Journal of educational equipment: Applied research, Volume 2, Issue 299 (October 2023)
ISSN 1859 - 0810

Tìm hiểu về một số phương pháp phát hiện ra Deepfake trong Deep learning

Nguyễn Thu Huyền*

**ThS. Học viện Hành chính Quốc gia*
Received: 9/9/2023; Accepted: 15/9/2023; Published: 22/9/2023

***Abstract**: Deep Learning can be considered a subfield of Machine Learning– where computers learn and improve themselves through algorithms. It has been successfully applied to solve complex problems ranging from big data analytics to computer vision and human-level control. However, Deep Learning advances have also been used to create software that can pose threats to privacy, democracy, and national security. One of the new applications that has appeared recently is "Deepfakes". Deepfake algorithms can create fake images and videos that humans cannot distinguish from real images or videos. Therefore, being able to detect and evaluate the correctness of images or videos is a current problem.*

***Keywords**: Deepfake, Deep learning*

1. Đặt vấn đề

Trong thời đại kỹ thuật số phát triển nhanh chóng, công nghệ trí tuệ nhân tạo (AI) ngày càng trở nên phổ biến. Trong đó Deepfake - một kỹ thuật tổng hợp hình ảnh, âm thanh hoặc video để tạo ra những nội dung giả mạo, nhằm gây hiểu nhầm cho người xem. Công nghệ deepfake đang phát triển nhanh chóng và trở nên ngày càng tinh vi. Hiện tại, kẻ xấu chủ yếu lợi dụng công nghệ này tạo ra những nội dung giả mạo, tuy nhiên chính phủ các nước đang lo ngại sẽ có nhiều video giả mạo được tung ra với mục đích dẫn dắt dư luận, gây bất ổn chính trị hoặc ảnh hưởng xấu đến tình hình kinh doanh của một công ty nào đó.

Vì vậy, việc nghiên cứu về deepfake là một vấn đề quan trọng và đa dạng, bao gồm nhiều khía cạnh từ công nghệ, xã hội học, an ninh mạng đến luật pháp và đạo đức. Trong đó, việc nghiên cứu và phát triển các công cụ và kỹ thuật để phát hiện và ngăn chặn deepfake trở thành một phần quan trọng của nỗ lực để đối phó với sự phổ biến của công nghệ này.

2. Nội dung nghiên cứu

2.1. Giới thiệu về Deepfake

Deepfake (bắt nguồn từ "Deep Learning" và "Fake") là một kỹ thuật có thể chồng hình ảnh khuôn mặt của người là mục tiêu vào video của người nguồn để tạo video người mục tiêu đang làm hoặc nói những điều mà người nguồn làm. Cơ chế cơ bản để tạo Deepfake là các mô hình học sâu ("Deep Learning") như bộ mã tự động và mạng đối phương chung, đã được áp dụng rộng rãi trong lĩnh vực thị giác máy tính. Công nghệ Deepfake sử dụng mạng nơ-ron sâu (deep neural networks) để học cách sao chép, thay đổi hoặc tạo ra nội dung số hóa có hình dáng, giọng điệu và hành vi của một người nào đó. Các mô hình này được sử dụng để kiểm tra các biểu hiện và chuyển động trên khuôn mặt của một người và tổng hợp các hình ảnh trên khuôn mặt của một người khác tạo ra các biểu hiện và chuyển động tương tự.

Các phương pháp Deepfake thường yêu cầu một lượng lớn dữ liệu hình ảnh và video để huấn luyện các mô hình tạo ra hình ảnh và video chân thực. Vì những nhân vật của công chúng như những người nổi tiếng hay những chính trị gia có thể có một số lượng lớn video và hình ảnh trên mạng, nên họ là mục tiêu ban đầu của những kẻ phá hoại. Deepfakes được sử dụng để hoán đổi khuôn mặt của những người nổi tiếng hoặc chính trị gia sang cơ thể trong các hình ảnh và video giả mạo.

Dưới đây là một số điểm quan trọng về deepfakes:

- *Sự phát triển của deepfake*: Deepfake đã phát triển rất nhanh trong vài năm gần đây nhờ vào sự tiến bộ trong lĩnh vực học máy và xử lý hình ảnh. Người ta đã thể hiện khả năng tạo ra deepfake ấn tượng với cả video và hình ảnh.

- *Ứng dụng của deepfake*: Ban đầu, deepfake được tạo ra để vui đùa hoặc giải trí, như đặt khuôn mặt của người nổi tiếng lên trên nhân vật trong video. Tuy nhiên, nó cũng có thể được sử dụng để tạo ra thông tin sai lệch, gian lận trong video, hoặc thậm chí gây hại cho mục tiêu như xâm phạm quyền riêng tư và danh tiếng.

- *Sử dụng trong nghệ thuật và giải trí*: Deepfake đã trở thành một phần quan trọng của ngành công

nghiệp giải trí. Nó được sử dụng để tạo ra các phim ảnh giả mạo với sự tham gia của những người nổi tiếng, hoặc để tái tạo các cảnh trong phim mà không cần diễn viên thật.

- *Quản lý và quy định*: Nhiều quốc gia và tổ chức quốc tế đã bắt đầu xem xét quy định và hướng dẫn về việc sử dụng Deepfake, đặc biệt là trong ngữ cảnh pháp lý và đạo đức.

- *Thách thức về đạo đức và bảo mật*: Deepfake đặt ra những thách thức đạo đức và bảo mật lớn. Nó có thể được sử dụng để lừa dối và đánh lừa người khác một cách dễ dàng, và việc xác định sự thật từ deepfake có thể khó khăn.

- *Phương pháp phát hiện:* Cùng với sự phát triển của deepfake, đã có sự phát triển của các phương pháp và công cụ phát hiện deepfake. Tuy nhiên, cuộc đua giữa việc tạo ra deepfake và phát hiện chúng vẫn đang diễn ra.

- *Quyền riêng tư và luật pháp*: Nhiều quốc gia đã bắt đầu xem xét và áp đặt luật pháp liên quan đến deepfake để bảo vệ quyền riêng tư và đảm bảo tính chân thực của thông tin truyền tải.

Deepfake là một ví dụ về cách công nghệ có thể mang lại lợi ích lớn nhưng cũng đặt ra nhiều thách thức đạo đức và bảo mật. Việc sử dụng nó nên được tiếp cận với sự cẩn thận và trách nhiệm

2.2. Cách tạo ra Deepfake

Deepfake đã trở nên phổ biến do chất lượng của các video bị giả mạo và cũng như khả năng dễ sử dụng của các ứng dụng của họ đối với nhiều người dùng có kỹ năng máy tính khác nhau từ chuyên nghiệp đến người mới. Các ứng dụng này hầu hết được phát triển dựa trên các kỹ thuật Deep Learning. Nó được biết đến với khả năng biểu diễn dữ liệu phức tạp ở nhiều chiều.

Một biến thể của Deep Learning với khả năng đó là bộ mã tự động sâu, đã được áp dụng rộng rãi để giảm kích thước và nén hình ảnh. Nỗ lực đầu tiên của việc tạo Deepfake là FakeApp, được phát triển bởi một người dùng Reddit bằng cách sử dụng cấu trúc ghép nối bộ mã hóa-giải mã tự động.

Trong phương pháp đó, bộ mã tự động trích xuất các đặc điểm tiềm ẩn của hình ảnh khuôn mặt và bộ giải mã được sử dụng để tái tạo lại hình ảnh khuôn mặt. Để hoán đổi khuôn mặt giữa hình ảnh nguồn và hình ảnh đích, cần có hai cặp bộ mã hóa - giải mã trong đó mỗi cặp được sử dụng để đào tạo trên một tập hợp hình ảnh và các thông số của bộ mã hóa được chia sẻ giữa hai cặp mạng. Nói cách khác, hai cặp có cùng bộ mã hóa mạng lưới. Chiến lược này cho phép bộ mã hóa chung tìm và học sự giống nhau giữa hai tập hợp hình ảnh khuôn mặt, tương đối không thay đổi vì các khuôn mặt thường có các đặc điểm giống nhau như vị trí mắt, mũi, miệng.

Sau đây là một phần của quy trình tạo deepfake để bạn hiểu cách nó hoạt động:

- *Chọn dữ liệu đầu vào:* Để tạo một deepfake, bạn cần có dữ liệu đầu vào, bao gồm video hoặc hình ảnh của người mà bạn muốn đưa vào deepfake. Điều này có thể là người nổi tiếng hoặc bất kỳ người nào bạn muốn.

- *Thu thập dữ liệu cho mô hình deep learning*: Để tạo deepfake, bạn cần một mô hình deep learning, thường là mạng neural học sâu. Bạn sẽ cần thu thập dữ liệu đủ lớn để đào tạo mô hình. Điều này bao gồm video và hình ảnh của người bạn muốn tạo deepfake, cũng như dữ liệu của người bạn muốn đặt khuôn mặt lên.

- *Xử lý và tiền xử lý dữ liệu:* Dữ liệu cần được xử lý để loại bỏ nhiễu, chuyển đổi sang định dạng phù hợp và chuẩn bị cho đào tạo mô hình.

- *Đào tạo mô hình deep learning*: Bạn sẽ cần sử dụng mô hình deep learning để học cách ghép khuôn mặt và biểu hiện của người bạn muốn deepfake vào video hoặc hình ảnh gốc. Quá trình này có thể mất thời gian và đòi hỏi nhiều tài nguyên tính toán.

- *Tạo deepfake*: Sau khi mô hình đã được đào tạo, bạn có thể sử dụng nó để tạo deepfake. Bạn cung cấp video hoặc hình ảnh gốc, và mô hình sẽ thay thế khuôn mặt và biểu hiện của người bạn muốn deepfake.

- *Tinh chỉnh và cải thiện*: Deepfake ban đầu có thể không hoàn hảo, vì vậy bạn có thể cần tinh chỉnh và cải thiện chúng bằng cách điều chỉnh các tham số của mô hình hoặc sử dụng các công cụ hậu kỳ để làm cho chúng trông thực tế hơn.

Chúng ta nên nhớ rằng việc tạo deepfake có thể phức tạp và đòi hỏi kiến thức về machine learning và xử lý hình ảnh. Ngoài ra, việc tạo deepfake có thể vi phạm quy định và luật pháp, vì vậy hãy sử dụng công nghệ này một cách đúng đắn và trách nhiệm.

2.3. Phát hiện deepfake

Deepfakes ngày càng gây phương hại đến quyền riêng tư, an ninh xã hội và dân chủ. Các phương pháp phát hiện Deepfakes đã được đề xuất ngay sau khi mối đe dọa này được đưa ra. Những nỗ lực ban đầu dựa trên các tính năng thủ công thu được từ các đồ tạo tác và sự mâu thuẫn của quy trình tổng hợp video giả mạo. Mặt khác, các phương pháp gần đây đã áp dụng học sâu để tự động trích xuất các đặc điểm nổi

bật và phân biệt để phát hiện các lỗi sâu.

Phát hiện Deepfakes thường được coi là một vấn đề phân loại nhị phân trong đó bộ phân loại được sử dụng để phân loại giữa video xác thực và video bị giả mạo yêu cầu một cơ sở dữ liệu lớn về video thật và giả để đào tạo các mô hình phân tích. Số lượng video giả mạo ngày càng nhiều, nhưng nó vẫn còn hạn chế về mặt thiết lập điểm chuẩn để xác thực các phương pháp phát hiện khác nhau.

- *Phân tích kỹ thuật đồ họa*: Một số deepfakes có thể bị phát hiện bằng cách phân tích các tính chất kỹ thuật của hình ảnh hoặc video. Điều này có thể bao gồm việc xem xét sự không khớp về sắc thái màu sắc, lỗi nhận dạng khuôn mặt, hoặc kiểu biểu đồ học cơ bản của video.

- *Kiểm tra phần lớn khuôn mặt*: Deepfake thường sử dụng kỹ thuật để ghép một khuôn mặt vào hình ảnh hoặc video gốc. Nếu bạn thấy rằng khuôn mặt chiếm phần lớn trong khung hình mà không có nhiều biểu đồ thay đổi xung quanh, có thể đây là dấu hiệu của deepfake.

- *Kiểm tra bất thường trong đôi mắt:* Một số deepfake có thể không hoàn hảo trong việc tái tạo đôi mắt và ánh mắt của người thật. Kiểm tra sự không bình thường trong ánh mắt, như việc không nhìn thẳng vào camera hoặc không có bóng đèn phản chiếu trong mắt, có thể là dấu hiệu của deepfake.

- *Kiểm tra âm thanh*: Nếu deepfake cũng bao gồm âm thanh, bạn có thể kiểm tra các biểu đồ sóng âm thanh để xem xét sự không phù hợp hoặc bất thường trong giọng điệu, ngữ điệu hoặc chất lượng âm thanh.

- *Sử dụng công cụ phát hiện deepfake*: Có nhiều công cụ và phần mềm phát hiện deepfake được phát triển bởi các tổ chức và nhóm nghiên cứu. Các công cụ này sử dụng các kỹ thuật máy học để phát hiện deepfake dựa trên các đặc điểm đặc thù của chúng. Ví dụ, các dự án như Deepware Scanner, Microsoft Video Authenticator và FaceForensics++ là một số ví dụ.

- *Phân tích đặc trưng cơ bản:* Các phương pháp sử dụng phân tích đặc trưng cơ bản của deepfake, chẳng hạn như việc xem xét dấu vết của các thuật toán deep learning, có thể giúp phát hiện sự can thiệp của deepfake.

- *So sánh với nguồn gốc*: So sánh kỹ lưỡng giữa deepfake và hình ảnh/video gốc có thể giúp bạn phát hiện sự không phù hợp. Tuy nhiên, điều này có thể đòi hỏi sự kỹ năng và thời gian.

Chúng ta thấy rằng các deepfake ngày càng phát triển và trở nên khó khăn để phát hiện. Do đó, việc kết hợp nhiều phương pháp phát hiện và cùng đánh giá cẩn thận là quan trọng để tăng khả năng phát hiện ra deepfakes.

3. Kết luận

Deepfakes đã bắt đầu làm mất lòng tin của mọi người đối với các nội dung truyền thông do việc nhìn thấy chúng không còn tương xứng với việc tin vào chúng. Chúng có thể gây ra đau khổ và tác động tiêu cực cho những người được nhắm mục tiêu, nâng cao thông tin sai lệch, và thậm chí có thể kích thích căng thẳng chính trị, kích động công chúng, bạo lực hoặc chiến tranh. Điều này đặc biệt quan trọng hiện nay khi các công nghệ để tạo ra Deepfakes ngày càng dễ tiếp cận và các nền tảng truyền thông xã hội có thể lan truyền những nội dung giả mạo đó một cách nhanh chóng.

Đôi khi, Deepfakes không cần phải được phổ biến rộng rãi đến khán giả để gây ra những tác động bất lợi. Những người tạo ra Deepfakes với mục đích xấu chỉ cần bỏ chúng để nhắm mục tiêu đến khán giả như một phần của chiến lược phá hoại của họ mà không cần sử dụng phương tiện truyền thông xã hội. Vì vậy, chúng ta cần hết sức tỉnh tóa trước những thông tin được cung cấp và lựa chọn những phương pháp phù hợp để phát hiện ra những thông tin giả mạo.

Tài liệu tham khảo

[1]. Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmenta- tion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481-2495.

[2]. Yang, W., Hui, C., Chen, Z., Xue, J. H., and Liao, Q. (2019). FV-GAN: Finger vein representation using generative adversarial networks. *IEEE Transactions on Information Forensics and Security*, 14(9), 2512-2524.

[3]. Tewari, A., Zollhoefer, M., Bernard, F., Garrido, P., Kim, H., Perez, P., and Theobalt, C. (2020). High-fidelity monocular face reconstruc- tion based on an unsupervised model-based face autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: 10.1109/TPAMI.2018.2876842.

[4]. Guo, Y., Jiao, L., Wang, S., Wang, S., and Liu, F. (2018). Fuzzy sparse autoencoder framework for single image per person face recognition. *IEEE Transactions on Cybernetics*, 48(8), 2402-2415.

[5]. Liu, F., Jiao, L., and Tang, X. (2019). Task-oriented GAN for PolSAR image classification and clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9), 2707-2719.