효율적인 모델 학습을 위한 심층 특징의 평균값을 활용한 의미 있는 비디오 프레임 추출 기법

*윤혁 **김영기 ***한지형

서울과학기술대학교

*titania7777@seoultech.ac.kr **kyg1552@seoultech.ac.kr ***jhhan@seoultech.ac.kr

Salient Video Frames Sampling Method Using the Mean of Deep Features for Efficient Model Training

*Yoon, Hyeok **Kim, Young-Gi ***Han, Ji-Hyeong

Seoul National University of Science and Technology

요약

최근 정보통신의 발달과 함께 인터넷에 접속하는 사용자 수와 그에 따른 비디오 데이터의 전송량이 늘어나는 추세이다. 이렇게 늘어나는 많은 비디오 데이터를 관리하고 분석하기 위해서 최근에는 딥 러닝 기법을 많이 활용하게 된다. 일반적으로 비디오 데이터에 딥 러닝 모델을 학습할 때 컴퓨터 자원의 한계로 인해 전체 비디오 프레임에서 균등한 간격 또는 무작위로 프레임을 선택하는 방법을 많이 사용한다. 하지만 학습에 사용되는 비디오 데이터는 항상 시간 축에 따라 같은 문맥을 담고 있는 Trimmed 비디오라고 가정할 수가 없다. 만약 같지 않은 문맥을 지닌 Untrimmed 비디오에서 균등한 간격 또는 무작위로 프레임을 선택해서 사용하게 된다면 비디오의 범주와 관련이 없는 프레임이 샘플링 될 가능성이 있기 때문에 모델의 학습 및 최적화에 전혀 도움이 되지 않는다. 이를 해결하기 위해 우리는 각 비디오 프레임에서 심층 특징을 추출하여 평균값을 계산하고 이와 각 추출된 심층 특징들과 코사인 유사도를 계산해서 얻은 유사도 점수를 바탕으로 Untrimmed 비디오에서 의미 있는 비디오 프레임을 추출하는 기법을 제안한다. 그리고 Untrimmed 비디오로 구성된 데이터셋으로 유명한 ActivityNet 데이터셋에 대해서 대표적인 2가지 프레임 샘플링 방식(균등한 간격, 무작위)과 비교하여 우리가 제안하는 기법이 Untrimmed 비디오에서 효과적으로 비디오의 범주에 해당하는 의미 있는 프레임 추출이 가능함을 보일 것이다. 우리가 실험에 사용한 코드는 https://github.com/titania7777/VideoFrameSampler에서 확인할 수 있다.

1. 서론

최근 코로나19의 영향으로 유튜브와 같은 동영상 플랫폼을 이용하는 사용자가 증가하면서 비디오 데이터의 양이 급속도로 증가하고 있다[1]. 또한 이에 대응하는 비디오 데이터의 관리 및 분석을 위해 전통적인 컴퓨터 비전 기술을 벗어나 딥 러닝 기법을 응용하는 추세이다[2]. 하지만 비디오 데이터는 이미지 데이터와 달리 시간 정보가 함께 포함되어 있어 매우 복잡하다. 그리고, 컴퓨터 자원의 한계로 전체 비디오 프레임을 사용하여 딥 러닝 모델을 학습하는 것은 현실적이지 않다. 따라서 비디오 데이터를 사용하여 딥 러닝 모델을 학습할 때는 전체 비디오 프레임의 일부를 선택해서 사용해야 한다. 일반적으로 전체 비디오 프레임에서 균등한 간격 또는 무작위로 프레임을 선택해서 사용한다. 그러나 균등한 간격 또는 무작위로 프레임을 선택하는 방식은 간단하면서 적은 컴퓨터 자원으로 모델 학습을 진행할 수는 있지만 비디오 범주와 관련이 없고 학습에 도움이 안 되는 프레임이 선택될 가능성이 높기 때문에 좋은 방법이라고 할 수 없다.

딥 러닝 모델 학습에 사용되는 비디오 데이터는 크게 Trimmed 비디오와 Untrimmed 비디오로 구분할 수 있다[3-5]. Trimmed 비디오는 비디오의 범주에 해당하는 핵심 프레임으로 구성된 비디오를 의미하고 Untrimmed 비디오는 비디오 범주에 해당하는 핵심 프레임과 의미 없는 프레임으로 구성된 비디오를 의미한다. 예를 들어 축구경기를 녹화한 비디오가 있다고 가정하면 축구 경기를 진행하는 장면뿐만 아니라 중간 광고도 같이 나오게 된다. 이때, 축구경기를 진행하는 장면을 잘 포착하고 정리된 비디오가 Trimmed 비디오이며 광고가 같이 포함되어있는 비디오가 Untrimmed 비디오라고 할 수 있다.

Trimmed 비디오의 경우는 하나의 비디오에 포함되어있는 프레임의 문맥이 대체적으로 비슷하게 구성이 되어있기 때문에 이를 이용하여 모델을 학습 할 때, 균등한 간격으로 프레임을 선택하거나 또는 무작위로 프레임을 선택해서 사용해도 모델 학습에는 심각한 문제를 일으키지 않는다. 반면에 Untrimmed 비디오의 경우에는 의미 없는 프레임이 선택될 가능성이 높기 때문에 모델 학습에 심각한 문제를 일으킬 수 있다. 또한, Trimmed 비디오의 경우에는 사람에 의해 편집이 이루어져 많은 비용이 들기 때문에 Untrimmed 비디오에서 의미 있는 비디오 프레임을 추출하는 방법은 매우 중요하다.

최근 위와 같은 문제를 해결하기 위해서 Wang et al. [6]는 전체 비디오 프레임을 k개의 세그먼트로 분리하고 각 세그먼트에서 랜덤으로 하나의 프레임을 선택하는 방법을 사용하여 비디오의 범주를 예측하는 모델을 제안하였다. Korbar et al. [7]는 비교적 가벼운 2D CNNs 모델

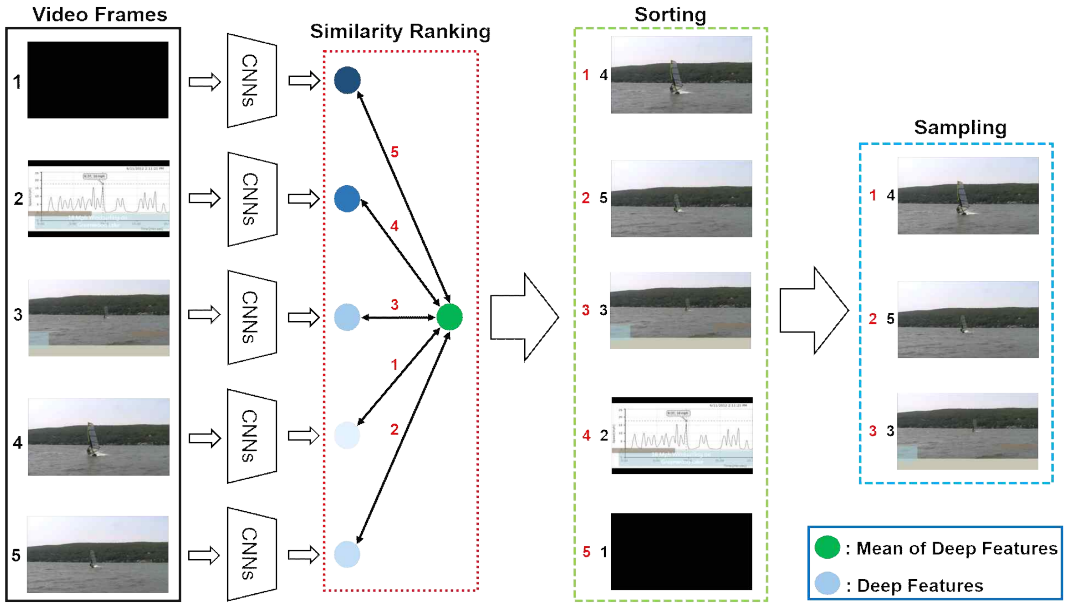

그림 1. 본 논문에서 제안하는 기법의 전체적인 과정을 나타낸 예시이다. 첫 번째 단계로 비디오의 모든 프레임(예시에서는 5개)을 RGB로 추출한다. 두 번째 단계로 2D CNNs인 VGG16을 사용하여 심층 특징을 추출한다. 세 번째 단계로 추출한 프레임의 심층 특징과 코사인 유사도 계산을 진행한다(빨간색 점선). 네 번째 단계로 계산된 코사인 유사도 점수로 계산된 유사도 순위(빨간색 숫자)를 바탕으로 비디오 프레임을 정렬한다. 마지막으로 정렬된 프레임에서 상위 k개를(예시에서는 k=3) 선택한다. 추가로 학습에 사용하기 위해 최종적으로 추출된 프레임을 한 번 더 원본 프레임 순서에 맞도록 다시 정렬을 해준다.

로 비디오 프레임 선택기를 구성하고 이를 비디오 프레임뿐만 아니라 비디오에 포함된 음성 데이터도 학습에 적극 이용하여 효과적인 프레임 추출 결과를 보여주었다. Wu et al. [8]는 비디오 프레임의 문맥 정보를 제공하는 증강된 전역 메모리를 이용하는 LSTM구조를 제안했으며 정책 경사(Policy Gradient) 기법을 이용하여 제안한 모델을 학습하고 유의미한 프레임을 추출하였다.

하지만 제안된 기법들은 유용한 비디오 프레임을 추출하기 위해 제안된 모델을 추가적으로 학습해야하는 문제점이 있기 때문에 그에 따른 비용이 발생한다. 따라서, 본 논문에서는 추가적인 모델 학습 없이 이미지넷[9]에 사전 학습된 2D CNNs 와 코사인 유사도만을 사용하여 비디오에서 의미 있는 프레임들을 추출한다.

본 논문의 2장에서는 본 논문에서 제안하는 기법에 대해서 자세히 설명한다. 3장에서는 제안하는 기법과 일반적으로 비디오 프레임 선택에 사용되는 균등한 간격 선택, 무작위 선택 방법들과 비교를 통해 본 논문에서 제안하는 기법이 실제로 잘 작동함을 보인다. 4장에서는 향후 연구에 대한 고찰과 정리로 마무리를 한다.

2. 심층 특징의 평균값을 활용한 비디오 프레임 추출 기법

이번 장에서는 본 논문에서 제안하는 기법에 대해서 자세히 설명한다. 본 논문에서 제안하는 기법은 전체적으로 다음과 같은 단계를 거친다.

1. 비디오 프레임 추출
2. 추출한 비디오 프레임에서 2D CNNs를 이용하여 심층 특징을 추출
3. 추출한 모든 심층 특징의 평균값을 계산하고 모든 심층 특징과 코사인 유사도를 계산
4. 계산한 코사인 유사도 점수를 바탕으로 내림차순으로 프레임들을 정렬
5. 정렬된 프레임에서 상위 k개의 프레임을 선택

첫 번째 단계는 비디오 데이터를 사용하기 전 진행하는 비디오 전처리 과정으로 추출한 비디오의 품질을 정하거나 프레임 사이즈 조절을 한다. 더 나아가서는 RGB 또는 광학 흐름(Optical Flow)으로 추출할지 여부 또한 정할 수 있다. 본 논문에서는 첫 번째 단계에서 기법의 간소화를 위해 간단하게 비디오 프레임을 RGB로 추출하였다. 두 번째 단계는 추출한 비디오 프레임에서 심층 특징을 추출하는 과정으로 보편적으로 많이 사용하는 2D CNNs중 하나인 이미지넷에 사전 학습된 VGG16[10] 모델을 사용하였다. 세 번째 단계는 추출한 심층 특징들의 평균값을 계산하고 추출된 모든 심층 특징과 평균값의 코사인 유사도를 계산하는 단계로 수식 1. 을 따른다.

$$\text{Smilarity Scores} = \frac{D \cdot \overline{D}}{\| D \| \ \| \overline{D} \|} \quad (1)$$

이때 D는 비디오 하나의 모든 프레임의 심층 특징을 나타내는 벡터로 $(n \times 4096)$사이즈를 가지며 n은 프레임의 개수를 나타낸다. 그리고 $\overline{D}$는 벡터 D에서 프레임 개수 축을 기준으로 계산한 평균값으로 (1×4096)사이즈를 가진다. 수식 1. 의 계산을 끝내면 모든 심층 특징의 평균값과 모든 심층 특징 사이의 코사인 유사도를 얻을 수 있으며, 이를 이용하여 네 번째 단계를 진행한다. 네 번째 단계에서는 앞서 구한

코사인 유사도를 이용하여 수식 2. 와 같이 프레임들의 인덱스를 구한다.

$$\text{Frame Indices} = \text{argsort}(\text{Similarity Scores}) \quad (2)$$

이때 argsort는 내림차순으로 값을 정렬하고 해당하는 값의 인덱스를 반환하는 함수이다. Frame Indices는 코사인 유사도를 기준으로 정렬이 되어있는 비디오 프레임 인덱스를 모아놓은 하나의 어레이다. 마지막으로 비디오 데이터 학습을 진행하기 위해서 프레임을 샘플링 할 때, Frame Indices 에서 상위 n개의 원소만 가져오고 다시 정렬 후 해당 인덱스에 해당하는 프레임을 샘플링한다. 이때, n은 학습에 사용할 비디오 프레임의 개수이다. 그림 1은 "Windsurfing" 범주를 가지는 비디오의 프레임에서 의미 있는 프레임을 추출하는 과정이며, 우리가 제안하는 기법은 1번 프레임과 2번 프레임과 같은 범주와 관련 없는 프레임을 제외한 의미 있는 프레임들을 추출 및 선택할 수 있다.

3. 실험 결과

이번 장에서는 Untrimmed 비디오 데이터셋 중에 하나인 ActivityNet 데이터셋을 이용해 본 논문에서 제안하는 기법과 일반적인 프레임 샘플링 방식 2가지 (균등한 간격, 무작위)를 간단히 비교 분석하였다.

그림 2는 3가지 범주에 해당하는 비디오에 각 기법을 적용하여 16개의 비디오 프레임을 추출한 결과이다. 그림 2에서 볼 수 있듯이 "Brushing hair"와 "Playing saxophone" 범주에 해당하는 비디오는 프레임마다 대체적으로 유사한 문맥을 유지하는 프레임이 많기 때문에 모든 결과가 비슷하게 나온다. 하지만 "Rollerblading" 범주에 해당하는 비디오의 경우에는 균등한 간격과 무작위 프레임 선택 방법은 모두 의미 없는 프레임을 선택했지만 본 논문에서 제안하는 기법은 범주와 관련된 의미 있는 프레임들을 추출하였다.

4. 결론 및 향후 계획

본 논문에서는 Untrimmed 비디오 데이터셋 중 하나인 ActivityNet 데이터셋을 이용하여 일반적인 프레임 샘플링 방식들과 비교를 통해 본 논문에서 제안한 기법이 Untrimmed 비디오에서도 의미 있는 비디오 프레임 추출이 가능함을 확인하였다.

하지만 본 논문에서 제안한 기법은 모든 프레임의 심층 특징의 평균값을 유용한 프레임의 기준으로 고정했기 때문에 문제가 있을 수 있다. 예를 들어, 비디오에 의미가 없는 프레임의 수가 범주와 관련된 프레임의 수 보다 많다면, 본 논문에서 제안한 기법은 잘 작동이 안 될 것이다. 따라서 이를 극복하기 위해 유용한 프레임의 기준으로 모든 심층 특징의 평균값이 아닌 해당 비디오를 잘 표현하는 새로운 심층 특징을 GAN 등을 이용하여 생성하면 어느 정도 문제를 유연하게 접근을 할 수 있을 것으로 기대를 한다. 또한 본 논문에서제안한 기법을 비디오 요약과 같은 학습뿐만 아니라 비디오 분석 분야에도 응용이 가능할 것으로 기대한다.

Acknowledgement

This work was supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (No. 2018R1C1B6007230).

참고문헌

[1] "YouTube During COVID-19", [Online]. Available: https://www.youtube.com/trends/articles/what-it-means-to-stayhome-on-youtube/ [Accessed: 2-Apr-2021].

[2] A. Alam, I. Ullah and Y. -K. Lee, "Video Big Data Analytics in the Cloud: A Reference Architecture, Survey, Opportunities, and Open Research Issues," *IEEE Access*, vol. 8, pp. 152377-152422, 2020, doi: 10.1109/ACCESS.2020.3017135.

[3] K. Soomro, A. R. Zamir and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," 2012, *arXiv:1212.0402*. [Online]. Available: https://arxiv.org/abs/1212.0402

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: A large video database for human motion recognition," in *Proceedings of the International Conference on Computer Vision*, Nov. 2011, pp. 2556-2563.

[5] F. C. Heilbron, V. Escorcia, B. Ghanem and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Conference on Computer Vision and Pattern Recognition*, June. 2015, pp. 961-970.

[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. V. Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in *European Conference on Computer Vision,* September, 2016, pp. 20-36.

[7] B. Korbar, D. Tran and L. Torresani, "SCSampler: Sampling Salient Clips From Video for Efficient Action Recognition," in *International Conference on Computer Vision*, October, 2019, pp. 6231-6241.

[8] Z. Wu, C. Xiong, C. Ma, R. Socher and L. S. Davis, "AdaFrame: Adaptive Frame Selection for Fast Video Recognition," in *Conference on Computer Vision and Pattern Recognition*, June, 2019, pp. 1278-1287.

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, June. 2014, pp. 1725-1732.

[10] K. Simonyan, A. Zisserman, Y. Bengio and Y. LeCcun, "Very Deep Convolutional Networks for Large-Scale Image Recognition," May, 2015, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

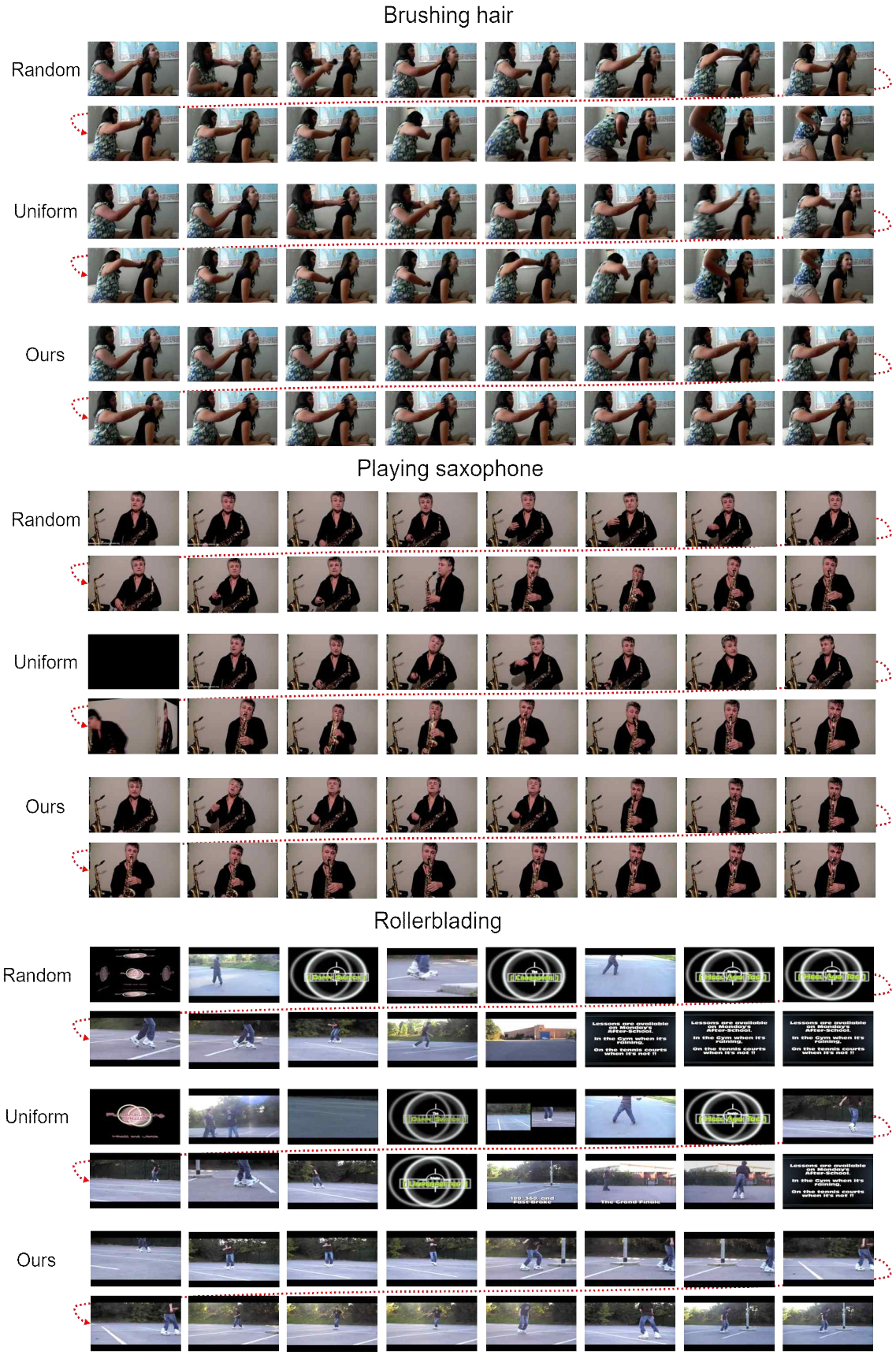

그림 2. ActivityNet 데이터셋에서 본 논문에서 제안한 기법과 균등한 간격 그리고 무작위로 프레임을 선택하는 방법과의 비교 결과로 총 3개의 비디오에 대해서 각 16개의 선택된 프레임의 결과를 보여준다.