RNN 기반 360°영상에서의 Saliency map 예측

이수정, 이주형, 이준영, 임가은, 김동호*

인천과학예술영재학교, *서울과학기술대학교

soojeong4513@gmail.com, jhlee18021@naver.com, goodjy2005@gmail.com, lge051121@gmail.com,
*dongho.kim@seoultech.ac.kr

Saliency map prediction in 360° video based on RNN

Lee Soojung, Lee Joohyung, Lee Joonyoung, Lim Gaeun, Kim Dongho*

Incheon Academy of Science and Arts, *Seoul National University of Science and Technology

요 약

본 연구는 convolution 신경망과 EMA를 적용한 RNN 신경망을 통해 global saliency와 local saliency를 계산하고, 이를 바탕으로 saliency map을 예측하는 모델에 관한 것이다. 모델은 global saliency를 구하기 위한 Attention stream과 local saliency를 구하기 위한 Expert stream으로 나뉜다. 이때 Expert stream은 영상을 방향에 따라 6개로 나누고 각각에 대해 EMA를 적용한 RNN 신경망을 통해 각 시점 별 local saliency map을 구한다. saliency360!을 학습 데이터셋으로 60회 학습한 결과 5개의 테스트 영상에 대해 평균 성능은 precision 0.678, recall 0.696, f-score 0.686를 얻었다.

Ⅰ. 서 론

최근 무선 네트워크 기술의 발전과 스마트폰 기반 콘텐츠 시장의 성장으로 사용자 만족을 위해 촉감을 비롯한 오감을 통해 정보를 제공하는 실감 미디어 기술이 주목받고 있다. 그 중 VR 기술은 현재 세계적으로 가장 넓은 시장을 보유하고 있다. 이에 따라 VR 콘텐츠 제작에 필요한 360 영상의 수요도 증가하고 있다. 360 영상은 모든 방향에서의 정보를 포함함으로써 한 쪽 방향의 정보만을 제공하는 기존의 영상보다 더 다양하고 완전한 정보를 사용자에게 제공한다. 그러나 시장화를 위해서는 대용량의 정보를 전송하는데 드는 비용으로 인한 낮은 경제성의 개선이 요구된다. 이러한 전송 비용 문제를 해결함과 동시에 사용자의 QoE (Quality of Experience)를 충족시킬 수 있는 해법은 tile-based viewport adaptive streaming이다. 이는 영상을 특정 구역으로 나누어 독립적으로 인코딩 해 사용자가 관심을 가질 것으로 예측되는 viewport에 해당하는 부분만 고화질의 영상을 전송하고 그 외 구역에 대해서는 저화질의 영상을 전송함으로써 효율적인 전송이 가능하다. 이를 위해서는 사용자가 관심을 가질 것으로 예측되는 viewport, 즉 시점을 예측하는 모델이 필요하다.

본 논문에서는 EMA 방식을 적용한 RNN 모델을 활용하여 360 영상의 saliency map을 예측하는 모델을 제시한다. 이때 saliency map을 local saliency와 global saliency로 나누어 예측하고자 한다.이때 local saliency의 계산에 이전 시점에서의 값을 포함하는 RNN 구조를 사용해 시계열 데이터인 동영상에 적절한 모델로 설계하고자 한다.

2. 선행 연구

1) Analyzing viewport prediction under different VR interactions (Tan Xu 외 2인, 2019) [1]

해당 연구에서는 다양한 예측 모델을 사용하여 VR 기기, PC, 모바일 3개의 서로 다른 플랫폼 환경에서의 사용자의 viewport를 예측하였다. 연구 결과, 심층학습을 통한 사용자의 viewport 예측 연구가 기존의 기계학습을 통한 연구보다 효과적임을 보였으며 심층학습을 통한 viewport 예측 모델 개발의 필요성을 실험적으로 입증했다.

2) Fixation Prediction for 360° Video Streaming in Head-Mounted Virtual Reality (Ching-Ling Fan 외 5인, 2017) [2]

해당 연구에서는 머리 장착형 VR 기기에서의 360 영상에 대한 타일 기반 전송 시스템을 제안하였다. 해당 연구의 시스템은 대량의 연산 자원을 소모하기 때문에 실시간 환경에서의 적용이 어렵다는 단점을 가진다. 또, 다양한 타일의 환경에서 적용할 수 없는 제한점을 가진다.

Ⅱ. 본론

1. Saliency map 예측 모델

1.1. 모델의 구조 (model architecture)

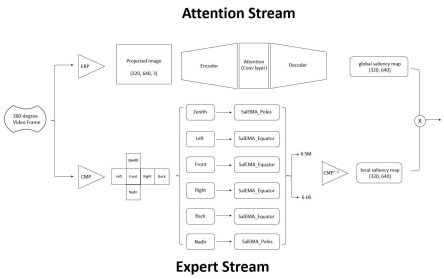

Fig 2. model architecture

영상 전체를 분석해 전반적인 관심도를 정량화한 global saliency와 세부적인 영역 별 관심도를 정량화한 local saliency로 나누어 2개의 saliency map을 산출한다. 2개의 saliency map을 합쳐 최종적인 saliency map을 만든다. 이때 global saliency를 예측하는 과정을 attention stream, local saliency를 예측하는 과정을 expert stream이라고 한다. 각 stream으로부터 각 pixel의 saliency 값을 0~1 사이의 값으로 예측한 결과를 얻는다. 최종 saliency map은 각 pixel에 대해 attention stream의 saliency 결괏값과 expert stream의 saliency 결괏값을 곱해 구한다.

1.2. Attention stream - global saliency map 예측

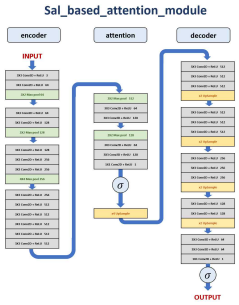

Fig 3. Attention stream architecture

Attention stream에서는 ERP(Equirectangular projection)을 사용한다. ERP를 통해 구면 좌표계로 입력된 영상을 직교 좌표계로 변환할 수 있다. Encoder를 거쳐 전처리된 벡터를 x라 하자. x가 attention module을 거치면 입력 이미지의 fixation map이 생성된다. 이때 fixation map의 결과를 전처리된 입력 데이터 x에 반영해 global saliency map을 구할 수 있다.

1.3. Expert stream - local saliency map 예측

CMP 변환된 데이터를 encoder를 통해 전처리하고, 이 데이터를 convolution layer를 통과시키면 시점과 관련된 특성 맵을 얻을 수 있다. 이 특성 맵을 decoder를 통과시키면 saliency map을 얻을 수 있다. 이때 특정 시점의 saliency map을 구하는데 이전 시점까지의 결과를 반영하기 위해 EMA(Exponential Moving Average)를 적용한 RNN 모델을 사용한다.

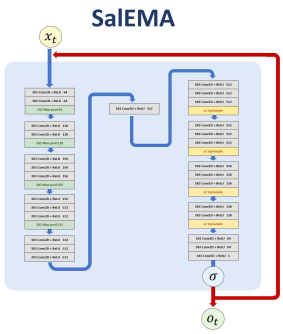

Fig 4. Expert stream architecture

1.4. 최종 Saliency map 반환

Local saliency map과 global saliency map을 합쳐 saliency map을 만든다. 입력된 비디오에 대해 각 frame별 saliency map을 구한다. ERP로 변환한 이미지를 attention stream의 모델의 input으로 넣어 얻은 saliency map를 first stream이라고 한다. 이때 first stream은 global saliency map에 해당한다. ERP 변환된 이미지를 CMP 변환을 통해 6개의 head state를 key로 하고 각 head state에 대한 이미지를 value로 하는 딕셔너리를 얻는다. 6개의 head state 각각에 대한 이미지 데이터를 SalEMA 모듈의 입력값으로 넣는다. Head state와 그 방향에 대한 local saliency를 담은 딕셔너리를 기반으로 역 CMP 변환을 진행해 local saliency map을 얻는다. 이를 second stream이라고 하자.

즉, (320, 640) 크기의 global saliency map과 local saliency map을 얻으며 각 셀의 값은 0부터 1까지의 값을 가진다. 각 셀의 saliency 값은 셀의 first stream 값과 second stream 값을 곱해 구한다.

2. 모델 학습

2.1. 학습 데이터셋 설명

학습 데이터셋은 salient360! 데이터셋을 사용한다. salient360! 데이터셋은 360°VR 영상과 그 영상에서 57명의 viewport의 움직임에 대한 정보(scanpath)를 제공한다.

2.2. Sal_based_attention_module의 학습

Sal_based_attention_module의 학습에서는 optimizer는 adam을 사용하며 loss 함수는 모듈에서 정의한 함수를 사용한다. 이때 loss를 다음과 같이 정의한다.

$$attention = 0.1 * (1 - NSS(s))$$

$$loss = 0.8 * KLD(s, gt) + 0.1 * KLD(f, gt) + attention$$

$$(s : \text{saliency map},\ f : \text{fixation map},\ gt : \text{ground truth})$$

2.3. SalEMA의 학습

SalEMA 모델의 학습에서도 optimizer 함수로 Adam을 사용하였다. Loss 함수는 BCE Loss(Binacy Cross Entropy Loss) 함수를 사용하였다.

3. 연구 결과

학습 데이터셋, 검증 데이터셋, 테스트 데이터셋 각각에 대한 loss와 accuarcy는 다음과 같다.

학습 데이터셋 : loss - 0.0187, accuracy - 84.28%, 검증 데이터 셋 : loss - 0.0193, accuracy - 84.25%, 테스트 데이터 셋 : loss - 0.0213, accuracy - 82.63%

각각의 테스트 데이터셋에 대한 모델의 성능 지표를 표 2에 나타냈다.

표 2. 테스트 데이터셋의 성능 지표

영상 제목	precision	recall	f-score
Cockpit	0.67	0.70	0.68
Turtle	0.79	0.80	0.80
UnderwaterPark	0.63	0.64	0.63
Bar	0.67	0.70	0.69
Touvet	0.63	0.64	0.63

Ⅲ. 결론

본 논문에서는 360 영상 전송 방식의 효율을 높이기 위해 RNN과 saliency map 기반의 사용자의 시점 예측에 대해 연구했다. 이에 따라 제작한 모델은 train loss 0.0187, train accuracy 84.28%, val loss 0.0193, val accuracy 84.25%의 성능을 보였다. 5개의 테스트 영상에 대해 평균 성능은 precision 0.678, recall 0.696, f-score 0.686이다.

해당 결과를 통해 정확도가 높은 saliency map을 예측하는 모델을 제작할 수 있었다. RNN을 사용했을 때 높은 정확도를 얻었다는 점에서 LSTM을 사용한 360° 영상의 saliency map 예측 모델을 통해 본 모델보다 정확도가 높은 모델을 만들 수 있을 것으로 기대한다. 추가적으로, 본 연구에서는 모델 설계에 RNN만을 활용하였으나 성능 향상을 위해 하이퍼 파라미터 최적화를 역강화학습 등 다양한 학습 방법을 도입할 수 있다. 향상된 예측 모델을 통해 비트레이트의 분배가 더 효율적으로 나타나면 360 VR 영상을 보는 사용자의 만족도를 더욱 향상시킬 수 있을 것이다.

참 고 문 헌

[1] Tan Xu, Bo Han, and Feng Qian. 2019. Analyzing viewport prediction under different VR interactions. In Proceedings of the 15th International Conference on Emerging Networking Experiments And Technologies (CoNEXT '19). Association for Computing Machinery, New York, NY, USA, 165 - 171.

[2] Ching-Ling Fan, Jean Lee, Wen-Chih Lo, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2017. Fixation Prediction for 360° Video Streaming in Head-Mounted Virtual Reality. In Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'17). Association for Computing Machinery, New York, NY, USA, 67 - 72.