ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO MÔN HỌC

THỊ GIÁC MÁY TÍNH NÂNG CAO

CS331.N12.KHCL

ĐỀ TÀI: DEEPFAKE DETECTION

GIẢNG VIÊN HƯỚNG DẪN: **MAI TIẾN DŨNG**

SINH VIÊN THỰC HIỆN: **NGUYỄN VŨ DƯƠNG – 20520465**

LÊ TRẦN QUỐC KHÁNH - 20520574

TP. HỒ CHÍ MINH, 12/2022

LỜI CẢM ƠN

Sức mạnh tính toán của các máy tính ngày càng tăng điều đó đã làm cho các kỹ thuật học sâu(deep learning) phát triển đến mức có thể được cho là không thể chỉ một vài năm trước đây. Sự lan rộng của deepfake trên các nền tảng truyền thông đã trở nên phổ biến dẫn đến thư rác và các thông tin sai sự thật trên mạng xã hội. Những kiểu deepfake này sẽ rất tồi tệ có thể dọa nạt và đánh lừa mọi người.

Là sinh viên ngành KHMT, trong đồ án môn Thị giác máy tính nâng cao này, nhóm chúng em đã chọn và thực hiện đồ án "Deepfake Detection" để thử mình cho vấn đề phát hiện Deepfake trên ảnh và video.

Nhóm xin gửi lời cảm ơn chân thành đến Thầy Mai Tiến Dũng đã tận tình giảng dạy, hướng dẫn chúng em trong suốt thời gian học vừa qua và các bạn học đã góp ý và giúp đỡ nhóm trong quá trình thực hiện đồ án này.

Do kiến thức và thời gian thực hiện hạn chế, đồ án của nhóm vẫn còn nhiều thiếu sót. Nhóm rất mong nhận được góp ý của Cô và các bạn để đồ án của nhóm được hoàn thiện.

TP. Hồ Chí Minh, tháng 12 năm 2022.

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC CHỮ VIẾT TẮT

DF	Deepfake
GAN	Generative Adversarial Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
SVM	Support-vector machine
AI	Artificial Intelligence
PRNU	Photo Response NonUniformity
LTRC	Long-Term Recurrent CNN
rPPG	remote PhotoPlethysmoGraphy
RGB	Red, Green and Blue
FF++	Face Forensics ++
2D	2 Dimension
BN	Batch Normalization

I) Giới thiệu

1) Tổng quan

- Deepfake là một thuật ngữ chỉ một kỹ thuật cải thiện các hình ảnh và video bởi trí tuệ nhân tạo và các công nghệ hiện đại để tạo ra các sản phẩm giả.
- Một trong những ví dụ nổi tiếng nhất liên quan đến deepfake là việc sử dụng xử lý hình ảnh để sản xuất video của những người nổi tiếng, chính trị gia hoặc những người khác nói hoặc làm những điều họ chưa bao giờ thực sự làm.
- Bất kỳ ai có kỹ năng dùng máy tính cơ bản đều có thể dễ dàng tạo ra deepfake
- Một số phương pháp phát hiện deepfake bằng việc sử dụng AI để phân tích video từ đó biết những đặc điểm mà deepfake không thể học (bắt chước) một cách giống nhất (Ví dụ: Chớp mắt hoặc giật cơ mặt)
- Tuy nhiên, việc phát hiện ra một video hay hình ảnh có dùng deepfake hay không là một thách thức rất lớn vì công nghệ deepfake phát triển ngày càng hiện đại.

2) Quá trình hình thành và phát triển của DeepFake

- Nguồn gốc của deepfake bắt đầu từ năm 1997. Đó là một video được tạo ra bởi chương trình của Bregler và cộng sự, lần đầu tiên công bố nghiên cứu về deepfake như là kết quả của một hội nghị về đồ họa và kỹ thuật tương tác máy tính. Các nhà nghiên cứu đã giải thích làm thế nào để sửa đổi cảnh quay video hiện có của một người đang nói bằng một giọng nói khác.
- Đó không phải là một khái niệm mới (việc chỉnh sửa ảnh đã xảy ra từ thế kỷ 19) nhưng các nghiên cứu liên quan dường như không đáng kể về việc thay đổi nội dung của video đây là nghiên cứu đầu tiên thuộc loại tự động quá trình tái tạo các đặc điểm trên khuôn mặt. Bằng việc sử dụng các thuật toán máy học đã giúp đạt được điều này. Đây là một cột mốc quan trọng và có thể coi đây là điểm khởi đầu của sự phát triển video deepfake trên thế giới.

- Tuy nhiên, phải đến nhiều năm sau, khái niệm về deepfake mới thực sự bắt đầu phổ biến. Giống như nhiều công nghệ mới, việc được sử dụng rộng rãi diễn ra khá lâu sau khi được phát minh ra. Face2Face lần đầu được xuất hiện vào năm 2016 , được viết bởi Thies và cộng sự, đã chỉ ra.
  - công nghệ sử dụng hình ảnh khuôn mặt trong thời gian thực có thể tái tạo lại 1 video một cách chân thực.

- Đến tháng 7 năm 2017 , nhiều người mới bắt đầu quan tâm đến các ứng dụng sử dụng deep learning. Một video khá nổi tiếng về việc sử dụng deepfake đó là video về cựu tổng thống Barack Obama. Video ấy do Suwajanahorn và cộng sự tạo ra, cho thấy rằng lần đầu tiên giọng nói có thể được nhái lại một cách chân thực đến đáng sợ như thế. Có thể thấy được rằng đây là một mối nguy hiểm có thể làm thay đổi cục diện tương lai chính trị.
- Sau đó mọi thứ diễn ra rất nhanh. Việc ứng dụng và giới thiệu cộng nghệ bùng nổ một cách nhanh chóng, khi ngày càng nhiều video deepfake và những thứ tương tự ứng dụng các kỹ thuật máy học (machine learning) được sử dụng trên toàn cầu. Deepfake Obama là một trong số những video phổ biến đó, và một năm sau đó nó đã được đăng lên youtube trên channel có tên là "BuzzFeed" và video có tiêu đề là "You Won't Believe What Obama Says In This Video!" .
- Video đó là một trong số rất nhiều video tương tự được ra mắt trong khoảng thời gian đó (mùa xuân năm 2018) là một trong những video deepfake đầu tiên giới thiệu cho mọi người về ý tưởng rằng video có thể là giả theo cách thực tế đến mức chúng ta không thể phân biệt. Ngay sau , ứng dụng "SnapChat" đã có "face swap filters" , trên các website nhiều có rất nhiều ứng dụng thay đổi giọng nói trực tiếp, hay thậm chí là phần mềm tạo video deepfake với mã nguồn mở và nhiều các ứng dụng, phần mềm khác.
- Deepfake hiện đã trở nên không thể kiểm soát được. Vì khi nhìn vào một mặt, đó là một điều tốt. Nhưng chúng ta không nên đánh giá thấp những tác động có hại mà deepfake có thể làm cho xã hội của chúng ta. Ngày nay chúng ta đang bước dần vào một kỷ nguyên mới, trong đó Deepfake sẽ dần bắt đầu trở thành một yếu tố liên quan trong nhiều ngành công nghiệp. Mối đe dọa của deepfake đang rõ ràng hơn bao giờ hết.

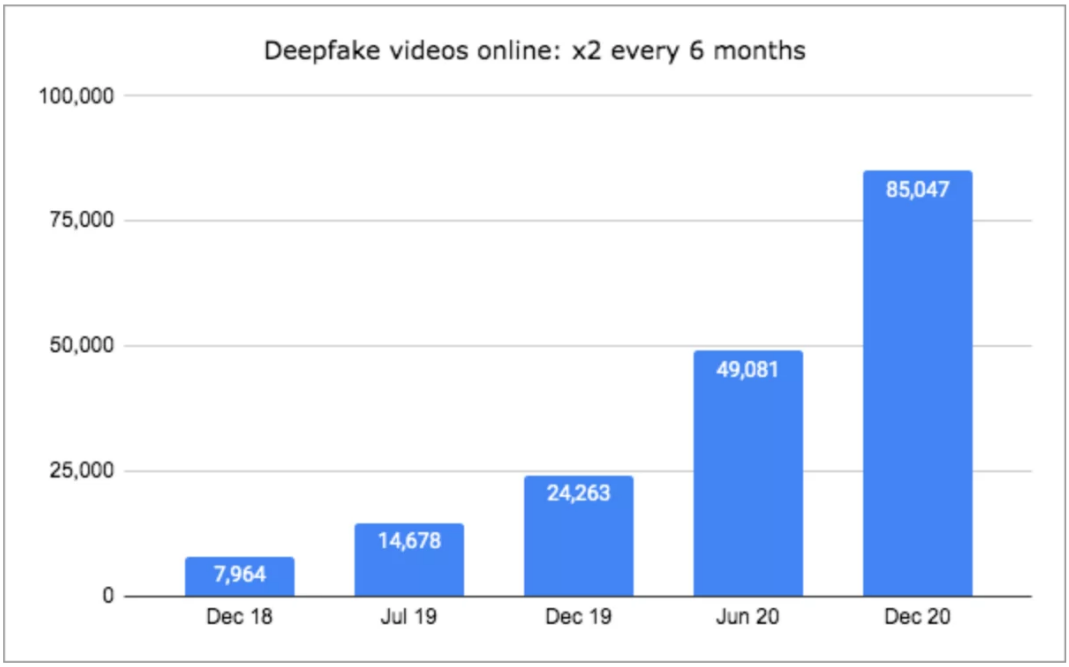
3) Vấn đề

- Sức mạnh tính toán của các máy tính ngày càng tăng điều đó đã làm cho các kỹ thuật học sâu (deep learning) phát triển đến mức có thể được cho là không thể chỉ một vài năm trước đây. Sự lan rộng của deepfake trên các nền tảng truyền thông đã trở nên phổ biến dẫn đến thư rác và các thông tin sai sự thật trên mạng xã hội. Những kiểu deepfake này sẽ rất tồi tệ có thể dọa nạt và đánh lừa mọi người.

- Deepfake đặt ra một vấn đề quan trọng đối với việc nâng cao kiến thức cộng đồng. Sự phát triển của chúng không phải là khởi nguồn của hình ảnh , âm thanh và video được chỉnh sửa đã tràn ngập trên Internet trong một thời gian dài, nhưng chúng sẽ đóng góp một phần đáng kể về việc mất niềm tin vào các nội dung kỹ thuật số. Các công cụ trí tuệ nhân tạo và nhiều công nghệ dành cho deepfake được phát hành ra thị trường với mã nguồn mở. Điều này có nghĩa là việc tao ra một nọi dung giả mạo dễ dàng hơn bao giờ hết và trở nên dễ tiếp cận hơn trong tương lai.
- Có lẽ sẽ có một video deepfake được đăng trên nền tảng mạng xã hội sẽ thuyết phục nhiều người điều đó có thể làm tạm thời sụp đổ nền kinh tế , gây bạo lực hoặc tổ chức một cuộc bạo loạn . Tuy nhiên đây không phải là những tình huống có khả năng xảy ra nhất. Nhiều khả năng một số lượng lớn deepfake được tải lên bởi những người nghiệp dư (với mục đích châm biếm hoặc chính trị) và các chiến dịch gây ảnh hưởng sẽ dần phân tán trên internet càng ngày càng làm che mờ tính xác thực trong thế giới kỹ thuật số.

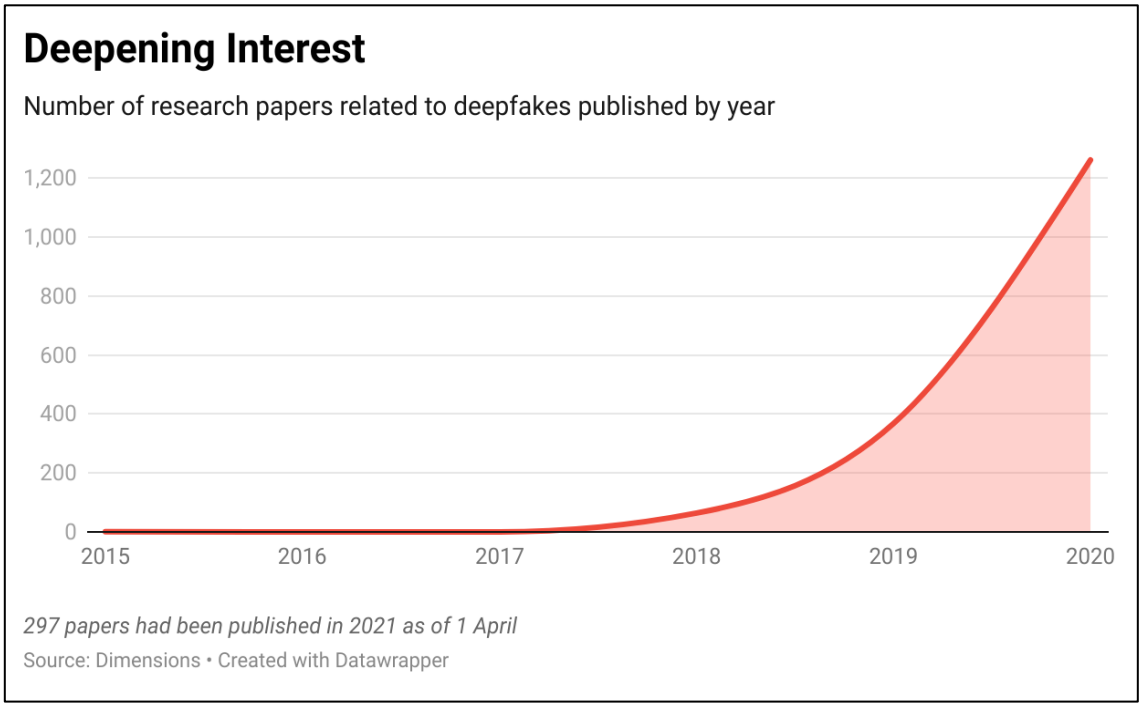
4) Sự phát triển của video deepfake

- Vào tháng 7 năm 2019 , một báo cáo đã cho biết rằng đã tìm thấy hơn 14678 video trực tuyến sử dụng deepfake , gấp 2 lần so với video deepfake vào năm 2018 là 7964 video trong số các video ấy có tới 96% video là phim khiêu dâm.
- Vào cuối năm 2019, đã có một sự gia tăng lớn trong vòng chưa đầy một năm! Xu hướng gia tăng này diễn ra mạnh mẽ khi deepfake trở thành trào lưu.
- Sự phát triển của các video deepfake được tăng lên đáng chú ý, kể từ tháng 6 năm 2020, các video deepfake trực tuyến được xác định đã tăng gấp đôi lên 49081 chỉ sau 6 tháng kể từ tháng 1, trích dẫn từ một báo cáo của Deep Trace Lab – một công ty công nghệ deepfake



Source: Sensity

Hình 1.1: Sự phát triển của video deepfake từ tháng 12 năm 2018



Hình 1.2: Sự phát triển của bài báo liên quan tới deepfake

5) Giải pháp

- Để khắc phục tình trạng như vậy, phát hiện deepfake là rất quan trọng. Vì thế , nhóm em sẽ giới thiệu các phương pháp để có thể phân biệt hiệu quả các video và hình ảnh do AI tao ra từ video và hình ảnh thực. Điều đó là cực kỳ quan trọng để phát triển công nghệ có thể phát hiện để ngăn chặn deepfake tràn lan trên mạng.

6) Lý do chọn đề tài?

- Do sự phát triển nhanh chóng và sự chính xác đến kinh ngạc của deepfake, nhiều công ty kỹ thuật đã hợp tác cùng nhau tạo ra bộ dữ liệu để có thể chống lại deepfake. Các video deepfake phổ biến đến mức nhiều đảng phái chính trị sử dụng công cụ này để tạo ra hình ảnh giả mạo về lãnh đạo của đảng đối lập nhằm tuyên truyền sai sự thật để chống lại họ. Video/hình ảnh giả mạo là một mối đe dọa cho chiến dịch bầu cử.
- Ngoài ra, nhiều thiết bị công nghệ đang sử dụng nhận dạng khuôn mặt hay là nhận dạng giọng nói . Các phương pháp này được cho là những cách an toàn nhất để bảo vệ thông tin cá nhân của bạn, nhưng với sự hùng mạnh của deepfake ngày nay thì tương lai các công cụ này liệu còn an toàn không vẫn là một câu hỏi.
- Ví dụ điển hình là “Face id” trên các thiết bị Iphone(Từ iphone X trở lên),… Nhưng điều gì sẽ xảy ra nếu ai đó có thể tạo ra chân dung và giọng nói của bạn? Ngày càng có nhiều lo ngại rằng công nghệ mới này sẽ ảnh hưởng lớn đến các công nghệ bảo mật. Để điều tra được sự giả mạo nét mặt chi tiết, thì ta cần phải kiểm tra từng hình ảnh trong quá trình tạo ra một bức ảnh deepfake.
- Mục tiêu của đồ án là tạo ra được mô hình có khả năng nhận dạng hình ảnh deepfake. Phân tích kỹ lưỡng các khung hình trong video để xác định những điểm không hoàn hảo trên khuôn mặt và model sẽ học cách có thể phân biệt hình ảnh thật với deepfake.

II) Cơ sở lý luận và tổng quan lý thuyết

1) Cơ sở lý luận

- Deepfake là video/hình ảnh giả được tạo ra bằng cách thao túng một người trong video/hình ảnh hiện có bằng các phương pháp máy học.
- Ví dụ cho cái này chính là video nói về cựu tổng thống Barack Obama mắng cựu tổng thống Donald Trump. Mục đích chính của video này là cho thấy được hậu quả của deepfake. Năm 2019, trên mạng cũng xuất hiện một video deepfake có sự xuất hiện của Mark Zuckerberg nói về cách Facebook kiểm soát dữ liệu của hàng tỷ người dùng.
- Các video đó được tạo bằng cách sử dụng Autoencoder và GANs để tạo ra một video chân thực nhất có thể.Deepfake(Được kết hợp từ 2 từ "deep learning" và "fake") được tạo ra bằng các kỹ thuật có thể thiết lập khuôn mặt của mục tiêu vào video nguồn .
- Ngày nay, video deepfake rất dễ dàng để tạo nên ai cũng có thể tạo ra chúng. Các phần mềm tạo deepfake ví dụ như là FakeApp , DeepFaceLab, FaceSwap và các phần mềm này dễ dàng sử dụng.

1.1) Các loại Deepfake

- Face-swapping: Hoán đổi khuôn mặt trong 2 bức hình với nhau để tạo hình ảnh giả mạo.
- Face reenactment: Thay đổi các đặc điểm trên khuôn mặt của một người.
- Audio Deepfakes: Giả mạo âm thanh của một người cụ thể.
- Lip-syncing Deepfakes: Video có cử động miệng nhất quán kém theo âm thanh.
- Puppet-master: Video của một người cụ thể (con rối) sử dụng chuyển động hành vi của một người khác đang ngồi trước máy quay.

1.2) Cách deepfake được tạo ra ?

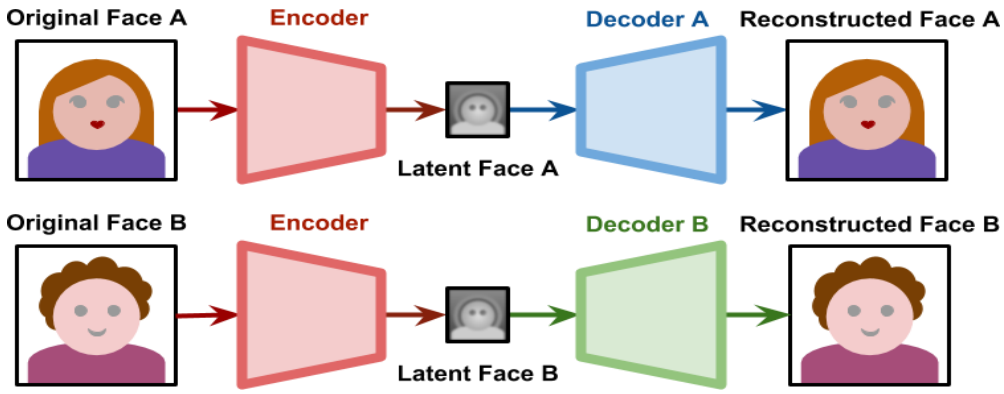
Autoencoder và GANs là hai kỹ thuật học sâu được áp dụng cho các ứng dụng deepfake ngày nay:

Autoencoder

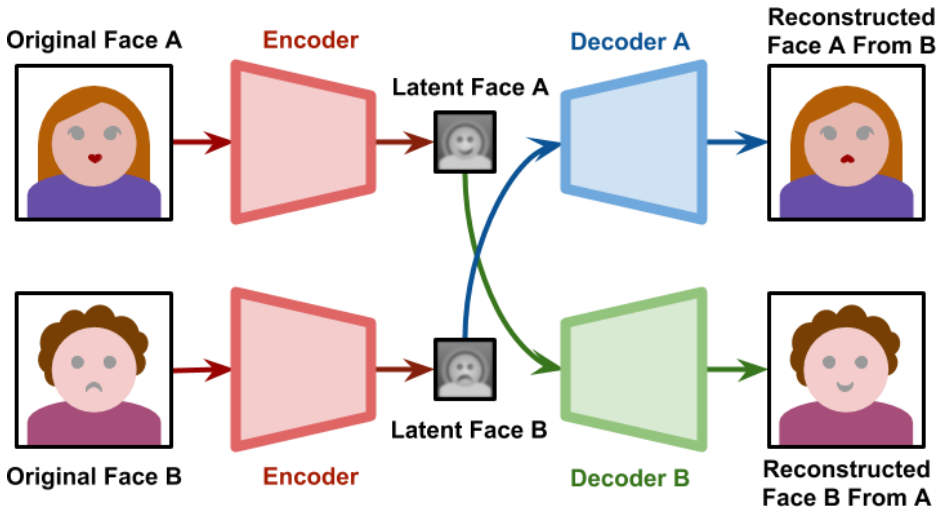
- Autoencoder chủ yếu sử dụng cho Face-swapping. Để tạo ra một video deepfake về một ai đó, bạn cần huấn luyện autoencoder gồm 2 phần : encoder và decoder. Đối với encoder , bạn cần phải cung cấp cho nó nhiều hình ảnh của hai người mà bạn muốn hoán đổi. Để làm cho những

hình ảnh này thực tế hơn, hình ảnh cần bao gồm các bức ảnh chụp khuôn mặt từ các góc độ và ánh sáng khác nhau.

- Trong quá trình huấn luyện, encoder sẽ trích xuất các đặc điểm trên từng khuôn mặt. Sau đó, decoder sẽ tái tạo và khôi phục các hình ảnh từ các đặc điểm này.
- Ví dụ: Giả sử bạn đã huấn luyện hình ảnh của người thứ nhất bằng decoder A. Sau đó bạn có thể sử dụng decoder B để khôi phục hình ảnh đó . Decoder A sau đó tái tạo hình ảnh của người thứ 2 sử dụng các đặc trưng của người số 1. Khi bạn hoàn thành xong quá trình huấn luyện, hoán đổi hai decoder để khôi phục hai hình ảnh khác nhau.



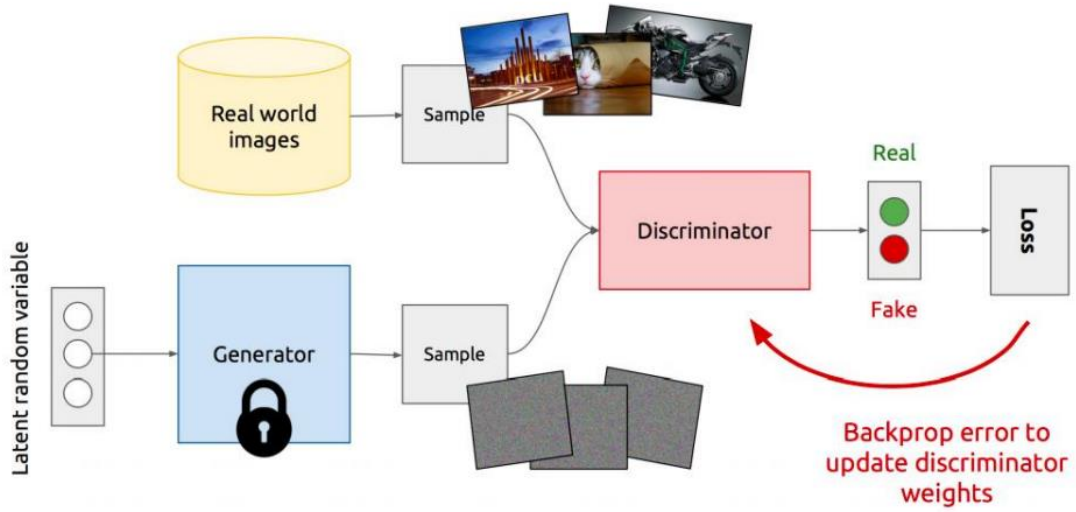
Hình 2.1: Quá trình huấn luyện



Hình 2.2: Quá trình tạo ra deepfake

Generative Adversarial Networks (GANs)

- Mục tiêu của mạng GANs là tạo ra một cái gì đó mới dựa trên dữ liệu trước đó. Ví dụ, nó có thể tạo ra khuôn mặt của con người sau khi nghiên cứu hàng nghìn bức ảnh. Hoặc nó có thể tạo ra một bức tranh giống với phong cách vẽ tranh của một họa sĩ cụ thể nào đó bằng cách sử dụng chính tác phẩm của họ làm tài liệu tham khảo.
- GANs bao gồm 2 neural network và cùng "cạnh tranh" trực tiếp với nhau - Generator và Discriminator . Generator tạo ra một hình ảnh mới dưới dạng là kiến thức mà generator được học. Discriminator sẽ xác định xem hình ảnh là thật hay giả.
- Cả hai thành phần này luôn tương tác liên tục với nhau . Generator tìm hiểu cách tạo hình ảnh sẽ đánh lừa được Discriminator và khiên Discriminator phân loại hình ảnh được tạo đó là hình ảnh thực. Mặt khác, Discriminator sẽ học cách làm sao cải thiện được cách phát hiện . Discriminator càng phân biệt rõ ràng và tốt bao nhiêu thì Generator sẽ càng tạo hình ảnh ngày càng chân thực bấy nhiêu. Phương pháp này để thể hiện trong hình 2.3
- Để minh họa một cách dễ hiểu. Ví dụ , một giáo viên (Discriminator) giảng dạy càng tốt thì càng giúp học sinh (Generator) học tập và cải thiện kết quả học tập của họ. Khi học sinh nộp bài tập của mình , và giáo viên đánh dấu những lỗi sai của họ. Chỉ khi đó học sinh mới nhận ra những gì họ đã làm sai và sửa lỗi của họ. Vì lý do này Generator sẽ tạo ra hình ảnh chân thực nhất có thể



Hình 2.3: GANs

1.3) Cần bao nhiêu bức ảnh để tạo ra deepfake?

Độ chính xác và chất lượng của hình ảnh deepfake phụ thuộc rất nhiều vào số lượng hình ảnh được sử dụng để đào tạo model. Tốt nhất là sử dụng khoảng từ 300-2000 bức hình của mục tiêu muốn fake để tạo ra hình ảnh giống nhất có thể. Các hình ảnh cũng cần phải có một loạt các đặc điểm của khuôn mặt.

2) Tổng quan lý thuyết

Các video deepfake ngày càng đe dọa đến quyền riêng tư và an ninh xã hội. Nhiều phương pháp đã được đề xuất để phát hiện các video này có bị thao túng hay không. Những nỗ lực ban đầu chủ yếu tập trung vào các đặc điểm không nhất quán do quá trình tổng hợp khuôn mặt gây ra trong khi các phương pháp hiện đại ngày nay chủ yếu tập trung vào các đặc điểm cơ bản. Những phương pháp này gồm 5 loại phụ thuộc vào đặc điểm của chúng.

Đầu tiên, việc phát hiện bằng việc sử dụng mạng nơ-ron nói chung thường được sử dụng trong rất nhiều tài liệu, trong đó nhiệm vụ phát hiện deepfake được coi là nhiệm vụ phân loại thông thường. Các đặc trưng nhất quán tạm thời cũng được khai thác để phát hiện sự gián đoạn của các khung hình liền kề của video fake.

Các phương pháp được đề xuất gần đây tập trung vào các tính năng cơ bản hơn, trong đó các sơ đồ dựa trên dấu vân tay và tín hiệu sinh học của máy ảnh cho thấy tiềm năng lớn cho việc phát hiện.

2.1) Phương pháp General-network-based

Những tiến bộ gần đây trong phân loại hình ảnh đã được áp dụng để cải thiện việc phát hiện video deepfake. Trong phương pháp này, hình ảnh khuôn mặt được trích xuất từ video được sử dụng để huấn luyện model. Sau đó, model được đào tạo áp dụng để đưa ra dự đoán cho tất cả khung hình của video. Dự đoán cuối cùng được tính toán bằng số lượng trung bình hoặc bỏ phiếu. Do đó, độ chính xác phụ thuộc nhiều vào mạng mà không cần thiết phải khai thác các đặc trưng cụ thể nào

Các phương pháp dựa vào mạng học sâu hiện có được chia thành hai loại:

- Phương pháp Transfer learning-based
- Phương pháp phát hiện dựa trên mạng được thiết kế đặc biệt

2.2) Phương pháp Temporal-consistency-based

Tính liên tục của thời gian là một đặc trưng độc nhất của video. Không giống như hình ảnh , video là một chuỗi bao gồm nhiều khung hình, trong đó các khung liền kề có mối liên quan chặt chẽ và liên tục. Khi các khung hình video bị thao túng, tính tương quan giữa các khung hình liền kề sẽ bị phá vỡ do lỗi của thuật toán deepfake, thể hiện cụ thể ở sự chuyển vị trí khuôn mặt và hiện tượng giật hình video. Do đó, các nhà nghiên cứu đã đề xuất kiến trúc CNN-RNN.

CNN-RNN

Theo tính liên tục thời gian trong video, lần đầu tiên đề xuất sử dụng RNN để phát hiện video deepfake. Trong công việc đó, autoencoder được phát hiện là hoàn toàn không biết về các khuôn mặt được tạo trước đó vì các khuôn mặt được tạo theo từng khung hình. Do đó dẫn đến nhiều điểm bất thường, là bằng chứng quan trọng để phát hiện deepfake.

Để kiểm tra tính liên tục giữa các khung hình liền kề, một hệ thống phát hiện video deepfake lặp lại có thể huấn luyện từ đầu đến cuối đã được đề xuất. Hệ thống được đề xuất chủ yếu bao gồm long short-term memory (LSTM) structure để xử lý các chuỗi khung.

Hai thành phần thiết yếu được sử dụng trong cấu trúc LSTM tích chập, trong đó CNN được sử dụng để trích xuất đặc trưng và LSTM được sử dụng để phân tích chuỗi thời gian. Cụ thể, InceptionV3 được điều chỉnh đề xuất ra một biểu diễn sâu cho từng khung hình. Các vector đặc trưng 2048 chiều được trích xuất bởi các lớp pooling làm đầu vào cho LSTM, đặc trưng cho tính liên tục giữa các chuỗi khung hình. Cuối cùng là lớp fully connected và lớp softmax được thêm vào để tính toán xác suất của chuỗi khung hình được kiểm tra.

2.3) Phương pháp Visual Artifacts-based

Trong hầu hết các phương pháp deepfake hiện có, khuôn mặt được tạo phải dựa trên một khuôn mặt hiện có, gây ra sự khác biệt hình ảnh trên các ranh giới . Các khuôn mặt và hình nền đến từ các hình ảnh nguồn khác nhau, dẫn đến các đặc điểm bất thường về ranh giới và độ sáng không nhất quán.

Có 3 Visual Artifacts chính:

- Tư thế đầu không nhất quán.
- Ranh giới "ghép ảnh".
- Độ tạo nét mặt.

2.4) Phương pháp Camera-fingerprints-based

Camera fingerprints là một loại nhiễu yếu , có vai trò quan trọng trong lĩnh vực pháp ý, đặc biệt là nhiệm vụ xác định nguồn gốc. Nói chung, phương pháp này đã phải trải qua ba quá trình: tín không đồng nhất của ảnh (PRNU), Noiseprint và mẫu nhiễu video gần nhất.

2.5) Phương pháp Biological-signals-based

Phát hiện dựa trên các tín hiệu sinh học là một phương pháp xuất hiện trong những năm gần đây. Mặc dù GANs có thể tạo ra các khuôn mặt có độ chân thực cao, tuy nhiên các tín hiệu sinh học tự nhiên ẩn không dễ dàng để sao chép được, gây khó khăn cho việc tổng hợp khuôn mặt con người với hành vi hợp lý. Do đó sử dụng các hành vi bất thường này, một số nghiên cứu đã được đề xuất.

Có 2 phương pháp tiếp cận dựa trên tín hiệu sinh học:

- Nháy mắt dựa trên tần số (Blinking frequency-based)
- Phát hiện dựa trên nhịp tim (Heart rate-based)

Eye Blinking(Nháy mắt)

Những bất thường về tần suất nháy mắt đã được xác định là đặc trưng có thể phân biệt được trong việc phát hiện deepfake. Điều này là do các thuật toán deepfake huấn luyện model bằng cách sử dụng một lượng lớn hình ảnh khuôn mặt.

Hầu hết các hình ảnh đều là hình ảnh của người mở mắt khiến cho khó có thể tạo ra lúc nhắm mắt trong một video giả. Dựa trên phát hiện này, một mô hình mạng được gọi là long-term recurrent CNN (LRCN), đã được giới thiệu để phân biệt trạng thái mở mắt và nhắm mắt.

Heart rate (Nhịp tim)

Bên cạnh dự chớp mắt, nhịp tim cũng được tìm thấy sự khác biệt giữa video thực và video giả mạo. Người ta đã chứng minh được rằng sự thay đổi màu sắc của da trong video có thể được áp dụng để suy ra nhịp tim. Dựa trên những phát hiện này, một máy dò tín hiệu sinh học có tên là FakeCatcher đã được thiết kế để phát hiện các video deepfake. Cụ thể, remote photoplethysmography (rPPG) đã được sử dụng để trích xuất tín hiệu nhịp tim theo những thay đổi nhỏ về màu sắc và chuyển động trong video RGB

3) Tóm lại

- Deepfake đã bắt đầu làm mất niềm tin của mọi người vào các video/hình ảnh/ âm thanh vì việc nhìn chúng ta không thể phân biệt cái nào là thật cái nào là giả. Chúng có thể gieo rắc nỗi đau khổ và tác động tiêu cực đến mục tiêu được nhắm tới , không chỉ thế nó còn làm tăng khối lượng thông tin sai sự thật , ngôn từ kích động thù địch thậm chí có thể kích động căng thẳng chính trị , kích động công chúng, bạo lực hoặc chiến tranh. Tất cả những điều đó đã khiến chúng em chọn đề tài về phát hiện deepfake.
- Ngày càng nhiều phương pháp dựa màng mạng học sâu bắt đầu được giới thiệu là cách học multitask, nghĩa là không chỉ phân loại khuôn mặt thật và khuôn mặt giả mà còn tạo “tampering masks” ở mức độ pixel.
- So sánh với các phương pháp dựa vào mạng học sâu nói chung, các phương pháp dựa trên tính nhất quả theo thời gian của từng khung ảnh cải thiện được khả năng phát hiện. Tuy nhiên nhiều model có xu hướng phá hủy cấu trúc không gian của các khung hình gốc khi tích xuất các đặc trưng trong khi mục đích thiết kế phương pháp này là để trích xuất sự không nhất quán của các đặc trưng không gian trong miền thời gian. Kiến trúc CNN-RNN gộp các đặc trưng trong khung thành vector do đó không thể nắm bắt các đặc điểm không gian trong khi phát hiện tính nhất quán theo thời gian.
- Phương pháp Visual-artifacts-based thường đạt được hiệu suất tổng quát hóa tốt hơn vì chúng nhắm tới các độ tái tổng quát hơn tồn tại trong hầu hết nội dung deepfake. Tuy nhiên, các thuật toán này chỉ có thể phát hiện dấu vết giả mạo . Với sự phát triển của các thuật toán deepfake, những độ tạo này đang dần biến mất. Tuy nhiên, các phương pháp này đạt được độ hiệu quả tốt hơn trong phiên bản mới nhất của bộ dữ liệu video deepfake.
- Camera fingerprints đã được chứng mình là có hiệu quả trong việc phát hiện deepfake. Tuy nhiên , ước tính chính xác Camera

fingerprints yêu cầu một số lượng lớn hình ảnh được chụp bởi các loại máy ảnh khác nhau. Do đó, độ chính xác sẽ giảm khi phát hiện hình ảnh được chụp bởi camera không xác định.

- Mặt khác, các phương pháp camera fingerprint-based không thực sử ổn định (hiệu quả) đối với hình ảnh xử lý hậu kỳ đơn giản như nén, nhiễu, mờ. Vì hình ảnh GANs được tạo mà không có bất kỳ quy trình chụp ảnh nào, nên sẽ không có camera fingerprint trong hình ảnh đầu ra, vì vậy các phương pháp camera fingerprint-based rất phù hợp các hình ảnh được tạo bởi GANs
- Mặc dù các phương pháp phát hiện dựa trên tín hiệu sinh học đã cho thấy hiệu suất tốt trên nhiều bộ dữ liệu khác nhau, những lỗ hổng tự nhiên của loại phương pháp này là quá trình phát hiện không thể được phát hiện theo cách từ đầu đến cuối. Ngoài ra, thông tin được phản ánh bởi tín hiệu sinh học bị ảnh hưởng nghiêm trọng bởi chất lượng video, do đó , có những sai sót tự nhiên và phạm vi ứng dụng hạn chế đối với các phương pháp tiếp cận dựa trên tín hiệu sinh học.

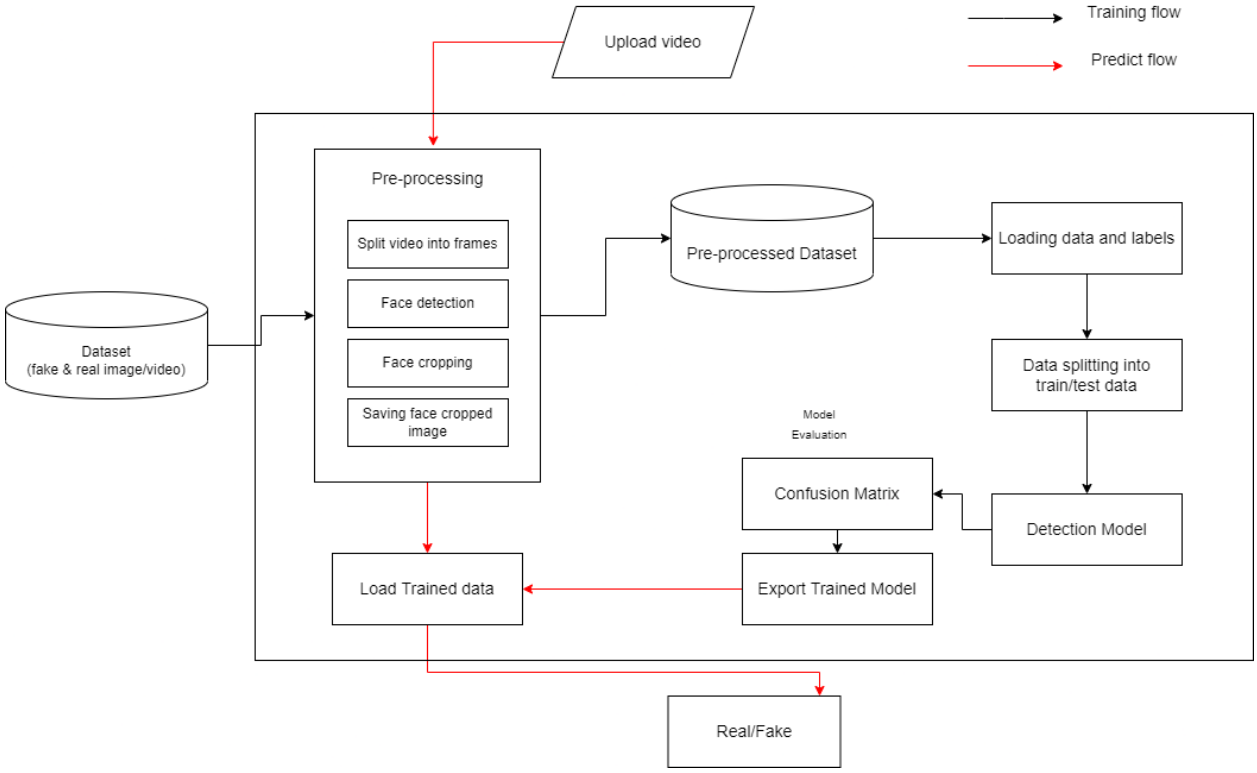
III) Giới thiệu hệ thống

1) Tổng quan

- Chúng em đã phát triển quy trình phát hiện các hình ảnh/ video deepfake một cách có hiệu quả nhất có thể bằng việc sử dụng công nghệ hoán đổi khuôn mặt của deepfake.
- Sau quá trình tiền xử lý, các khuôn mặt sẽ được cắt ra khỏi ảnh gốc từ tập dữ liệu, trong khi tập dữ liệu video ban đầu sẽ biến thành các khung hình, sau khi các khuôn mặt được cắt và gán label(nhãn) cho khuôn mặt đó label là 0 nếu đó là khuôn mặt giả còn label là 1 cho khuôn mặt thật, giúp giảm đáng kể khối lượng công việc của quá trình phát hiện và cải thiện hiệu quả.
- Bộ dữ liệu video được sử dụng là bộ Face Forensics(FF++) được sử dụng phổ biến để phát hiện video có là giả hay không, và bộ dữ liệu hình ảnh là 140k Real and Fake Faces. Việc đào tạo và đánh giá các mô hình được thực hiện trên hai bộ dữ liệu lớn này.

2) Thiết kế hệ thống

- Phương pháp của chúng em là tập trung vào việc phát hiện các loại deepfake ví dụ là Replacement deepfakes. Hình 3.1 biểu diễn cho kiến trúc của hệ thống

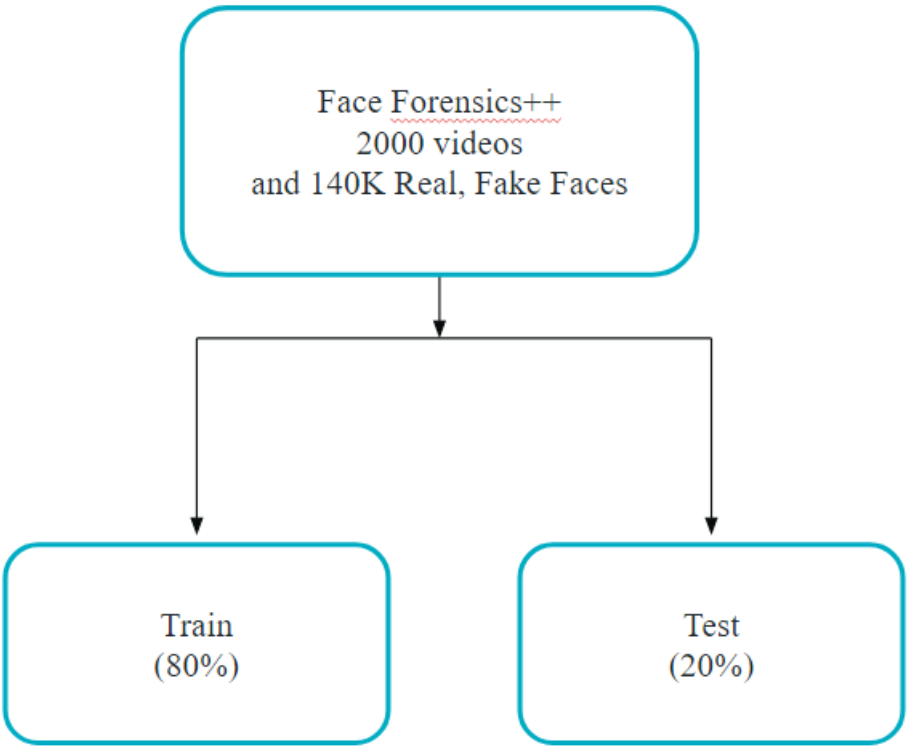


Hình 3.1: Kiến trúc hệ thống

3) Quá trình thực nghiệm

3.1) Dataset

- Tập dữ liệu video: chúng em sử dụng hỗn hợp bao gồm số lượng từ các nguồn dữ liệu khác nhau như youtube, FaceForensics++. Video dài trung bình là 10 giây với tốc độ khung hình là 30 khung hình /s
- Kích thước khung hình trung bình là 720*1280. Bộ dữ liệu của chúng em chứa 50% video gốc và 50% video deepfake. Tập dữ liệu chia thành 80% cho quá trình huấn luyện và 20% cho quá trình kiểm tra.
- Với tập dataset hình ảnh: tụi em sử dụng tập dataset 140k Real and Fake Faces bao gồm 70k hình ảnh khuôn mặt là thật từ Flickr dataset được thu nhập bởi Nvidia , và 70k hình ảnh khuôn mặt là giả từ the 1 Million FAKE faces(được tạo bởi Style GAN) .Trong tập dataset này, tất cả hình ảnh đều có cùng kích thước là 256 pixels. Tập dataset này cũng chia ra thành 80% lượng dataset cho train và 20% cho việc test.



Hình 3.2: Dataset

3.2) Tiền xử lý

Sau khi chọn được bộ dữ liệu cho quá trình training và xác định được model sẽ sử dụng, chúng em chuẩn bị bộ dữ liệu bằng cách xử lý chúng trước.

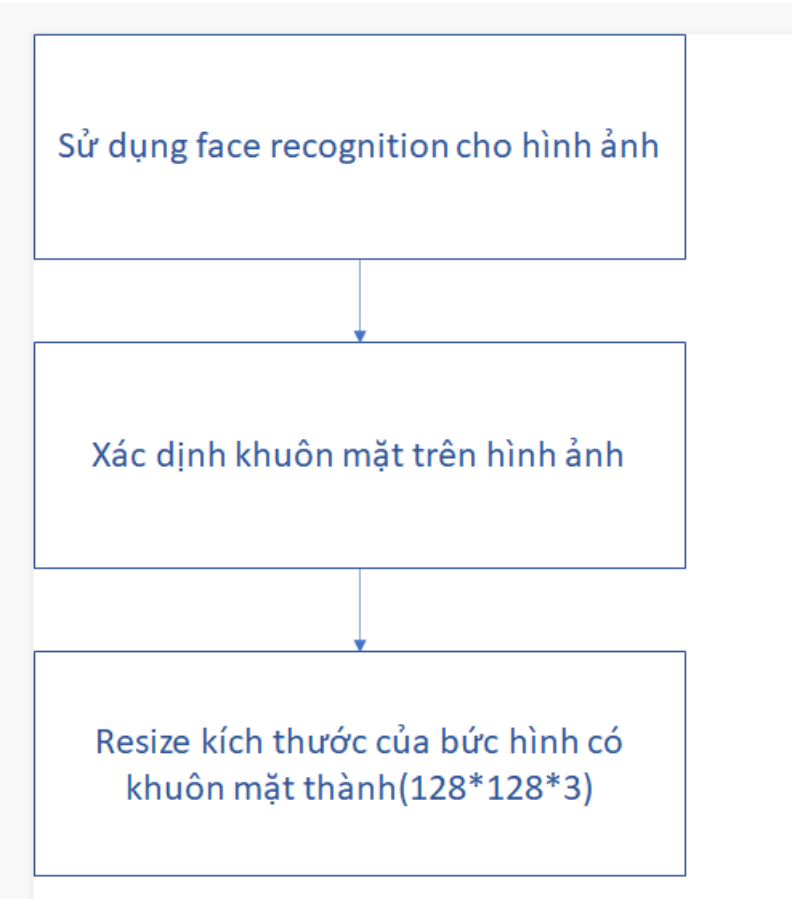
Đầu tiên, chúng em gán label cho tập dữ liệu. Sau đó, tập dữ liệu video sẽ được trích xuất thành các khung hình bằng việc sử dụng thư viện cv2.

Để nâng cáo được tính đa dạng và phong phú của tập dữ liệu ảnh cũng như việc tối ưu việc sử dụng bộ nhớ và thời gian chạy, chúng em chọn trích xuất một khung ảnh sau mỗi n khung hình khi video được chia thành các khung hình.(n là số lượng khung hình /1s của video)

Để cho model học một cách đầy đủ các khuôn mặt thật và giả, việc nhận dạng và cắt khuôn mặt được thực hiện trên hình ảnh/ khung hình của video. Để nhận diện thì em đã sử dụng thư viện Face Recognition.

Cuối cùng, ảnh khuôn mặt thật và giả đã được xử lý sẽ có cùng tỉ lệ là (128*128*3) và được đánh số lần lượt, được lưu vào 2 thư mục tương ứng là ảnh thật và ảnh giả.

Hình 3.3: Quá trình xử lý video



Hình 3.4: Quá trình xử lý hình ảnh

3.3) Model

3.3.1) EfficientNet

Phương pháp the multidimensional mixed model scaling (Seri mạng EfficientNet) do Google đề xuất vào năm 2019 đã thu hút sự chú ý của các chuyên gia. Mạng EfficientNet là một mạng tích chập giúp chia tỉ lệ đồng đều giữa tất cả các kích thước độ sâu/chiều rộng/độ phân giải bằng cách sử dụng hệ số gộp.

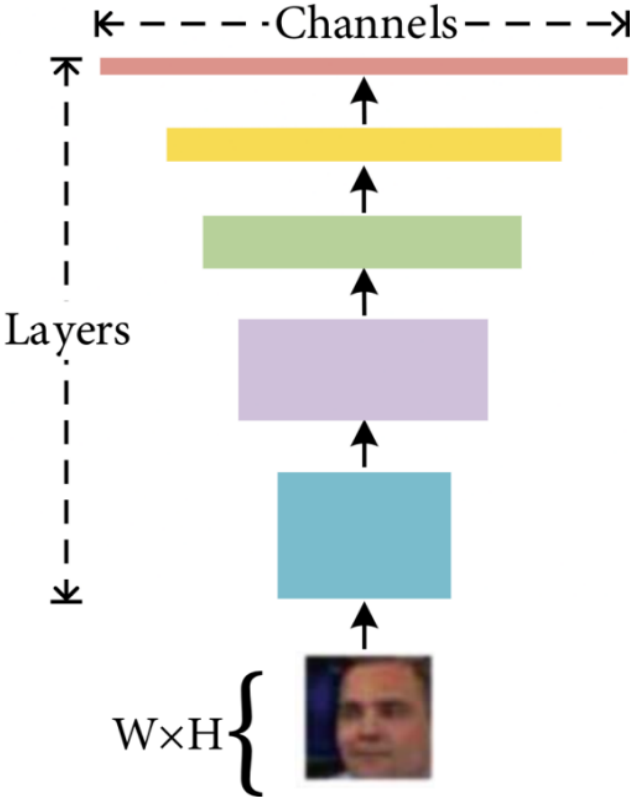
Như được biểu diễn trong hình 3.5 , hình 3.5(a) biểu thị mô hình mạng có cơ sở từ mạng tích chập. Đầu vào là hình ảnh 3 kênh màu có chiều rộng là W và chiều cao là H.

Sau khi tích chập từng lớp, model có thể tìm ra các đặc trưng tương ứng trong ảnh, Hính 3.5(b)-3.5(d) lần lượt biểu thị ba phương pháp phổ biến để mở rộng quy mô của model. Hình 3.5(b) là cải thiện model theo độ phân giải của hình ảnh đầu vào và cải thiện hiệu quả học của model theo tỷ lệ phóng to hoặc thu nhỏ kích thước của ảnh đầu vào. Hình 3.5(c) là cải thiện hiệu suất của model bằng cách thay đổi số lượng kênh của mỗi lớp model. Hình 3.5(d) là tăng hoặc

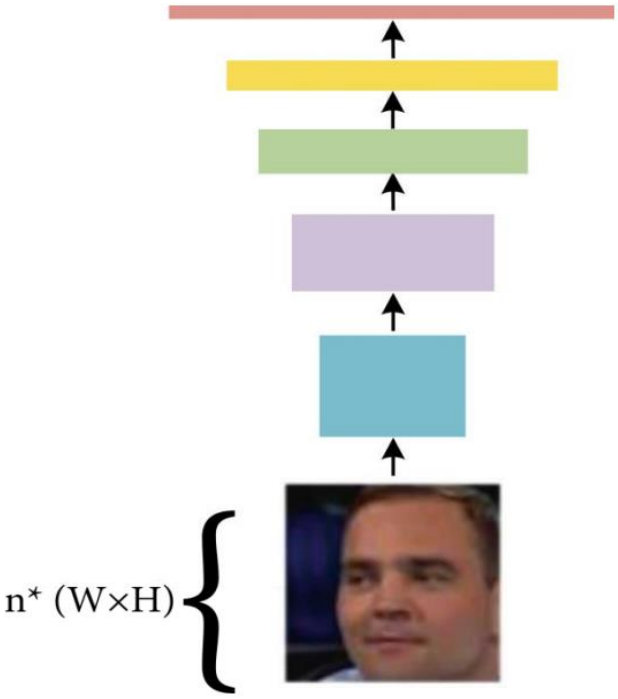
giảm số lớp của model để model có thể học đặc đặc trưng cụ thể và nâng cao độ hiệu quả của model.

Hình 3.5(e) cho thấy, trong EfficientNet, các tham số tổng hợp được sử dụng để thực hiện đồng thời việc chia tỷ lệ của 3 chiều trên, tử đó cải thiện hiệu suất tổng thể của mạng.

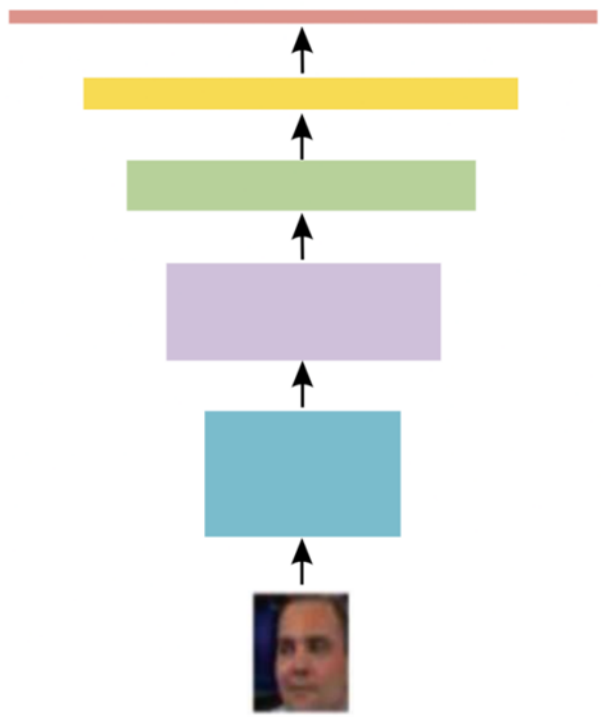
Hình 3.5: EfficientNet



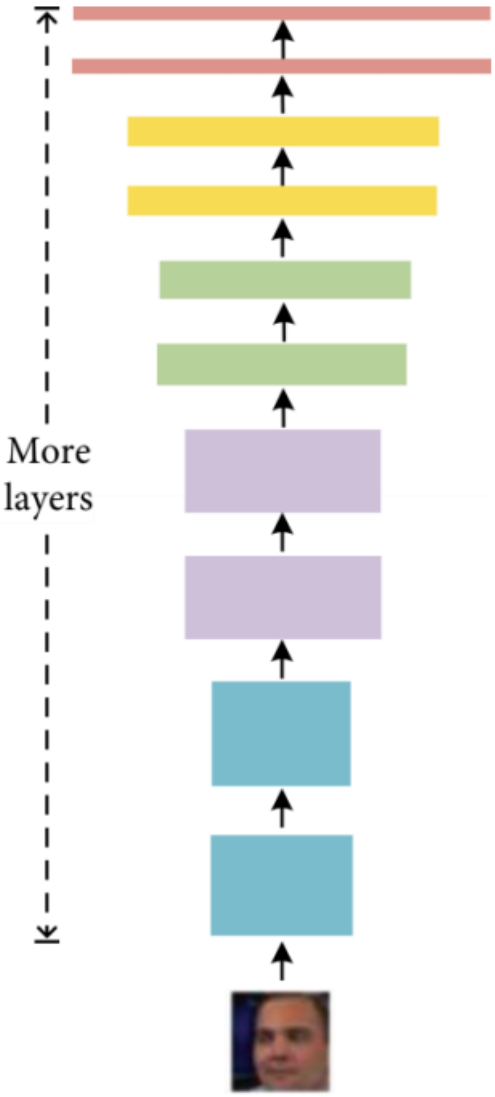
Hình 3.5(a)



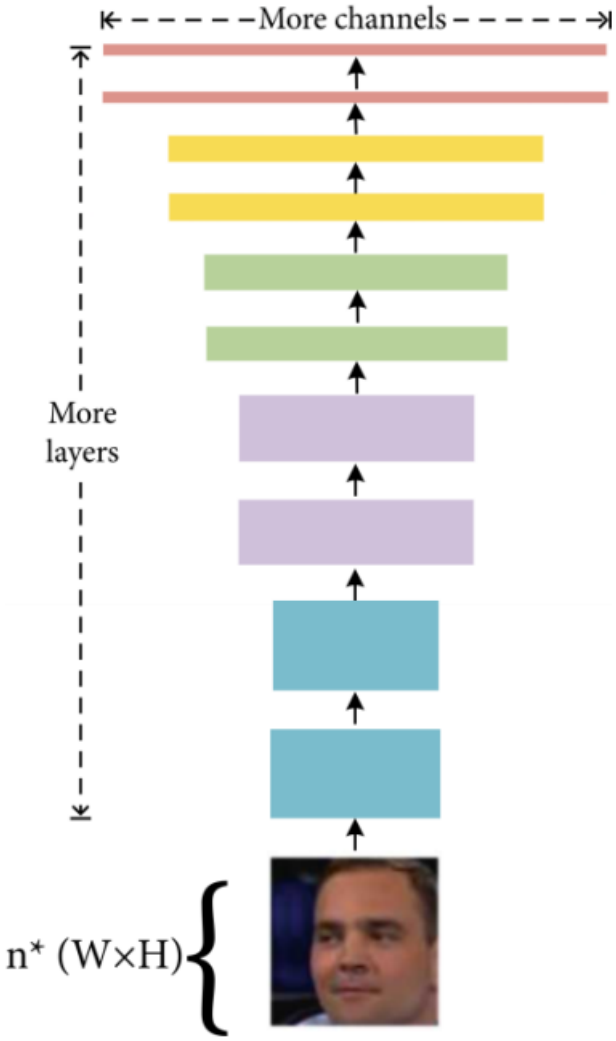
Hình 3.5(b)



Hình 3.5(c)



Hình 3.5(d)



Hình 3.5(e)

EfficientNet tương tự như mạng nơ-ron tích chập truyền thống. Bảng 3-1 cho thấy cấu trúc của EfficientNet-B0. Có thể thấy rằng mạng được chia thành 9 lớp.

Lớp thứ nhất là một lớp tích chập thông thường với kích thước của kernel là 3*3 và stride là 2, bao gồm chuẩn hóa toàn bộ(Batch Normalization) và hàm kích hoạt.

Lớp 2 đến lớp 8 đều là có cấu trúc MBConv xếp chồng lên nhau lặp đi lặp lại. Các lớp trong cột cuối cùng của bảng cho biết số lần lặp lại cấu trúc MBConv, trong khi lớp 9 bao gồm 1 lớp tích chập 1*1 thông thường, một lớp average pooling và một lớp fully-connected bao gồm cả chuẩn hoán toàn bộ và hàm kích hoạt.

Trong bảng 3.1 , mỗi MBConv sẽ được theo sau mợi một số là 1 hoặc 6 trong đó 1 hoặc 6 là hệ số nhân n, nghãi là lớp tích chập 1*1 đầu tiên của

MBConv sẽ được mở rộng các kênh của ma trận đặc trưng đầu vào lên n lần, trong đó k3*3 hoặc k5*5 biểu thị kích thước của kernel tích chập được Depthwise Conv sử dụng trong MBConv.

Các kênh đại diện cho các kênh xuất ra ma trận đặc trưng sau khi qua các lớp.

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

Bảng 3.1: Cấu trúc của EfficientNet-B0

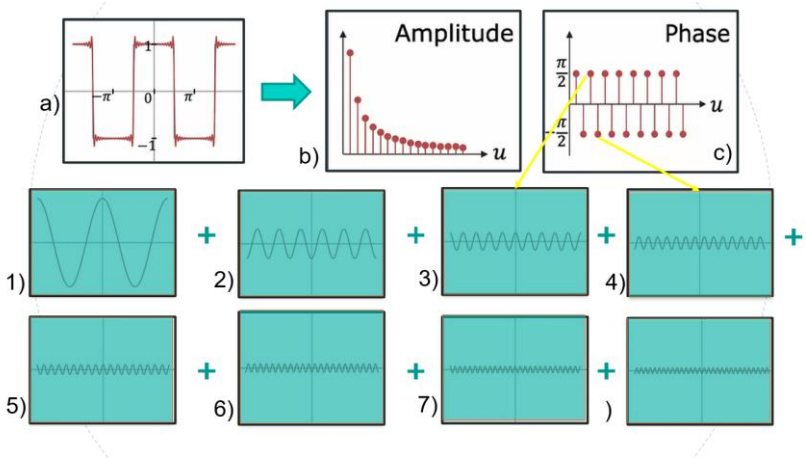
Biểu thức của hàm kích hoạt Switch được biểu diễn trong công thức sau(1), trong đó là hằng số hoặc là một tham số có thể huấn luyện. Hàm kích hoạt Switch có đặc điểm là có giới hạn dưới, mượt và không đơn điệu.

$$f(x) = x \cdot sigmoid(\beta x) \quad (1)$$

3.3.2) DFT + SVM

a) Feature extraction (DFT)

Fourier Transform (FT – Biến đổi Fourier) cho phép hàm không tuần hoàn có thể biểu diễn thành tích của sin hoặc cosin nhân với hàm trọng số. Đặc điểm riêng của Fourier Transform là làm việc trên miền tần số (đặc trưng bởi Amplitude: Biên độ và Phase: Pha của hàm số)



Hình 3.6: Biến đổi Fourier

Như hình 3.6, với hàm số ban đầu (a) chúng ta có thể biểu thị thông qua biến đổi Fourier bằng tổng các hàm (1), (2), (3), (4), (5), (6), (7), (8), …. Với mỗi hàm có 2 đặc trưng được biểu diễn tương ứng trong hàm (b) (đặc trưng Biên độ) và hàm (c) (đặc trưng Pha)

Ta có thể thấy nếu tổng các hàm (1), (2), (3), (4), (5), … càng chi tiết thì có thể tương đương với hàm ban đầu (a). Tuy nhiên chỉ cần một số hàm đầu tiên: (1), (2), (3), (4) chúng ta vẫn có thể mô tả hết những đặc trưng quan trọng của hàm (a).

Tuy nhiên do ảnh là hàm không tuần hoàn (giới hạn bởi độ rộng và độ cao ảnh) và các giá trị pixel trên ảnh là riêng lẻ nhau và ảnh được biểu diễn theo độ rộng và độ cao vì vậy em sử dụng 2D DFT (2D Discrete Fourier Transform) được biểu diễn bằng công thức:

2D Discrete Fourier Transform:

$$F[p,q] = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f[m,n]e^{-i2\pi pm/M}e^{-i2\pi qn/N} \qquad \begin{matrix} p = 0 \ldots M-1 \\ q = 0 \ldots N-1 \end{matrix}$$

$\boldsymbol{p}$, $\boldsymbol{q}$ tương ứng là giá trị tần số dọc theo $\boldsymbol{M}, \boldsymbol{N}$
với ảnh đầu vào có kích thước $\boldsymbol{M} \times \boldsymbol{N}$

2D Inverse Discrete Fourier Transform:

$$f[m,n] = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} F[p,q]e^{i2\pi pm/M}e^{i2\pi qn/N} \qquad \begin{matrix} \text{m}= 0 \ldots M-1 \\ n = 0 \ldots N-1 \end{matrix}$$

Hình 3.7: 2D DFT

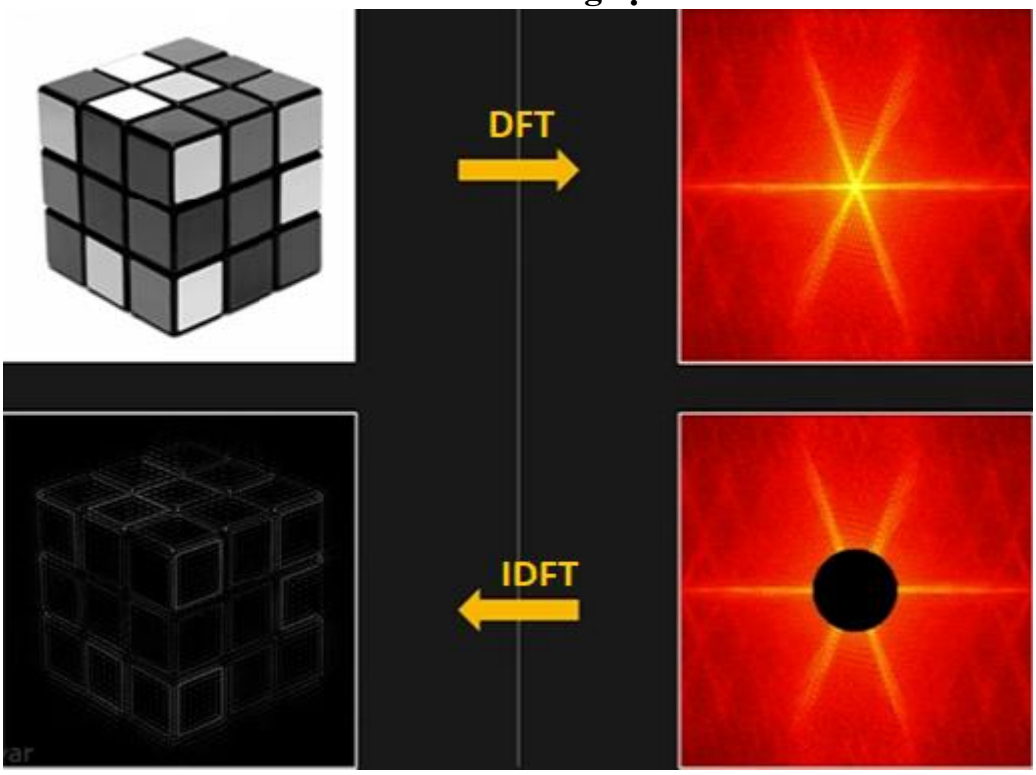
Trong đó 2D DFT (2D Discrete Fourier Transform) và 2D IDFT (2D Inverse Discrete Fourier Transform) được sử dụng 2 quá trình trong quá trình:

Hình 3.8.1: Biến đổi 2D DFT

Trong đó hàm $f(x)$ được chuyển sang miền tần số được biểu diễn dưới dạng $F(u)$

Sau khi sử dụng các phương pháp trích chọn đặc trưng ở $F(u)$ chúng ta sẽ dùng 2D Inverse DFT để trả về $f(x)$ chỉ còn lại những đặc trưng quan trọng:

Hình 3.8.2: Biến đổi ngược 2D DFT



Hình 3.9: Bộ lọc cạnh dùng 2D DFT

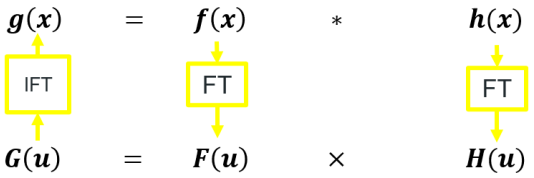
Các lợi ích của DFT:

- Các bộ lọc (Filtering) dễ dàng biểu thị (tần số cao, tần số thấp) hơn trong miền tần số (Hình 3.9)

- Phép toán convolution trong miền tần số đơn giản hơn trong miền không gian thông thường vì trong miền tần số, phép toán convolution tương tương với phép nhân ma trận:

Convolution và 2D DFT

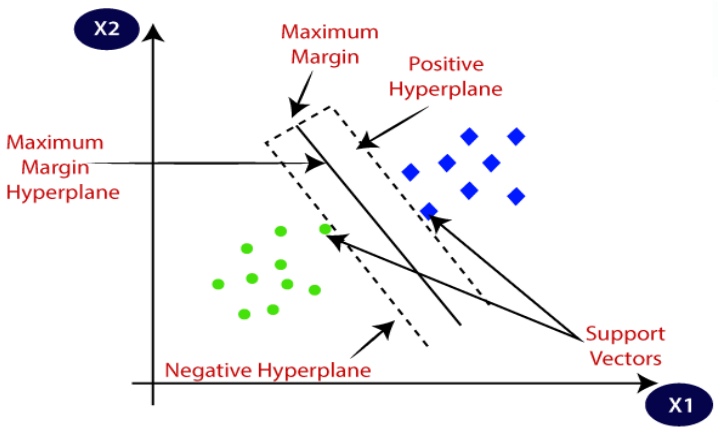
Miền không gian	Miền tần số
$g(x) = f(x) * h(x)$ *Phép tính chập*	$G(u) = F(u)\,H(u)$ *Phép nhân*
$g(x) = f(x)\,h(x)$ *Phép nhân*	$G(u) = F(u) * H(u)$ *Phép tính chập*



Hình 3.10: Convolution và 2D DFT

b) Support vector machine (SVM)

Support vector machine là thuật toán học có giám sát, thuật toán tìm ra một Hyperplane để phân chia các lớp dữ liệu thành các phần riêng biệt, bằng cách đi tìm Hyperplane phân cách sao cho margin giữa Hyperplane đó và các support vectors là lớn nhất.



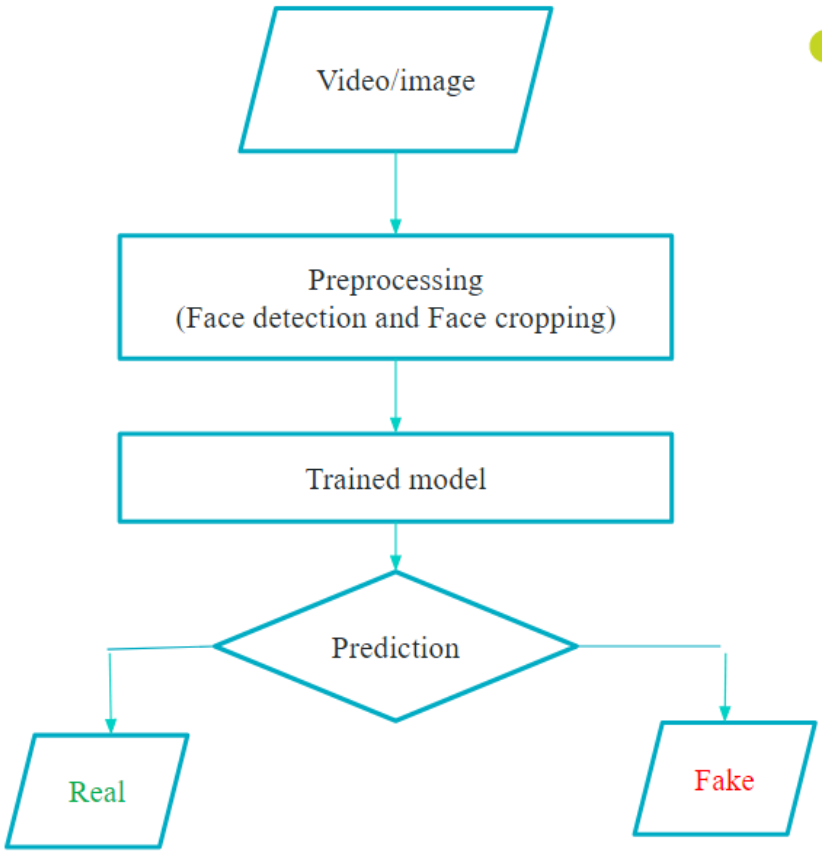
Hình 3.11: Support Vector Machine

Đặc biệt SVM cho kết quả tốt trên bài toán Binary Classification, vì vậy đối với dữ liệu của em chỉ gồm 2 label Real và Fake nên SVM thích hợp dùng làm bộ Classification cho bài toán.

Sau qua nhiều lần training với các tham số của model, và tham khảo nhiều bài toán sử SVM có liên quan, em chọn sử dụng $kernel = 'rbf'$, $C = 6{,}37$, $gamma = 0{,}86$.

4) Dự đoán

Như trong hình 3.6, một video/ hình ảnh được đưa vào model(đã được trian) để dự đoán . Video cũng được xử lý trước sau đó đưa tới model để dự đoán. Video được chia thành các khung hình sau đó là nhận dạng khuôn mặt, sau đó cắt khuôn mặt đó ra và thay vì lưu trữ video vào bộ nhớ thì các khung hình được cắt sẽ được đưa vào model để dự đoán. Dự đoán video bằng cách dự đoán tất cả các khung hình của video và nếu số lượng khung hình thật lớn hơn thì predict là real còn không thì ngược lại. Trong trường hợp là hình ảnh , thì coi nó là một video khung hình đơn.



Hình 3.12 Quá trình dự đoán

5) Kết quả

a) EfficientNet

Lúc đầu thì em sử dụng ResNet nhưng độ chính xác thấp và tiêu tốn quá nhiều tài nguyên, sau đó em tìm tòi xem model nào hiệu quả vì

các lý do ở trên em đã chọn EfficientNet trên bộ dataset em đã trình bày ở trên và nó đạt được kết quả như em mong muốn.

Deep learning Model	Dataset	Accuracy
ResNet	FaceForensics++	47.3%
EfficientNetB0	140k real and fake faces	92.45%
	FaceForensics++	91.32%
	Combination of 140k real and fake faces and FaceForensics++	96.77%

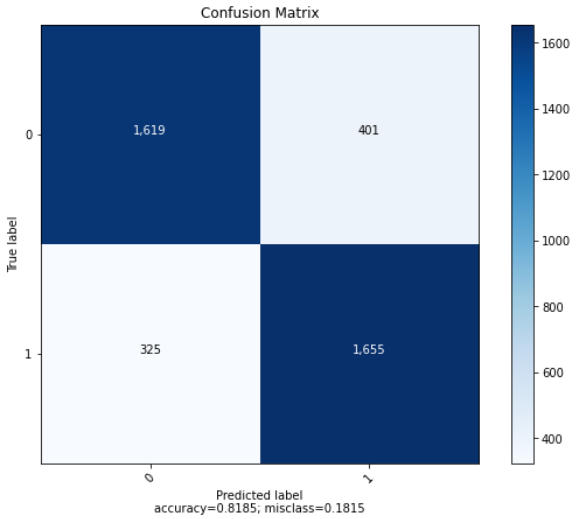
Hình 3.13: Kết quả thử nghiệm trên EfficientNet

b) DFT + SVM

- Accuracy:

Hình 3.14: Accuracy thử nghiệm trên DFT + SVM

- Confusion matrix:



Hình 3.15: Confusion matrix trên DFT + SVM

IV) KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1) EfficientNet

1.1) Kết luận

- Model dự đoán khá chính xác với những người châu Âu và châu Mỹ.
- Tuy là dự đoán với độ chính xác trên tập test nhưng khi lấy các bức ảnh bên ngoài tập dataset thì kết quả chỉ ở mức tạm được.
- Model dự đoán với những video sử dụng deepfake tiên tiến thì predict tương đối đúng.

1.2) Hướng phát triển

- Mở rộng Dataset (mặt người từ nhiều quốc gia, dùng Deepfake công nghệ cao để tạo thêm fake data)
- Thêm tính năng xác định khuôn mặt nào trong video sử dụng công nghệ deepfake còn khuôn mặt nào trong video là thật.

2) DFT + SVM

2.1) Kết luận

- Tuy model SVM đạt accuracy: 81,85% nhưng trong quá trình thực nghiệm, em nhận thấy model có xu hướng cho accuracy cao ở những khuôn mặt có những điểm nhòe, bất thường dễ dàng quan sát được bằng mắt.
- Do đó khi detect những video/ ảnh deepfake dùng kỹ thuật cao, model đạt kết quả không được hiệu quả

2.2) Hướng phát triển

- Mở rộng dataset với nhiều khuôn mặt từ nhiều vùng quốc gia, khu vực ví dụ: Châu Á, Châu Phi (vì dataset của chúng em đa phần là khuôn mặt Châu Âu).
- Sử dụng lại các mô hình deepfake kỹ thuật cao để tạo data fake trong quá trình training model.

V) TÀI LIỆU THAM KHẢO

[1] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, Janis Keuper (2019), *Unmasking DeepFakes with simple Features*, Data and Web Science Group, University Mannheim, Germany: https://arxiv.org/pdf/1911.00686.pdf

[2] Ngô Quốc Việt (2012), BÀI GIẢNG XỬ LÝ ẢNH SỐ - 5. XỬ LÝ ẢNH TRONG MIỀN TẦN SỐ, TPHCM: https://www.thuvientailieu.vn/tai-lieu/bai-giang-xu-ly-anh-so-5-xu-ly-anh-trong-mien-tan-so-41775/

[3] Trung Nguyen (2022), Xử lý miền tần số: https://www.academia.edu/32457117/Xử_lý_trong_miền_tần_số

[4] Nguyễn Thùy Dương (2017), Biến đổi Fourier rời rạc(DFT) trong nhận diện mặt người sử dụng Matlab, Viblo: https://viblo.asia/p/bien-doi-fourier-roi-racdft-trong-nhan-dien-mat-nguoi-su-dung-matlab-YWOZrM0EKQ0

[5] First Principles of Computer Vision (2021), Image Processing I| Image Processing II, link Youtube: https://www.youtube.com/playlist?list=PL2zRqk16wsdorCSZ5GWZQr1EMWXs2TDeu

[6] Deepfake detection challenge: https://www.kaggle.com/c/deepfake-detection-challenge/data

[7] FaceForensics: https://github.com/ondyari/FaceForensics

[8] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu "Deep Video Portraits" in arXiv:1901.02212v2: https://www.ijraset.com/research-paper/fake-face-detection-using-cnn

[9] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin "Detection of Synthetic Portrait Videos using Biological Signals" in arXiv:1901.02212v2: https://www.researcher-app.com/paper/3359070

[10] F. Song, X. Tan, X. Liu, and S. Chen, "Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients," Pattern Recognition, vol. 47, no. 9, pp. 2825–2838, 2014: https://link.springer.com/article/10.1007/s44196-022-00108-2

[11] R. Raghavendra, Kiran B. Raja, Sushma Venkatesh, and Christoph Busch, "Transferable deep-CNN features for detecting digital and print-scanned morphed face images," in CVPRW. IEEE, 2017: https://ieeexplore.ieee.org/document/8014962

[12] Yuezun Li, Siwei Lyu, “ExposingDF Videos By Detecting Face Warping Artifacts,” in arXiv:1811.00656v3: https://www.ijraset.com/best-journal/deepfake-detection-a-survey-of-countering-malicious-deepfake

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

[14] Y. Qian et al. Recurrent color constancy. Proceedings of the IEEE International Conference on Computer Vision, pages 5459–5467, Oct. 2017. Venice, Italy: https://www.isroset.org/journal/IJSRCSE/full_paper_view.php?paper_id=2342

[15] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018: https://gangw.web.illinois.edu/class/cs598/papers/AVSS18-deepfake.pdf

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014: https://papers.nips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html