GAN을 이용한 하이라이트 영상 예측 모델의 성능 개선

이한솔, 이계민
서울과학기술대학교 IT미디어공학과
{han8200, gyemin}@seoultech.ac.kr

Improving Highlight Prediction Models Using GAN

Hansol Lee, Gyemin Lee
Seoul National University of Science and Technology

요약

최근 다양한 개인방송 플랫폼에 의해 엄청난 양의 콘텐츠가 업로드 되고 있으며 그 중 축구와 야구와 같은 스포츠 영상이 차지하는 비율이 상당하다. 방송사에서는 시청자들의 편의를 위해 경기 영상 중 흥미를 끌거나 또는 중요한 장면을 모아 하이라이트 영상을 만들어 제공하는데, 이는 시간과 비용이 많이 소요되는 문제가 있다.

이에 본 논문에서는 스포츠 영상에서 자동으로 하이라이트를 예측하는 모델을 제안한다. 우리의 모델은 오디오와 이미지 정보를 함께 사용하며, 영상의 단기적 전후관계와 중장기적 흐름을 동시에 파악하는 모델을 제시한다. 또한 더 좋은 특징벡터를 추출하기 위해 GAN을 결합하는 방법을 설명한다. 제안하는 모델들은 야구 경기 영상을 이용하여 평가한다.

1. 서론

최근 온라인에서는 Youtub, Kakao TV[1]와 같은 여러 개인방송 플랫폼을 통해 엄청난 양의 콘텐츠가 업로드 되고 있으며, 특히 축구와 야구 같은 스포츠 영상이 주가 된다. 이러한 스포츠 영상의 수요가 증가함에 따라 방송사에서는 시청자들의 편의를 위해 편집된 하이라이트 서비스를 제공한다. 하지만 하이라이트 영상을 제작하는 것은 전문적인 기술과 장비를 요구하기 때문에 시간과 비용 면에서 문제가 있다. 이에 본 논문에서는 자동으로 하이라이트를 예측하는 모델을 제안한다.

이와 관련된 다양한 연구들이 진행되고 있다. [2]는 LSTM과 Determinantal Point Process (DPP)를 결합한 모델을 제안하였고, [3]은 Generative Adversarial Network (GAN)[4]을 이용하는 비지도 학습 알고리즘을 소개하였다. 또한 [5]는 encoder-decoder 알고리즘에 retrospective encoder를 추가한 계층적 모델을 제시하며, [6]은 오디오와 이미지 정보를 같이 사용함과 동시에 adversarial network를 응용하는 방법을 제안했다. 개인방송의 경우, 채팅과 오디오 데이터를 이용하여 하이라이트를 검출한 [7, 8]이 있다.

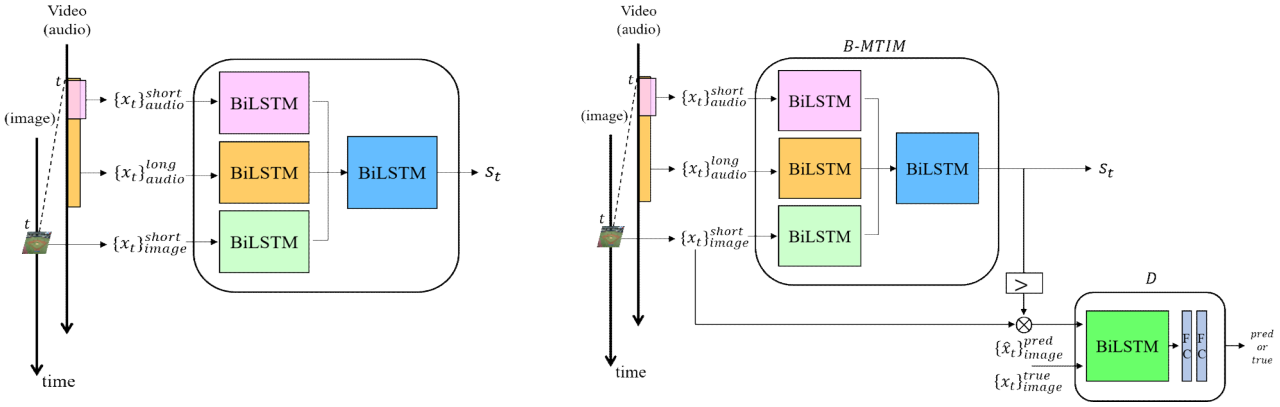
대부분의 연구는 이미지 정보만을 이용하여 하이라이트를 추출한다. 하지만 스포츠와 같은 경기 영상에서는 관중들의 호응과 해설자의 목소리 크기가 경기를 이해하는데 큰 도움이 된다. 따라서 우리는 오디오와 이미지 정보를 함께 사용하는 모델을 제안한다. 또한 스포츠 경기는 보통 한 순간의 이벤트만 봐서는 그 이벤트가 득점으로 이어지는가에 대한 판단이 어렵다. 이에 우리는 단기적 전후관계와 중장기적 흐름을 같이 파악하는 다중 시구간 모델를 이용한다. 이 때, 모델의 성능은 추출된 특징벡터가 얼마나 좋은가에 영향을 받게 된다. 따라서 우리는 GAN을 이용하여 더 좋은 특징 벡터를 추출할 수 있도록 하는 모델 개선 방법을 제시한다. 제안하는 모델들의 평가에는 직접 수집한 야구 경기 영상을 이용하였다.

2. 제안하는 알고리즘

이 장에서는 하이라이트를 자동으로 검출하기 위해 제안하는 모델들을 설명한다. 먼저 단기적 흐름과 중장기적 흐름을 동시에 파악하면서 오디오와 이미지 정보를 모두 이용하는 B-MTIM을 설명한다. 그 다음, GAN을 결합한 우리의 최종 모델을 제안한다.

2.1 B-MTIM

제안하는 B-MTIM 모델은 짧은 시구간의 흐름을 파악하는 양방향 LSTM과 중장기적 흐름을 파악하는 양방향 LSTM을 결합한다. B-MTIM 구조에 대한 그림이 그림 1의 (a)에 나타나있다. 첫 번째 LSTM은 짧은 시구간의 전후관계를 파악한다. 그러나 콘텐츠의 종류에 따라 단기적 흐름과 중장기적 흐름을 같이 고려해야 되는 경우가 있다. 예를 들어, 축구와 야구 같은 전통적인 스포츠는 선수들의 플레이가 이 후 득점으로 이어질지는 오래 지켜봐야 알 수 있다. 이를 위해 두 번째 LSTM이 이러한 중장기적 흐름을 고려하는 역할을 한다. 그리고 오디오와 이미지 정보를 모두 사용하기 위해 세 번째 LSTM이 추가된다. 즉, 짧은 구간의 오디오 특징벡터 x_{audio}^{short}, 긴 구간의 오디오 특징벡터 x_{audio}^{long}, 그리고 짧은 구간에 해당하는 이미지 특징벡터 x_{image}^{short}가 각각의 LSTM을 통과한 후 결합된다. 이어서 또 다른



(a) B-MTIM (b) B-MTIM-iGAN

그림 1. (a) B-MTIM (b) B-MTIM-iGAN

LSTM과 FC layer를 거쳐 하이라이트 스코어 s_t을 만든다. 이 때, B-MTIM은 손실함수

$$L_{CE} = \frac{1}{T}\sum_{t} \text{cross-entropy}(s_t, y_t) \quad (1)$$

을 최적화 한다. 여기서 y_t은 ground truth label을 의미한다.

2.2 B-MTIM-iGAN

이 절에서는 GAN을 결합한 하이라이트 예측 모델을 제안한다. GAN은 generator와 discriminator로 구성된 알고리즘으로, generator는 discriminator를 속이기 위해 실제와 유사한 가짜 데이터를 생성하고, discriminator는 generator로부터 만들어진 가짜 데이터와 실제 데이터를 구분해내면서 서로 대립 관계를 가진다.

우리의 모델에서는 generator 대신 앞에서 제안한 B-MTIM을 이용한다. 그림 1의 (b)는 B-MTIM과 discriminator D가 결합된 모델의 구조를 보여준다. 이 때, GAN은 대체적으로 이미지 데이터에 더 효과적이기 때문에, discriminator D가 이미지 특징벡터에 결합되어 실제 하이라이트와 모델이 만들어낸 하이라이트를 구별하면서 더 중요한 이미지 특징 벡터를 찾아낼 수 있도록 돕는다. x^{true}는 실제 하이라이트이고 x^{pred}는 B-MTIM에 의해 얻어진 score에서 상위 5%의 score를 가지는 frame의 x들을 나타낸다. B-MTIM은 discriminator D가 구분하지 못하도록 최대한 ground truth와 유사한 하이라이트를 만든다. discriminator D는 x^{true}와 x^{pred}을 정확히 구분하기 위해 학습한다. 따라서 우리의 모델은 다음의 최적화 문제를 풀게 된다.

$$\min_{B\text{-}MTIM}\max_{D} \; L_{CE} + \log D(x^{true}) + \log(1 - D(x^{pred})) \quad (2)$$

3. 실험 및 결과

제안하는 모델들을 평가하기 위해 2018년 4월부터 5월 초 기간 중에 Kakao TV에서 중계된 한국 프로 야구 경기영상 28개를 직접 수집하였다. 이 중 5개의 경기 영상을 테스트 데이터로 이용하였고 ground truth는 Naver-sports[9]에서 제작한 하이라이트 영상을 활용하였다. 이에 대한 요약을 표 1에 나타내었다. ground truth의 평균 길이는 약 600초로, 이는 전체 경기 영상의 대략 평균 5% 비율이며 실험에서도 전체 영상 길이의 5%를 하이라이트로 검출하였다. 짧은 구간은 1초, 긴 구간은 2분을 기준으로 실험을 진행하였다.

또한 학습에 이용된 데이터는 시간과 메모리의 효율적인 사용을 위해 미리 특징을 추출하였다. 정해진 구간(1초 또는 2분) 단위로 데이터를 나눈 다음, 오디오는 Mel Frequency Cepstral Coefficient (MFCC)[10]를 이용하여 각 구간 별로 500차원의 특징벡터 x_{audio}을 추출하였고, 이미지는 ResNet-34[11]를 이용하여 512차원의 특징벡터 x_{image}을 추출하였다.

정량적 평가를 위해 F-score를 이용한다. F-score는 비디오 요약에 많이 사용되는 성능지표로, 정밀도(precision)와 재현율(recall)의 조화평균으로 구할 수 있다. 수식은 다음과 같다.

$$P = \frac{|H_{gt} \cap H_{pred}|}{|H_{pred}|}, \quad R = \frac{|H_{gt} \cap H_{pred}|}{|H_{gt}|} \quad (3)$$

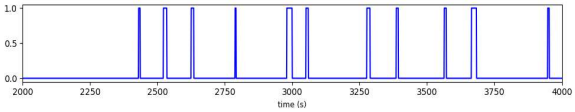
$$F\text{-}score = \frac{2PR}{P+R} \times 100\,\% \quad (4)$$

여기서 H_{gt}와 H_{pred}는 각각 ground truth와 모델에 의해 예측된 결과를 의미한다.

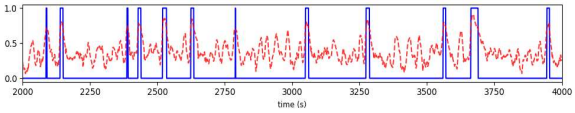
그림 2는 제안하는 모델들의 결과의 일부(2000~4000초)를 시각적으로 보여준다. 그림 2의 (a)는 ground truth이고 파란 선은 하이라이트의 유무를, 빨간 점선은 하이라이트 score를 나타낸다. (b)

표 1. 야구 데이터 요약 정보

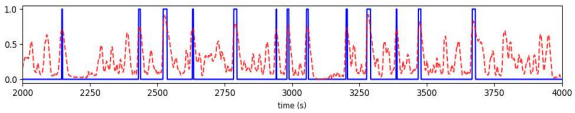
Type	Statistics	Video length (sec)	Length of highlights (sec)	Highlight ratio (%)
Baseball	mean (±std)	12,175.39 (±1,176.13)	599.25 (±225.34)	4.95 (±1.93)
	max	14,866	1,361	12.59
	min	9,909	76	0.61



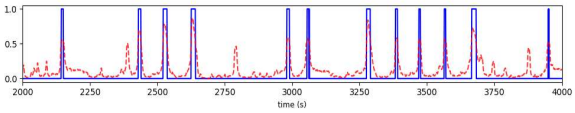
(a) ground truth



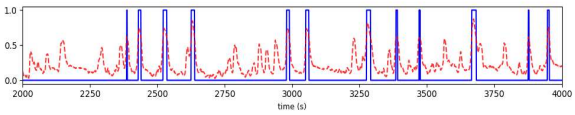
(b) STIM (image)



(c) MTIM (audio)



(d) B-MTIM (audio+image)



(e) B-MTIM-iGAN (proposed)

그림 2. 실험 결과 비교 (2000~4000초)

Single-Time Interval Model (STIM)와 (c) MTIM는 각각 오디오와 이미지 정보를 모두 사용하는 (d) B-MTIM에서 오디오에 대한 다중 시구간 LSTM을 제외하고 단일 시구간의 이미지 정보만을 이용하는 모델과, 반대로 오디오에 대한 다중 시구간 LSTM만을 사용한 모델을 의미한다. (b) STIM은 2000초에서 2200초 구간을 잘못 예측 하였으며, (c) MTIM은 3900초를 예측하지 못하였다. 반면에 (d) B-MTIM의 2900초에서 3200초 구간은 ground truth와 더 유사함을 알 수 있다. 그리고 GAN을 결합한 우리의 최종 모델 (e) B-MTIM-iGAN은 ground truth와 가장 근접한 결과를 보인다.

실험에 대한 정량적 결과는 표 2를 통해 확인할 수 있다. 실험 결과를 보면 STIM과 MTIM이 각각 53.65, 57.57을 가진다. 그리고 B-MTIM은 61.90로 하나의 정보만을 이용하는 모델들보다 더 높은 F-score를 가진다. 이 결과로부터 오디오와 이미지 정보를 같이 이용하는 것이 영상의 하이라이트를 예측하는데 더 효과적임을 확인하였다. 그리고 우리의 최종 모델 B-MTIM-iGAN은 63.57로 가장 높은 F-score를 갖는다. 따라서 GAN이 성능 향상에 효과적임을 알 수 있다.

4. 결론

우리는 야구경기와 같은 콘텐츠의 단기적 흐름과 중장기적 흐름을 함께 파악하는 B-MTIM을 제안하였고 오디오와 이미지 정보를 동시에 사용하였다. 또한 하이라이트 예측에 있어 더 좋은 특징벡터를 얻기 위해 GAN을 결합한 최종 모델을 설명하였으며 다른 모델들과 비교하였을 때 가장 높은 성능을 가짐을 정량적, 시각적으로 확인하였다.

표 2. 실험 결과

Data	Model	F-score
Image	STIM	53.65
Audio	MTIM	57.57
Image + Audio	B-MTIM	61.90
	B-MTIM-iGAN (proposed)	**63.57**

Acknowledgments

본 논문은 한국 연구 재단(NRF-2017R1E1A1A03070596)의 지원을 받아 수행되었음.

참고문헌

[1] Kakao TV, https://tv.kakao.com/

[2] K. Zhang, W.-L. Chao, F.Sha, and K.Grauman, “Video Summarization with Long Short-term Memory,” In ECCV, 2016.

[3] B. Mahasseni, M. Lam, and S. Todorovic, “Unsupervised Video Summarization with Adcersarial LSTM Networks,” In CVPR, pp. 2982-2991, 2017.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” In NIPS, pp. 2672-2680, 2014.

[5] K. Zhang, K. Grauman, and F. Sha, “Retrospective Encoders for Video Summarization,” In ECCV, 2018.

[6] H. Lee, G. Lee, “Summarizing Long-Length Videos with GAN-Enhanced Audio/Visual Features,” In ICCV workshop, 2019. (accepted)

[7] 김은율, 이계민, “채팅 트래픽 분석을 통한 개인방송 하이라이트 검출,” 방송공학회논문지, Vol. 23, No. 2, pp. 218-226, 2018.

[8] 김은율, 이계민, “채팅과 오디오의 다중 시구간 정보를 이용한 영상의 하이라이트 예측,” 방송공학회논문지, Vol. 24, No. 4, pp. 553-563, 2019.

[9] Naver-sports, https://sports.news.naver.com/

[10] S. Davis, P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol, 28, No. 4, pp. 357-366, 1980.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” In CVPR, 2016.