The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 22, No. 5, pp.17-22, Oct. 31, 2022. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2022.22.5.17
JIIBC 2022-5-3

온디바이스 AI 비전 모델이 탑재된 지능형 엣지 컴퓨팅 기기 개발

Development of an intelligent edge computing device equipped with on-device AI vision model

강남희*

Namhi Kang*

요 약 본 논문에서는 지능형 엣지 컴퓨팅을 지원할 수 있는 경량 임베디드 기기를 설계하고, 영상 기기로부터 입력되는 이미지에서 객체를 실시간으로 빠르게 검출할 수 있음을 보인다. 제안하는 시스템은 산업 현장이나 군 지역과 같이 사전에 설치된 인프라가 없는 환경에 적용되는 지능형 영상 관제 시스템이나 드론과 같은 자율이동체에 탑재된 영상 보안 시스템에 적용될 수 있다. 지능형 비전 인지 시스템이 확산 적용되기 위해 온디바이스 AI(On-Device Artificial intelligence) 기술 적용 필요성이 증대되고 있다. 영상 데이터 취득 장치에서 가까운 엣지 기기로의 컴퓨팅 오프 로딩은 클라우드를 중심으로 수행되는 인공지능 서비스 대비 적은 네트워크 및 시스템 자원으로도 빠른 서비스 제공이 가능하다. 또한, 다양한 해킹 공격에 취약한 공격 표면의 감소와 민감한 데이터의 유출을 최소화 할 수 있어 다양한 산업에 안전하게 적용될 수 있을것으로 기대된다.

Abstract In this paper, we design a lightweight embedded device that can support intelligent edge computing, and show that the device quickly detects an object in an image input from a camera device in real time. The proposed system can be applied to environments without pre-installed infrastructure, such as an intelligent video control system for industrial sites or military areas, or video security systems mounted on autonomous vehicles such as drones. The On-Device AI(Artificial intelligence) technology is increasingly required for the widespread application of intelligent vision recognition systems. Computing offloading from an image data acquisition device to a nearby edge device enables fast service with less network and system resources than AI services performed in the cloud. In addition, it is expected to be safely applied to various industries as it can reduce the attack surface vulnerable to various hacking attacks and minimize the disclosure of sensitive data.

Key Words : AI, Computer Vision, Deep Learning, Ondevice-AI, Lightweight Device

*정회원, 덕성여대 사이버보안전공
접수일자 2022년 8월 30일, 수정완료 2022년 9월 26일
게재확정일자 2022년 10월 7일

Received: 30 August, 2022 / Revised: 26 September, 2022 / Accepted: 7 October, 2022
*Corresponding Author: kang@duksung.ac.kr
Dept. of Cybersecurity, Duksung Women's University, Korea

Ⅰ. 서 론

물리 보안시스템은 인프라 시설 및 정보 등의 자산(Asset)과 인명을 보호하기 위해 물리적 취약성(Vulnerability)에 대응하여 침해 위협을 통제할 수 있는 전반적인 보안 시스템을 의미한다. 산업 현장 및 서비스 공간에 적용되는 영상 감시 시스템, 출입 통제 시스템, 침입 경보 시스템, 물리 도난 대응 시스템 등이 포함된다. 최근 인공지능 기술을 활용한 지능형 물리 서비스 산업 및 지능형 엣지 컴퓨팅 시장이 확대되고 있다[1, 2].

본 논문에서는 물리 보안 시스템 중 영상이나 이미지를 기반하여 객체를 검출하고 이상징후를 탐지하는 지능형 비전 인지 시스템을 대상으로 한다. 물리 보안 산업에서 지능형 비전 인지 시스템이 확산 적용되기 위해 온디바이스 AI(On-Device Artificial intelligence) 기술 적용 필요성이 증대되고 있다[3]. 물리 공간에 인프라로 설치된 보안시스템은 클라우드 및 온프레미스 관제 시스템과 연계되어 침해 위협에 대응할 수 있겠지만, 건설 현장이나 군 지역과 같이 인프라 없이 애드혹으로 설치되는 지능형 영상 관제 시스템이나 드론과 같은 자율이동체에 탑재된 영상 보안 시스템은 준 실시간으로 기능을 처리할 수 있는 지능형 엣지 컴퓨팅 기술이 필요하다.

온디바이스 엣지 컴퓨팅 기술은 클라우드에 전송된 데이터를 중앙 집중식으로 처리하는 기존의 방안에서, 데이터가 생성되는 기기에서 가까운 위치에 있는 네트워크 엣지에서 데이터를 처리하는 기술을 의미한다. 또한 이러한 엣지에서 지능형 서비스를 동작시키는 온디바이스 AI는 사용자 주변의 상황 정보를 클라우드나 서버로 전송하여 인공지능 모델을 학습하고 추론하는 기존의 인공지능 서비스를, 사용자 주변의 네트워크 엣지에서 수행할 수 있도록 해주는 기술이다[3]. 미국 MIT에서는 온디바이스 AI를 향후 5년 안에 사회에 큰 영향과 변화를 가져올 10대 혁신 기술 중 하나로 선정하기도 했다.

엣지 기기로의 컴퓨팅 오프 로딩은 클라우드를 중심으로 수행되는 인공지능 서비스 대비 적은 네트워크 및 시스템 자원으로도 빠르고 안전한 서비스 제공을 특징으로 한다. 즉, 기존 인공지능 서비스에서 발생되었던 데이터 트래픽 양의 폭증, 개인정보의 프라이버시 문제, 서비스 지연 시간 등의 문제를 해결할 수 있는 방안이 된다. 따라서 엣지의 경량 기기의 제한된 시스템 자원을 최적 활용하여 성능 한계를 극복할 수 있는 기술과 온디바이스에 특화된 지능형 기술 개발이 필요하다.

이러한 관점에서, 본 논문에서는 경량 임베디드 기기에서 인공지능 추론 서비스를 효율적으로 수행할 수 있는 방안을 제시하고 시험한다. 경량 기기에서 인공지능 추론 서비스를 제공하기 위해 하드웨어 가속 모듈을 탑재하여 CPU 기반 처리속도와 하드웨어 모듈을 장착했을 때의 속도를 비교하여 엣지에서 인공지능 비전 서비스가 원활하게 동작될 수 있음을 검증한다.

Ⅱ. 관련 연구

인공지능 서비스는 다양한 산업에 적용되어 활용되고 있다[11, 12]. 특히, 컴퓨터가 사람처럼 이미지나 영상에서 물체를 식별하고 장면을 이해하는 컴퓨터 비전 기술이 인공지능 기술의 하나인 딥러닝 기술의 적용으로 급속도로 발전하고 있다. 2015년 이미지 인식 경진대회인 ILSVRC 대회에서는 사람의 인식률인 95% 정도를 넘어선 96.4%의 인식률을 보이는 딥러닝 모델이 발표되었고, 2020년에는 98% 이상의 인식률로 사람보다 월등하게 높은 인공지능 기술이 발표되기도 했다[4].

이미지에서 객체를 인식하는 방식도 빠르게 발전하고 있다. 초기 모델에서 가장 많이 알려진 R-CNN은 이미지에서 물체가 존재한다고 예상되는 영역을 제안하고 컨볼루션 신경망을 이용하여 각 제안된 영역에서 물체를 검출한다. 이러한 2단계 처리방식은 처리 속도의 문제가 있는데 이를 개선하기 위해 Fast RCNN과 Faster R-CNN모델 등이 제안되었다. 또한, 객체 검출의 정확도는 떨어지지만 2단계 처리방식이 아닌 1단계 방식으로 동작되는 YOLO나 SSD와 같은 모델 등도 제안되었다[5].

본 논문에서는 이 중 SSD Mobilenet을 적용하였다. 1단계 처리로 동작되는 YOLO와 SSD는 다음 그림 1과 같은 차이를 갖는다.

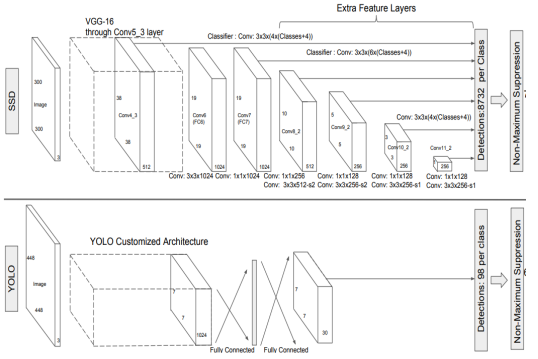

그림 1. SSD와 YOLO 모델 비교
Fig. 1. A comparison between SSD and YOLO[6]

SSD는 컨벌루션과 풀링 계층에서 나타나는 위치정보를 활용하여 특징을 추출하고 앵커를 활용하여 위치와 크기가 임의로 존재하는 물체를 검출하는 방안이다[6]. MobileNet은 기존 CNN 방식에서 계층과 노드가 많아질수록 연산량이 증가하는 문제를 해결하기 위해 Depth-wise 컨벌루션과 Point-wise 컨벌루션을 적용하여 연산량을 줄인 네트워크이다[7].

일반적인 인공지능 기반 컴퓨터 비전 시스템의 경우 사전에 확보된 데이터셋을 기반으로 학습을 수행하고, 학습된 모델을 적용하여 서비스에 필요한 분석이나 추론을 수행한다. 모델의 구성과 구현에 따라 학습과 추론에 많은 컴퓨팅 자원을 요구하게 된다. 따라서, 제안된 많은 기술들이 산업 현장에 설치된 카메라를 통해 입력되는 이미지나 영상을 클라우드 시스템이나 백엔드에 구축된 이미지 분석 모듈로 전송하여 기능을 수행하도록 구현하고 있다[8, 9, 10].

하나의 예로 스마트폰과 아두이노를 활용하여 이동형 홈 CCTV 시스템이 제안되기도 했다[10]. 해당 제안에서는 기존에 많이 사용되고 있는 CCTV의 고정 방식 문제를 개선하고 비용을 절감하기 위해 구형 스마트폰을 활용하고 있다. 그러나 엣지에서 객체를 인식하지 못하고 서버로 이미지를 전송하고, 서버에서 머신러닝을 통한 인식이 수행된 후 인식률과 라벨이 포함된 객체 인식 결과를 보여주는 방식으로 통신 네트워크를 사용하고 있다. 이렇게 동작하는 서비스는 네트워크를 이용하여 분석 모듈로 데이터를 전송하고 추론이 이루어진 후 결과가 전송되기 때문에 실시간으로 추론이 필요한 현장에 반영이 어렵다. 또한 서비스 시나리오에 따라 민감도가 높은 데이터가 서버로 전송되므로 통신 과정과 시스템에서 다양한 보안 위협에 노출되는 문제가 있다. 이를 해결하기 위해서는 경량 기기에서 인공지능 추론이 가능해질 수 있는 온디바이스 엣지 컴퓨팅 기술이 필요하다.

Ⅲ. 제안시스템 설계 및 구현

본 논문에서는 온디바이스 AI 응용에서 요구되는 실시간 추론을 제공하기위한 시스템을 구현하고 성능을 제시한다. 2장에 기술한 것처럼 인공지능 응용은 학습 단계와 학습이 완료된 모델을 사용하는 추론이나 분석 단계로 구분된다. 제안 시스템에서는 데이터셋의 크기가 크고 많은 컴퓨팅 자원이 요구되는 학습은 GPU 자원이 확보된 서버에서 수행하였고, 학습 완료된 비전 인식 모델을 임베디드 기기에 탑재하여 실시간으로 객체가 인식되는지를 검증하였다. 즉, GPU나 TPU와 같은 하드웨어 가속 모듈을 지능형 응용에서 요구되는 연산 능력에 맞게 커스터마이징할 수 있는 하드웨어 가속 모듈을 탑재한 지능형 엣지 컴퓨팅 기기 개발을 목표로 한다.

제안 시스템은 드론과 같은 자율 이동체나 기존에 설치되어 운영되고 있는 CCTV와 같은 보안 시스템에 연결하여 지능형 서비스를 엣지에서 수행할 수 있도록 한다. 이를 위해 범용 임베디드 시험 기기로 많이 사용하는 라즈베리파이 컴퓨트 모듈을 활용하여 기존의 시스템을 포함하여 TPU 모듈과 연결될 수 있는 다양한 통신 인터페이스를 보드에 장착하였다. 그림 2는 TPU와 타 기기와 연결될 수 있는 인터페이스를 장착한 개발 보드를 나타낸다.

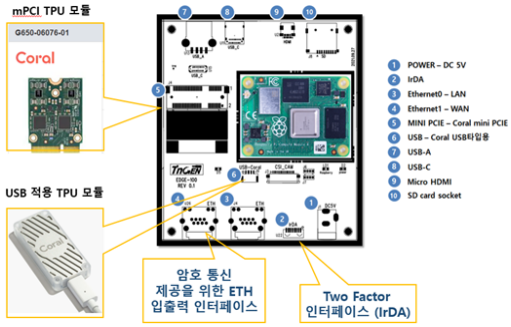

그림 2. 제안시스템 구조
Fig. 2. Proposed System Architecture

온디바이스 AI 비전 기술이 탑제된 임베디드 컴퓨팅 보드의 딥러닝 처리 속도는 보드에 적용한 하드웨어 가속 모듈인 Google Coral TPU의 처리 속도에 의존한다. 본 제안에서 사용한 구글의 Coral USB 가속 모듈은 구글에서 자체적으로 만든 프로세서를 사용하고 있고, x86-64 또는 ARMv8 명령어 집합을 포함하는 ARM32/64 기반의 프로세서에서 사용 가능하도록 지원하고 있다.

그림 2에 나타낸 것처럼, 본 논문에서 시험용으로 구현한 임베디드 보드는 Google Coral TPU 모듈을 2개 이상 적용할 수 있도록 USB와 miniPCI를 통신인터페이스로 사용할 수 있도록 하였다. 본 논문에서 진행한 시험에서는 USB 인터페이스를 적용하여 1개의 모듈을 적용했다. USB를 사용하는 Coral TPU 모듈의 주요 하드웨어 스펙은 다음과 같다. 하드웨어 스펙으로 판단할 때, Coral TPU에서 제공되는 딥러닝 가속 속도는 4TOPS로 float 연산인 FLOPS로 변경 시 1FLOPS에 해당한다. 기

타 모듈의 스펙은 다음 표 1과 같다.

표 1. 구글 TPU 하드웨어 스펙
Table 1. Google TPU Hardware Spec.

ML Accelerator	TPU Processor: 4TOPS (int8)
Energy Consumption	2 TOPS per watt
Connector	USB 3.0 Type-C (data/power)
Dimensions	65mm X 30mm

Ⅳ. 성능 평가

1. AI 모델 및 시험 환경

인공지능 객체인식 추론 처리 속도 검증을 위한 시험 평가에서는 구글의 텐서플로우 라이트(TensorFlow Lite)에서 경량화된 객체 인식 AI 모델인 SSD Mobilenet v1(코드: coco_ssd_mobilenet_v1_1.0_quant_2018_06_29)을 사용하여 실시간 입력되는 영상에서 객체를 인식하는 처리 속도를 측정하였다. 객체 인식에 사용되는 영상은 온디바이스 AI가 지원되는 임베디드 기기에 USB 카메라를 연결하여 유투브(YouTube) 영상이 출력되는 모니터 화면을 실시간 촬영하여 해당 영상에 있는 사람을 인식하는 속도를 측정하였다. 시험하기 위한 시스템 구성은 다음 그림 3과 같다.

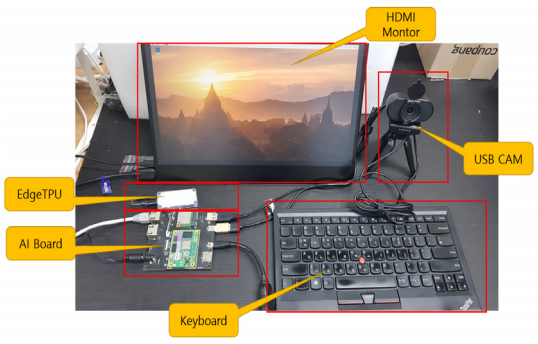

그림 3. 시험 환경
Fig. 3. Test Environment

시험 환경의 주요 항목은 다음과 같이 구성된다.

- AI Board: 온디바이스 AI 비전 기술이 동작되는 임베디드 보드
- Edge TPU: AI 비전인지 Object Detection을 위한 하드웨어 가속 모듈
- USB Cam: Object detection에 사용되는 영상 데이터의 송신원

온디바이스 AI 응용은 경량 임베디드 기기를 대상으로 하고 있어 영상 추론을 위한 AI 모델도 경량화되어야 한다. 본 논문에서 적용한 Mobilenet SSD는 객체 인식을 수행할 수 있는 경량 AI 모델로 CPU를 위한 모델의 크기와 TPU를 사용할 수 있도록 한 모델의 크기는 다음과 같다.

- CPU 기반으로 동작되는 Object Detection 모델 크기: 4.2Mbyte
- TPU 기반으로 동작되는 Object Detection 모델 크기: 6.9Mbyte

2. 인공지능 객체인식 추론 처리 속도

온디바이스 AI가 지원되는 임베디드 기기에 하드웨어 가속 모듈인 구글 Coral TPU를 사용하여 객체 인식하는 경우 CPU만을 사용한 것 대비 우수한 성능을 보였고, 실시간 입력 영상에서 객체를 인식하는데 무리가 없는 결과를 얻었다.

하드웨어 가속기의 사용 여부의 특성을 확인하기 위해 TPU를 장착하지 않고 보드의 CPU만을 사용하여 시험했다. 그림 4처럼 유투브 영상을 USB 카메라로 입력 받아 AI 보드의 CPU를 사용하여 사람을 인식하는 장면을 나타내고, 사람을 인식하는 화면을 크게 출력하여 결과를 보였다. 실행된 화면의 상단에 객체 인식을 위해 필요한 성능 지표를 나타내었다. 그림에서 FPS는 초당 객체 인식이 처리되는 영상 프레임 수이고, inference는 객체 인식 응용의 처리 속도를 나타내는데 다음 3가지 지연 시간의 총합을 의미한다.

시험의 평균 추론 속도(inference 속도로 처리되는 모든 지연 시간의 합)는 210msec 정도로, 1초에 4~4.5개 정도의 영상 이미지에서 Object를 검출해낼 수 있었고 평균 4.23 FPS의 결과를 보였다.

그림 4처럼 유투브 영상을 USB 카메라로 입력 받아 AI 보드의 CPU를 사용하여 사람을 인식하는 장면을 나타내고, 사람을 인식하는 화면을 크게 출력하여 결과를 보였다. 실행된 화면의 상단에 객체 인식을 위해 필요한 성능 지표를 나타내었다. 그림에서 FPS는 초당 객체 인식이 처리되는 영상 프레임 수이고, inference는 객체 인식 응용의 처리 속도를 나타내는데 다음 3가지 지연 시간의 총합을 의미한다.

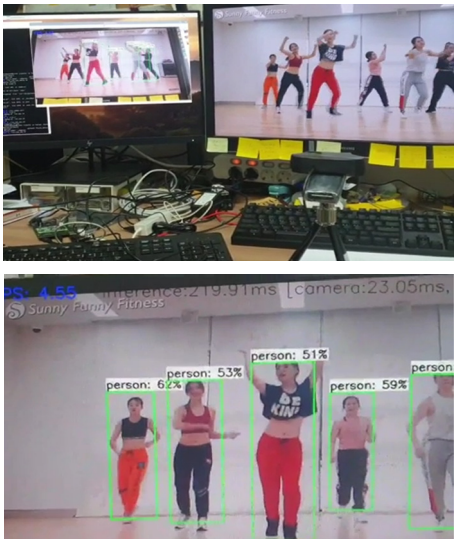

그림 4. CPU를 사용하는 경우의 추론 속도
Fig. 4. Inference time in case of using CPU

- camera: 영상 송신원으로부터 이미지를 로딩하는 속도
- calc: 영상에서 object detection을 처리하는 속도
- Other: 메모리 관리 등 Object detection을 위해 요구되는 기타 프로세싱 속도

임베디드 보드에서 온디바이스 AI의 추론 속도 향상을 위해 하드웨어 가속기 (Google Coral TPU)를 장착하여 시험을 진행하였다. CPU만 사용했던 시험과 동일한 영상을 USB 카메라로 촬영하여 객체 인식의 추론을 수행하였다. 다음 그림 5의 수치로 확인할 수 있듯 CPU만을 활용한 경우보다 좋은 성능을 보였는데 평균 추론 속도는 88msec로 1초에 11~13개 정도의 영상 이미지에서 Object를 검출해낼 수 있었다.

그림 5. TPU 모듈을 사용하는 경우의 추론 속도
Fig. 5. Inference time in case of using TPU module

V. 결 론

본 논문에서는 인공지능 비전 모델이 경량 임베디드 기기에서도 동작될 수 있는 방안을 제시하고 시험을 통해 객체 인식이 실시간 처리가 가능함을 보였다. 엣지 기기에서 인공지능 기술이 동작될 경우 적은 네트워크 및 시스템 자원으로도 비용 효율적인 서비스 제공이 가능하고, 해킹 공격 표면의 감소로 보안에 강한 서비스 제공이 가능하여 다양한 영역에서 사용될 수 있다.

References

[1] Byung-Wan Kim, "Temperature Measurement and Intelligent Access Management System Service Platform Advancement Research using AI Facial Recognition Technology," Journal of the Korea Entertainment Industry Association(JKEIA), Vol. 15, No. 7, Oct. 2021.
DOI : https://doi.org/10.21184/jkeia.2021.10.15.7.24

[2] Xingwei Sun, Yalan Ning, Decheng Yang, "Research on the application of deep learning in campus security monitoring system," Journal of Physics: Conference Series, Vol. 1744. No. 4. pp. 042035, Oct. 2021.
DOI: https://doi.org/10.1088/1742-6596/1744/4/042035

[3] Guangxu Zhu, Dongzhu Liu, Yuqing Du, Changsheng You, Jun Zhang, Kaibin Huang, "Toward an Intelligent Edge: Wireless Communication Meets Machine Learning," IEEE Communications Magazine, Vol. 58, Is. 1, pp. 19-25, Jan. 2020.
DOI: https://doi.org/10.48550/arXiv.1809.00343

[4] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, Quoc V. Le, "Self-training with Noisy Student improves ImageNet classification," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10687-10698, Jun. 2020.
DOI: https://doi.org/10.48550/arXiv.1911.04252

[5] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, Rong Qu, "A Survey of Deep Learning-Based Object Detection", IEEE Access, Vol. 7, pp. 128837-128868, Sep. 2019.
DOI: https://doi.org/10.1109/ACCESS.2019.2939201

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, "SSD: Single Shot MultiBox Detector," European conference on computer vision. Springer, Cham, LNCS Vol. 9905, pp. 21-37, Sep. 2016.
DOI: https://doi.org/10.48550/arXiv.1512.02325

[7] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, "MobileNets: Efficient

Convolutional Neural Networks for Mobile Vision Applications,"
DOI: https://doi.org/10.48550/arXiv.1704.04861

[8] Geunmo Kim, Jinsung Cho, Sungmin Kim, Seunghwan Beak, Seunghoon Ryu, Jaejong Koh, Bongjae Kim, "Deep Learning-based Real-time Traffic Accident Type and Fault Information Provision Service", The Journal of The Institute of Internet, Broadcasting and Communication (IIBC), Vol. 21, No. 3, pp. 1-6, Jun. 2021.
DOI: https://doi.org/10.7236/JIIBC.2021.21.3.1

[9] Jeong-Hoon Lee, Ro-Woon Lee, Seung-Taek Hong, Young-Gon Kim, "Image Processing System based on Deep Learning for Safety of Heat Treatment Equipment", The Journal of The Institute of Internet, Broadcasting and Communication (IIBC), Vol. 20, No. 6, pp. 77-83, Dec 2020.
DOI: https://doi.org/10.7236/JIIBC.2020.20.6.77

[10] Dong-Ju Kim, Chae-Won Lim, Hyun-Ho Choi, "Development of Remote-Controlled Object- Recognizing Mobile Home CCTV Using Smartphone and Arduino", Journal of the Korea Institute of Information and Communication Engineering, Vol. 24, No. 11, pp. 1546-1549, Nov 2020.
DOI: http://doi.org/10.6109/jkiice.2020.24.11.1546

[11] Songhee Kim, Sunhye Kim, Byungun Yoon, "Deep Learning-Based Vehicle Anomaly Detection by Combining Vehicle Sensor Data", Journal of the Korea Academia-Industrial cooperation Society, Vol. 22, No. 3, pp. 20-29, Mar. 2021.
DOI: https://doi.org/10.5762/KAIS.2021.22.3.20

[12] Seok-Jin Kwon, Min-Soo Kim, "Flaw Evaluation of Bogie connected Part for Railway Vehicle Based on Convolutional Neural Network", Journal of the Korea Academia-Industrial cooperation Society, Vol. 21, No. 11, pp. 53-60, Nov. 2020.
DOI: https://doi.org/10.5762/KAIS.2020.21.11.53

저 자 소 개

강 남 희(정회원)

- 1999년 : 숭실대학교 정보통신공학과 (공학사)
- 2001년 : 숭실대학교 정보통신대학원 (공학석사)
- 2005년 : University of Siegen 컴퓨터공학과 (공학박사-인터넷및시스템보안)
- 2009년 ~ 현재 : 덕성여대 과학기술대학 사이버보안전공교수
- 2016년 ~ 현재 : 덕성여대 학교기업 DCS 대표
- 관심분야 : 유무선 인터넷통신, 인터넷보안, 시스템 보안, 사물인터넷 보안, 온디바이스 AI