Deep Learning for Anomaly Detection: A Review

Author	GUANSONG PANG and CHUNHUA SHEN
read date	@2022년 3월 25일
Journal	ACM
PDF	[2021 ACM] Deep Learning for Anomaly Detection, A Review.pdf
Published Date	March 2021
detail	Review 논문 - Major Problem Complexities - Main Challenges Tackled - Categorization of Deep Anomaly Detection 1. Deep Learning for Feature Extraction 2. Learning Feature Representations of Normality 3. End-to-end Anomaly Score Learning
keyword	Anomaly Detection survey
link	https://dl.acm.org/doi/pdf/10.1145/3439950
status	Finished!

Abstract

- Anomaly Detection 조사를 아래와 같이 하였다.
  1. 포괄적 분류
  2. 발전 : 3 high-level 카테고리
  3. 방법론 : 11개 카테고리로 나눈

- 논의 내용
  1. key intuitions : 핵심 직관
  2. objective functions : 목적 함수
  3. underlying assumptions : 기본적 전제 조건
  4. advantages and disadvantages : 장점과 단점
  5. a set of possible future opportunities : 가능한 미래 기회들의 모음
  6. new perspectives on addressing the challenges

INTRODUCTION

- 이상치 탐지는 다수의 인스턴스 데이터로부터 매우 벗어난 인스턴스를 검출하는 프로세스라고 불리며, 1960년대 초기 연구로부터 수십 년 동안 활발하게(active) 연구되었다.
- 적용 도메인 : 넓은 분야에서 적용 필요성이 증가 되고 있다.

> 💡 risk management, compliance(규정 준수), security, financial surveillance(재정 감시), health and medical risk, and AI safety

- 다양한 커뮤니티 : 점점 더 중요한 역할을 하고 있다.

> 💡 data mining, machine learning, computer vision, and statistics

- 딥러닝은 다음과 같은 데이터에 대해 대단한 표현 학습을 보인다.

> 💡 high-dimensional, temporal(시간), spatial(공간), graph → pushing the boundaries of different learning tasks(다양한 학습들의 경계를 허물다)

- deep 이상치탐지 aim(목표) : NN를 통해 feature representations or anomaly scores 학습
  - 많은 deep anomaly detection이 실무에서 기존의 이상 탐지보다 훨씬 더 좋은 성능을 보임

- 5개의 주요 기여점 : Why we need deep anomaly detection?
  1. Problem nature and challenges (문제의 본질과 과제)
  2. Categorization and formulation (분류 및 공식화)
     a. deep learning for generic(일반) feature extraction
     b. learning representations of normality(정규성)
     c. end-to-end anomaly score learning : end-to-end 이상 점수 학습
  3. Comprehensive literature review(포괄적 문헌 리뷰)
  4. Future opportunities : discuss about implication(영향)
  5. Source codes and datasets : real anomalies data-set

ANOMALY DETECTION : PROBLEM COMPLEXITIES AND CHALLENGES

- 명백한 패턴 task와는 달리 이상치 감지는 소수, 예측 불가능/불확실성 및 희귀 이벤트를 해결하기에 모든(심층 또는 얕은) 감지 방법에 고유한 문제가 복잡해진다.

1. **Major Problem Complexities (주요 문제)**
   a. Unknownness (알려지지 않음)
   b. Heterogeneous anomaly classes (여러 종류의 클래스들)
   c. Rarity and class imbalance (희귀성과 불균형)
   d. Diverse types of anomaly (다양한 유형의 이상치)
      i. Point anomalies : 개별 인스턴스
      ii. Conditional anomalies : in a specific context (ex. 급격한 온도 강하/상승)
      iii. Group anomalies : 대규모로 발생 함 (ex. 가짜 계정들, 디도스 공격)

2. **Main Challenges Tackled by Deep Anomaly Detection (주요 과제)**
   a. CH1: Low anomaly detection recall rate. : 낮은 이상치 감지율 (이상치 수 자체가 적기 때문)
      - false positives ↓, detection recall rates ↑
      - significant expense of failing to spotting anomalies.
   b. CH2: Anomaly detection in high-dimensional and/or not-independent data : 고차원, 비독립
      - 저 차원에서 명백한 이상치의 특성을 볼 수 있지만, 고차원에서는 눈에 뛰지 않는다.
      - solution : subspace-based, feature selection-based methods
      - preserved proper information : 미지/이질성 때문에 보장이 어려움

- temporal, spatial, graph-based : 종속적인 인스턴스는 이상 감지가 어려움

c. CH3: Data-efficient learning of normality/abnormality : 데이터-효율적인 학습
   - fully supervised anomaly detection은 비현실적임 비용 多
   - Semi-supervised anomaly detection : normal data만 갖고 있음
   - unsupervised methods do not have any prior knowledge(사전 지식) of true anomalies.
     ㄴ assumption on the distribution of anomalies에 크게 의존하고 있다. 노이즈에 민감
   - weakly supervised anomaly detection : 정밀하지 않거나 부정확한 클래스 라벨이 있음

d. CH4: Noise-resilient anomaly detection : 노이즈에 강함
   - 레이블이 잘못된 데이터 또는 레이블이 지정되지 않은 이상치

e. CH5: Detection of complex anomalies : 복잡한 이상치 검출
   - 기존 방법론의 대다수인 Point anomalies를 Conditional & Group에 적용하기 어려움

f. CH6: Anomaly explanation : 이상치 설명
   - 블랙박스 모델은 리스크가 크다. 드문(=레어) 데이터를 학습하여 편향을 배울 수 있음
   - main challenge : well balance the model's interpretability and effectiveness : 검출 능력이 다가 아님

     → Ranking Model 유일하게 해결하며, 결론 5.를 보면 이 문제를 추는 전용 DL 모델 필요성에 대해 대두하고 있음

3. PROBLEM COMPLEXITIES AND CHALLENGES 정리

Table 1. Deep Learning Methods vs. Traditional Methods in Anomaly Detection

Method	End-to-end Optimization	Tailored Representation Learning	Intricate Relation Learning	Heterogeneity Handling
Traditional	×	×	Weak	Weak
Deep	✓	✓	Strong	Strong
Challenges	CH1-6	CH1-6	CH1, CH2, CH3, CH5	CH3, CH5

- Deep methods 쓰면 기존과 달리 end-to-end 최적화 & 특화된 표현 학습이 가능
- Intricate Relation Learning : 복잡한 관계 학습
- Heterogeneity Handling : 여러 다른 종류 데이터 다루는 것

ADDRESSING THE CHALLENGES WITH DEEP ANOMALY DETECTION

1. Preliminaries (예비)
   - Activation functions
     - linear, sigmoid, tanh, Rectified Linear Unit (ReLU)
   - layer
     - fully connected(MLP), convolutional + pooling(CNN), recurrent(RNN) layers
   - dataset

$$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

   - representation space

$$\mathbf{x}_i \in \mathbb{R}^D, \text{ let } \mathcal{Z} \in \mathbb{R}^K \ (K \ll N)$$

- feature representation mapping function : deep anomaly detection이 학습 목표로 하는 것

$$\phi(\cdot) : \mathcal{X} \mapsto \mathcal{Z}$$

  - 학습을 위해 각 인스턴스의 이상치 점수를 계산하기 위한 추가 단계 필요
- anomaly score learning function

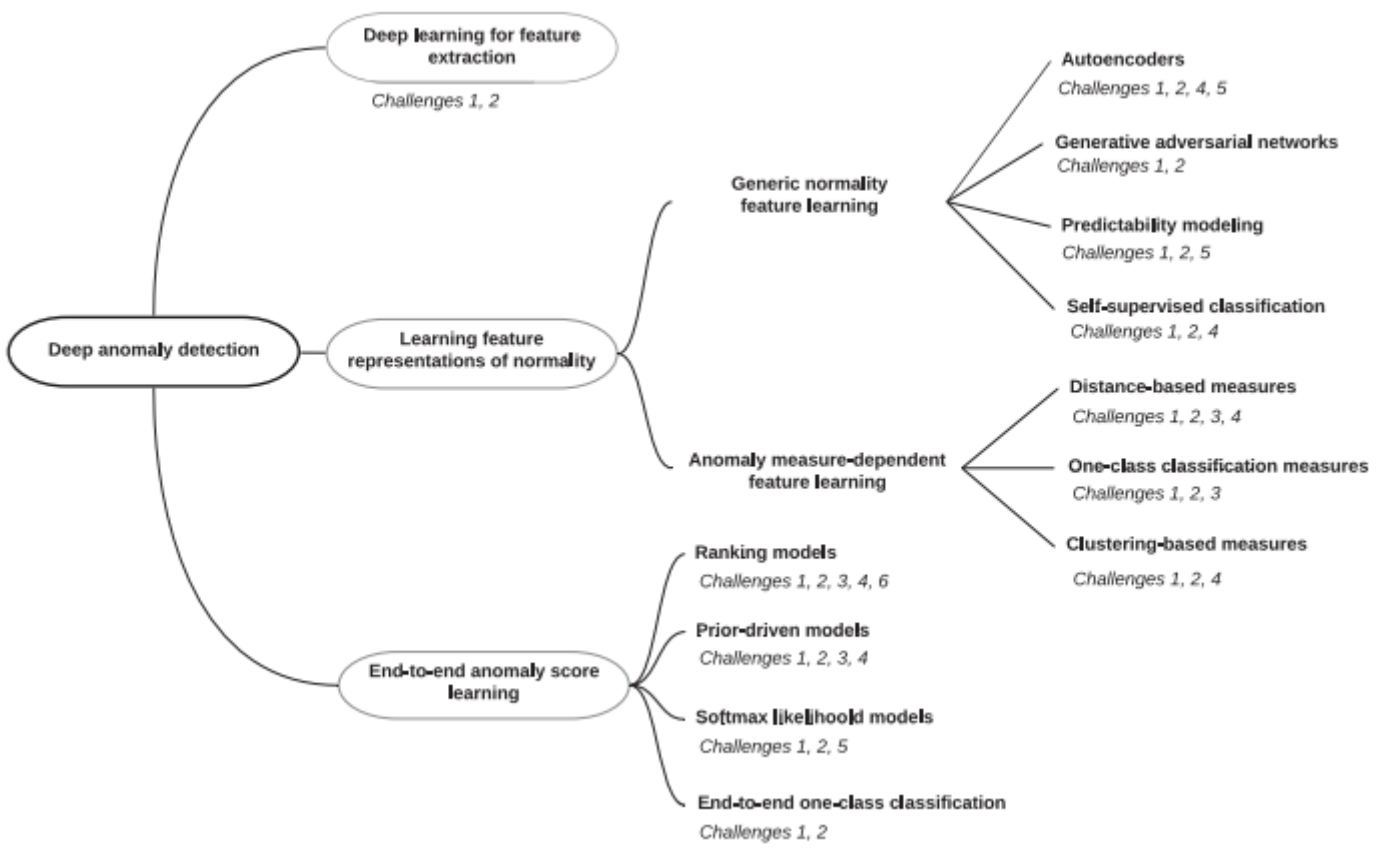
$$\tau(\cdot) : \mathcal{X} \mapsto \mathbb{R}$$

  - raw data(원시 데이터) 입력으로 직접 이상치 점수를 예측할 수 있음
  - 큰 τ는 더 큰 이상 정도를 나타냄
- ϕ 와 τ는 hidden layer 와 아래의 가중치 행렬이 있는 신경망 지원 mapping function이다.

$$\Theta = \{\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^H\}$$

2. Categorization of Deep Anomaly Detection (카테고리 화)

- 3개의 메인 카테고리와 11개의 세밀한 카테고리로 모델의 관점에서 분류함



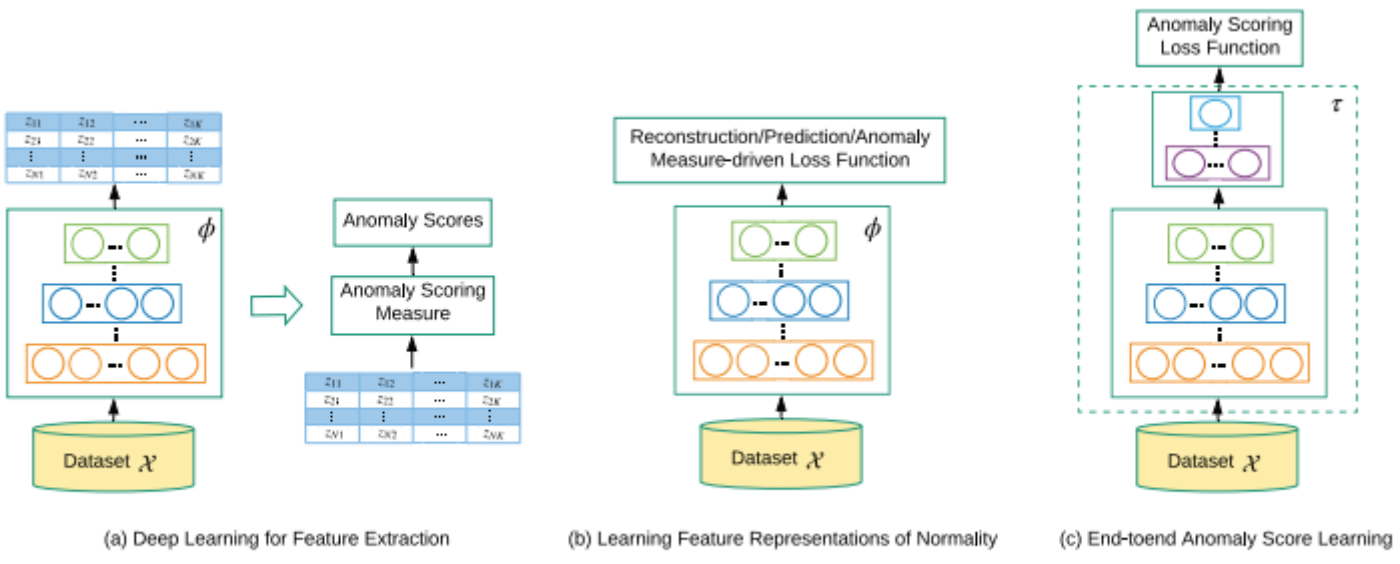


Fig. 2. Conceptual frameworks of three main deep anomaly detection approaches.

1) Deep Learning for Feature Extraction : hybrid의 경우 이 부분을 선 처리 후 ML 진행 됨

: independent(독립적) feature extractors only

2) Learning Feature Representations of Normality : 아래 2개 방법이 서로 의존적

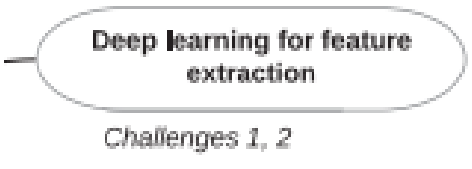
- 아래 2개의 방법은 거리 및 클러스터링 기반 측정(measure)을 하는 지에 따라 나뉜다.

1. Generic normality feature learning : 일반적인 정상 핏쳐 학습
2. Anomaly measure-dependent feature learning : : 비정상 종속-측정 핏쳐 학습

3) End-to-end Anomaly Score Learning

: end-to-end 방식으로 이상치 점수를 학습하는데 전념함

DEEP LEARNING FOR FEATURE EXTRACTION



- extract low-dimensional feature representations from high-dimensional and/or non-linearly separable data for downstream anomaly detection

  : 최종(downstram) 이상치 문제를 해결하기 위해 분리할 수 있는 비선형 데이터 또는 고차원으로부터 feature representation)을 저차원으로 뽑아낸다.

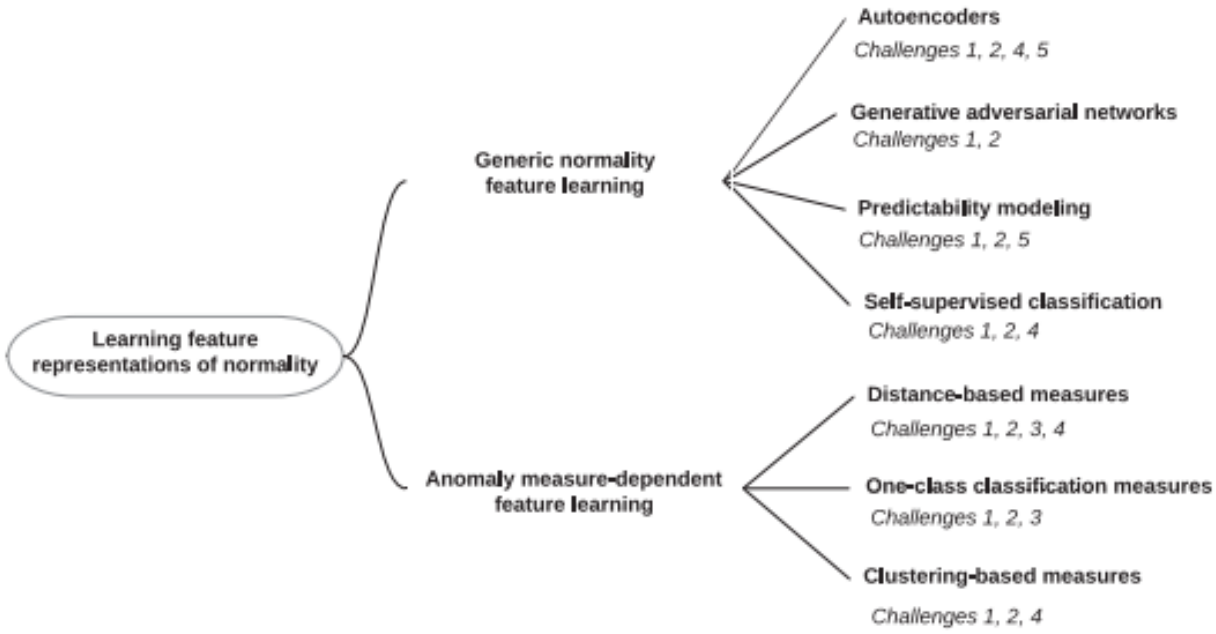
- feature extraion 과 anomaly scroing은 독립적으로 분리됨

  → 딥러닝 구성 요소는 순전히 차원축소로 작동함. anomaly scroing은 새로운 공간에 적용 산출

$$\phi : \mathcal{X} \mapsto \mathcal{Z} \qquad \mathcal{X} \in \mathbb{R}^D, \mathcal{Z} \in \mathbb{R}^K \quad D \gg K$$

- Assumptions.
  - The feature representations preserve discriminative information : 식별 정보 보존
- 차원 축소 방법(PCA, random projection)들에 비해 딥러닝 기법이 더 나은 능력을 갖고 있다.
- 종류 : AlexNet / VGG / ResNet / Unmasking framework / a linear one-class SVM
- Advantages
  1. A large number of state-of-theart (pre-trained) deep models and off-the-shelf anomaly detectors are readily available. : 바로 적용 가능한 DL model이 많다
  2. Deep feature extraction offers more powerful dimensionality reduction than popular linear methods. : 유명한 선형 방법보다 Deep feature extration이 차원 축소에 더 강하다.
  3. It is easy-to-implement given the public availability of the deep models and detection methods. : 감지 방법과 딥 모델의 공개를 고려했을 때 구현이 용이하다.
- Disadvantages.
  1. The fully disjointed feature extraction and anomaly scoring often lead to suboptimal anomaly scores. : 핏쳐 추출과 이상치 점수가 완전히 분리되어 최적이 아닌 점수가 나오는 경우가 많다.
  2. Pre-trained deep models are typically limited to specific types of data : 사전 훈련된 딥 모델은 일반적으로 특정 타입의 데이터로 제한된다.
- Challenges Targeted(대상 과제) :
  1. CH2(고차원/비독립성) : 고차원/비독립 데이터를 낮은 차원 공간에 투영하여 단순한 데이터 공간에서 AD 가능.
  2. CH1(낮은 이상치 감지율) : 여러 유형의 핏쳐를 활용하고 표현이 풍부한 검출 모델을 학습해 false positive를 줄이는데 도움이 됨

LEARNING FEATURE REPRESENTATIONS OF NORMALITY



1. Generic Normality Feature Learning

   : 일반 기능 학습 목적 함수를 최적화하여 데이터 인스턴스의 표현을 학습

   AE / GAN / Predictability modeling / SSL classfication

$$\{\Theta^*, \mathbf{W}^*\} = \arg\min_{\Theta,\, \mathbf{W}} \sum_{\mathbf{x}\in\mathcal{X}} \ell\Big(\psi\big(\phi(\mathbf{x};\Theta);\mathbf{W}\big)\Big),$$

$$s_{\mathbf{x}} = f(\mathbf{x}, \phi_{\Theta^*}, \psi_{\mathbf{W}^*}),$$

- 데이터 재구성, 생성 모델링, 예측 가능성 모델링, 자체 지도 분류 방법 등 여러 관점이 포함됨
  - ϕ : 표현 공간 Z에 원본데이터 x를 매핑함
  - ψ : W에 의해 파라미터화 되었으며, Z공간에서 동작하고, 데이터 규칙성 학습을 시행하는데 전념하는 대용 학습 작업이다.
  - ℓ : 기본 모델링 접근 방식 관련 손실 함수
  - f : ϕ 와 ψ를 확용하여 s_x를 계산하는 scoring funtion

- Autoencoders (AE)

  : 주어진 데이터가 잘 재구성 될 수 있는 일부 저 차원 기능 표현 공간을 배움. reconstrucion error를 최소화하기 위해 데이터의 중요 규칙성을 학습하도록 강제 됨. 이상치는 이런 에러가 큼

  - Assumption

    : 정상 인스턴스는 이상치보다 공간에서 더 잘 재구성 되고 압축된다.

  - network : encoding(기존 데이터 → 저차원 공간) + decoding (저차원 → 되돌리기 시도)
  - 이상데이터의 경우 reconstructed 잘 안됨(reconstrucion error를 anomlay score로 사용 가능)

$$\mathbf{z} = \phi_e(\mathbf{x};\Theta_e),\ \hat{\mathbf{x}} = \phi_d(\mathbf{z};\Theta_d),$$

    - ϕ_e : θ_e 를 가진 인코딩 네트워크, ϕ_d : θ_d를 가진 디코딩 네트워크

$$\{\Theta_e^*, \Theta_d^*\} = \arg\min_{\Theta_e,\Theta_d} \sum_{\mathbf{x}\in\mathcal{X}} \Big\|\mathbf{x} - \phi_d\Big(\phi_e(\mathbf{x};\Theta_e);\Theta_d\Big)\Big\|^2,$$

    - 인코딩과 디코더는 동일한 가중치 파라미터를 공유하여 파라미터 감소와 정규화할 수 있다.

$$s_{\mathbf{x}} = \Big\|\mathbf{x} - \phi_d\Big(\phi_e(\mathbf{x};\Theta_e^*);\Theta_d^*\Big)\Big\|^2,$$

    - s_x : x의 reconstruction error 기반의 이상치 점수

- AE 종류 : sparse/Denoising/Contractive/Variational/RandNet/RDA
- 복잡한 데이터에서 AE 사용 시

  : CNN-AE, LSTM-AE, Conv-LSTM-AE, graph convolutional network-AE
- Advantages.
  1. The idea of AEs is straightforward and generic to different types of data.

     : AE의 개념은 다양한 타입의 데이터에 대해 간단하고 일반적임
  2. Different types of powerful AE variants can be leveraged to perform anomaly detection. : 다양한 유형의 AE 변형 모델들이 존재
- Disadvantages.
  1. The learned feature representations can be biased : 편향될 수 있음
  2. 목적함수가 기본 규칙성의 일반적 요약이고 이상 검출 목적이 아니라서 불규칙성 검출에 최적화가 아님
- Challenges Targeted (Challenge : 해결 내용)
  1. CH2(고차원/비독립성) : 다양한 AE 아키텍쳐 사용으로 graph & multivariate sequence data에 적용 가능
  2. CH1(낮은 이상치 감지율) : 기존 방법론 보다 표현력이 높기 때문
  3. CH4(노이즈에 강함) : AE는 일반적으로 약함, RPCA+AE로 사용하면 강해질 수 있음

- Generative Adversarial Networks (GAN)
  - DAD로 유명함. latent space의 normality 잘 포착하는 것 목표.
  - Anomaly score = 실제 인스턴스와 생성된 인스턴스의 residual
  - 종류 : AnoGAN / EBGAN / BiGAN / ALAD / GANomaly / Wasserstein GAN / Cycle GAN
  - GAN 목적 함수 : min-max game

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_X}\Big[\log D(\mathbf{x})\Big] + \mathbb{E}_{\mathbf{z}\sim p_Z}\Big[\log\Big(1 - D\Big(G(\mathbf{z})\Big)\Big)\Big],$$

    - 왼쪽 : 원본 데이터 중에 여러 개를 뽑고 로그 후 평균 값을 취하겠다.
    - 오른쪽 : 노이즈 벡터를 뽑고 생성자에 넣어서 가짜 이미지를 만들고 그걸 판별한 뒤 확률 값을 로그 후 평균 값
    - D(discriminator), G(generator), V(value function), 각 x에 대해 최고의 z를 찾음
    - x의 GAN loss 기반의 이상치 점수

$$s_{\mathbf{x}} = (1-\alpha)\ell_R(\mathbf{x}, \mathbf{z}_{\gamma^*}) + \alpha\ell_{fm}(\mathbf{x}, \mathbf{z}_{\gamma^*}).$$

      - α : hyperparameter, γ^* : 마지막 단계
    - 손실 함수 1) residual loss

$$\ell_R(\mathbf{x}, \mathbf{z}_\gamma) = \Big\|\mathbf{x} - G(\mathbf{z}_\gamma)\Big\|_1,$$

    - 손실 함수 2) discrimination loss

$$\ell_{fm}(\mathbf{x}, \mathbf{z}_\gamma) = \Big\|h(\mathbf{x}) - h\Big(G(\mathbf{z}_\gamma)\Big)\Big\|_1,$$

      - γ = index of search iteration, h = feature mapping 수행 (D의 중간 레이어에서)
  - Advantages.
    1. 사실적 인스턴스을 생성하는 탁월한 능력으로 잠재 공간에서 재구성 되지 않는 비정상적인 인스턴스를 탐지할 수 있음

2. 많은 종류의 GAN 방법들이 이상치 탐지에 적용될 수 있음

- Disadvantages.
  1. 수렴 실패와 모드 붕괴와 같은 문제로 모델 훈련에 큰 어려움을 겪을 수 있음
  2. real data 분포가 복잡하거나 특이치가 포함된 경우, 생성 네트워크는 잘못된 길로 이끌 수 있다.
  3. data synthesis(합성)을 위해 설계되었기 때문에 AD에는 최선이 아닐 수 있다.
- Challenges Targeted (Challenge : 해결 내용)
  1. CH1(낮은 이상치 감지율) : 잠재 공간이 주요 이상치 판단 정보를 보존할 때, 기존 데이터 공간을 넘어 정확도 감지가 향상됨
  2. CH2(고차원/비독립성) : 학습된 저차원 잠재 공간의 reconstruction을 조사하여 고차원의 이상치들을 검출할 수 있음

- Predictability Modeling (예측가능 모델링)

  : 이전 인스턴스 표현을 문맥으로 사용하여 현재를 예측 (ex. video frames in a video sequence. 비디오 이상 검출에서 good)
  - 정상은 종속성에 잘 종속되어 있어 예측 가능, 비정상은 종종 종속성을 위반하여 예측 불가
  - Assumption.

    : 정상치는 일시적으로 예측이 용이
  - 비디오 프레임의 충분한 예측을 보장하기 위해 prediction objective function에서 모양과 동작 특징에 대한 다른 제약이 필요
  - 종류 : U-Net / autoregressive model
  - general objective function for video fram

$$\alpha \ell_{pred}\left(\hat{\mathbf{x}}_{t+1}, \mathbf{x}_{t+1}\right) + \beta \ell_{adv}\left(\hat{\mathbf{x}}_{t+1}\right),$$

    - ℓ_{pred} : frame prediction loss by mse, ℓ_{adv} : adversarial loss
    - $\hat{x}_{t+1}$: predict future frame, x_{t+1} : ground truth

$$\hat{\mathbf{x}}_{t+1} = \psi\left(\phi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t; \Theta); \mathbf{W}\right)$$

      - ψ: function for the frame generation
  - Peak Signal-to-Noise Ratio를 사용해 $\|x_i = \hat{x}_i\|_2$ 로 Anomaly score를 정의함
  - 프레임 예측을 더 높이기 위해 AE 기반의 reconstruction network가 추가되기도 한다.
  - Advantages.
    1. 다수의 시퀀스 학습 기법이 이 접근 방식에 적용 될 수 있음
    2. 이 접근 방식은 다양한 시간적, 공간적 의존성 학습할 수 있게 함
  - Disadvantages.
    1. 이 방식은 연속 데이터에서의 이상치 탐지에 한정된다.
    2. 연속 예측 계산 소요 多
    3. 순차적 예측을 기초하기 때문에 이상 검출에 차선일 수 있음
  - Challenges Targeted
    1. CH1 and CH2 : 고차원 및 시간 데이터의 이상 탐지의 false positive를 해결하는데 도움
    2. CH5(복잡한 이상치 검출) : 시간적 컨텍스트 기반 조건부 이상 탐지 가능

- Self-supervised Classification

  : 분류 모델을 생성하면서 정상의 표현을 배우고, 분류 모델이 일치 되지 않는 인스턴스를 이상치로 구분한다.
  - 전통적 기법인 Cross-Feature Analysis, feature models에 기반함

- 전통적인 모델은 원본의 tabular data를 사용하는 반면, deep consistency-based anomaly detection은 augmented 된 image data에 초점을 맞춰서 예측 모델을 구축함
- augmented instance를 효과적으로 판별하기 위해 훈련 데이터의 패턴 묘사가 중요
- Assumptions.
  - 정상 인스턴스는 자기지도 분류기에서 이상치 보다 더 강하게 일치 됨
- normality score loss function

$$L_{cons} = CE\Big(\psi\big(\mathbf{z}_{T_j};\mathbf{W}\big),\mathbf{y}_{T_j}\Big),$$

  - CE = standard cross-entropy loss function.
  - ψ = multi-class classifier parameterized with W
  - $y_{T\,j}$ = 변환 유형 T_j에 의해 argumented 된 synthetic class의 one-hot encoding
  - z = instace x가 T(변환 유형)에 따라서 argumented된 저차원 feature representation

$$\mathbf{z}_{T_j} = \phi\Big(T_j(\mathbf{x});\Theta\Big)$$

    - ϕ = 하이퍼파라미터를 통한 feature learner
- classification score 합산하여 anomaly score 계산(일관성 위해 가정 : Dirichlet distribution)
  - negative entropy-based anomaly scores perform better than average, maximum
- Advantages.
  1. They work well in both the unsupervised and semi-supervised settings.

     : 비지도, 준지도 학습에서 잘 작동됨
  2. Anomaly scoring is grounded by some intrinsic properties of gradient magnitude and its updating : 이상치 점수는 기울기 크기와 업데이트에 의해 기반 된다.
- Disadvantages
  1. above transformation은 이미지 데이터에만 적용되며, 종종 데이터에 의존적이다.
  2. classification model은 end-to-end 방식이기 때문에, an integrated module in the optimization(최적화 통합 모듈)이 아니라 분류 점수에 따라 도출됨으로 이상치 점수가 최적화가 되지 않을 수 있다.
- Challenges Targeted
  1. CH1 and CH2 : 정규성 표현적 저차원 표현이 기존 고차원 공간보다 이상 감지에 더 도움
  2. CH4(노이즈) : 정상과 이상 인스턴스 사이의 본질적 차이로 unsupervised에서 작동 가능하며, 노이즈에 대해 good robustness

2. Anomaly Measure-dependent Feature Learning

   : 기존의 anomaly measure 하나에 특별히 최적화된 feature representation을 학습하는 것을 목표로 함

   Distance-based Measure / One-class Classification-based Measure / Clustering-based Measure

- Distance-based Measure
  - Distance-based methods are straightforward and easy-to-implement.
  - 종류 : DB outliers / k-nearest neighbor distance / average k-nearest neighbor distance / relative distance

    → 차원의 저주 때문의 전통적 방법들은 고차원 데이터에서 잘 작동되지 않았음
  - deep distance-based anomaly detection은 저차원 공간 투영으로 한계를 잘 극복할 수 있음
    - 종류 : random nearest neighbor distance
  - Assumption.

        - 이상치는 정상치로부터 밀도 주변에서 멀리 떨어져 있다.
    - hinge loss function

$$L_{query} = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{A}, \mathbf{x}' \in \mathcal{N}} \max\left\{0, m + f(\mathbf{x}', \mathcal{S}; \Theta) - f(\mathbf{x}, \mathcal{S}; \Theta)\right\},$$

        - S = subset of x
        - A & N = Anomlay & Normal instance Set
        - m = $f(x, S; \theta)$에서 산출된 두 거리 사이의 여백 관련 정의 된 constant(상수)

$$f(\mathbf{x}, \mathcal{S}; \Theta) = \min_{\mathbf{x}' \in \mathcal{S}} \left\| \phi(\mathbf{x}; \Theta), \phi(\mathbf{x}'; \Theta) \right\|_2.$$

            - ϕ = projected representation space
    - Advantages.
        1. 거리기반 이상 탐지는 간단하고 이론이 잘 정의되어 있음. 그러므로 deep distance-based anomaly detection methods은 충분한 근거가 될 수 있다.
        2. 저차원 표현 공간에서 작동하고 기존에 실패한 고차원 데이터에 효과적일 수 있음
        3. 이상치에 특별하게 맞춰진 표현을 배울 수 있다.
    - Disadvantages.
        1. 광범위한 연산은 표현 학습 과정에서 거리 기반 이상 측정 통합에 장애가 될 수 있음
        2. distance-based anomaly measures 고유 약점으로 기능(capabilities)이 제한 될 수 있다.
    - Challenges Targeted.
        1. CH1 and CH2 : distance-based detection의 차원의 저주를 해결하고 저 차원 표현을 학습 할 수 있음
        2. CH3(효율적 학습) : 라벨링 된 몇 가지의 이상치를 이용하여 효과적 정규성 표현을 학습하기 위해 고안될 수 있다.
        3. CH4(노이즈) : pseudo 이상치 라벨의 이점은 잠재적 오염된 이상치에 로버스트하고 비지도 학습 환경에서 효과적으로 작동함

- One-class Classification-based Measure
    - one-class classification-based 맞춤화 된 핏쳐 표현 학습 목표
    - 종류 : one-class SVM / SVDD / Conventional one-class SVM / deep SVDD
    - 모든 정상치는 single class이며, 이상치에 일치하지 않는 간단한 모델로 요약할 수 있다.
    - NN와 one-class SVM을 결합하려는 수많은 시도가 있었다.
    - deep one-class SVM의 알반적 공식

$$\min_{r, \Theta, \mathbf{w}} \frac{1}{2}||\Theta||^2 + \frac{1}{\upsilon N} \sum_{i=1}^{N} \max\left\{0, r - \mathbf{w}^\intercal \mathbf{z}_i\right\} - r,$$

        - r = margin parameter
        - θ = 표현 네트워크의 parameters
        - ν = 훈련 데이터에서 이상 비율의 상한선으로 볼 수 있는 하이퍼파라미터
    - Advantages.
        1. The one-class classification-based anomalies는 학문적으로 잘 연구되어 왔고, deep one-class classification-based methods의 강한 기반을 제공한다.
        2. 표현 학습과 one-class classification models을 통합하여 맞춤형으로 최적의 표현을 학습 할 수 있다.
        3. one-class models에서 적절한 커널 함수를 수동으로 선택하지 않아도 된다.

- Disadvantages.
  1. The one-class models은 정상 클래스 안 복잡한 분포의 데이터셋에는 효과적으로 동작하지 않을 수 있다.
  2. 감지 성능은 one-class classification-based anomaly measures에 의존 된다.
- Challenges Targeted.
  1. CH1 and CH2 : 감지 정확도를 one-class classification models의 최적화된 저차원 표현 공간을 학습하여 향상 시킨다.
  2. CH3(효율적 학습) : 적게 라벨링된 정상과 이상치를 활용하여 단지 알려진 이상뿐 아니라 새로운 이상 클래스를 감지 할 수 있고 더 효과적으로 one-class 모델 학습을 위해 소수의 레이블 된 정상, 비정상 데이터를 사용할 수 있다.

- Clustering-based Measure

  : Deep clustering-based anomaly detection은 새롭게 학습된 표현 공간의 클러스터에서 분명하게 벗어난 이상치의 표현을 학습하는 것을 목표로 한다.

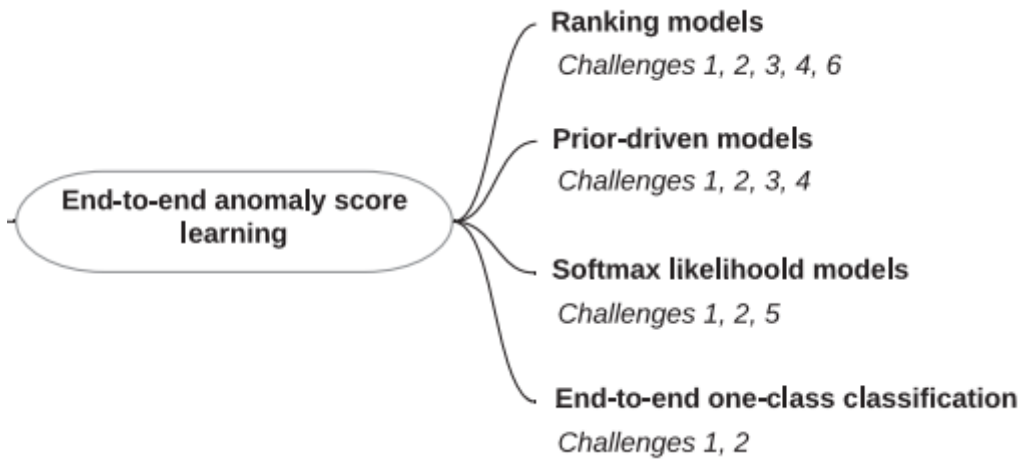
  - 클러스터링과 이상감지는 서로 연관되어 있으므로 클러스터 크기/중심까지의 거리/중심사이의 거리/구성원과 같은 이상치를 정의하기 위해 많은 연구가 있었다.
  - 전통적 방법 종류 : K-means, GMM, spectral clustering, agglomerative clustering
  - Assumptions.
    - 정상 인스턴스는 이상치보다 클러스터에 더 잘 준수한다.
    - 클러스터링 성능이 입력데이터에 크게 의존한다.
    - 주요 2개의 KEY
      1. 좋은 표현은 더 나은 클러스터링을 가능하게 하고, 클러스터링 결과가 좋으면 표현 학습에 효과적인 감시 신호를 제시할 수 있다.
      2. 하나의 클러스터링 알고리즘을 위해 최적화된 표현은 클러스터링 알고리즘에 의해 기본 가정의 차이점 때문에 다른 클러스터링 알고리즘에 반드시 유용하지 않다.
  - The deep clustering methods = forward pass(클러스터링 실행) + backward pass(클러스터 할당)
  - deep clustering methods loss function

$$\alpha \ell_{clu}\Big(f\left(\phi(\mathbf{x};\Theta);\mathbf{W}\right), y_x\Big) + \beta \ell_{aux}(\mathcal{X}),$$

    - ℓ_{clu} = clustering loss function, ϕ = feature learner parameterized by θ
    - f = clustering assignment function parameterized by W
    - y_x = pseudo class labels yielded(산출된) by the clustering
    - ℓ_{aux} = 손실 함수에 추가적인 제한을 주기 위해 사용되는 클러스터링이 아닌 손실 함수
    - α, β = 두 손실의 중요도를 통제하기 위한 hyperparameters
  - After the deep clustering, 클러스터 할당을 사용하여 f 함수 기반 Anomaly score 계산 가능
  - 훈련 데이터셋에서 오염이 생기면 편향이 될 수 있기 때문에 비지도 학습에서 잠재적 이상치의 영향 제거를 위해 몇 가지 추가적 제한이 필요하다.
  - reconstruction error-based handcrafted features는 이상 탐지에 deep 클러스터링 보다 결과표현이 더 적합하다.
  - Advantages.
    1. 많은 deep clustering methods을 AD에 사용 가능
    2. 기존의 클러스터링 방법들과 비교하여 deep clustering-based methods은 복잡한 데이터 세트의 이상치를 쉽게 발견할 수 있도록 특별히 최적화된 표현을 학습함
  - Disadvantages.
    1. 클러스터링 결과에 따라 이상 감지 성능이 크게 의존된다.
    2. 클러스터링 과정이 훈련 데이터 중 오염된 이상치에 의해 편향 되었을 수 있다.
  - Challenges Targeted.

1. CH1 and CH2 : The clustering-based anomaly measure은 입력된 데이터의 새롭게 학습된 저차원 표현들에 적용된다. 판별 정보가 충분히 표현 공간에 보존된다면 기존 데이터 공간보다 좋은 감지 정확도를 달성할 수 있음
3. CH4(노이즈) : 어떤 클러스터링 알고리즘은 이상치들에 민감하기에 딥 클러스터링과 순차적 이상 탐지는 주어진 데이터가 이상치에 의해 오염될 때, 크게 잘못된 길로 이끌 수 있다. AE의 재건설 에러로부터 수 작업 된 핏쳐가 사용된 Deep clustering은 오염을 포함한 보다 로버스트한 모델을 학습하는데 도움이 될 수 있다.

END-TO-END ANOMALY SCORE LEARNING



- end-to-end 방식으로 스칼라 AD score를 학습하는 것을 목표로 한다.
- end-to-end anomaly score learning network:

$$\Theta^* = \arg\min_{\Theta} \sum_{\mathbf{x}\in\mathcal{X}} \ell\Big(\tau(\mathbf{x};\Theta)\Big),$$

$$s_{\mathbf{x}} = \tau(\mathbf{x};\Theta^*).$$

1. Ranking Models
   - ranking model을 직접 학습하는 것을 목표로 하며, 이상치 점수 신경망은 관측 가능 순서형 변수에 의해 실행된다.
   - Assumptions.
     - 데이터 이상을 포착한 관측 가능한 순서형 변수가 존재한다.
     - anomaly scoring neural network을 실행하기 위해 순서형 regression-based loss function이 고안되다. (self-trained deep ordinal regression model for unsupervised video AD)
     - self-trained deep ordinal regression model의 objective function

$$\arg\min_{\Theta} \sum_{\mathbf{x}\in\mathcal{G}} \ell\Big(\tau(\mathbf{x};\Theta), y_{\mathbf{x}}\Big),$$

     - Advantages.
       1. 이상치 점수는 조정된 손실 함수를 사용하여 즉시 최적화 할 수 있다.
       2. 이들은 보통 이상과 정상의 경우 사이에 순서 형의 약한 가정을 함으로 이상치에 대한 정의에서 자유롭다.
       3. Ranking Models 접근 방식은 순위 매기기 학습과 같은 분야에서 확립된 순위 매기기 기술과 이론을 기반으로 할 수 있다.
     - Disadvantages.
       1. 라벨이 지정된 이상치들이 필요하며, 라벨이 지정된 이상치를 사용할 수 없는 응용(application)에는 적용되지 않을 수 있다.
       2. 모델은 소수의 라벨링 된 이상 징후를 감지하는 데만 적합되기 때문에 라벨링 된 이상 징후에 대해 다른 비정상적인 특징을 나타내는 보이지 않는 이상 징후는 일반화하지 못할 수 있다.
     - Challenges Targeted

1. CH1 and CH2 : 임의 라벨이나 노이즈 라벨과 같은 weak supervision을 사용하면. 의심 가는 이상치의 중요한 지식을 제공하여 보다 표현적인 저 차원 표현 공간과 더 나은 탐지 정확도를 학습할 수 있게 한다.
2. CH3(효율적 학습) : The MIL scheme & pairwise relation prediction은 거칠고/제한적인 이상치 라벨을 통합할 수 있는 쉬운 방법을 제시함
3. CH4(노이즈) : 데이터의 노이즈 라벨 또는 오염된 이상치가 Deep weakly-supervised & Self-trained deep ordinal regression for end-to-end video 방법에서 잘 작동한다.
4. CH6(이상 설명) : the end-to end anomaly score learning은 이상치 활성화 가중치 또는 이상 점수의 기울기를 역전파하여 큰 이상 점수를 담당하는 특징을 찾아냄으로서 간단한 이상 설명을 제공함.

2. Prior-driven Models
   - 사전 분포(Prior distribution)를 사용하여 anomaly score learning을 인코딩하고 구동함
   - 이상 점수가 end-to-end 로 학습 되기 때문에, 우선은 내부 모듈 또는 score learning function τ의 산출물 중 하나에 부과될 수 있음
   - 종류 : Bayesian inverse reinforcement learning (IRL)
   - Assumptions.
     - 적용된 우선은 데이터셋의 근본적인 이상/정상을 포착한다.
   - IRL optimization function

$$\max_{\Theta} \mathbb{E}_{\mathbf{s}\sim\mathcal{S}}\Big[\log p(\mathbf{s}|\Theta) + \log p(\Theta)\Big],$$

     - latent reward function parameterized by θ

$$p(\mathbf{s}|\Theta) = \frac{1}{Z}\exp\Big(\sum_{(o,a)\in\mathbf{s}} \tau_{\Theta}(o,a)\Big),$$

       - $\tau_\theta(o,a)$ = latent reward function parameterized by θ
       - $\theta, (o,a)$= a pair of state and action in the sequence s
       - Z = 마크코프 결정 프로세스 역학과 일치하는 모든 시퀀스에 대한 $\exp(\sum_{(o,a\in s)} t_\theta(o,a))$의 적분의 분할 함수
       - $p(\theta) = \theta$ 의 사전 분포
       - S = 관찰된 시퀀스의 집합
   - prior은 contrastive loss를 정의하는데 활용 됨
   - Advantages.
     1. 이상 점수는 주어진 사전(given prior)으로 최적화 될 수 있다.
     2. 다양한 사전 분포를 이상 점수 학습에 통합하기 위한 유연한 프레임 워크 제공. 다양한 Bayesian deep learning techniques가 적용 될 수 있음
     3. 사전은 다른 방법보다 더 해석 가능한 이상 점수를 초래할 수 있음
   - Disadvantages.
     1. 다양한 이상 감지 적용 시나리오에 대해 보편적으로 효과적 사전 설계를 하는 것이 불가능 하지는 않지만 어렵습니다.
     2. 모델은 사전이 분포에 잘 맞지 않는 다면 덜 효과적으로 작동할 수 있다.
   - Challenges Targeted
     - CH1 and CH2 : 사전은 고차원 및 시퀀스 같은 다양한 복잡한 데이터의 저차원 표현을 모델이 학습할 수 있도록 지원한다.
     - CH1 and CH3(낮은 감지율 효율적 학습) : 이상 점수에 대해 사전을 부과함으로, 한정된 양의 라벨된 이상 데이터를 활용해 정상과 이상 표현을 향상 시켜 감지율을 큰폭으로 올리는 성능을 나타낸다.
     - CH4(노이즈) : 감지 모델은 이상 점수 함수의 사전 분포에 의해 구동 되고, 훈련 데이터의 이상인 오염된 데이터에서 잘 작동한다.

3. Softmax Likelihood Models

- 훈련 데이터의 사건 발생 가능도를 극대화하여 이상 점수를 학습하는 것을 목표한다.
- 정상치는 높은 확률 사건으로 추정되고, 이상치는 낮은 확률의 사건인 경향이 있다.

  → -사건 가능도(=the negative of the event likelihood)는 자연히 이상 점수로 정의된다.
- 종류 : Softmax likelihood(via noise contrastive estimation (NCE))
- Assumptions.
  - 이상과 정상 인스턴스는 각각 낮은 확률과 높은 확률의 사전이다.
- 사건 가능도를 모델링하여 이상 점수를 학습하는 함수

$$\Theta^* = \arg\max_{\Theta} \sum_{\mathbf{x} \in X} \log p(\mathbf{x}; \Theta),$$

  - $p(x;\theta)$ = 학습될 파라미터 θ인 인스턴스 x의 확률 = is modeled with a softmax function

$$p(\mathbf{x}; \Theta) = \frac{\exp\left(\tau(\mathbf{x}; \Theta)\right)}{\sum_{\mathbf{x} \in X} \exp\left(\tau(\mathbf{x}; \Theta)\right)},$$

  - $\tau(x;\theta)$ = 핏쳐 쌍으로 이루어진 상호작용을 포착하도록 설계된 이상 스코어링 함수

$$\tau(\mathbf{x}; \Theta) = \sum_{i,j \in \{1,2,\ldots,K\}} w_{ij} \mathbf{z}_i \mathbf{z}_j,$$

    - z = x의 i번째 핏쳐 값의 저차원 임베딩
    - w_{ij} = 상호작용에 가중치를 더한 것으로 트레이닝 가능한 파라미터
- 주로 범주형 데이터에서 이상을 탐지하도록 설계 됨
- Advantages.
  1. 상호작용의 다양한 종류가 이상 점수 학습 프로세스로 통합될 수 있다.
  2. 이상 점수는 우리가 포착하려는 특정 이상 상호작용 관하여 충실하게 최적화 된다.
- Disadvantages.
  1. 상호작용의 계산은 각 데이터의 인스턴스의 핏쳐/요소가 많은 경우 많이 소요 될 수 있음
  2. 이상 점수 학습은 네거티브 샘플 생성 품질에 크게 좌우된다.
- Challenges Targeted
  1. CH2 and CH5(고차원/비독립 and 복잡) : 여러 종류의 데이터 소스를 가진 데이터셋의 저차원 표현을 학습할 수 있는 유망한 방법 제공
  2. CH1(낮은 감지율) : 학습된 표현은 기존 방법보다 더 많은 정상/비정상성 정보를 포착함으로 더 나은 감지를 할 수 있게 한다.

4. End-to-end One-class Classification
   - end-to-end 방식에서 인스턴스가 정상인지 아닌지를 판별하는 방법을 학습하는 one-class classifier를 훈련하는 것을 목표로 함
   - ex) GAN + one-class classification 즉, adversarially learned one-class classification
   - GAN-based methods는 우선 실제 데이터 분포와 근사하게 생성 분포를 학습하는 것을 목표로 하며, 정규 인스턴스와 적대적으로 생성된 인스턴스를 분리하기 위해 차별적 모델을 최적화하는 것을 목표로 함
   - 종류 : adversarially learned one-class classification(ALOCC), One-class adversarial networks (OCAN)
   - Assumptions.
     1. 이상치에 근사된 데이터 인스턴스는 효과적으로 합성할 수 있다.
     2. 모든 정규 인스턴스를 discriminative one-class model로 요약할 수 있다.
   - The one-class model은 판별자 네트워크에 구축되고, 생성자 네트워크는 노이즈 제거 AE에 기반한다.

- The objective of the AE-empower GAN

$$\min_{AE}\max_{D} V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_X}\Big[\log D(\mathbf{x})\Big] + \mathbb{E}_{\hat{\mathbf{x}}\sim p_{\hat{X}}}\Big[\log\Big(1 - D\Big(AE(\hat{\mathbf{x}})\Big)\Big)\Big],$$

  - $p_{\hat{X}}$ = 가우시안 노이즈에 의해 손상된 X의 데이터 분포
  - 위의 함수는 아래의 data construction error in AE와 같이 최적화 됨

$$\ell_{ae} = \|\mathbf{x} - AE(\hat{\mathbf{x}})\|^2.$$

- Advantages.
  1. 이상 분류 모델은 end-to-end 방식으로 적대적으로 최적화된다.
  2. 적대적 학습과 one-class 분류기의 풍부한 기술과 이론으로 발전, 지원될 수 있다.
- Disadvantages.
  1. 알려지지 않은 이상치와 유사하게 생성을 보증 하는건 어려울 수 있다.
  2. GAN의 불안정성으로 인해 댜양한 품질을 가진 인스턴스가 발생하여 이상 분류 성능이 불안정해질 수 있다.
  3. 준지도 이상치 탐지 시나리오 응용에 제한된다.
- Challenges Targeted
  1. CH1 and CH2(낮은 감지율과 고차원 비독립) : 적대적 학습된 one-class classifiers은 현실적인 경계 인스턴스를 생성하는 방법을 학습하고, 저차원 정규성 표현을 학습 가능하게 한다.

ALGORITHMS AND DATASETS

1. Representative Algorithms (30가지 대표적인 알고리즘의 주요 특징)

Table 2. Key Characteristics of 30 Representative Algorithms

Method	Ref.	Sup.	Objective	DA	DP	PT	Archit.	Activation	# layers	Loss	Data
OADA	[65] (4)	Semi	Reconstruction	Yes	No	No	AE	ReLU	3	MSE	Video
Replicator	[57] (5.1.1)	Unsup.	Reconstruction	No	No	No	AE	Tanh	2	MSE	Tabular
RandNet	[29] (5.1.1)	Unsup.	Reconstruction	No	Yes	Yes	AE	ReLU	3	MSE	Tabular
RDA	[175] (5.1.1)	Semi	Reconstruction	No	No	No	AE	Sigmoid	2	MSE	Tabular
UODA	[91] (5.1.1)	Semi	Reconstruction	No	No	Yes	AE & RNN	Sigmoid	4	MSE	Sequence
AnoGAN	[138] (5.1.2)	Semi	Generative	No	No	No	Conv.	ReLU	4	MAE	Image
EBGAN	[170] (5.1.2)	Semi	Generative	No	No	No	Conv. & MLP	ReLU/lReLU	3-4	GAN	Image & Tabular
FFP	[86] (5.1.3)	Semi	Predictive	Yes	No	Yes	Conv.	ReLU	10	MAE/MSE	Video
LSA	[1] (5.1.3)	Semi	Predictive	No	No	No	Conv.	lReLU	4-7	MSE & KL	video
GT	[48] (5.1.4)	Semi	Classification	Yes	Yes	No	Conv.	ReLU	10-16	CE	Image
E^3Outlier	[157] (5.1.4)	Semi	Classification	Yes	Yes	No	Conv.	ReLU	10	CE	Image
REPEN	[112] (5.2.1)	Unsup.	Distance	No	No	No	MLP	ReLU	1	Hinge	Tabular
RDP	[155] (5.2.1)	Unsup.	Distance	No	No	No	MLP	lReLU	1	MSE	Tabular
AE-1SVM	[104] (5.2.2)	Unsup.	One-class	No	No	No	AE & Conv.	Sigmoid	2-5	Hinge	Tabular & image
DeepOC	[161] (5.2.2)	Semi	One-class	No	No	No	3D Conv.	ReLU	5	Hinge	Video
Deep SVDD	[132] (5.2.2)	Semi	One-class	No	No	Yes	Conv.	lReLU	3-4	Hinge	Image
Deep SAD	[133] (5.2.2)	Semi	One-class	No	No	Yes	Conv. & MLP	lReLU	3-4	Hinge	Image & Tabular
DEC	[162] (5.2.3)	Unsup.	Clustering	No	Yes	Yes	MLP	ReLU	4	KL	Image & Tabular
DAGMM	[179] (5.2.3)	Unsup.	Clustering	No	Yes	No	AE & MLP	Tanh	4-6	Likelihood	Tabular
SDOR	[117] (6.1)	Unsup.	Anomaly scores	No	No	Yes	ResNet & MLP	ReLU	50 + 2	MAE	Video
PReNet	[114] (6.1)	Weak	Anomaly scores	Yes	No	No	MLP	ReLU	2-4	MAE	Tabular
MIL	[145] (6.1)	Weak	Anomaly scores	No	Yes	Yes	3DConv. & MLP	ReLU	18/34 + 3	Hinge	Video
PUP	[107] (6.2)	Unsup.	Anomaly scores	No	No	No	MLP	ReLU	3	Likelihood	Sequence
DevNet	[115] (6.2)	Weak	Anomaly scores	No	No	No	MLP	ReLU	2-4	Deviation	Tabular
APE	[30] (6.3)	Unsup.	Anomaly scores	No	No	No	MLP	Sigmoid	3	Softmax	Tabular
AEHE	[45] (6.3)	Unsup.	Anomaly scores	No	No	No	AE & MLP	ReLU	4	Softmax	Graph
ALOCC	[135] (6.4)	Semi	Anomaly scores	Yes	No	No	AE & CNN	lReLU	5	GANs	Image
OCAN	[174] (6.4)	Semi	Anomaly scores	No	No	Yes	LSTM-AE & MLP	ReLU	4	GANs	Sequence
Fence GAN	[103] (6.4)	Semi	Anomaly scores	No	Yes	No	Conv. & MLP	lReLU/Sigmoid	4-5	GANs	Image & Tabular
OCGAN	[120] (6.4)	Semi	Anomaly scores	No	No	No	Conv.	ReLU/Tanh	3	GANs	Image

DA, DP, PT, and Archit. are short for data augmentation, dropout, pre-training, and architecture, respectively. # layers account for all layers except the input layer. lReLU represents leaky ReLU.

a. 대부분의 방법은 비지도 또는 준지도 모드로 운영됨

b. 데이터 증강, 드롭 아웃 및 사전 훈련과 같은 딥러닝 트릭이 충분히 연구되지 않았음

c. 네트워크 아키텍처가 깊지 않다. 대부분 방법이 5개 이하의 네트워크 레이어를 가짐

d. ReLu는 가장 인기 있는 활성화 함수이다

e. 다양한 타입의 인풋 데이터를 처리하기 위해 다양한 backbone 네트워크 사용 가능.

2. Datasets with Real Anomalies : 실제 이상치에 대한 공개 가능 21개의 실제 데이터 세트

Table 3. 21 Publicly Accessible Real-world Datasets with Real Anomalies

Domain	Data	Size	Dimension	Anomaly (%)	Type	Reference
Intrusion detection	KDD Cup 99 [13]	4,091-567,497	41	0.30%-7.70%	Tabular	[57, 103, 104, 179]
Intrusion detection	UNSW-NB15 [100]	257,673	49	≤9.71%	Streaming	[114, 115]
Excitement prediction	KDD Cup 14	619,326	10	6.00%	Tabular	[114, 115]
Dropout prediction	KDD Cup 15	35,091	27	0.10%-0.40%	Sequence	[91]
Malicious URLs detection	URL [93]	2.4m	3.2m	33.04%	Streaming	[112]
Spam detection	Webspam [160]	350,000	16.6m	39.61%	Tabular/text	[112]
Fraud detection	Credit-card-fraud [34]	284,807	30	0.17%	Streaming	[114, 115, 174]
Vandal detection	UMDWikipedia [76]	34,210	N/A	50.00%	Sequence	[174]
Mutant activity detection	p53 Mutants [13]	16,772	5,408	0.48%	Tabular	[112]
Internet ads detection	AD [13]	3,279	1,555	14.00%	Tabular	[112]
Disease detection	Thyroid [13]	7,200	21	7.40%	Tabular	[114, 115, 133, 179]
Disease detection	Arrhythmia [13]	452	279	14.60%	Tabular	[116, 133, 179]
Defect detection	MVTec AD	5,354	N/A	35.26%	Image	[15]
Video surveillance	UCSD Ped 1 [81]	14,000 frames	N/A	28.6%	Video	[117, 161]
Video surveillance	UCSD Ped 2 [81]	4,560 frames	N/A	35.9%	Video	[117, 161]
Video surveillance	UMN [106]	7,739 frames	N/A	15.5%-18.1%	Video	[117]
Video surveillance	Avenue [90]	30,652 frames	N/A	12.46%	Video	[161]
Video surveillance	ShanghaiTech Campus	317,398 frames	N/A	5.38%	Video	[86]
Video surveillance	UCF-Crime	1,900 videos (13.8m frames)	N/A	13 crimes	Video	[145]
System log analysis	HDFS Log [164]	11.2m	N/A	2.90%	Sequence	[40]
System log analysis	OpenStack log	1.3m	N/A	7.00%	Sequence	[40]

CONCLUSIONS AND FUTURE OPPORTUNITIES

- 이상치 감지를 위한 딥러닝 기술 활용에 대한 12가지의 다양한 모델링 관점을 검토 했다.

1. Exploring Anomaly-supervisory Signals : 이상 감지 신호 조사
   - 5.1 : objective functions이 일반적이지만, 이상치 감지에 특히 최적화 되지 않는 문제
   - 5.2 : 제약을 가함으로 위의 문제를 해결하는데 도움을 주려 했지만, 한계가 있음
   - data reconstruction and GANs의 형식을 벗어나 이상 분포의 약한 가정을 가진 새로운 소스를 탐색하는 것이 중요하다.

2. Deep Weakly Supervised Anomaly Detection
   - Deep weakly supervised anomaly detection

     : 일부/부정확/불확실하게 라벨이 부착된 이상 데이터 즉, weakly supervised anomaly signals를 deep neural networks를 활용하여 이상-정도 감지 모델을 학습하려고 함.
   - unknown anomaly detection하는 모델을 구축하는 것을 목표로 함
   - data-efficient anomaly detection & few-shot anomaly detection을 발전시키는 것이 중요함

3. Large-scale Normality Learning
   - Large-scale 비지도 representation learning은 downstream learning tasks에서 거대한 성공을 거두었다.
   - 2.1 : 충분한 라벨된 데이터를 얻기 힘든 이상치 탐지인 학습 과제에서 특히 중요하다.
   - 목표 : 비지도 학습에서 레이블이 없는 대규모 데이터를 통해 표현 학습 모델을 사전 학습하고, 준지도 학습을 통해 감지 모델을 미세 조정하는 것이다. (단 데이터에 오염이 없어야 함)

4. Deep Detection of Complex Anomalies
   - 대부분의 이상치 감지 방법들은 Point anomalies가 연구되어 왔고, conditional/group anomalies은 덜 연구되어졌다.
   - 딥러닝은 복잡한 시간적/공간적 의존과 unordered data points 세트의 표현을 포착하고 학습하는 뛰어난 능력을 가졌다. → 새로운 NN 또는 objectives functions가 필요할 수 있음
   - Multimodal anomaly detection은 미개척 연구 영역이다.

- 다른 데이터 소스 표현을 연결하여 통합 표현을 학습하는 등 deep 접근은 multimodal anomaly detection에 중요한 기회를 제공한다.

5. Interpretable and Actionable Deep Anomaly Detection
   - 해석 가능 & 실행 가능 DAD는 결과와 모델 결정 이해를 위해 필수적이다.
   - 이상을 가장 비상적으로 만드는 핏쳐의 서브셋을 찾아서 이상치 설명 문제를 푸는 연구가 있었다.
   - 하지만 해석성과 조치성이 약하기 때문에 이상 설명을 제공하는 고유 기능을 가진 심층 모델이 필요함

6. Novel Applications and Settings (새로운 응용 및 설정)
   a. out-of-distribution (OOD) detection

      : 훈련 분포로부터 먼 인스턴스를 감지하는 방법으로 ML system이 실제 환경에 새로운 클래스 인스턴스를 처리할 수 있도록 하는 필수 기술이다. OOD를 통해 이상 감지 뿐만 아닌 세분화된 정상 등급 클래스 라벨을 사용할 수 있다고 가정한다.

   b. curiosity learning

      : 보너스 보상 함수를 강화 학습에서 희박한 보상과 함께 배우는 것
      - 강화학습은 보상이 희박한 환경에서 잘 작동하지 않기 때문에 보너스 보상을 통해 환경을 강화함으로써 강화학습 문제를 해결한다.
      - 보너스 보상은 신규성 또는 희귀성에 기초하여 정의 됨. 즉 novel/rare states를 발견하면 큰 보너스 보상을 받음

   c. non- independent and identically distributed (IID) : 독립적/동일한 분포가 아닌
      - 실제 인스턴스들은 non-IID 하다. 이러한 인스턴스를 학습하기 위해 non-IID anomaly detection이 필요함.
      - 복잡한 상황에서 non-IID 특성을 고려하는 것은 매우 중요하다.

   d. other interesting applications
      - detection of adversarial examples
      - anti-spoofing in biometric systems : 생체 인식 시스템의 스푸핑(거짓 IP) 방지
      - early detection of rare catastrophic events : 희귀 재앙적 사건에 대한 조기 감지
        - e.g., financial crisis(금융 위기) and other black swan events(흑조 이론)

	특징	예시
머신러닝	1. 비정형 데이터에 대해서 작동할 수 없음 2. SVM/Tree 등 이미 많은 anomaly detection 방법론이 개발되었고, 정형 데이터에 대해 준수한 성능을 보장함	1. OC-SVM, SVDD 2. Isolation Forest 3. Clutering
딥러닝	1. 비정형 데이터 혹은 시계열 데이터에 대해서도 anomaly detection을 수행할 수 있음 2. 모델 구조 및 학습을 위한 loss 함수를 어떻게 설계하는지가 관건이라고 할 수 있음	1. AE 계열 2. Word2Vec 계열 3. GAN 계열 4. Deep SVDD
Hybrid	1. 딥러닝 모델을 feature extractor로 활용하여 비정형 데이터에 대해서도 머신러닝 모델을 적용할 수 있도록함 2. End-to-End 학습이 불가능하기 때문에 엉뚱한 feature로 인해 성능이 저하될 우려가 있음	1. AE 계열 + 머신러닝 2. Word2Vec 계열 + 머신러닝

- 비정형 데이터 : Vision, NLP
- 딥러닝은 인풋 형태가 상관없다
- 하이브리드 : 딥러닝 피셔 익스트랙터 + 머신러닝 : END - END 불가능

변수 개수	변수 형태	그래프
일변량 (변수 1개)	연속형 테이터	· 히스토그램 (Histogram) · 커널 밀도 곡선 (kernel Density Curve) · 박스 그래프 (Box Plot) · 바이올린 그래프(Violin Plot)
	범주형 데이터 (명목형, 순서형)	· 막대 그림 (Bar Chart) · 원 그림 (Pie Chart)
다변량 (변수 2개 이상)	연속형 테이터	· 산점도 (행렬) (Scatter Plot) · 선 그래프 (Line Plot) · 시계열 그래프 (Time Series Plot)
	범주형 데이터	· 모자이크 그림 (Mosaic Chart)