고밀도 비디오 캡션을 위한 객체 검출 및 속성 분석에 관한 연구

윤은지, 정호영*

*경북대학교 인공지능학과

(yooneunji, hoyjung)@knu.ac.kr

An Object Detection and Attribute analysis for Dense Video Captioning

Yoon Eun Ji, Jung Ho Young*

*Kyunpook Univ.

요 약

본 논문에서는 비디오로부터 자연어 캡션을 자동으로 생성하기 위한 새로운 접근 방식을 제안한다. 제안된 모델은 객체 감지와 속성 분석을 통합하여 비디오의 시각적 정보를 더욱 세부적으로 분석하고, 이를 자연어로 표현하는 것을 목표로 한다. 제안한 모델은 비디오 인코더와 캡션 디코더로 구성되며, 입력 비디오에서 시각적 특징을 추출하기 위해 YOLOv5를 사용하여 객체를 탐지한 후, 사전 학습된 ResNet과 ViT를 통해 각각 로컬 및 글로벌 시각적 특징을 캡처한다. 이러한 특징들은 비디오의 맥락을 정교하게 반영하여, 자연스러운 캡션을 생성하는 데 활용된다. MSVD 데이터셋을 이용한 다양한 실험을 통해, MSE 손실 함수를 적용한 ResNet과 ViT 결합 모델이 CIDEr 점수 0.727을 달성하며, 기존 연구 대비 0.21 이상의 성능 향상을 확인할 수 있었다. 이는 제안된 모델이 비디오 내 시각적 정보의 연결성과 연속적 맥락을 효과적으로 통합하여, 캡션 생성 성능을 크게 향상시킬 수 있음을 확인하였다.

Ⅰ. 서론

인공지능의 급속한 발전으로 인해 이미지와 자연어의 동시 분석이 필요한 복잡한 인지 능력과 예측 능력에 관한 관심이 높아졌다. 고밀도 비디오 캡션은 입력 비디오에서 이벤트 구간을 찾아내고, 각 구간에 대한 자연어 설명 문장을 생성하는 작업이다. 고밀도 비디오 캡션 생성에는 여러 동작과 이벤트를 고려하는 긴 비디오를 처리하는 작업을 포함한다. 효과적인 캡션 생성을 위해 분할되지 않은 비디오를 적절하게 나타내는 특징 추출, 이벤트 또는 동작이 있는 시간적 영역 식별, 감지된 이벤트 시간 영역을 설명하는 자연어 문장 생성이 요구된다. 또한, 비디오 콘텐츠에 대한 이해를 높이기 위해 객체 검출을 통합하여 각 프레임 내의 객체를 식별하고 분석하여 공간 관계와 객체 동작에 대한 자세한 표현을 제공할 수 있다. 이러한 로컬(Local) 의미 기능은 비디오 전반에 걸쳐 동작을 나타내는 글로벌(Global) 의미 기능과 결합하여 입력 특징으로 활용한다. 그리고 주의집중을 적용한 캡션 생성 네트워크를 사용하여 시각 특징들을 중요도에 따라 선택적으로 활용하여 캡션 생성 성능을 향상시킨다. 따라서 본 논문에서 제안하는 모델의 성능과 효과를 분석하기 위해 대용량 벤치마크 데이터셋인 MSVD와 MSR-VTT 데이터셋을 이용한 다양한 실험을 수행하고, 제안하는 모델의 성능과 효과를 분석하여 소개하고자 한다.

Ⅱ. 본론

본 논문에서 객체 검출과 속성 분석을 통합하여 자연어 캡션을 자동으로 생성하는 접근 방식을 제안하고, 이를 위한 고밀도 비디오 캡션 생성 모델의 구조를 그림 1을 통해 설명한다. 본 논문에서 제안하는 비디오 캡션 생성 모델은 비디오 인코더와 캡션 디코더의 두 가지 모델로 구성되어 있으며, 먼저 시각적 특징 추출 네트워크를 통해 입력 영상을 대표하는 특징을 추출한다. 모델 시작 부분에 위치한 비디오 인코더는 시각적 분석 기술을 사용하여 입력 비디오에서 동작 및 객체 특징을 추출한다.

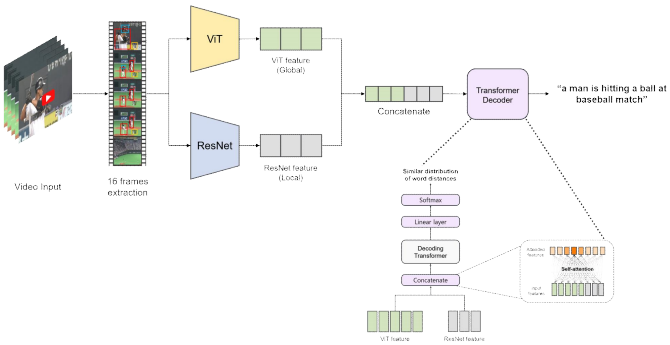

그림 1 제안된 비디오 캡션 생성 모델 아키텍처

본 논문에서는 시각적 특징을 추출하기 위하여 입력 비디오를 16개의 프레임 클립 단위로 나누어 사용하며, 각 클립 내에서 객체를 효과적으로 검출하기 위해 YOLOv5 모델을 활용한다. YOLOv5는 비디오 내에서 다수의 객체를 동시에 탐지하고, 각 객체의 위치 정보를 추출한다. 이 정보를 바탕으로 입력 비디오의 시각적 특징을 잘 드러내기 위해 합성곱 신경망의 종류 중 하나인 ResNet과 ViT를 활용한다. ResNet은 잔여 연결로 구성되어 있는 네트워크로, 입력 비디오에 대한 로컬 특징을 추출할 수 있다는 장점이 있다. 또한 ViT는 입력 비디오의 프레임에 대해 별도의 패치로 처리하는 주의 메커니즘을 사용하는 네트워크로, 장거리 종속성을 캡처함으로써 다양한 지역 간의 글로벌 특징을 추출할 수 있다는 장점이 있다. 본 논문에서 사용한 ResNet과 ViT는 각각 ImageNet과 ImageNet-21k 데이터셋으로 사전 학습된 네크워크이다. 추출된 시각 특징은 이후 캡션 생성 네트워크를 통해 캡션을 생성하는 데 사용된다. 본 논문에서는 추출된 시각적 특징을 바탕으로 캡션 생성의 효율성을 높이기 위해 주의집중 메커니즘을 적용한 캡션 생성 네트워크를 제안한다. 먼저 입력 시각 영역에 대해 추출한 로컬 특징과 글로벌 특징을 단순 연결하여 입력 특징으로

활용한다. 이후 캡션 단어의 정보를 기반으로 주의집중 메커니즘을 통해 단어의 순서와 배열을 고려할 수 있도록 한다. 식 (1)과 같이 프레임 간에 비디오 특징과 실제 비디오에 대해 예측값과 실제값의 MSE(Mean Squared Error)[1]를 기준으로 학습한다. 이는 비디오 내에서 동일한 객체나 시각적 특징을 추적하고 시각적 정보의 일관성 학습을 통해 비디오 프레임 간에 유사한 시각적 특징을 보존하도록 한다. 이후 식 (2)와 같이 학습 과정에서 타겟과 모델이 생성한 출력의 차이를 측정하기 위해 손실 함수[2]를 사용한다. 이를 통해 각 시점에서 모델이 예측한 단어와 실제 타겟 단어 사이의 유사성을 평가하고, 모델이 어떤 단어를 높은 확률로 예측하는지 분석할 수 있다.

$$Loss = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \quad (1)$$

$$Loss = -\sum_{i=1}^{C} Y_i \log(\widehat{Y}_i) \quad (2)$$

본 논문에서 MSVD(Microsoft Research Video Description) 및 MSR-VTT(Microsoft Research Video-to-Text)와 같은 벤치마크 데이터셋을 사용하여 비디오 캡셔닝 작업을 위한 실험을 진행했다. MSVD 데이터셋에는 1,970개의 짧은 비디오 클립과 각 비디오에 대한 여러 캡션이 함께 포함되어 있다. MSVD 데이터셋의 비디오 클립은 1,200개, 100개, 670개로 각각 학습, 검증 및 테스트에 활용된다. 다양한 주제와 활동을 다루는 MSR-VTT 데이터셋에는 20초 이하의 길이를 가진 10,000개의 비디오 클립이 포함되어 있다. 이들은 학습, 검증 및 테스트 세트로 각각 6,513개, 497개, 2,990개로 나뉘어 사용된다.

Ⅲ. 연구 결과

본 논문에서는 BLEU@N, METEOR, ROUGE-L, CIDEr를 포함한 여러 평가 지표를 사용하여 이 방법의 성능을 평가한다. BLEU@N은 생성된 캡션을 하나 이상의 참조 캡션과 비교하는 변형된 형태의 N-gram 정밀도이며, METEOR는 생성 자막과 참조 자막 간의 유사성을 평가하기 위한 Recall에 초점을 맞춘 측정 지표이다. ROUGE-L은 참조 문장에서 가장 긴 일치하는 단어 시퀀스와 비교하여 캡션의 품질을 측정한다. CIDEr는 TF-IDF 가중치를 사용하여 N-gram 유사성을 고려한다.

Pre-trained Models for Video Feature Extraction	Dataset			
	MSVD			
	BLEU@4	METEOR	ROUGE-L	CIDEr
SA[3]	0.419	0.296	-	0.517
ResNet	0.272	0.299	0.613	0.525
ViT	0.330	0.298	0.644	0.681
MSE ResNet+ViT	0.408	0.331	0.685	0.727

[표 1] MSVD 데이터셋에서 비디오 캡션 생성 생성 모델 성능 비교

[표 1]에서 확인할 수 있듯이, 로컬 특징만을 활용한 모델보다 글로벌 특징만을 활용한 모델의 성능이 더 우수한 성능을 보이는 것을 알 수 있다. 이는 글로벌 특징이 비디오의 시각적 정보를 더 정밀하게 표현하기 때문이다. 특히, MSE 손실 함수를 사용하여 ResNet과 ViT를 결합한 모델이 가장 뛰어난 성능을 나타냈으며, CIDEr 점수에서 0.727로 가장 높은 성능을 기록하였다. 이는 비디오 프레임 간 시각적 정보의 연결성을 효과적으로 강화하고, 비디오의 연속적 맥락을 보다 정확하게 반영하여 캡션을 생성하고 있음을 보여준다.

Pre-trained Models for Video Feature Extraction	Dataset			
	MSR-VTT			
	BLEU@4	METEOR	ROUGE-L	CIDEr
HMN[4]	0.435	0.290	0.627	0.515
ResNet	0.235	0.277	0.590	0.380
ViT	0.303	0.295	0.627	0.496
MSE ResNet+ViT	0.357	0.257	0.549	0.558

[표 2] MSR-VTT 데이터셋에서 비디오 캡션 생성 모델 성능 비교

[표 2]의 결과를 살펴보면, ResNet 단독 사용 시 CIDEr 점수는 0.380으로 낮았으며, ViT 단독 사용 시에도 0.496에 그쳤다. 이러한 결과를 통해 본 논문에서 제안한 모델이 글로벌 특징을 활용하여 캡션을 효과적으로 생성하지만, 추출된 로컬 특징의 정확도가 낮으면 전체 캡션 성능에도 영향을 미칠 수 있음을 확인하였다. 반면, 본 논문에서 제안한 MSE 손실 함수를 사용하여 ResNet과 ViT를 결합한 모델은 CIDEr 점수에서 0.558로 가장 높은 성능을 나타냈다. 이는 MSE 손실 함수를 통해 시각적 특징 추출 네트워크와 주의집중 메커니즘을 효과적으로 적용하고 있음을 확인할 수 있다.

Ⅳ. 결론

본 논문에서는 고밀도 비디오 캡션을 위해 객체 검출 및 속성 분석 모델을 제시하였다. YOLOv5를 통해 비디오 내 객체를 탐지하고 위치 정보를 추출하며, 사전 학습된 ResNet과 ViT를 활용하여 비디오의 시각적 특징을 캡처한다. Transformer 기반의 인코더-디코더 구조를 통해 장기 종속성 문제를 해결하며, 의미 특징 추출 네트워크와 주의집중 메커니즘을 적용하여 캡션을 생성한다. 제안된 모델이 중요한 단어를 효과적으로 생성하나, 관사와 같은 기능어의 생성에는 제한이 있는 것으로 나타났다. 향후 연구에서는 비디오의 문장 구조 분석을 통해 캡션 생성 성능을 향상시키도록 할 것이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음(IITP-2024-2020-0-01808)

참 고 문 헌

[1] Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." arXiv preprint arXiv:1511.05440 (2015).

[2] Zhang, Zhilu, and Mert Sabuncu. "Generalized cross entropy loss for training deep neural networks with noisy labels." Advances in neural information processing systems 31 (2018).

[3] Pan, Yingwei, et al. "Video captioning with transferred semantic attributes." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[4] Ye, Hanhua, et al. "Hierarchical modular network for video captioning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.