# 오토인코더 모델을 이용한 리그 오브 레전드 게임 동영상 하이라이트 추출

윤성훈O 이승진 김경중*
세종대학교 컴퓨터공학과
kiboyz2@naver.com, cilabjin1819@gmail.com, kimkj@sejong.ac.kr

# League of Legends Game Video Highlight Extraction using Autoencoder

Seonghun YoonO SeungJin Lee Kyung-Joong Kim*
Department of Computer Science and Engineering, Sejong University

**요 약**

최근, 길이가 긴 동영상을 시청하는 대신 영화나 드라마의 예고편을 보거나 스포츠 경기의 하이라이트를 보여주는 동영상이 인기를 끌고 있다. 하지만, 긴 동영상을 축약하기 위해 수작업에 의존하고 있으며, 이는 많은 시간과 노력을 필요로 한다. 이를 해결하기 위해, 본 논문에서는 딥 러닝을 이용하여 동영상에서 하이라이트를 자동으로 추출하기 위한 오토인코더 기반의 방법을 제안하며, 리그 오브 레전드 월드 챔피언쉽 게임 대회 영상으로 실험을 수행하였다. 기존 딥 러닝을 이용한 하이라이트 추출 방법을 제안하는 논문들은 이미지 데이터, 음성 데이터 또는 텍스트 데이터만을 이용하였지만[1-3], 본 연구에서 제안하는 모델에서는 동영상에서 추출해낸 이미지 데이터와 음성 데이터를 입력 데이터로 적용한 것으로, 동영상의 여러 데이터를 입력으로 사용한 동영상 하이라이트 추출을 제안하는 논문이다. 실험 결과 제안하는 모델의 성능은 기존의 동영상 데이터를 통해 정확도를 측정했을 때, 72.44%의 성능을 보여주었다.

## 1. 서 론

최근 동영상을 동영상의 업로드 수는 시청하는 사람들의 수만큼이나 증가하고 있다. 동영상 데이터 시청은 시간에 비례하는 시간이 소요된다. 이로 인해, 긴 동영상의 경우 영상을 축약한 형태의 편집된 영상이 점점 많아지고 있다. 편집된 영상은 하이라이트 영상이라고 불리며, 하이라이트 영상을 편집하기 위해 많은 시간과 노력이 필요로 한다. 편집에 대한 시간과 노력을 경감시키기 위해, 하이라이트 추출 연구는 예로부터 많이 진행되어 왔다. 그러나, 이전의 하이라이트 추출 연구들은 데이터에 의존적인 도메인 지식을 사용하는 경우가 많았으며 여러 종류의 데이터를 사용하지 않았다. 본 연구에서는 이러한 도메인 지식을 이용하지 않고 이미지와 음성 데이터를 통해 딥 러닝을 이용한 하이라이트 추출 모델을 제안한다.

딥 러닝을 이용한 하이라이트 추출 연구는 이전부터 다양하게 시도되어 왔었다[1-3]. 그러나 기존의 논문들은 이미지 데이터, 음성 데이터 또는 텍스트 데이터만을 통해 추출하려고 시도하였다. 한 종류의 데이터를 사용하는 것은 모든 데이터를 사용하지 않아, 하이라이트 영상을 추출해내지 못하는 결과를 보여줄 수 있다.

그림 1을 보면 다른 일반적인 장면과는 확연히 대비되는 음성 데이터가 나오는 것을 볼 수 있다. 이는 하이라이트 영상에서 동영상의 모든 데이터들이 하이라이트를 추출하기 위해 필요하다는 것을 말한다.

본 연구에서는 리그 오브 레전드 게임 동영상을 수집하여 연구를 진행했다. 이 게임 동영상은 일반적인 장면과는 대비되는 하이라이트 장면의 확연히 다른 화면 이미지와 게임을 중계하는 해설자들의 목소리 톤에서의 차이가 있어 하이라이트 영상의 공통점이 많은 장점이 있다. 그 외에, 원본 영상과 하이라이트 영상의 데이터 수가 다른 데이터보다 많아 학습을 위한 데이터로 선정하였다.

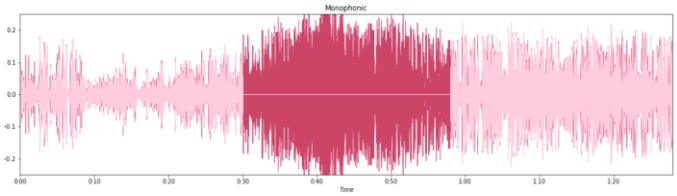



그림 1 하이라이트(진한 색)과 일반 장면 음성 데이터 차이

## 2. 관련 연구

### 2.1 하이라이트 추출

하이라이트를 추출하는 방법에는 여러 종류의 데이터를 사용하는 다양한 연구들이 있다.

H. Yang et al.은 액션 캠으로 찍은 스포츠 영상들의 이미지만을 사용하여 하이라이트를 추출하는 모델을 제안하였다[1].

H. J. Woo et al.은 음성 데이터만을 사용하여 여러 장르의 노래 하이라이트를 추출하는 모델을 제시하였다[2].

H. G. Nam (2018)의 경우 스트리밍 서비스의 유저 텍스트를

기반으로 하이라이트를 추출하는 모델을 제시하였다[3].

다양한 종류의 데이터를 사용하여 하이라이트를 추출하는 연구들이 있지만, 여러 종류의 데이터를 통해 하이라이트를 추출하는 연구는 활성화가 안되어 있다. 본 연구에서는 이미지와 음성 데이터를 사용한 모델을 제시한다.

## 3. 오토인코더 기반 게임동영상 하이라이트

제안한 모델의 학습할 데이터는 전 세계적으로 유명한 게임 리그 오브 레전드의 월드 챔피언십 대회의 하이라이트 영상이다. 데이터 수집을 위해 유튜브에서 동영상을 다운로드하기 위해 youtube-dl을 사용하여 리그 오브 레전드 월드 챔피언십 2017년도 8강, 4강, 결승의 학습을 위한 하이라이트 영상과, 원본 영상들을 수집하였다.

### 3.1. 데이터 전처리

동영상 데이터의 크기는 학습에 일반적으로 사용하는 이미지보다 크고 연속적이다. 따라서, 동영상 데이터 전체를 사용하기 보다는 전처리를 통해 변형한 것을 학습에 사용하도록 한다.

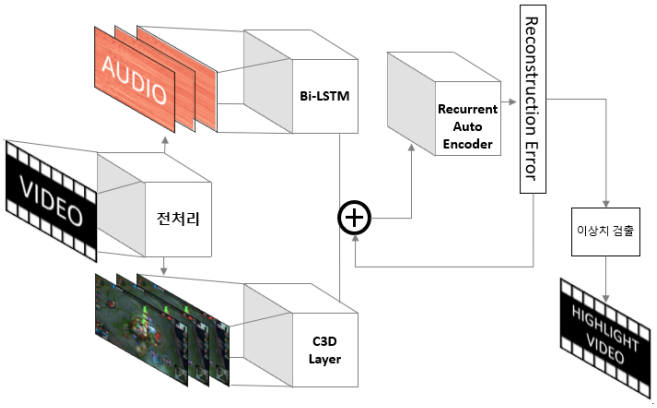
본 연구에서는 이미지의 크기는 동영상의 크기에 관계없이 크기 조절을 통하여 64x64 RGB이미지로 구성하였으며, 음성 데이터는 스펙트럼 형식으로 표현된 2차원 데이터로 구성하였다. 이미지와 음성 데이터의 경우 0.1초마다 추출하도록 하였다. 추출된 이미지와 음성 데이터는 학습 시 저장된 파일에서 데이터를 추출하여 처리한다. 일반적으로 동영상 데이터는 순차적인 데이터이므로 시간적으로 배열된 Sequence 구조를 사용하여야 한다. 이를 위해, 본 연구에서는 0.9초동안 추출된 정보를 하나로 묶어 하나의 입력 값으로 사용한다.


그림 2 데이터 확장 방식

### 3.2 데이터 확장

데이터를 학습하는데 있어 데이터의 양을 늘리는 것은 학습 성능을 높여주고, 학습 모델의 신경망을 좀 더 강건하게 해주는 효과가 있다. 이러한 방법을 데이터 확장(Data agumentation)이라고 부르며, 이를 통해 동일 영상에 대한 데이터를 여러 개로 늘림으로써 좀 더 학습이 잘 되도록 할 수 있다. 본 연구에서는 좌우반전(Fliplr)과 확대(Scale), 부분 지우기(CoarseSalt)의 세 가지 데이터 확장을 사용하였다. 본 연구에서 사용한 좌우반전은 1.0의 확률 즉, 항상 좌우반전이 이루어지도록 하였으며, 크기 조절 방식에 있어서는 가로, 세로 모두 최소 1.1배에서 최대 1.7배로 확장하도록 하였다. 발생 확률은 0.1이다.

데이터 확장으로 모은 데이터들은 병합하여 저장하기 전, 모든 데이터들은 학습 시에 처리되어야 할 Sequence 구조를 유지하기 위해 0.9초 동안의 데이터들을 묶어 하나의 Record로 만들어 저장한다.




| Video Data<br>게임 비디오 데이터 | Input<br>전처리 | C3D & Bi-LSTM<br>오디오 & 이미지에서 하이라이트 피쳐 추출 | Recurrent AutoEncoder<br>하이라이트 피쳐를 복원하는 방향으로 학습 | Output |
|---|---|---|---|---|
| 학습 단계<br>하이라이트 비디오 | 오디오 & 이미지 추출 및 정규화<br>오디오 Filter-Bank TF Record | Seqeunce Length : 9<br>활성화 함수 : Relu | 입력 : 오디오 & 이미지 피쳐<br>출력 : 오디오 & 이미지 피쳐 | Reconstruction Error로 학습 |
| 테스트 단계<br>풀 경기 영상 비디오 | 오디오 & 이미지 추출 및 정규화<br>오디오 Filter-Bank TF Record | Seqeunce Length : 9<br>활성화 함수 : Relu | 입력 : 오디오 & 이미지 피쳐<br>출력 : 오디오 & 이미지 피쳐 | 풀 영상 하이라이트 비디오 |
| Shape | Audio : [Data, Filters]<br>Image : [Width, Height, RGB] | [Batch, Sequeunce, Data] | [Batch, Sequeunce, Data] | [Batch, Reconstruction Error] |

그림 3 하이라이트 추출 모델 구조

### 2.3 하이라이트 추출 모델 구조

하이라이트 추출을 위해 기존에 신경망에서 사용하는 이미지 처리 방식에 쓰이는 Convolutional Neural Network(CNN)를 사용한다. 단, Sequence 데이터를 효율적이고 빠르게 처리하기 위해 동영상 데이터를 처리하는 3D Convolution layer(C3D)를 사용한다[4, 5]. 하이라이트 학습 데이터들만의 특징들을 표현할 C3D 계층을 거쳐 나온 특징들을 통해 오토인코더의 학습을 위한 입력 값으로 사용한다. 오토인코더는 C3D 계층에서 나온 하이라이트 특징들을 입력 값으로 가지며, 오토인코더에서 출력된 모델 예측 값과 실제 입력 값 사이의 유클리디안 거리(Euclidean Distance)를 Reconstructed error로 정의한다. 오토인코더는 Sequence 데이터와 같이 시간의 흐름에 따라 재귀적으로 반복적으로 학습하는 Recurrent AutoEncoder를 사용하였다.

음성 데이터는 이산적인 데이터를 추출하기 위해 사용하는 Sample rate는 피아노 옥타브의 2배인 8374, power-spectrum을 구하기 위한 FFT는 2048, Filter-Bank를 사용하여 총 128개의 특징을 0.1초마다 추출하고, 이를 Bi-directional LSTM 모델에 넣어 나온 값을 이미지와 병합하여 Recurrent Autoencoder의 입력 값으로 사용한다.

## 3. 실험 방식

동영상 데이터의 전처리를 위해 텐서플로우(Tensorflow)에서 제공하는 데이터 저장 형식인 TFRecord 형식을 사용하여 동영상에서 이미지 데이터와 음성 데이터 두 가지를 추출해낸다. TFRecord 형식은 규모가 큰 데이터 처리를 효율적으로 처리해준다.

### 3.1 평가 방법

하이라이트 추출 모델의 성능을 평가하기 위해, 원본 영상에서 하이라이트 영상과 유사한 부분을 체크하는 레이블링(Labeling) 작업을 하였다. 이를 통해 레이블링 된 원본 영상과 제안하는 모델의 예측 값을 비교하여 레이블링 된 하이라이트 영상과 예측 값의 차이를 통해 정확도를 측정한다. 레이블링 작업은 수작업으로 입력하였다.

### 3.2 실험 평가

실험은 학습 데이터 중 결승전 2경기 영상의 하이라이트 영상을 epoch를 100, learning rate는 0.05로 학습한 뒤 원본 영상으로 테스트한 것이다. 그림 4를 보면 하이라이트 추출 모델 예측 값을 확인할 수 있다. X축은 시간에 따른 프레임이며, Y축은 빨간 점은 하이라이트이며 파란 점은 하이라이트가 아닌 장면이다. 왼쪽 그래프가 레이블링된 기존 영상이며, 오른쪽이 모델에서 나온 예측 값이다. 성능은 레이블링된 기존 영상과 비교하여 72.44%의 성능을 보여주었다. 제안된 모델의 성능을 비교하기 위해, 이미지만 사용한 모델과 제안된 모델의 성능 차이를 비교하였다. 표 1에서 이미지만 사용한 모델보다 제안된 모델의 성능이 뛰어난 것을 확인할 수 있다. 즉, 음성 데이터를 통해 유의미한 하이라이트 특징을 추출했다는 것을 의미한다. 하지만, 음성 데이터만으로 하이라이트를 추출하였을 때는 유의미한 결과를 얻지 못하였다.

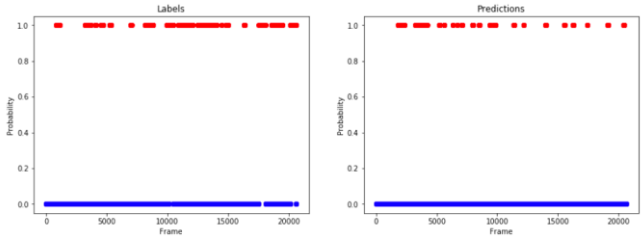



그림 4 하이라이트 추출 모델 예측 값

표 1 이미지+음성 모델과 이미지만 사용한 모델의 성능

| 모델 입력 데이터 구조 | 정확도 |
|---|---|
| 이미지 | 46.42% |
| 이미지+음성 | 72.44% |

### 3.3 하이라이트 영상 추출

하이라이트 추출을 위해 모델에서 나온 값을 학습에서 사용되었던 하이라이트 영상이 아닌 이상치(하이라이트 영상이 아닌 부분)을 검출한다. 검출 조건은 다음과 같다:

$$Reconstrunction\ Error = Euclidean\ Distance(x - \phi(x))$$

$$Outlier = \frac{max(RE) - min(RE)}{2} + \min(RE)$$

$$Highlight = RE < Outlier$$

이상치를 검출해 내면, 원본 동영상으로부터 이상치가 아닌 값들로 이루어진 부분적인 영상들을 병합한 영상이 모델에서 추출된 최종 하이라이트 영상이 된다.

## 4. 결론 및 향후 연구

현재 연구의 실험 대상은 리그 오브 레전드(League of Legends) 게임 영상으로 진행하였다. 기존의 연구 방식인 한가지의 데이터 종류만을 사용한 모델이 아닌, 이미지와 음성 데이터를 입력 값으로 사용하였다. 오토인코더를 사용한 제안된 모델의 성능은 레이블링된 기존의 영상과 비교하여 72.44%의 성능을 보여주었으며, 이미지만을 사용한 모델의 성능보다 월등한 성능을 보여주었다.

하이라이트 영상을 통해 추출해 낼 수 있는 데이터의 종류는 다양하다. 스트리밍 서비스를 통해 얻어낸 동영상의 경우 텍스트 데이터도 넣을 수 있을 것이라 생각한다. 추후, 다양한 종류의 데이터를 통해 하이라이트 추출을 시도할 예정이며, 이를 위해 더 많은 종류의 데이터와 좋은 성능을 위한 모델 구조 개선이 필요할 것으로 보인다.

## 5. 감사의 글

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초 연구 사업임 (2017R1A2B4002164). * 교신 저자

## 참 고 문 헌

[1] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, B. Guo, “Unsupervised Extraction of Video Highlights Via Robust Recurrent,” *IEEE International Conference on Computer Vision (ICCV)*, 2015

[2] H. J. Woo, A. Kim, C. Kim, J. Park and S. Kim, “Automatic Music Highlight Extraction using Convolutional Recurrent Attention Networks,” *CoRR abs/1712.05901*, 2107

[3] H. G. Nam, “Automatic Generation of Titled Video Highlights From Mass Interaction,” *Masters dissertation. Chungnam National University, Daejeon, Korea.* 2018

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks,” *ICCV,* 2015

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional Architecture for Fast Feature Embedding,” *arXiv,* 2014.