엣지 디바이스 기반 자기 지도 학습을 활용한 환편기 원단 결함 검출 시스템의 경량화 및 배포 파이프라인 설계

문서영, 김다현, 김동주, 서영주, 황병일*

*포항공과대학교 인공지능연구원

moondau@postech.ac.kr, kdhyun8011@postech.ac.kr, kkb0320@postech.ac.kr,

yjsuh@postech.ac.kr, *bihwang@postech.ac.kr

Lightweight Edge Deployment Pipeline for Self-Supervised Fabric Defect Detection in Circular Knitting Machines

Moon Seo Young, Kim Da Hyun, Kim Dong Ju, Suh Young Joo, Hwang Byeong Il*

*POSTECH Institute of Artificial Intelligence

요 약

환편기는 원통형 바늘 실린더가 고속 회전하며 실을 편직하는 설비로, 미세한 부품 노후화나 이물 유입에도 원단 결함이 발생하며 탐지 지연 시 대량 폐기로 이어진다. 국내 현장에서는 작업자 육안 검사에 의존하여 검출 누락이 빈번하며, 수백 종의 결함으로 라벨 확보가 어렵다. 또 다수의 환편기에서 동시 수집되는 대용량 데이터를 일관되게 처리하려면 학습을 서버에서 수행한 뒤 각 설비로 배포하는 방식이 필수적이다. 본 연구는 서버에서 정상 데이터 기반 MSFlow를 학습하고 임계치를 산정한 후 ONNX(Open Neural Network Exchange)로 변환하고 TensorRT의 FP16/INT8 양자화로 경량화하여 젯슨 오린 나노(Jetson Orin Nano)로 배포하는 시스템을 제시한다. 각 환편기 카메라에서 취득한 영상은 100×100 표준 패치로 분할되어 자동화된 파이프라인의 학습 데이터로 활용되며, 배포된 엔진은 현장에서 프레임 단위 이상 점수 맵을 실시간으로 산출하고 임계치 초과 시 즉시 경보와 기계 정지를 수행한다. 이 구조는 엣지 디바이스 환경에서 메모리 사용량과 추론 시간을 감소시켜 결함의 연속 발생을 차단하고 경제적 손실이 최소화될 것으로 기대된다.

Ⅰ. 서 론

환편기는 원통형 바늘 실린더가 고속 회전하며 실을 편직하는 설비이다. 부품의 미세한 노후화나 이물 유입과 같은 사소한 요인에도 결함이 발생하며, 원인이 제거될 때까지 동일 결함이 반복된다. 탐지가 지연되면 원단 결함 발생 시점 이후의 생산물이 대량 폐기로 이어져 직접적 비용 손실을 초래한다. 국내 현장은 환편기 2~4대 당 1명이 관리하는 고밀도 인력 배치가 일반적이며, 현재는 작업자의 육안에 크게 의존한다. 이로 인해 장시간 근무에 따른 피로 누적과 주관적 판단 편차로 검출 누락과 오류가 빈번히 발생한다. 기존의 인력 중심에서 벗어난 스마트 관리 체계의 필요성이 대두되고 있다. 또한, 동일 공장 내 다양한 원단이 생산되고, 환편기별로 생산 품목과 특성이 상이하여 단일 모델로 포괄하기 어렵다. 원단 결함 유형은 수백 종에 이르지만 발생 빈도가 낮아 결함별 충분한 라벨을 축적하기 어렵다. 따라서 관측되지 않은 결함 패턴에도 견고하게 작동하는 자기 지도 학습(Self-Supervised Learning) 이상탐지가 요구된다.

환편기는 수십 RPM으로 운전되며 장비 내부의 공간 및 비용 문제로 고성능 PC 도입이 어렵다. 이에 카메라 탑재 엣지 디바이스를 활용해 결함 탐지 시 즉시 기계를 정지시킬 수 있다. 이는 모델의 연산량과 메모리 사용을 낮춘 경량화가 필수이다. 또 환편기는 제조 공정에 따라 생산하는 원단 변경이 자주 일어나 엣지 환경 학습 시 느린 속도로 신속한 대응이 어렵다. 본 연구는 정상 데이터 기반 비지도 이상탐지를 채택하고, 데이터를 서버로 전송하여 모델 학습 후, 이를 ONNX(Open Neural Network Exchange)로 변환한 뒤 TensorRT로 FP16/INT8 양자화하여 젯슨 오린 나노(Jetson Orin Nano)에 배포함으로써, 엣지 환경에서 실시간 추론과 안정적 운영이 가능한 원단 결함 검출 시스템 구조 제시한다.

Ⅱ. 관련 연구

스마트 제조 비전에서 결함 검출 문제에 대응하기 위해, 정상 데이터만으로 표현 공간을 학습해 이상도를 산출하는 자기 지도 학습 기반 이상탐지가 활발히 연구되고 있다. 특히, MSFlow는 다중 스케일 특징 추출과 융합 흐름을 통해 스케일 변동이 큰 이상 패턴에 견고하도록 설계된 모델로 보고되었다.[2] 선행 연구에서는 EfficientAD와 MSFlow를 동일 조건에서 비교한 결과, AUROC는 유사했으나 MSFlow가 180 FPS로 더 높은 추론 속도를 달성했으며, 이상맵 기반 결함 구획화에서 보다 안정적인 영역을 재현했다.[1]

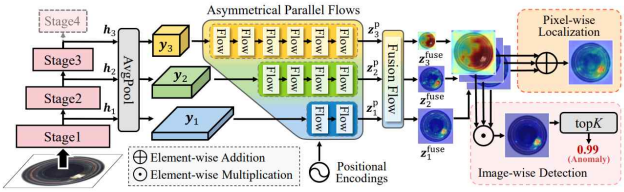
한편, 기존 연구는 주로 GPU 기반 서버 환경을 전제로 했으나, 메모리 제약이 큰 엣지 디바이스에 직접 적용하기에는 한계가 있다. 이를 해결하기 위해 모델 경량화 및 최적화 기법이 활발히 연구되고 있다. 객체 검출 분야에서는 모델 경량화와 ONNX 형식 변환 후 TensorRT 최적화를 결합할 경우, 젯슨 오린 나노 계열과 같은 엣지 디바이스에서 추론 시간이 단축되고 전력 대비 처리량이 향상된다는 결과가 보고되었다.[3]

특히, 편직 공정은 제품과 패턴 전환 주기가 짧아 모델의 신속한 학습과 즉시 가동이 요구된다. 따라서 다수의 환편기에서 동시 수집되는 대용량 영상과 메타 데이터를 집약하고 학습 파이프라인을 운영하기 위해서는 서버 인프라가 필수적이다. 이에 본 연구는 정상 데이터 기반 MSFlow를 서버에서 학습하고 임계치를 산정한 뒤 ONNX 변환과 TensorRT의 양자화로 경량화한 엔진을 각 환편기에 배포하여 실시간 추론을 수행하는 서버 학습과 엣지 배포 전략의 시스템 구조를 제안하고자 한다.

Ⅲ. 관련 기술 및 경량화 기반 시스템 구조

1. MSFlow 기반 이상 탐지 모델

MSFlow는 정상 데이터만으로 특징 분포를 학습하는 비지도 이상탐지로, 다중 스케일 특징을 추출해 정규화 흐름으로 각 스케일 분포를 모델링하고 이를 결합해 이상 점수를 산출한다. 크기 변동이 큰 결함에 견고하며, 추론 단계에서 프레임 단위 이상 점수 맵을 생성하고 임계치로 결합 영역을 구획화한다. 라벨 확보가 어려운 환편기 환경에 적합하고 저연산 조건에서도 이상 데이터 검출 성능이 우수하다.



[그림1] MSFlow 구조도

2. 모델 변환 및 추론 최적화 기법

MSFlow 모델은 일반적으로 PyTorch 환경에서 학습되므로, 이를 엣지 디바이스에서 효율적으로 실행하려면 ONNX 형식으로의 변환이 요구된다. ONNX는 딥러닝에 사용되는 여러 가지 프레임워크를 지원하며 학습된 모델을 배포하거나 프레임워크 간에 이동을 용이하게 한다.

TensorRT는 다양한 딥러닝 프레임워크에서 학습된 모델을 최적화하여 NVIDIA GPU 상에서 추론 속도를 최대 수십 배까지 향상시킬 수 있는 모델 최적화 엔진이다. ONNX 모델을 실행 엔진으로 변환하면서 레이어 통합과 연산 경로 최적화를 적용해 추론 속도와 연산 효율을 향상시킨다.

3. MSFlow 모델을 활용한 환편기 원단 결함 검출 시스템 설계

3-1. 영상 수집과 전송

영상 수집은 환편기 내부 카메라를 통해 수행되며, 원본 해상도 1024×1024 영상을 RTSP(Real-Time Streaming Protocol) 스트리밍으로 서버에 전송한다. 서버는 스트림을 수신하여 프레임 시퀀스로 변환한 뒤 저장소에 적재되며, 각 프레임에는 위치, 타임스탬프, 장비 ID 등 메타정보를 포함하여 처리한다. 이후 프레임으로 변환된 이미지에 대해 다양한 패치 크기를 비교 평가하여 원단 패턴과 결함이 가장 선명하게 드러나는 100×100을 표준 패치로 절단한다. 생성된 패치는 MSFlow 학습 파이프라인으로 자동 연계된다.

3-2. 정상 데이터 기반 학습 자동화

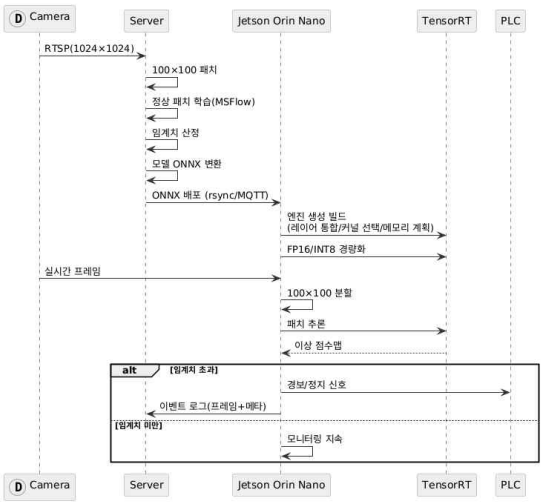
MSFlow는 정상 데이터만으로 특징 분포를 학습하므로 학습 단계에는 정상 패치만 사용한다. 검증 구간에서 이상 점수의 분포를 분석해 임계치를 산정한다. 학습 파이프라인은 서버에서 자동으로 구성하여 데이터 적재, 배치 생성, 학습 및 검증, 체크포인트 저장까지 일괄 수행한다. 임계치 산정 결과와 모델 가중치는 버전 정보와 함께 자동 저장된다.

3-3. 모델 전송 및 TensorRT 엔진 생성

학습이 완료된 MSFlow 모델은 ONNX 형식으로 변환한다. 변환된 ONNX 모델은 rsync, MQTT 등 적합한 전송 방식으로 젯슨 오린 나노에 배포한 뒤 장치에서 TensorRT로 실행 엔진을 생성한다. 이때 레이어 통합과 커널 선택, 메모리 계획 등 최적화를 적용하고, FP16/ INT8로 정밀도 변환하여 추론 속도와 연산 효율을 향상시킨다.

3-4. 실시간 추론 및 결과 처리

최적화된 엔진은 젯슨 오린 나노에서 카메라 프레임을 수신하는 즉시 100×100 표준 패치로 분할한 뒤 TensorRT로 패치별 이상 점수를 계산하여 프레임 단위 이상 점수 맵을 실시간으로 생성한다. 이상 점수가 임계치를 초과하면 즉시 경보를 발생시키고 설비 인터록 신호를 출력하여 기계를 정지한다. 이벤트 발생 시 프레임과 메타 데이터를 서버에 기록하여 추후 분석과 재학습에 활용한다.



[그림2] TensorRT 기반 환편기 원단 결함 탐지 시스템 구조도

Ⅳ. 결론

본 연구에서는 환편기 원단 결함 검출을 위해 MSFlow 기반 비지도 이상탐지 모델을 구축하고, 영상을 서버로 전송해 학습된 모델의 ONNX 변환 및 TensorRT 경량화를 통해 젯슨 오린 나노에 배포하는 시스템 구조를 제시하였다. 제안 시스템으로 메모리 사용량과 추론 시간이 크게 감소하여, 환편기에서 연속으로 수집되는 영상에 대해 프레임 단위 이상 점수 맵을 실시간 산출할 수 있다. 이상 점수가 설정한 임계치를 초과할 경우 즉시 경보를 발생시키고 기계 정지를 포함한 후속 절차를 수행함으로써, 결함의 연속 발생을 차단하고 경제적 손실이 최소화될 것으로 기대된다.

ACKNOWLEDGMENT

이 연구는 2025년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업(RS-2022-NR070870) 지원을 받아 수행되었으며, 과학기술정보통신부·경찰청이 공동 지원한 '폴리스랩 2.0 사업'(RS-2023-00281072)과 2025년도 교육부의 재원으로 수행된 경상북도 지역혁신중심 대학지원체계(RISE) - 특화산업 Scale-UP 사업(B0080527002567)의 지원을 받아 수행된 연구입니다.

참 고 문 헌

[1] 김다현, 황병일, 황도경, 김호연, 권현섭, 이창엽, 안서희, 김동주, 서영주. (2024). 자기 지도 학습 기반 환편기 원단 결함 실시간 고속 탐지 시스템 개발. 한국통신학회 학술대회논문집.

[2] Zhou, Y., Xu, X., Song, J., Shen, F., & Shen, H. T. (2023). MSFlow: Multi-Scale Flow-based Framework for Unsupervised Anomaly Detection. arXiv preprint arXiv:2308.15300.

[3] Alqahtani, D.K., Cheema, A., Toosi, A.N.(2024). Benchmarking deep learning models for object detection on edge computing devices. arXiv preprint arXiv:2409.16808