THỊ GIÁC MÁY TÍNH
Chương 12: Các kỹ thuật học sâu cho nhận dạng chuỗi video

Thi-Lan Le, Hong-Quan Nguyen
Thi-Lan.Le@mica.edu.vn

Nội dung buổi học

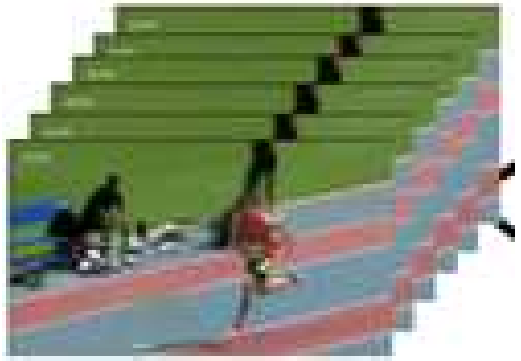
- Dữ liệu chuỗi
- Các cách tiếp cận trong nhận dạng video
- RNN (Recurrent Neural Network)
- LSTM (Long Short Term Memory)
- Định danh lại người trong mạng camera

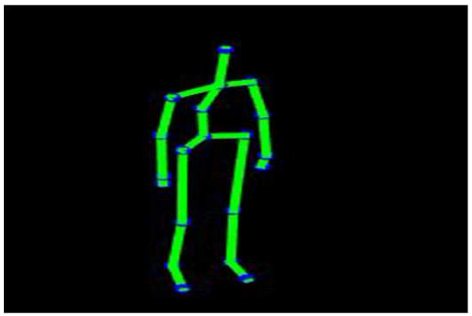
Nội dung buổi học

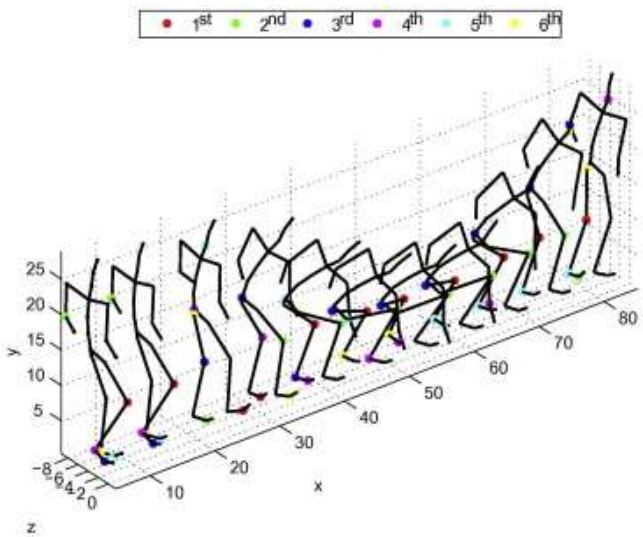
- Dữ liệu chuỗi
- Các cách tiếp cận trong nhận dạng video
- RNN (Recurrent Neural Network)
- LSTM (Long Short Term Memory)
- Định danh lại người trong mạng camera

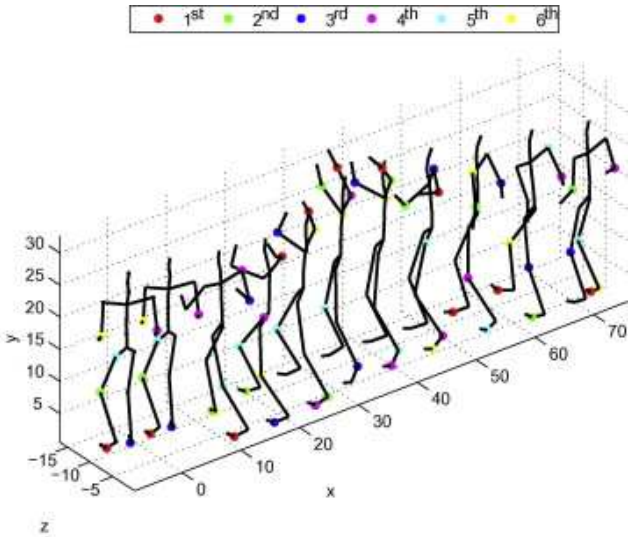
Dữ liệu chuỗi (1)

- Giá chứng khoán
- Chuỗi DNA
- Dữ liệu video
- Chuỗi từ/ký tự
- ……



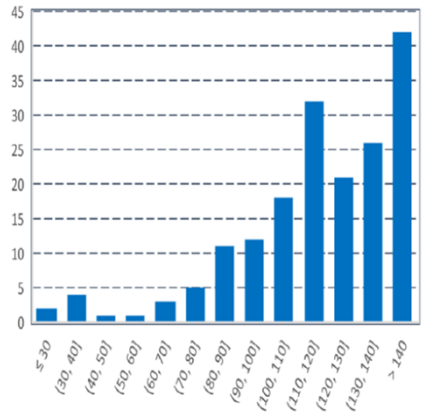






Dữ liệu chuỗi (2)

- Các đặc điểm chính khi làm việc trên dữ liệu chuỗi
  - Thường có **yếu tố thời gian/trật tự**
  - Dữ liệu lớn, **số lượng các phần tử có thể không bằng nhau** (ví dụ: số frame trong cùng một hoạt động thực hiện bởi nhiều người khác nhau thường khác nhau)
  - Có mối quan hệ giữa các phần tử (ví dụ: từ đứng trước và từ đứng sau trong 1 câu)



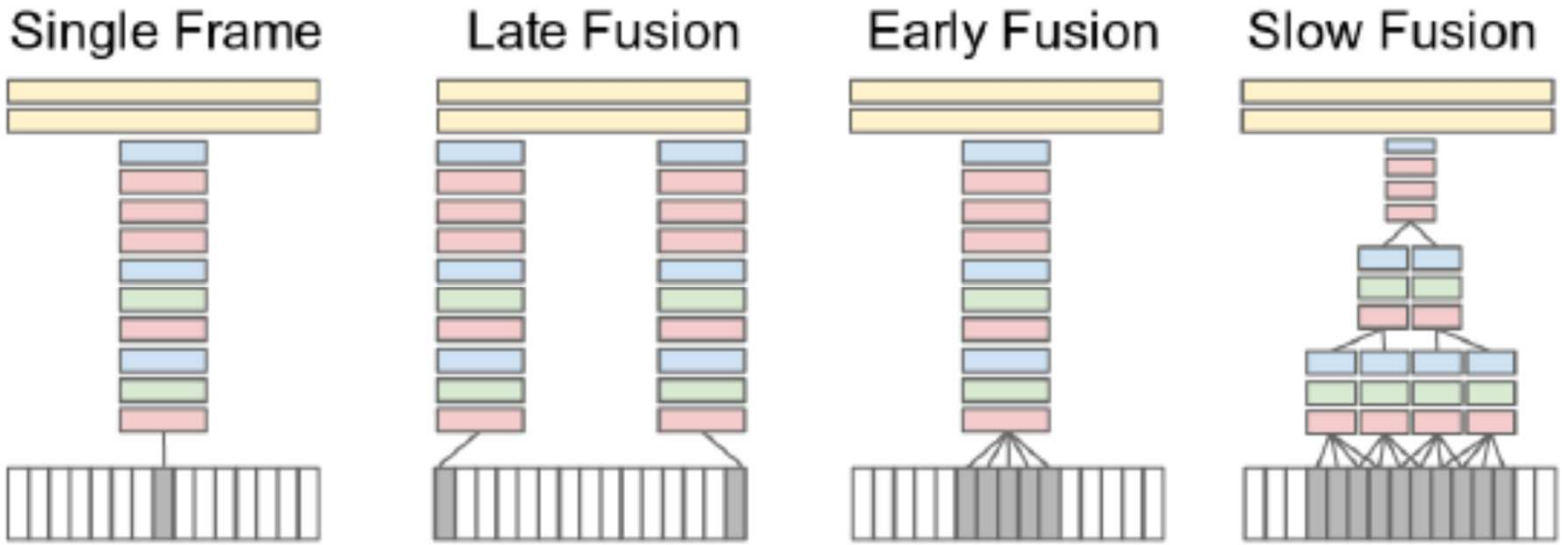
Phân bố số lượng ảnh của từng người trong CSDL PRID 2011

Nội dung buổi học

- Dữ liệu chuỗi
- Các cách tiếp cận trong nhận dạng video
- RNN (Recurrent Neural Network)
- LSTM (Long Short Term Memory)
- Định danh lại người trong mạng camera

Các cách tiếp cận trong nhận dạng video

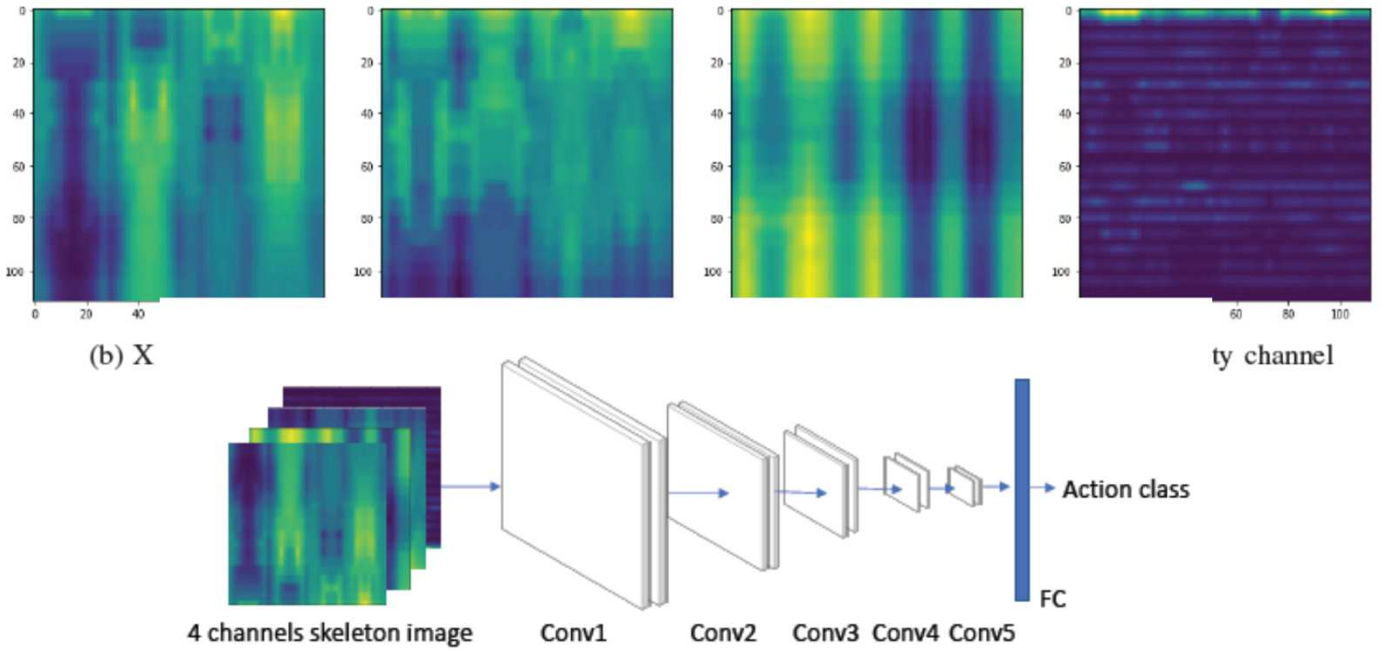
- **<u>Cách tiếp cận 1</u>**: Bổ sung thêm **các cơ chế/bước xử lý** để áp dụng được **các mạng CNN /phương pháp có sẵn làm việc trên dữ liệu ảnh**



Karpathy et al. Large-scale Video Classification with Convolutional Neural Networks.

Các cách tiếp cận trong nhận dạng video

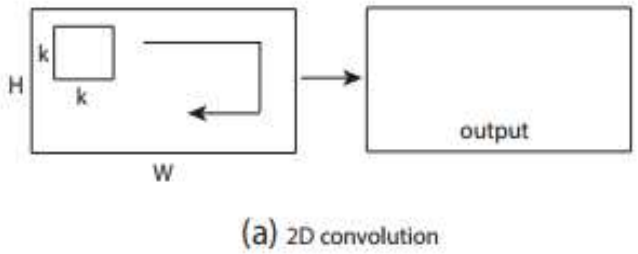
- **<u>Cách tiếp cận 2</u>**: Biến đổi dữ liệu chuỗi thành dữ liệu ảnh để có thể áp dụng được các phương pháp đang có trên ảnh



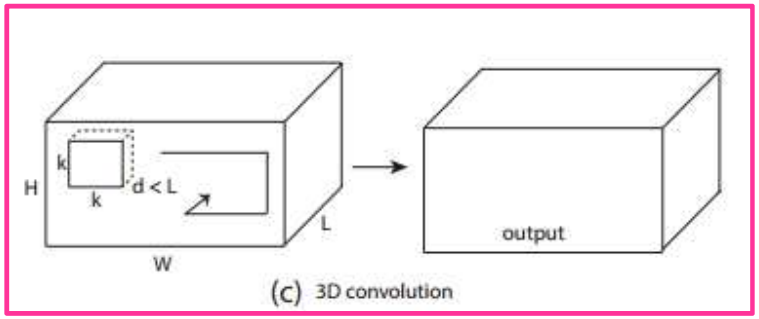
Van-Nam Hoang, Van-Toi Nguyen, Thi-Lan Le, Thanh-Hai Tran, Hai Vu, Activity recognition from skeleton using deep neural networks, Multimedia Analysis and Pattern Recognition (MAPR) 2019

Các cách tiếp cận trong nhận dạng video

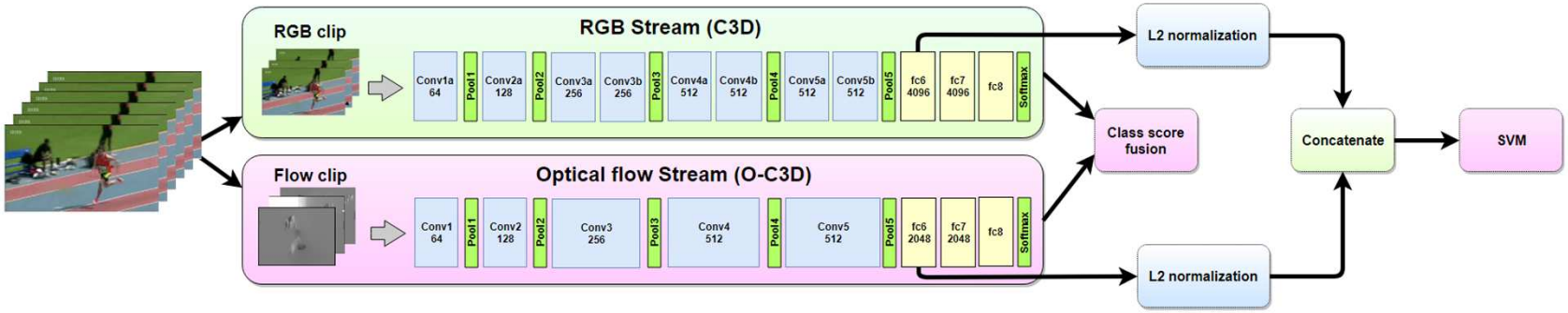
- **<u>Cách tiếp cận 3</u>**: Xây dựng mạng 3D



(a) 2D convolution



(c) 3D convolution



Các cách tiếp cận trong nhận dạng video

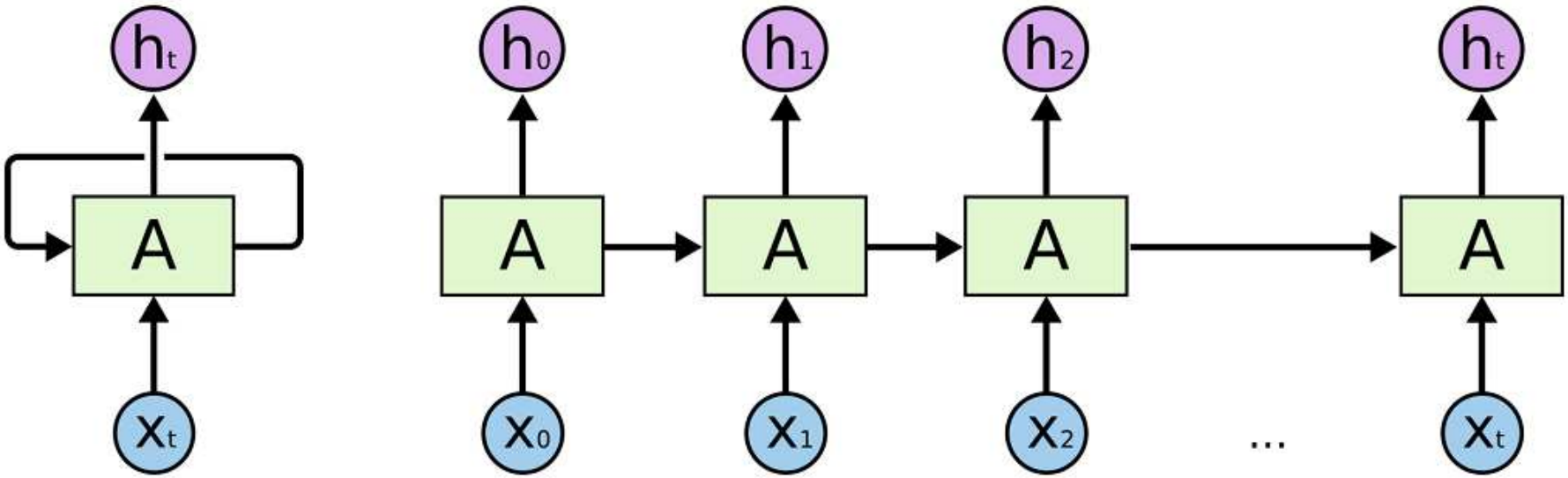
- **<u>Cách tiếp cận 4</u>**: Xây dựng mạng cho phép biểu diễn tường minh mối liên hệ về mặt thời gian RNN, LSTM→ **nội dung của buổi học**

Nội dung buổi học

- Dữ liệu chuỗi
- Các cách tiếp cận trong nhận dạng video
- RNN (Recurrent Neural Network)
- LSTM (Long Short Term Memory)
- Định danh lại người trong mạng camera

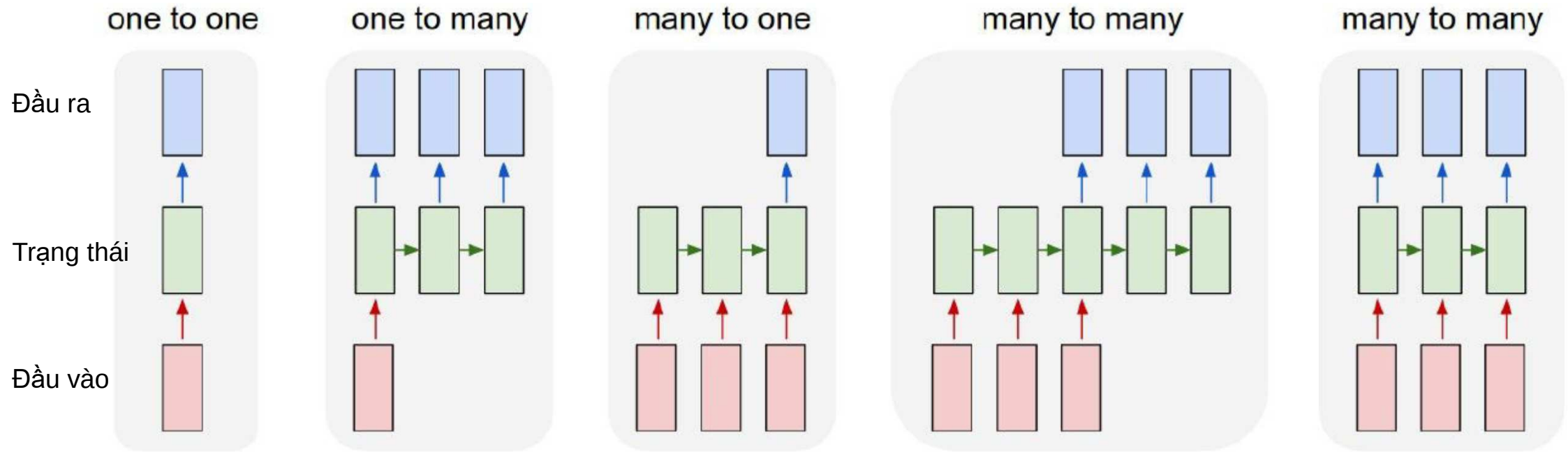
RNN (Recurrent Neural Network)

- X_t : đầu vào ở thời điểm t
- h_t : đầu ra ở thời điểm t



Nguồn hình ảnh: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

RNN (Recurrent Neural Network)



Nguồn: CS231n Lecture 10

RNN (Recurrent Neural Network)

- **“one to one”**
- Ví dụ: phân loại ảnh
  - Đầu vào: 1 ảnh
  - Đầu ra: lớp mà ảnh thuộc vào

Steel drum

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

✔

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

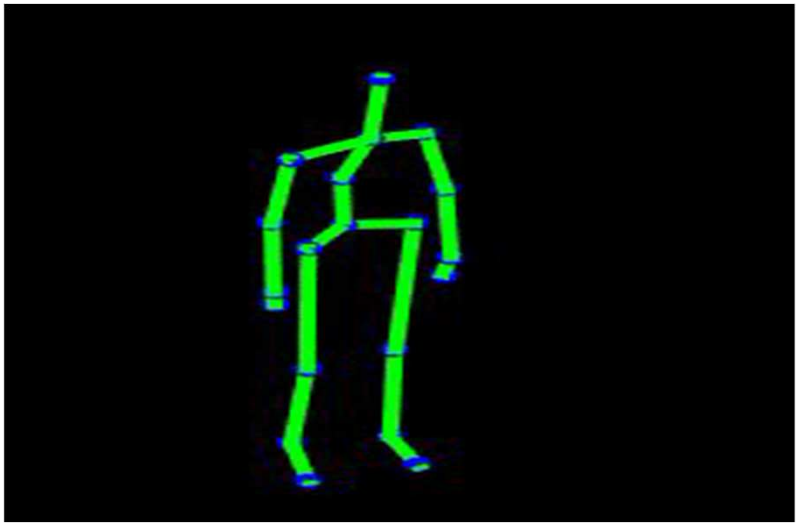
RNN (Recurrent Neural Network)

- **“one to many”**
- Ví dụ “image captioning”
  - Đầu vào: 1 ảnh
  - Đầu ra: một chuỗi các từ

https://www.captionbot.ai/: it's a close up of a flower

RNN (Recurrent Neural Network)

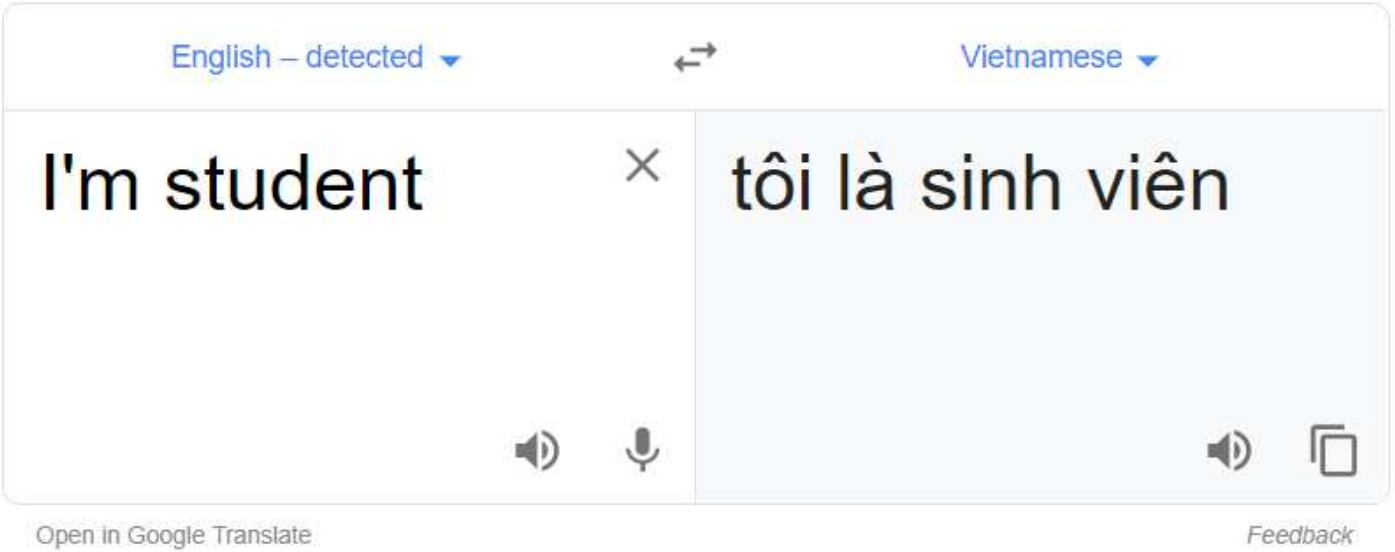
- **“many to one”**
- Ví dụ: nhận dạng hoạt động trên video
  - Đầu vào: một chuỗi các khung hình/khung xương
  - Đầu ra: nhãn của hoạt động



Hoạt động: ném

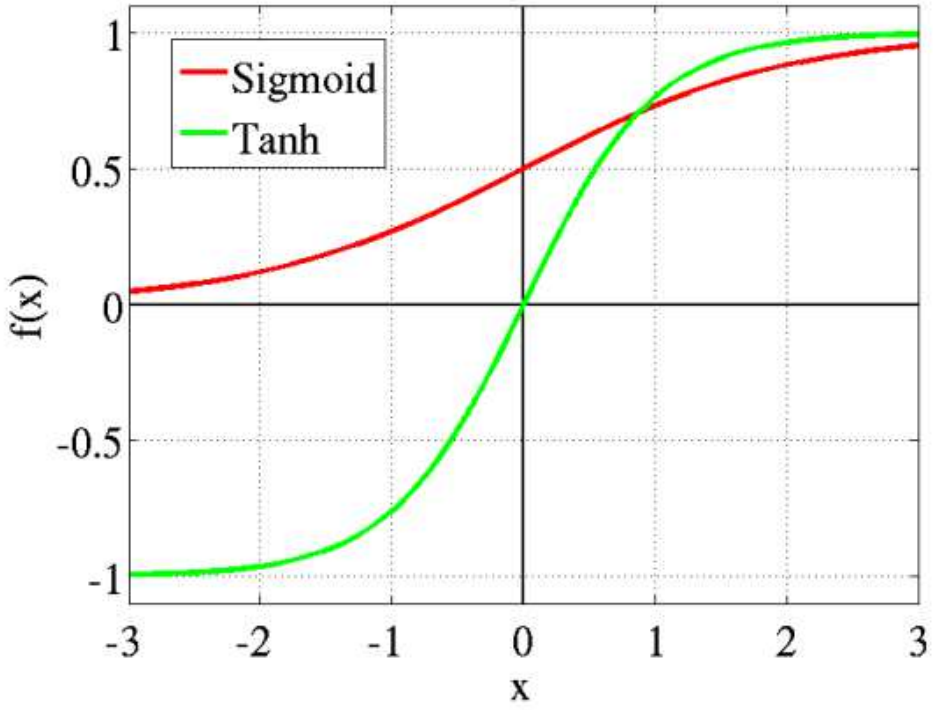
RNN (Recurrent Neural Network)

- **“many to many”**
- Ví dụ: dịch tự động
  - Đầu vào: 1 câu – 1 chuỗi các từ
  - Đầu ra: 1 câu – 1 chuỗi các từ



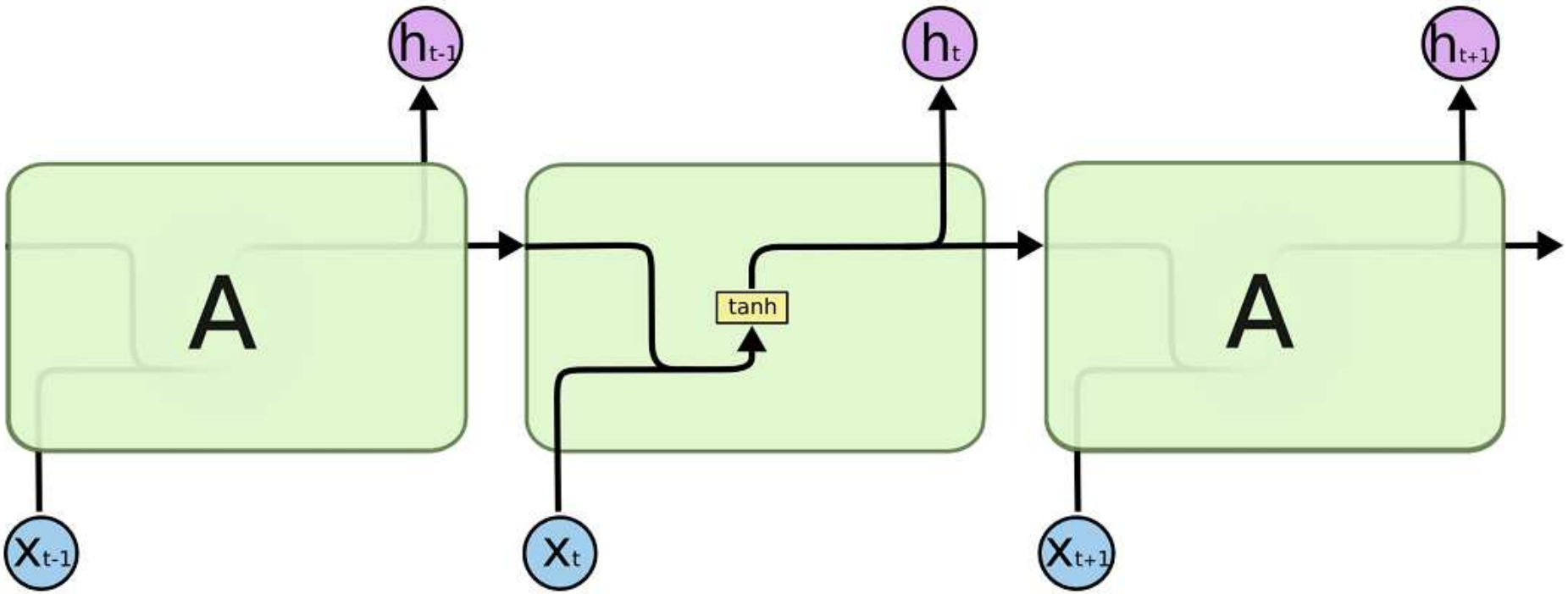
RNN (Recurrent Neural Network)

- Hàm tanh và hàm sigmoid



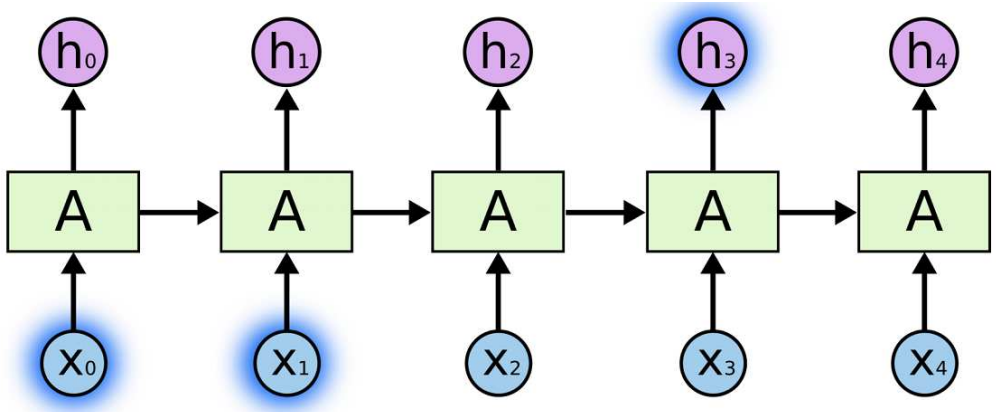
RNN (Recurrent Neural Network)

- Mỗi nút của RNN bao gồm 1 mạng duy nhất:

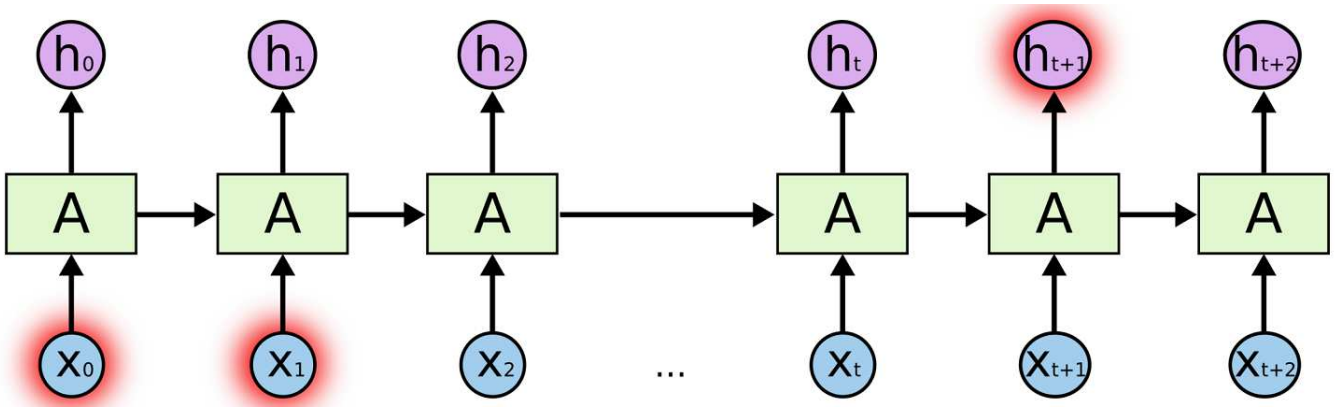


RNN (Recurrent Neural Network)

- Vấn đề “phụ thuộc lâu dài”



Trường hợp RNN có khả năng xử lý



Trường hợp RNN không có khả năng xử lý

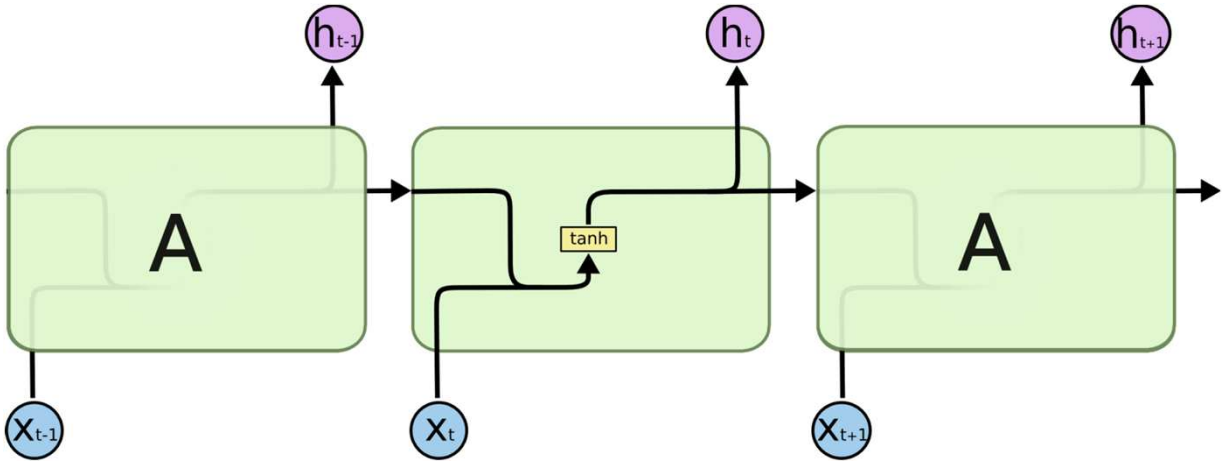
Nguồn hình ảnh: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Nội dung buổi học

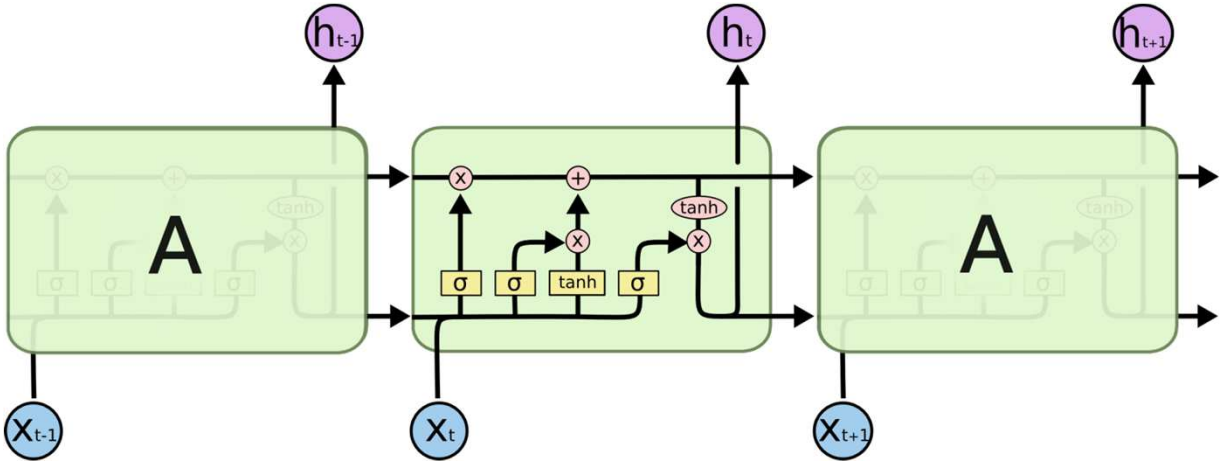
- Dữ liệu chuỗi
- Các cách tiếp cận trong nhận dạng video
- RNN (Recurrent Neural Network)
- LSTM (Long Short Term Memory)
- Định danh lại người trong mạng camera

LSTM (Long Short Term Memory)

- LSTM được đưa ra vào 1997

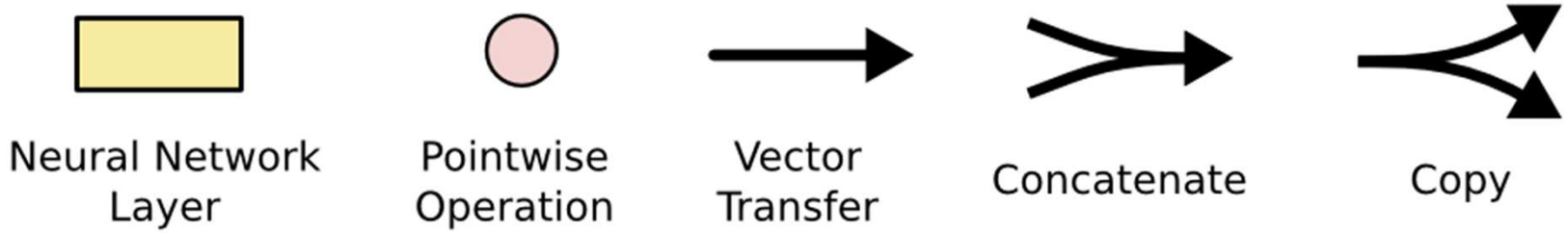


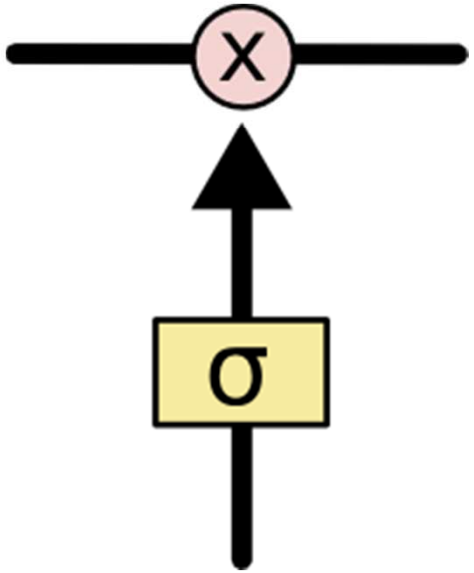
Cấu trúc chung của một RNN



Cấu trúc chung của một LSTM

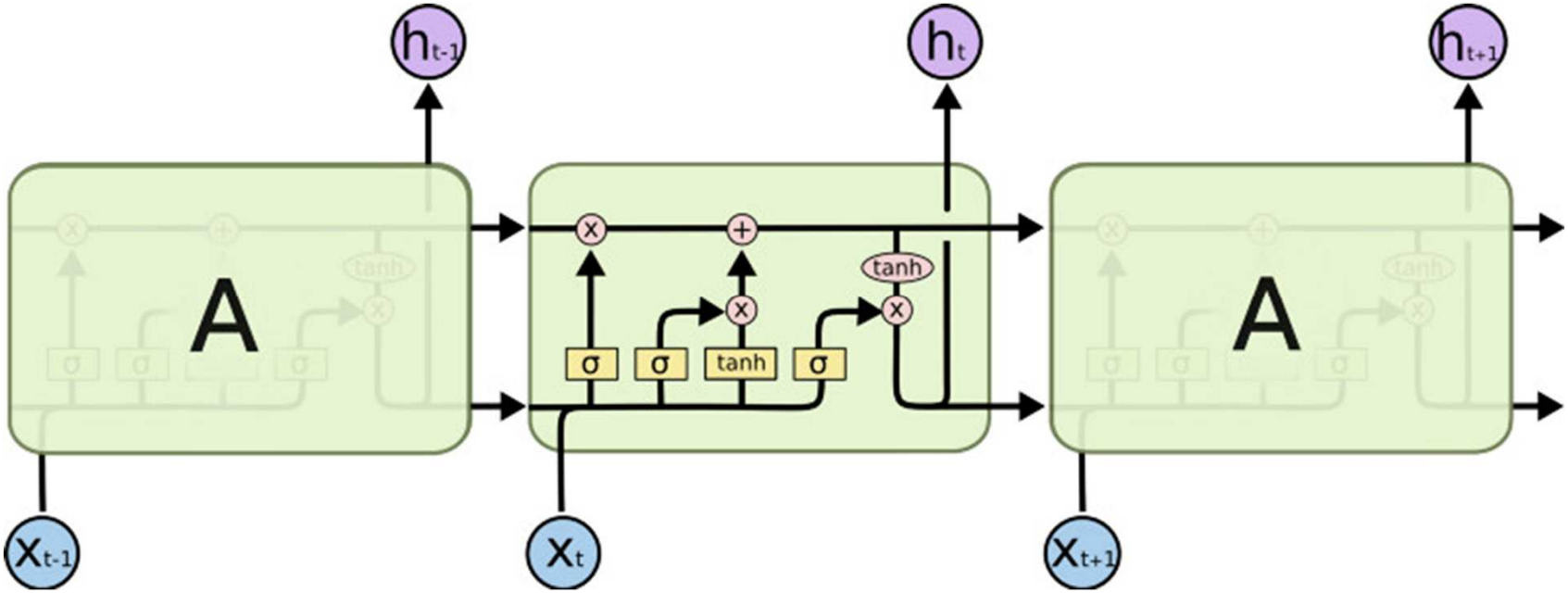
LSTM (Long Short Term Memory)





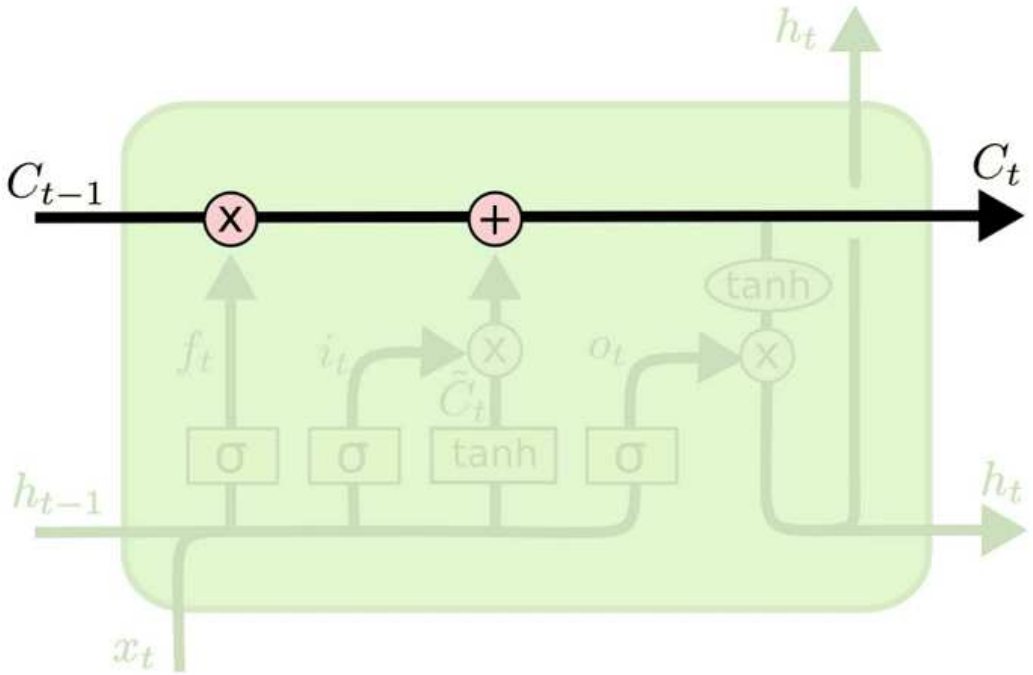
Ví dụ:
Một mạng sigmoid (đầu ra từ 0 đến 1)
Cổng X thực hiện phép nhân

LSTM (Long Short Term Memory)



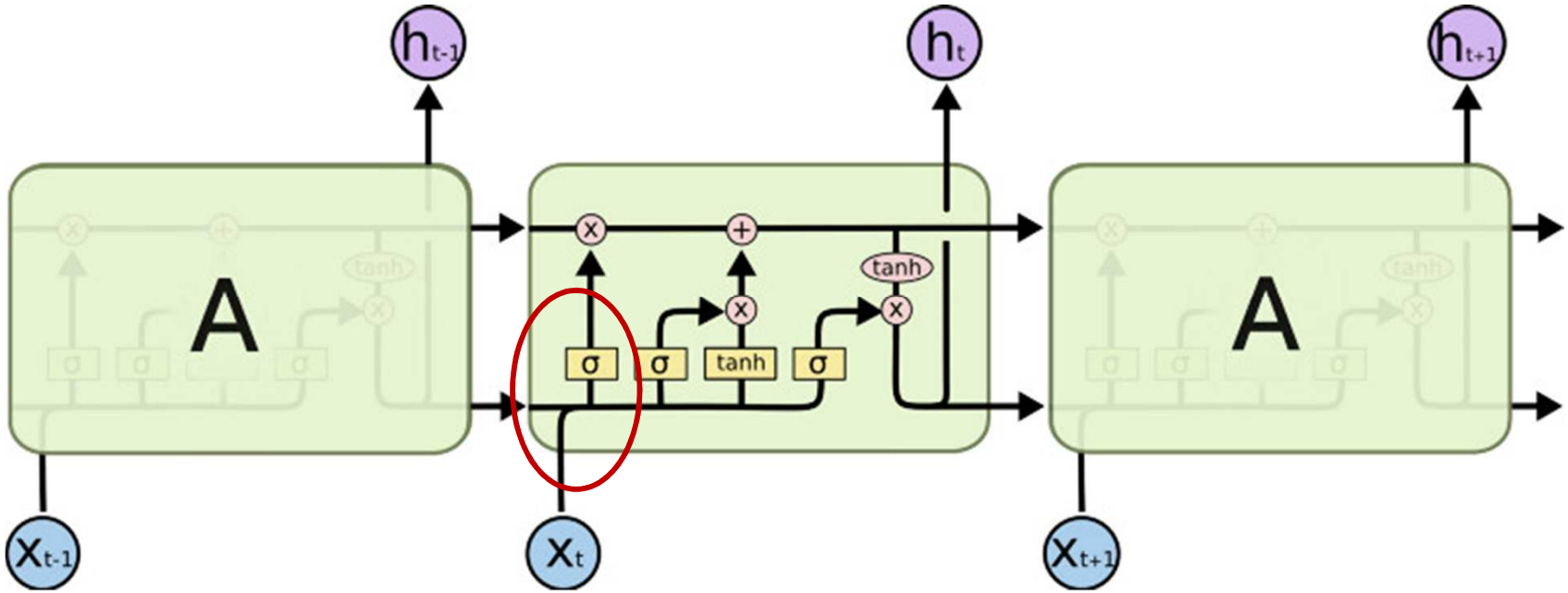
LSTM (Long Short Term Memory)

- “Cell state”
- C_{t-1} : trạng thái ở thời điểm t-1
- C_t : trạng thái ở thời điểm t



LSTM (Long Short Term Memory)

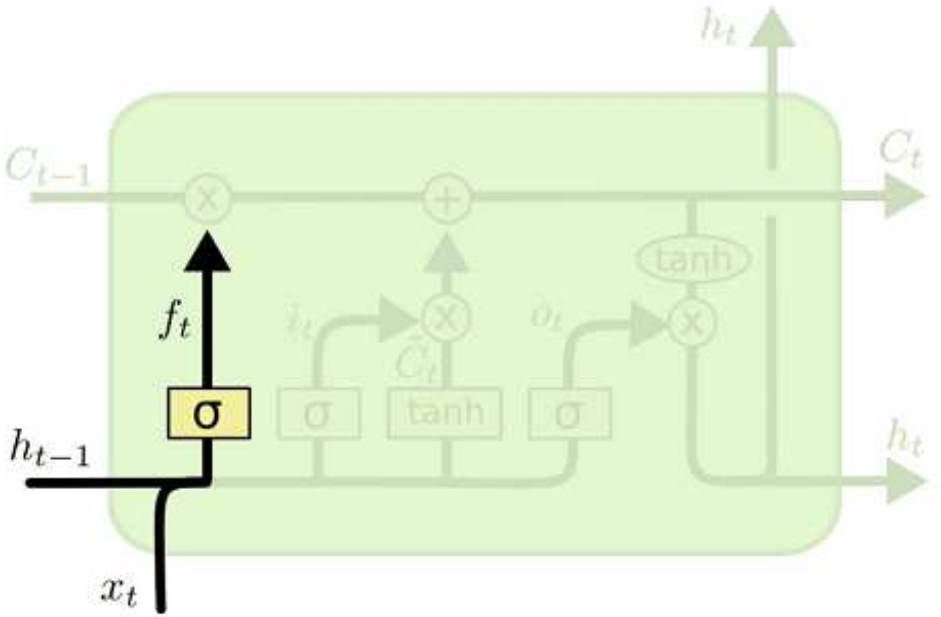
- Cổng quên – forget gate:
  - Quyết định sẽ quyết định thông tin gì sẽ được giữ lại từ trạng thái trước đó



LSTM (Long Short Term Memory)

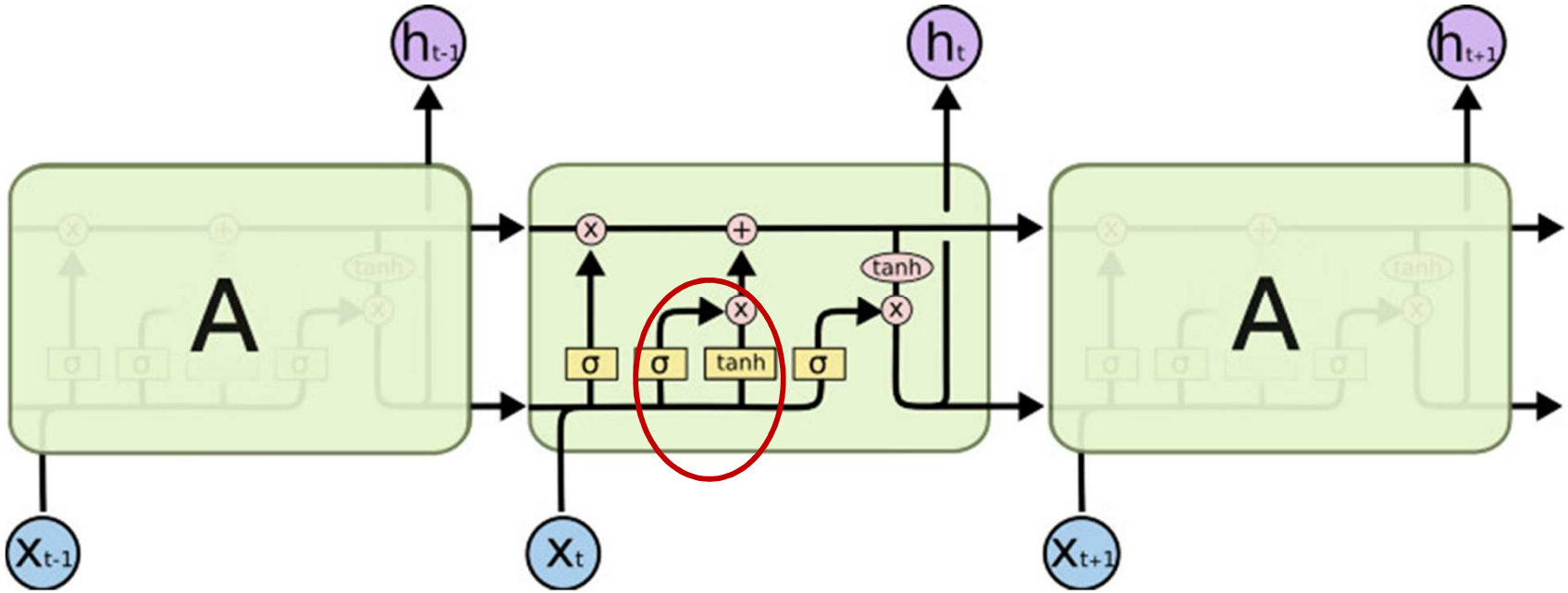
- Cổng quên – forget gate:

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



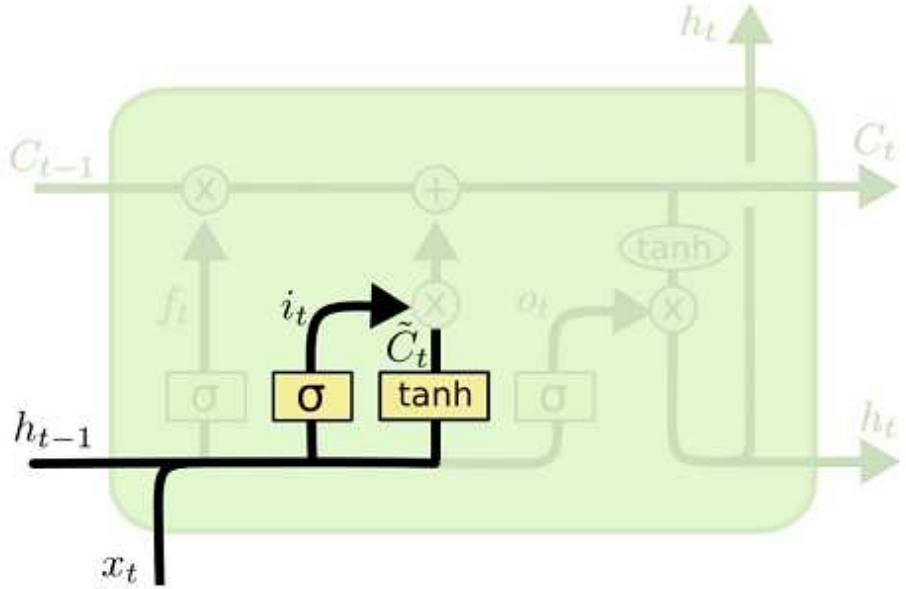
LSTM (Long Short Term Memory)

- Cổng đầu vào – input gate:
  - Quyết định những thông tin nào từ input đầu vào sẽ được sử dụng



LSTM (Long Short Term Memory)

- Cổng đầu vào – input gate:

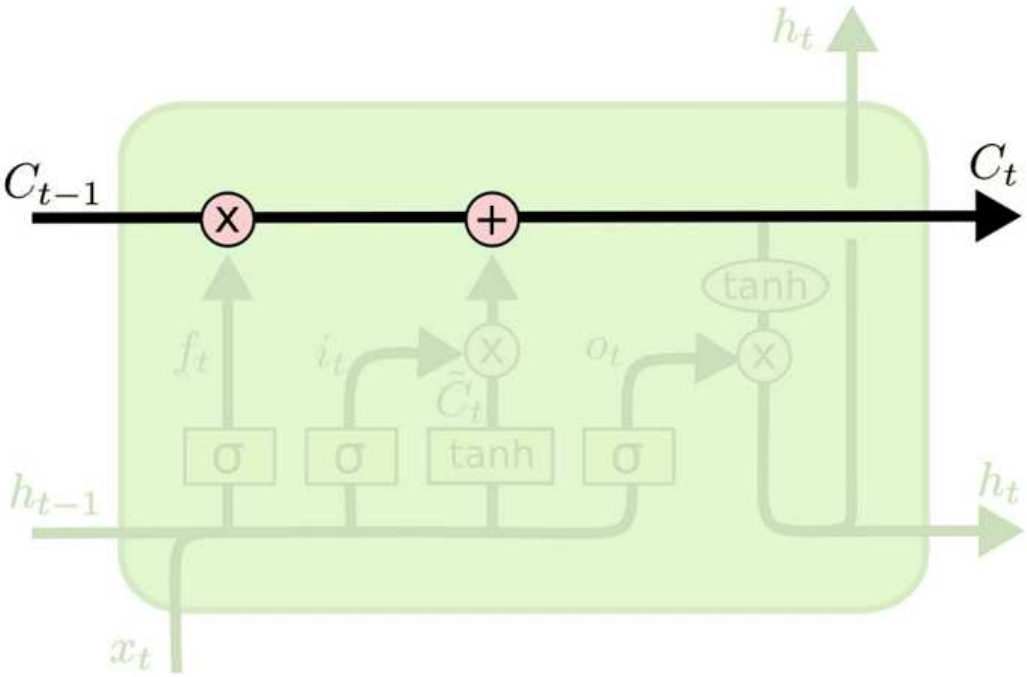


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

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

LSTM (Long Short Term Memory)

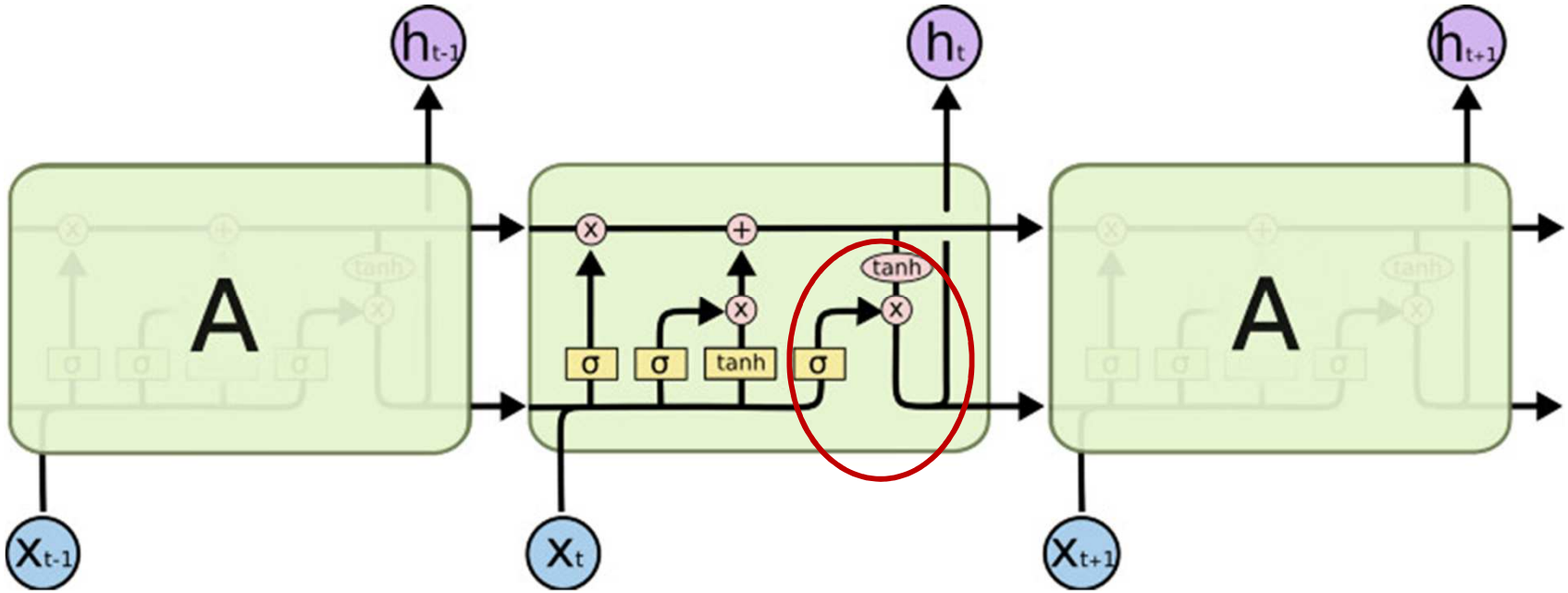
- Cập nhật trạng thái



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

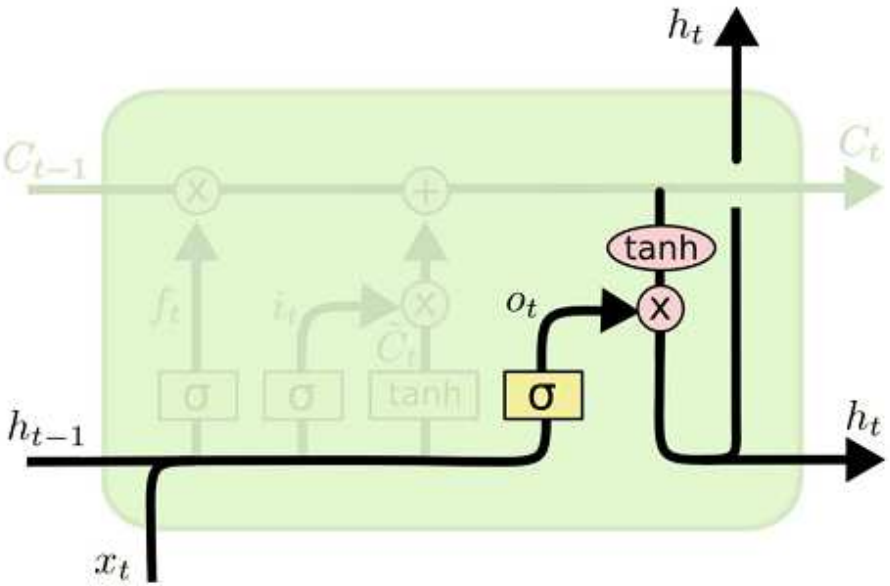
LSTM (Long Short Term Memory)

- Cổng đầu ra– output gate:
  - Quyết định đầu ra ở bước hiện tại



LSTM (Long Short Term Memory)

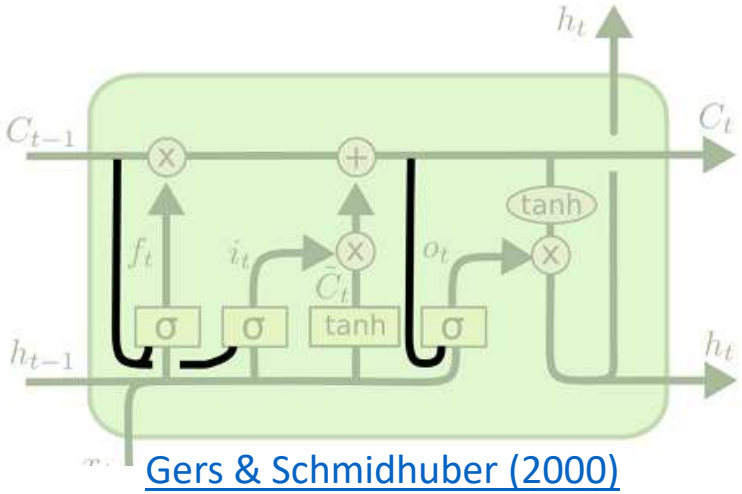
- Cổng đầu ra – output gate:



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

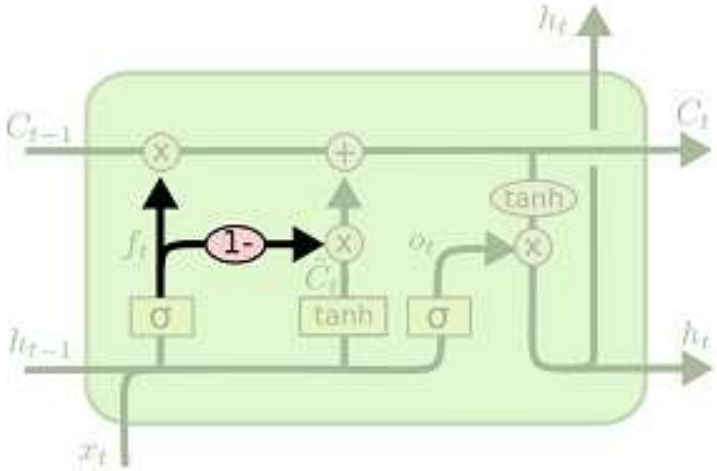
$$h_t = o_t * \tanh\left(C_t\right)$$

Các biến thể của LSTM



Gers & Schmidhuber (2000)

$$f_t = \sigma\left(W_f \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [\boldsymbol{C_t}, h_{t-1}, x_t] \; + \; b_o\right)$$



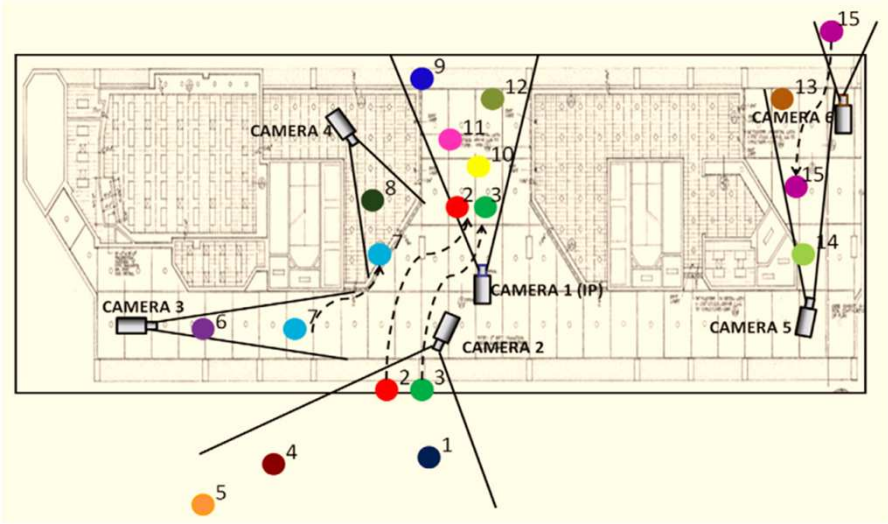
$$C_t = f_t * C_{t-1} + (1 - \boldsymbol{f_t}) * \tilde{C}_t$$

Nội dung buổi học

- Dữ liệu chuỗi
- RNN (Recurrent Neural Network)
- LSTM (Long Short Term Memory)
- Định danh lại người trong mạng camera

Định danh lại

- Định danh lại (tái định danh): nhằm kết nối các thể hiện của cùng 1 người khi người này di chuyển trong 1 mạng camera
- Định danh lại <> theo vết người



[1] Bedagkar-Gala, Apurva, and Shishir K. Shah. "A survey of approaches and trends in person re-identification." Image and Vision Computing 32.4 (2014): 270-286.

Định danh lại

Định danh lại

❖ Ứng dụng của định danh lại

Airports, railway-stations

Supermarkets

Surveillance

Human-robot interaction

Hospitals

Định danh lại

❖ Khó khăn và thách thức

- Chỉ dựa trên hình dáng, trang phục của đối tượng → thay đổi do thay đổi ánh sáng, góc nhìn, hình trạng
- Bài toán đối sánh nhiều-nhiều
- Bài toán định danh lại là bước tiếp theo của phát hiện và theo vết đối tượng→ phụ thuộc vào chất lượng của phát hiện và theo vết đối tượng

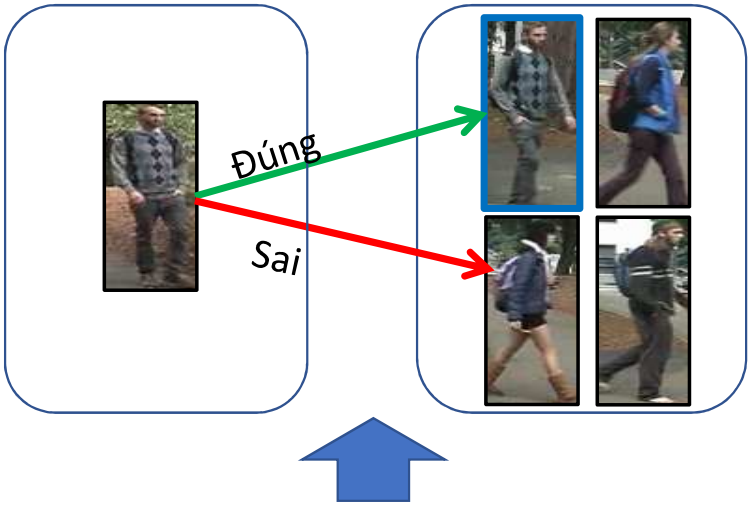
Thay đổi điều kiện chiếu sáng

Thay đổi hướng nhìn

Định danh lại

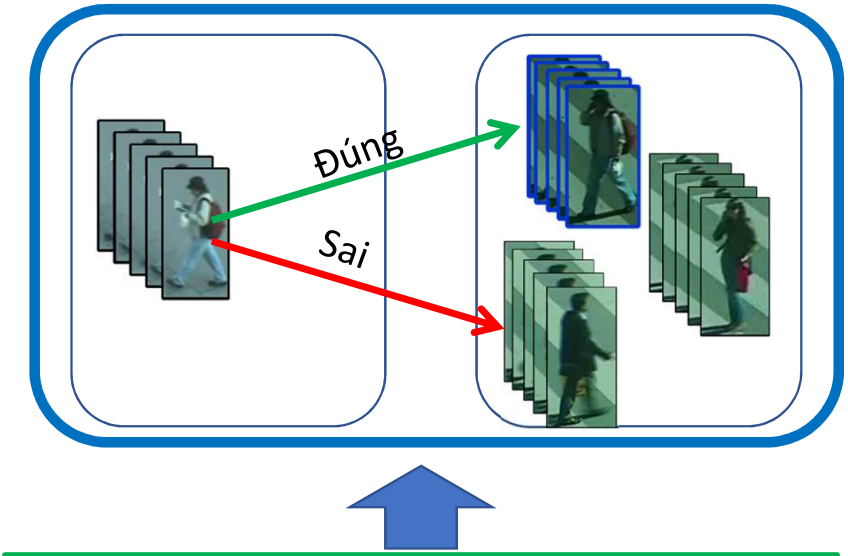
❖ Các hướng tiếp cận

- **Chỉ sử dụng 1 thể hiện/ ảnh (single shot)**



+ Đặc trưng mức ảnh
+ Tính toán sự tương tự giữa 2 ảnh

- **Sử dụng nhiều ảnh/thể hiện (multiple shot)**



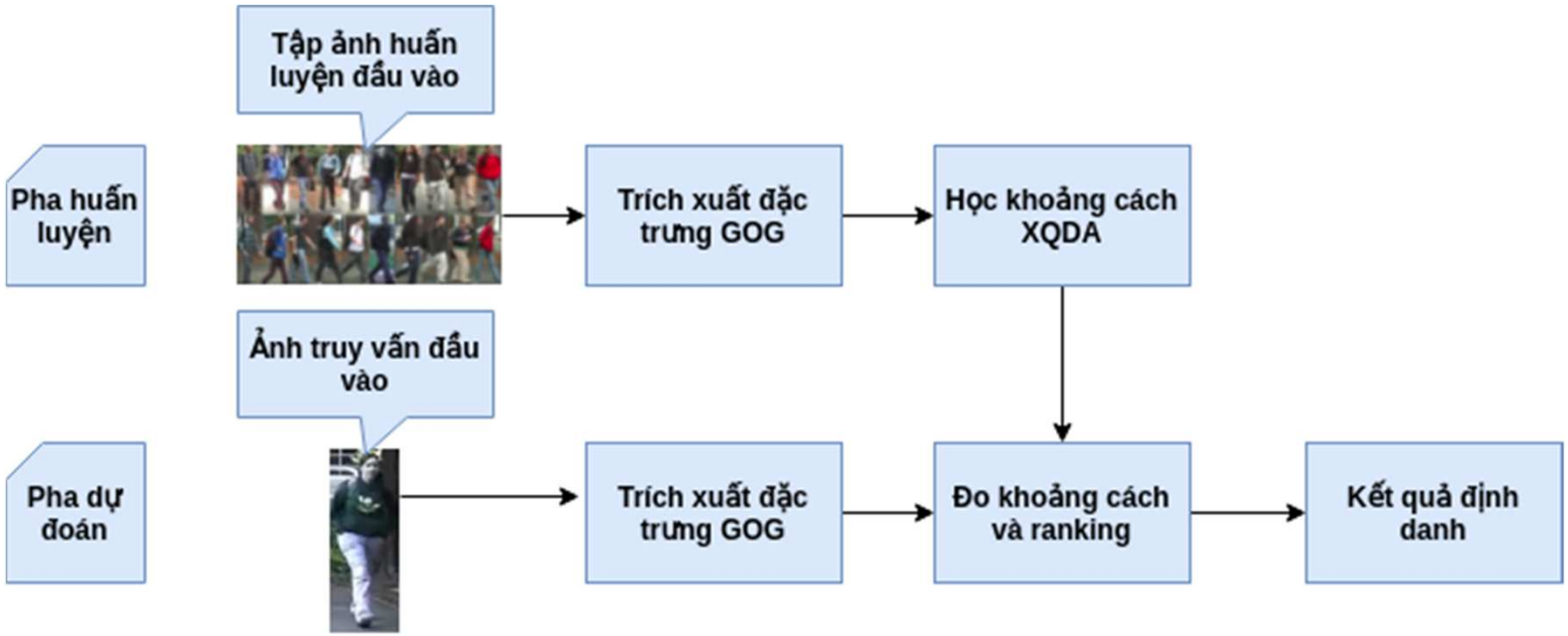
+ Đặc trưng theo không gian và thời gian
+ Chuyển từ mức ảnh sang mức tập ảnh hay mức chuỗi ảnh
+ So sánh các đặc trưng

Định danh lại 1 thể hiện

❖ Các đặc trưng sử dụng

- **Các đặc trưng:**
  - Đặc trưng thiết kế: **GOG descriptor** [Matsukawa, 2016], Saliency signature [Zhao, 2013], Salient color name [Yang, 2014], LOMO [Liao, 2015],…
  - Đặc trưng dựa trên học sâu: FNN [Wu, 2016]; Multi-channel parts-based CNN [Cheng, 2016]; …
- **Học độ đo (Metric learning):** học các độ đo để khoảng cách của các thể hiện cùng 1 người bé hơn khoảng cách của các thể hiện ở các người khác nhau
  - KISSME (Keep it simple and straightforward) [Koestinge, 2012]
  - LDA (Linear Discriminant Analysis) [T. Hastie, 2009]
  - LMNN(Large Margin Nearest Neighbor) [K. Weinberger, 2006]
  - **XQDA [Liao, 2015]**

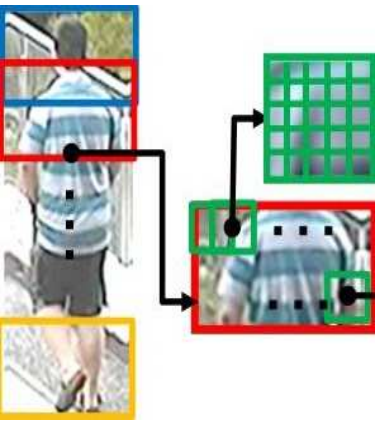
Định danh lại 1 thể hiện

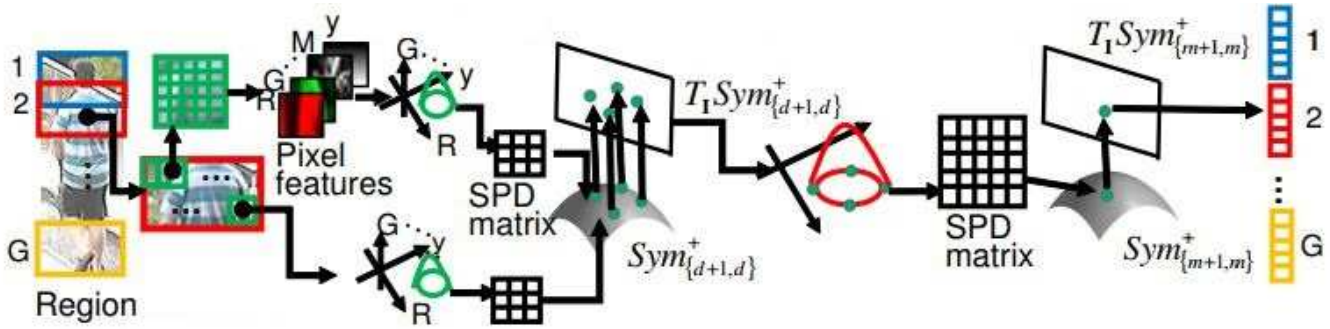


TB Nguyen, DL Tran, TL Le, TTT Pham, HG Doan, An effective implementation of Gaussian of Gaussian descriptor for person re-identification, NICS 2018

Định danh lại 1 thể hiện

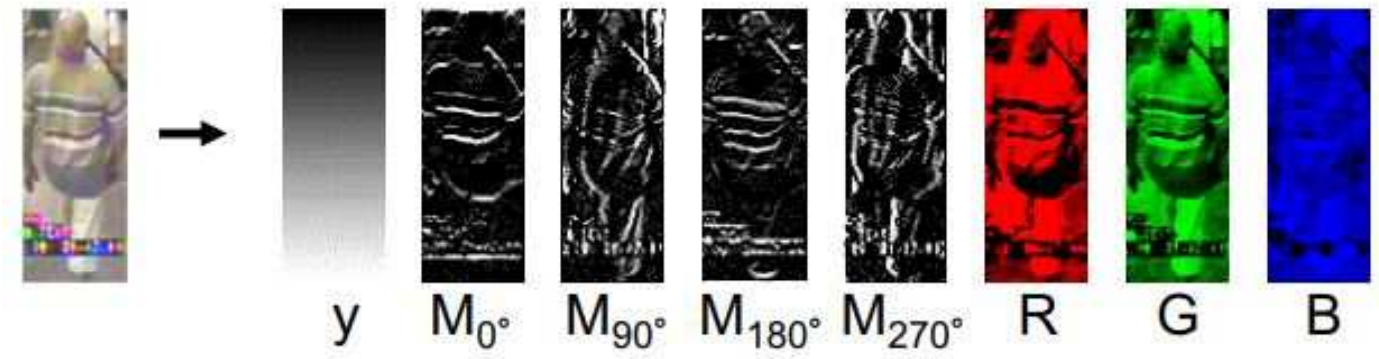
- **Đặc trưng Gaussian of Gaussian (GOG)**
  - Trích đặc trưng từng phần cơ thể bằng cách biểu diễn các đặc trưng cục bộ.
  - Tổng hợp lại đặc trưng với trọng số tăng dần ở gần trục cơ thể.
- **Biểu diễn qua hai phân cấp**
  - Mảnh ghép (patch): Mỗi patch là một cửa sổ vuông có kích thước nhỏ
  - Khu vực (region): Mỗi region đại diện cho một phần cơ thể người





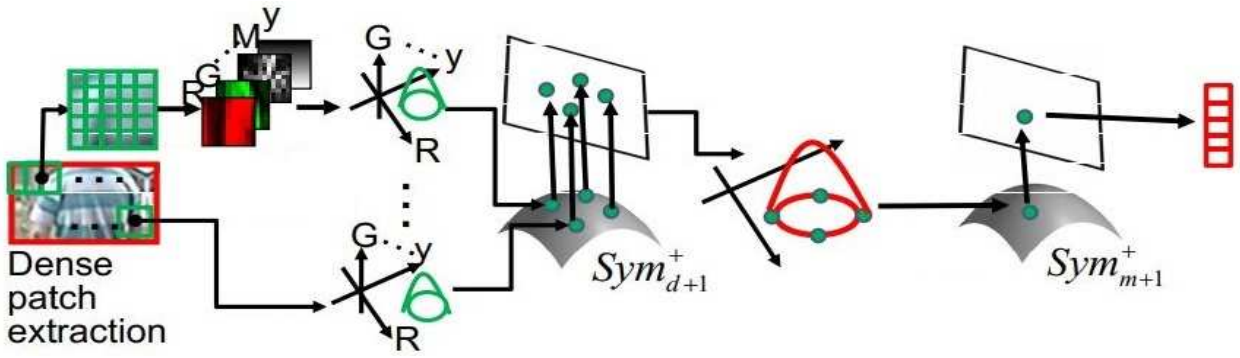
Định danh lại 1 thể hiện

- **Đặc trưng pixel**: Kết hợp nhiều đặc trưng mức thấp thành một vector đặc trưng với mỗi điểm ảnh như tọa độ, biên ảnh, màu sắc, độ sáng, …
- Vì số lượng điểm ảnh trên patch thường nhỏ, chọn kích thước vector đặc trưng pixel nhỏ để không lấn át đặc trưng mảnh ghép.



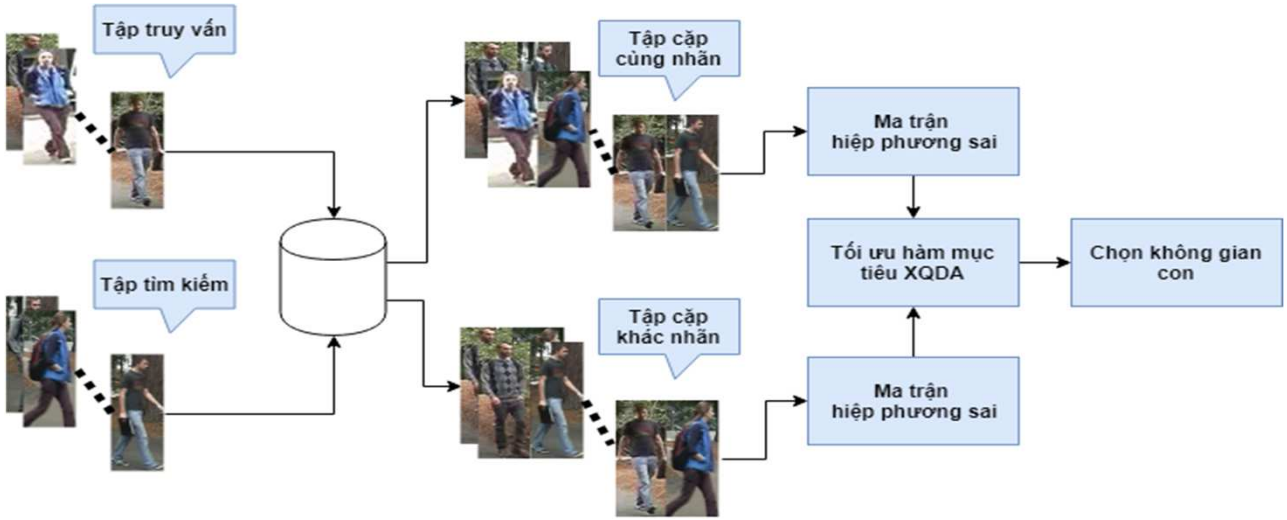
Định danh lại 1 thể hiện

- **Tư tưởng chính của GOG**
  - Mỗi region như một phân phối Gaussian của các patch.
  - Mỗi patch là một phân phối Gaussian của các pixel.
- **Biến đổi phân phối Gaussian**
  - Nhúng qua không gian ma trận xác định dương
  - Ánh xạ qua không gian tiếp tuyến và vector hóa



Định danh lại 1 thể hiện

- XQDA là phương pháp phổ biến để đánh giá độ tương tự của các cặp/chuỗi ảnh trong bài toán định danh lại.
- Mục tiêu của XQDA là học một không gian con với số chiều nhỏ hơn số chiều của đặc trưng, từ đó xây dựng được một ma trận khoảng cách.



Định danh lại 1 thể hiện

Cơ sở dữ liệu

- **VIPeR**

  + captured by 2 static cameras

  + 1264 images of 632 perssons

  + Resolution: 128 x 48

Cam_a

Cam_b

- **VIPeR-foreground**

  + Eliminate background by applying Interative Segmentation algorithm

Cam_a

Cam_b

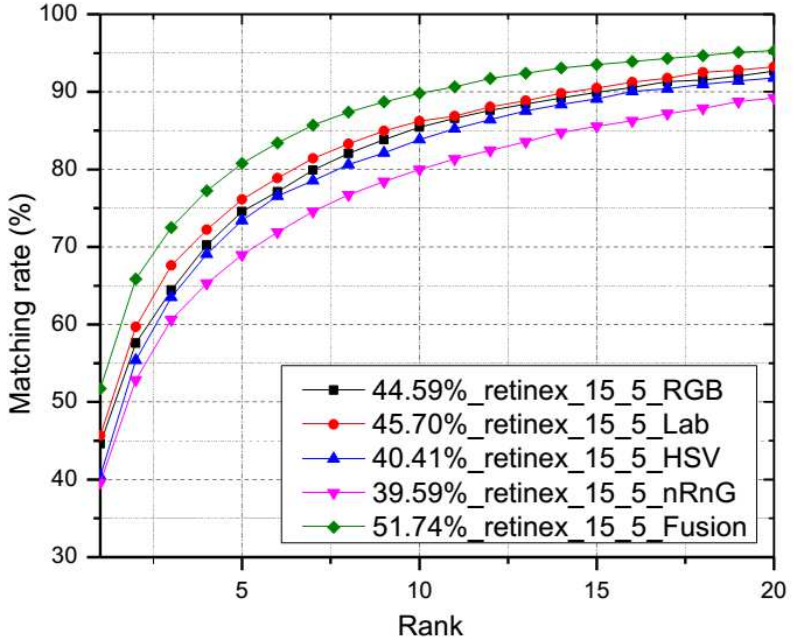
Định danh lại 1 thể hiện

Evaluation metric

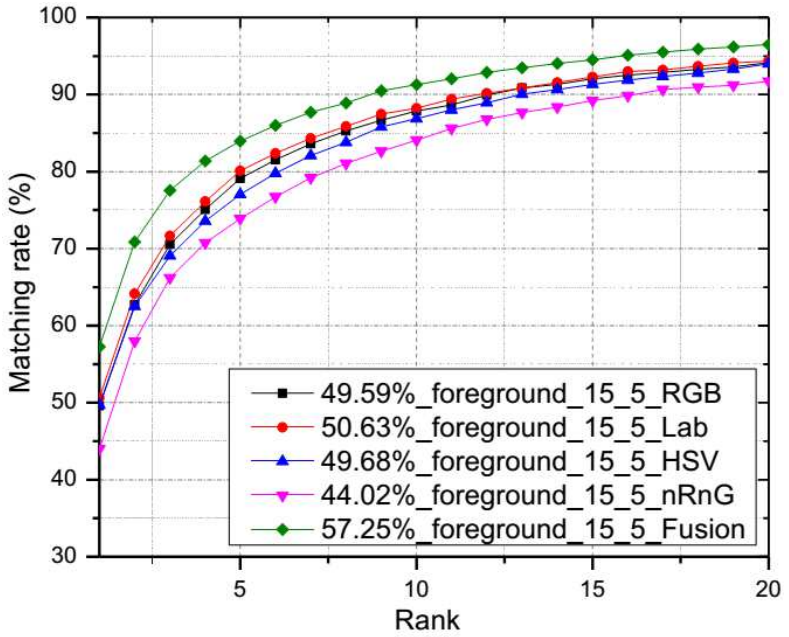
- Đường cong CMC (Cumulative Matching Characteristic): Giá trị thứ k trên đường cong tính bằng xác suất nhìn thấy mẫu đúng (cùng một người) trong k thứ hạng đầu tiên.
- Một thuật toán được đánh giá là hiệu quả hơn khi đường cong của thuật toán đó cao hơn đường cong của các thuật toán khác.

Định danh lại 1 thể hiện

Kết quả thử nghiệm



The obtained results when applying Retinex algorithm on VIPeR dataset.



Obtained results on VIPeR-Foreground dataset.

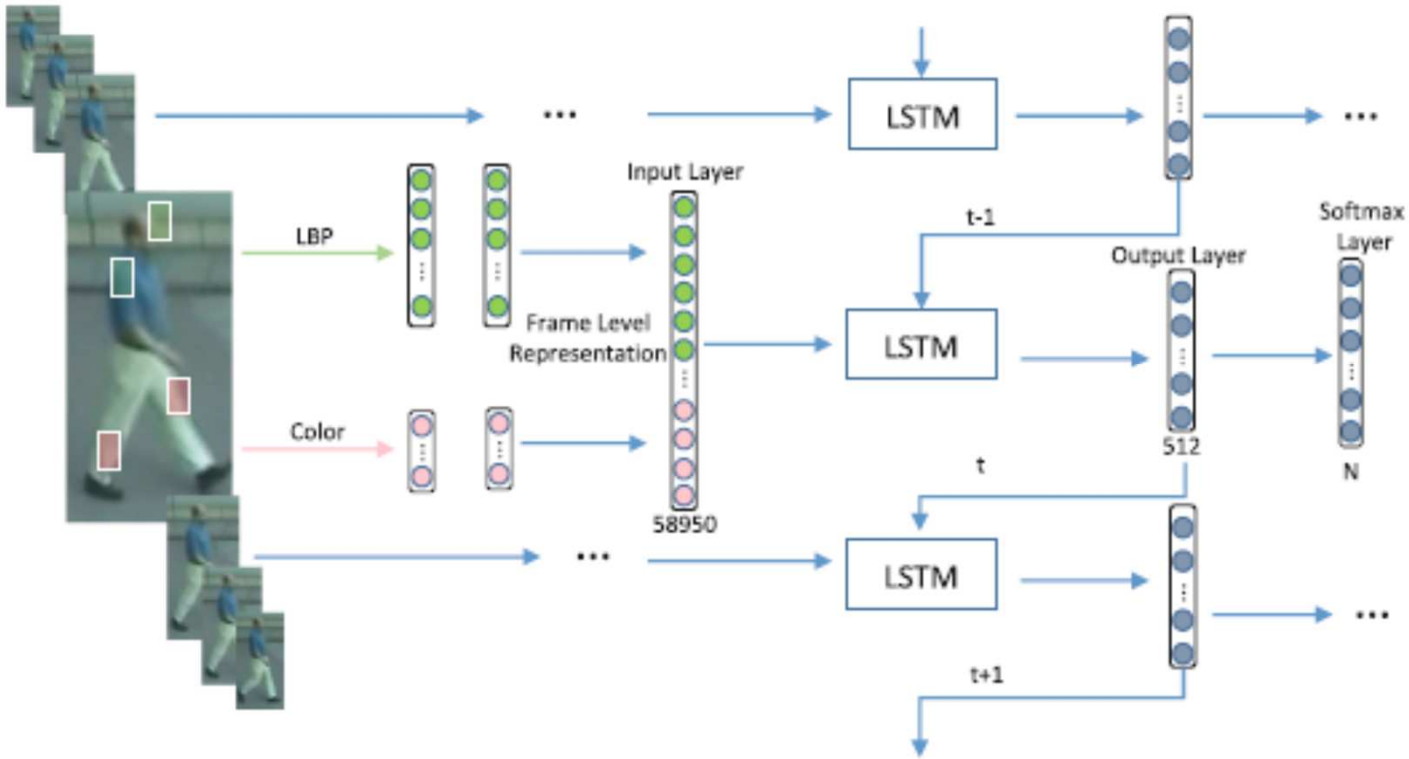
Fusion: + 30.07% compared to obtained result in [19] (27.18%)

Định danh lại 1 thể hiện

So sánh với các phương pháp trước đó

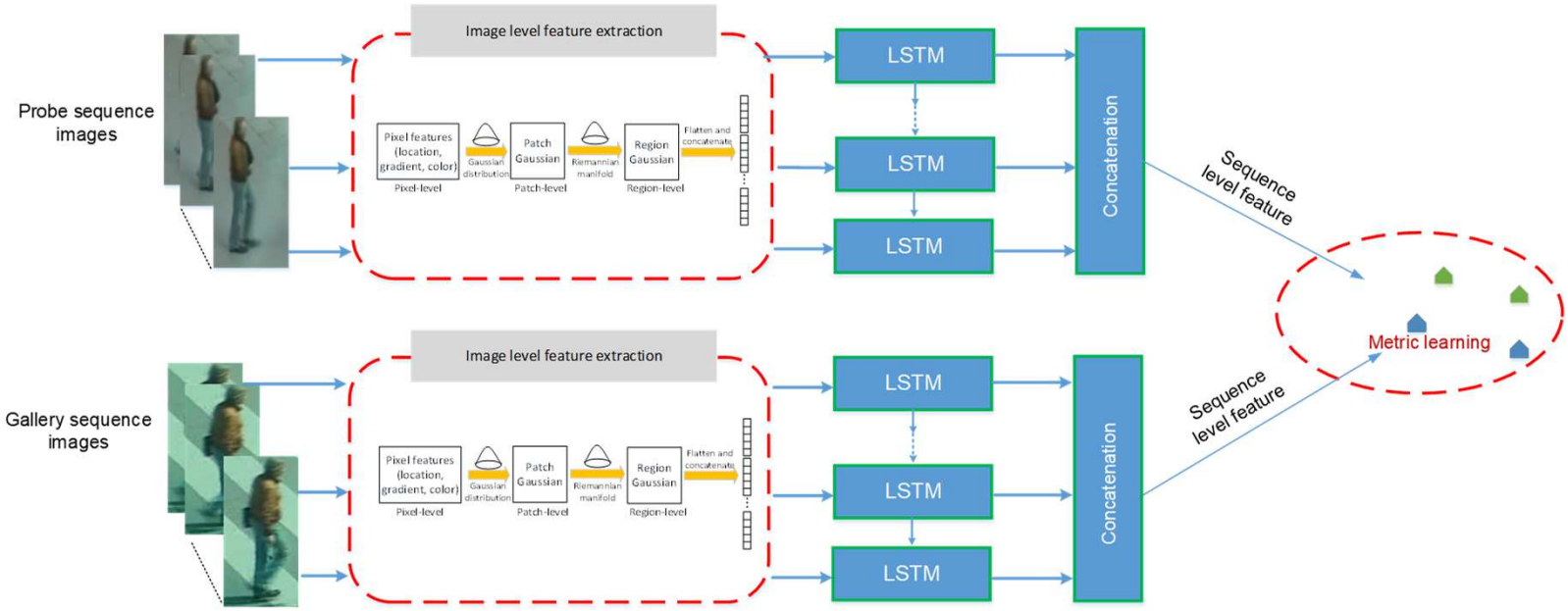
Methods		Rank1	Rank 5	Rank 10	Rank 20
Hand-designed	IRWPS [3]	23.2	45.3	58.3	68.7
	SalMatch [4]	30.2	52.0	65.0	-
	SCNCD [5]	37.8	68.5	81.2	90.4
	LOMO+XQDA [6]	40.0	-	80.5	91.1
Deep-learning	FNN [7]	51.1	81.0	**91.4**	**96.9**
	MCCNN [8]	47.8	74.7	84.8	91.1
	RDC [9]	40.5	60.8	70.4	84.4
Proposed method	$GOG_{OptimalParameters}$	51.0	**81.1**	89.1	94.6
	$GOG_{Retinex}$	**51.7**	80.8	89.8	95.3
	$GOG_{Foreground}$	**57.2**	**84.0**	**91.3**	**96.5**

Định danh lại nhiều thể hiện



Yan et al. Person Re-identification via Recurrent Feature Aggregation, ECCV 2016

Định danh lại nhiều thể hiện

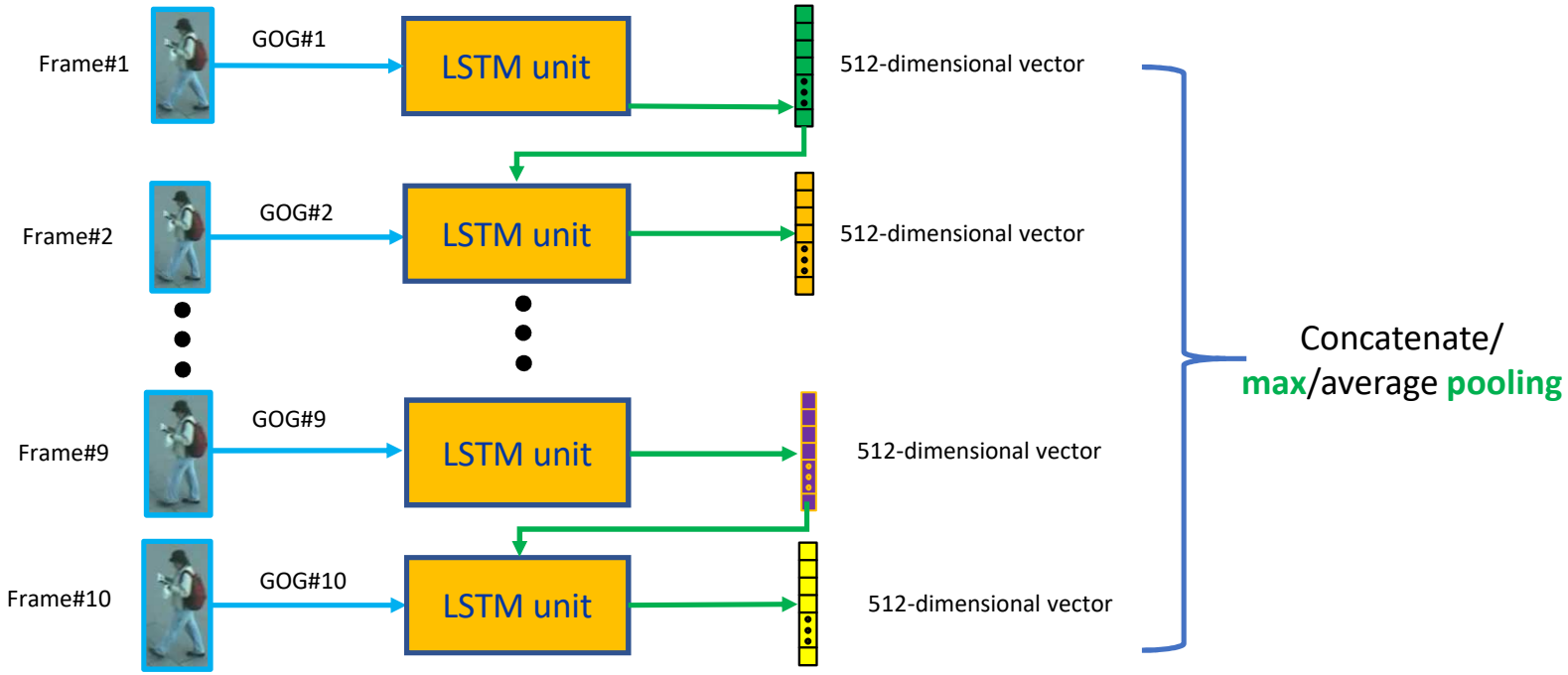


- **Contributions**:
  - Gaussian of Gaussian (GOG) as image-level feature.
  - Apply Cross-View Quadratic Discriminant Analysis (XQDA) for metric learning

QN Hong, TB Nguyen, TL Le, **Enhancing Person Re-Identification Based on Recurrent Feature Aggregation Network**, Multimedia Analysis and Pattern Recognition (MAPR) 2018

Định danh lại nhiều thể hiện

Trích chọn đặc trưng mức chuỗi bằng LSTM



Định danh lại nhiều thể hiện

Cơ sở dữ liệu

- **PRID2011**

  + captured by 2 static cameras

  + 385 persons in camera view A and 749 persons in camera view B

  + Only 200 persons appear on both cameras

- **iLIDS-VID**

  + captured by 2-indoor and 2-outdoor cameras

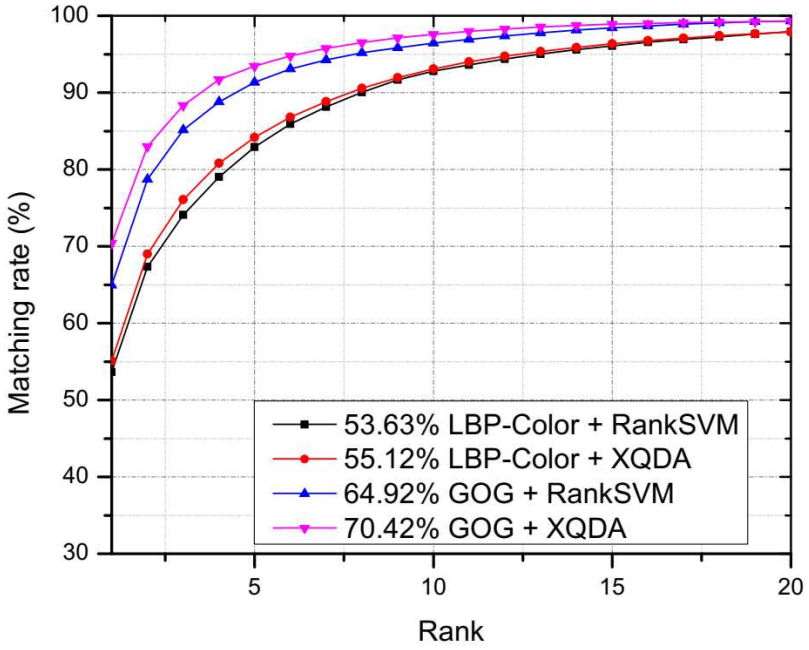
  + 600 image sequences of 300 people

  + **Challenges:** clothing similarities, background complexity, and occlussion.

Định danh lại nhiều thể hiện

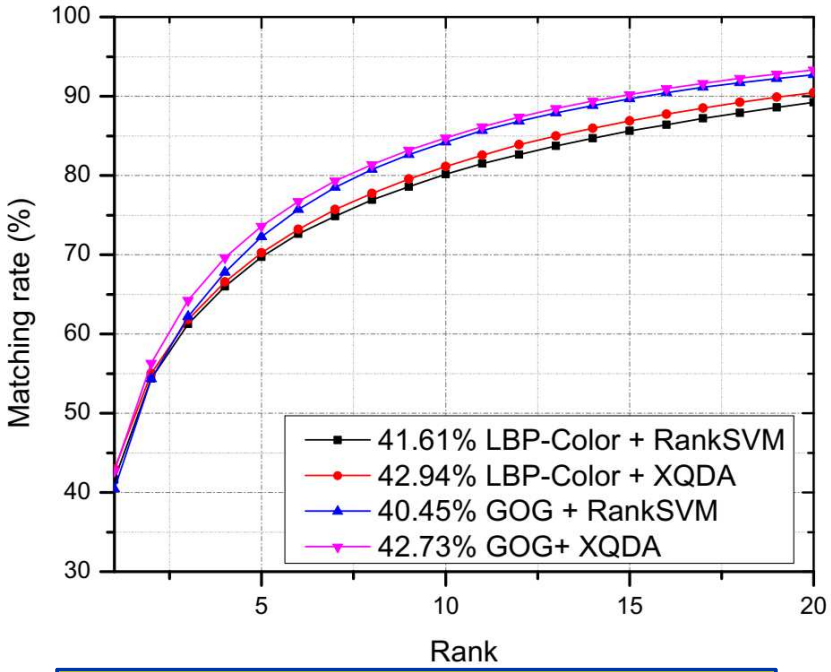
So sánh với phương pháp trước đó [Yan,2016]

- **PRID2011**



GOG vs LBP-Color : **+11.29%** (Rank SVM)
+ 15.3% (XQDA)
XQDA vs RankSVM: **+ 5.5%** (GOG)
+1.49% (LBP-Color)

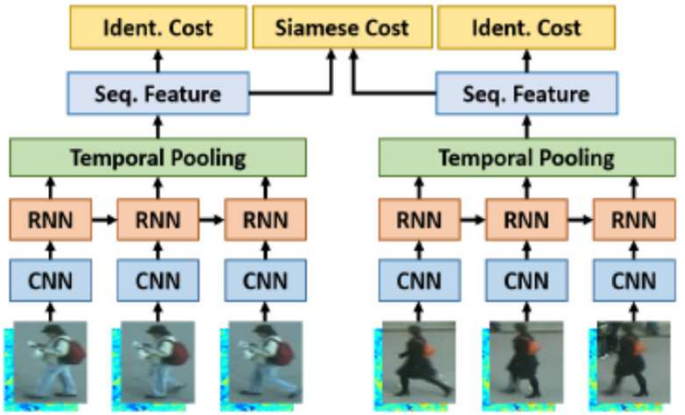
- **iLIDS-VID**



Ours vs the original results: **+1.1%**

Định danh lại nhiều thể hiện

Method	PRID 2011				iLIDS-VID			
	Rank=1	Rank=5	Rank=10	Rank=20	Rank=1	Rank=5	Rank=10	Rank=20
HOG3D + DVR [2]	28.9	55.3	65.5	82.8	23.3	42.2	55.3	68.4
STFV3D + KISSME [3]	64.1	87.3	89.9	92.0	44.3	71.7	83.7	91.7
TAPR [4]	68.6	**94.6**	**97.4**	**98.9**	55.0	**87.5**	**93.8**	**97.2**
FAST3D [5]	31.2	60.3	76.4	88.6	28.4	54.7	66.7	78.1
RNN [6]	**70.0**	90.0	95.0	97.0	**58.0**	84.0	91.0	96.0
DFCP [8]	51.6	83.1	91.0	95.5	34.5	63.3	74.5	84.4
TDL [10]	56.7	80	87.6	93.6	**56.3**	**87.6**	**95.6**	**98.3**
RFA-Net [7]	53.6	82.9	92.8	97.9	41.6	69.7	80.2	89.2
Ours	**70.4**	**93.4**	**97.6**	**99.3**	42.7	73.6	84.7	93.3



- [6] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification, CVPR 2016

Các tài liệu tham khảo

- Slides của GS. **Ming Li,** CS 898: Deep Learning and Its Applications, https://cs.uwaterloo.ca/~mli/cs898-2017.html
- Hong-Quan Nguyen, Thuy-Binh Nguyen, Thi-Lan Le, **Enhancing Person Re-Identification Based on Recurrent Feature Aggregation Network,** 2018 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR)**.**
- http://karpathy.github.io/2015/05/21/rnn-effectiveness/
- http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Lời cảm ơn

- Cảm ơn nhóm nghiên cứu về Định danh lại tại Viện nghiên cứu quốc tế MICA (NCS Nguyễn Hồng Quân, NCS Nguyễn Thúy Bình) đã cung cấp dữ liệu và chương trình phục vụ cho buổi học

THỊ GIÁC MÁY TÍNH

AI Academy Vietnam

CẢM ƠN!