심층 학습 기반 정지 영상 분류와 행동 인식을 이용한 음란 동영상 탐지

이원재, 김정학, 이남경
한국전자통신연구원

russell@etri.re.kr, junghak@etri.re.kr, nklee@etri.re.kr

Pornographic Video Detection with Image Classification and Action Recognition based on Deep Learning

Wonjae Lee, Junghak Kim, Nam Kyung Lee
Electronics and Telecommunications Research Institute

요 약

본 논문은 심층 학습 기반 정지 영상 분류와 행동 인식을 이용한 음란 동영상 탐지 시스템에 대해 논의한다. 개발한 시스템은 정지 영상 분석 네트워크와 동작 분석 네트워크의 Two-Stream 네트워크 융합을 통해 정지 영상 분석 결과와 동작 분석 결과를 효과적으로 활용하여 음란성을 판단한다. 정지 영상 분석 네트워크는 121 개 레이어를 가지고 있는 DenseNet 을 기반으로 한다. 동작 분석 네트워크는 5 개의 광학 흐름을 입력으로 가지며, DenseNet 을 기반으로 한 구조를 가진다. 네트워크 융합을 위해 각 네트워크의 마지막 ReLU 출력을 병합하고 2D 합성곱과 2D 전역 평균 풀링을 적용하였다.

Ⅰ. 서론

인터넷 상에서의 음란물 유포 문제는 과거부터 있어 왔었고, 일반적으로 사람에 의해 탐지 및 삭제되었다. 그러나 최근 심층 학습(deep learning) 기술의 급속한 발전으로 자동화된 음란물 탐지가 가능하게 되었고, 이를 통해 신속한 모니터링이 가능하다. 본 논문에서는 심층 학습 기반 정지 영상 분류와 행동 인식을 이용한 음란 동영상 탐지 시스템에 대해 기술한다.

관련 선행 연구로는 심층 합성곱 신경망 기반 음란 정지 영상 분류 [1], 심층 학습 기법과 움직임 정보를 활용한 음란 동영상 탐지 [2], 동영상 행동 인식을 위한 합성곱 two-stream 네트워크 융합 [3]이 있다.

Ⅱ. 본론

본 연구에서는 행동 인식 연구 [3]와 유사한 방식을 적용하여 음란 동영상을 탐지한다. 해당 방식에서는 정지 영상 분류 네트워크와 동작 분석 네트워크를 각각 학습시킨 후 융합하고, 융합된 네트워크에 대해 추가 학습을 수행한다. 사용한 심층 학습 프레임워크는 Keras 이며 백엔드는 TensorFlow 이다.

1. 정지 영상 분류 네트워크

정지 영상 분류 네트워크로는 121 개 레이어를 가지고 있는 심층 합성곱 신경망(convolutional neural network)인 DenseNet [4]을 사용하였다. DenseNet 은 dense connectivity pattern 을 활용하여 적은 수의 파라미터로 효과적인 학습이 가능하기에 선택하였다. ImageNet 으로 사전 학습된 네트워크를 사용하여 전이 학습을 수행하였으며, 일부 계층(layer)을 고정시키고 학습하는 경우 달성할 수 있는 최대 정확도가 떨어져서 모든 계층을 학습 가능하게 설정하고 학습시켰다. 입력 영상 크기는 224×224 이다.

학습에는 392,533 개의 정지 영상을, 검증에는 3,707 개의 정지 영상을 사용하여 검증 정확도 99.7%를 달성하였다.

학습/검증 데이터 분류와 관련하여 방송통신심의위원회 인터넷내용등급서비스 노출 4 등급에 해당하면 음란물로 판단하였다. 정지 영상 학습 및 검증을 위한 데이터에는 동작 분석 네트워크 및 융합 네트워크 학습을 위해 수집한 동영상에서 추출한 정지 영상이 일부 포함되어 있다.

2. 동작 분석 네트워크

동작 분석을 위해 Gunner Farnebäck 알고리즘 [5]을 이용하여 고밀도 광학 흐름(dense optical flow)을 생성하였다. 선행 연구 [2][3]의 경우 변위 벡터(displacement vector) 방식으로 광학 흐름을 표현했지만, 본 연구에서는 각도와 크기로 표현했다. 학습 속도를 향상시키기 위해 학습 전 미리 광학 흐름을 생성한 후 각도와 크기를 HSV 색공간의 색상(hue)과 명도(value)로 변환하여 JPEG 으로 저장한 후, 학습 시에 읽어 각도와 크기로 변환하였다.

신경망 네트워크로는 121 개 레이어를 가지고 있는 DenseNet 을 5 개의 광학 흐름을 입력으로 받도록 수정하여 사용하였다. 입력 형상이 변경되어 전이학습은

사용하지 못했다. 7 프레임 간격으로 인접한 동영상 프레임을 비교하여 광학 흐름을 생성하였으며, 입력 크기는 224×224 이다.

학습 데이터 1,273,832 개와 검증 데이터 160,098 개를 사용하여 검증 정확도 97.1%를 달성하였다.

학습/검증 데이터 분류와 관련하여 동영상 클립이 방송통신심의위원회 인터넷내용등급서비스 노출 4 등급 또는 성행위 4 등급에 해당하면 음란물로 판단하였다.

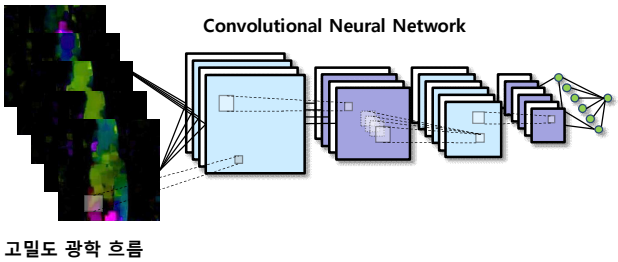

그림 1 동작 분석 네트워크 개념도

3. 네트워크 융합

정지 영상 분류 네트워크와 동작 분석 네트워크의 two-stream 네트워크 융합을 위해 각 네트워크의 마지막 ReLU 출력을 병합하고, 3×3 필터를 가지는 2D 합성곱(convolution)과 2D 전역 평균 풀링(global average pooling)을 적용하였다. 융합 네트워크에 대한 학습 시 5 개의 광학 흐름과 1 개의 정지 영상을 입력으로 제공하였다.

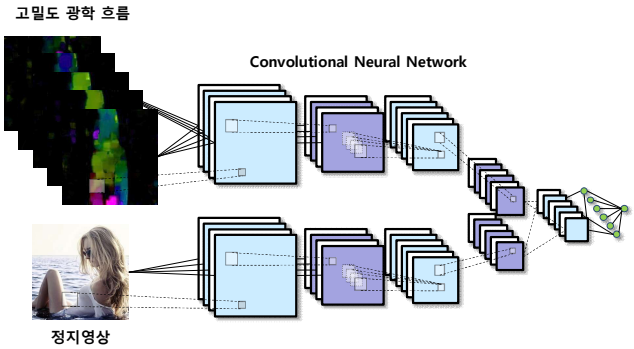

그림 2 융합 네트워크 개념도

기존의 학습된 정지 영상 분류 네트워크와 동작 분석 네트워크에서 가져온 계층들을 모두 고정시키고 융합 시 추가된 계층들에 대해서만 학습을 진행할 경우 정확도가 특정 값 이상으로 올라가지 않는 현상이 발생하였다. 이에 따라 상대적으로 떨어지는 정확도를 보인 동작 분석 네트워크에서 가져온 계층들은 학습 가능하게 설정하고 학습시켰다. 병합 후 적용하는 2D 합성곱의 필터 수가 64 개, 128 개일 때는 검증 정확도 99.5%, 필터 수가 1024 개일 때는 검증 정확도 99.6%를 달성하였다.

네트워크 융합 시 3×3×3 필터를 가지는 3D 합성곱과 3D 전역 평균 풀링을 적용한 경우, 3D 합성곱 필터 수가 16, 32, 128 개인 경우에 대해 모두 검증 정확도 약 97%를 달성하였다. 선행 연구 [3]와 달리 데이터 증대 기법(data augmentation)을 적용하지 않아서 과적합(overfitting)이 일어난 것으로 판단된다.

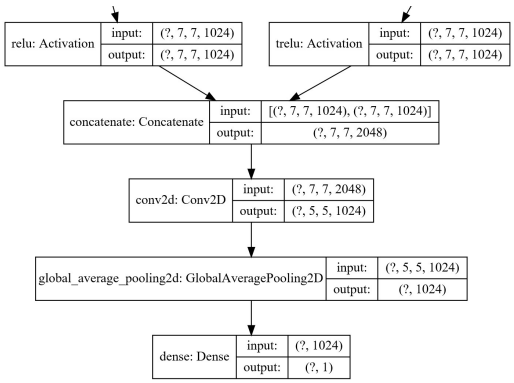

그림 3 융합된 네트워크 중 융합되는 부분

Ⅲ. 결론

본 논문에서는 심층 학습 기반 정지 영상 분류와 행동 인식을 이용한 음란 동영상 탐지 시스템에 대해 기술하였다. 각기 학습된 정지 영상 분석 네트워크와 동작 분석 네트워크를 융합하고 추가적인 학습을 수행하였다. 이를 통해 정지 영상 분석 결과와 동작 분석 결과의 효과적 활용이 가능하였으며 높은 검증 정확도를 달성하였다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신·방송 연구개발 사업의 일환으로 수행하였음. [2019-0-00287, 인공지능 기반 유해미디어(음란성) 분석·검출 시스템 개발]

참 고 문 헌

[1] Kailong Zhou, Li Zhuo, Zhen Geng, Jing Zhang, Xiaoguang Li, “Convolutional neural networks based pornographic image classification,” IEEE Second International Conference on Multimedia Big Data (BigMM), 2016, pp. 206-209.

[2] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, A. Rocha, “Video pornography detection through deep learning techniques and motion information,” Elsevier Neurocomputing 230, 2017, pp. 279-293.

[3] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” IEEE International Conference on Computer Vision and Pattern Recognition CVPR, 2016.

[4] G. Huang, Z. Liu, K. Q. Weinberger, L. Maaten, “Densely connected convolutional networks,” IEEE Conference on Computer Vision and Pattern Recognition CVPR, 2017.

[5] G. Farnebäck, “Two-frame motion estimation based on polynomial expansion,” SCIA, 2003.