영상의 배경 정보를 활용한 영화 지문 생성 방법

곽창욱, 손정우, 이호재, 김선중
미디어지능화연구실, 미디어연구본부, 통신미디어연구소,
한국전자통신연구원
cukwak@etri.re.kr, jwson@etri.re.kr, lhjalex@etri.re.kr, kimsj@etri.re.kr

Movie captioning using background information of video

Kwak Chang-Uk, Son Jeong-Woo, Lee Alex, Kim Sun-Joong
Media Intellectualization Research Section, Media Research Division,
Telecommunications & Media Research Laboratory,
Electronics and Telecommunications Research Institute

요 약

영상 컨텐츠를 활용한 다양한 서비스와 농인들의 부족한 컨텐츠 접근성을 지원하기 위해 영상의 내용을 정확하게 설명하는 문장을 생성하는 연구들이 이루어지고 있다. 영화 지문은 영상의 스토리를 파악할 수 있을 정도로 영상 지문 데이터보다 더 묘사적으로 서술되어 있다. 하지만, 일반적인 영상 지문 생성 모델을 적용할 경우, 적은 분포로 나타난 단어들을 지문에서 생성하기 어렵다. 본 논문에서는 이러한 문제를 해결하기 위해 장소와 시간 정보를 활용하여 영화의 지문을 생성하는 방법을 제안한다. 이를 위해, 인코딩 과정에서는 객체와 행위로 구성된 이벤트 벡터가 장소와 시간의 조건에 따라 다른 공간에 임베딩되며, 디코딩 과정에서는 장소와 시간 정보에 따라 생성되는 단어 분포가 달라지도록 모델을 설계한다. 실험에서는 한국영화를 대상으로 영화 지문 생성을 수행하였고, 장소, 시간 조건에 따라 생성된 지문의 성능을 정량적, 정성적으로 평가하였다. 그 결과, 제안하는 방법이 더 정확하고 묘사적인 문장을 생성할 수 있었다.

Ⅰ. 서 론

영상 지문 생성 방법은 컴퓨터 시각 지능 분야와 자연어 처리 분야를 융합한 것으로써, 영상과 관련된 내용을 문장으로 생성하는 것을 목적으로 한다. 지문 생성 대상이 영화 컨텐츠일 경우, 영화 지문 생성 방법으로 불리우며, 영화 지문의 경우에 영상 지문 생성 문장보다 더 묘사적이다. 우리는 더 정확한 영화 지문 생성을 위해 각본 형식을 참조하여 지문 생성 모델을 구성한다. 각본은 영상이 제작되기 전에 작성된 구조화된 텍스트로써, 장면 단위로 장소, 시간, 배경 인물, 대사, 지문 등이 작성되어 있다. 이러한 정보들은 각각의 장면 안에서 서로 유기적으로 연결되어 있으며, 각각의 정보의 구성에도 영향을 미친다. 따라서, 영화 지문 생성에서 배경 정보, 즉 장소와 시간 정보를 반영하면 더 정확한 문장들을 생성할 수 있다. 따라서, 본 논문에서는 이러한 구조적 형식을 반영할 수 있는 조건부 지문 생성 모델을 제안한다. 제안하는 방법은 인코더-디코더로 구성되어 있으며, 인코더는 장소와 시간 정보에 따라 이벤트 벡터를 다른 공간에 임베딩하며, 디코더에서는 장소와 시간 정보에 따라 생성되는 단어 분포가 달라진다. 실험에서는 한국영화 201편의 지문을 학습하고 24편의 영화에서 테스트한다. 또한, 장소와 시간 정보를 반영하지 않은 지문 생성 모델과 비교했으며, 제안하는 방법이 장소와 시간 조건에 따라 더 정확하고 묘사적인 문장을 생성할 수 있었다.

Ⅱ. 배경 정보를 활용한 영화 지문 생성 방법

본 논문에서 제안하는 영화 지문 생성 방법은 기본적으로 영상의 내용을 벡터로 압축하여 표현하는 인코더, 압축된 벡터를 기반으로 문장을 생성하는 디코더 구조를 가진다. 특히, 제안하는 방법에서는 각본의 구조를 참조하여, 각본에서 장소, 시간, 지문이 서로 유기적으로 연관되어 있음에 집중했다. 영화 각본을 분석하여 장소와 시간 정보에 따라 지문에서 사용되는 단어의 분포가 달라지는 것을 확인했고, 이러한 특징을 지문 생성 모델에 반영했다. 장소와 시간 정보들의 조건에 따라 인코딩되는 벡터와 디코딩되는 단어가 적응되도록 설계했다.

인코더는 프레임에서 추출된 객체, 행위, 장소, 시간 특징 정보를 입력받으며, 각각의 입력 특징 정보를 인코딩하는 특징 정보 인코더, 조건부 인코딩 계층으로 구성된다. 특징 정보 인코더는 Linear Layer - Attention Layer - Relu Activation Layer로 구성된다. Attention Layer에서는 이전 단계에서 생성된 문장의 Hidden state를 참조하여 각각의 특징 정보를 임베딩한다. 조건부 인코딩 계층에서는 장소와 시간 정보를 조건적으로 반영한다. 먼저, 객체-행위 특징 정보를 기반으로 객체의 행위를 표현하는 이벤트 벡터를 만든다. 생성된 이벤트 벡터에 장소와 시간 특징 정보를 각각 부여하여 적응된 조건부 이벤트 벡터를 생성한다. 이는 장소와 시간에 따라 모델이 참조하는 정보를 조건적으로 임베딩 할 수 있도록 한다. 최종적으로 이벤트 벡터와 장소, 시간의 조건부 이벤트 벡터들을 서로 연결하여 영상 인코딩 벡터 ve를 생성한다.

디코더는 생성된 영상 인코딩 벡터를 기반으로 문장을 생성하며, LSTM 기반으로 언어 모델을 학습한다. 문장 S가 $\{w_0, w_1, \cdots, w_n\}$으로 구성되어 있다고 할 때, 언어 모델에서는 t-단계에서 단어 시퀀스

그림 1 영화 '파파-62' 장면의 프레임

표 1 영화 '파파-62' 장면의 지문 생성 결과

구분	생성 지문
정답	사람과 대화를 하는 사람
OA	사람들이 이야기한다
OAPT	교회에서 대화하는 사람들

$\{w_0, \cdots, w_{t-1}\}$를 부여하고 순차적으로 생성될 단어 w_t를 학습한다. 인코더에서 인코딩된 영상 인코딩 벡터 ve와 LSTM Cell을 통해 배출된 출력 벡터 h_t를 서로 연결하여 softmax 함수를 통해 단어 생성 확률을 추정한다. 추가적으로 장소와 시간에 따른 단어 분포 조정을 위해 h_t와 각각의 장소, 시간 정보 인코딩 벡터를 연결하며, 각각의 벡터에 대해 softmax 함수를 통해 장소, 시간별 단어 생성 확률을 추정한다. 최종적으로 세 가지의 단어 확률을 서로 곱하여 생성될 단어 확률을 도출한다. 제안하는 방법에서 손실 함수는 Categorical Crossentropy를 사용한다.

Ⅲ. 실험

실험에서는 한국 영화 224편을 20초 단위의 장면 클립으로 분할하고, 분할된 장면 영상에 마다 지문 문장을 태깅하였다. 201개의 영화의 67,715개의 장면 지문을 학습하였고, 23개의 영화에서 7,573개의 장면의 지문을 테스트하였다.

영상 인코딩을 위해 각각의 장면 영상마다 1초에 하나의 프레임을 추출했다. 객체, 행위, 장소, 시간 특징 정보는 프레임 단위로 추출하였다. 객체 정보는 미리 학습된 Resnet-152[1] 모델을 사용하여 프레임마다 2,048 차원의 벡터를 추출했고, 행위 정보는 미리 학습된 ResNeXT-101[2] 모델을 사용하여 1초당 1.5개의 2,048 차원 벡터를 추출했다. 장소와 시간 특징 정보는 VGG-16[3]과 Dual Attention[4]을 활용하여 분류 모델을 구축하고, 모델 네트워크의 마지막 계층에서 각각 512 차원의 벡터를 추출하였다.

지문 문장은 어절 단위로 분리하여 단어 사전을 구축하였고, 총 8,727개의 단어로 구성되어 있다. 단어 임베딩 모델은 미리 학습된 300 차원의 FastText 모델[1]을 사용하였다. 문장에서 최대로 생성 가능한 단어는 20개로 한정하였다.

성능 평가를 위해, 객체-행위 정보만을 사용하여 인코딩하고 LSTM으로 문장을 생성한 OA 모델과 비교하였다. 장소와 시간 정보를 반영하는 제안하는 방법은 OAPT로 정의했다.

본 논문에서는 영화 지문 생성의 성능을 확인하기 위해, 생성된 지문에 대해 정량적, 정성적 평가를 수행하였다. 전체 테스트 데이터에 대한 BLEU 점수는 표 3에서 확인할 수 있다. 또한, 위의 그림 1과 그림 2에

1) https://fasttext.cc/docs/en/crawl-vectors.html

그림 2 영화 '야수-199' 장면의 프레임

표 2 영화 '야수-199' 장면의 지문 생성 결과

구분	생성 지문
정답	사람들이 앉아 있다
OA	사람들이 이야기한다
OAPT	사람이 술을 마신다, 사람이 사람에게 말한다, 사람이 말한다

대해 지문 문장 생성 결과는 표 1과 표 2과 같다. 본 논문에서는 어절 단위로 사전을 구축하였기 때문에, BLEU@2 점수가 일반적으로 생성 모델의 평가 지표로 사용되는 BLEU@4 점수와 비슷하다고 볼 수 있다. 표 3에서 볼 수 있듯이, 장소와 시간을 모두 활용한 방법이 그렇지 않을 때보다 좋은 성능을 보였다. 이는 장소와 시간 정보 반영이 영화 지문 생성에서 효과적으로 작용하고 있음을 나타낸다.

실제로 표 1, 표 2에서 생성된 지문 문장을 확인해보면, 장소와 시간을 모두 사용한 모델이 정답 문장과 비교해볼 때 가장 정확하고 묘사적인 문장을 생성했다. 예를 들어, 표 1의 경우에 정답 문장은 "*사람들과 대화하는 사람*", 또한, OA 모델은 "*사람들이 이야기한다*"라는 문장을 생성했으나, OAPT 방법은 "*교회에서 대화하는 사람들*"과 같이 장소를 포함하는 문장들을 생성해냈다. 또한 표 2에서 볼 수 있듯이, OAPT 방법이 "*사람이 술을 마신다, 사람이 사람에게 말한다, 사람이 말한다*"와 같이 더 자세하고 풍부한 내용을 생성할 수 있었다. 이는, OAPT에서는 장소가 바(Bar), 시간은 저녁이라는 것을 인지했고, 이에 따라 "*술을 마신다*"라는 문장을 생성한 것으로 유추할 수 있다.

표 3 생성된 지문의 BLEU 성능

구분	BLEU@1	BLEU@2	BLEU@3	BLEU@4
OA	0.1286	0.0363	0.0092	0.0018
OAPT	0.1328	0.0410	0.0090	0.0037

Ⅳ. 결론

본 논문에서는 장소와 시간 정보를 활용한 영화 지문 생성 방법을 제안한다. 이를 위해, 제안하는 방법은 장소와 시간 정보에 따라 적응된 이벤트 벡터로 영상을 인코딩하고, 디코더에서는 장소와 시간 정보에 따라 단어 분포가 적응되도록 설계했다. 실험에서는 한국 영화 224편을 대상으로 지문 생성 실험을 수행했다. 그 결과, 장소와 시간 정보를 사용하였을 때, 더 정확하고 묘사적인 문장을 생성할 수 있었다.

ACKNOWLEDGMENT

본 연구는 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음. [20ZH1200, 초실감 입체공간 미디어·콘텐츠 원천기술 연구]

참 고 문 헌

[1] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. "In Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 770-778, 2016.

[2] Kataoka, H., Wakamiya, T., Hara, K., and Satoh, Y. "Would Mega-scale Datasets Further Enhance Spatiotemporal 3D CNNs?", arXiv preprint arXiv:2004.04968. 2020.

[3] Woo, S., Park, J., Lee, J. Y., and So Kweon, I., "Cbam: Convolutional block attention module", In Proceedings of the European conference on computer vision (ECCV), pp. 3-19, 2018

[4] Simonyan, K., and Zisserman, A., "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556., 2014.