Video Stream Analytics: Phát hiện khẩu trang và đối tượng

Nguyễn Ngọc Quí - 18520410[1], Trang Hoàng Nhựt - 18520123[1], Nguyễn Thị Phương - 18520135[1], and Lê Thị Minh Hiền - 18520049[1]

Đại học Công nghệ Thông tin - Đại học Quốc gia Thành phố Hồ Chí Minh
{18520410, 18520123, 18520135, 18520049}@gm.uit.edu.vn

Tóm tắt nội dung : Video stream analytics đã là một chủ đề nổi tiếng và đang không ngừng nâng tầm giá trị về nó. Bên cạnh đó, trong bối cảnh đại dịch hiện tại thì việc dùng khẩu trang khi tiếp xúc với cộng đồng là điều bắt buộc, nghiên cứu về việc nhận diện đeo mặt nạ được quan tâm nhiều hơn. Nắm rõ được điều này, nhóm muốn đặt cho mình một bài toán về việc phân tích nhận diện thực thể bên trong video kết hợp xác định người có đeo khẩu trang làm chủ đề cho bài báo cáo. Nhóm sẽ thực hiện sử dụng kết hợp 3 model là phân tích nhận diện vật thể, xác định khuôn mặt và xác định khuôn mặt có đeo khẩu trang nhằm phân tích giải quyết bài toán đặt ra. Sau thời gian nghiên cứu, nhóm đã hoàn thành việc Streaming video và nhận diện được hơn 80 vật thể với đối tượng "person" được nhóm đặc biệt chú trọng. Nhóm cũng xây dựng được thành công một ứng dụng HTML đơn giản để có thể hiển thị kết quả dưới dạng hình ảnh. Đây được xem là một kết quả khả quan với việc phân tích nhận diện khuôn mặt trong video thời gian thực.

Keywords: Spark Streaming · Video stream analytics · Nhận diện khẩu trang · Nhận diện khuôn mặt· Nhận diện vật thể

1 Giới thiệu

Với nhiều ứng dụng từ video stream analytics trong thời điểm ở quá khứ đến hiện tại – thì đây được xem như một cái nôi cho các vấn đề thiết thực của xã hội. Kết hợp những tính năng từ việc phân tích video vào đời sống ngày nay không còn quá xa lạ với chúng ta. Những ứng dụng như: phân tích video chứa nội dung không phù hợp, phân tích video có chứa âm thanh bản quyền, phân tích các phương tiện vận chuyển qua video, ... qua điều này chúng ta có thể thấy sự hiện hữu vô hình cũng như tác động không hề nhỏ trong cuộc sống tương lai của việc video stream analytics.
Bên cạnh đó, với bối cảnh đại dịch Coronavirus (Covid-19) đã trở thành một vấn đề cấp thiết cho toàn thể nhân loại trên thế giới. Đại dịch này có sức tàn phá khủng khiếp đối với xã hội và nền kinh tế thế giới gây ra cuộc khủng hoảng sức khỏe toàn cầu. Nguyên nhân cụ thể tạo nên căn bệnh truyền nhiễm này là do Coronavirus 2 (SARS-CoV-2) gây ra. Đến thời điểm hiện tại, vi rút đã lây lan nhanh chóng với phần lớn các quốc gia trên toàn thế giới. Theo Trung tâm

Kiểm soát và Phòng ngừa Dịch bệnh (CDC), việc nhiễm coronavirus lây truyền chủ yếu qua các giọt ở đường hô hấp tạo ra khi con người hít thở, nói chuyện, ho hoặc hắt hơi và sự giải phóng phạm vi lây nhiễm sẽ tăng hơn khi con người nói và hét lớn. Vì lẽ đó, trên toàn thế giới, đặc biệt là trong làn sóng thứ ba, COVID-19 đã và đang là một thách thức đáng kể trong việc chăm sóc sức khỏe cũng như ở nhiều phương diện khác. Nhiều vụ đóng cửa trong các ngành công nghiệp khác nhau là do đại dịch này gây ra. Tuy nhiên, vẫn có một số ngành phải thực hiện và phát triển trong giai đoạn dịch bệnh này như: y tế, cơ sở hạ tầng, an ninh, một số ngành có liên quan và ảnh hưởng đến đời sống sinh hoạt người dân, ... điều này làm nháy lên sự lo lắng khi phải làm việc tại cơ sở trong tình trạng lây nhiễm bởi đại dịch.

Do đó, để ngăn chặn sự lây nhiễm Covid-19 nhanh chóng, nhiều giải pháp đã được đặt ra. Chẳng hạn như dùng biện pháp 5K: không tập trung đông người, khẩu trang, khoảng cách, khử khuẩn, khai báo y tế. Đây là một trong những cách phòng chống cơ bản cần thiêt nhất trong thời điểm trước đó cũng như sau này.

Bên cạnh những khó khăn, trong thời điểm này là một bước tiến lớn cho ngành nghiên cứu từ y học đến công nghệ. Với lĩnh vực công nghệ, biết được vi rút lây lan qua đường không khí khi một người bị nhiễm bệnh hắt hơi hoặc giao tiếp với người khác. Do đó, đeo khẩu trang là điều cần thiết, đặc biệt đối với những người có nguy cơ bị bệnh nặng bởi COVID-19. Người ta thấy rằng sự lây lan của COVID-19 chủ yếu là ở những người tiếp xúc trực tiếp với nhau, nó có thể lây lan bởi những người không có triệu chứng và không biết thực tế là họ bị nhiễm bệnh. Vì vậy, Trung tâm Kiểm soát và Phòng ngừa Dịch bệnh (CDC) khuyến cáo tất cả mọi người từ 2 tuổi trở lên đeo khẩu trang ở các khu vực công cộng. Bằng cách giảm nguy cơ lây truyền loại vi-rút chết người này từ người bị bệnh sang người khỏe mạnh, mức độ lây lan và mức độ nghiêm trọng của bệnh có thể giảm rất nhiều. Tuy nhiên, hiện nay bằng việc ý thức cộng đồng với phương diện đeo khẩu trang ở mỗi người là điều hơi khó khăn và nan giải.

Vì sự việc ấy mà tính năng nhận diện khuôn mặt có mặt nạ đã trở thành chủ đề nghiên cứu được cập nhật và phát triển trong thời gian đại dịch vừa qua bởi đeo khẩu trang trong đại dịch này là một biện pháp phòng ngừa quan trọng và là điều kiện bắt buộc trong thời điểm khó duy trì giãn cách xã hội. Nhận định được điều đó, với sự phát triển không ngừng nghỉ của phân tích video thời gian thực cùng nhận diện khuôn mặt. Nhóm chúng tôi lựa chọn cho mình việc nghiên cứu về đề tài video stream analytics nhằm nhận diện trực tiếp thời gian thực với video được phân tích. Trong quá trình nghiên cứu, nhóm thực hiện đặt ra bài toán cho mình là nhận diện các thực thể bên trong video sau đó sẽ dựa trên phân tích ấy mới tiến hành nhận diện con người có đeo khẩu trang hay không. Nghĩa là thực hiện nhận diện các vật thể với điều kiện, nếu là người (person) thì sẽ bắt đầu thực hiện việc phân tích người ấy có đeo mặt nạ hay không và ngược lại nếu là vật thì sẽ không tiếp tục phân tích tình huống đeo khẩu trang. Bằng việc tiếp cận công nghệ và sự phát triển nền dữ liệu lớn, nhóm đã chọn chủ đề thiết thực này với sự kết hợp lưu trữ big data cho dự án của mình.

2 Công trình liên quan

Tháng 6 năm 2014, Yangqing Jia và các cộng sự với công trình "Caffe: Convolutional Architecture for Fast Feature Embedding" [2] đã giới thiệu về việc cung cấp một framework hoàn chỉnh để đào tạo, kiểm tra, sàng lọc và triển khai với các công trình nghiên cứu liên quan đến học sâu, phù hợp với các ngành tính toán với GPU đạt đến tốc độ xử lý hơn 40 triệu hình ảnh mỗi ngày. Đồng thời đây là một mã nguồn mở, trong đó có Deep Neural Networks (DNN) phát hiện khuôn mặt dựa trên thuật toán Single Shot Detector (SSD) sử dụng ResNet-10. Bao gồm "deploy.prototxt" xác định kiến trúc mạng và "res10_300x300_ssd_iter_140000.caffemodel" xác định trọng số của các lớp được xây dựng.

Năm 2018, Joseph Redmon Ali Farhadi với công trình "YOLOv3: An Incremental Improvement" [3] giới thiệu về bước cải tiến của YOLO so với các phiên bản trước, trong việc nhận dạng vật thể. Tại phiên bản YOLOv3 mức thang đo .5 IOU ngang bằng với Focal Loss nhưng nhanh hơn gấp 4 lần. Nó có ưu điểm hơn so với các hệ thống phân loại, đưa ra kết quả nhanh hơn 1000 lần so với R-CNN và 100 lần so với Fast R-CNN. YOLOv3 sử dụng một số thủ thuật để cải thiện đào tạo và tăng hiệu suất, bao gồm: Thay softmax bằng các logistic classifier rời rạc, Multi-scale prediction, Backbone mới - Darknet-53, ... Nó có thể phát hiện khẩu trang của nhóm vật thể, phân loại chính xác đối tượng, và được xử lý thời gian thực hiên để nhận biết những người đeo khẩu trang.

3 Bộ dữ liệu

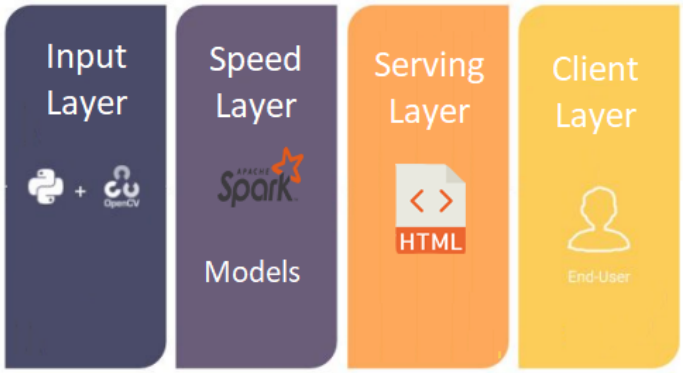
Bộ dữ liệu nhóm chọn để phụ vụ cho công việc phân loại khuôn mặt có đeo khẩu trang hay không là bộ dữ liệu Face Mask 12K Images Dataset [1] từ Kaggle với tổng số 11,792 hình ảnh. Bộ dữ liệu được tạo ra bởi tác giả Ashish Jangra. Đây là bộ dữ liệu sử dụng cho Face Mask Detection Classification với hình ảnh. Hình ảnh trong bộ dữ liệu là hình ảnh có đeo khẩu trang (WithMask) và không đeo khẩu trang (WithoutMask) được cắt theo vùng khuôn mặt. Bộ dữ liệu được chia làm 3 tập bao gồm Train, Test và Validation. Bảng 1 cho thấy sự phân bổ của bộ dữ liệu trong 3 tập.

Bảng 1: Phân bố bộ dữ liệu trong 3 tập

	WithMask	WithoutMask	Tổng
Train	5000	5000	10000
Test	483	509	992
Validation	400	400	800
Tổng	5883	5909	11792

3.1 Phương pháp

Ứng dụng được chia làm 4 Layer. Bao gồm: Input Layer, Speed Layer, Serving Layer và Client Layer.



Hình 1: 4 layer

Input Layer Đầu tiên, tại Input Layer, nhóm lấy hình ảnh từ một video làm hình ảnh đầu vào. Hình ảnh sẽ được ghi tuần tự và chuyển sang định dạng base64 để có thể phù hợp với cấu trúc dữ liệu lớn và sẽ được lưu thành file .csv.

Speed Layer Tiếp theo, ở Speed Layer, nhóm sử dụng Spark để đọc các hình ảnh đầu vào từ file .csv tương ứng. Sau khi dữ liệu được đọc vào, nhóm sử dụng các model để tiến hành xử lý hình ảnh. Đầu tiên thì hình ảnh sẽ được thực hiện phát hiện đối tượng bởi YOLO model. Nếu phát hiện đối tượng "person" thì sẽ tiếp tục sử dụng Caffe model để xác định khuôn mặt của người đó. Sau cùng là gọi model phát hiện đeo khẩu trang do nhóm training để xem xét người đó có đeo khẩu trang không.

Serving Layer Sau khi xử lý, kết quả sẽ được truyền Serving Layer. Kết quả này sẽ là hình ảnh kết hợp giữa hình ảnh đầu vào và kết quả đầu ra của quá trình xử lý để người dùng cuối có thể thấy được một cách tổng quan và dễ hiểu.

4 Thực nghiệm

Giai đoạn đầu, nhóm sử dụng pre-trained model VGG19 của Keras để áp dụng cho việc phát hiện khẩu trang. Nhóm tiến hành train model phân loại khuôn mặt có đeo khẩu trang hay không trên bộ dữ liệu Face Mask 12K Images

Dataset được lấy từ Kaggle với các thông số: steps_per_epoch = 9, epochs = 20, validation_steps = 0. Vì khối lượng hình ảnh sử dụng cho việc training khá lớn nên nhóm tiến hành giảm giá trị mỗi pixel của hình ảnh từ giá trị 0 đến 255 xuống còn 0 đến 1 để mô hình xử lí dễ dàng hơn. Ngoài ra, nhóm còn thực hiện resize kích thước tất cả các hình ảnh đầu vào về kích thước (128,128).

Sau khi đã có model nhận diện khẩu trang, nhóm tiến hành đến bước stream hình ảnh từ video. Mỗi 1s thì nhóm sẽ tiến hành lấy hình ảnh một lần nhờ vào thư viện OpenCV. Để phù hợp với cấu trúc dữ liệu lớn, điểm dữ liệu đầu vào sau khi lấy được lưu với cấu trúc gồm: hình ảnh dưới dạng base64, thời gian lưu hình ảnh, tên sourse video và index tương ứng với thứ tự được lấy của mỗi tấm hình. Và nó được lưu lại dưới dạng file .csv. Cứ 20s thì dữ liệu đầu vào sẽ được gửi đến Speed Layer để tiến hành xử lí.

Tiếp theo, nhóm sử dụng Spark Stream để thực hiện đọc dữ liệu từ thư mục input (nơi chứa các file .csv của hình ảnh đầu vào). Sau khi các điểm dữ liệu được đọc vào, các hình ảnh sẽ được xử lí bởi các model. Đầu tiên thì hình ảnh sẽ được thực hiện phát hiện đối tượng bởi YOLO model. Sau quá trình đó thì ta sẽ nhận được danh sách các đối tượng có trong hình ảnh. Nếu phát hiện "person" thì sẽ tiếp tục thực hiện nhận diện khuôn mặt với "deploy.prototxt" và "res10_300x300_ssd_iter_140000.caffemodel". Hình ảnh khuôn mặt sau khi nhận diện sẽ được resize về kích thước (128, 128) để phù hợp với model nhận diện khẩu trang. Cuối cùng, hình ảnh khuôn mặt đã được nhận diện sau khi resize sẽ được xử lí bởi model nhận diện khẩu trang mà nhóm xây dựng để xác định người đó có đeo khẩu trang hay không.

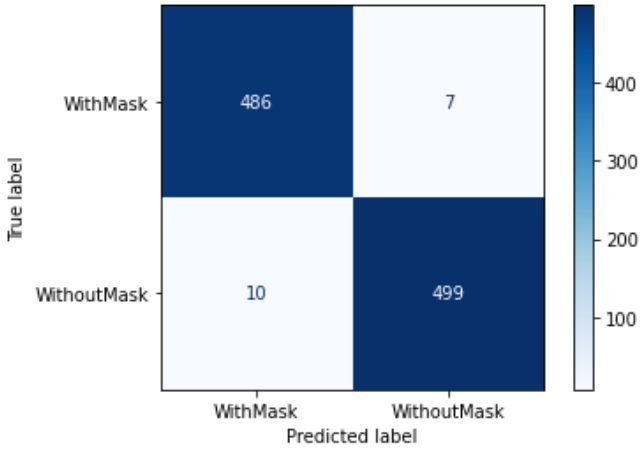
Sau quá trình xử lí, hình ảnh sẽ được lưu lại với các dữ liệu đầu ra nhận được như là đối tượng đó là gì? Nếu là người thì người đó có đeo khẩu trang không? Để người dùng cuối có thể nhận được hình ảnh đó, nhóm đã tiến hành xây dựng một ứng dụng html đơn giản. Ở đó, hình ảnh đầu vào và đầu ra sẽ được hiển thị.

5 Kết quả thực nghiệm

Về model nhận diện khẩu trang, model mà nhóm thực hiện training có Accuracy lên đến 98.3%. Ngoài thông số Accuracy, nhóm còn đánh giá model trên 3 thống số Precision, Recall, F1-score, số liệu được trình bày trong Bảng 2

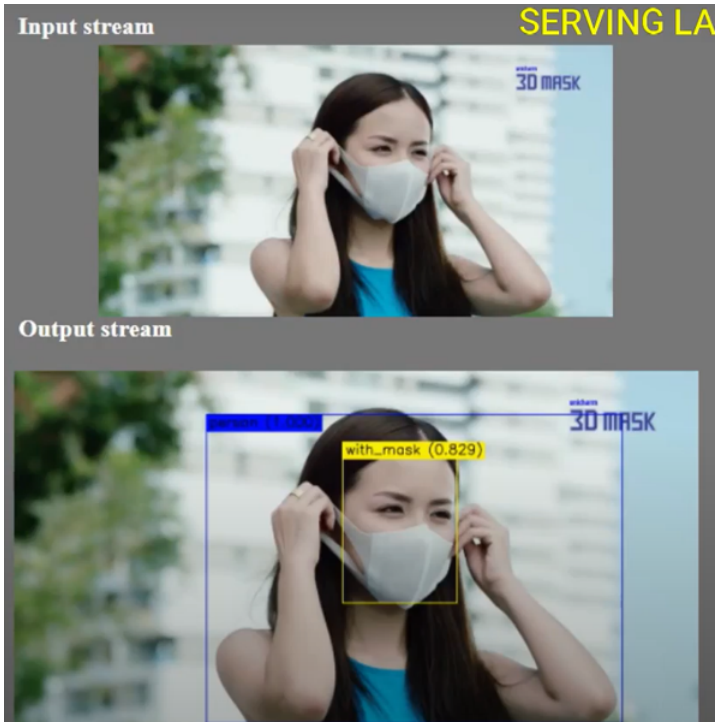
Bảng 2: Precision, Recall, F1-score của model trên bộ dữ liệu

Nhãn	Precision	Recall	F1-score
WithMask	98%	99%	98%
WithoutMask	99%	98%	98%



Hình 2: confusion matrix model nhận diện khẩu trang

Bên cạnh đó, khi nhìn vào Hình 2 ta thấy được dự đoán sai cao hơn đối với WithoutMask với 10 nhãn dự đoán sai trên tổng số 509 nhãn.
Về ứng dụng, nhóm đã thực hiện Stream thành công và nhận diện được 80 vật thể và cái mà nhóm chú trọng nhiều nhất là phát hiện person. Sau khi phát hiện persion thì nhóm thành công xác định được persion đó có đang mang khẩu trang hay không. Mỗi hình ảnh đầu vào mất 5s để có thể hoàn thành quá trình xử lý. Đồng thời, nhóm cũng xây dựng được thành công một ứng dụng HTML đơn giản để có thể hiển thị kết quả dưới dạng hình ảnh.



Hình 3: Kết quả hiển thị cho người dùng cuối

6 Kết luận và hướng phát triển

Trong thời gian làm đề tài, nhóm chúng em đã áp dụng được nhiều kiến thức giảng dạy của Thầy về Big data đặc biệt là PySpark. Bên cạnh đó, để giải quyết bài toán, chúng em còn phải tìm hiểu thêm các thông tin ngoài luồn tài liệu giảng dạy và kết hợp cùng kiến thức đã có từ giảng viên để hoàn thành vấn đề đặt ra.

Sau quá trình thực hiện và nghiên cứu thì nhóm đã thực hiện chạy được cả ba model là phân tích nhận diện vật thể, xác định khuôn mặt và xác định khuôn mặt có đeo khẩu trang với độ tin cậy được xem là khá ổn. Kết quả trả về với độ chính xác tương đối cao.

Ngoài kết quả đạt được như mong đợi ban đầu đề ra, thì nhược điểm của bài nghiên cứu là cho ra kết quả được đánh giá là khá chậm. Mỗi hình ảnh mất 4-5s để xử lí và mất khá nhiều thời gian cho các thao tác xử lý trong Spark. Tuy nhiên thời gian này cũng được xem là chấp nhận được cho việc nhận diện với trích xuất kết quả thời gian thực.

Trong thời gian tới, nhóm sẽ thực hiện cải tiến cũng như học hỏi thêm hoặc khai thác tìm hiểu thuật toán nhằm cải thiện model của mình để cho ra kết quả nhanh hơn thời điểm hiện tại. Bởi các dữ liệu cập nhật vào liên tục cần phải nhận được trả kết quả nhanh chóng để đáp ứng nhu cầu cũng như các đặt tính cần thiết cho người dùng trong tương lai.

Ngoài ra, để mang tính chất vận dụng và thiết thực, nhóm có thể sẽ thực hiện tạo một ứng dụng về vấn đề nhận diện này và tìm một đối tác cùng phát triển ra sản phẩm để cung ứng sản phẩm trong tương lai cho đối tượng khách hàng.

Lời cảm ơn Nhóm chúng em xin chân thành cảm ơn Trường Đại học Công Nghệ Thông Tin đã tạo cho các sinh viên có môi trường học tập thoải mái. Cảm ơn ThS. Đỗ Trọng Hợp - giảng viên phụ trách môn Công nghệ dữ liệu lớn đã chỉ dạy chúng em các kiến thức từ cơ bản đến nâng cao của môn học. Bên cạnh đó, Thầy còn hỗ trợ và giúp đỡ chúng em rất nhiều, đưa ra những góp ý để chúng em có thể hoàn thành trọn vẹn đề tài báo cáo của mình. .

Tài liệu

1. Ashish Jangra. Face mask detection 12k images dataset, 2020. Retrieved December 1, 2021 from https://www.kaggle.com/ashishjangra27/face-mask-12k-images-dataset/activity.
2. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014.
3. Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.