일반논문 (Regular Paper)
방송공학회논문지 제25권 제4호, 2020년 7월 (JBE Vol. 25, No. 4, July 2020)
https://doi.org/10.5909/JBE.2020.25.4.620
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

하이라이트 검출을 위한 구간 분할 앙상블 모델

이 한 솔[a], 이 계 민[a)‡]

Subdivision Ensemble Model for Highlight Detection

Hansol Lee[a] and Gyemin Lee[a)‡]

요 약

하이라이트를 자동으로 예측 하는 문제는 영상을 사람이 직접 편집하는 시간과 비용 문제를 해결하기 위해 필요한 기술이다. 본 논문에서는 하이라이트 구간 내에서 하이라이트 판단 여부에 영향을 주는 특정 부분에 집중하기 위해 앙상블 모델을 제안한다. 우리의 모델은 하나의 단일 모델만으로는 충분히 학습하기 어려운 중요한 정보를 앙상블을 통해 더 많은 유용한 특징들을 얻을 수 있다. 앙상블을 이루는 단일모델들은 오디오와 이미지 정보를 결합하여 다양한 영상의 특징들을 추출한다. 직접 수집한 e스포츠 경기 영상과 야구 경기 영상을 통해 하이라이트 예측 성능이 개선됨을 확인한다.

Abstract

Automatically predicting video highlight is an important task for media industry and streaming platform providers to save time and cost of manual video editing process. We propose a new ensemble model that combines multiple highlight detectors with each focusing on different parts of highlight events. Therefore, our model can capture more information-rich sections of events. Furthermore, the proposed model can extract improved features for highlight detection particularly when the train video set is small. We evaluate our model on e-sports and baseball videos.

Keyword : Video highlight, Ensemble model, BiLSTM, Event subsection, Event subdivision

a) 서울과학기술대학교 일반대학원 미디어IT공학과(Dept. of Media IT Engineering, Graduate School, Seoul National University of Science and Technology)
‡ Corresponding Author : 이계민(Gyemin Lee)
E-mail: gyemin@seoultech.ac.kr
Tel: +82-02-970-6416
ORCID: https://orcid.org/0000-0001-6785-8739
※ This study was supported by the Research Program of Seoul National University of Science and Technology.
· Manuscript received June 1, 2020; Revised July 23, 2020; Accepted July 23, 2020.

Ⅰ. 서 론

하이라이트 영상은 방송사 또는 동영상 스트리밍 서비스를 하는 기업에서 시청자들의 편의와 네트워크의 원활한 운영을 위해 제공된다. 특히 축구와 야구같이 경기 시간이 매우 긴 영상의 경우, 시청자들은 원본의 긴 영상보다는 짧은 하이라이트 영상들을 더 많이 선호한다. 그러나 하이라

이트 영상을 제작하기 위해서는 전문가들의 기술과 편집 장비가 필요하며 시간과 비용적인 문제가 따른다. 이에 본 논문에서는 자동으로 하이라이트를 예측하는 모델을 제안한다.

하이라이트 영상을 만들기 위해서는 시청자들의 흥미를 끄는 장면 또는 중요한 이벤트를 찾을 수 있어야 한다. 하이라이트 영상 구간은 해당 이벤트에 대한 정보를 포함하게 된다. 이 때 하나의 이벤트 구간은 여러 개의 소구간으로 나뉠 수 있으며, 이벤트에 대한 정보는 각 소구간에 불균일하게 분포 할 수 있다. 다시 말해, 같은 이벤트를 구성한다 하더라도 이벤트의 특징을 잘 보여주는 부분이 일부 구간에 집중되어 있을 수 있다. 예를 들어, 야구 경기에서 안타에 해당하는 이벤트는 투수가 준비동작을 한 뒤 공을 던지는 초반부, 타자가 투수의 공을 쳐내고 베이스 또는 홈을 향해 달려 나가는 중반부, 홈으로 들어와 점수를 득점하거나 공을 잡아 아웃시키는 후반부로 나눌 수 있다. 여기서 타자가 투수의 공을 받아치고 달려 나가는 중반부가 가장 중요하다고 볼 수 있다. 이처럼 어떤 장면에 더 집중하느냐에 따라 해당 구간의 하이라이트 포함 여부를 보다 더 잘 결정할 수 있다.

따라서 본 연구에서는 어떤 소구간이 하이라이트 결정에 많은 영향을 미치는지를 파악하고 하이라이트 각 구간 내의 특정 부분에서 더 많은 특징들을 추출하기 위한 앙상블 모델을 제안한다. 또한 앙상블 모델이 학습 데이터 수가 적은 경우에도 효과적으로 정보를 획득하는지 확인하기 위해 데이터 개수에 따른 앙상블 결과를 비교한다. 앙상블을 이루는 우리의 단일모델은 오디오 정보와 이미지 정보를 결합하여 하이라이트 검출에 이용한다. 제안하는 앙상블 모델은 직접 수집한 e스포츠 경기 영상과 야구 경기 영상을 통해 평가하며, 데이터 개수에 따른 모델의 성능을 확인하였다.

Ⅱ. 관련 연구

비디오 데이터를 활용하는 많은 연구들이 진행되고 있으며 하이라이트 검출과 관련도가 높은 연구는 비디오 요약과 영상에서의 특정 이벤트 검출이 대표적이다. Zhang 등은 비디오 요약 문제를 해결하기 위해 Long Short-Term Memory (LSTM)를 사용하며 Determinantal Point Process (DPP)와 결합한 지도 학습 알고리즘을 제안하였다[1]. 강화학습 모델을 제안한 Zhou 등은 영상에서 다양성과 대표성을 띄는 프레임을 찾기 위해 reward를 부여하고 그에 따른 우선순위가 높은 프레임들을 선택하는 알고리즘을 제시하였다[2]. Zhao 등은 비디오의 계층적 특징을 이용하여 장면이 전환되는 경계에 위치하는 프레임들을 검출하고 shot단위로 묶어 요약을 하는 모델을 설명하였다[3].

한편 기계번역에서 처음 도입되어 주로 사용되는 encoder-decoder 구조에 기반을 둔 연구도 있다. Mahasseni 등은 encoder로 비디오를 압축한 정보를 가지고 비디오를 요약한 후 선택된 프레임들의 정보를 가지고 decoder로 비디오를 복원한다[4]. 이때 복원된 비디오와 원본 비디오를 비교하기 위해 Generative Adversarial Network (GAN)[5]를 이용한 비지도 학습 알고리즘을 소개하였다. 또한 Zhang 등은 encoder-decoder에 또 다른 retrospective encoder를 추가하여 LSTM의 의존성을 보완한 계층적 모델을 제안하였다[6].

위의 방법들은 비디오에서 시각적인 정보만을 추출하여 연구를 진행했다면, 오디오 또는 텍스트 정보를 활용한 연구도 최근 제안되었다. Lee 등은 오디오와 이미지 정보를 함께 사용하면서 GAN을 결합한 모델을 제시하였고[7], 영상의 짧은 전후관계와 동시에 중장기적 흐름을 파악하는 다중 시구간 모델을 설명하였다[8]. 또한 개인방송에서 얻을 수 있는 채팅내역을 텍스트 정보로 이용하여 오디오 정보와 함께 하이라이트를 검출한 연구도 이루어졌다[9,10].

Ⅲ. 하이라이트 예측 알고리즘

제안하는 앙상블 알고리즘은 오디오 정보와 이미지 정보를 함께 이용하는 단일 모델들로 구성된다. 각각의 개별 모델들은 서로 다른 3 종류의 ground truth 비율로 학습이 이루어지고, 모두 결합되어 하나의 앙상블 모델을 만든다.

1. 하이라이트 예측 단일모델

비디오 요약 또는 장면 검출과 같이 비디오 데이터를 다

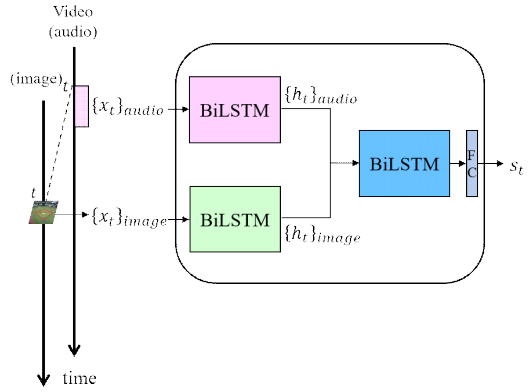


그림 1. 오디오/이미지 정보를 결합한 단일모델
Fig. 1. Single model that combines audio/image information

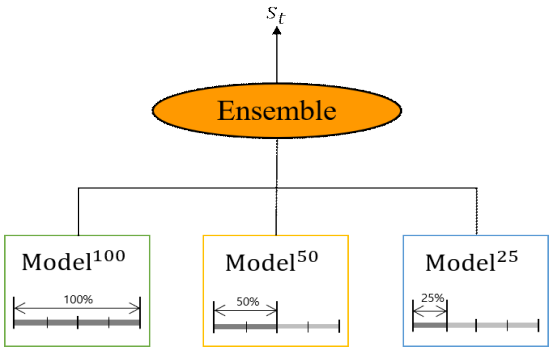


그림 2. 단일모델을 결합한 최종 앙상블 모델
Fig. 2. The proposed ensemble model

루는 기존의 연구들은 대부분 영상의 이미지 정보를 주로 이용하지만 이미지 정보와 같이 오디오 정보 또한 영상의 특징을 파악할 수 있는 유용한 정보를 많이 포함한다. 이에 본 논문에서는 이미지 정보와 함께 오디오 정보를 이용하는 모델을 앙상블의 단일모델로 사용한다.

그림 1은 오디오와 이미지 정보를 결합하는 단일모델의 구조를 나타낸다. 오디오의 특징 벡터 x_{audio}와 이미지의 특징 벡터 x_{image}는 각각 BiLSTM의 입력으로 들어가서 h_{audio}와 h_{image}를 출력한다. h_{audio}와 h_{image}는 함께 결합되어 또 다른 BiLSTM을 통해 하이라이트 스코어 s_t를 생성한다. 하이라이트 스코어 s_t는 해당 frame이 얼마나 하이라이트에 포함될 만한 장면인지를 판단하는 지표가 된다.

제안하는 하이라이트 예측 모델은 이 단일모델의 앙상블로 구성된다.

2. 제안하는 하이라이트 예측 앙상블 모델

하이라이트를 대부분의 시청자들이 흥미를 느끼는 이벤트들의 집합으로 볼 때, 각 이벤트들은 하이라이트로 분류되는 중요한 부분을 포함하고 있다. 일례로 e스포츠의 경우 캐릭터간의 싸움이 상대적으로 가장 치열한 중반부나 어느 한 팀에 치명적인 결과를 가져오는 마지막 순간이 하이라이트 내에서 가장 중요한 순간임을 예상해 볼 수 있다. 이렇게 각 부분에 포함된 이벤트에 대한 정보는 불균일하므로 중요한 구간에서 더 많은 정보를 획득할 수 있다면 보다 우수한 하이라이트 예측 알고리즘을 만들 수 있을 것이다.

이에 본 논문에서는 이벤트 구간을 세부 구간으로 나누고 이벤트에 대한 정보를 보다 많이 포함하는 부분을 더욱 중점적으로 고려하는 앙상블 모델을 제안한다. 우리의 앙상블 모델은 각 하이라이트 이벤트의 구간 비율을 조절한 다수의 단일모델로 구성된다.

제안하는 앙상블 모델은 그림 2와 같이 총 3개의 단일모델을 가지며, 각 단일 모델들은 서로 주목하는 부분을 달리 하기 위해 ground truth에 해당되는 이벤트의 길이의 비율을 다르게 하여 학습한다.

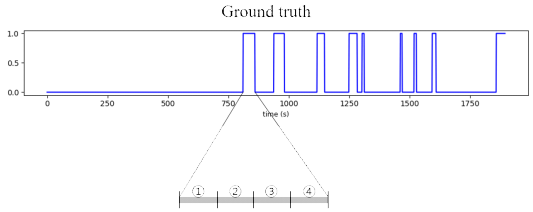


그림 3. 각 하이라이트 구간을 4등분하는 예시
Fig. 3. A highlight event divided into four subsections

Ground truth 비율을 나누는 기준은 그림 3을 통해 설명할 수 있다. 각각의 하이라이트 구간을 4등분을 하여 ①, ②, ③, ④라고 할 때, 해당 이벤트 전체(100%)를 보는 첫 번째 단일모델 $Model^{100}$(①,②,③,④)과 50%만을 보는 두 번째 단일모델 $Model^{50}$(초반부①,② 또는 중반부②,③ 또는 후반부③,④), 그리고 $Model^{50}$에서 선택된 50% 중에 25%

만 보는 마지막 단일모델 Model^{25}(① 또는 ② 또는 ③ 또는 ④)로 앙상블을 구성한다. 이벤트의 전반부, 중반부, 후반부를 다양하게 고려하여 어느 부분에 더 많이 집중 할 때 가장 성능이 우수한지를 파악한다. 각 단일모델로 하이라이트 스코어를 구한 후 이들의 평균으로 최종 하이라이트 스코어$s_t^{ensemble}$을 생성하며 다음의 수식과 같다.

$$s_t^{ensemble} = \frac{1}{3}(s_t^{100} + s_t^{50} + s_t^{25}) \quad (1)$$

여기서 s_t^{100}은 이벤트 전체를 고려하는 첫 번째 단일모델, s_t^{50}은 50%만을 고려하는 두 번째 단일모델, 그리고 s_t^{25}는 25%만을 보는 세 번째 단일모델이 생성한 각각의 하이라이트 스코어를 의미한다. 하이라이트 영상은 최종 스코어 $s_t^{ensemble}$가 높은 프레임들을 모아 만들 수 있다.

Ⅳ. 실험 및 결과

실험을 위해 Twitch[11]에서 2017년도에 중계한 'League of Legends' 대회 5개의 일부 경기 영상과 Kakao TV[12]에서 2018년 4월부터 5월 사이에 중계한 한국 프로 야구 경기 영상을 수집하였다. 실험 시간을 단축하기 위해 모든 경기영상에서 미리 특징 벡터를 추출하여 실험을 진행하였다. 오디오 정보는 Mel-Frequency Cepstral Coefficients (MFCC)를 이용하여 40ms당 20차원의 특징을 추출한 뒤 25개씩 묶어 1초당 500차원을 가지는 특징벡터를 얻었다. 이미지 정보는 ImageNet[13]으로 사전 학습된 ResNet-34[14]로 1fps 당 512차원의 특징벡터를 생성하였다. 단일 모델의 모든 BiLSTM은 512개의 hidden unit을 가지고 학습하였으며, 모델은 Pytorch로 구현하였다.

정량적 평가는 F-score를 활용하였다. F-score는 비디오 데이터에 많이 사용되는 성능 평가 방법으로, 정밀도(precision)와 재현율(recall)의 조화 평균으로 구해진다.

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

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

위 수식에서 H_{gt}와 H_{pred}는 각각 ground truth와 모델에 의해 선택된 하이라이트 구간을 의미한다.

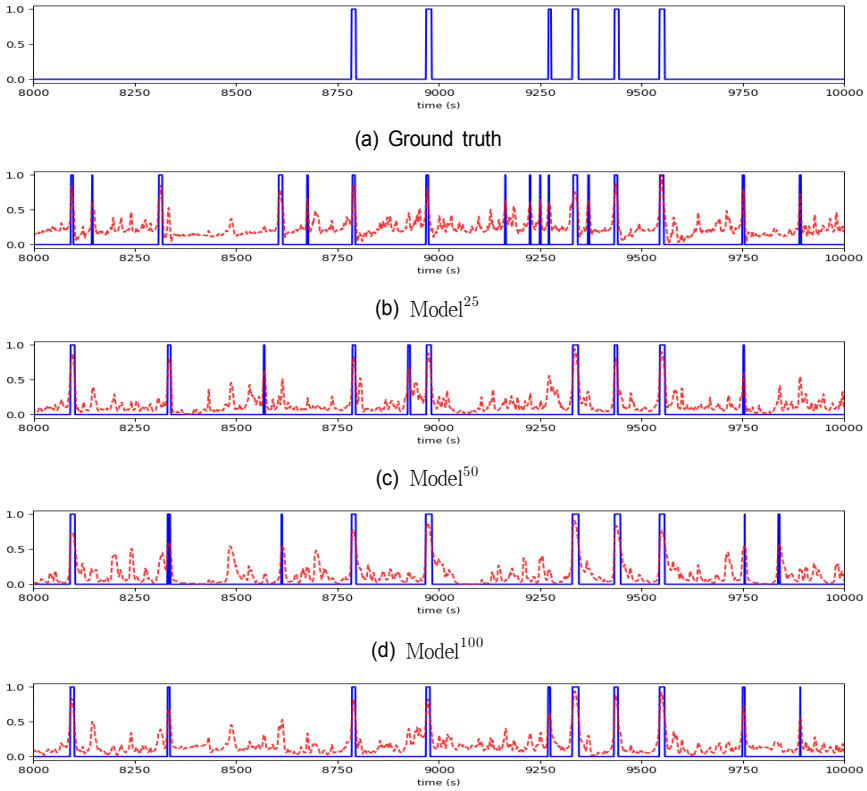
1. Baseball dataset

2018년 4월부터 5월 사이에 열린 한국 프로 야구 경기 영상 28개 중 5개를 테스트로 나머지 23개를 학습 데이터로 사용하였다. 표 1에 데이터에 대한 자세한 통계가 나타나있다. 경기 영상의 평균 길이는 약 200분, 하이라이트 영상의 평균은 약 10분이므로 전체 영상의 5% 정도의 비율이 하이라이트로 만들어지며, 실험에서도 상위 5%의 프레임들을 하이라이트로 검출하였다. Ground truth는 Naver-sports[15]에서 제공하는 하이라이트 영상을 이용하였다.

그림 4를 통해 시각적 결과를 볼 수 있다. 파란색 실선은 하이라이트 구간이 아니면 0, 하이라이트 구간이면 1을 나타내고 빨간색 점선은 하이라이트 스코어 s_t값을 의미한다. 야구 경기는 e스포츠에 비해 경기 시간이 매우 길기 때문에 8000초에서 10000초에 해당하는 일부 구간의 결과만을 표시하였다. 그림 4(a)는 ground truth를 나타낸다. 그림 4(b)와 그림 4(c)는 ground truth의 일부만 보는 단일모델 Model^{25}와 Model^{50}을 의미하고 그림 4(d)는 ground truth 전체를 보는 단일모델 Model^{100}의 결과이다. 이 때 빨간 점

표 1. e스포츠와 야구경기 데이터 요약 정보
Table 1. Summary of e-Sports and baseball data sets

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
e-Sports	mean (±std)	2,096.76 (±599.10)	213.27 (±70.99)	10.55 (±3.78)
	max	4,785	469	22.30
	min	1,483	146	9.84



(a) Ground truth

(b) Model^{25}

(c) Model^{50}

(d) Model^{100}

(e) Ensemble (proposed)

그림 4. 야구 영상에서 Type3 모델별 실험 결과 (파란 실선: 하이라이트 구간, 빨간 점선: 하이라이트 스코어)
Fig. 4. Experiment results on a Baseball test video (blue: highlight labels, red: highlight scores)

선은 각각 스코어 s_t^{25}, s_t^{50}, s_t^{100}을 의미한다. Model^{100}이 Model^{25}, Model^{50}보다 잘못 예측한 부분이 적으며 ground truth와 근접함을 확인할 수 있다. 최종 앙상블 모델의 결과는 그림 4(e)에 나타내었으며 빨간 점선은 $s_t^{ensemble}$이다. 앙상블 모델(그림 4(e))은 9250~9600초 구간의 4개의 하이라이트 구간을 ground truth와 매우 유사하게 예측하였으며 Model^{100}(그림 4(d))이 잘못 예측한 8600초 구간의 하이라이트 구간을 제거하였다.

정량적 결과는 표 2를 통해 F-score값으로 비교 할 수 있다. 표 2에서는 각각 이미지와 오디오 정보 하나의 정보만을 사용한 모델(BiLSTM+FC layer)의 결과도 보여주고 있다. 하나의 정보만을 사용하는 모델들은 각각 55.33과 53.65 값을 가지는데 반해서 두 정보를 모두 사용하는 Model^{100}은 57.74로 약 2~4% 향상된 F-score를 가진다. 특히 우리의 최종 앙상블 모델은 59.17로 가장 우수한 성능을 나타낸다. 한편 비디오 요약 연구에서 잘 알려진 DPP[1]는 우리의 야구 데이터에서 10%미만의 결과를 보였고, SUM-GAN[4]은 모델의 메모리 문제로 학습이 도중에 중단되었다.

표 3을 통해 하이라이트 구간의 어떤 부분에 집중했을 때 도움이 되었는가를 확인 할 수 있다. Type 3에서 가장 높은 결과를, Type 6에서 가장 낮은 결과를 가지며 이 때, Type 3은 각각의 하이라이트 구간에서 초중반 부분(그림 3에서 ②와 ③)에 집중하는 앙상블 모델이고 Type 6은 후반부(그림 3에서 ⑤와 ⑥)에 집중하는 경우이다. 즉, 야구 경기에서 타자가 공을 치고 달려 나가는 초중반부가 하이라이트 예측에 좋은 영향을 주는 것으로 해석된다. 반면에 선수가 득점에 성공하거나 아웃되어 이벤트가 마무리되어 정리되는 후반부는 하이라이트 선택에 있어 많은 도움이 되

표 2. e스포츠와 야구 데이터에 대한 실험 결과 (F-score)
Table 2. Experiment results (F-score) on e-sports and baseball data sets

Data type	Model	Baseball (%)	e-Sports 28 videos (%)	e-Sports 56 videos (%)
Image	DPP [1]	5.46	-	19.72
	SUM-GAN [4]	model breakdown	-	34.08
	$Model^{100}$ (i)	55.33	61.16	68.94
Audio	$Model^{100}$ (a)	53.65	59.23	66.55
Image + Audio	$Model^{100}$ (i+a)	57.74	67.09	69.65
	Ensemble	59.17	68.47	64.94

표 3. 하이라이트 이벤트 서브구간별 앙상블 모델의 비교 (F-score)
Table 3. Comparison of ensemble models using different subdivisions of highlight events

Ensemble Type \ Dataset		Baseball (%)	e-Sports 28 videos (%)	e-Sports 56 videos (%)
$Model^{100}$ (i+a) only		57.74	67.09	69.65
Type 1	$Model^{50}$: ①②, $Model^{25}$: ①	56.26	64.63	61.91
Type 2	$Model^{50}$: ①②, $Model^{25}$: ②	59.15	65.91	64.51
Type 3	$Model^{50}$: ②③, $Model^{25}$: ②	59.17	68.47	64.94
Type 4	$Model^{50}$: ②③, $Model^{25}$: ③	58.28	65.60	63.69
Type 5	$Model^{50}$: ③④, $Model^{25}$: ③	55.67	62.48	63.42
Type 6	$Model^{50}$: ③④, $Model^{25}$: ④	51.74	62.71	54.81

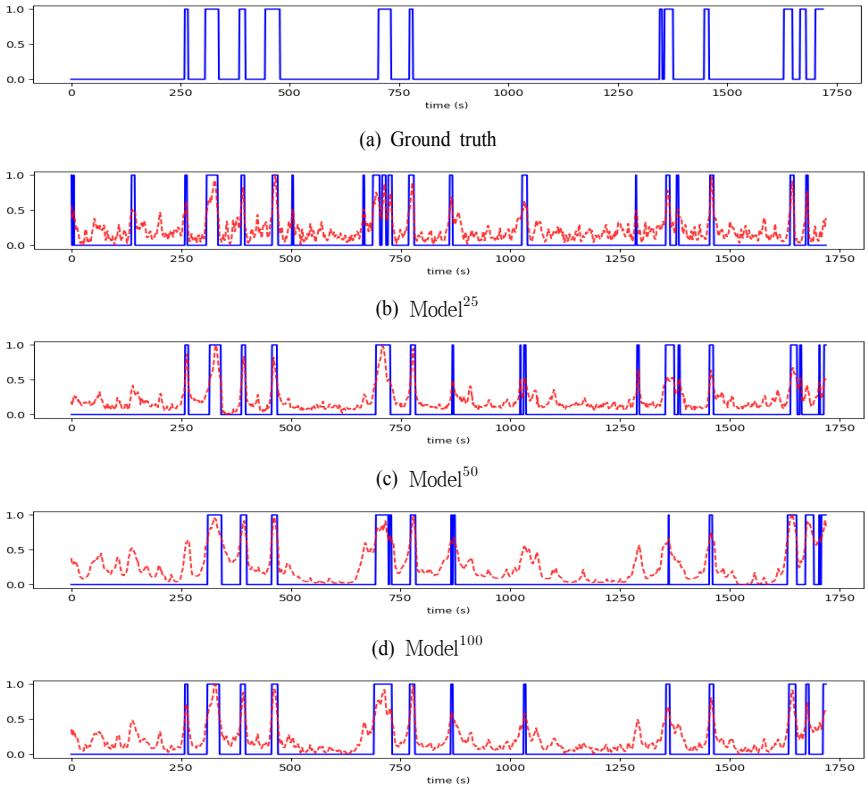
지 않는 것으로 판단할 수 있다.

2. e-Sports dataset

2017년에 개최된 'League of Legends' 대회 5개(IEM World Championship Katowice 2017, 2017 LoL World Championship, LoL All Star 2017, 2017 LoL Champions Korea Spring, 2017 LoL Champions Korea Summer)에서 수집한 63개의 경기 영상 중에서 7개의 경기 영상을 테스트로, 나머지 56개를 학습 데이터로 이용하였다. 표 1을 보면, 영상의 평균길이가 약 30분이고 하이라이트 영상의 평균길이는 약 3분이므로 하이라이트 비율이 대략 10%이다. 따라서 우리의 모델에서도 상위 10%의 하이라이트 스코어를 가지는 프레임들을 하이라이트로 선택하였다. 또한 모든 경기 영상의 ground truth는 OGN[16]에서 제공하는 하이라이트 영상을 활용하였다.

e스포츠 데이터의 경우 야구 데이터보다 학습 데이터의 개수가 더 많기 때문에 학습 데이터 개수에 차별을 주어 실험을 진행하였다. 56개의 학습데이터 중에서 절반에 해당하는 28개의 학습데이터에 대한 시각적 결과가 그림 5에 있다. 야구 결과와 마찬가지로 그림 5(a)는 ground truth이고 그림 5(b)~(d)는 각각 단일 모델 $Model^{25}$, $Model^{50}$, 그리고 $Model^{100}$이다. 우리의 최종 앙상블 모델(그림 5(e))은 $Model^{100}$이 예측하지 못한 250초 구간의 하이라이트를 예측하면서 전체적으로 $Model^{100}$보다 더 ground truth에 근접함을 확인 할 수 있다.

이러한 결과는 정량적으로도 확인할 수 있다. 표 2의 28개의 학습데이터를 사용한 e스포츠 결과를 보면, 하나의 정보만을 사용한 두 모델의 경우 F-score가 각각 61.16과 59.23이고 두 정보를 모두 활용한 단일모델 $Model^{100}$은 약 6~8% 증가한 67.09 값을 가진다. 모든 학습 데이터 56개를 사용한 결과 또한 단일모델이 69.65로 약 1~3% 향상된 결과를 보인다. 그리고 최종 앙상블 모델은 28개의 학습데이터로 진행한 실험에서는 68.47로 $Model^{100}$보다 더 좋은 성능을 보이지만, 반면에 56개의 학습데이터를 사용한 실험에서는 64.94로 $Model^{100}$보다 많이 부족한 결과를 보인다. 한편 DPP와 SUM-GAN은 우리의 e스포츠 데이



(a) Ground truth

(b) Model^{25}

(c) Model^{50}

(d) Model^{100}

(e) Ensemble (proposed)

그림 5. e스포츠 영상에 대한 Type3 모델별 실험 결과 (파란 실선: 하이라이트 구간, 빨간 점선: 하이라이트 스코어)
Fig. 5. Experiment results on an e-Sports test video (blue: highlight labels, red: highlight scores)

터에 대해 35%미만의 결과를 보였다.

28개의 학습데이터로 학습한 앙상블 모델이 영상의 어느 부분을 집중적으로 보았을 때 도움이 되었는가를 확인하기 위해 앙상블 타입 별 결과를 표 3에 나타내었다. 가장 높은 F-score를 보이는 경우는 야구 데이터의 결과와 동일하게 Type 3이며 초중반 부분을 집중적으로 보는 앙상블 모델이다. 반면에 Type 5와 6은 중후반 부분을 집중적으로 보는 경우인데 단일모델 Model^{100}에 비해 약 4% 정도 감소한 F-score값을 볼 수 있다. 즉, e스포츠 경기에서 캐릭터들끼리의 전투가 종료되는 후반부 보다는 전투가 격렬하게 이루어지고 있는 초중반 부분이 하이라이트를 예측하는데 좋은 영향을 준다고 해석 할 수 있다.

데이터 크기에 따른 성능을 확인하기 위해 학습 데이터 56개 중 7개, 14개, 21개, 35개, 42개를 선택하여 Type 3에 해당하는 앙상블 모델을 학습하였다. 학습 데이터 개수에 따른 결과가 표 4에 있다. 그리고 그림 6에서도 확인 할 수 있다. 파란색 커브가 Model^{100}의 결과를, 빨간색 커브가 앙상블 모델의 결과를 가리킨다. 단일모델 Model^{100}과 앙상블을 적용한 모델(Type3)의 F-score를 비교하였을 때, 28개의 학습 데이터까지는 앙상블 모델이 더 성능이 우수하지만 35개의 학습데이터부터는 성능이 많이 저하되는 결과를 보인다. 즉, 앙상블은 학습 데이터가 적을 때 보다 우수한 결과를 이끌어냄을 확인 할 수 있다. 이는 모든 단일 모델들의 학습시간이 Model^{100}을 기준으로 정해져있기 때문에, Model^{50}과 Model^{25}이 데이터 개수에 따라 학습이 덜되거나 과하게 되는 경향이 있어 이러한 결과가 나타난 것으로 보인다.

표 4. 앙상블 Type 3에서의 e스포츠 학습 데이터 개수 별 실험 결과
Table 4. Experiment results (F-score) as the number of e-sports train video is varied

	7 videos	14 videos	21 videos	28 videos	35 videos	42 videos	56 videos
$Model^{100}$ (i+a)	65.19	65.56	67.21	67.09	67.22	67.50	69.65
Ensemble Type 3	65.61	67.55	69.09	68.47	62.88	64.61	64.94

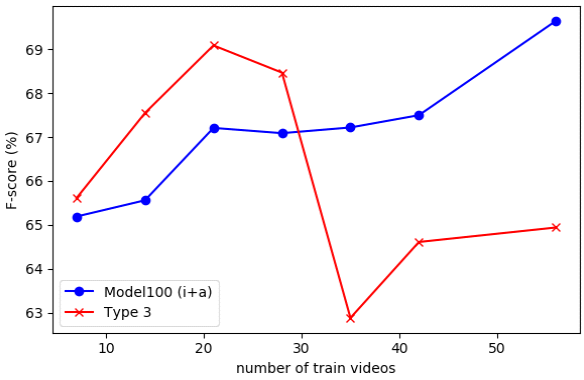


그림 6. 앙상블 Type 3에서의 e스포츠 학습 데이터 크기 별 결과 커브 (파란선: 단일모델 결과, 빨간선: 앙상블 모델 결과)
Fig. 6. Experiment results (F-score) as the number of e-sports train video is varied

V. 결 론

본 논문에서는 영상에서 불균일한 정보의 분포를 극복하기 위해 하이라이트 각 구간 내에서 특정 부분을 더욱 집중적으로 탐색하는 앙상블 모델을 제안하였다. 앙상블의 각 단일모델들은 하이라이트 이벤트의 ground truth 비율을 서로 다르게 조절하여 영상을 이해한다. 따라서 제안하는 앙상블 모델은 특정 구간에서 하이라이트 예측에 중요한 영향을 미치는 특징들을 많이 확보할 수 있다. 우리의 앙상블 모델은 e스포츠 데이터와 야구 경기 데이터를 통해 앙상블 모델이 성능 개선에 효과가 있음을 확인하였다. e스포츠와 야구 데이터 모두 하이라이트 이벤트의 초중반 부분을 집중적으로 파악하는 것이 하이라이트 검출에 있어 가장 좋은 영향을 줌을 알 수 있었다. 또한 앙상블 모델이 데이터 개수가 적은 경우에 성능을 향상시키는 것을 검증하였다.

제안한 앙상블 모델의 단일 모델들은 서로 다른 ground truth 구간을 파악하는데, 이때 동일한 가중치를 가지고 합쳐진다. 그러나 특정 구간에 더 높은 가중치를 부여하여 앙상블을 구성하면 더욱 향상된 성능을 보일 수 있을 것이라 기대한다. 또한 각 하이라이트 구간을 사전에 나누지 않고 attention 모델을 적용하여 가중치를 높여야하는 특정 부분을 자동으로 찾아내는 것도 차후 연구 방향이 될 수 있을 것이다.

참 고 문 헌 (References)

[1] K. Zhang, WL. Chao, F. Sha, and K. Grauman, “Video Summarization with Long Short-term Memory,“ European Conference on Computer Vision, Amsterdam, Netherlands, pp. 766-782, 2016, https://doi.org/10.1007/978-3-319-46478-7_47

[2] K. Zhou, Y. Qiao, and Tao Xiang, “Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward,” In Thirty-Second AAAI Conference on Artificial Intelligence, pp. 7582-7589, 2018.

[3] B. Zhao, X. Li, and X. Lu, “HSA-RNN: Hierarchical Structure-Adaptive RNN for Video Summarization,” The IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 7405-7414, 2018, https://doi.org/10.1109/cvpr.2018.00773

[4] B. Mahasseni, M. Lam, and S. Todorovic, “Unsupervised Video Summarization with Adversarial LSTM Networks,” In CVPR, pp. 2982-2991, 2017, https://doi.org/10.1109/cvpr.2017.318.

[5] 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.

[6] K. Zhang, K. Grauman, and F. Sha, “Retrospective Encoders for Video Summarization,” In ECCV, pp. 383-399, 2018, https://doi.org/10.1007/978-3-030-01237-3_24.

[7] H. Lee, G. Lee, “Summarizing Long-Length Videos with GAN-Enhanced Audio/Visual Features,” In ICCV workshop, 2019, https://doi.org/10.1109/iccvw.2019.00462

[8] H. Lee, G. Lee, “Video Highlight Prediction Using GAN and Multiple Time-Interval Information of Audio and Image,” Journal of Broadcast Engineering, Vol. 25, No. 2, pp. 143-150, 2020, https://doi.org/10.5909/JBE.2020.25.2.143

[9] E. Kim, G. Lee, “Highlight Detection in Personal Broadcasting by Analysing Chat Traffic : Game Contests as a Test Case," Journal of Broadcast Engineering, Vol. 23, No. 2, pp. 218-226, 2018, http://dx.doi.org/10.5909/JBE.2018.23.2.218.

[10] E. Kim, G. Lee, “Video Highlight Prediction Using Multiple Time-Interval Information of Chat and Audio,” Journal of Broadcast Engineering, Vol. 24, No. 4, pp. 553-563, 2019, https://doi.org/

10.5909/JBE.2019.24.4.1.
[11] Twitch, https://www.twitch.tv/ (accessed May. 20, 2020).
[12] Kakao TV, https://tv.kakao.com/ (accessed May. 20, 2020).
[13] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet Classification with Deep Convolutional Neural Networks,” In NIPS, 2012, https://doi.org/10.1145/3065386.
[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” In CVPR, pp. 770-778, 2016, https://doi.org/10.1109/cvpr.2016.90.
[15] Naver-sports, https://sports.news.naver.com/ (accepted May. 20, 2020).
[16] OGN, http://ogn.tving.com/ (accepted May. 20, 2020).

저 자 소 개

이 한 솔

- 2019년 : 서울과학기술대학교 전자IT미디어공학과 학사
- 2019년 ~ 현재 : 서울과학기술대학교 일반대학원 미디어IT공학과 석사과정
- ORCID : https://orcid.org/0000-0002-1127-976X
- 주관심분야 : 머신러닝, 딥러닝, 신호처리

이 계 민

- 2001년 : 서울대학교 전기공학부 학사
- 2007년 : University of Michigan EECS 석사
- 2011년 : University of Michigan EECS 박사
- 2011년 ~ 2012 년 : University of MichiganResearch Fellow
- 2013년 ~ 현재 : 서울과학기술대학교 전자 IT 미디어공학과 부교수
- ORCID : https://orcid.org/0000-0001-6785-8739
- 주관심분야 : 머신러닝, 신호처리, 의료정보학