2023년도 한국멀티미디어학회 추계학술발표대회 논문집 제26권 2호

감시 시스템을 위한 새로운 캡션 생성 기법

나광호*, 이진경*, 전민성, 최경주
충북대학교 소프트웨어학과
email : kjcheoi@cbnu.ac.kr

New Caption Generation Techniques for Surveillance Systems

Gwangho Na*, Jingyeong Lee*, Minseong Jeon, Kyung Joo Cheoi
Dept. of Computer Science, Chungbuk, National University

요 약

최근 딥러닝 기술의 발전으로 객체 탐지 성능이 향상되어, 지능형 감지 시스템의 수요가 증가하고 있다. 본 논문에서는 이미지 캡셔닝 기술을 사용한 감지 시스템을 제안한다. 보다 효과적인 감시 시스템을 만들기 위해 객체의 특징 정보, 공간 정보, 행동 정보 간 관계를 유기적으로 해석하는 캡션을 생성하는 방법을 제안한다. 감시 시스템 활용에 적절한 새로운 이미지와 캡션 쌍의 데이터 세트를 구축하여 BLIP모델을 새로 구축한 데이터세트로 파인 튜닝을 진행하였다. 실험 결과 90% 성능으로 위험 상황, 사고에 대한 캡션을 생성함을 확인할 수 있었으며, 제안하는 방법이 위험 상황과 사고를 중심으로 캡션을 생성하며 객체 뿐만 아니라 행동과 공간 정보를 포함한 구체적인 묘사가 가능함을 확인할 수 있었다.

1. 서론

최근 컴퓨터비전 영역의 딥러닝 기술의 급속한 발전으로 인해 객체 탐지 및 추적 등의 성능이 향상되어, 지능형 CCTV는 차세대 CCTV로 인정받으며 그 수요가 증가하고 있다. 행정안전부의 발표 자료[1]에 따르면 전국의 CCTV 53,000대를 2027년까지 지능형 CCTV로 전면 교체할 예정이며 현재 13,000대가 교체되어 있다. 이러한 사회적 요구에 따라 사회의 안전 위협으로부터 대응하기 위해 이상 행동 및 위험 상황을 탐지하려는 시도가 활발하다.

박준태[2] 등은 객체탐지, 객체추적, 포즈 예측을 통해 '쓰러짐', '배회', '침입' 등 위험 사건을 인식하는 연구를 수행하였다. KNN을 적용하여 특정 행동을 인식하고 검출된 객체의 위치와 시간을 추출한다. 해당 위치정보가 ROI(Region of Interest)에 포함되는지의 여부와 ROI에 머무르는 시간을 계산하여 '침입'과 '배회' 상황을 인식한다. Khan[3]은 CNN을 사용하여 교통 감시 카메라 비디오로부터 교통사고와 같은 이상 상황을 탐지하고 해당 프레임을 사고로 분류하여 감지하는 연구를 수행하였다. ZHANG[4] 등은 영향지도(influence maps)와 CNN을 사용하여 객체 궤적을 생성하여 각 프레임이 사고를 나내는지 여부를 판단하여 교통사고를 감지하는 연구를 수행하였다. 이 연구는 CCTV 프레임을 이용한 교통사고 감지 연구로, 도로 위 물체들의 시공간적인 관계를 분석한다. 이러한 기존의 이상 행동 감지 연구는 객체를 인식하고 이상 행동이 탐지되면 경고 알림을 발생한다. 본 논문에서는 기존의 객체 탐지에 기반하지 않고, 이미지 캡셔닝 기술을 사용한 새로운 접근 방법의 감시 시스템을 제안하고자 한다. 이미지 캡셔닝 기술을 활용하면 장면의 복합적인 상황을 인간의 언어로 구체적으로 표현할 수 있으며, 이를 통해 감시자는 빠르게 상황을 파악할 수 있게 된다.

이미지 캡셔닝 분야에서 주로 사용되는 이미지-캡션 쌍 데이터 세트 [5],[6]은 감지 시스템에 적용하기에는 다음과 같은 문제점이 있다. 첫째, 현재 공개된 대부분의 캡션 데이터 세트는 이미지 내 큰 객체에 중점을 두고 설명하므로, 항상 사람 중심의 해석을 보장하지 않는다. 둘째, 같은 행동을 포함한 객체 정보가 공간 정보를 활용하지 않을 경우 위험 상황을 검출하지 못할 수 있다. 예를 들어, 의자에 앉는 행동은 정상적인 상황인 반면, 난간 위에 걸터앉는 행동은 위험 상황으로 볼 수 있다. 셋째, 대규모 데이터를 사용하여 학습된 대규모 멀티모달 모델은 일반적인 상황의 이미지와 캡션이 대다수를 차지하기 때문에 감시 시스템이 검출하고자 하는 상황을 정확하게 설명하는 데에 한계가 있다.

이러한 문제점을 해결하기 위하여 객체의 특징 정보, 공간 정보, 행동 정보 간 관계를 유기적으로 해석하는 감시 시스템에 적절한 캡션을 생성하는 방법을 제안한다. 이를 위해 감시 시스템에 적용 가능한 이미지와 캡션 쌍의 데이터 세트를 구축하였고, 이 데이터 세트를 활용하여 BLIP[7] 모델을 활용한 이미지 캡셔닝 기반의 감시 시스

※ 이 논문은 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업(2019-0-01183)의 지원을 받아 작성되었음

템을 제안한다.

2. 제안 방법

본 논문에서는 감시 시스템을 위한 이미지 캡션의 새로운 구조를 제안한다. 데이터세트 구축 시 캡션의 문장 구조에는 객체의 종류와 객체정보, 객체의 행동, 행동이 발생한 공간 정보가 모두 포함되도록 하였다. 그림 1은 데이터 세트 구축 시 캡션을 어떻게 작성해야 하는지, 캡션 작성 규칙에 대한 예시를 보여준다. 그림 1의 캡션에서 빨간색으로 표시된 부분은 객체와 종류와 객체 정보를 나타낸다. 객체 정보는 사람, 운송 수단, 화재와 같은 객체를 포함하고 사람이 성인인지 어린이인지 구분해야 하며, 인상착의도 표현한다. 초록색으로 표시된 부분은 객체의 행동을 나타낸다. 그림 1은 제안한 캡션 구조에 맞춰 제작한 데이터 세트 예시이다. 그림 1의 (a), (b)는 동일하게 앉아 있는 행동이지만, 그림 1(a)는 공원에 있는 의자에 안전하게 앉아 있는 정상적인 상황인 반면, 그림 1(b)는 난간 위에 걸터앉아 있는 위험 상황으로 볼 수 있다.

(a) (b)

(a) A man wearing a green jacket is sitting on a bench in the park. (b) A woman wearing a gray cardigan is sitting on a rooftop railing.

그림 1. 데이터세트의 캡션 형식

캡션 제작을 위한 다양한 이미지는 구글, 유튜브, 공공데이터 등 다양한 곳에서 수집하였다. 표 1은 제작한 데이터 세트의 클래스별 이미지 수이다. 총 2,193장의 이미지를 수집하였고, 각각의 이미지는 모두 제안한 구조에 맞춘 캡션을 가지도록 구축되었다. 데이터 세트는 '싸움', '쓰러짐', '화재', '교통사고' 및 주의 행동 등의 5개의 클래스로 구분되어져 있다. 각 클래스는 빈번하게 발생하는 위험 상황과 동일 행동의 공간에 따른 변화를 반영하기 위해 선정되었다.

표 1. 클래스별 이미지 수

클래스	싸움	쓰러짐	화재	교통 사고	위험 행동
개수	505	601	205	338	544

3. 실험 및 결과

본 논문에서는 제안 방법을 통해 캡션이 효과적으로 학습되고, 생성됨을 증명하기 위해 대규모 언어 모델(LLM; Large Language Model) 기반 대규모 멀티모달 모델(Large Multi-modal Model)인 BLIP 모델을 새로 구축한 데이터세트로 파인 튜닝을 진행하였다. BLIP 모델은 시각적 정보와 언어 정보 간의 상호작용을 통합하여 멀티모달 태스크를 효과적으로 수행할 수 있다. 실험에 사용한 데이터 세트는 총 2193장이며, 이 중 학습 데이터는 1883개, 검증 데이터는 210개를 사용하였다. 테스트 세트는 각 클래스 별로 20장씩 추출하여 총 100장의 데이터를 사용하여 실험하였다. 옵티마이저(Optimizer)로 AdamW[12], 로스(loss)는 크로스 엔트로피(Cross Entropy), 학습률은 0.0005, 에포크(epoch)는 20으로 설정하였다.

그림 2는 테스트 데이터세트를 사용하여 기존의 BLIP의 모델[7]과 BLIP 모델을 새로 구축한 데이터 세트로 파인 튜닝을 한 결과를 비교한 것이다. 그림 2(a1)는 전체적인 상황을 잘 표현하였지만, 객체 간 행동에 대한 묘사가 부족하여 위험 상황과 일반 상황을 판단하기 어렵다. 그림 3(a2)는 생성된 캡션에 객체와 공간에 대한 정보뿐만 아니라 행동 정보인 "punching"을 포함하고 있어 위험 상황임을 알 수 있다. 그림 2(b1)의 경우에도 행동 정보의 대한 묘사가 부족한 캡션을 생성하였다. 그림 2(b2)는 두 객체와 특징, 위험 행동, 공간에 대한 구체적인 캡션을 생성하였다. 그림 2(c1)의 경우, 작은 객체인 쓰러진 사람보다 더 큰 객체인 tv와 사람의 손에 중점을 두고 해석한 캡션이 생성되었다. 반면 그림 2(c2)는 쓰러진 사람에 대해 캡션을 생성하였고 행동과 공간 정보를 포함한 문장 구조로 생성되었다. 이를 통해 제안 방법으로 구축한 데이터세트는 기존 캡션 모델에서 수행하기 어려웠던 위험 상황에 대한 캡션 생성뿐만 아니라 객체와 특징, 행동, 공간에 대한 정보를 제안한 형식에 맞게 문장을 효과적으로 생성하는 것을 확인하였다.

생성된 결과를 평가하기 위해 정량적 평가도 진행하였다. 생성된 결과가 이미지의 위험 상황을 정확하게 표현한 경우와 그렇지 못한 경우의 비율을 측정하였고, 그렇지 못한 경우의 이유를 분석하였다. 표 2는 정량적 평가의 결과이다. 총 100장의 테스트세트의 대한 모델의 위험 요소 검출 성공률을 나타낸다.

표 2. 정량적 평가 결과

위험 상황 표현	
성공	90%
실패	10%

90%의 성능으로 위험 상황, 사고에 대한 캡션을 생성하였다. 이는 제안하는 방법이 위험 상황과 사고를 중심으로 캡션을 생성하며 객체 뿐만 아니라 행동과 공간 정보를 포함한 구체적인 묘사가 가능함을 보여준다. 다만, 위험 상황 표현에 실패한 경우는 자주 등장하지 않는 단어들이 포함된 경우 문장이 완벽하게 생성되지 않는 문제가 발생

(a1) **BLIP(기존)** : two boys in a parking

(a2) **제안방법** : a man in black clothes is punching a man in white clothes on the road.

(b1) **BLIP(기존)** : a man doing a trick on a car

(b2) **제안방법** : a man in pink clothes is flying after crashing into a gray car on the road.

(c1) **BLIP(기존)** : a woman is playing a video game on a tv

(c2) **제안방법** : a man wearing a black top is lying down next to a blue board in the office.

그림 2. 비교 실험 결과 예시

하였다. 단어의 빈도 수를 고려하여 더 많은 데이터 세트를 구축한다면 더욱 좋은 성능이 나올 것으로 예상된다.

4. 결론 및 향후 연구

본 논문에서는 이미지 캡셔닝 기술을 사용한 새로운 접근 방법의 감지 시스템을 제안하였다. 이미지 캡셔닝 기술을 감시 시스템에 효과적으로 적용하기 위해 객체의 특징 정보, 공간 정보, 행동 정보 간 관계를 유기적으로 해석하는 캡션을 생성하는 방법을 제안하였고, 이 방법에 따라 이미지와 캡션 쌍의 새로운 데이터 세트를 구축하였다.

제안하는 방식으로 구축한 데이터 세트는 기존의 벤치마크 데이터 세트가 이미지 내 큰 객체에 중점을 두고 설명하여 항상 사람 중심의 해석을 보장하지 않는다는 한계점과 같은 행동을 포함한 객체 정보가 공간 정보를 활용하지 않을 경우 위험 상황을 검출하지 못하는 단점을 해결하여 감시 시스템에 적용할 수 있는 캡션 데이터 세트의 기준을 제시하였다.

향후 연구에서는 객체 정보, 행동 정보, 공간 정보가 포함된 캡션을 이미지 정보와 조합하여 해당 장면의 위험도를 측정하는 방안에 대해 연구를 진행할 것이다.

참고문헌

[1] 국가안전시스템 개편 종합대책(2023). https://www.mois.go.kr/plan2023/sub_01_02.html (accessed March 27, 2023).

[2] 박준태, 한규필, 박양우, “인간 행동 분석을 이용한 위험 상황 인식 시스템 구현,” 멀티미디어학회논문지, 제24권, 3호, pp. 345-354, 2021.

[3] Khan, Sardar Waqar, et al. "Anomaly detection in traffic surveillance videos using deep learning," Sensors, vol.22, pp. 6563-6590, 2022.

[4] Zhang, Yihang, and Yunsick Sung, "Traffic Accident Detection Method Using Trajectory Tracking and Influence Maps," Mathematics, vol.11, pp.1743-1756, 2023.

[5] LIN, Tsung-Yi, et al. "Microsoft coco: Common objects in context," ECCV, vol.8693, pp. 740-755, 2014.

[6] Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," IEEE, vol.123, pp. 2641-2649, 2015.

[7] Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," PMLR, vol.162, pp. 12888-12900, 2022.

[8] 지하철 역사 내 CCTV 이상행동 영상(2022), https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=174 (accessed April 3, 2023).

[9] 이상행동 CCTV 영상(2023), https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=171 (accessed April 3, 2023).

[10] 실내(편의점, 매장) 사람 이상행동 데이터(2022), https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71550 (accessed April 3, 2023).

[11] NEW_Trimmed_compressed(Violence-Detection-Dataset)(2021), https://www.kaggle.com/datasets/engmohamedsshubber/new-trimmed-compressedviolencedetectiondataset (accessed April 3, 2023).

[12] I. Loshchilov, F. Hutter, "Decoupled weight decay regularization," International Conference on Learning Representations, arXiv preprint arXiv:1711.05101, 2018.