트랜스포머 기반 판별 특징 학습 비전을 통한 얼굴 조작 감지

Van-Nhan Tran[1], 김민수 [1], 최필주 [1], 이석환 [2], Hoanh-Su Le[3], 권기룡 [1]
[1] 부경대학교 인공지능융합학과, [2] 동아대학교 컴퓨터공학부
[3] Faculty of Information Systems, University of Economics and Law, Ho Chi Minh City and Vietnam National University

tvnhanpk@pukyong.ac.kr, whan7808@naver.com, pjchoi@pknu.ac.kr, skylee@dau.ac.kr, sulh@uel.edu.vn krkwon@pknu.ac.kr

Facial Manipulation Detection with Transformer-based Discriminative Features Learning Vision

Van-Nhan Tran[1], Minsu Kim[1], Philjoo Choi[1], Suk-Hwan Lee[2], Hoanh-Su Le[3], Ki-Ryong Kwon[1]
[1]Dept. of Artificial Intelligence Convergence, Pukyong National University
[2]Division of Computer and AI Engineering, Dong-A University, Busan
[3]Faculty of Information Systems, University of Economics and Law, Ho Chi Minh City and Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

Due to the serious issues posed by facial manipulation technologies, many researchers are becoming increasingly interested in the identification of face forgeries. The majority of existing face forgery detection methods leverage powerful data adaptation ability of neural network to derive distinguishing traits. These deep learning-based detection methods frequently treat the detection of fake faces as a binary classification problem and employ softmax loss to track CNN network training. However, acquired traits observed by softmax loss are insufficient for discriminating. To get over these limitations, in this study, we introduce a novel discriminative feature learning based on Vision Transformer architecture. Additionally, a separation-center loss is created to simply compress intra-class variation of original faces while enhancing inter-class differences in the embedding space.

1. Introduction

With substantial deep learning breakthroughs. Face manipulation techniques [1,2] based on Generative Adversarial Networks (GAN) [3] enables normal people to create high-quality forged faces without the need for specialized abilities or knowledge. Over time, a variety of solutions have been offered in response to this problem. Early research has focused on altering the design of existing neural networks or using new properties [4]. The focus of mainstream research then gradually shifted to methods that build backbone networks using a variety of information and knowledge [5].

2. Proposed Method

Data preprocessing. The first stage in our proposal is data preparation. The open dataset FaceForensics++ [6] is used in our proposal. The FaceForensics++ contains both authentic and manipulated videos. To obtain images, these videos are sampled. After that, we use the MTCNN package [7] to crop out faces from the sampled images. An attention neural network is the first tool our approach uses to generate attention features for either genuine or synthetic facial images. An overview of data preprocessing is shown in Figure 1.

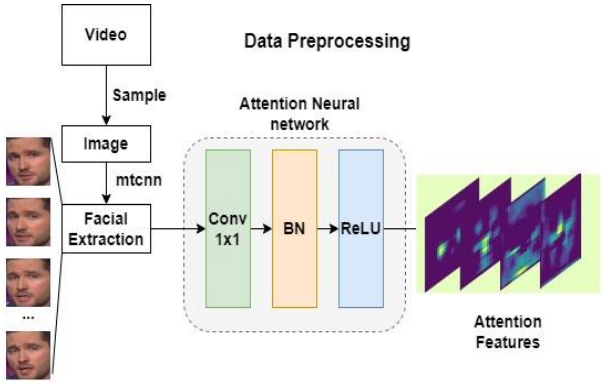

Fig. 1. Overview of face image preprocessing.

Separation-center loss. Softmax loss is frequently used in deep learning-based face forgery detection methods currently in use. The objective of softmax loss is to identify a decision

boundary that can be utilized to distinguish between multiple classes, though. Fundamentally, the learned properties under softmax loss supervision are insufficient for discrimination. It is challenging to assemble all the altered faces because samples created by various alteration procedures have different feature compositions. due to the difficulties of optimization, leads to a less-than-ideal solution and even worsens performance through overfitting. To address this problem, we offer a separability-center loss. Figure 2 displays the separability-center loss visualization in embedding space.

$$L_{sep} = \frac{1}{R}\sum_{i=1}^{R}(F_{ri} - C) + \max\left(\frac{1}{R}\sum_{i=1}^{R}(F_{ri} - C) - \frac{1}{M}\sum_{i=1}^{M}(F_{fi} - C), 0\right) \quad (1)$$

Where F_{ri}, F_{fi} are the corresponding embedding features of real and fake samples. At each iteration, the centers C are updated. R and M are number of real and fake points

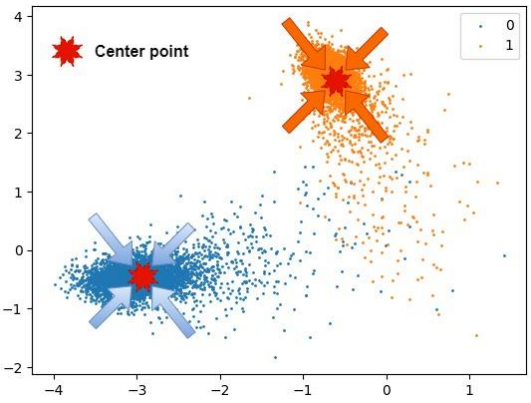

Fig. 2. The sample feature distribution in embedding space. Blue points represent original faces and orange points represent manipulated faces.

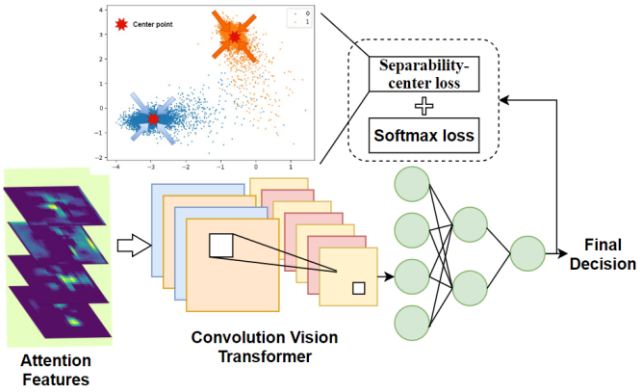

Fig. 3. Architecture of our proposal.

Our main architecture is displayed in Figure 3. The Attention features from data processing is used as input. ResNet-50 [8] is used as the main backbone of the comvolution vision transformer block [9]. Figure 4 shows the detailed components of convolutional vision transformer. The center point C of separation-center loss is randomly established and modified based on the mini-batches. Additionally, we combine softmax loss with separation-center loss to direct the center points.The total loss can be calculated as follows:

$$L_{total} = L_{softmax} + \lambda L_{sep} \quad (2)$$

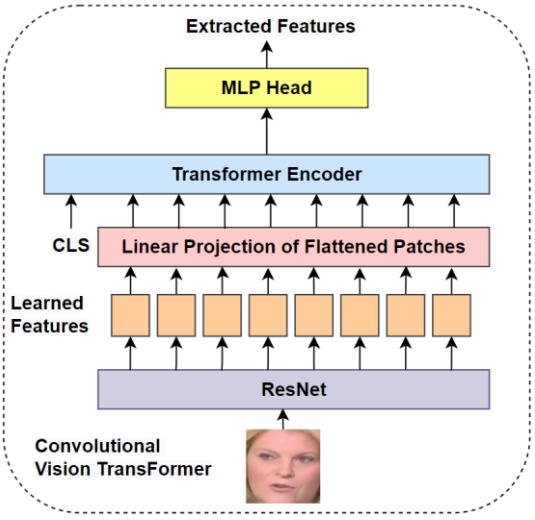

Fig. 4. Convolution Vision Transformer.

3. Experiment

Implementation detail. we use the FaceForensics++ [6] dataset for our evaluation. The FaceForensics++ contains 1000 original videos that have all been modified utilizing four different face modification methods. Three distinct versions of the FaceForensics++ dataset are available: c0 (raw), c23 (high quality image), and c40 (low quality image), each with a different amount of compression. In our experiments, we exclusively use low quality images c40 for evaluation. We set λ to 0.5.

Evaluation metric. Area under the receiver operating characteristic curve (AUC) is utilized as evaluation metric. When choosing a classification threshold, a classifier is displayed using the receiver operating characteristic (ROC). AUC is a region that is under the ROC curve. Additionally, accuracy score (ACC) is used to assess categorization models.

Table 1. Comparison results using the FaceForensics++ dataset.

Methods	c40	
	Acc	AUC
Two-Branch [10]	-	86.59
Xception [6]	81.00	-
MesoNet [4]	70.46	-
ResNet-50 [8]	85.59	87.62
Belhassen et .al [11]	66.84	-
Face-X-ray [12]	-	61.6
Ours	**88.11**	**91.7**

Acknowledgments

This research was supported by the MSIT (Ministry of Scien ce and ICT), Korea, under the ITRC (Information Technolog

y Research Center) support program (IITP-2023-2020-0-01797) supervised by the IITP (Institute for Information & C ommunications Technology Planning & Evaluation) and the MSIT (Ministry of Science and ICT), Korea, under the ICT Consilience Creative program (IITP-2023-2016-0-00318) supervised by the IITP (Institute for Information & co mmunications Technology Planning & Evaluation).

References

[1] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE transactions on image processing,* vol. 28, no. 11, pp. 5464-5478, 2019.

[2] R. Durall, J. Jam, D. Strassel, M. H. Yap, and J. Keuper, "FacialGAN: Style transfer and attribute manipulation on synthetic faces," *arXiv preprint arXiv:2110.09425,* 2021.

[3] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine,* vol. 35, no. 1, pp. 53-65, 2018.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*, 2018: IEEE, pp. 1-7.

[5] H. Qi *et al.*, "Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 4318-4327.

[6] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1-11.

[7] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters,* vol. 23, no. 10, pp. 1499-1503, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[9] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126,* 2021.

[10] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, 2020: Springer, pp. 667-684.

[11] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, 2016, pp. 5-10.

[12] L. Li *et al.*, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001-5010.