「정보시스템연구」 제32권 제4호
한국정보시스템학회
2023년 12월, pp. 119～136

http://dx.doi.org/10.5859/KAIS.2023.32.4.119

이미지 캡셔닝 기반의 새로운 위험도 측정 모델*

전민성** · 고재필*** · 최경주****

〈목 차〉

Ⅰ. 서 론

최근 컴퓨터비전 영역의 딥러닝 기술의 비약적인 발전으로 객체 탐지 및 추적 등의 성능이 크게 향상되었으며, 지능형 CCTV 시스템이 차세대 감시 기술로 인식됨에 따라 그 수요가 증가하고 있다. 행정안전부의 “국가안전시스템 개편 종합대책”(2023)에 따르면 2027년까지 전국 CCTV 53,000대를 모두 지능형 CCTV로 전면 교체할 예정이며, 현재 13,000대가 교체 완료되었다고 한다. 이러한 사회적 요구에 따라 국민의 안전 문제를 해결하기 위한 수단으로 이상 행위 및 보안 위협 가능성을 탐지하려는 노력이 활발해지고 있다.

감시 시스템은 위험한 행동, 비정상적인 행동, 사고와 같은 패턴을 탐지하도록 설계되었다. 이러한 감시 시스템은 주로 객체 탐지 및 인식, 추적, 포즈 추정, 움직임 탐지 등을 통해 및 ‘싸움’, ‘실신’, ‘교통사고’ 등과 같은 이상치를 탐지하는 방향으로 연구되어 왔다(Alairaji et al., 2021; Chang et al., 2022; Jha et al., 2021; Perez et al., 2019; Sultani et al., 2018; Wu et al., 2020). 이러한 다양한 연구들은 이상 행동과 관련된 다양한 특징 정보를 추출하여 사용함으로써 지속적으로 성능을 향상시켜왔다. 하지만, 기존의 접근법들은 다음과 같은 제약들로 인해 어려움을 겪고 있다. 첫째, 기존의 접근법은 변칙적인 상황과 관련된 공간적 맥락

* 이 논문은 2022학년도 충북대학교 학술연구영역 사업의 연구비 지원에 의하여 연구 되었음(This work was supported by a funding for the academic research program of Chungbuk National University in 2022).
** 충북대학교 대학원 컴퓨터과학전공 박사과정, mywjsalstjd@naver.com(주저자)
*** 금오공과대학교 컴퓨터공학과 교수, nonzero@kumoh.ac.kr
**** 충북대학교 전자정보대학 소프트웨어학부 교수, kjcheoi@chungbuk.ac.kr(교신저자)

에 대한 포괄적인 고려를 제공하기에 매우 부족하다. 동일한 행동이라도 공간적 맥락에 따라 해석이 달라질 수 있다. 예를 들어, 공원에서 뛰는 행동과 도로를 가로질러 뛰는 행동은 모두 같은 '뛰는' 행동이지만, 위험의 정도는 다르다. 공원에서 뛰는 행동은 안전하다고 간주되는 반면, 도로를 가로질러 뛰는 행동은 무단횡단으로 위험한 상황을 나타낸다. 둘째, 감시 시스템은 특정 장소나 특정 행동이 아닌 불특정한 유형의 위험 요소와 사고를 검출하고 상황을 해석할 수 있어야 한다. 예를 들어, 수영장의 익사 사고를 감지하도록 설계된 시스템은 수영장의 물과, 물에 빠진 사람의 행동 패턴의 학습을 통해 객체 검출 기술이나 움직임에 따른 이상치 감지 등을 이용하여 충분히 위험 상황을 검출할 수 있다. 그러나 공공장소나 거리에 설치되는 CCTV는 특정 행동, 특정 공간이 아닌 다목적 감지 능력이 필요하다. 안전에 관련된 위험 요소는 다양한 환경에서 발생하며, 다양한 행위를 포함하기 때문이다. 결국 감시 시스템이 위험도에 대한 해석의 정확성과 신속성을 높이기 위해서는 객체의 행동뿐만 아니라 주변 환경까지 고려해야 하며, 이를 기반으로 전반적인 위험수준을 평가하여 감시자에게 구체적인 해석 정보를 제공해야 할 필요가 있다.

사람은 수준 높은 표현을 할 수 있는 언어라는 정의된 구조를 사용한다. 자연어를 사용하면 장면에 포함된 위험 상황을 객체에 대한 정보, 행동 정보, 행동이 발생한 장소, 근처의 위험 요소 등 다양한 정보를 포함한 문장 형태로 표현할 수 있다. 이미지 캡셔닝(image captioning)은 자연어로 영상의 내용을 기술하는 텍스트를 생성하는 자연어 처리 분야이다. 일반화 성능이 높아진 대규모 언어 모델을 활용한 이미지 캡셔닝 기술을 통해 객체의 특성, 행동 및 공간적 맥락에 대한 정보를 포괄하여 장면을 문장으로 상세하게 표현할 수 있게 된다면, 위에 제기된 문제들을 해결할 수 있다.

본 논문에서는 객체 중심의 행동 분석에 중점을 둔 기존의 감시 시스템이 보여준 근본적인 한계를 뛰어넘도록 설계된 새로운 접근 방법의 감시 시스템을 소개한다. 제안하는 시스템은 감시 대상 장면에 대해 객체의 속성, 행동 및 공간적 맥락을 포함하는 포괄적인 정보를 사용하여 구체적으로 장면을 설명하는 캡션을 생성하고, 이러한 캡션은 이후 관찰된 장면의 위험 수준을 평가하는 데 사용된다. 본 논문의 전체적인 구성은 다음과 같다. 다음 2장에서는 대규모 멀티모달 모델과 이미지 캡셔닝을 적용한 기존의 감시 시스템 연구에 대해 간단히 설명한다. 이어 3장에서는 새롭게 구축한 데이터세트에 대한 설명과 전체 시스템 구조에 대해 자세하게 설명한다. 4장에서는 제안하는 시스템의 성능 평가를 위한 실험과 결과를 제시하고, 결과 분석을 통해 제안하는 시스템의 발전 가능성을 논한다. 마지막으로 5장에서는 제안하는 시스템에 대한 요약과 발전 가능 방향에 대해 기술하면서 결론을 맺는다.

Ⅱ. 이론적 배경

2.1 대규모 멀티모달 모델과 감시 시스템

대규모 언어 모델(LLM; large language model)은 방대한 양의 데이터로 사전 학습된

트랜스포머(Vaswani et al., 2017) 기반의 초대형 딥러닝 모델로, 수많은 파라미터를 보유한 인공 신경망으로 구성되어 있다. 이러한 LLM의 등장은 자연어 처리(NLP) 분야의 상당한 발전을 가져왔으며, 단어 유사성, 맥락적 관계를 구별하고 문장 구조, 문법 및 의미를 효과적으로 처리함으로써 인간 언어를 이해하고 생성하는 능력을 향상시켰다. 2018년 구글(google)이 개발한 BERT(bidirectional encoder representations from transformers)(Devlin et al, 2019)는 양방향 언어모델을 사용하여 문장 내의 모든 단어를 고려하여 문맥을 이해하고 학습하기 때문에 문장 내 단어의 의미를 파악하는데 뛰어나고, 언어에 대한 풍부한 정보를 확보할 수 있다. BERT는 양방향 문맥을 포착하므로, 개체 인식(entity recognition), 감성 분석(sentiment analysis)과 같은 작업에서 뛰어난 성능을 보인다. OpenAI가 개발한 GPT (generative pre-trained transformer) (Radford et al., 2018)는 단방향 또는 '왼쪽에서 오른쪽으로' 문맥을 학습하며 이전 단어를 기반으로 다음 단어를 예측하도록 학습된다. GPT는 문장 생성(generative task)에 적합하여 텍스트 생성, 요약, 기계 번역 등의 작업에 사용된다. GPT는 GPT-2, GPT-3로 진화되면서 점점 더 많은 양의 데이터로 학습되어 인간과 유사한 텍스트를 만들어내는 능력을 보여주었다.

그러나, 이러한 LLM은 학습 과정에 이미지 데이터가 없기 때문에 이미지를 이해할 수 없다는 제약이 있었다. 이러한 문제를 해결하기 위해 등장한 모델이 대규모 멀티모달 모델(LMM; large multi-modal model)이다. 대규모 멀티모달 모델은 텍스트와 이미지를 통합적으로 이해하고 처리할 수 있는 대규모 언어 모델로, 방대한 양의 텍스트, 이미지, 오디오 등의 다양한 형태의 데이터를 모두 학습하여 이미지와 텍스트 사이의 경계를 지우고 이미지와 텍스트를 통합적으로 이해하고 처리할 수 있다. LLM이 주로 텍스트 데이터를 기반으로 언어를 이해하고 생성하는 데 사용되는 반면, LMM은 텍스트와 이미지 사이의 관계를 파악하거나, 이미지를 설명하는 텍스트를 생성하거나, 또는 텍스트를 입력받아 이미지를 생성하는 등의 작업을 수행할 수 있다. 여기서, 입력된 이미지에 대한 설명을 생성하는 기술을 이미지 캡셔닝(image captioning)이라고 한다. 2023년 6월 발표된 BLIP-2(Li et. al., 2023)는 웹에서 수집한 대규모의 이미지-텍스트 쌍으로 이루어진 데이터 세트를 학습한 멀티모달 모델이다. 이 모델은 인코더와 디코더 모두에 동결 사전학습 모델(frozen pretrained model)을 도입하고, 쿼리 트랜스포머(q-former)를 통해 인코더와 LLM과의 다중 모달 차이(modality gap)을 해결하여 다양한 비전-자연어 작업(vision-language task)에서 SOTA(state of the art) 성능을 달성하였다. 2023년 3월 발표된 OpenAI가 개발한 GPT 시리즈 4번째 모델인 GPT-4는 문자, 음성, 이미지를 생성하는 '멀티모달' 언어모델로, 전작들과 달리 GPT-4 (OpenAI, 2023)는 문자 외에도 이미지를 입력으로 받을 수 있다.

2.2 선행 연구

Dilawari 등(2021)과 Chen 등(2023)은 이미지 캡셔닝 기술을 적용한 감시 시스템을 개발하였다. Dilawari 등은 VGG-16 모델(Simonyan

& Zisserman, 2015)을 통해 영상의 특정 상황에 대한 시각 정보를 추출하고 양방향(bidirectional)-LSTM(Graves et al., 2005)을 활용하여 캡션을 생성하는 시스템을 제안하였는데, 이 시스템은 사람의 연령, 의상, 표정과 같은 객체 정보 위주의 캡션을 생성할 수 있도록 훈련되었다. Chen 등(2023)은 SwinBERT (Lin et al., 2022)를 통해 생성한 캡션과 ResNet-50 (He et al., 2016)을 통해 추출한 영상의 특징을 조합하여 이상치 점수(anomaly score)를 계산하는 시스템을 제안하였다. 이러한 연구들에서 이상 행동 감지에 주로 사용된 데이터 세트는 UCF Crime(Sultani et al., 2018), NTU CCTV-Fights(Perez et al., 2019), XD-Violence(Wu et al., 2020) 등으로 '싸움', '기절', '배회', '유기'와 같은 다양한 행동들이 포함되어 있는 비디오 데이터들이다. 그런데 이러한 기존의 데이터 세트는 캡션 정보를 제공하지 않거나, 제공하는 캡션이 있더라도 객체에 대한 간단한 설명만 존재할 뿐 실제 감시 시스템에 적용할 만큼 객체의 행동을 묘사하거나 공간적 맥락을 고려한 캡션은 제공하지 않는다. 결국 이러한 연구들은 이미지 캡셔닝을 감시 시스템에 적용하였다 하더라도, 생성되는 캡션이 객체 중심의 해석이기 때문에 공간 정보에 대한 정보를 활용하지 못한다는 제한이 있어 생성된 캡션만으로 장면에 대한 위험도를 판단하기 매우 어렵다.

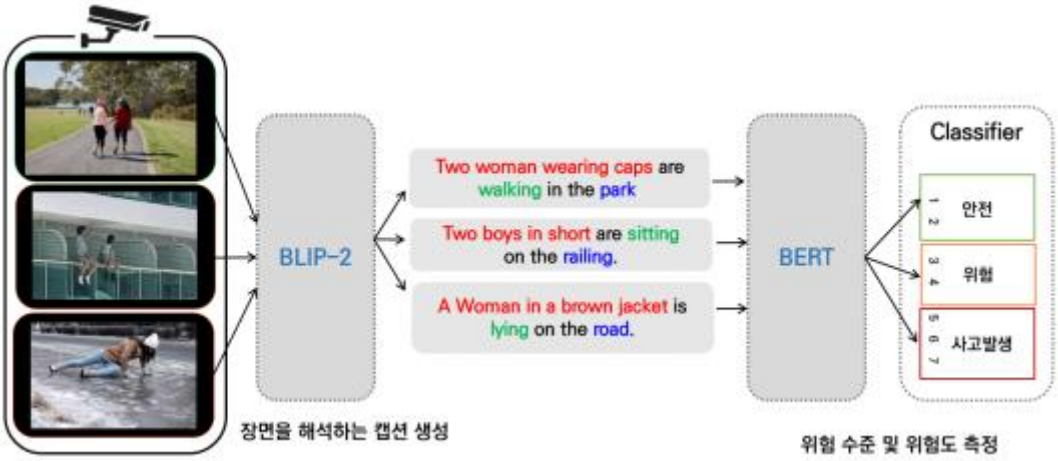
본 논문에서는 기존의 감시 시스템이 가지는 근본적인 제약 조건을 해결하기 위해 혁신적인 감시 시스템을 제안한다. 제안하는 시스템이 장면을 해석할 때 객체, 행동, 공간에 대한 모든 정보를 포함하여 구체적으로 장면을 해석한 캡션을 생성하도록 하기 위해 감시 시스템에 적절한 새로운 데이터 세트를 수집하고 캡션을 생성하는 문법을 정의하여 제안한다. 제안하는 시스템은 대규모 웹 데이터로 학습한 BLIP-2를 자체 구축한 데이터 세트로 미세 조정하여 입력 장면의 상황을 정확히 해석한 캡션을 생성하고, BERT를 통해 생성된 캡션의 의미를 해석하여 이를 기반으로 위험도를 측정한다.

Ⅲ. 이미지 캡셔닝 기반의 새로운 감시 시스템

<그림 1>은 제안하는 시스템의 전체적인 흐름도를 보여준다. CCTV로부터 입력받은 장면은 대규모 멀티모달 모델인 BLIP-2를 사용하여 입력된 장면의 내용을 설명하기 위한 캡션을 생성한다. 이러한 캡션은 객체 속성, 행동 및 공간 정보를 포괄하는 다양한 세부 정보를 포함한다. 이렇게 생성된 캡션을 대규모 언어 모델인 BERT를 사용하여 캡션에 내장된 정보를 심층 분석을 수행하여 장면의 위험 수준을 측정하게 된다.

3.1 감시 시스템에 적절한 새로운 형태의 캡션 문장 구조 형식

제안하는 시스템의 핵심인, 감시 대상 장면을 구체적으로 해석하는 캡션을 생성해내려면 캡션이 달린 데이터 세트가 필요하다. 여기서 필요한 캡션은 일반적인 캡션 형태와는 다르게 객체의 속성, 행동, 공간적 맥락이 포함된 포괄적인 정보를 사용하여 장면을 구체적으로 해석한 캡션이어야 한다. 앞장에서 기술했듯이 일반



<그림 1> 시스템의 전체적인 흐름도

적인 감시 시스템에서 사용하는 학습용 데이터 세트는 해당 장면에 대한 해석이 담긴 캡션 정보를 제공하지 않거나, 제공하는 캡션이 있다 하더라도 객체의 대한 간단한 설명 외에 실제 감시 시스템에 적용할 만큼의 객체의 행동을 묘사하거나 주변의 공간을 고려한 캡션을 제공하지 않는다. 이에 본 논문에서는 감시 시스템을 위한 이미지 캡션의 새로운 문장 구조 형식을 제안한다. 감시 시스템에 필요한 캡션은 감시 시스템에 필요한 정보인 객체 정보, 행동 정보, 그리고 공간적인 맥락까지 모두 표현하여 작성되어져야 한다. 또한, 안전 사고의 감시 대상인 사람에 초점을 맞추어, 다른 객체보다 먼 거리에 있는 경우에도 사람을 감지하고 해석할 수 있도록 데이터 세트를 구축한다. 이는 단순히 시각적 요소를 나열하는 것을 넘어, 감시하고자 하는 상황을 정확하게 이해하고 분석할 수 있는 충분한 정보를 전달할 수 있게 된다.

<표 1>은 캡션을 어떻게 작성해야 하는지, 캡션 작성 규칙에 대한 예시를 보여준다. <표 1>의 캡션에서 빨간색으로 표시된 부분은 객체와 종류와 객체의 속성을 나타낸다. 객체의 종류는 '사람', '운송 수단', '화재'와 같은 것을 말하는데, 사람일 경우에는 성인인지 어린이인지 구분해야 한다. 일반적으로 성인보다 어린이일 경우 위험에 대한 가능성이 커지기 때문이다. 객체에 대한 종류 뿐 아니라, 인상착의 등과 같은 속성도 같이 표현한다. 초록색으로 표시된 부분은 객체의 행동을 나타낸다. <표 1>의 1번 이미지의 캡션에는 '검은 운동복을 입은 여성'이라는 객체에 대한 정보, '걷기'라는 행동 정보, 그리고 '공원'이라는 공간 정보가 모두 포함되어 있음을 확인할 수 있다. 마찬가지로 2번과 3번의 이미지 역시 객체, 행동, 공간에 대한 구체적인 정보가 모두 캡션에 포함되어 있다. 특히, 2번의 경우에는 크기가 큰 자동차를 중심 객체로 해석하는 것이 아닌 사람에 초점을 맞추어 설명하는 캡션을 확인할 수 있다.

<표 1> 감시 시스템에 적절한 캡션 문장 구조 형식 예

	이미지	캡션
1		**A woman wearing black track suit is walking in the park.**
2		**Four men and women are walking on the road.**
3		**A man with his top off is fainting on the grass.**

3.2 위험 수준 및 위험도 정의

보다 효율적인 감시 시스템을 만들기 위하여 제안하는 시스템은 위험도를 측정한다. 이를 위해 우선적으로 감시 대상이 되는 장면에 대한 위험수준과 위험 수준 별 위험 강도에 따른 위험도 점수를 구분하여 정의하였다.

<표 2>는 본 논문에서 정의한 위험 수준 및 위험도 분류 기준이다. 감시 대상이 되는 장면에 대한 위험 수준은 크게 '안전', '위험', '사고 발생'의 3개로 구분하였고, 각각의 위험 수준 별 위험 강도에 따라 다른 위험도 점수를 부여하였다. '안전'과 '위험'을 나타내는 위험 수준은 위험 강도에 따라 각각 2단계로 구분되는 데 비해, '사고 발생'을 나타내는 위험 수준은 위험 강도에 따라 3단계로 더 세분화된다. 위험 수준 '안전'은 일반적으로 보행로나 공원, 실내 등과 같은 안전한 공간이라 할 수 있는 장소에서 '걷기', '앉기', '뛰기' 등과 같은 일상적인 행동을 하는 경우에 해당된다. 그런데, 동일한 위험수준을 가지고 있더라도 성인이 아닌 어린이가 대상자일 경우에는 성인보다는 상대적으로 위험할 수 있다는 판단하에 위험도 점수가 높은 높게 부여된다. <표 2>의 위험도 2가 이 사례에 해당되는데, 위험 수준이 '안전'이라 하더라도 성인이 행동했을 경우 1점의 위험도 점수를 부여하고, 어린이의 경우에는 성인보다는 조금 더 위험할 수 있다는 의미에서 2점의 위험도 점수를 부여한다. 위험 수준 '위험'은 '안전' 상황에서도 일어나는 일상적인 행동을 하고 있지만, 행동이 일어나는 공간 상황에 따라 '위험' 상황으로 판단되는 경우에 해당된다. 예를 들어 똑같이 앉아 있는 행동이라 하더라도 벤치에 앉아 있는 경우에는 안전한 상황이라 판단하고 1점의 위험도 점수를 부여하지만, 난간, 다리, 지붕과 같이 위험한 장소에 앉아 있는 경우에는 위험한 상황이라 판단하고 3점의 위험도 점수를 부여한다. 위의 <표 1>의 1번과 2번의 이

미지에는 성인의 ‘걷기’ 행동이 들어있는데, 1번은 ‘걷기’라는 행동이 ‘공원’에 위치하고, 2번은 차가 많은 ‘도로’에 위치한다. 따라서 1번은 ‘안전’ 상황으로 1점의 위험도 점수를 부여하고, 2번은 ‘위험’ 상황으로 3점의 위험도 점수를 부여한다. 여기서 만일 성인이 아닌 어린이가 공원을 걷고 있다면 ‘안전’ 상황으로 위험도 점수 2를 부여하고, 어린이가 무단횡단한다고 하면 ‘위험’으로 성인보다 높은 4의 위험도 점수를 부여하게 된다. 위험 수준 ‘사고발생’은 검출된 행동이 사고로 분류되는 경우로, ‘실신’, ‘싸움’, ‘화재’, ‘교통사고’ 등과 같은 행위가 포함된 경우에 해당된다. <표 1>의 3번은 검출된 행동이 ‘실신’으로 긴급한 상황이기에 위험 수준은 ‘사고발생’으로 분류되고, ‘공원’에 쓰러져 있기 때문에 5점의 위험도 점수를 부여한다. 그런데 만일 행위가 발생한 공간이 난간, 절벽, 도로 위, 공사장과 같은 위험한 장소인 경우 위험도 점수가 더 높아지며, 객체가 어린이라면 이보다 더 높은 위험도 점수를 부여한다.

이러한 체계적인 위험 평가 접근법은 다양한 공간적 맥락에서 환경적 요인과 관련된 개인의 연령의 영향을 고려하여 위험의 정도를 보다 미묘하고 정확하게 평가할 수 있게 해준다.

3.3 데이터 세트의 구축

본 연구에서는 <표 2>의 각 위험도 별 분류 기준에 맞는 영상으로 ‘안전’ 상황에 해당하는 영상 557개, ‘위험’ 상황에 해당하는 영상 546개, ‘사고발생’ 상황에 해당하는 영상 1,616의 총 2,719개의 영상을 수집하였다. 수집한 영상에 대한 캡션은 객체의 종류를 중심으로 객체의 특징 정보를 설명하고 객체의 행동과 객체

<표 2> 위험 수준 및 위험도 분류 기준

위험 수준	위험도	분류 기준	객체	행동	공간
안전	1	영상 내의 객체가 안전한 장소에 있는 경우	성인	걷기, 앉기, 뛰기, 타기	보행로, 공원, 실내
	2		어린이	걷기, 앉기, 뛰기, 타기	보행로, 공원, 실내
위험	3	영상 내의 객체가 위험한 장소에 있는 경우	성인	걷기, 앉기, 뛰기, 타기	난간, 절벽, 도로 위, 공사장
	4		어린이	걷기, 앉기, 뛰기, 타기	난간, 절벽, 도로 위, 공사장
사고 발생	5	영상 내의 객체가 위험 행동을 하는 경우	성인 이동 수단 화재	쓰러짐, 싸움, 화재, 교통사고	보행로, 공원, 실내
	6	영상 내의 객체가 위험 행동을 하면서 위험한 공간에 있는 경우	성인 이동 수단 화재	쓰러짐, 싸움, 화재, 교통사고	난간, 절벽, 도로 위, 공사장
	7	영상 내의 객체가 어린이이고, 위험 행동을 하면서 위험한 장소에 있는 경우	어린이 이동 수단 화재	쓰러짐, 싸움, 화재 교통사고	난간, 절벽, 도로 위, 공사장

가 포함된 공간 정보를 캡션에서 설명할 수 있도록 <표 1>과 같은 규칙에 따라 작성하였고, 각 상황에 맞는 위험도 점수를 부여하여 [이미지-캡션-위험도점수]가 포함되어 있는 데이터 세트를 구축하였다.

3.4 감시 장면을 해석한 캡션 생성 및 위험도 측정

제안하는 시스템은 대규모 멀티모달 모델인 BLIP-2를 사용하여 장면을 해석하는 캡션을 생성하였고, 해석된 장면을 분석하기 위해 대규모 언어모델인 BERT를 사용하였다. 이러한 언어 모델들은 대규모의 데이터로 사전 학습되어 있어 다양한 자연어 처리 작업에 사용될 수 있지만, 특정 작업에 최적화되어 있지는 않기 때문에 이들을 감시 시스템에 적용하여 적절한 결과를 얻기 위해서는 반드시 미세 조정(fine-tuning) 과정을 거쳐야 한다. BLIP-2 모델을 자체 구축한 데이터 세트로 미세 조정하여 새로 정의된 문장 구조에 맞는 캡션을 출력하도록 유도하였고, 이후 대규모 언어 모델인 BERT를 사용하여 생성된 문장의 의미를 해석하여 장면에 대한 위험 수준 별 위험도를 1~7 단계로 분류하여 측정하였다.

<그림 1>의 가운데에 보이는 BLIP-2는 '이미지 인코더(image encoder)', '쿼리 트랜스포머(q-former)', '대규모 언어 모델(LMM)'의 3가지 구성 요소로 이루어져 있다. 쿼리 트랜스포머는 고정된 가중치의 이미지 인코더와 대규모 언어 모델 사이의 격차를 줄이기 위한 학습 가능한 모듈로 사용되며, 이미지 인코더는 ViT-G(Dosovitskiy et al., 2021)를, 대규모 언어 모델은 OPT 2.7B(Zhan et al, 2022)를 사용하였다. 이렇게 사전 학습된 모델을 기반으로, 구축된 데이터 세트로 BLIP-2를 미세 조정하였다. 대규모 멀티모달 모델을 미세 조정하는 경우 엄청난 비용이 발생하는데, 이를 해결하기 위해서 Hu가 제안한 LoRA(low-rank adaptation)를 적용하였다. LoRA는 기존 사전 학습된 가중치는 고정하고, 업데이트 해야 하는 매개 변수를 낮은 등급(low rank)으로 분해(decomposition)된 가중치를 학습한다(Hu et al., 2021). 이는 학습에 필요한 매개 변수를 줄여주는 효과와 동시에, 새로운 작업을 학습할 때 기존 작업에서 학습한 지식을 잊어버리는 현상 재앙성 망각(catastrophic forgetting) 현상을 방지할 수 있다. 결과적으로 자체 구축한 데이터 세트로 미세 조정된 BLIP-2를 통해 객체의 속성, 행동, 공간적 맥락이 포괄적으로 포함된 캡션을 생성한다.

이렇게 생성된 캡션은 <그림 1>의 우측에 보이는 BERT를 통해 생성된 캡션의 의미가 분석되고, 이후 분류기를 통해 장면에 대한 위험도가 측정된다. BERT는 사전 훈련 과정에서 단어를 수치 벡터로 표현한 양질의 임베딩(embedding)을 학습하고, 주어진 문장의 앞과 뒤 문맥을 모두 고려하기 위해 트랜스포머(transformer)의 인코더를 12층으로 쌓았다. [CLS] 토큰은 BERT에서 항상 시퀀스의 시작 부분에 위치하는 특수 토큰이다. [CLS] 토큰은 입력 시퀀스에 대한 전체 맥락을 나타내기 때문에 문서의 주제 또는 클래스를 나타내는 용도로 사용되므로 분류 작업에 사용된다. 제안하는 시스템에서 BLIP-2로부터 생성된 캡션은 BERT를 통해 [CLS] 토큰이 만들어지고, [CLS]

토큰으로 출력된 벡터는 선형 층(linear layer)을 거쳐 소프트맥스(softmax) 함수를 통해 장면의 위험 수준과 위험도를 분류하도록 하였다. 입력 데이터로부터 BERT를 통과하여 출력으로 얻은 [CLS] 토큰의 벡터를 h라고 정의하면, 선형 층의 출력 벡터 z는 아래의 식 (1)과 같이 정의된다.

$$z = W \bullet h + b \quad (1)$$

여기서 W는 선형 층의 가중치 행렬이며, b는 편향 벡터이다. 식 (2)는 선형 층으로부터 생성된 출력 벡터가 소프트맥스 함수를 통해 각 클래스에 대한 확률 분포를 계산하는 수식이다. e는 자연상수를 나타내며 제안하는 시스템은 위험도의 단계가 1~7이기 때문에 1부터 7까지 누적한 것을 볼 수 있다.

$$Softmax(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{7} e^{z_j}} \quad (2)$$

Ⅳ. 실험 및 결과

4.1 실험 환경

BLIP-2는 224×224의 입력 영상 크기로 구성하였으며, 16개의 고정된 배치 사이즈를 가진다. Adam(Kingma et al., 2015) 옵티마이저를 사용하여 최적화를 수행하였고, 미세 조정은 50 에포크(epoch)에 걸쳐 진행하였다. 학습률은 10^{-5}로 설정하였다. 내용이 풍부한 문장을 생성해야 하기 때문에 모델의 최대 출력 길이는 30으로 설정하였다. BERT의 경우 학습률은 10^{-5}로 설정하였고, 배치 사이즈는 32이다. BERT는 100 에포크에 걸쳐 미세 조정을 진행하였다. 실험 데이터 세트는 총 2,719개로, 훈련 데이터, 검증 데이터, 테스트 데이터를 각각 8 : 1 : 1의 비율로 분할하여 실험을 진행하였으며, 모든 실험은 단일 A100 GPU에서 수행되었다.

4.2 실험 환경

4.2.1 장면을 해석하는 캡션의 생성과 위험도 측정

제안하는 시스템이 감시 장면에 대해 객체의 속성, 동작 및 공간적 맥락을 포함하면서, 지정된 구조 형식을 준수하는 캡션을 생성하고 있는지 확인하기 위하여 기존의 이미지 캡셔닝 모델이 만들어 낸 캡션과 비교하는 실험을 수행하였다. 비교 대상 모델로는 트랜스포머(transformer)를 기반으로 연구하기 시작한 이후 성능이 높아진 모델인 OFA(Wang et al., 2022)와 LAION 데이터 세트(Schuhmann et al., 2022)로 미세 조정한 모델 BLIP-2(base)와 비교하였다. OFA는 COCO(Lin et al., 2014) 데이터 세트 외 4가지 다양한 데이터 세트로 미세 조정한 모델이다.

<표 3>은 위험도 점수가 1~7에 해당하는 이미지를 대상으로 실험했을 때 기존의 이미지 캡셔닝 모델과 제안하는 시스템에서 만들어진 캡션과 이를 기반으로 위험도를 측정한 결과이다. 기존의 이미지 캡셔닝 모델은 장면에 대한 캡션만 생성할 뿐 위험도 판단을 하지 않기 때문에 위험도 점수의 결과에 X로 표시하였다.

<표 3> 각 위험도 케이스 별 생성된 캡션과 위험도 측정 결과

	이미지	생성된 캡션		위험도 점수	
				정답	결과
(a)		OFA	three people walking down a path in a park	1	X
		BLIP-2 (base)	a man and woman walking down a path		X
		Ours	a man and two women in red are walking in the park		1
(b)		OFA	two women and a child walking down a road with a stroller	2	X
		BLIP-2 (base)	a woman and two children walking down a path		X
		Ours	two women and a boy are with strollers walking in the park.		2
(c)		OFA	two people crossing a street in front of a school bus	3	X
		BLIP-2 (base)	a school bus is stopped on the side of the road		X
		Ours	two men are running on the road next to a gray car		3
(d)		OFA	a young boy is running in a parking lot	4	X
		BLIP-2 (base)	a little boy is walking across the street		X
		Ours	a boy wearing blue shirt is walking on the road		4
(e)		OFA	two girls laying on the floor in a kitchen	5	X
		BLIP-2 (base)	a woman in a kitchen with a cat on the floor		X
		Ours	a woman wearing black shorts fell on the floor		5
(f)		OFA	a man is kneeling on the ground in front of a window	6	X
		BLIP-2 (base)	a man is seen in the window of his house		X
		Ours	a man wearing a jeans collapsed on the snow.		6
(g)		OFA	a child is playing on a swing at a playground	7	X
		BLIP-2 (base)	a todd is playing on a swing in the playground		X
		Ours	a boy wearing blue T-shirt is lying on the ground		7

<표 3(a)>와 <표 3(b)>에서 보여지는 이미지들은 모두 '안전' 상황에 해당하는 영상으로 사람들이 공원에서 뛰거나 걷고 있는 경우이다. 사람이 걷고 있다는 캡션은 3개의 모델 모두 잘 생성했으나, 캡션의 구체성을 보면 결과가 다르다. <표 3(a)>의 경우 2명의 여성과 남성이라는 객체의 종류, 빨간 옷이라는 객체의 속성까지 가장 잘 표현한 캡션을 생성한 것은 제안하는 시스템이다. <표 3(b)>의 경우에도 제안하는 시스템만이 공원에서 2명의 여성과 1명의 어린이, 그리고 유모차라는 정보를 캡션에 녹여내고 있다. OFA는 '도로'라는 공간 정보를 표현했는데, '도로'는 차가 다닐 수 있는 곳이기 때문에 '위험' 상황으로 판단할 수도 있는 공간이다. 하지만 실제 이미지는 차가 다니지 않는 공원 산책로이며, 이를 제안하는 시스템이 잘 해석하여 캡션으로 생성했다. 제안하는 시스템은 <표 3(a)>~<표 3(b)>의 2가지 상황 모두 '안전' 상황으로 잘 측정하였는데, <표 3(b)>은 어린이가 장면에 포함되어 있기 때문에 <표 3(a)>에 비해 1단계 높은 2점의 위험도 점수가 부여되었다.

<표 3(c)>와 <표 3(d)>에서 보여지는 이미지는 무단횡단을 하는 장면이다. <표 3(c)>의 경우 감시 카메라에서 가장 중요한 부분이 사람인데, BLIP-2(base)는 사람이 아닌 '스쿨버스'를 중심으로 캡션을 생성하여 사람이 무단횡단하고 있다는 것을 해석하지 못했다. 제안하는 시스템은 2명의 남자가 도로를 뛰고 있다고 정확히 캡션을 생성해 냈고, 옆에 회색 자동가 있다는 것도 잘 해석하였다. 뛰는 행동은 '안전' 상황에서도 발생할 수 있는 행동이지만, <표 3(c)>는 안전한 공간이 아닌 차가 다니는 '도로'에서 이루어졌기 때문에 '위험' 상황으로 측정되었고, 위험도 점수는 3점 부여되었다. <표 3(d)>의 경우 제안하는 시스템과 BLIP-2(base)가 '도로' 위를 걷고 있는 어린이로 올바르게 캡션으로 생성하였다. 제안하는 시스템은 4점의 위험도 점수를 부여했는데, 행동의 대상자가 어린이기 때문에 성인보다 위험도 점수가 1단계 높아졌다. 이는 제안하는 시스템이 다양한 상황에서 시스템이 정확한 위험 평가를 수행할 수 있는 능력을 강조하면서, 유사하게 보이지만 다양한 위험 수준을 수반하는 행동을 구별하는 시스템의 숙련도를 보여준다.

<표 3(e)>~<표 3(g)>에서 보여지는 이미지는 '사고 발생' 상황에 해당하는 장면으로, 제안하는 시스템에서 쓰러진 행동은 5점 이상의 위험도 점수를 가지며, 행동이 발생한 공간 정보와, 성인과 어린이의 구분을 통해 위험도가 달라진다. <표 3(e)>의 이미지는 건물 실내 바닥에 사람이 쓰러져 있는 장면이다. 제안하는 시스템만이 사고를 가장 잘 표현한 캡션을 생성했음을 확인할 수 있으며, '실내 바닥'이라는 공간은 추가로 사고가 날 위험이 없는 곳이기 때문에, 5점의 위험도 점수가 부여되었다. <표 3(f)>의 이미지에도 동일한 쓰러진 행동이 포함되어 있지만, 행동이 일어난 장소가 실외의 '눈길'이다. 제안하는 시스템만이 사고를 정확히 설명하는 캡션을 생성했음을 확인할 수 있으며, '눈길'이라는 공간이 '실내 바닥'보다 더 위험하다고 판단하여 6점의 위험도 점수가 부여되었다. 표 3(e)의 이미지는 어린이가 놀이기구에서 떨어져 쓰러져있는 장면이다. 역시 제안하는 시스템만이 사고를 정확히 설명하는 캡션을 생성하였으며 어린이라는 객체 정보, 쓰러진 행

<표 4> 테스트 데이터세트에 대한 위험 수준 별 위험도 측정 정확도

분류결과 / 정답	안전	위험	사고 발생	총 영상 개수	정확도
안전	48	3	1	52	92.3%
위험	2	53	4	59	89.8%
사고 발생	1	5	154	160	96.2%
				271	94.0%

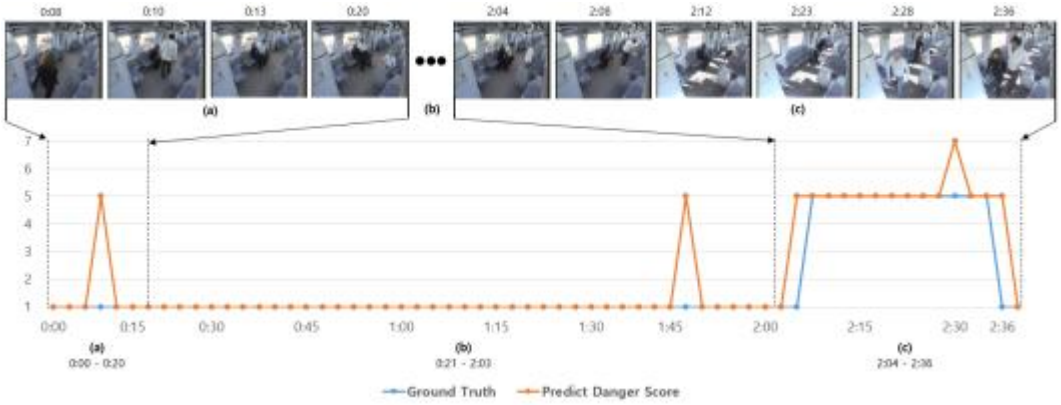
동, 그리고 이 행동이 일어난 공간인 야외를 구분하여 7점의 위험도 점수가 부여되었다.

<표 4>는 위험도 측정 결과에 대한 정확도를 위험 수준 별로 구분하여 나타낸 것이다. 제안하는 시스템은 테스트 데이터 271개에 대하여 안전, 위험, 사고 발생 3가지 위험 수준 별 각각 92.3%, 89.8%, 96.2%의 정확도를 보인 결과를 보여주었다. 이러한 결과는 제안하는 데이터 세트가 위험도를 측정하기 매우 용이한 정보를 제공한다는 것을 확인해주고 있다. 모든 위험 범주에 걸쳐 달성된 높은 정확도는 시스템이 견고하다는 것을 나타내며, 이는 실제 응용 분야에서 신뢰성을 준다.

4.2.2 비디오에서의 위험도 변화 측정

<그림 2>는 다양한 이상 행동이 포함된 비디오에 대하여 실제 정답과 제안 시스템이 측정한 위험도 점수의 결과를 그래프로 나타낸 것이다. 제안 시스템은 이미지를 입력으로 받아들이는 이미지 캡셔닝 기술을 활용한다. 비디오도 여러 개의 이미지로 구성되어 있으므로 비디오에서는 어떤 결과가 나올지 확인하기 위하여 비디오에서의 위험도 변화를 측정하였다.

<그림 2>의 파란색 선은 실제 위험도 점수의 정답(ground truth)을 나타내며, 주황색 선은 제안 시스템이 측정한 위험도 점수를 나타낸다. 테스트 데이터로는 BOSS 데이터 세트(Duan et al., 2022)에 포함된 비디오 영상을 사용하였다. BOSS 데이터 세트는 자세, 동작 및 상호 작용이 포함된 영상으로 '서다', '앉다', '걷다', '싸우다'와 같은 객체의 행동을 감지하는 비디오 감시 시스템 개발을 위해 만들어진 데이터 세트이다. 이 테스트 비디오는 기차 안에서 촬영된 비디오로, 총 2분 36초의 영상이다. 실험을 위해 초당 6 프레임의 이미지 추출하여 총 933개의 이미지에 대해 실험하였다. 그림 2(a)의 00:00~00:20 구간에는 3명의 사람들이 순서대로 걸어서 입장하고, 의자에 앉는 행동이 포함되어 있다. 이어 00:21~02:03 구간에는 사람들이 의자에 앉은 상태로 서로 이야기를 주고 받는다. 02:04~02:33 구간에는 두 남성이 '싸움'을 하는 장면이 포함되어 있으며, 사람들이 영상 밖으로 나가면서 끝이 난다. 실제 정답(ground truth)과 비교했을 때, 비디오임에도 불구하고 대부분의 상황에서 정답과 일치하는 결과를 보였다. 특히 02:04~02:33 구간의 싸움하는 장면에서 두 남성의 '싸움' 장면을 올바르게 해석하여 위험도를 정확하게 측정하였다. 실험에서 위험도를 제대로 측정하지 못한 부분은 세 부분이다. 제안하는 시스템은 <그림 2(a)>의 00:10 구간에서 11개의 이미지 프레임 속의 남자가 큰 동작으로 걷는 모습을 발차기하는 모



<그림 2> 폭력이 포함되어 있는 비디오 데이터에 대한 위험도 변화 측정 결과
(a) 사람들이 입장하는 구간 (b) 사람들이 의자에 앉는 구간 (c) 사람들이 싸우고 있는 구간

습으로 오해석하여 위험도를 '5'로 측정하였고, >그림 2(b)>의 00:13 구간에서 나란히 의자에 앉아 있는 남성과 여성이 서로 마주 보며 가까이 붙어있는 찰나의 순간을 두 사람이 싸움을 하는 장면으로 오해석하여 15개의 이미지 프레임의 위험도를 '5'로 측정하였다. <그림 2(c)>의 02:28 구간의 사람이 쓰러진 장면에서는 사람이 쓰러지면서 신체의 일부가 가려져서 원래 성인을 남자 어린이로 캡션을 잘못 생성하여 6개의 이미지 프레임의 위험도를 '7'로 측정하였다.

이렇게 시스템이 위험도를 오측정하는 경우는 이미지에 보여지는 객체의 크기가 너무 작거나, 학습되지 않은 단어가 포함된 경우이다. 이러한 오류는 제안 시스템이 다양한 환경에서 더 견고하게 작동할 수 있도록 다양한 크기의 객체와 텍스트를 포함하는 추가 데이터를 수집하여 훈련시켜 많은 상황에 대한 캡션을 생성하게 하면 해결할 수 있다. 그러나, 단일 프레임 수준에서 검출할 수 있는 행동 패턴의 범위는 한계가 있다. 따라서 상황의 전후를 상황을 고려하도록 기존의 시스템을 비디오 데이터에 대해 적용하여 확장하면, 위와 같은 오류가 발생하는 문제를 해결할 수 있을 것이다.

V. 결 론

인공지능 기술이 점점 발전하고 하드웨어의 성능이 향상됨에 따라 감시 시스템은 더욱 많은 일을 해낼 수 있게 되었다. 제안하는 감시 시스템은 위험 상황을 판단함에 있어 시스템 스스로 폭넓은 정보를 해석하고 구체적으로 상황을 분석하여 정보를 제공하는 새로운 접근 방식을 채택하였다. 제안된 시스템은 대규모 멀티모달 모델을 사용하여 위험 상황에 대한 캡션을 생성해내고, 생성된 캡션의 의미 분석을 통해 위험도를 측정하는 새로운 형태의 감시 시스템이다. 감시하고자 하는 다양한 상황을 객체 정보, 행동 정보, 공간 정보를 이용하여 포괄적인 캡션을 생성하고, 생성된 캡션의 다양한

내부 정보를 분석하여 각기 다른 위험도를 가지도록 자체 데이터 세트를 구축하였다. 다양한 실험을 통해 자체 구축한 데이터 세트가 위험도를 측정하기에 용이한 정보를 제공함을 확인하였다. 생성된 캡션은 기존의 다른 데이터 세트로 사전 학습된 모델들에 비해 감시 시스템에서 필요한 객체, 행동, 공간 정보를 충분하게 표현하였으며, 새로운 문장 형태에 적응하여 새로운 문맥에서도 유연하고 작동할 수 있음을 보여주었다. 향후 연구로 제안하는 시스템이 다양한 환경에 적용할 수 있도록 더 많은 행동과 공간 정보를 고려한 데이터를 추가한다면 제안하는 감시 시스템의 성능은 더욱 향상되고 많은 장소에서도 사용할 수 있는 범용적인 시스템으로의 발전이 가능해질 것이다. 또한 VQA (vision question answering) 기술의 적용을 고려할 수도 있다. VQA는 시각적 데이터에서 질문에 대한 답변을 제공하는 기술로, 감시 시스템에 통합될 경우 효율성을 높일 수 있다. 감시자는 질문을 설정하고, 시스템은 영상 데이터를 분석하여 빠르게 직접적인 답변을 제공함으로써 실시간 감시 및 분석의 효율성을 높여 대규모 감시 시스템에서 유용하게 사용될 수 있다. 현재 연구의 한계점으로 단일 프레임 수준에서 특정 순간의 이미지에서만 정보를 얻기 때문에, 상황의 전체적인 맥락과 연속성을 포착하는 데에는 부족함이 있다. 예를 들어 단순한 걷기로 보이는 행동도, 연속된 프레임에서는 빠르게 달리는 동작일 수 있다. 이를 해결하기 위해서는 제안하는 시스템을 시간적 차원으로 확장하여 상황의 전후 관계를 고려할 수 있도록 비디오 캡셔닝으로 확장한 연구가 필요하다.

참고문헌

행정안전부, “국가안전시스템 개편 종합대책”, 2023.1.27.

Alairaji, R. A., Aljazaery, I. A., ALRikabi, H. S., “Abnormal behavior detection of students in the examination hall from surveillance videos,” In Advanced Computational Paradigms and Hybrid Intelligent Computing, 2021, pp.113-125.

Chang, C. W., Chang, C. Y., and Lin, Y. Y., “A hybrid CNN and LSTM-based deep learning model for abnormal behavior detection,” *Multimedia Tools and Applications,* Vol. 81, No. 9, 2022, pp.11825-11843.

Chen, W., Ma, K. T., Yew, Z. J., Hur, M., and Khoo, D. A., “TEVAD: Improved video anomaly detection with captions,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2023, pp. 5548-5558.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., “Bert: Pre-training of deep bidirectional transformers for language understanding,” *Proceedings of NAACL-HLT,* 2019, pp.4171-4186.

Dilawari, A., Khan, M. U. G., Al-Otaibi, Y. D., Rehman, Z. U., Rahman, A. U., and Nam, Y. “Natural language description of videos for smart surveillance,” *Applied Sciences,* Vol. 11, No. 9, 2021,

pp.3730-3741.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., and Unterthiner, T, "Transformers for image recognition at scale," *The International Conference on Learning Representations,* 2021, arXiv:2010.11929.

Duan, J., Yu, S., Tan, N., Yi, L., and Tan, C, "BOSS: A Benchmark for Human Belief Prediction in Object-context Scenarios," 2022, arXiv:2206.10665.

Graves, A., Fernández, S., and Schmidhuber, J., "Bidirectional LSTM networks for improved phoneme classification and recognition," *International conference on artificial neural networks,* 2005, pp. 799-804.

He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," *IEEE conference on computer vision and pattern recognition,* 2016, pp. 770-778.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wand, L., Chen, W., "LoRA : Low-rank adaptation of large language models," 2021, arXiv: 2106. 09685v2.

Jha, S., Seo, C., Yang, E., and Joshi, G. P., "Real time object detection and tracking system for video surveillance system," *Multimedia Tools and Applications,* Vol. 80, 2021, pp.3981-3996.

Kingma, D. P., and Ba, J., "Adam: A method for stochastic optimization," *International Conference on Learning Representations,* 2015, arXiv:1412. 6980.

Li, J., Li, D., Savarese, S., and Hoi, S., "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, 10.48550/arXiv.2301.12597.

Lin, K., Li, L., Lin, C. C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., and Wang, L., "Swinbert: End-to-end transformers with sparse attention for video captioning," *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2022, pp. 17949-17958.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P., "Microsoft coco: Common objects in context," *European Conference on Computer Vision,* 2014, pp. 740-755.

OpenAI. Gpt-4 technical report, 2023.

Perez, M., Kot, A. C., and Rocha, A., "Detection of real-world fights in surveillance videos," *IEEE International Conference on Acoustics, Speech and Signal Processing,* 2019, pp. 2662-2666.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I., "Improving language understanding by generative pre-

training," 2018.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J., "LAION-5B: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems,* Vol. 35, 2022, pp.25278-25294.

Simonyan, K., and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations,* 2015, pp. 1-14.

Sultani, W., Chen, C., and Shah, M., "Real-world anomaly detection in surveillance videos," *IEEE conference on computer vision and pattern recognition,* 2018. pp. 6479-6488.

Vaswani, A., Shazeer, N., Parmer N., Uszkoreit, J., Jones, L., Gomes, A.N., Kaiser, L., Polosukhin, L., "Attention is all you need," *Advances in neural information processing systems,* Vol. 30, 2017.

Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., and Yang, Z., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," *16th European Conference on Computer Vision,* 2020, pp. 322-339.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P, S., Sridhar, A., Wang, T., and Zettlemoyer, L., "OPT: Open pre-trained transformer language models," 2022, arXiv:2205.01068v4.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhouand, J., and Yang, H, "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," *International Conference on Machine Learning,* 2022, pp. 23318-23340.

전 민 성 (Jeon, Min Seong)

충북대학교 공학사와 석사 학위를 취득하였다. 현재 충북대학교 박사과정 중이며, 주요 관심분야는 딥러닝, 컴퓨터비전 등이다.

고 재 필 (Ko, Jae Pil)

연세대학교 공학사, 석사와 박사학위를 취득하였다. 현재 금오공과대학교 컴퓨터공학과교수로 재직하고 있으며, 주요 관심분야는 패턴인식, 컴퓨터비전, 영상처리 등이다.

최 경 주 (Cheoi, Kyung Joo)

충북대학교 이학사와 연세대학교 석사와 박사학위를 취득하였다. 현재 충북대학교 소프트웨어학부 교수로 재직하고 있으며, 주요 관심분야는 딥러닝, 컴퓨터비전, 영상처리 등이다.

<Abstract>

A Novel Image Captioning based Risk Assessment Model

Jeon, Min Seong · Ko, Jae Pil · Cheoi, Kyung Joo

Purpose

We introduce a groundbreaking surveillance system explicitly designed to overcome the limitations typically associated with conventional surveillance systems, which often focus primarily on object-centric behavior analysis.

Design/methodology/approach

The study introduces an innovative approach to risk assessment in surveillance, employing image captioning to generate descriptive captions that effectively encapsulate the interactions among objects, actions, and spatial elements within observed scenes. To support our methodology, we developed a distinctive dataset comprising pairs of [image-caption-danger score] for training purposes. We fine-tuned the BLIP-2 model using this dataset and utilized BERT to decipher the semantic content of the generated captions for assessing risk levels.

Findings

In a series of experiments conducted with our self-constructed datasets, we illustrate that these datasets offer a wealth of information for risk assessment and display outstanding performance in this area. In comparison to models pre-trained on established datasets, our generated captions thoroughly encompass the necessary object attributes, behaviors, and spatial context crucial for the surveillance system. Additionally, they showcase adaptability to novel sentence structures, ensuring their versatility across a range of contexts.

Keyword: Intelligent Surveillance Systems, Image Captioning, Risk Level Assessment

* 이 논문은 2023년 11월 9일 접수, 2023년 11월 20일 1차 심사, 2023년 12월 6일 게재 확정되었습니다.