어텐션 기반 비디오 하이라이트 예측 알고리즘의 개선

윤원빈, 황준규, 이계민
서울과학기술대학교 전자 IT 미디어공학과
{noizbeen, dr1nkc0ckt6il, gyemin} @seoultech.ac.kr

Improving Attention-based Video Highlight Prediction

Wonbin Yoon, Junkyu Hwang, Gyemin Lee
Department of Electronic and IT Media Engineering
Seoul National University of Science and Technology

요 약

하이라이트 영상은 원본 영상의 중요한 장면들을 짧은 시간 안에 감상할 수 있게 도와준다. 특히나 경기 시간 긴 축구나 야구 그리고 e-스포츠의 시청자들에게 있어, 하이라이트 영상의 효용성은 더욱 증가한다. 하이라이트 영상 추출의 자동화로 방송사나 온라인 플랫폼은 비용 절감과 시간 절약의 이점을 얻을 수 있다. 따라서 본 논문에서는 스포츠 영상에서 자동으로 하이라이트 구간을 추출하는 모델을 제안한다. 제안하는 모델은 멀티 헤드 어텐션 매커니즘과 LSTM 네트워크의 결합으로 구성된다. 해당 매커니즘의 여러 헤드를 통해 어텐션을 다양한 관점에서 진행한다. 이로 인해 영상의 전체적인 맥락과 장면 간의 유기적 관계를 다양한 관점에서 파악할 수 있다. 또한 오디오와 이미지 정보를 함께 이용하여 모델을 학습한다. 학습한 모델의 평가는 e-스포츠 경기 영상을 이용하여 평가한다.

1. 서론

각종 스포츠들 중 축구와 야구 그리고 e-스포츠의 경우 경기 시간이 길다. e-스포츠의 경우 평균 경기 길이가 30 분이고 야구의 경우 3 시간이 넘어간다. 이에 경기를 중계하는 방송사, 온라인 플랫폼은 시청자들의 편의를 돕고자 경기 종료 후 하이라이트 영상을 제공한다. 이러한 하이라이트 영상은 경기의 중요한 장면을 제공하는 것과 동시에 경기의 전반적인 맥락을 파악할 수 있게 한다. 하지만 30 분이 넘는 길이의 영상의 하이라이트를 제작하는 것은 시간과 비용이 적지 않게 소모되는 작업이다. 이에 착안하여, 본 논문을 통해서 긴 길이의 영상의 하이라이트 구간을 자동으로 추출하는 모델을 구현했다.

과거 BiLSTM 과 어텐션 매커니즘을 접목한 하이라이트 예측 모델에 관한 연구가 진행됐다[1]. 해당 매커니즘을 통해 영상의 전체적인 맥락을 파악하고 한 장면과의 관계를 파악할 수 있다. 또한 기존 BiLSTM 기반의 모델에 비해 높은 성능을 보여주었다. 어텐션이란 결국 중요한 부분을 더욱 강조해서 이용하기 위한 연산을 하는 것으로 이해할 수 있다. 하지만 스포츠 경기의 경우 다양한 요인이 경기의 흐름에 영향을 끼치기에, 이를 다양한 관점에서 바라볼 필요성이 있다. 따라서 어텐션 기반 모델의 성능을 높이기 위해 어텐션을 멀티 헤드 어텐션으로 대체해볼 수 있다.

본 논문은 [1]의 결과를 바탕으로 멀티 헤드 어텐션의 접목을 통해 모델의 성능을 향상시키기 위한 연구이다. 모델을 학습하고 평가하기 위한 데이터는 e-스포츠 경기 영상이다. 이미지 정보뿐만 아니라 오디오 정보도 함께 학습에 이용하여 영상의 다양한 특징을 고려해 하이라이트 구간을 추출한다.

2. 관련 연구

하이라이트 구간 추출과 관련된 연구는 꾸준히 진행되어 왔다. 대표적인 지도 학습 모델은 [2]의 LSTM 과 DPP (Determinan tal Point Process)를 결합한 모델이다. [3]은 비지도 학습 모델로, LSTM 을 이용하여 VAE (Variational Auto-Encoder)와 GAN (Generative Adversarial Networks)의 결합된 구조를 형성했다. 앞의 연구는 짧은 영상을 데이터셋으로 활용한 경우이다. [4]는 다중 시구간 데이터를 이용하는 LSTM 모델로, 야구 경기의 오디오와 채팅 데이터를 이용했다. [1]는 어텐션 매커니즘을 이용한 LSTM 기반의 모델이다. 영상의 한 장면이 전체 흐름에 끼치는 영향을 어텐션을 통해 파악했다.

한편 [5]는 트랜스포머와 멀티 헤드 어텐션 매커니즘을 소개한 연구이다. 트랜스포머는 RNN 기반 인코더-디코더 구조의 기존 기계번역 모델에서, RNN 을 사용하지 않은 모델이다. 대신 멀티 헤드 어텐션을 통한 병렬 연산으로 정확하고 신속한 연산을 수행했다.

3. 제안하는 알고리즘

해당 장에서는 모델의 구조를 두 경우로 나누어 설명한다. 먼저 오디오와 이미지 정보 중 하나만 하이라이트 구간 추출에 이용하는 경우 모델 구조를 설명한다. 다음으로 모델에 멀티 헤드 어텐션 매커니즘을 접목시켜, 영상의 한 장면과 전체 맥락의 유기적 관계를 다양한 관점에서 파악하는 방법을 설명한다. 그리고 두 정보를 모두 이용하는 모델을 설명한다. 이를 통해 다양한 특징으로 하이라이트를 예측한다.

3.1 멀티 헤드 어텐션 기반 예측 모델 (MHA)

영상의 한 장면이 전체적인 흐름에 끼친 영향이 크다면, 그 장면은 하이라이트에 포함될 가능성이 높아진다. 이를 파악하기 위해선 영상의 맥락과 장면 사이의 유기적 관계를 읽어낼 수 있어야 한다. 또한 스포츠 경기는 다양한 요인에 의해 흐름이 결정된다. 따라서 위의 유기적 관계를 다양한 관점에서 해석해 볼 필요성이 있다.

먼저 오디오와 이미지 중 하나의 정보만을 이용하는 모델을 그림 1 을 통해 설명한다. 오디오 혹은 이미지의 프레임 x_t 는 BiLSTM 네트워크의 입력으로 들어간다. 해당 네트워크를 통해 출력된 결과는 f_t이다. f_t는 이후 두 LSTM 네트워크의 입력으로 들어간다. 첫 번째 LSTM 네트워크로 f_t가 들어간다. 따라서 출력 결과인 z_t 는 영상의 단기적 맥락에서 전후 관계 해석 정보와 특징 을 가진다. 두 번째 LSTM 네트워크로는 f_t의 샘플링 된 데이터가 입력으로 들어간다. 샘플링은 일정한 간격(예, 1 분)으로 진행

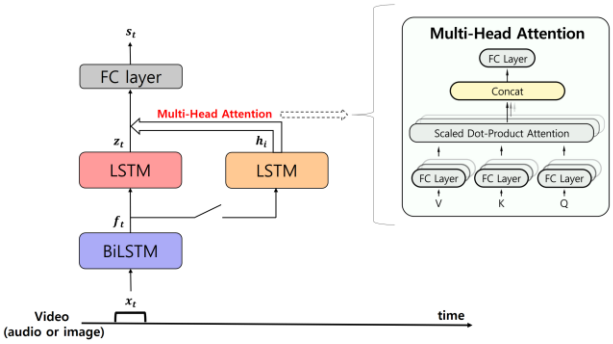

그림 1. 멀티 헤드 어텐션 기반의 하이라이트 예측 모델(MHA)
Fig 1. Our Multi-head attention-based video highlight prediction model

된다. 출력 결과인 h_i는 영상의 중장기적 맥락에서 전후 관계 해석 정보와 특징을 가진다. 따라서 h_i는 영상의 전체 맥락에 대한 정보를 가졌다고 이해할 수 있다.

z_t와 h_i를 통해서 멀티 헤드 어텐션을 진행한다. 멀티 헤드 어텐션 연산 과정은 그림 1 의 우측에 나타냈다. 여기서 어텐션 매커니즘의 query 는 z_t, key 와 value 는 h_i에 해당한다. 먼저 query 와 key 그리고 value 는 각각의 FC layer 에 들어가 연산된다. 이때 FC layer 의 출력 차원은 정해진 임베딩 차원에 헤드의 수를 곱한 값이다. 출력 값으로 나온 Q, K 그리고 V 를 이용하여 Scaled Dot-Product Attention 을 수행해 Attention value 를 구한다. 수식은 다음과 같다. d_k는 key 의 임베딩 차원이다.

$$\text{Attention}(Q,\ K,\ V) = \text{softmax}(\frac{QK^T}{\sqrt{d_K}})V$$

이후 attention value 를 모두 합친 다음 FC layer 의 입력으로 넣는다. FC layer 의 결과값은 z_t와 합쳐져 최종 FC layer 에 들어간다. 이의 결과값은 하이라이트 스코어 s_t이다. 하이라이트 스코어가 높은 일부의 프레임을 적절한 비율로 선택하여 최종 하이라이트를 형성한다. 모델은 아래의 손실 함수를 최적화한다.

$$L_{CE} = \frac{1}{T}\sum_t cross - entropy(s_t, y_t)$$

여기서 y_t는 ground truth 의 label, T 는 동영상의 총 길이이다.

3.2 멀티 헤드 어텐션 기반 멀티모달 예측 모델 (M-MHA)

동영상의 오디오와 이미지는 각기 다른 정보를 가지고 있다. 스포츠 경기의 중요한 상황에서 시청자들은 격앙된 해설과 관중들의 환호성을 들을 수 있다. 따라서 우리는 하이라이트 구간을 추출하기 위해 두 정보를 모두 활용하는 모델을 제안한다.

그림 2 는 두 정보를 모두 이용하기 위한 모델이다. 오디오와 이미지의 프레임 x_t^{audio}, x_t^{image} 는 각각의 BiLSTM 네트워크의 입력으로 들어간다. 출력된 결과는 각각 f_t^{audio}, f_t^{image} 이다. 이후 두 결과가 합쳐져 첫 번째 LSTM 으로 들어가 z_t 출력한다.

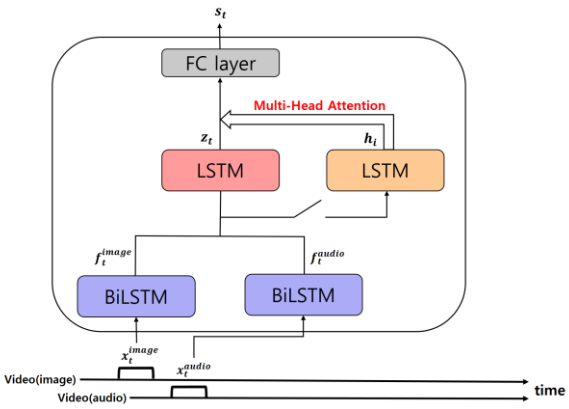

그림 2. 오디오/이미지 정보를 동시에 이용하는 멀티 헤드 어텐션 기반 하이라이트 예측 모델

Fig 2. Our multi-head attention-based highlight prediction model that uses both audio and image information

f_t^{audio}, f_t^{image}는 각각 샘플링 된 후 합쳐져 두 번째 LSTM 으로 들어가 h_i를 형성한다. 이후 위의 멀티 헤드 어텐션 연산 과정을 거치고 FC layer 통과 후 하이라이트 스코어 s_t를 추출한다.

4. 실험 및 결과

구현한 모델을 평가하기 위한 데이터는 Twitch[6]에서 2017 년도에 중계한 'League of Legends' 대회의 경기 영상이다. 총 5 개의 대회(IBM World Championship Katowice 2017, 2017 LoL World Championship, LoL All Star 2017, 2017 LoL Champions Korea Spring, 2017 LoL Champions Korea Summer)에서 진행된 63 개의 경기 영상으로 데이터를 구성했다. 이 중 56 경기를 학습 데이터로, 7 개의 경기를 테스트를 위한 데이터로 나누어 이용했다. 하이라이트 영상의 ground truth 은 OGN[7]에서 제공한 영상이다. 표 1 에 데이터의 요약 정보가 경기 길이의 평균과 하이라이트 길이의 평균의 비가 10 대 1 정도이다. 따라서 하이라이트 스코어 s_t 값이 상위 10%에 해당하는 프레임을 하이라이트로 추출했다.

수집한 데이터들은 전처리 과정을 통해 특징 벡터를 미리 추출한 후 이용했다. 이미지 정보의 특징 벡터 추출에는 ImageNet[8] 데이터셋으로 사전 학습된 ResNet-34[9] 모델을 활용했다. 이를 통해 1fps 당 512 차원의 특징을 갖는 벡터 x_t^{image}를 추출하였다. 오디오 정보의 경우 Mel-Frequency Cepstral Coefficients (MFCC)를 이용, 40ms 당 20 차원의 특징 벡터를 추출하였다. 이후 이를 25 개씩 묶어 1 초당 500 차원의 특징을 갖는 벡터 x_t^{audio}를 형성하였다.

모델은 Pytorch 를 통해 구현했다. 모델을 구성하는 LSTM 네트워크는 256 개의 hidden unit 을 가진다. 멀티 헤드 어텐션의 헤드 개수는 8, key 와 value 의 임베딩 차원은 64 이다.

표 1. e-스포츠 데이터 요약 정보

Table 1. Summary of e-Sports data set

Statistics	Video Length (sec)	Length of Highlights (sec)	Highlight ratio (%)
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

표 2. e-스포츠 데이터에 대한 실험 결과 (F-score)

Table 2. Experiment results on e-Sports data set

Data type	Model	F-score (%)
Image	$ATTN_i[1]$	70.53
	MHA_i	66.93
Audio	$ATTN_a[1]$	69.23
	MHA_a	69.80
Image + Audio	$ATTN_{i+a}[1]$	73.93
	MHA_{i+a}	73.70

학습한 모델의 평가는 F-score 를 통해 이루어졌다. F-score 는 다음 식으로 표현된다.

$$P = \frac{|H_{gt} \cap H_{pred}|}{|H_{pred}|}, R = \frac{|H_{gt} \cap H_{pred}|}{|H_{gt}|}$$

$$F\text{-}score = \frac{2PR}{P+R} \text{ X } 100\%$$

여기서 H_{gt}는 ground truth, H_{pred}는 모델에 의해 추론된 하이라이트를 나타낸다.

표 2 에 F-score 를 계산한 정량적 결과를 나타냈다. 표에서 ATTN 은 단일 어텐션 모델을, MHA 는 멀티 헤드 어텐션 모델을 의미한다. 이미지 정보를 학습에 활용한 어텐션 모델의 스코어는 70.53%, MHA 는 66.93%이다. 3.6% 가량 스코어가 감소하였다. 오디오 정보의 경우 어텐션 모델은 69.23%, MHA 는 69.80%로 0.57% 가량 스코어가 상승하였다. 두 정보를 모두 활용하였을 때 어텐션 모델은 73.93%, MHA 는 73.70%로 0.2% 가량 감소하였다. 이는 오디오 정보를 활용했을 때의 성능 향상과 이미지 정보를 활용했을 때의 성능 하락이 맞물린 결과로 보인다.

현재 모델은 시간축에 따라 정보를 구성하고 멀티 헤드 어텐션을 진행한다. 실험 결과 오디오의 경우 시간 축에 대해 맥락을 파악하는 것이 유용한 것으로 보인다. 반면 이미지 데이터의 경우 시간 축 상에서만 맥락을 파악하는데 한계가 있는 것으로 판단된

다. 따라서 이미지 정보의 경우 시간에 따른 정보가 아닌 공간적 정보에 대한 멀티 헤드 어텐션을 적용해볼 수 있고 이에 대한 방법을 모색해볼 수 있을 것이다.

5. 결론

기존에 연구되었던 어텐션 매커니즘 기반의 하이라이트 예측 모델을 멀티 헤드 어텐션으로 대체하여 성능 향상을 도모하였다. 이를 통해 더욱 다양한 관점의 어텐션을 진행하고 오디오 데이터를 이용하는 모델의 성능을 향상시켰다. 이는 정량적 결과를 통해 확인했다. 하지만 이미지 데이터를 이용하는 모델의 성능은 오히려 떨어졌다. 이미지와 오디오 두 정보를 동시에 이용하는 멀티모달 모델의 경우 성능에 크게 변화가 없었다. 향후 이미지의 경우 공간적 정보에 대한 어텐션을 이용해 볼 필요가 있어 보인다.

참고문헌

[1] 이한솔, 이계민, “Attention 기반의 문맥을 이용한 하이라이트 예측 알고리즘,” 대한전자공학회 학술대회, pp. 886-890, 2020

[2] K. Zhang, WL. Chao, F. Sha and K. Grauman, “Video Summarization with Long Short-term Memory,” In European Conference on Computer Vision (ECCV), Amsterdam, Netherlands, pp.766-782, 2016.

[3] B. Mahasseni, M. Lam and S. Todorovic, “Unsupervised Video Summarization with Adversarial LSTM Networks,” In Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2982-2991, 2017.

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

[5] A. Vaswani, N. Shazeer, N. parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., “Attention is all you need,” In Conference on Neural Information Processing Systems (NIPS), 2017.

[6] Twitch, https://www.twitch.tv/ (accessed July.04, 2020).

[7] OGN, https://www.ogn.tving.com/ (accessed July.04, 2020)

[8] A. Krizhevsky, I. Sutskever and G. Hinton, “Imagenet Classification with Deep Convolutional Neural Networks,” In Conference on Neural Information Processing Systems (NIPS), 2012.

[9] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” In Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.