HIỂU BIẾT MÔI TRƯỜNG XUNG QUANH DÙNG MẠNG NƠRON TÍCH CHẬP

SCENE UNDERSTANDING USING CONVOLUTIONAL NEURAL NETWORK

Huỳnh Thu Thảo[1], Lê Mỹ Hà[1]

[1] Trường đại học Sư phạm Kỹ thuật TP.HCM

TÓM TẮT

Trong bài báo này, tác giả sẽ sử dụng công nghệ học sâu, cụ thể ở đây là sử dụng các kỹ thuật trong mạng nơron tích chập để nhận dạng đối tượng, đa đối tượng và vị trí của đối tượng trong ảnh. Các kỹ thuật CNN, RCNN, Fast RCNN, Faster RCNN trong mạng nơ ron tích chập sẽ được giới thiệu và mô phỏng, trích dẫn kết quả so sánh để đánh giá hiệu quả trong việc nhận dạng đối tượng trong ảnh. Faster RCNN được đánh giá là thích hợp hơn nhất để nhận dạng và xác định vị trí của đối tượng và đa đối tượng trong ảnh.

Từ khóa: *Học Sâu; Mạng Nơ Ron Tích Chập; Nhận Dạng Đơn Đối Tượng Trong Ảnh; Xác Định Vị Trí Của Đối Tượng Trong Ảnh; Nhận Dạng Đa Đối Tượng Trong Ảnh.*

ABSTRACT

In this paper, the author writes about Deep Learning Algorithm. A class of Deep Learning Network, Convolution Neural Network applied to detecting and locating the single object or multi object in input image. The methods such as RCNN, Fast RCNN and Faster RCNN with Convolution Neural Network will be introduct and get result from simulation and other paper to compare and evaluate. Faster RCNN is the best to choose for detection and location the single object or multi objects in input image.

Keywords: *Deep Learning; Convolution Neural Network; Single Object Detection; Object Localization; Multi Objects Detection.*

1. GIỚI THIỆU

Trí tuệ nhân tạo (Artificial Intelligence - AI) giờ xuất hiện ở khắp mọi nơi. Nó là thứ được sử dụng để trả lời email tự động trên Gmail, học cách lái xe cho chúng ta ngồi chơi, sắp xếp lại ảnh của những chuyến đi chơi thành từng album riêng biệt, thậm chí còn giúp quản lý ngôi nhà hay đi mua sắm. Trí tuệ nhân tạo có thể được hiểu đơn giản là được cấu thành từ các lớp xếp chồng lên nhau, trong đó mạng thần kinh nhân tạo nằm ở dưới đáy, Machine Learning nằm ở tầng tiếp theo và Deep Learning nằm ở tầng trên cùng.

Năm 2011, Google khởi tạo dự án Google Brain với mục đích tạo ra một mạng thần kinh được huấn luyện bởi các thuật toán Deep Learning. Dự án này sau đó đã chứng minh được khả năng tiếp nhận được cả những khái niệm bậc cao của Deep Learning. Facebook cũng thành lập AI Research Unit, đơn vị nghiên cứu về AI sử dụng Deep Learning vào việc tạo ra các giải pháp hiệu quả hơn giúp nhận diện khuôn mặt và sự vật trên 350 triệu bức ảnh và video được đăng tải lên Facebook mỗi ngày. Một ví dụ tiêu biểu khác về Deep Learning trong thực tế là khả năng nhận diện giọng nói của các trợ lý ảo Google Now và Siri.

Deep Learning đang ngày càng cho thấy một tương lai đầy hứa hẹn với ứng dụng vào điều khiển xe tự lái hay robot quản gia. Mặc dù các sản phẩm này vẫn còn nhiều hạn chế nhưng những thứ chúng làm được hiện nay thực sự rất khó tưởng tượng nổi chỉ vài năm trước đây; tốc độ nâng cấp cũng cao chưa từng thấy. Khả năng phân tích dữ liệu lớn và sử dụng Deep Learning vào các hệ thống máy tính có thể tự thích nghi với những gì chúng tiếp nhận mà không cần đến bàn tay lập trình của con người sẽ nhanh chóng mở đường cho nhiều đột phá trong tương lai. Những đột phá này có thể là việc thiết kế ra những trợ lý ảo, các hệ thống xe tự lái hay sử dụng vào thiết kế đồ họa, sáng tác nhạc, cho đến phát triển các nguyên liệu mới giúp robot thấu hiểu thế giới

xung quanh hơn. Chính vì tính thương mại cao mà các công ty lớn, đặc biệt là Google, luôn ưu tiên các startup về robot và Deep Learning trong danh sách thâu tóm của mình.

Deep Learning nói riêng hay trí tuệ nhân tạo nói chung thực sự có rất nhiều ứng dụng tuyệt vời, nhưng chúng ta hiện mới chỉ đang ở giai đoạn đầu phát triển nó nên những hạn chế là không thể tránh khỏi. Có lẽ còn phải chờ khá lâu nữa những hệ thống AI "có tri giác" mới thực sự xuất hiện, nhưng những gì các công ty lớn như Google, Facebook, IBM đang làm hiện nay cũng tương tự với việc đặt những viên gạch đầu tiên mở đường cho kỷ nguyên AI trong những thập kỷ tới.

CNN là một trong những thuật toán Deep Learning cho kết quả tốt nhất hiện nay trong hầu hết các bài toán về thị giác máy như phân lớp, nhận dạng. Kể từ khi thành công của LeNet của Yann LeCun cùng các cộng sự [3] công bố năm 1998 và ImageNet do Alex Krizhevsky cùng cộng sự công bố vào năm 2012 [4], mạng nơ ron tích chập (CNNs) đã trở thành tiêu chuẩn vàng để phân loại hình ảnh. Từ thời điểm đó đến nay, độ sai lệch của CNN đã được cải thiện đến mức xấp xỉ bằng con người.

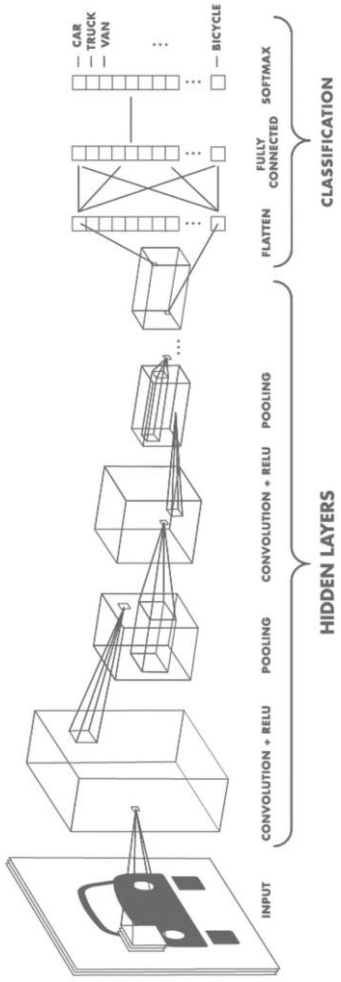
Nếu như chỉ cần xác định một đối tượng đơn trong ảnh thì ta đơn giản chỉ cần sử dụng một mạng CNN đơn giản. Nhưng bài toán đặt ra là khi ảnh có nhiều đối tượng, vật thể trong cùng một ảnh, bài toán trở nên rất phức tạp, ta phải tìm vị trí của vật thể trong ảnh rồi mới tiến hành phân loại. Vị trí của các đối tượng có thể chồng chéo nhau ở các bối cảnh khác nhau, ngoài tìm hình ảnh đòi hỏi chúng ta phải xác định ranh giới, sự khác biệt và mối quan hệ với nhau.

Phần này chúng ta sẽ tìm hiểu qua các kỹ thuật chính được sử dụng trong phát hiện phân loại và xác định vị trí đối tượng. Cụ thể sẽ lần lượt giới thiệu các kỹ thuật ban đầu của mạng CNN và các kỹ thuật phát triển sau đó như RCNN, Fast RCNN, Faster RCNN. Dưới đây là các kỹ thuật được sử dụng để xác định vị trí và đối tượng trong mạng CNN:

– Xác định vị trí đối tượng đơn dùng mạng CNN đơn giản.
– Kỹ thuật R-CNN.
– Kỹ thuật Fast R-CNN.
– Kỹ thuật Faster R-CNN.

2. CÁC KỸ THUẬT SỬ DỤNG ĐỂ XÁC ĐỊNH VỊ TRÍ VÀ ĐỐI TƯỢNG TRONG MẠNG CNN

2.1 Mạng nơ ron tích chập



Hình 1. *Kiến trúc mạng CNN cơ bản.*

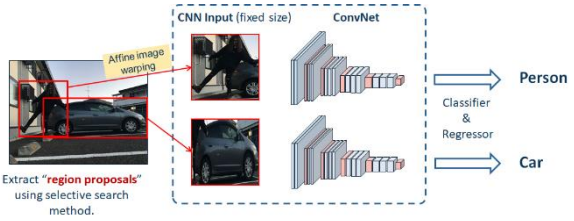
CNN có kiến trúc được hình thành từ các thành phần cơ bản bao gồm Convolution (CONV), Pooling (POOL), ReLU, Fully-connected (FC) về mặt xây dựng kiến trúc tổng quát CNN được mô tả như hình 1. CNN là thuật toán có kiến trúc bao gồm nhiều tầng có chức năng khác nhau trong đó tầng chính hoạt động thông qua cơ chế tích chập. Trong suốt quá trình huấn luyện, CNN sẽ tự động học được các thông số cho các bộ lọc – tương ứng là các đặc trưng theo từng cấp độ khác nhau. Ví dụ trong bài toán phân lớp ảnh, CNN sẽ cố gắng tìm ra các thông số tối ưu cho các

bộ lọc tương ứng theo thứ tự pixel > edges > shapes > facial > high-level features. Đây chính là lý do mà CNN có được kết quả vượt trội so với các thuật toán trước đây.

2.2 Xác định vị trí đối tượng đơn dùng mạng CNN đơn giản.

Ở đây, thuật toán hồi qui được sử dụng để tăng độ chính xác cho các bounding box để xác định vị trí của đối tượng, các chỉ số (x0, y0, chiều cao, chiều rộng) của bounding box được hồi qui. Ta huấn luyện mạng với ảnh đối tượng đã được xác định các chỉ số gốc chính xác về bounding box và so sánh chỉ số hồi qui với các chỉ số gốc để tính ra độ sai lệch của các bounding box. Thông thường chức năng xác định vị trí sẽ được thêm vào lớp fully connected của mạng tích chập.

2.3 Kỹ thuật R-CNN



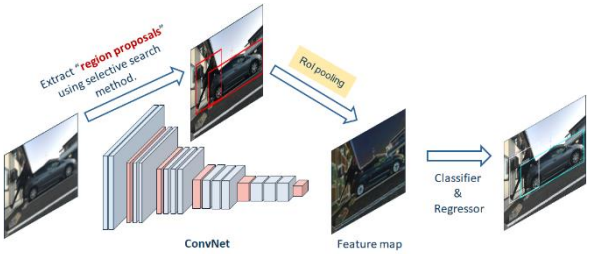
Hình 2. *Mạng CNN sử dụng kỹ thuật R-CNN.*

RCNN (Regions + CNN) là phương pháp dùng các thuật toán tìm kiếm có chọn lọc để đoán được các vị trí có thể có đối tượng trong ảnh, gọi là các vùng đề xuất. Các vùng đề xuất này sẽ được điều chỉnh lại kích cỡ bằng các thuật toán xử lý ảnh và cho vào các mạng CNN đã huấn luyện sẵn. Số lượng vùng đề xuất này có thể lên tới 2000 vùng trên một ảnh.

Sau khi được đưa vào đầu vào của mạng CNN đã huấn luyện sẵn để tính toán feed forward sẽ thu được các đặc tính chập của mỗi vùng đề xuất, sau đó tiếp tục huấn luyện SVM để xác định được vật thể nào được chứa trong vùng đề xuất đó. Các vùng đề xuất này sẽ được lưu vào bộ nhớ. Cuối cùng sẽ dùng các thuật toán hồi qui tuyến tính để hiệu chỉnh các giá trị của các vị trí của đỉnh của các vùng đề xuất.

Vì phải đưa một số lượng lớn các vùng đề xuất vào mạng CNN và xử lý một cách tuần tự nên tốc độ của RCNN là rất chậm.

2.4 Kỹ thuật Fast R-CNN

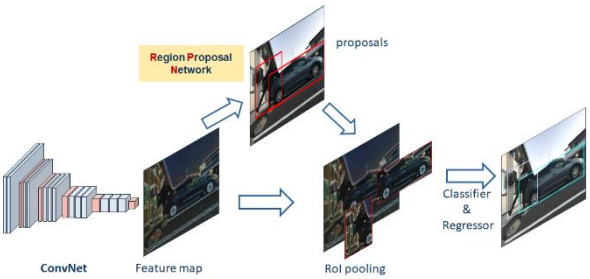


Hình 3. *Mạng CNN sử dụng kỹ thuật Fast R-CNN.*

Sử dụng các mạng huấn luyện sẵn để feed-forward các vùng dự đoán, sẽ tốn nhiều thời gian bởi với mỗi ảnh thuật toán tìm kiếm có chọn lọc sẽ cho ra hàng nghìn vùng dự đoán. Ta sẽ chỉ feed-forward một lần đối với ảnh gốc, thu được các đặc tính tích chập của ảnh đó. Dựa vào kích thước cùng vị trí của các vùng dự đoán đối với ảnh gốc, ta sẽ tính toán được vị trí của vùng dự đoán trong đặc tính tích chập. Sử dụng giá trị đặc tính tích chập của vùng dự đoán, ta dự đoán được vị trí các đỉnh của đường bao cũng như vật thể nằm trong đường bao là gì. Đối với Fast RCNN, do chia sẻ tính toán giữa các vùng trong ảnh, tốc độ thực thi của thuật toán đã được giảm từ 120s mỗi ảnh xuống 2s. Phần tính toán gây ra nghẽn chính là phần đưa ra các vùng dự đoán đầu vào, chỉ có thể thực thi tuần tự trên CPU.

Faster RCNN giải quyết vấn đề này bằng cách sử dụng Region Proposal Network để tính toán các vùng dự đoán có đối tượng này.

2.5 Kỹ thuật Faster R-CNN



Hình 4. *Mạng CNN sử dụng kỹ thuật Faster R-CNN.*

Điểm khác biệt chính giữa faster RCNN và fast RCNN là fast RCNN sử dụng các thuật toán tìm kiếm có chọn lọc để khởi tạo vùng đề xuất còn faster RCNN dùng mạng RPN để khởi tạo vùng đề xuất này.

Thời gian để khởi tạo các vùng đề xuất được rút ngắn lại rất nhiều khi sử dụng RPN so với thời gian sử dụng thuật toán tìm kiếm có chọn lọc.

RPN sử dụng một thuật toán sắp xếp các ô khu vực gọi là các anchor và kiểm tra các anchor này có khả năng chứa các đối tượng trong đó hay không.

Tùy theo đối tượng cần xác định trong mạng mà ta có thể tùy biến kích thước cũng như số lượng anchor để tăng tính hiệu quả của của kỹ thuật này.

RPN thường có hai bước chính:

Bước 1: Feed-forward ảnh qua mạng thu được các đặc tính tích chập.

Bước 2: Sử dụng các cửa sổ trượt lên các ảnh đã lấy được các đặc tính tích chập.

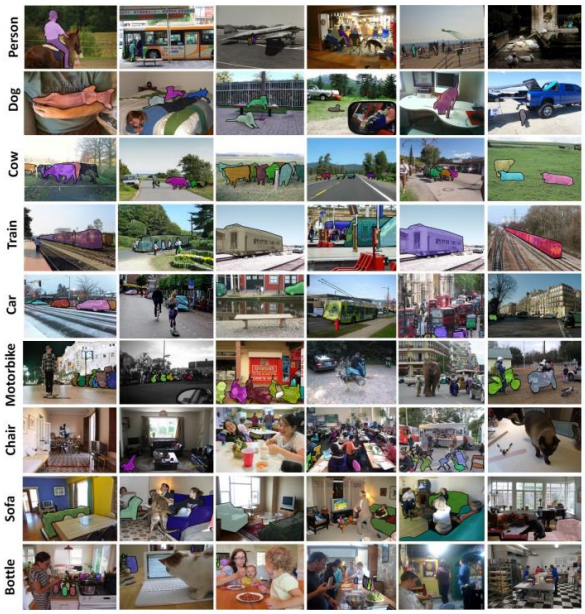
3. MÔ HÌNH MẠNG VÀ TẬP DỮ LIỆU

Mô hình mạng VGGNet được phát triển bởi Karen Simonyan và Andrew Zisserman [1]. VGGNet đã cho thấy hiệu suất của mạng sẽ phụ thuộc vào chiều sâu của mạng. Mô hình mạng tốt nhất của họ bao gồm 16 lớp CONV/FC và điều đặc biệt là nó đồng nhất về mặt kiến trúc với 3x3 convolutions và 2x2 pooling từ đầu đến cuối. Một điểm yếu của mô hình mạng này là sử dụng nhiều bộ nhớ và thông số, với hơn 138 triệu trọng số và phải cần đến 24Mb cho mỗi ảnh cần xử lý. Các thông số chi tiết được trình bày dưới bảng 1.

Bảng 1. *Thông số mạng CNN sử dụng để nhận diện đa đối tượng.*

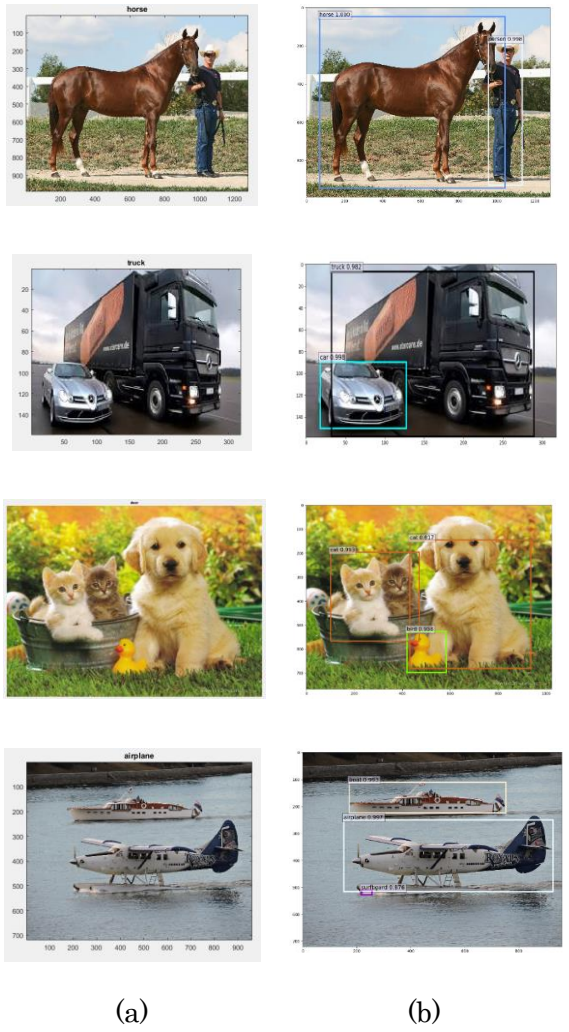
Lớp	Kích thước	Bộ nhớ	Trọng số
Input	224x224x3	224*224*3	0
Conv	224x224x64	224*224*64	1728
Conv	224x224x64	224*224*64	36864
Pool	112x112x64	112*112*64	0
Conv	112x112x128	112*112*128	73728
Conv	112x112x128	112*112*128	147456
Pool	56x56x128	56*56*128	0
Conv	56x56x256	56*56*256	294912
Conv	56x56x256	56*56*256	589824
Conv	56x56x256	56*56*256	589824
Pool	28x28x256	28*28*256	0
Conv	28x28x512	28*28*512	1179648
Conv	28x28x512	28*28*512	2359296
Conv	28x28x512	28*28*512	2359296
Pool	14x14x512	14*14*512	0
Conv	14x14x512	14*14*512	2359296
Conv	14x14x512	14*14*512	2359296
Conv	14x14x512	14*14*512	2359296
Pool	7x7x512	7*7*512	0
FC	1x1x4096	4096	102760448
FC	1x1x4096	4096	16777216
FC	1x1x1000	1000	4096000
Tổng số trọng số:			138,344,128

Ở phần này ta sẽ sử dụng tập dữ liệu MS COCO [7], các thuộc tính của tập dữ liệu này như sau: 80 đối tượng với 330000 ảnh. Tập dữ liệu được chia làm 3 phần với 83 nghìn ảnh để huấn luyện, 41 nghìn ảnh để kiểm tra chéo và 41 nghìn ảnh để đánh giá độ chính xác.



Hình 5. *Một số hình ảnh của tập dữ liệu MS COCO.*

4. KẾT QUẢ MÔ PHỎNG



(a) (b)

Hình 6. *Ảnh kết quả giữa mô hình CNN đơn thuần (a) và mô hình dùng kỹ thuật Faster RCNN (b).*

Từ 4 mẫu ảnh kết quả trên ta có thể nhận xét được ưu điểm vượt trội của mạng CNN sử dụng kỹ thuật Faster RCNN so với mạng CNN không sử dụng kỹ thuật Faster RCNN. Các ảnh được xử lý qua mạng CNN không sử dụng kỹ thuật Faster RCNN hầu như nhận dạng sai hình ảnh đầu vào chứa nhiều đối tượng.

Từ các kết quả mô phỏng với kỹ thuật CNN, Faster RCNN đã trình bày ở trên và dựa trên các nghiên cứu ở tài liệu tham khảo số [5], [6], [9], [10] đối với kỹ thuật RCNN và Fast RCNN, ta có thể xây dựng các bảng đánh giá theo các tiêu chí như khả năng nhận dạng, khả năng xác định vị trí và thời gian xử lý như sau:

Bảng 2. *Đánh giá khả năng nhận dạng đối tượng.*

Kỹ thuật	Khả năng nhận dạng	
	Ảnh chứa một đối tượng	Ảnh chứa nhiều đối tượng
CNN	✓	Không
R-CNN	✓	✓
Fast R-CNN	✓	✓
Faster R-CNN	✓	✓

Bảng 3. *Đánh giá khả năng xác định vị trí.*

Kỹ thuật	Xác định vị trí	
	Ảnh chứa một đối tượng	Ảnh chứa nhiều đối tượng
CNN	Không	Không
R-CNN	✓	✓
Fast R-CNN	✓	✓
Faster R-CNN	✓	✓

Bảng 4. *Đánh giá thời gian xử lý.*

Kỹ thuật	Thời gian xử lý	
	Ảnh chứa một đối tượng	Ảnh chứa nhiều đối tượng
CNN	~ 0,1	NA
R-CNN	~ 50	~ 50
Fast R-CNN	~ 2	~ 2
Faster R-CNN	~ 0,3	~ 0,3

5. KẾT LUẬN

Từ các kết quả ở phần 4, ta có thể đưa ra các kết luận như sau:

- Kỹ thuật CNN đã phần nào giải quyết được bài toán nhận dạng đối tượng trong ảnh. Tuy nhiên với kỹ thuật CNN chỉ dừng lại ở ảnh chứa một đối tượng với ảnh chứa đa đối tượng thì gần như không thực hiện được.

- Nhưng kỹ thuật Faster RCNN ra đời sau đó đã không những khắc phục được khuyết điểm này mà còn tăng được tốc độ xử lý và khả năng mở rộng đối tượng nhận dạng của mạng. Qua đó cũng đã cho thấy tiềm năng của kỹ thuật này trong ứng dụng hiểu biết môi trường xung quanh, ở đây là mạng CNN dùng kỹ thuật Faster RCNN hiểu được đối tượng trong ảnh là gì và biết được vị trí của đối tượng trong ảnh từ ảnh đầu vào.

TÀI LIỆU THAM KHẢO

[1]. Karen Simonyan, Andrew Zisserman, *Very Deep Convolutional Networks For Large-Scale Image Recognition*, 2015.

[2]. Y. LeCun and Y. Bengio, *Convolutional networks for images, speech, and time-series,* 1995.

[3]. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, *Imagenet classification with deep convolutional neural networks*, 2012.

[4]. Vedaldi, Andrea and Karel Lenc, *MatConvNet-convolutional neural networks for MATLAB,* 2014.

[5]. Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation Tech report (v5)*, 2014.

[6]. Ross Girshick, *Fast R-CNN*, In Microsoft Research, 2015.

[7]. Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollar, "*Microsoft COCO: Common Objects in Context*", 2015.

[8]. Ian Goodfellow and Yoshua Bengio, *Deep Learning*, 2016.

[9]. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, 2016.

[10]. Fei-Fei Li, Andrej Karpathy, Justin Johnson, *Spatial Localization and Detection*, 2016.

Tác giả chịu trách nhiệm bài viết:

Họ tên: Huỳnh Thu Thảo

Đơn vị: Trường Đại Học Sự Phạm Kỹ Thuật TP.HCM

Điện thoại: 0123 975 0909

Email: huynhthuthao90@gmail.com

BÀI BÁO KHOA HỌC

THỰC HIỆN CÔNG BỐ THEO QUY CHẾ ĐÀO TẠO THẠC SỸ

Bài báo khoa học của học viên

có xác nhận và đề xuất cho đăng của Giảng viên hướng dẫn

Thực hiện theo MTCL & KHTHMTCL Năm học 2018-2019 của Thư viện Trường Đại học Sư phạm Kỹ thuật Tp. Hồ Chí Minh.