Multi-head Self-Attention을 이용한 비디오 캡션 생성

박다솔[1O], 손정우[2$], 김선중[2#] ,차정원[1*]
창원대학교[1], 한국전자통신연구원[2]
{dasol_p [O],jcha[*]}@changwon.ac.kr, {jwson[$],kimsj[#]}@etri.re.kr

Video Caption Generation Using Multi-head Self-Attention

Da-Sol Park[1O], Jeong-Woo Son[2$], Sun-Joong Kim[2#], Jeong-Won Cha[1*]
Changwon National University[1], Electronic Telecommunications Research Institute[2]

요 약

비디오 캡션 생성은 짧은 비디오를 설명하는 문장을 생성하는 태스크이다. 본 논문에서는 멀티 헤드 셀프 어텐션을 이용하여 캡션을 생성하는 방법을 제안한다. 비디오를 벡터로 표현하고 범주 정보를 단어 임베딩을 통해 벡터를 생성하여 동시에 입력 받는다. 비디오 정보와 범주 정보는 셀프 어텐션을 통해 중요한 비디오 내 프레임을 선택하게 된다. 이 과정을 반복함으로써 중요한 프레임들은 더욱 강조하게 된다. 이를 입력으로 하여 마지막 생성기를 통해 문장을 생성한다. MSR-VTT 데이터셋을 이용하여 비디오 캡션 생성을 진행한다. 제안 방법은 BLEU 4 0.313, CIDEr 0.237의 성능을 보였다. 제안한 방법은 3D CNN과 같은 추가정보를 사용하지 않고도 좋은 성능을 얻을 수 있었으며 비디오 캡션 생성 뿐만 아니라 비쥬얼 QA 등과 같은 분야에도 사용할 수 있을 것이다.

1. 서 론

비디오 캡션 생성이란 입력으로 사용되는 비디오에 대해 설명하는 자연어 문장을 생성하는 태스크이다. 일반적으로 2가지 과정을 거쳐 문장을 생성한다. 첫 번째 과정은 입력 비디오로부터 분산 표현인 특징을 추출하는 과정이다. 일반적으로 비디오를 1초당 n개의 프레임 이미지로 생성하여 특징을 추출하는 과정을 포함하고, 이를 이용하여 분산표현 자질로 생성하는 과정이다. 두 번째 과정은 추출한 특징을 이용하여 문장을 생성하는 과정이다.

비디오를 분산표현 자질로 생성하기 위해서는 각 프레임(frame)을 CNN(Convolution Neural Network)에 이용한다. 이미지를 처리하기 위한 2D CNN은 단일 이미지에만 적용되고, 시간 정보를 사용하지 않는다. 이를 해결하기 위해 3D CNN이 연구 되었다. 연속 프레임에 대한 정보를 가질 수 있으며 시간 정보를 인코딩하는 학습이 가능하게 되었다.

3D CNN을 적용하기 위해 미리 학습된 모델을 사용하여 미세조정(finetuning)도 학습 속도가 너무 느리고 비용 소모가 많이 든다. 그리고 은닉층이 깊은 네트워크는 학습하기가 어렵다. 오히려 2D CNN을 미세조정하는 것이 3D CNN을 사용하는 것보다 성능이 더 좋은 연구도 있었다[1,2]. 비디오의 특징을 추출하기 위해서 시간 정보가 필요하며, 2D CNN과 RNN(Recurrent Neural Network)을 이용하여 비디오 벡터를 생성하여 문장을 생성하기도 하였다[3]. 시간 정보를 인코딩할 수 있는 3D CNN을 사용하지 않고 2D CNN과 멀티 헤드 셀프 어텐션(Multi-head Self-attention)을 이용하여 비디오 캡션에 사용하는 연구를 진행한다. 본 논문에서는 MRC(Machine Reading Comprehension)에서 아이디어를 차용하여 비디오 캡션 생성 문제를 해결한다. 따라서 비디오를 문서로 간주하고 중요한 부분을 선택하고 이를 이용해 캡션을 생성하는 알고리즘을 제안한다.

2. 관련 연구

비디오를 하나의 이미지와 같은 표현으로 바꾸는 과정은 주로 CNN을 이용하고, 문장 생성을 위하여 RNN을 이용하는 연구가 진행되었다. [4]는 비디오 클립의 이미지 프레임들을 CNN을 거쳐 특징을 추출하되 이들의 평균값을 계산하여 비디오 벡터로 사용하였다. 비디오 벡터를 매 시간 단계마다 입력으로 받아 LSTM(Long-Short Term Memory) 2개의 층을 이용하여 비디오를 설명하는 문장을 생성하였다.

멀티모달을 이용한 비디오 캡션 연구도 진행되고 있다. [5]은 문장 생성 과정에서 특정 요소에 선택적으로 어텐션을 취하기 위해 프레임, 모션, 오디오를 이용한 멀티모달 스트림을 활용한다. 시퀀스-투-시퀀스(Sequence-to-Sequence) 프레임워크를 기반으로 하고 있으며 3개의 LSTM 모델을 이용하여 비디오 프레임, 비디오 모션, 오디오의 자질을 각각 인코딩한다. 모달 스트림을 결합하여 디코더의 초기 상태를 생성한다. 멀티레벨 어텐션 메커니즘(Multi-level attention mechanism)은 비디오의 핵심 단서를 효과적으로 포착하는 데 사용되며 시간 순서와 결합된 스트림에서 중요한 자질을 출력한다. 이를 이용하여 LSTM 2개의 층으로 각 단어들을 예측하여 비디오 캡션을 생성한다.

3. 제안 방법

[그림 1]은 본 논문에서 제안한 시스템을 보여준다. 인코더에서 3가지 과정(임베딩, 스택 임베딩 인코더 블록, 비디오-범주 정보(Category) 어텐션)을 거친 후 디코더 과정에서 비디오에 대한 문장을 생성한다.

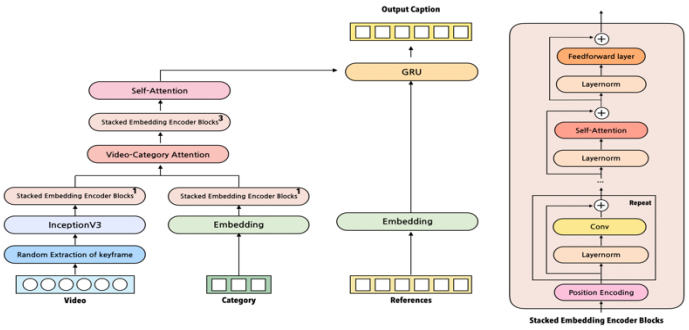

그림 1 제안 모델의 구조

3.1 임베딩

각 비디오의 프레임 중 100개를 임의 추출하여 키프레임(keyframe)이라고 가정한다. 키프레임을 입력하여 InceptionV3의 결과를 출력한다[6]. 범주 정보는 단어 임베딩과 문자 임베딩을 이용하여 분산 표현을 생성한다. 단어 임베딩은 학습된 fastText[7]를 사용하고 학습되지 않도록 설정한다. 문자 임베딩은 학습된 fastText를 사용하고 학습되도록 설정한다. CNN와 max-over-time-pooling을 거친 후 Highway network를 통한 결과를 사용한다. 단어 임베딩과 문자 임베딩을 연결하여 사용한다.

3.2 스택 임베딩 인코더 블록

포지션 인코딩(Position Encoding)은 임베딩에서 추가적인 위치 정보가 없기 때문에 sine, cosine 함수를 이용하여 위치 정보를 가질 수 있도록 한다. 이후 레이어 정규화(Layer Normalization)을 진행하고, 분리가능한 깊이별 컨볼루션 층(Depthwise separable convolution Layer)은 설정한 층 수만큼 반복한다. 레이어 정규화를 진행하고 셀프 어텐션을 거친다. 셀프 어텐션은 각 입력으로 들어오는 비디오와 범주 정보에 대해 자기 자신을 잘 표현할 수 있는 비디오와 범주 정보 쌍을 찾아 이를 이용하여 임베딩을 생성한다. 레이어 정규화를 거친 후 순방향 층(FeedForward Layer)를 거친다. 각 헤드(Head)가 만들어낸 셀프 어텐션이 치우치지 않도록 균등하게 섞는 역할을 한다. 그림 내 명시된 숫자는 해당 모듈의 반복 횟수를 의미한다.

3.3 비디오-범주 정보 어텐션

총 4개의 벡터를 구하여 연결하여 사용한다. 인코딩된 비디오 벡터와 범주 정보 벡터를 이용하여 유사도 매트릭스를 구한다. 유사도 매트릭스를 이용하여 각 행에 대해 softmax를 취해 정규화한 유사도 매트릭스를 구한다. 정규화한 유사도 매트릭스와 범주 정보 벡터를 이용하여 어텐션을 계산한다. 유사도 매트릭스와 정규화한 유사도 매트릭스, 비디오 벡터를 이용하여 어텐션을 구한다. 스택 임베딩 인코더 블록을 3번 반복하여 나온 출력 벡터에 셀프 어텐션을 거쳐 최종 인코딩 결과물이 생성된다.

3.4 디코더

인코더의 결과물은 비디오와 범주 정보의 유익한 자질들을 생성한 벡터이며 GRU(Gated Recurrent Unit)의 초기 상태로 설정하고 비디오에 대한 캡션을 생성한다. 문장을 생성하기 위해서 단어 임베딩과 문자 임베딩을 이용하여 사용한다.

4. 데이터

실험에는 MSR-VTT 데이터셋[8]을 이용하였다. MSR-VTT 데이터셋은 Microsoft사에서 2017년에 공개한 데이터셋이다. 한국어 번역 작업을 통해 한국어 데이터셋을 구축한 후 실험을 진행하였다. [표 1]은 MSR-VTT 데이터셋 내 비디오와 레퍼런스에 대한 통계이다. 하나의 클립에 대해 20개의 레퍼런스가 존재한다. 20개의 문장 표현은 다르지만 해당 클립을 설명하는 문장의 의미는 같다. 비디오를 다운받아서 사용해야함으로 계정이 삭제되었거나 게재된 비디오가 삭제되었을 경우 다운받지 못한다는 문제점이 존재한다.

‘음악, 사람, 게임, 스포츠/액션, 이벤트/뉴스/정책, 교육, TV, 영화/코미디, 애니메이션, 차량/운송수단, 방법, 여행, 과학/기술, 동물/애완동물, 가족/아동, 다큐멘터리, 음식/음료, 요리, 미용, 패션, 광고’로 총 20개의 범주 정보를 가지고 있다. 여러 단어로 구성되어 있는 범주 정보는 가장 처음 나오는 단어를 대표 범주 정보라고 가정하여 이용한다.

표 1 MSR-VTT 데이터 통계

	비디오 수	클립 수	총 레퍼런스 수
공식 데이터	7,180	10,000	200,000
수집 데이터	5,706	7,825	156,500

한국어 번역 작업 후 형태소 분석을 통해 품사를 제거 후 사용하였으며 캡션 내 형태소 갯수를 최대 15개로 설정하였다. 형태소 갯수가 15개 이상인 캡션은 제외하여 진행하였다. [표 2]는 실험 데이터 정보이다.

표 2 실험 데이터 정보

코퍼스 분류	비디오 수	레퍼런스 수
학습 코퍼스	6,329	70,527
검증 코퍼스	713	9,087
평가 코퍼스	783	9,661
총합	7,825	89,275

5. 실험 방법 및 설정

[그림 2]는 [표 4]에 기본(baseline) 모델의 구조이다. 동일하게 100개의 키프레임을 임의 추출하고 InceptionV3의 결과물을 LSTM을 거쳐 128차원의 인코더 벡터를 생성한다. LSTM의 초기 상태로 설정하고 비디오에 대한 캡션을 생성한다. 단어 임베딩만 사용했으며 단어 임베딩의 차원은 128차원이다. 이 모델은 학습 데이터를 3,500개를 사용하였으며 평가 데이터는 동일하다.

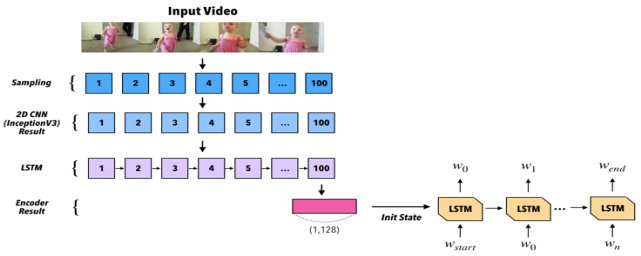

그림 2 기본 모델(2D CNN+LSTM) 구조도

제안 모델을 이용하여 비디오 프레임을 임의로 100개를 추출하여 실험을 진행하였다. [표 3]은 실험 파라미터이고 [표 4]는 성능 결과표이다. 분류 내 괄호는 성능을 측정하기 위해 레퍼런스의 사용한 형태소 갯수를 의미한다.

표 3 실험 파라미터

파라미터 분류	설정 값	파라미터 분류	설정 값
비디오 프레임 수	100	Dropout	0.1
비디오 임베딩 차원	2,048	학습률	0.001
단어 임베딩 차원	300	사용한 헤드 수	2
문자 임베딩 차원	200		

표 4 성능 결과표

분류	B@1	B@2	B@3	B@4	ROUGE_L	CIDEr
기본 모델 (전체 형태소)	0.631	0.323	0.198	0.121	-	-
제안 모델 (형태소 15개 이하)	0.640	0.490	0.392	0.313	0.500	0.237
제안 모델 (전체 형태소)	0.663	0.506	0.404	0.321	0.494	0.116

제안 모델에서 비디오를 설명하는 문장의 생성이 잘된 경우는 총 3가지로 분류할 수 있다. (1) 부가적 수식을 통해 새로운 단어가 생성된 경우이다. (2) 영상에 대해 더 포괄적으로 설명해주는 행위를 생성한 경우이다. (3) 복합 컨텍스트의 영상에서 단일 컨텍스트를 찾아 문장을 생성한 경우이다.

제안 모델에서 문장의 생성이 잘 되지 않은 경우는 총 2가지로 분류할 수 있다. (1) 비디오를 잘 못 인식한 경우이다. (2) 사전 내 존재하지 않은 단어가 많이 발생한 경우이다.

기본 모델과 제안 모델의 결과를 비교해보았을 때 기본 모델이 사전 내 존재하지 않은 단어를 많이 생성했으며 어두운 화면이나 장면의 변화에 대해 인식률이 낮은 경향을 보였다. 제안 모델은 기본 모델에 비해 사전 내 존재하지 않은 단어를 적게 출력하며 복합 컨텍스트에 대한 출력 또한 비교적 잘된다고 판단된다.

6. 결론 및 향후 연구

본 논문에서는 MRC에서 영감을 받아 비디오 캡션 생성을 진행한다. 비디오 캡션을 생성하기 위해 비디오 내 프레임을 이용하여 중요한 자질을 생성해야 한다. 이를 해결하기 위해 학습을 통해 비디오를 벡터로 표현하고 범주 정보를 단어 임베딩을 통해 벡터를 생성하여 동시에 입력 받는다. 비디오 정보와 범주 정보는 셀프 어텐션을 통해 중요한 비디오 내 프레임을 선택하게 된다. 이 과정을 반복함으로써 중요한 프레임들은 더욱 강조하게 된다. 이를 입력으로 하여 마지막 생성기를 통해 문장을 생성한다.

MSR-VTT 데이터를 이용하여 실험을 진행하였고 비디오를 설명하기 위해 추가적인 단어 생성과 포괄적인 행위의 문장 생성, 복합 컨텍스트에서 단일 컨텍스트를 찾아 문장을 생성한 경우가 존재했으며, 비디오의 행동을 오인식한 경우와 사전 내 존재하지 않은 단어를 많이 출력하는 경우가 있었다.

제안 방법이 3D CNN과 같이 추가 정보를 이용하지 않아도 좋은 성능을 얻을 수 있었으며 2D CNN과 멀티 헤드 셀프 어텐션을 이용함으로써 비디오를 표현하기 위한 자질을 생성하는데 도움을 주었다고 할 수 있다.

향후 연구로는 제안한 구조를 개선하여 보다 정확한 문장을 생성하는 연구를 진행할 예정이다.

Acknowledgement

본 연구는 한국전자통신연구원 연구운영비지원사업의 일환으로 수행되었음. [19ZH1300, 오픈시나리오 기반 프로그래머블 인터랙티브 미디어 창작 서비스 플랫폼 개발]

참고 문헌

[1] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[2] Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." *Proceedings of the IEEE international conference on computer vision*. 2015.

[3] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[4] Venugopalan, Subhashini, et al. "Translating videos to natural language using deep recurrent neural networks." *arXiv preprint arXiv:1412.4729*. 2014.

[5] Xu, Jun, et al. "Learning multimodal attention LSTM networks for video captioning." *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017.

[6] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[7] Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *Transactions of the Association for Computational Linguistics* 5. 2017.

[8] Xu, Jun, et al. "Msr-vtt: A large video description dataset for bridging video and language." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.