ứng dụng học sâu cho  
bài toán tự động mô tả hình  
ảnh

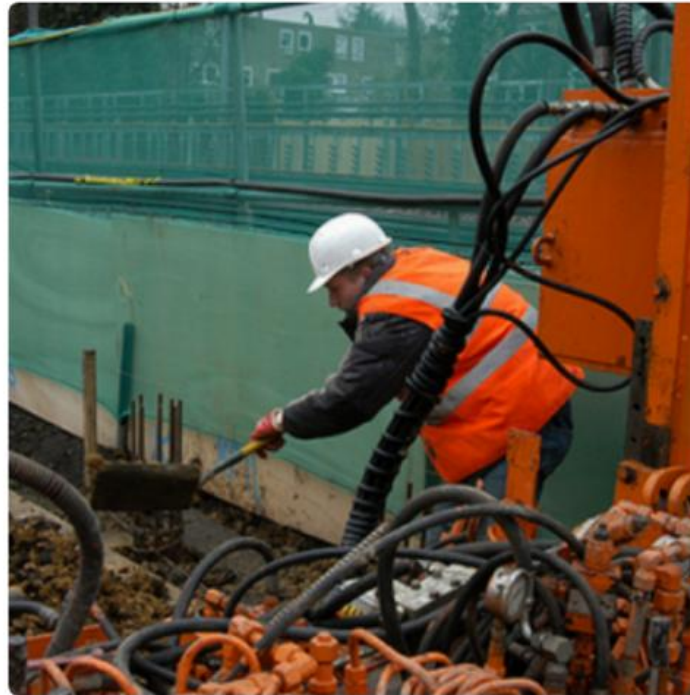
# Mục lục

- 1 Giới thiệu bài toán tự động mô tả hình ảnh
- 2 Cơ sở lý thuyết
- 3 Xây dựng mô hình
- 4 Đánh giá và kết quả
- 5 Kết luận

# Bài toán mô tả hình ảnh



"man in black shirt is playing guitar."

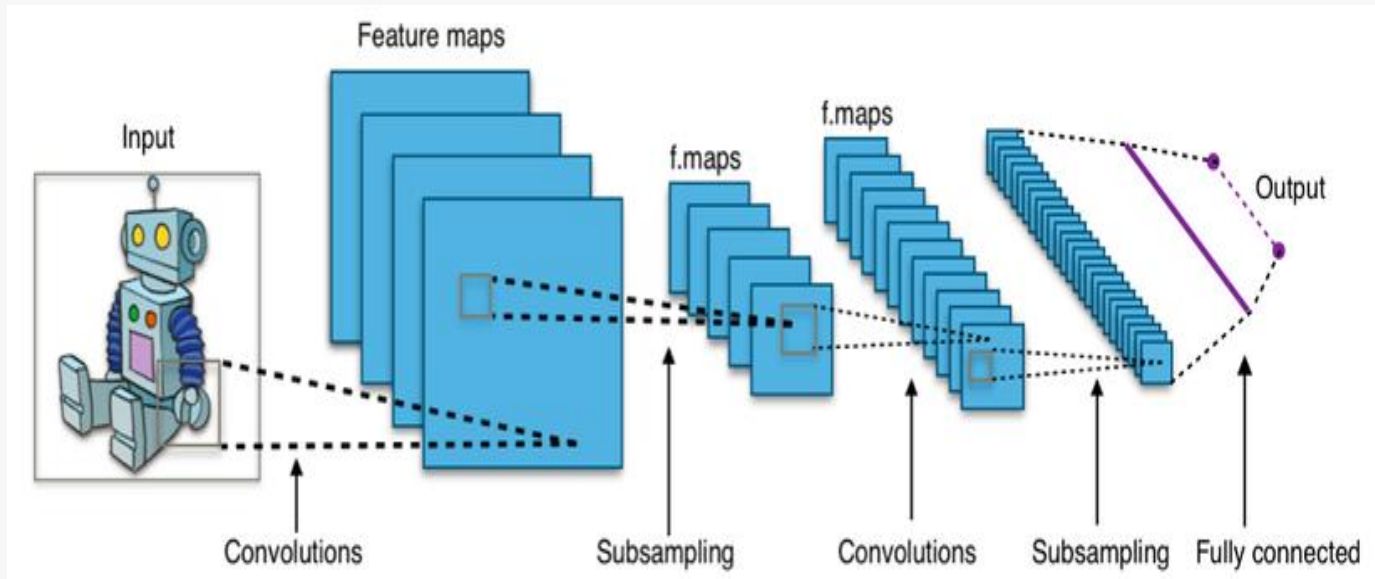


"construction worker in orange safety vest is working on road."



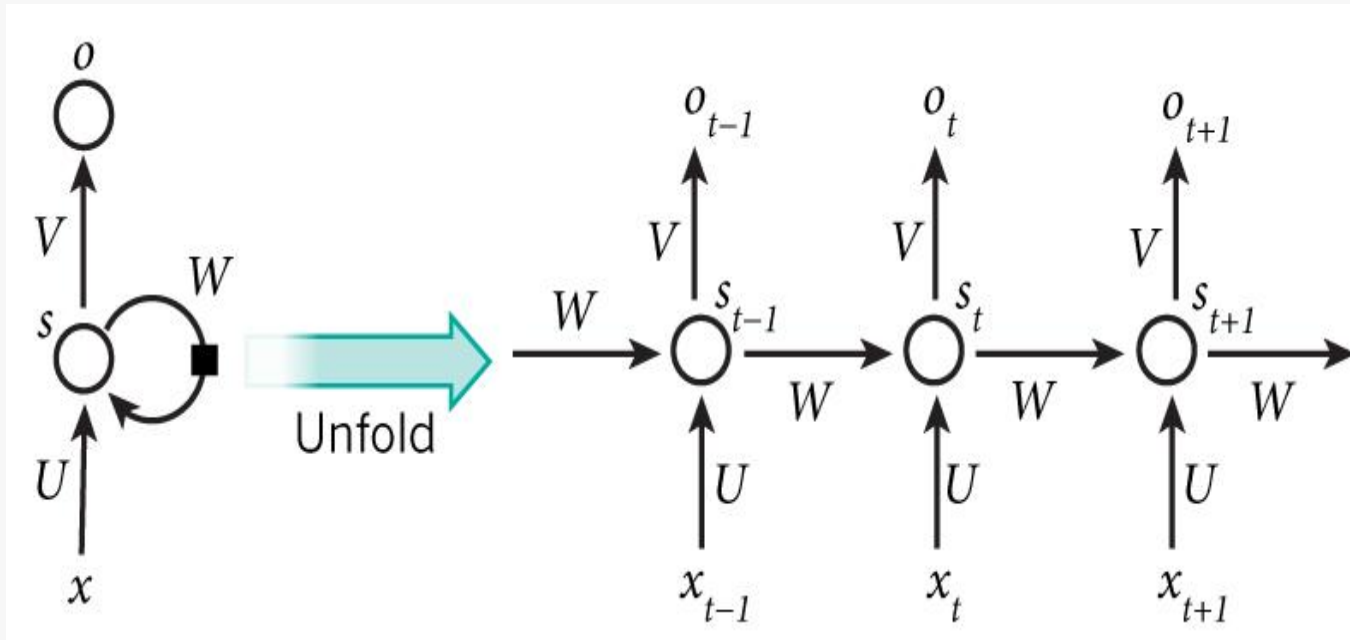
"two young girls are playing with lego toy."

# Mạng CNN



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.

# Mạng hồi quy RNN



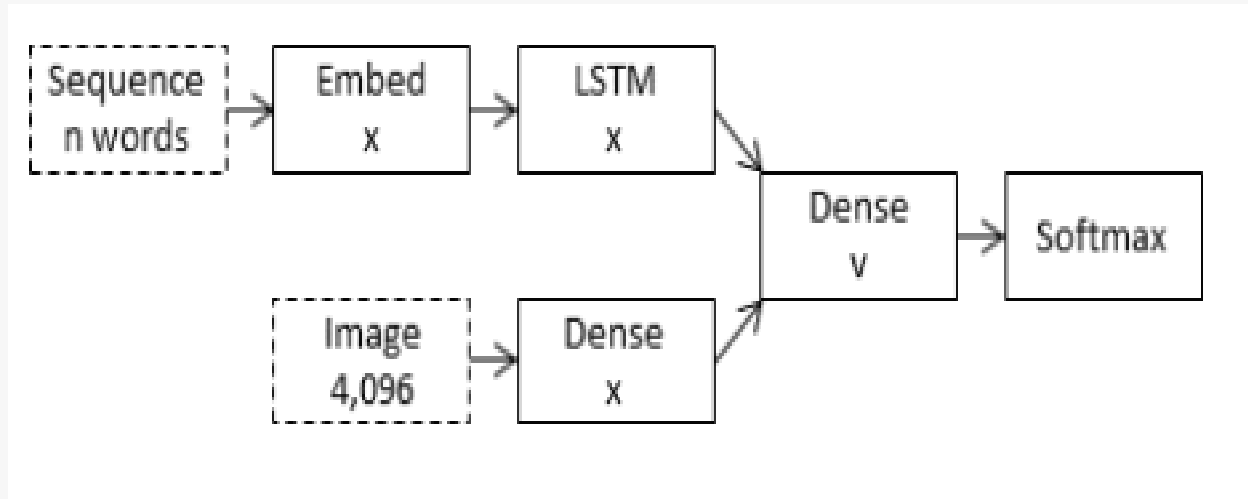
Quá trình tính toán bên trong một RNN, ta xét tại trạng thái  $t$ :

$x_t$  là đầu vào tại bước thứ  $t$ .

$s_t$  là trạng thái ẩn tại bước thứ  $t$ . Nó chính là bộ nhớ của mạng.  $s_t$  được tính toán thông qua cả đầu vào tại chính bước đó  $t$  và trạng thái ẩn trước đó  $t-1$ .  $s_t = f(Ux_t + Ws_{t-1})$  trong đó, hàm  $f$  có thể là hàm tanh hoặc ReLU. Để có thể tính toán được cho bước đầu tiên, thường khởi tạo  $s = 0$ .

$o_t$  là đầu ra tại bước thứ  $t$ . Ví dụ, ta muốn dự đoán từ tiếp theo có thể xuất hiện trong câu thì  $o_t$  chính là một véc tơ xác suất các từ trong danh sách từ vựng  $o_t = \text{softmax}(Vs_t)$ .

# Mô hình sinh ngôn ngữ

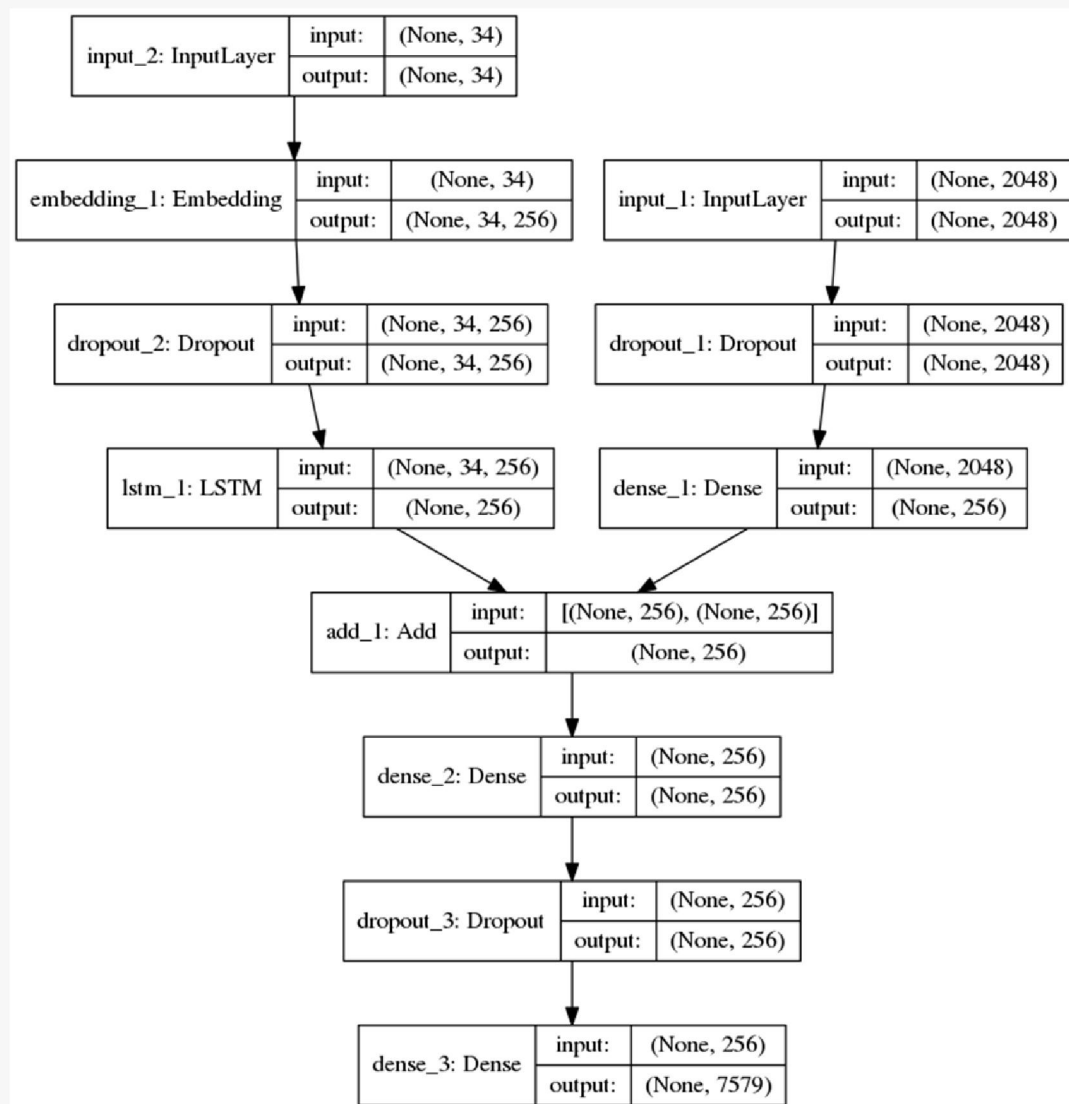


Dựa trên các mô hình đã công bố, ta đưa ra một vài ý tưởng đặc trưng xuyên suốt:

- Kết hợp mạng tích chập CNN và mạng hồi quy LSTM
- CNN trích xuất dữ liệu ảnh (encoder)
- LSTM sinh câu miêu tả (encoder)
- Decoder thông qua một tầng fully connected, end-to-end model: image -> sentence
- Maximize  $P(S|I)$  - với  $P(S|I, w)$  là xác suất của câu miêu tả  $S$  nếu biết ảnh  $I$ , và tham số  $w$ .



# Mô hình sinh ngôn ngữ

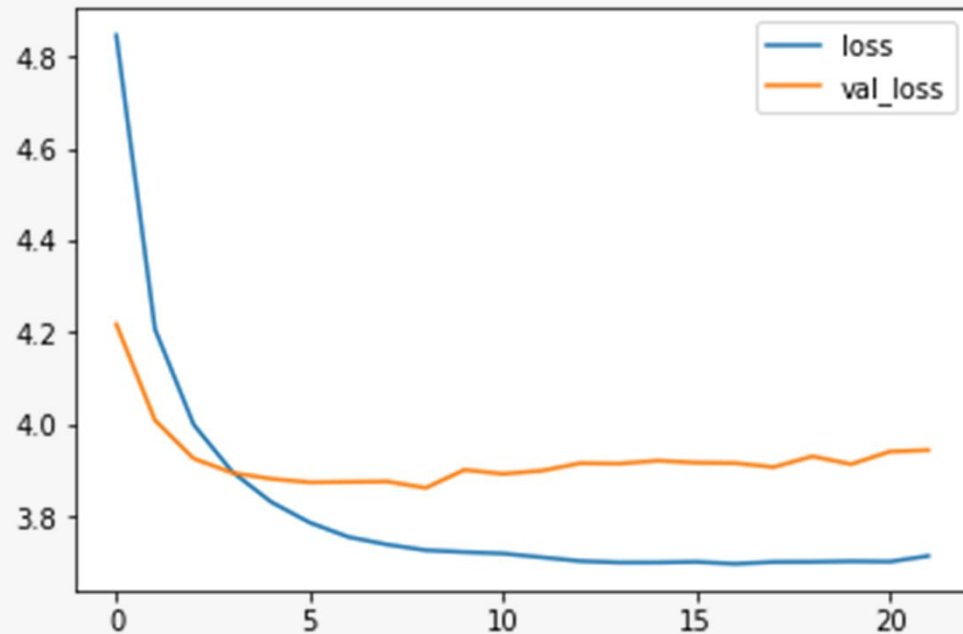


Cụ thể hóa, ta xây dựng mô hình mô tả hình ảnh trên nền keras như dưới hình vẽ:

- Tầng Photo Feature Extractor nhận đầu vào là một véc-tơ 4096 thành phần, được xử lý qua một lớp fully connected và đưa về một véc-tơ 256 thành phần.
- Tầng Sequence Processor nhận đầu vào là các câu với độ dài cố định (34 từ), sau khi đưa qua tầng embedding, rồi đưa qua tầng LSTM với 256 đơn vị nhớ (memory units) để đưa ra một véc-tơ 256 thành phần.
- Các tầng Dropout được thêm vào giúp mô hình có thể tránh được overfitting và quá trình huấn luyện nhanh hơn.
- Cuối cùng, các véc-tơ output 256 thành phần của Photo Feature Extractor và Sequence Processor được merge vào một véc-tơ (sử dụng phép cộng véc-tơ). Véc-tơ tổng đi qua dense layer cuối với activate function là softmax tính xác suất của từ tiếp theo.



# Mô hình sinh ngôn ngữ



Mô hình sử dụng hàm mất mát là `categorical_crossentropy`, optimizer Adam.

Sau khi train thử nghiệm trên 24 epochs, trên cloud của kaggle mất ~5 tiếng, biểu đồ loss theo epochs của mô hình như sau:

# Kết quả



BLEU-1: 0.542805  
BLEU-2: 0.301714  
BLEU-3: 0.207351  
BLEU-4: 0.095704



Thank you!!!

