다중 양식 융합과 개선된 E-MTN을 활용한 비디오 폭력 탐지 성능 향상

나광호, 최경주
충북대학교 소프트웨어학과
e-mail : gh369ho@naver.com, kjcheoi@cbnu.ac.kr

Enhancing Video Violence Detection Performance Using Multi-modal Fusion and E-MTN

Gwangho Na, Kyung Joo Cheoi
Dept. of Computer Science, Chungbuk, National University

요 약

본 논문에서는 다중 양식 융합과 개선된 E-MTN을 활용하여 비디오 폭력 탐지 성능을 향상한 새로운 모델을 제안하였다. 제안된 모델은 기존의 이중 양식 모델인 TEVAD를 확장한 것으로, 시각, 음성, 문자 정보의 3중 양식을 활용하였고, 다양한 실험을 통해 모든 양식의 특징을 통합한 후에 비디오 스니펫 간의 다중 시간 스케일 특징 추출 네트워크(MTN)를 적용하는 것이 가장 우수한 성능을 보임을 확인하였으며, 특히 각 특징을 단순히 연결하는 방법이 가장 효과적임을 입증하였다. 또한, 개선된 MTN을 사용하여 성능이 향상됨을 확인하였다.

1. 서 론

최근 몇 년간 비디오 콘텐츠의 소비가 급증하면서 폭력적인 장면을 포함한 비디오를 자동으로 감지하는 기술에 대한 필요성이 증가하고 있다. 기존의 폭력 탐지 연구는 주로 시각 정보에 의존하여 개발되었다. 하지만 이러한 접근 방식은 복잡한 실제 상황에서 발생하는 다양한 형태의 폭력을 정확하게 탐지하는 데 한계가 있다. 비디오를 정확하게 해석하기 위해서는 시각 정보뿐만 아니라 다중 양식(Multi-modal)을 고려할 필요가 있다. 시각 정보만으로는 부족한 의미론적 의미를 포함하고 있으므로 다중 양식은 비디오를 효과적으로 해석하는 데 도움을 줄 수 있다.

최근의 비디오 폭력 탐지 연구들은 완전히 라벨이 달린 데이터가 아닌, 부분적으로 또는 간접적으로 라벨이 제공된 데이터를 사용하여 모델을 학습하는 접근법인 약한 지도 학습(weakly supervised learning)[1]을 사용한다. 약한 지도 학습 방법은 대부분 다중 인스턴스 학습(Multiple Instance Learning, MIL) 프레임워크를 기반으로 하는데, MIL은 클립 내의 각 구간에 대한 세부 라벨은 사용하지 않고, 비디오 클립 전체에 대한 라벨을 사용해 클립 내의 프레임 혹은 짧은 구간마다 라벨을 추정한다.

Chen 등[2]은 기존의 시각 특징만으로는 파악하기 어려운 의미론적 정보를 포착하기 위해 생성한 문자 특징을 시각 특징과 결합한 문자 기반 비디오 이상 탐지 시스템인 TEVAD(Text Empowered Video Anomaly Detection)를 제안하였다. TEVAD는 비디오 내의 의미론적 의미를 포착한 문자 정보를 함께 사용함으로써 기존의 시각적 특징 기반 방법보다 향상된 성능을 보여주었다.

본 논문에서는 시각 정보와 문자 정보를 활용한 TEVAD을 확장하여 시각 특징만으로는 추출하기 어려웠던 의미론적 의미를 비디오 캡셔닝 모델을 사용해 생성한 문자(캡션)를 통해 포착하고, 음성 정보를 융합하는 통합적인 접근 방법을 제안한다. 기존 TEVAD와는 음성 정보를 추가로 사용한다는 점과 특징 융합의 방법 및 시기가 다르다는 차이가 있다. 또한 다중 시간 스케일에서 시간 특징을 더욱 효과적으로 추출하기 위해 기존의 MTN(Multi-scale Temporal Network)을 수정한 개선된 E-MTN을 제안한다. 이러한 차이로 비디오 내 폭력을 탐지하는 성능을 극대화시켰다.

2. 제안 방법

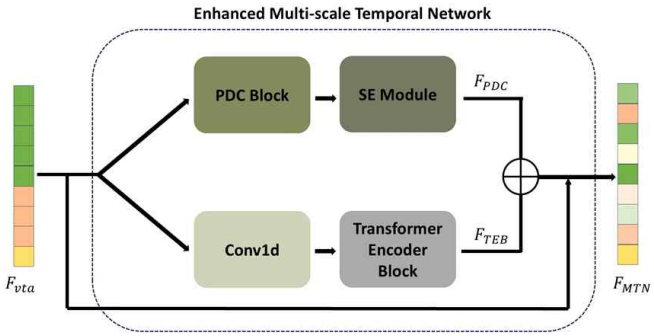
2.1 제안된 구조

제안된 모델의 전체 구조를 설명하면 다음과 같다. 먼저 비디오가 모델에 입력되면 T개의 스니펫(snippets)으로 분할되고, 분할된 각 스니펫에 대해 시각 특징 F_v, 음성 특징 F_a, 문자 특징 F_t를 추출한다. 이렇게 추출된 각 특징들은 연결(Concatenation) 방법을 통해 하나의 특징으로 융합되어 F_{vta}를 생성한다. 이후 융합된 특징은 개선된 E-MTN(Enhanced Multi-scale Temporal Networks)을 거쳐 다중 시간 스케일 특징 F_{E-MTN}을 생성한다. 이렇게 생성된 F_{E-MTN}은 각 스니펫의 특징 크기(magnitude) 계산에 사용되어진다. 정상과 폭력 비디오로부터 만들어진 상위 k개의 가장 큰 특징 크기가 스니펫 수준에서 폭력인지 아닌지를 탐지하기 위한 폭력 탐지 분류기의 훈련을 위해 전달되어 폭력 탐지를 위한 2진 분류기를 훈련한다.

2.2 개선된 E-MTN

기존의 MTN[2]은 다양한 시간 스케일에서 시간 특징을 추출하기 위해 피라미드 확장 컨볼루션(PDC) 블록과

Non-local 블록(NLB)을 사용하였다. 본 논문에서는 SE 모듈(Squeeze and Excitation Module)을 추가하고, NLB를 트랜스포머 인코더 블록(TEB)으로 대체한 개선된 E-MTN을 제안한다, 그림 1은 개선된 E-MTN의 전체적인 구조도를 보여준다. F_{E-MTN}을 추출하는 절차는 다음과 같다. 먼저 융합된 특징 F_{vta}는 PDC 블록과 SE 모듈을 통과하여 F_{PDC}를 생성한다. 동시에 F_{vta}는 1D-컨볼루션을 거친 후 TEB를 통과하여 F_{TEB}를 생성한다. 이후 F_{PDC}와 F_{TEB}를 연결(Concatenation)하고, 원래 특징인 F_{vta}를 더하여(Addition) 최종 출력 F_{E-MTN}을 생성한다.



(그림 1) 개선된 E-MTN 구조도

3. 실험 및 결과

제안하는 시스템의 성능을 평가하기 위해 다중 양식을 제공하는 XD-Violence[3] 데이터세트를 사용하였다. 표 1은 XD-Violence 데이터세트에 대한 프레임 수준의 AP 성능을 비교한 결과이다. 본 연구에서 제안한 다중 양식 접근법의 효과를 평가하기 위해, 성능이 우수한 기존의 단일 양식 및 이중 양식 모델들과 성능을 비교하였다. 본 연구에서 제안된 다중 양식 접근법은 시각, 음성, 문자 정보를 모두 결합하여 사용함으로써 83.9%의 AP 성능을 기록하였다.

(표 1) XD-Violence 데이터세트에 대한 프레임 수준 AP 성능 비교

Method	Modality	AP(%)
RTFM (2021) [4]	V	77.8
TEVAD (2023) [2]	V, T	79.8
Zhang et al. (2023) [5]	V, A	81.4
UR-DMU (2023) [6]	V, A	81.7
Ours (Original MTN)	**V, T, A**	**82.2**
Ours (Enhanced MTN)	**V, T, A**	**83.9**

4. 결 론

본 논문에서는 TEVAD를 확장하여 시각, 음성, 문자 정보를 모두 활용한 다중 양식 비디오 폭력 탐지 시스템을 제안하였다. 다양한 실험을 통해, 모든 양식의 특징을 하나의 통합된 특징으로 융합한 후 MTN을 거치도록 하는 방법이 가장 우수한 성능을 보임을 확인하였다. 다양한 특징 융합 방법 중에서는 각 특징들을 단순하게 이어 붙이는 연결(Concatenation) 방법이 가장 효과적임을 입증하였으며 다중 양식을 사용하는 경우가 기존 단일 양식 또는 이중 양식을 사용하는 경우보다 성능이 더 우수함을 확인하였다. 또한 대부분의 실험에서 기존 MTN을 사용한 경우보다 제안된 개선된 E-MTN을 사용했을 때 성능이 더욱 향상됨을 확인하였다. 향후 연구에서는 다중 양식을 더욱 효율적으로 사용할 수 있는 융합 방식과 모델 구조의 개발을 통해 비디오 폭력 탐지 시스템의 정확성과 효율성을 더욱 높이는 방향으로 진행할 예정이다.

Acknowledgment

이 논문은 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업(2019-0-01183)의 지원을 받아 작성됨.

참 고 문 헌

[1] W. Sultani, C. Chen, M. Shah, "Real-world anomaly detection in surveillance videos," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6479-6488.

[2] W. Chen, K. T. Ma, Z. J. Yew, M. Hur, D. A. A. Khoo, "TEVAD: Improved video anomaly detection with captions," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5549-5559.

[3] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," Computer Vision - ECCV 2020: 16th European Conference, Springer International Publishing, Part XXX 16, Glasgow, UK, August 23 - 28, 2020, pp. 322-339.

[4] Y. Tian, G. Pang, Y. Chen, R. Singh, J.W. Verjans, G. Carneiro, "Weakly-Supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning," Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4975-4986.

[5] C. Zhang, G. Li, Y. Qi, S. Wang, L. Qing, Q. Huang, M. H. Yang, "Exploiting Completeness and Uncertainty of Pseudo Labels for Weakly Supervised Video Anomaly Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16271-16280.

[6] H. Zhou, J. Yu, W. Yang, "Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video Anomaly Detection," Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 3, 2023, pp. 3769-3777.