

FAS sử dụng YOLOv9 và Resnet+Att

Tôn Nữ Mai Khanh
Nguyễn Khánh Huyền

Đại học Khoa học Tự nhiên - ĐHQGHN

12/2024



Outline

- 1 Giới thiệu
- 2 Cơ sở lý thuyết
- 3 Phương pháp đề xuất
- 4 Quá trình huấn luyện
- 5 Kết quả thực nghiệm
 - Đánh giá mô hình
 - So sánh mô hình
 - Ứng dụng
- 6 Tài liệu tham khảo

Đặt vấn đề

- Face anti-spoofing là yếu tố then chốt để bảo vệ các hệ thống nhận diện khuôn mặt khỏi các cuộc tấn công giả mạo như ảnh in, video phát lại, mặt nạ 3D, hay deepfake.
- Khi công nghệ nhận diện khuôn mặt được sử dụng rộng rãi, nguy cơ bị xâm nhập ngày càng cao.
- Dưới đây là minh họa cho thấy các ảnh khuôn mặt đã được cắt trong các trường hợp thật - giả.



(a) Mặt thật



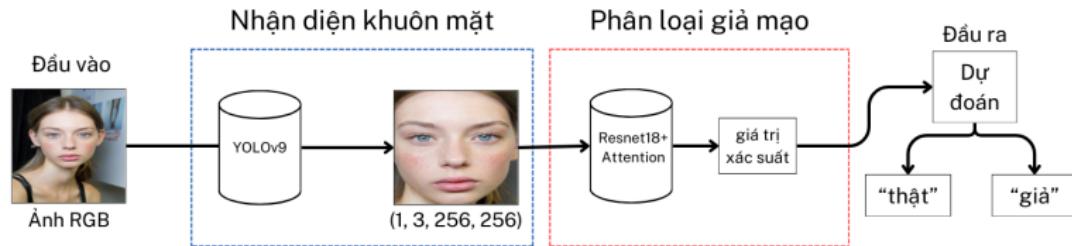
(b) Mặt giả mạo

Tổng quan đề tài

- Các phương pháp và mô hình nhận diện khuôn mặt và phân loại thật giả dựa trên CNN truyền thống chưa đủ tốt đối với các thách thức giả mạo hiện nay
- Nhóm em đề xuất phương pháp kết hợp:
 - Mô hình YOLOv9 [1]: Nhận diện khuôn mặt.
 - Mô hình Resnet18 kết hợp cơ chế Attention [2]: Phân loại mặt thật - giả.

Quy trình tổng thể

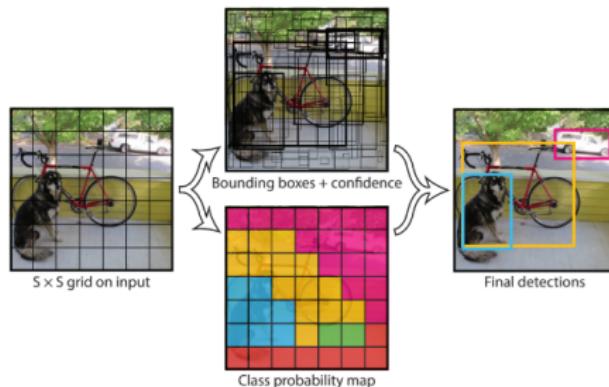
Quy trình tổng thể do chính nhóm tự đề xuất với ý tưởng dựa trên bài báo nghiên cứu khoa học DeePixBis [3].



Hình: Quy trình chính

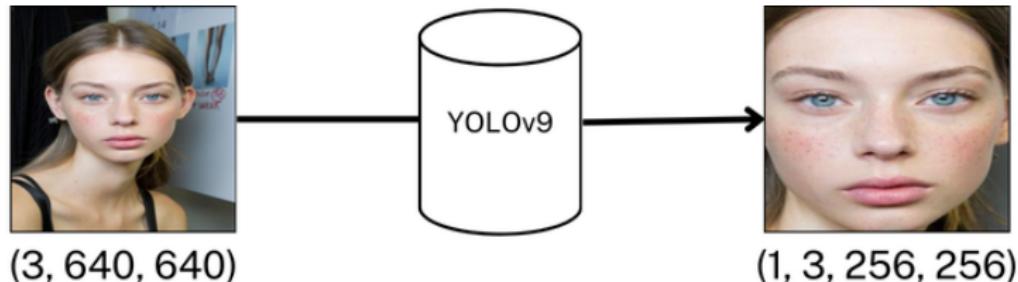
- Sử dụng mô hình YOLO để nhận diện khuôn mặt.
- Với output từ mô hình YOLO, sử dụng ResNet18 kết hợp Attention để phát hiện giả mạo.

Kiến trúc YOLO



- YOLO (You Only Look Once) [4] là một kiến trúc mạng nơ-ron tích chập (CNN) tiên tiến.
- Đầu vào của mô hình là một ảnh, mô hình sẽ nhận dạng ảnh đó có đối tượng nào hay không, sau đó sẽ xác định tọa độ của đối tượng trong bức ảnh.

Nhận diện khuôn mặt bằng YOLOv9



Hình: Quá trình huấn luyện YOLOv9

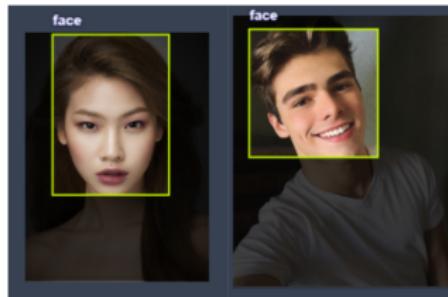
- Sử dụng thư viện Ultralytics, trọng số pre-train là yolov9c.pt
- Ảnh gốc được cắt theo vùng bounding box và điều chỉnh kích thước về 256x256.

Dữ liệu huấn luyện

Dữ liệu huấn luyện YOLOv9 là Face Recognition for CV (phiên bản 1):

Đã tăng cường dữ liệu: lật ngang, xoay ảnh (từ -15° đến $+15^\circ$), chỉnh độ bão hoà (-25% đến 25%)

Gồm 2400 ảnh các ảnh trong bộ dữ liệu được gán nhãn với một nhãn duy nhất là "face".



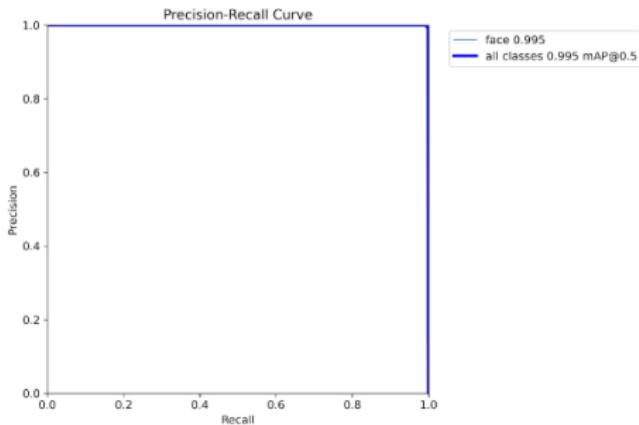
	Số lượng mẫu
Tập huấn luyện	2100
Tập xác minh	200
Tập kiểm tra	100

Huấn luyện mô hình

- Huấn luyện YOLOv9: Được huấn luyện trên colab T4 GPU.
Quá trình huấn luyện YOLOv9 được chạy qua 50 epoch với thời gian huấn luyện là 1.435 tiếng.

Đánh giá mô hình YOLOv9

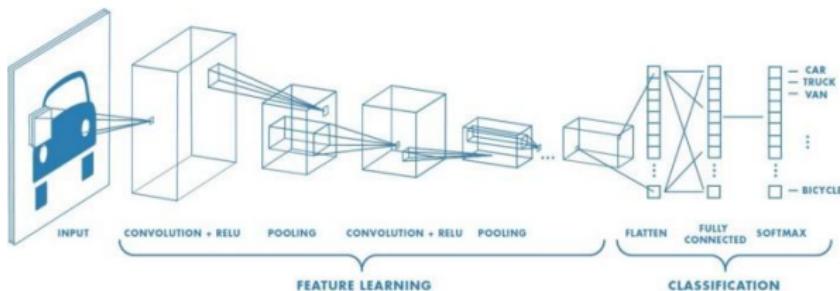
Precision	Recall	mAP50	mAP50-95
0.994	1	0.995	0.864



Mô hình YOLOv9 đã huấn luyện có thể cho ra kết quả dự đoán tốt trên tập xác minh.

Kiến trúc Resnet18

Mạng nơ ron tích chập (CNN)



Hình: Luồng CNN

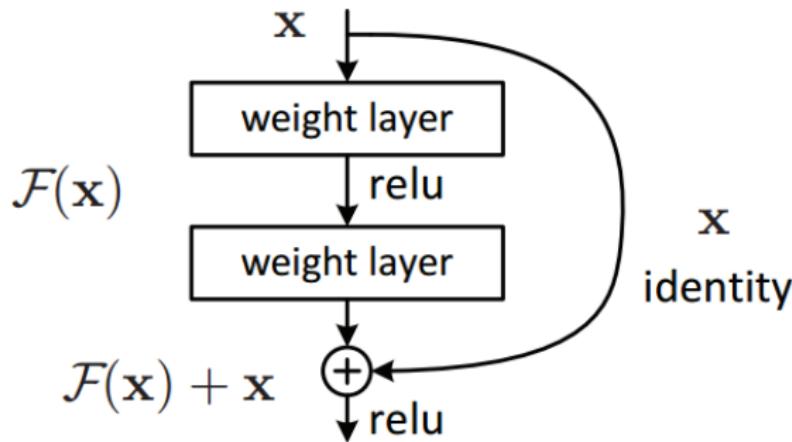
- Trong mạng neural, mô hình mạng neural tích chập (CNN) là 1 trong những mô hình để nhận dạng và phân loại hình ảnh.
- Mỗi hình ảnh đầu vào sẽ được chuyển qua 1 loạt các lớp tích chập với các bộ lọc, tổng hợp lại các lớp được kết nối đầy đủ (Full Connected) và áp dụng hàm Softmax để phân loại đối tượng có giá trị xác suất giữa 0 và 1.

Mạng ResNet

- Với các mạng nơ ron sâu, khi độ sâu của mạng tăng lên, sẽ xảy ra vấn đề Vanishing Gradient, làm cho quá trình huấn luyện trở nên khó khăn và không ổn định.
- Hiện tượng Vanishing Gradient khiến các lớp ở sâu không nhận được cập nhật đủ mạnh để cải thiện trọng số, dẫn đến mô hình không hội tụ tốt.
- ResNet giải quyết vấn đề này bằng cách sử dụng các kết nối tắt (skip connections), bổ sung đầu vào ban đầu ($\text{input } x$) vào đầu ra của một lớp nhất định.

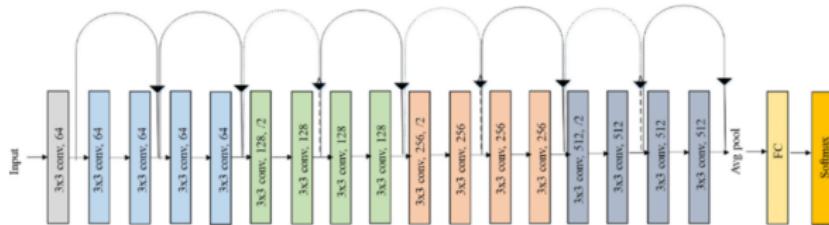
Kiến trúc Resnet18

Mạng ResNet



Hình: Residual Block trong ResNet

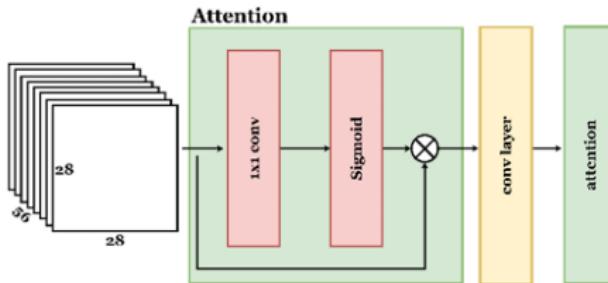
Kiến trúc ResNet18



Hình: Kiến trúc gốc ResNet18 [5]

- ResNet sử dụng các khối dư (Residual Blocks) để giúp mạng học các đặc trưng hiệu quả hơn.
- Resnet18 có tổng cộng 18 lớp: 17 lớp tích chập, lớp fully-connected, và lớp softmax bổ sung để thực hiện nhiệm vụ phân loại.

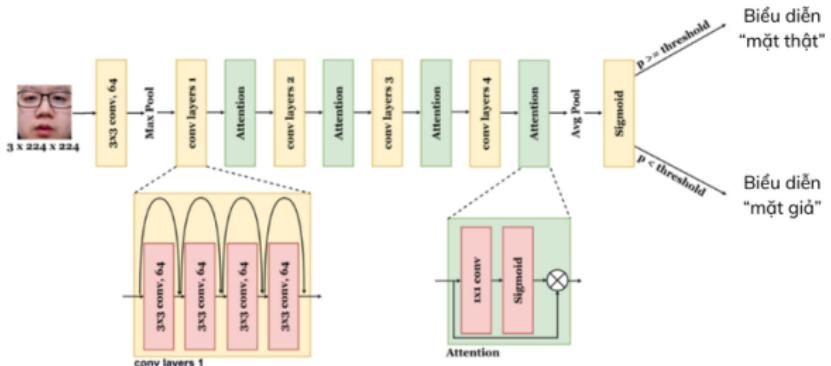
Cơ chế Self-Attention



Hình: Cơ chế Attention đơn giản [2]

- Cơ chế cho phép mô hình tập trung vào các phần quan trọng của hình ảnh, giúp cải thiện hiệu suất của các tác vụ như phân loại, nhận dạng và phát hiện đối tượng.
- Cơ chế self-attention hoạt động bằng cách tính toán sự tương quan giữa các phần khác nhau của hình ảnh. Điều này giúp mô hình hiểu được mối quan hệ giữa các phần của hình ảnh và tập trung vào những phần quan trọng nhất.

Phân loại giả mạo bằng Resnet18+Att



Hình: Kiến trúc cho phương pháp đề xuất Resnet18 kết hợp Attention [6]

- Đầu vào: Ảnh kích thước $3 \times 224 \times 224$ đã qua xử lý.
- Đầu ra: Giá trị xác suất p trong khoảng $[0,1]$

Dữ liệu huấn luyện

Dữ liệu huấn luyện Resnet18+Attention là NUAA Photograph

Imposter: Được gán 2 nhãn là client (mặt thật) và imposter (mặt giả mạo).



	Mặt thật	Mặt giả mạo
Tập huấn luyện	1743	1748
Tập kiểm tra	3362	5761

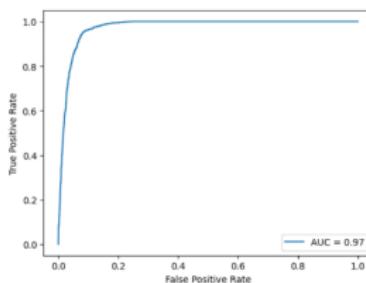
*

Huấn luyện mô hình

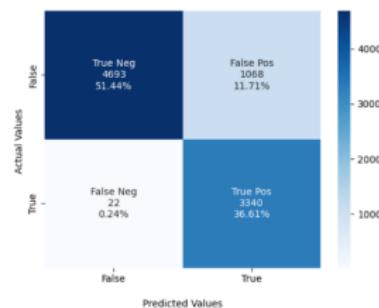
- Huấn luyện ResNet18+Attention: Resnet18+Attention được huấn luyện trên Kaggle, CUDA T4 GPU
- Huấn luyện mô hình ResNet18+Attention với 100 epoch trong thời gian là 6 tiếng.

Đánh giá mô hình Resnet18+Attention

Độ đo	Kết quả
ACER	0.09596410291730659
FRR	0.006543723973825104
FAR	0.18538448186078807
ROC AUC score	0.9728075230676312
EER	0.07255684776948447
F1 score	0.8597168597168597
Accuracy	0.8805217581935767



(a) Đường cong ROC



(b) Ma trận nhầm lẫn

So sánh mô hình YOLOv9 và phương pháp Haar Cascade

Dữ liệu:

- ① Tập kiểm tra Face Recognition: 100 mẫu với 100 box "face".
- ② Tập dữ liệu Face Detection Dataset trên Kaggle: 800 mẫu với 2323 box "face".

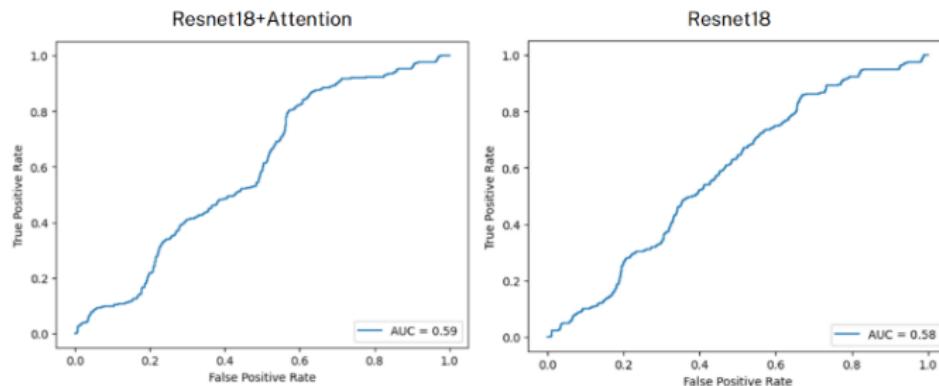
Face Recognition				Face Detection		
	Số box được dự đoán	IOU trung bình	Thời gian chạy (s)	Số box được dự đoán	IOU trung bình	Thời gian chạy (s)
YOLOv9	100	0.8917	107.87	705	0.4607	343.44
Haar Cascade	96	0.5566	46.90	1491	0.4750	541.94

- YOLOv9 vượt trội trong nhận diện khuôn mặt đơn lẻ với độ chính xác IOU cao và khả năng khớp nhãn hoàn hảo.
- Hiệu suất giảm khi xử lý ảnh phức tạp với nhiều khuôn mặt cần nhận diện đầy đủ.

So sánh mô hình Resnet18+Attention và Resnet18

Dữ liệu LCC_FASD: gồm 389 mẫu "client" và 3377 "imposter"

Mô hình	FRR	FAR	ACER
Resnet18+Attention	0.077	0.799	0.438
Resnet18	0.139	0.698	0.418



Hình: Biểu đồ ROC hai mô hình

Demo camera

- Đầu vào: Luồng video trực tiếp từ camera.
- Đầu ra: Hiển thị video có bounding box, nhãn (Fake/Real) và xác suất.

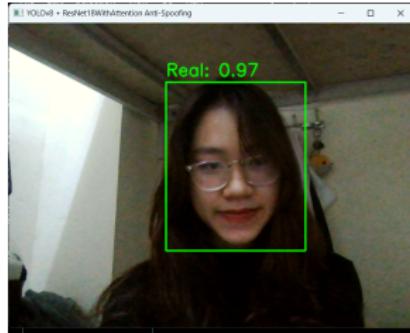
Quy trình:

- ① YOLOv9 phát hiện khuôn mặt, trả về bounding box.
- ② Cắt và xử lý vùng khuôn mặt (resize, crop, normalize).
- ③ Khuôn mặt đã xử lý được đưa vào mô hình ResNet18+Attention để dự đoán thật/giả.

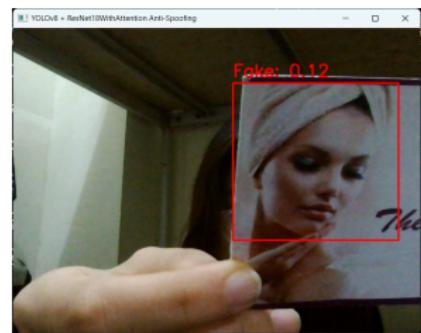
Demo camera



(a) Phát hiện ảnh giả
mạo (1)



(b) Phát hiện với ảnh
thật



(c) Phát hiện ảnh giả
mạo (2)

Tài liệu tham khảo

-  C.-Y. Wang, I.-H. Yeh, H.-Y. M. Liao. Yolov9: Learning what you want to learn using programmable gradient information[C]. , 2024,
-  A. Pikul. A method for improving presentation attack detection in biometric face recognition systems using a convolutional neural network with an attention mechanism[C]. , ,
-  A. George, S. Marcel. Deep pixel-wise binary supervision for face presentation attack detection[C]. , 2019,
-  J. Redmon, S. Divvala, R. Girshick, et al. You only look once: Unified, real-time object detection[C]. , 2016,
-  F. Ramzan, M. U. Khan, A. Rehmat, et al. A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks[J]. Journal of Medical Systems, 2019, 44:
-  A. Pikul. A method for improving presentation attack detection in biometric face recognition systems using a convolutional neural network with an attention mechanism[C]. , ,

Thank you!

Cảm ơn thầy và các bạn đã lắng nghe