

BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học

**CS519 - PHƯƠNG PHÁP LUẬN
NGHIÊN CỨU KHOA HỌC**

Lớp học

CS519.011

Giảng viên

PGS.TS. LÊ ĐÌNH DUY

Thời gian



09/2023 - 02/2024

THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo:

<https://www.youtube.com/watch?v=VWJKArE2rW0>

- Link slides: <https://github.com/NgKhTr/CS519.O11/blob/main/Slides.pdf>

<ul style="list-style-type: none">• Họ và Tên: Phan Tiến Quân• MSSV: 21522502 	<ul style="list-style-type: none">• Lớp: CS519.O11• Tự đánh giá (điểm tổng kết môn): 9/10• Số buổi vắng: 0• Số câu hỏi QT cá nhân: 11/11• Link Github: https://github.com/NgKhTr/CS519.O11.git• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng cho đề tài○ Làm poster○ Làm slide
<ul style="list-style-type: none">• Họ và Tên: Nguyễn Khánh Trình• MSSV: 21522717 	<ul style="list-style-type: none">• Lớp: CS519.O11• Tự đánh giá (điểm tổng kết môn): 9/10• Số buổi vắng: 1• Số câu hỏi QT cá nhân: 11/11• Link Github: https://github.com/NgKhTr/CS519.O11.git• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng cho đề tài○ Viết đề cương○ Làm video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

ÁP DỤNG CƠ CHẾ ĐA CHÚ Ý VÀO NHẬN DẠNG ẢNH KHUÔN MẶT
DEEPPFAKE

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

USING MULTI-ATTENTION MECHANISM TO THE DEEPPFAKE FACE IMAGE
RECOGNITION

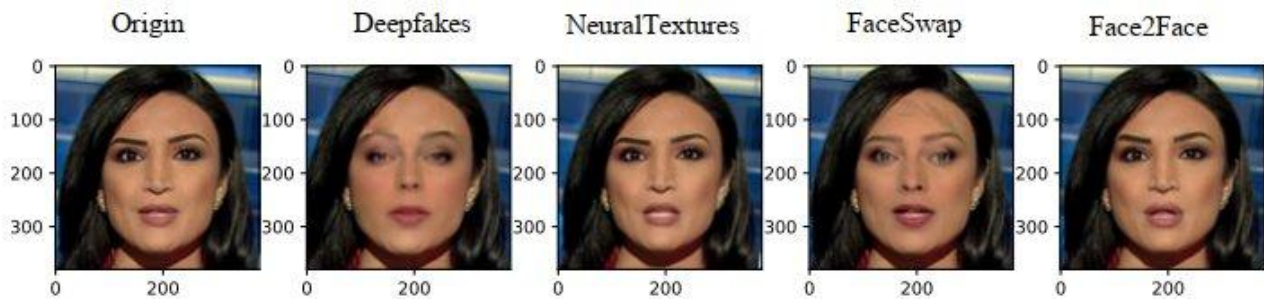
TÓM TẮT

Hiện nay bài toán nhận dạng ảnh khuôn mặt deepfake đang được cấp thiết nghiên cứu do những vấn đề về xã hội, chính trị ngày càng nghiêm trọng gây ra bởi các thành phần độc hại sử dụng mô hình deepfake trong việc giả mạo khuôn mặt. Chúng tôi xem xét bài toán này dưới dạng bài toán phân loại ảnh chi tiết (fine-grained image classification) thay vì bài toán phân loại ảnh nhị phân thông thường vì các chi tiết khác biệt giữa ảnh khuôn mặt thật và ảnh khuôn mặt tạo bằng deepfake ngày càng tinh vi, khó phân biệt bằng mắt thường. Từ đó, chúng tôi đề xuất áp dụng cơ chế đa chú ý (multi-attention mechanism) vào mô hình giải quyết bài toán với mục đích cung cấp cho mô hình khả năng chú ý tới các vùng cục bộ khác nhau chứa thông tin quan trọng cho việc phân loại. Hơn nữa, chúng tôi đề xuất cần có một phương pháp huấn luyện mới để đảm bảo mô hình chú ý tới những vùng độc lập khác nhau (multi) trong ảnh. Chúng tôi sẽ nghiên cứu tìm mô hình tối ưu cho cơ chế đa chú ý cũng như chứng minh tính hiệu quả của mô hình tìm được trên các bộ dữ liệu nhận dạng ảnh khuôn mặt deepfake hiện có, so sánh hiệu quả với các mô hình nhận dạng ảnh khuôn mặt deepfake hiện nay. Chúng tôi cho rằng mô hình chúng tôi tìm được sẽ đạt được hiệu suất cao nhất. Để chứng minh mô hình có khả năng chú ý tới các vùng chứa thông tin phân loại khác nhau của bức ảnh, chúng tôi sẽ minh họa các bản đồ chú ý (attention map) có được từ mô hình khi đưa vào hình ảnh khuôn mặt của một người. Chúng tôi cho rằng các bản đồ chú ý sẽ khác biệt và chú ý vào các phần quan trọng của hình ảnh.

GIỚI THIỆU

- Hiện nay nhờ sự phát triển nhanh chóng của các mô hình tạo sinh, các kỹ thuật deepfake khuôn mặt khác nhau [1, 2, 3, 4, 5] đã đạt được những kết quả đột phá. Điều này lại bị những cá nhân, tổ chức độc hại lạm dụng để gây ra các vấn đề xã hội và chính trị nghiêm trọng. Do đó, việc phát hiện deepfake nói riêng

và deepfake khuôn mặt nói chung là bài toán đang được nghiên cứu rộng rãi và đã có nhiều phương pháp được đề xuất.



Ảnh 1: Ảnh khuôn mặt thật và deepfake

- Bài toán: nhận dạng ảnh khuôn mặt deepfake.
 - Input: ảnh rgb khuôn mặt người.
 - Output: thật/giả.
- Tuy nhiên, hầu hết các mô hình được đưa ra đều xem xét bài toán này là bài toán phân loại nhị phân (thật/giả) thông thường [6, 7, 8, 9, 10], tức là mô hình sẽ sử dụng 1 mạng backbone để trích xuất đặc trưng toàn cục rồi đưa vào một bộ phân loại nhị phân. Nhưng sự khác biệt giữa ảnh thật và giả được tạo ra bằng các kỹ thuật mới hiện nay chỉ là những vùng nhỏ cục bộ trong ảnh và chúng rất tinh vi, khó phát hiện. Do đó chúng tôi cho rằng giải pháp trên là không tối ưu, chúng tôi đề xuất xem xét việc phát hiện deepfake là một bài toán phân loại chi tiết.
- Do đó chúng tôi nghiên cứu vấn đề: Mô hình như thế nào là tối ưu cho bài toán nhận dạng ảnh khuôn mặt deepfake khi ta xem xét nó dưới dạng bài toán phân loại ảnh chi tiết ?

MỤC TIÊU

- Thực hiện nghiên cứu để tìm kiếm kiến trúc phù hợp cho 3 khối được đề ra bên dưới.
- Nghiên cứu tìm kiếm phương pháp huấn luyện và tăng cường dữ liệu tối ưu để các bản đồ chú ý (attention map) độc lập và chứa các thông tin khác nhau.
- Thực hiện thử nghiệm mô hình được phát triển với các bộ dữ liệu nhận dạng deepfake hiện tại. Chứng minh lập luận đưa ra là chính xác.

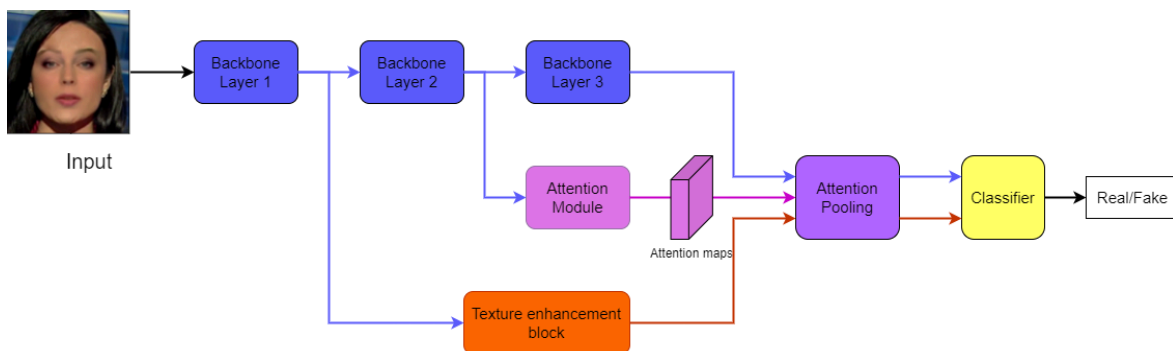
NỘI DUNG VÀ PHƯƠNG PHÁP

a. NỘI DUNG:

- Phân loại ảnh chi tiết (fine-grained image classification) là một bài toán đầy thách thức trong thị giác máy tính. Các nghiên cứu trong lĩnh vực này chủ yếu

tập trung vào việc xác định vị trí các vùng mang thông tin phân loại. Đối với ảnh khuôn mặt, các chi tiết nhỏ như mắt, miệng, kết cấu khuôn mặt, ... chứa thông tin quyết định đến kết quả phân loại.

- Hiện nay cơ chế attention đã và đang được áp dụng rộng rãi trong nhiều loại bài toán khác nhau và đạt được kết quả vượt trội [11, 12, 13]. Chúng tôi áp dụng cơ chế multi-attention vào mô hình giải quyết bài toán, nhờ đó mô hình có khả năng nhận dạng và chú ý vào đa dạng (multi) vùng cục bộ trong ảnh chứa thông tin để phân biệt thật và giả.
- Chúng tôi giả thuyết rằng, những đặc trưng cấp thấp ở những lớp đầu của bức ảnh sẽ vẫn còn thông tin về các vùng cục bộ chứa thông tin quan trọng. Do đó, bộ phân loại cần phải có thông tin ở các đặc trưng nông cần được chú ý mang tính cục bộ và thông tin ngữ nghĩa cấp cao cần được chú ý tương ứng mang tính toàn cục ở các đặc trưng sâu. Do đó, chúng tôi đề xuất mô hình có kiến trúc sau với 3 thành phần quan trọng:



Ảnh 2: Khái quát mô hình

- **Texture enhancement block** để tăng cường các đặc trưng nông ở các lớp đầu. Các đặc trưng này là quan trọng vì thông tin này vẫn còn chi tiết, chưa được toàn cục, tổng quát hoá. Do đó, đặc trưng này cần được tăng cường để khai thác thông tin cục bộ bên trong.
- **Module attention** tạo ra các attention map từ đặc trưng ở các lớp giữa, mỗi attention map chứa thông tin xác định các vùng cục bộ cần chú ý đối với các đặc trưng nông cũng như các đặc trưng ngữ nghĩa cấp cao.
- **Module attention pooling** để gộp thông tin từ attention map với đặc trưng cấp thấp được tăng cường và với đặc trưng ngữ nghĩa cấp cao có được ở các lớp cuối để chỉ các thông tin mang tính phân biệt được đưa vào bộ phân loại.
- Backbone layer: lớp trích xuất đặc trưng.
- Để đảm bảo các attention map chú ý tới các vùng cục bộ khác nhau (multi), chúng tôi đề xuất cần có một phương pháp huấn luyện mới. Trong quá trình huấn luyện, thông tin ở các vùng đã được mô hình chú ý (attention map) sẽ

được được ẩn đi (tăng cường dữ liệu), để mô hình chú ý vào các vùng khác có thể chứa thông tin để phân biệt.

b. PHƯƠNG PHÁP:

- Tiến hành khảo sát các lớp của các mạng phân loại ảnh tốt hiện nay như EfficientNet, XceptionNet, MobileNet, ... để đưa vào backbone layer.
- Nghiên cứu, tìm kiếm kiến trúc mạng phù hợp cho 3 khối được chúng tôi đề ra dựa trên những mô hình xử lý ảnh có áp dụng cơ chế chú ý như [14].
- Nghiên cứu, tìm kiếm phương pháp huấn luyện phù hợp cho mô hình.
- Chứng minh hiệu quả của mô hình chúng tôi đề xuất bằng cách:
 - Thực nghiệm trên các bộ dữ liệu cho bài toán nhận dạng khuôn mặt deepfake: FaceForencis++ [17], DFDC dataset [15], Celeb-DF [16] và so sánh kết quả với các mô hình hiện tại.
 - Minh họa bản đồ chú ý để kiểm chứng mô hình sẽ chú ý vào các vùng khác nhau trong ảnh chứa thông tin quan trọng cho việc phân loại ảnh khuôn mặt thật/giả

KẾT QUẢ MONG ĐỢI

- Chúng tôi hi vọng sẽ tìm ra được kiến trúc mạng phù hợp cho 3 khối được đề ra trong kiến trúc trên cũng như phương pháp huấn luyện phù hợp. Cùng với đó, kết quả của mô hình chúng tôi sẽ tốt hơn so với các mô hình hiện tại cho bài toán này.
- Các bản đồ chú ý thu được khi đưa một hình ảnh khuôn mặt vào kiến trúc phù hợp sẽ cho ra các kết quả độc lập đồng nghĩa với việc mô hình chú ý tới các vùng khác nhau của ảnh.

TÀI LIỆU THAM KHẢO

- [1]. Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, Matthias Nießner: Face2Face: Real-time Face Capture and Reenactment of RGB Videos. CoRR abs/2007.14808 (2020)
- [2]. Supasorn Suwajanakorn, Steven M. Seitz, Ira Kemelmacher-Shlizerman: Synthesizing Obama: learning lip sync from audio. ACM Trans. Graph. 36(4): 95:1-95:13 (2017)
- [3]. Albert Pumarola, Antonio Agudo, Aleix M. Martínez, Alberto Sanfeliu, Francesc Moreno-Noguer: GANimation: Anatomically-Aware Facial Animation from a Single Image. ECCV (10) 2018: 835-851
- [4]. Yuval Nirkin, Yosi Keller, Tal Hassner: FSGAN: Subject Agnostic Face Swapping and Reenactment. ICCV 2019: 7183-7192

- [5]. Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, Chen Change Loy: ReenactGAN: Learning to Reenact Faces via Boundary Transfer. ECCV (1) 2018: 622-638
- [6]. Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner: FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV 2019: 1-11
- [7]. Feng Wang, Weiyang Liu, Hanjun Dai, Haijun Liu, Jian Cheng: Additive Margin Softmax for Face Verification. ICLR (Workshop) 2018
- [8]. Xin Yang, Yuezun Li, Siwei Lyu: Exposing Deep Fakes Using Inconsistent Head Poses. ICASSP 2019: 8261-8265
- [9]. Yuezun Li, Ming-Ching Chang, Siwei Lyu: In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. WIFS 2018: 1-7
- [10]. Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, Baining Guo: Face X-Ray for More General Face Forgery Detection. CVPR 2020: 5000-5009
- [11]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention is All you Need. NIPS 2017: 5998-6008
- [12]. Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, Dacheng Tao: A Survey on Visual Transformer. CoRR abs/2012.12556 (2020)
- [13]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186
- [14]. Heliang Zheng, Jianlong Fu, Tao Mei, Jiebo Luo: Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. ICCV 2017: 5219-5227
- [15]. Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, Cristian Canton-Ferrer: The Deepfake Detection Challenge (DFDC) Preview Dataset. CoRR abs/1910.08854 (2019)
- [16]. Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, Siwei Lyu: Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. CVPR 2020: 3204-3213
- [17]. Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner: FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV 2019: 1-11