자가 지도 웨이브릿-트랜스포머 융합을 통한 산업용 기계음 이상 탐지

FusionWave: A Self-Supervised Wavelet-Transformer Fusion Framework for Industrial Machine Sound Anomaly Detection

요 약

본 연구에서는 산업용 기계 소리 데이터에서 이상 상태를 탐지하기 위한 새로운 비지도 학습 프레임워크를 제안한다. 제안된 모델은 학습 가능한 웨이브렛 기반 변환과 트랜스포머 인코더, 그리고 벡터 양자화 오토인코더(VQ-VAE)로 구성되며, 적은 데이터 환경에서도 높은 일반화 성능을 달성하도록 설계되었다. 완전 학습형 웨이브렛(DeSpaWN) 변환을 통해 입력 음향 신호를 다중 해상도의 시간-주파수 특징으로 변환하고, 마스킹 자기지도 학습 기반 오디오 트랜스포머가 전역 문맥 정보를 학습한다. 마지막으로 VQ-VAE 디코더가 두 종류의 특징을 융합하여 입력을 복원하며 이상 점수를 산출한다. MIMII 공개 데이터셋의 4 가지 기계 유형(Fan, Pump, Slider, Valve)에 대한 실험 결과, 제안 모델은 높은 ROC-AUC 를 기록하였고 특히 복잡한 음향 패턴에서도 견고한 이상 탐지 성능을 보였다.

1. 서론

제조 업계에서는 기계 상태 모니터링을 통한 효율적인 정비가 중요하다. 지도학습은 이상 음향 데이터 부족으로 한계가 있어, 비지도 학습 기반 오토인코더가 대안으로 등장했다[1]. 그러나 단순 오토인코더는 이상 음향까지 재구성하여 탐지 민감도가 떨어진다. 최근에는 딥러닝을 활용한 고도화된 오디오 이상 탐지 기법이 연구되고 있다. 웨이블릿 변환과 트랜스포머 기반 자기지도학습 방법이 주목받지만, 산업 환경의 복잡하고 다양한 이상 상태를 단일 패턴 방식으로 탐지하기 어렵다. 본 연구에서는 DeSpaWN 기반 시간-주파수 변환[2], Masked Audio Transformer 기반 시계열 음향 패턴 학습[3], VQ-VAE 기반 이상 탐지[4] 모듈을 결합한 새로운 모델을 제안한다. 전역 필터 레이어로 음향 신호의 장기적 의존성을 학습하고, 다중 패턴 생성기로 정상 패턴을 정확하게 모델링한다. MIMII 데이터셋에서 다양한 기계 유형의 이상 탐지에 우수한 성능을 보인다.

2. MIMII 데이터셋

본 연구에서는 산업용 기계의 이상 감지를 위한 음향 기반 데이터로 MIMII(Malfunctioning Industrial Machine Investigation and Inspection) 데이터셋을 활용하였다[5]. 해당 데이터셋은 팬(Fan), 펌프(Pump), 슬라이더(Slider), 밸브(Valve) 등 총 네 가지 기계 장비의 정상 및 이상 작동 소리를 수집하여 구성되었으며, 각 기계는 고유 식별자(id_00, id_02, id_04, id_06)를 통해 구분된다. 각 샘플은 약 10 초 길이의 오디오 클립으로, 주변 소음을 포함한 현실적인 환경에서 기록되었다. 데이터는 정상 상태와, 모터 이상 등 구조적 결함이 반영된 상태로 나뉘며, 학습에는 정상 음향만을 사용하고 테스트 단계에서 이상 여부를 판별한다.

표 1. MIMII 데이터셋의 wav 파일의 개수. 한 개의 wav 파일은 10 초 동안의 기계 소리 녹음 데이터이다.

모델	이상 여부	기계 종류 4종			
		Fan	Pump	Slider	Valve
id_00	정상	1011	1006	1068	991
	이상	407	143	356	119
id_02	정상	1016	1005	1068	708
	이상	359	111	267	120
id_04	정상	1033	702	534	1000
	이상	348	100	178	120
id_06	정상	1015	1036	534	992
	이상	361	102	89	120

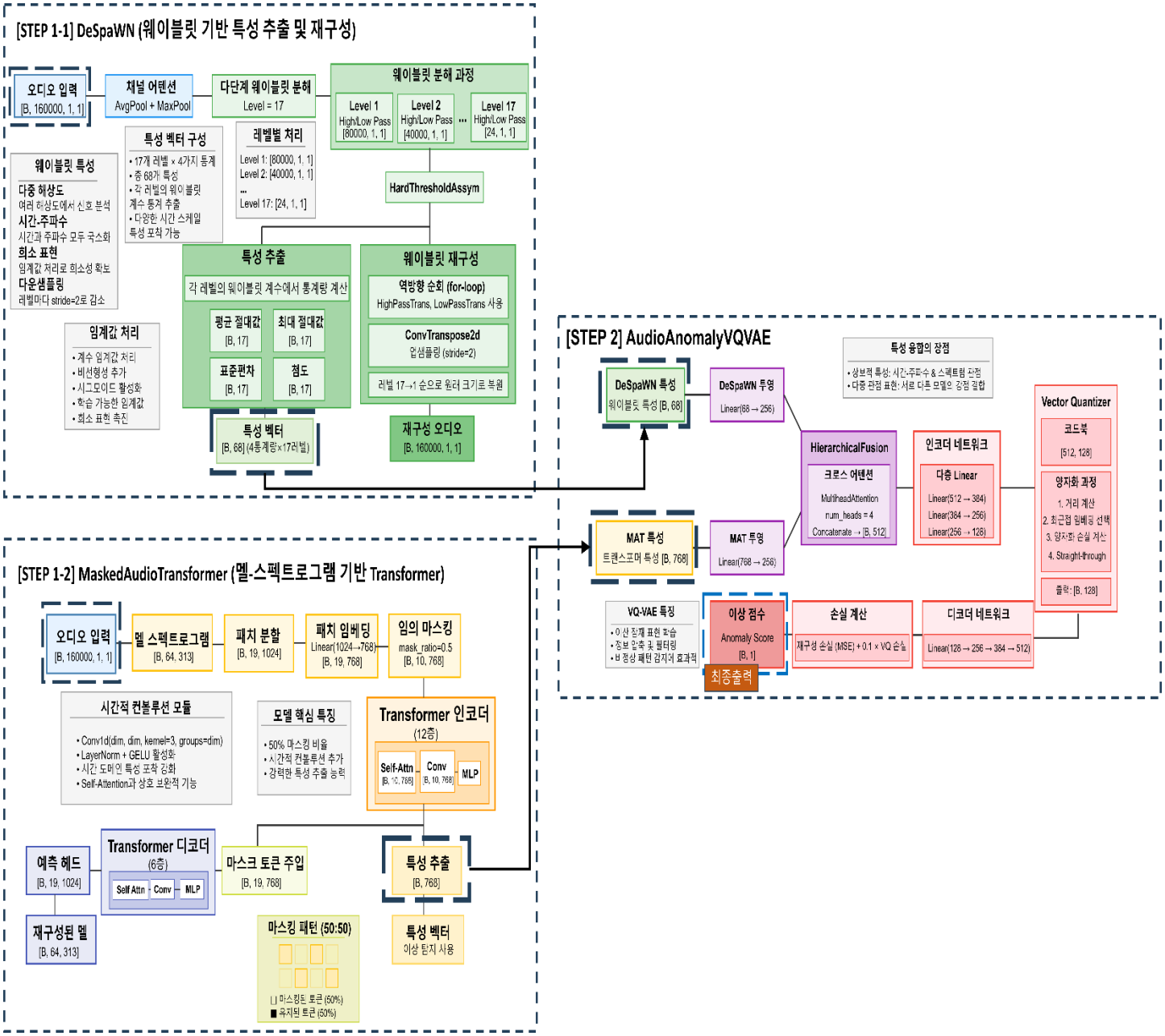

그림 1. 제안한 오디오 이상 탐지 파이프라인의 전체 구조

3. 방법

3.1 모델 개요

본 연구에서는 웨이브렛 변환 기반 신호 세부 패턴 추출과 마스킹 자기지도 트랜스포머 기반 전역 특징 학습, 그리고 VQ-VAE 기반 복원 및 이상검출 모듈을 결합한 새로운 모델 FusionWave를 제안한다[그림 1]. 모델은 2개의 단계로 구성된다. 첫번째 단계[STEP 1-1, STEP 1-2]에서는 입력으로 주어지는 원시 기계음 파형을 두 경로로 처리하여 특징을 추출하고, 두번째 단계[STEP 2]에서는 첫번째 단계에서의 특징을 잠재 공간에서 융합하여 복원한다. 학습은 정상 데이터에 대한 자기지도적 복원 과제(masked reconstruction)로 이루어지므로, 별도의 이상 레이블 없이도 정상 패턴을 모델링할 수 있다.

3.2 [STEP 1-1] DeSpaWN 변환

원시 1차원 오디오 신호 (16 kHz, 길이 160k 샘플)는 우선 완전 학습형 웨이브렛 패킷 변환인 DeSpaWN 인코더로 입력된다. DeSpaWN은 학습 가능한 웨이브릿 필터뱅크를 구현한 심층신경망으로, 일반적인 이산 웨이브릿 변환(DWT)의 계층적 여파망을 모사한다[5]. 각 인코딩 레벨에서는 두 개의 컨볼루션 필터를 사용하여, 입력 신호를 저역 성분과 고역 성분으로 병렬 분해한다. 하드임계값 활성화를 적용하여 웨이브릿 변환 계수들의 노이즈 성분을 제거하고 희소성을 높인다. 각 서브밴드에 대해 통계적 특성치(평균 에너지, 표준편차, 첨도 등)를 산출하여 웨이브릿 특징 벡터를 생성한다.

3.3 [STEP 1-2] 마스킹 기반 오디오 트랜스포머 인코더

입력 오디오 신호는 STFT 기반 멜-스펙트로그램 변환을 적용하고, 이를 패치(patch) 단위로 분할하여 트랜스포머의 입력 토큰으로 사용한다. 각 패치는 주파수 16 × 시간 16 프레임의 영역에 해당한다. 전체 멜-스펙트로그램 패치 중 50%정도를 무작위로 선택하여 마스킹 처리한다. 마스킹된 입력 시퀀스는 8개 레이어의 Transformer 인코더에 입력되어 은닉 표현을 산출한다. 인코더의 최종 출력에서 평균 풀링을 통해 하나의 특징 벡터를 추출한다.

3.4 [STEP 2] 특징 융합 및 VQ-VAE 기반 이상 점수 산출

웨이브릿 특징 벡터와 Transformer 특징 벡터는 연결(concatenation) 연산을 통해 하나의 통합 잠재 벡터를 구성한다. 이 벡터는 VQ-VAE의 인코더 부분에 입력되어 저차원 잠재공간으로 압축된다. VQ-VAE는 벡터 양자화를 통해 연속적인 잠재 표현을 이산적인 코드북 벡터로 변환한다. 선택된 코드북 벡터는 디코더로 전달되어 원본 입력의 스펙트로그램을 재구성한다. 학습은 입력 스펙트로그램 재구성 오차와 벡터 양자화 손실의 가중합을 최소화하는 방향으로 진행된다. 테스트 단계에서는 재구성 오류가 이상 점수로 사용되며, 정상 데이터는 작은 오류, 이상 데이터는 큰 오류를 보일 것으로 예상된다. 이를 통해 웨이브렛 필터, Transformer 가중치, VQ-VAE 인코더/디코더 및 코드북 등 모든 구성요소가 협력적으로 최적화된다. 학습 완료 후에는 정상 패턴을 높은 정확도로 재구성할 수 있게 되고, 테스트 단계에서 입력 신호의 이상 여부를 재구성 오류로 판단하게 된다.

표 2. MIMII 데이터셋에서 모델별 최고 ROC-AUC 평균 비교

Method	Fan	Pump	Slider	Valve
Baseline	0.658	0.729	0.848	0.663
VQ-VAE	0.739	0.788	0.794	0.522
DeSpaWN[2]	0.865	0.845	0.910	0.928
MAT[3]	0.792	0.681	0.873	0.564
GRLNet[6]	0.839	0.843	0.823	0.581
FusionWave	0.885	0.909	0.928	0.815

4. 실험 결과

[표 2]은 실험 결과를 Fan, Pump, Slider, Valve의 4가지 종류에 대해 각 ROC-AUC를 정리한 것이다. 제안 모델은 모든 기계에서 기준 모델보다 우수한 성능을 보였다. VQ-VAE만 사용한 경우 평균 AUC ~0.794, MAT만 사용한 경우 ~0.873, DeSpaWN만 사용한 경우 ~0.928을 기록했으나, 두 특징을 모두 융합한 경우 기존 보다 상승하였고, 결합 모델이 개별 모델을 능가함을 확인하였다. 또한 다른 기존 fusion방법인 GRLNet(평균 AUC ~0.843)과 비교하여도 높은 성능을 확인하였다.

5. 결론

본 논문에서는 MIMII 산업 기계음 데이터셋을 대상으로, DeSpaWN과 마스킹된 오디오 트랜스포머를 융합한 VQ-VAE 이상 탐지 모델을 제안하였다. 제안된 모델은 정상 기계음의 세부 특성과 장기 패턴을 동시에 학습함으로써 본 모델이 다양한 기계에서 기존 방법 대비 향상된 이상 탐지 정확도를 달성함을 확인하였다. 이러한 접근법은 이상 음향 데이터가 부족한 상황에서도 효과적인 모델 구축이 가능함을 보여주며, 향후 다른 산업 음향 데이터셋이나 다양한 이상 탐지 시나리오에도 활용될 수 있을 것으로 기대된다.

6. 참고문헌

[1] Suzuki, Yuma, et al. “Anomalous sound detection based on probabilistic modeling of normal sounds." DCASE 2019.

[2] Michau, Gabriel, and Olga Fink. “Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series.” PNAS 2021.

[3] Pengfei, Cai, et al. “MAT-SED: A Masked Audio Transformer with Masked-Reconstruction Based Pre-training for Sound Event Detection.” Interspeech 2024

[4] van den Oord, Aaron, Oriol Vinyals, and Koray Kavukcuoglu. “Neural Discrete Representation Learning.” In International Conference on Learning Representations (ICLR), 2018. arXiv:1711.00937v2.

[5] Purohit, Harsh, et al. "MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection", Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), pp. 209-213, 2019.

[6] Sha, Yu, et al. “Regional-Local Adversarially Learned One-Class Classifier for Anomalous Sound Detection in Global Long-Term Space.” KDD 2022