ĐẠI HỌC TÔN ĐỨC THẮNG
TON DUC THANG UNIVERSITY

KHOA
CÔNG NGHỆ THÔNG TIN

Báo cáo giữa kì

Môn: Học Sâu_N01

GV hướng dẫn: PGS. TS. Lê Anh Cường
Người thực hiện: 52200167 – Lưu Hữu Trí
52200149 – Hồ Thu Yến Ngọc
52200193 – Nguyễn Quang Trung

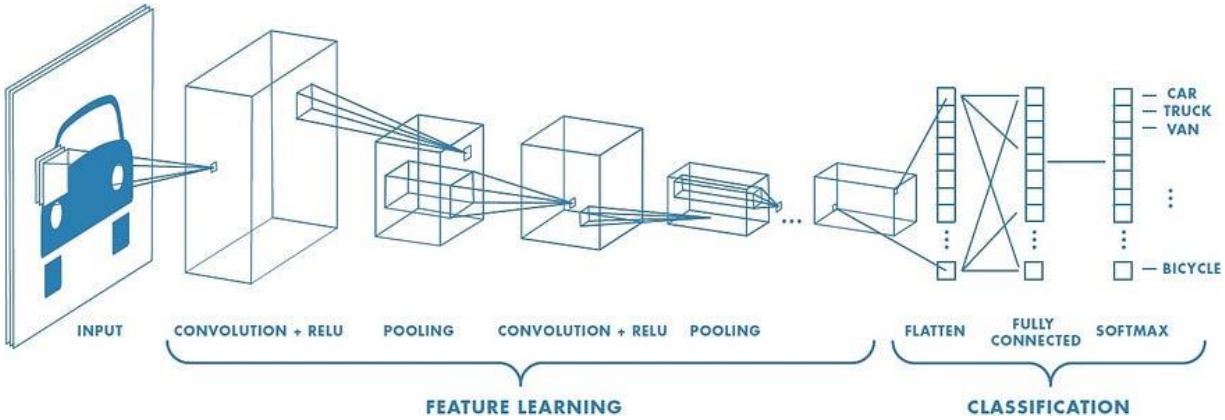
Convolutional Neurel-Network

- Tổng quan
- Phép tích chập
- Pooling
- Các kiến trúc nổi bật

Convolutional Neural-Network

- **CNN** là một loại mạng thần chuyên xử lí dữ liệu dạng lưới, điển hình là hình ảnh.
- Lấy cảm hứng từ cơ chế thị giác của con người, xử lý từng khu vực nhỏ để phát hiện đặc trưng của đối tượng.
- CNN thường được sử dụng để nhận diện vật thể, nhận diện khuôn mặt,

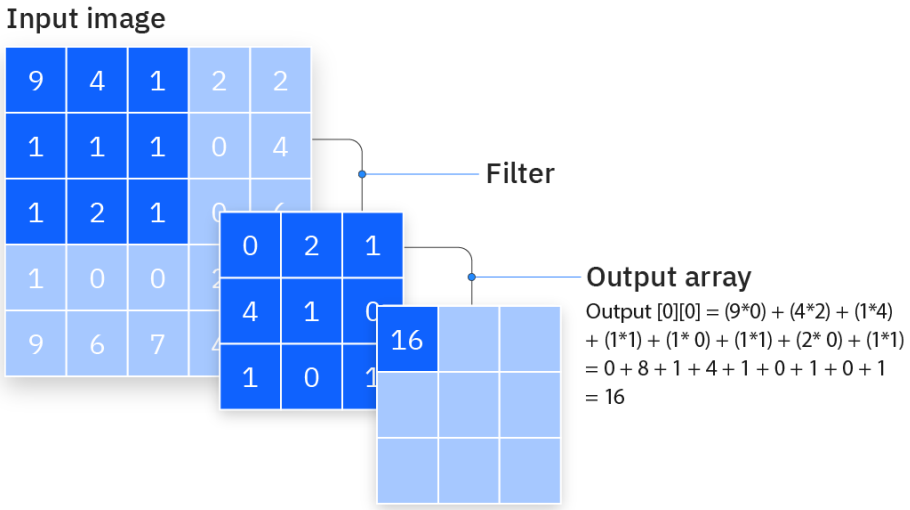
Kiến trúc CNN cơ bản



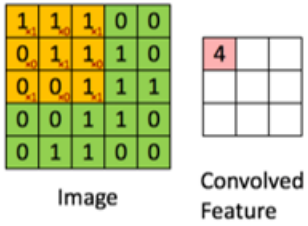
- **CNN** hoạt động dựa trên phép tích chập.
- Thường bao gồm các lớp: Convolution layer, pooling layer, activation layer và fully connected layer.

Phép tích chập

- Trích xuất đặc trưng của ảnh như cạnh, màu sắc, ...
- Dùng một cửa sở trượt, gọi là kernel (hoặc filter) để trượt trên dữ liệu và dùng phép nhân ma trận để trích xuất ra bản đồ đặc trưng (feature map).

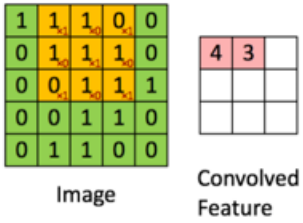


Phép tích chập



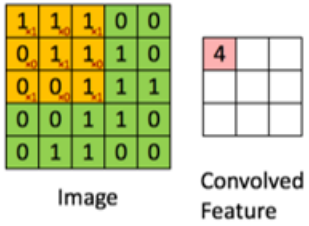
$1_{\times1}$	$1_{\times0}$	$1_{\times1}$	0	0
$0_{\times0}$	$1_{\times1}$	$1_{\times0}$	1	0
$0_{\times1}$	$0_{\times0}$	$1_{\times1}$	1	1
0	0	1	1	0
0	1	1	0	0

4		



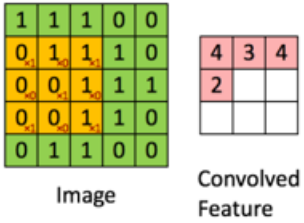
1	$1_{\times1}$	$1_{\times0}$	$0_{\times1}$	0
0	$1_{\times0}$	$1_{\times1}$	$1_{\times0}$	0
0	$0_{\times1}$	$1_{\times0}$	$1_{\times1}$	1
0	0	1	1	0
0	1	1	0	0

4	3	



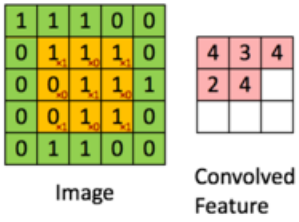
$1_{\times1}$	$1_{\times0}$	$1_{\times1}$	0	0
$0_{\times0}$	$1_{\times1}$	$1_{\times0}$	1	0
$0_{\times1}$	$0_{\times0}$	$1_{\times1}$	1	1
0	0	1	1	0
0	1	1	0	0

4		



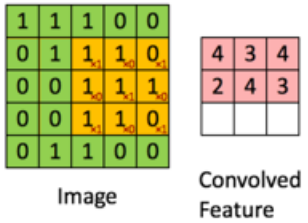
1	1	1	0	0
$0_{\times1}$	$1_{\times0}$	$1_{\times1}$	1	0
$0_{\times0}$	$0_{\times1}$	$1_{\times0}$	1	1
$0_{\times1}$	$0_{\times0}$	$1_{\times1}$	1	0
0	1	1	0	0

4	3	4
2		



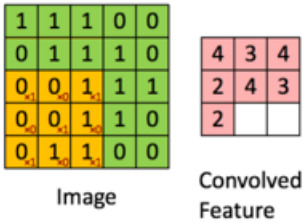
1	1	1	0	0
0	$1_{\times1}$	$1_{\times0}$	$1_{\times1}$	0
0	$0_{\times0}$	$1_{\times1}$	$1_{\times0}$	1
0	$0_{\times1}$	$1_{\times0}$	$1_{\times1}$	0
0	1	1	0	0

4	3	4
2	4	



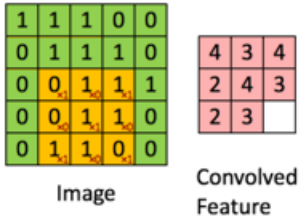
1	1	1	0	0
0	1	$1_{\times1}$	$1_{\times0}$	$0_{\times1}$
0	0	$1_{\times0}$	$1_{\times1}$	$1_{\times0}$
0	0	$1_{\times1}$	$1_{\times0}$	$0_{\times1}$
0	1	1	0	0

4	3	4
2	4	3



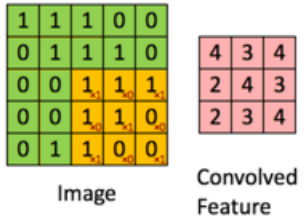
1	1	1	0	0
0	1	1	1	0
$0_{\times1}$	$0_{\times0}$	$1_{\times1}$	1	1
$0_{\times0}$	$0_{\times1}$	$1_{\times0}$	1	0
$0_{\times1}$	$1_{\times0}$	$1_{\times1}$	0	0

4	3	4
2	4	3
2		



1	1	1	0	0
0	1	1	1	0
0	$0_{\times1}$	$1_{\times0}$	$1_{\times1}$	1
0	$0_{\times0}$	$1_{\times1}$	$1_{\times0}$	0
0	$1_{\times1}$	$1_{\times0}$	$0_{\times1}$	0

4	3	4
2	4	3
2	3	



1	1	1	0	0
0	1	1	1	0
0	0	$1_{\times1}$	$1_{\times0}$	$1_{\times1}$
0	0	$1_{\times0}$	$1_{\times1}$	$0_{\times0}$
0	1	$1_{\times1}$	$0_{\times0}$	$0_{\times1}$

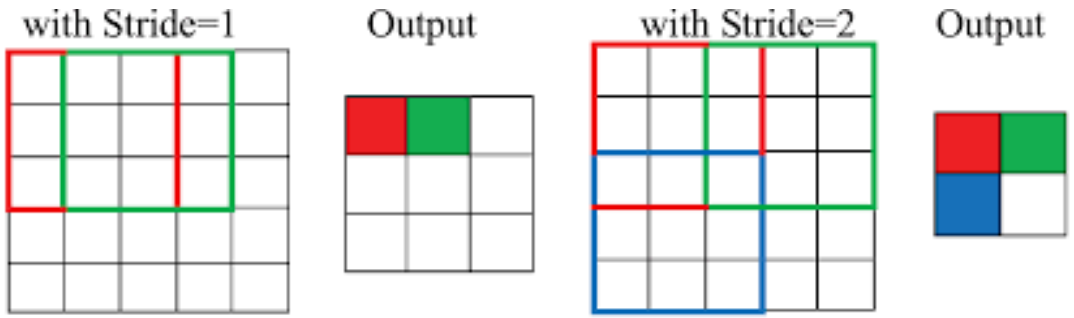
4	3	4
2	4	3
2	3	4

Padding (Đệm)

- Giảm mất mát thông tin ở viền ảnh.
- Giữ nguyên kích thước đầu ra.

0	0	0	0	0	0	0
0	1	1	1	0	0	0
0	0	1	1	1	0	0
0	0	0	1	1	1	0
0	0	0	1	1	0	0
0	0	1	1	0	0	0
0	0	0	0	0	0	0

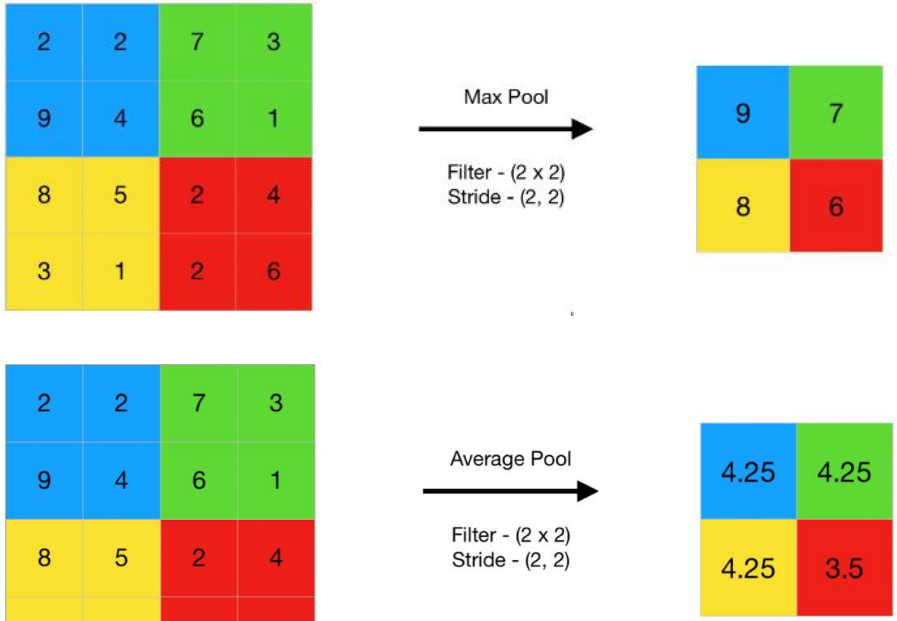
Stride (Bước nhảy)



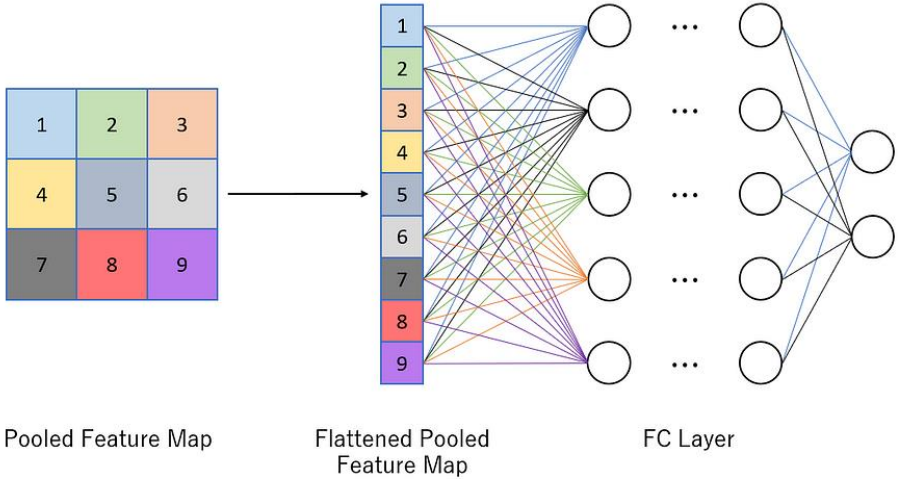
- Giảm số lượng phép toán cần tính.
- Giảm kích thước của feature map.

Pooling layer (Lớp gộp)

- Giảm số chiều của dữ liệu và giảm số lượt tính toán.
- Chỉ giữ lại thông tin quan trọng.
- Giảm overfit.
- Có hai loại pooling phổ biến: Max Pooling và Average Pooling



Fully connected layer



- Tổng hợp các đặc trưng từ các lớp tích chập trước để đưa ra quyết định cuối.
- Thông tin được truyền đến các nơ-ron (tương tự MLP).
- Thường được sử dụng ở cuối để đưa ra dự đoán

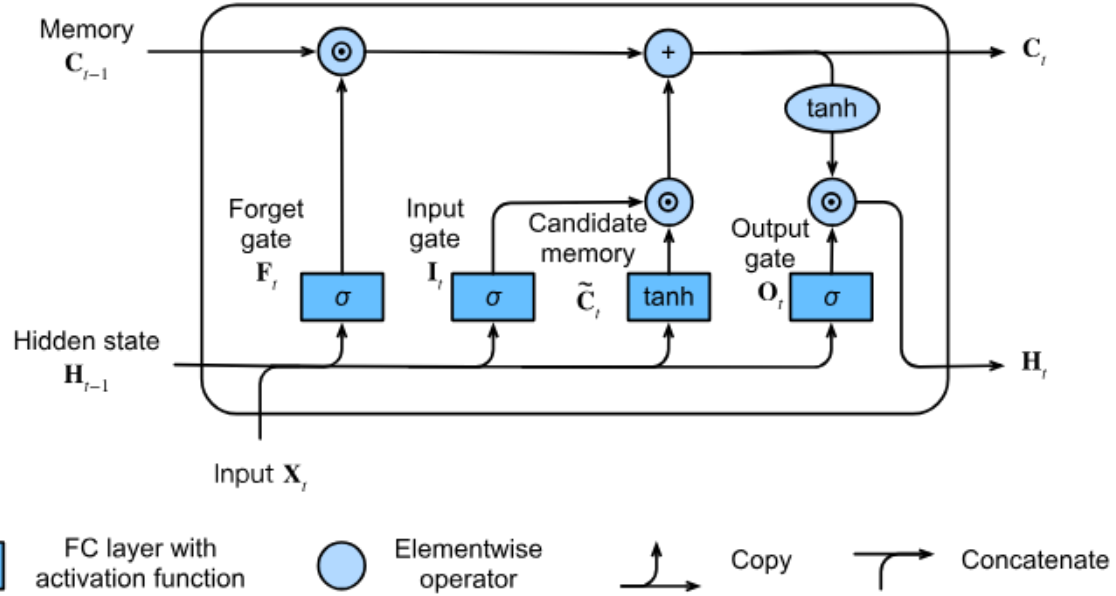
Long-Short Term Memory

- Tổng quan
- Cell state
- Các cổng

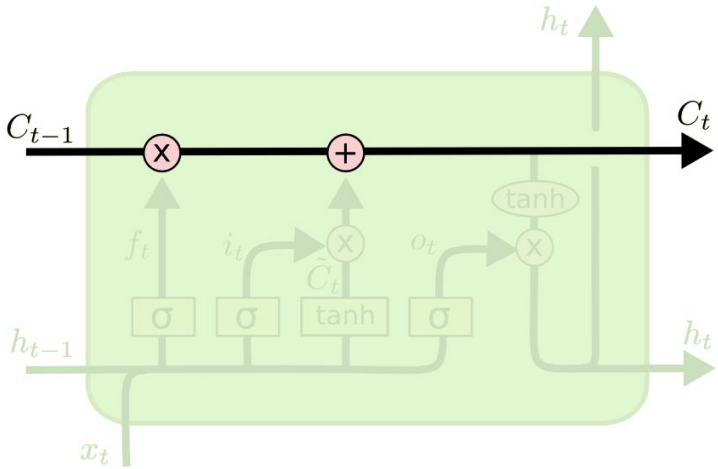
Tổng quan về LSTM

- **LSTM** là một cải tiến của mạng nơ-ron hồi tiếp RNN, có khả năng ghi nhớ thông tin trong khoảng thời gian dài hơn
- **Khắc** phục vấn đề gradient vanishing/exploding bằng cơ chế cổng (gate).
- Ứng dụng trong xử lí dữ liệu dạng chuỗi thời gian, văn bản, giọng nói, ...

Tổng quan về LSTM

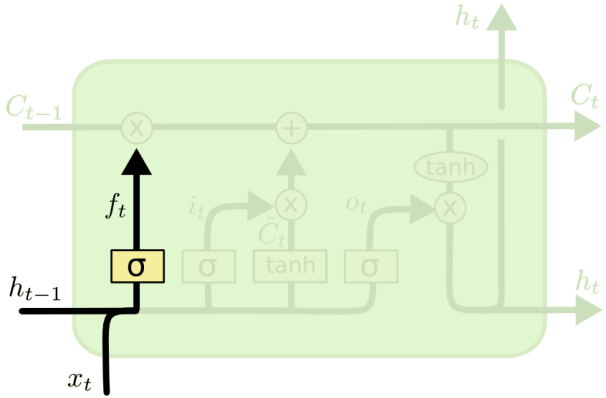


Cell state



- Là bộ nhớ chính của LSTM, giúp lưu trữ thông tin dài hạn.
- Được điều chỉnh bởi các cổng để thêm, giữ hoặc quên thông tin.
- **Khắc** phục vấn đề gradient vanishing/exploding.

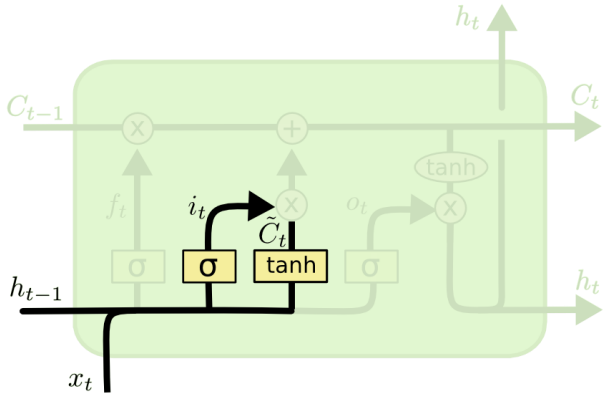
Forget gate



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

- Quyết định giữ lại hay loại bỏ thông tin khỏi cell state.
- Sử dụng hàm sigmoid để quyết định lượng thông tin được giữ lại, từ 0 là quên hoàn toàn cho đến 1 là giữ lại hoàn toàn.
- Giúp mô hình giữ lại thông tin quan trọng.

Input gate

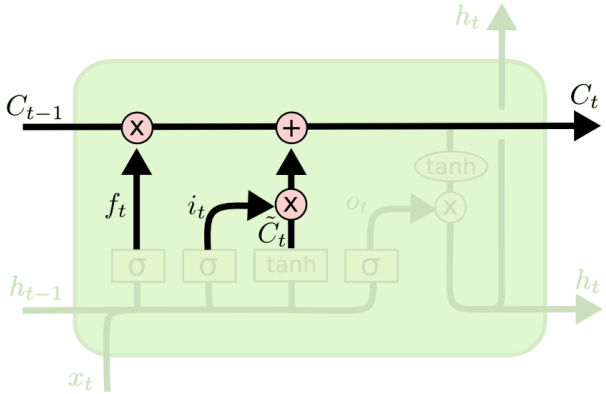


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

- Quyết định thêm thông tin mới vào cell state.
- Hàm sigmoid quyết định lượng thông tin được them vào cell state
- Hàm tanh tạo ra thông tin tiềm năng mới
- Giúp mô hình giữ lại thông tin quan trọng từ đầu vào mới.

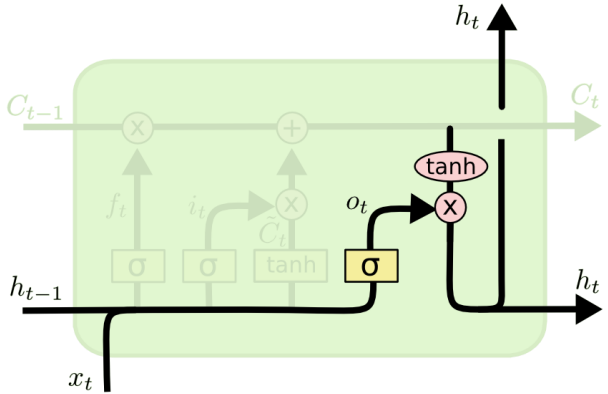
Cập nhật cell state



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

- Quyết định thêm thông tin mới vào cell state.
- Hàm sigmoid quyết định lượng thông tin được them vào cell state
- Hàm tanh tạo ra thông tin tiềm năng mới
- Giúp mô hình giữ lại thông tin quan trọng từ đầu vào mới.

Output gate



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

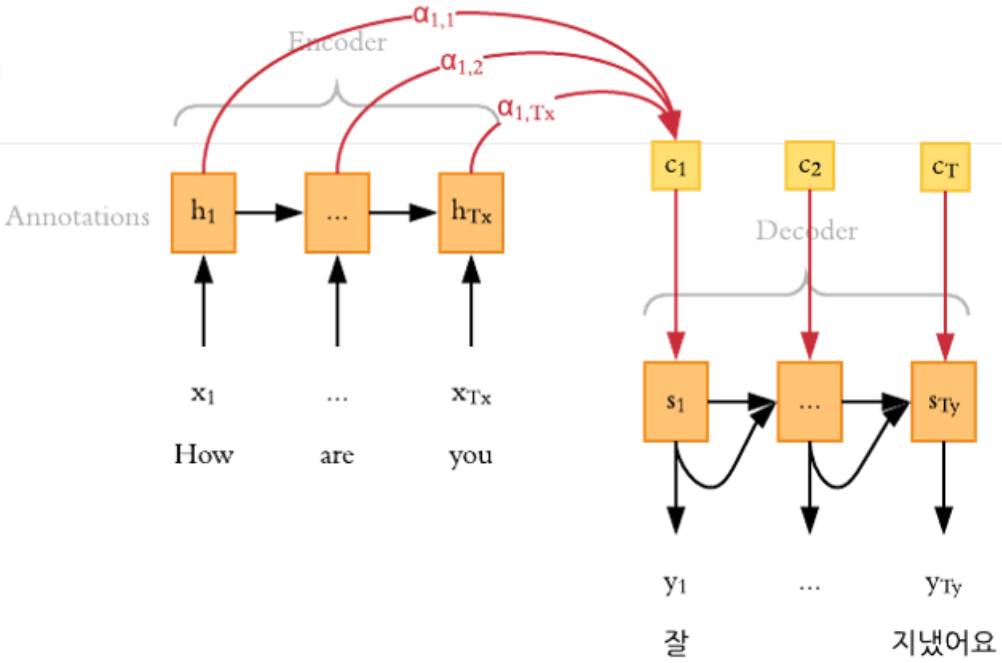
$$h_t = o_t * \tanh\left(C_t\right)$$

- Quyết định phần thông tin của cell state làm đầu ra.
- Dùng sigmoid để xác định thông tin cần truyền đi.
- Kết hợp với hàm tanh để tạo đầu ra phù hợp.

Cơ chế Attention

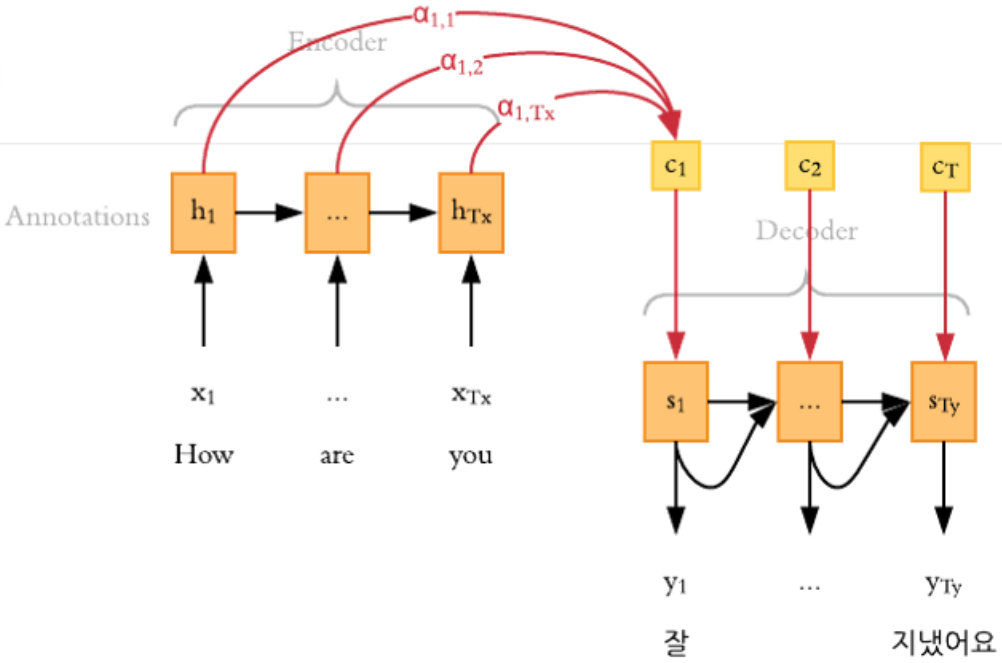
- Tổng quan
- Nguyên lí hoạt động
- Kết quả

Tổng quan



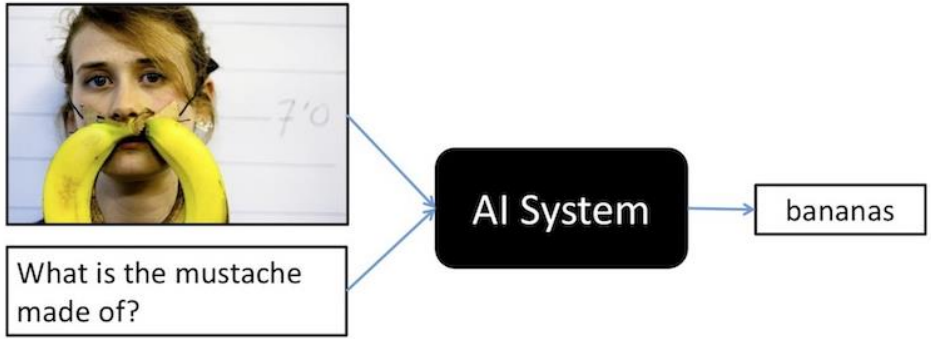
- Tập trung vào **thông tin quan trọng,** bỏ qua thông tin ít liên quan.
- Giúp mô hình xử lý chuỗi dài hiệu quả hơn bằng cách **gán trọng số** cho từng phần của đầu vào.
- Ứng dụng rộng rãi trong **NLP** (dịch máy, tóm tắt văn bản) và thị giác máy tính (nhận diện ảnh, **VQA**)

Tổng quan



- Tập trung vào **thông tin quan trọng,** bỏ qua thông tin ít liên quan.
- Giúp mô hình xử lý chuỗi dài hiệu quả hơn bằng cách **gán trọng số** cho từng phần của đầu vào.
- Ứng dụng rộng rãi trong **NLP** (dịch máy, tóm tắt văn bản) và thị giác máy tính (nhận diện ảnh, **VQA**)

Visual Question Answering



- Visual Question Answering (VQA) là một bài toán trong AI, trong đó mô hình cần trả lời câu hỏi dựa trên nội dung của một bức ảnh.
- Ứng dụng làm trợ lý ảo, hỗ trợ người khiếm thính, truy xuất thông tin trong ảnh.

Thực nghiệm

- Dataset
- Kiến trúc mô hình
- Kết quả

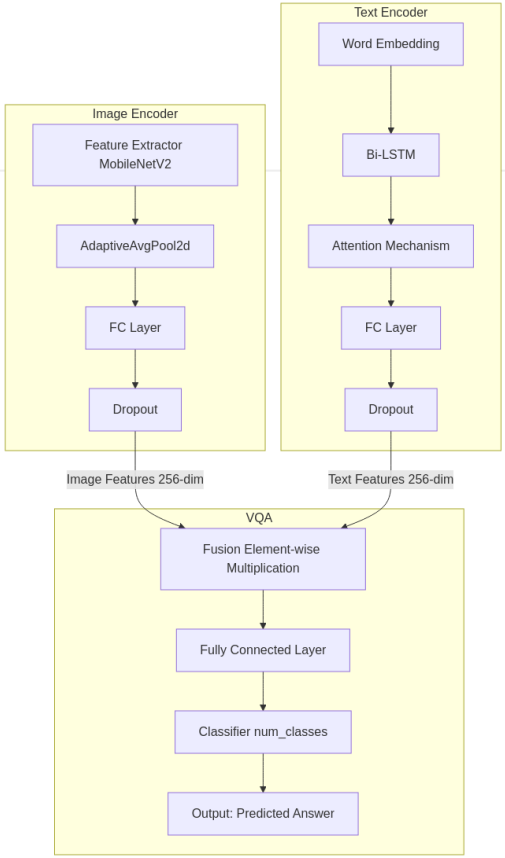
Dataset

Bộ dữ liệu gồm 65,331 mẫu, chia thành Train (41,613), Validation (4,839), Test (10,879), mỗi mẫu gồm câu hỏi, câu trả lời và hình ảnh

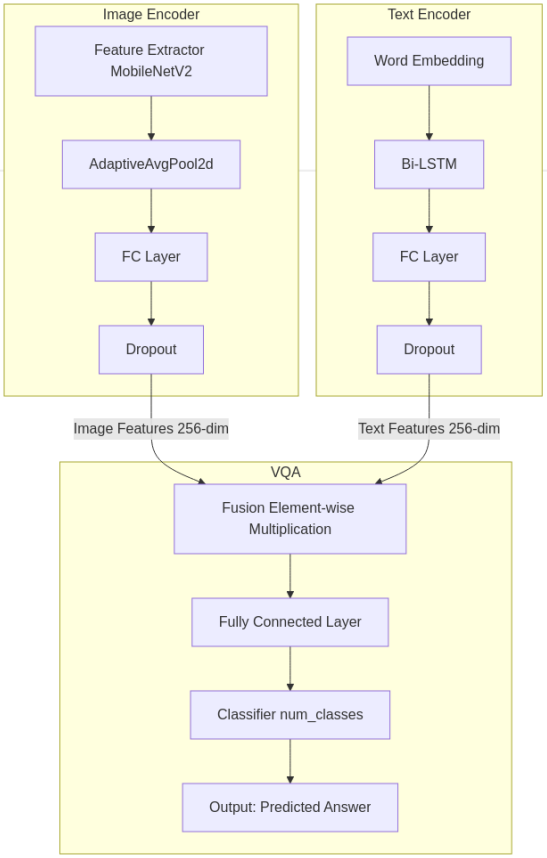
Q: món này là món gì
A: bánh bèo
Q: đây có phải là bánh bèo không
A: có
Q: màu chủ đạo của bánh bèo là gì
A: trắng

Q: món này là gì
A: bánh bèo
Q: màu chủ đạo của món ăn là gì
A: trắng
Q: đây có phải là món bánh bèo không
A: có

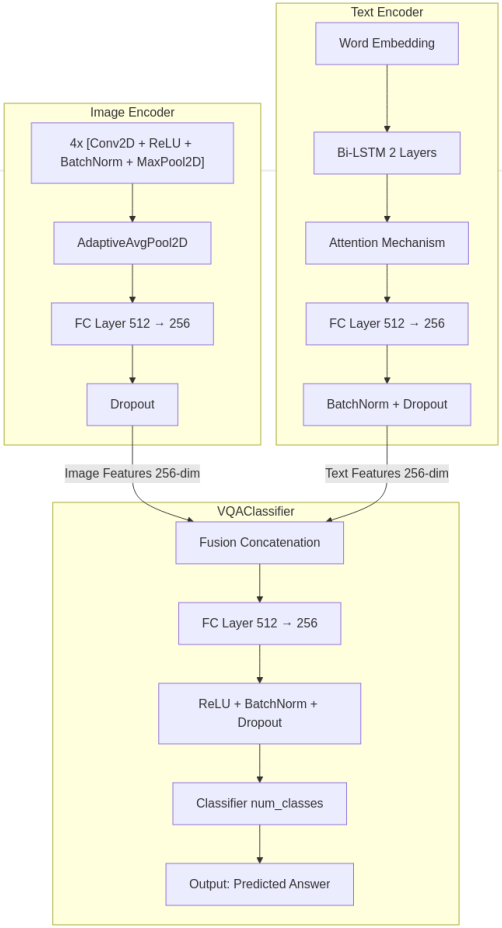
Q: món này là món gì
A: bánh bèo
Q: đây là món bánh bèo phải không
A: có
Q: màu sắc chủ đạo của món ăn là gì
A: trắng



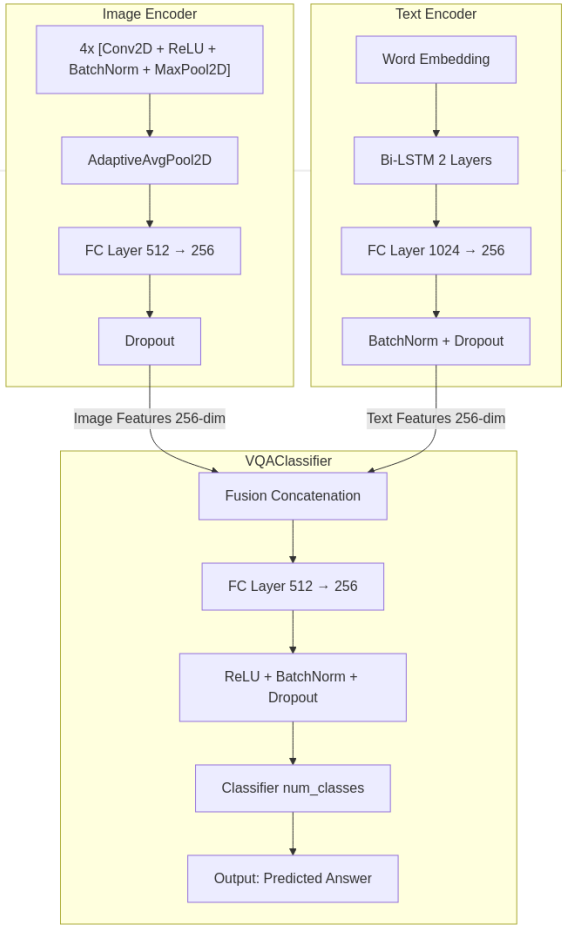
MobileNet_v2 + LSTM Attention



MobileNet_v2 + LSTM no Attention



CNN + LSTM Attention



CNN + LSTM no Attention

Kết quả

Model	Loss	Accuracy
MobileNet_v2 + LSTM Attention	0.4503	0.8317
MobileNet_v2 + LSTM no Attention	0.4682	0.8288
CNN + LSTM Attention	0.5589	0.8032
CNN + LSTM no Attention	0.6600	0.7703