2022 하계 세미나

An Introduction to Anomaly Detection

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

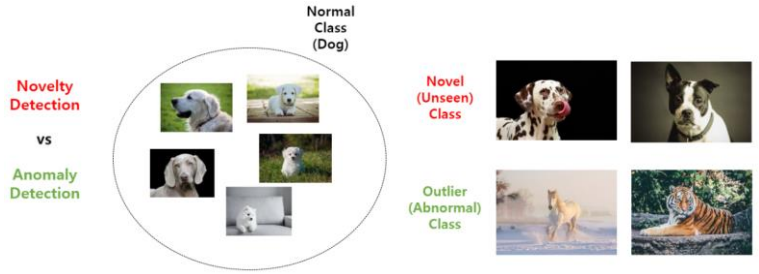
Presented By
나송주

Outline

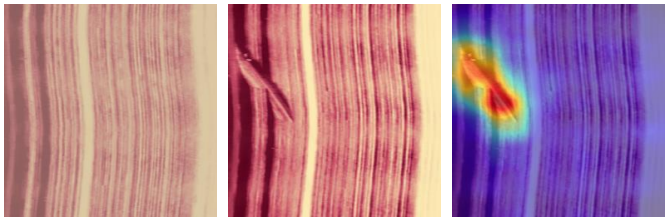
- Background
  - Anomaly Detection
    - What is anomaly detection?
    - Basic methods
    - Metric
- Papers
  - PANDA
  - Reverse Distillation

Background

- Anomaly Detection



Outlier / Novelty detection



Anomaly Segmentation

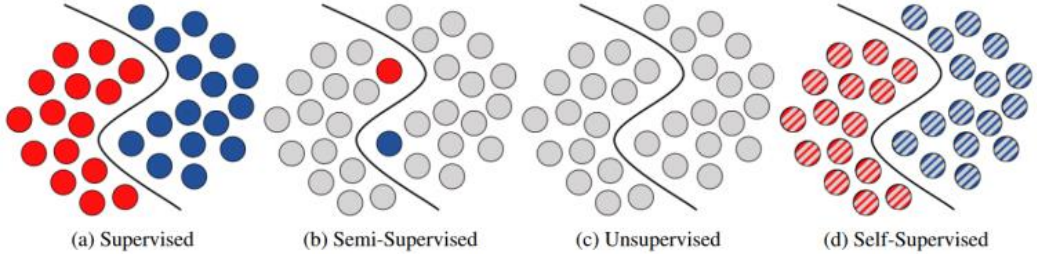
▪ 학습된 정상 데이터에서 벗어난 데이터를 구분하는 기술

– Semantic anomaly detection

⁘ Outlier

⁘ Novelty

– Anomaly segmentation



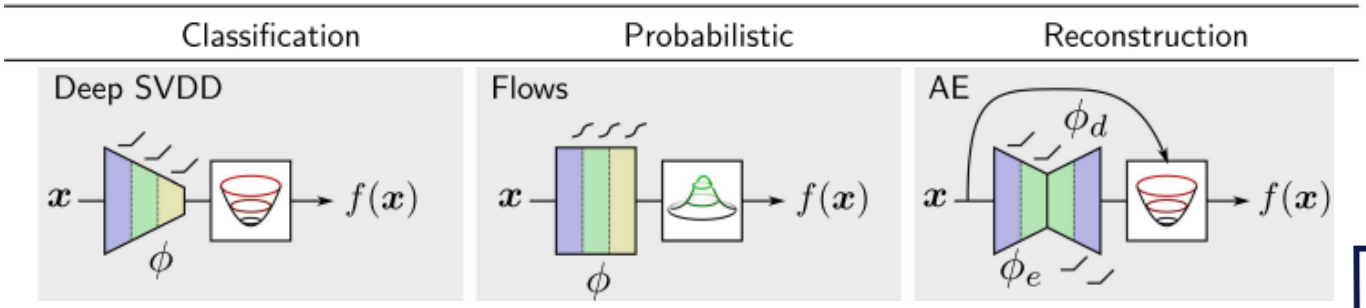
▪ Self-supervised / Unsupervised learning

1) RUFF, Lukas, et al, "Deep one-class classification.", *ICML*, 2018.
2) REISS, Tal, et al, "Panda: Adapting pretrained features for anomaly detection and segmentation.", *CVPR*, 2021.
3) YU, Jiawei, et al, "Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows.", *arXiv*, 2021.
4) ZAVRTANIK, Vitjan, et al, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection.", *ICCV*, 2021.
5) DENG, Hanqiu, et al, "Anomaly Detection via Reverse Distillation from One-Class Embedding.", *CVPR*, 2022.

Background

- Anomaly Detection
  - Classification based method
    - Normal data와 anomaly 사이에 decision boundary를 학습하여 anomaly 분류
    - Deep SVDD[1], PANDA[2]
  - Probabilistic method
    - Deep neural network를 이용하여 정상 sample들의 feature를 추출하고 통계적 방법으로 이들의 분포를 학습
    - 테스트 단계에서 정상 sample의 분포에서 벗어난 데이터를 anomaly로 분류
    - FastFlow[3]
  - Reconstruction based method
    - 학습 단계에서 정상 sample을 reconstruct하도록 모델을 학습시킴
    - 테스트 단계에서 reconstruction error가 큰 데이터를 anomaly로 분류
    - DRAEM[4], Reverse Distillation[5]

x : input data
ϕ : weights
$f(x)$: prediction



Background

- Anomaly Detection
  - Metric
    - AUROC
      - $TPR = \frac{TP}{TP+FN}$ (Sensitivity)
      - $FPR = 1 - \frac{TN}{TN+FP}$ (1-Specificity)
      - Decision threshold
      - ROC curve
      - Sample / Pixel AUROC
      - Anomaly score / Anomaly map
      - AUPRO

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Fig1. Confusion matrix

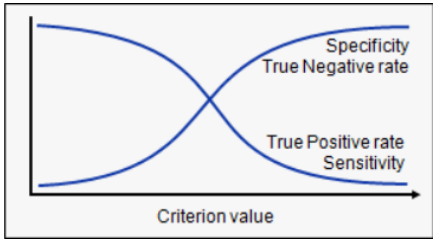


Fig2. Decision threshold

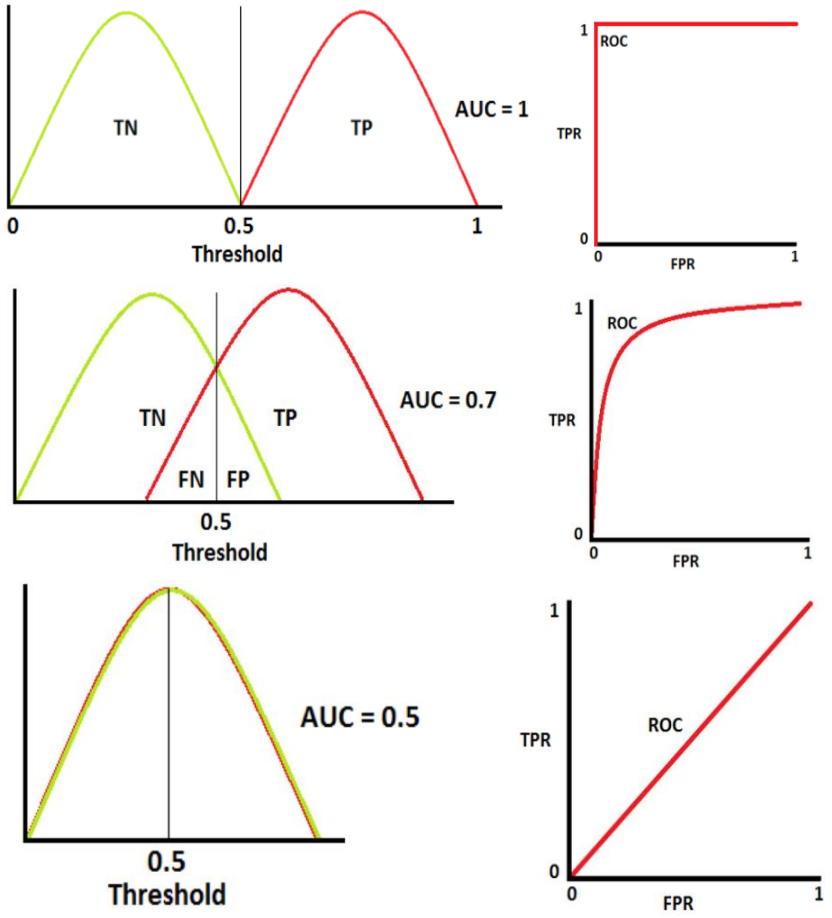
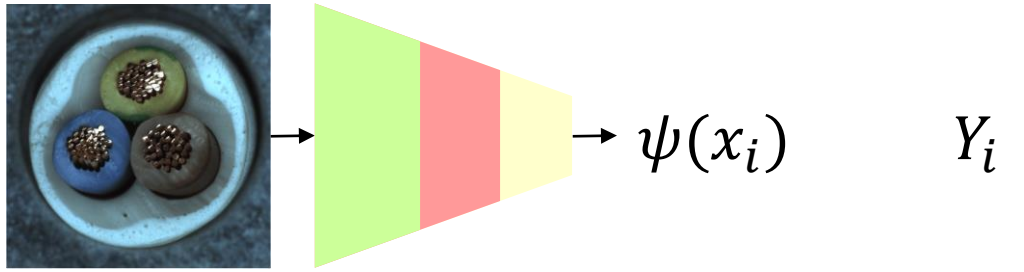


Fig3. ROC curve

Background

- Anomaly Detection
  - Metric
    - Sample AUROC / Anomaly Score
      - 각각의 test sample이 얼마나 anomalous한지 나타내는 anomaly score를 출력



      - Decision threshold를 0부터 1까지 조금씩 변화시키며 그 때 마다의 FPR과 TPR을 계산

$[0.01, 0.10, 0.95, \dots, 0.03, 0.98]$ → $[0, 1, 1, \dots, 1, 1]$ → FPR, TPR

Test sample x_i

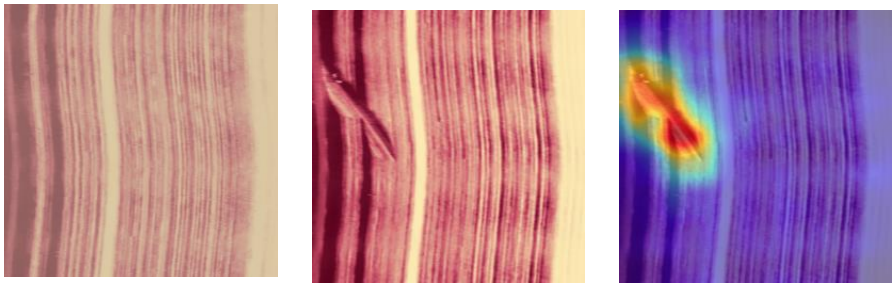
Decision threshold =0.02

$[0, 0, 1, \dots, 0, 1]$ → FPR, TPR

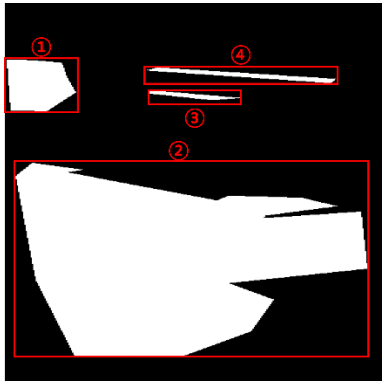
Decision threshold =0.116

Background

- Anomaly Detection
  - Metric
    - Pixel AUROC / Anomaly map
      - Anomaly segmentation task에서 네트워크는 input image의 각각에 pixel이 얼마나 anomalous한지 나타내는 anomaly map을 출력함
      - Anomaly map을 input image와 같은 크기로 resizing하여 anomaly segmentation을 수행
    - AUPRO
      - GT mask의 모든 anomalous region을 bounding box로 구분하고, 각각의 bounding box에서 pixel AUROC를 계산하고 이들의 평균을 계산



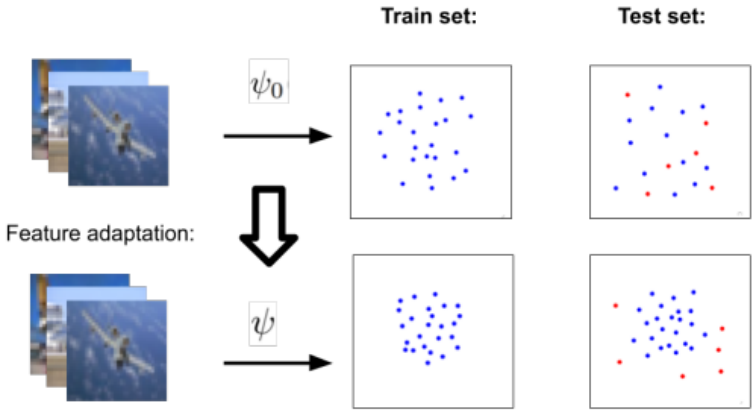
Anomaly Segmentation



AUPRO

Papers

- PANDA[1)]
  - Abstract
    - Pre-trained network를 사용한 간단한 unsupervised 기반의 anomaly detection이 기존의 self-supervised 기반의 방법론보다 뛰어난 성능을 가지는 것을 실험을 통해 증명

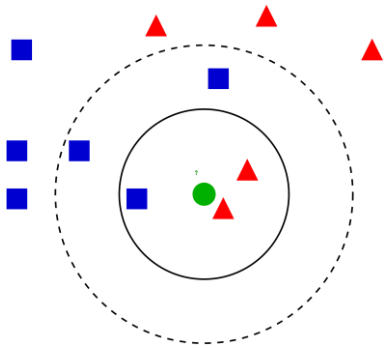


An illustration of feature adaptation procedure

    - Anomaly detection task에 pre-trained network를 fine-tuning하는 방법론을 제안
      - Compactness loss
      - Early stopping
      - Fine tuning with elastic weight consolidation (EWC)

Papers

- PANDA[1)]
  - Simple Baseline for Anomaly Detection (DN2[2)])
    - ImageNet pre-trained ResNet parameter ψ
    - Extract features of all train images $\psi(x_1), \psi(x_2), \dots \psi(x_N)$
    - Measure the density of normal features around the test sample feature $\psi(x_{test})$ (KNN)



An illustration of KNN

$$Anomaly\ score = \frac{1}{K}\frac{1}{HW}\sum_{\psi \in N_K(\psi(x_{test}))} \|\psi - \psi(x_{test})\|^2$$

H, W : $\psi(x_i)$의 size
$N_K(\psi(x_{test}))$: x_{test}의 K-nearest neighbor

  - Simple Baseline for Anomaly Segmentation (SPADE[3)])
    - K번째 Res block의 feature f_k
    - $F(x_i) \coloneqq concat\big(f_1(x_i), f_2(x_i), \dots, f_L(x_i)\big), (L : \#of\ Res\ blocks, f_L = \psi(x_i))$
    - $F(x_i)$를 image size로 resize
    - Let $F(x_{test})$ be F_{test}

$$Anomaly\ map(h,w) = \frac{1}{K}\sum_{F \in N_k(F_{test}(h,w))} \|F - F_{test}(h,w)\|^2$$

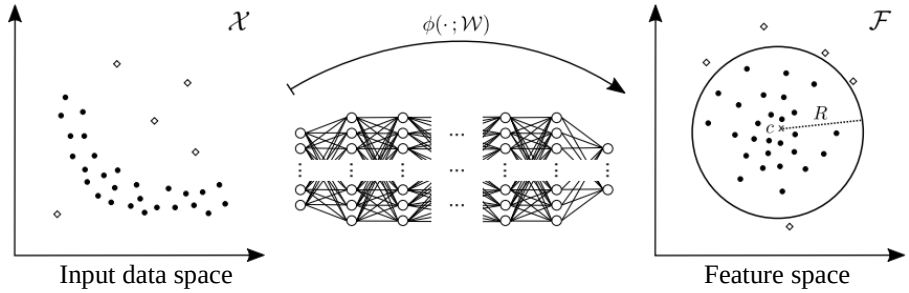
h, w : image의 pixel 좌표
$N_K(F_{test}(h,w))$: F_{test}의 (h,w)픽셀의 KNN 픽셀들

Papers

- PANDA[1)]
  - Feature Adaptation for Anomaly Detection
    - Existing feature adaptation method : DeepSVDD[2)]

$$L_{compact} = \sum_{x \in \mathcal{D}_{train}} \|\psi(x) - c\|^2$$

$L_{compact}$: Compactness loss
ψ_0 : Initial ImageNet pretrained ResNet
ψ : Fine-tuned ResNet
c : Center of initial normal features (average of $\psi_0(x)$)



An illustration of Deep SVDD

문제점

✓Catastrophic Collapse (for all x, $\psi(x) = c$)

✓Lose good discriminative representations of the pretrained features

Papers

- PANDA[1)]
  - Feature Adaptation for Anomaly Detection
    - Existing feature adaptation method : JO2)
      - Pre-trained network를 fine-tuning할 때 pretraining data (여기선 ImageNet) 함께 사용
      - Miss-classification에 대한 penalty를 부여
      - 처음 pretrained network가 가지고 있던 discriminative representation을 잃지 않도록 하는데 기여

$$L_{Joint} = \sum_{(x,y)\in\mathcal{D}_{pretrain}} \ell_{pretrain}(SMax(W\psi(x)), y) + \alpha \cdot \sum_{x\in\mathcal{D}_{train}} \|\psi(x) - c\|^2$$

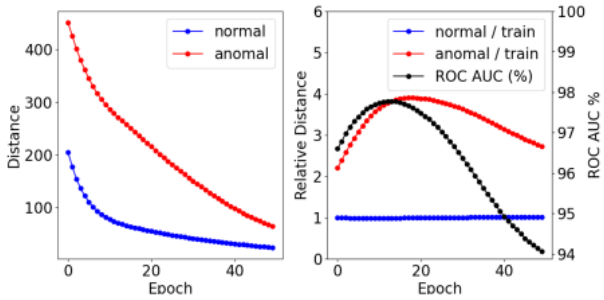
$D_{pretrain}$: ImageNet data
D_{train} : Anomaly detection data
$l_{pretrain}$: pretraining에 사용되었던 loss
W : pretraining에 linear classification layer
α : 두 loss의 weight를 조절하는 hyper parameter

      - 문제점
        - ✓Anomaly detection task에 온전하게 fine-tuning되지 못할 수 있음

Papers

- PANDA[1)]
  - Feature Adaptation for Anomaly Detection
    - Early Stopping
      - Compactness loss

$$L_{compact} = \sum_{x \in \mathcal{D}_{train}} \|\psi(x) - c\|^2$$



Fine-tuning with early stopping

      - Early stopping
        - ✓Simple early stopping
          - Fixed fine-tuning epoch (hyper parameter)
        - ✓Sample-wise early stopping
          - Save checkpoints $(\psi_1, \psi_2, \ldots \psi_T)$ during fine-tuning at fixed interval
          - $s_t = mean(\|\psi_t(x_{train}) - c\|^2)$
          - Test 단계에서 모든 check point weight에 대해 $s_t^{target} = \|\psi_t(x_{test}) - c\|^2$ 계산
          - s_t^{target}/s_t의 최대값을 가질 때 ψ_t를 test sample x_t에 대한 모델 파라미터로 사용

Papers

- PANDA[1)]
  - Feature Adaptation for Anomaly Detection
    - Elastic weight consolidation (EWC) loss
      - Fisher-information matrix

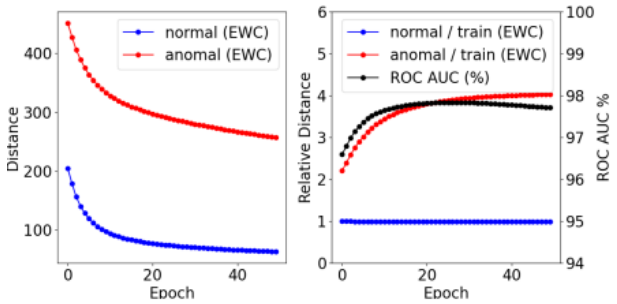
✓Log likelihood function의 2nd order derivative에 비례함

✓F_{θ_i}는 θ_i가 변화할 때 likelihood가 얼마나 급격하게 변하는지를 나타냄

$$F_\theta = \mathbb{E}_{(x,y)\in\mathcal{D}_{pretrain}}\left[\left(\frac{\partial}{\partial\theta}L_{pretrain}(x,y)\right)^2\right]$$

$$L_\theta = L_{compact}(\theta^*) + \frac{\lambda}{2}\cdot\sum_i F_{\theta_i}(\theta_i - \theta_i^*)^2$$

F_θ : Fisher-information matrix
F_{θ_i} : i번째 Fisher-information matrix의 대각 성분
θ : initial weight
θ^* : fine-tuned weight
i : weight의 index



Fine-tuning with EWC loss

Papers

- PANDA[1)]
  - Experiments
    - A Comparison of self-supervised and pre-trained features (One-class classification)

Dataset	Self-Supervised			Pretrained	
	OC-SVM	DeepSVDD	MHRot	DN2	PANDA
CIFAR10	64.7	64.8	90.1	92.5	**96.2**
CIFAR100	62.6	67.0	80.1	**94.1**	**94.1**
FMNIST	92.8	84.8	93.2	94.5	**95.6**
CatsVsDogs	51.7	50.5	86.0	96.0	**97.3**
DIOR	70.7	70.0	73.3	93.0	**94.3**

One-class classification performance (AUROC(%))

⁛ ImageNet pretrained features have significant advantages over self-supervised features

Dataset	Self-Supervised			Pretrained
	OCSVM	DeepSVDD	MHRot	DN2
Birds	62.0	60.8	64.4	**95.3**
Flowers	74.5	78.1	65.9	**94.1**
MvTec	70.8	77.9	65.5	**86.5**
WBC	75.4	71.2	57.7	**87.4**

One-class classification performance on small datasets (AUROC(%))

⁛ Self-supervised 방식은 데이터셋이 작아지면 성능이 크게 하락함

Papers

- PANDA[1)]
  - Experiments
    - A Comparison of self-supervised and pre-trained features (Anomaly segmentation)

	AE_{SSIM} [2]	AE_{L2} [2]	AnoGAN [28]	CNN Dict [22]	CAVGA-R_u [32]	Student [3]	SPADE
ROCAUC	87	82	74	78	89	-	**96.2**
PRO	69.4	79	-	51.5	-	85.7	**92.1**

Comparison of anomaly segmentation methods (pixel ROCAUC and PRO (%)

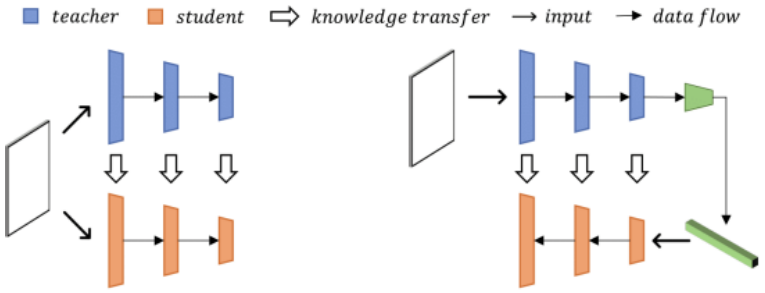
    - A Comparison of feature adaptation methods

Dataset	Baseline	PANDA		
	JO	Early	SES	EWC
CIFAR10	93.2	**96.2**	95.9	**96.2**
CIFAR100	91.1	**94.8**	94.6	94.1
FMNIST	94.9	95.4	95.5	**95.6**
CatsVsDogs	96.1	91.9	95.7	**97.3**
DIOR	93.1	95.4	**95.6**	94.3

Comparison of different feature adaptation methods

Papers

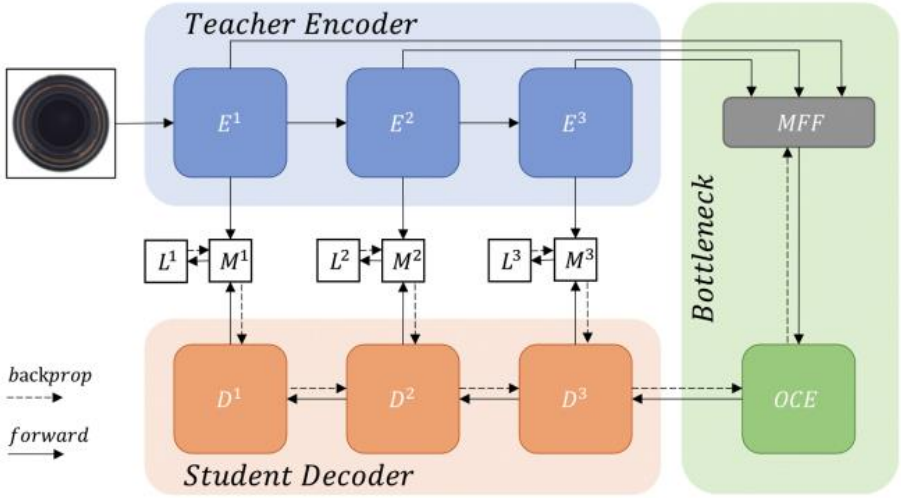
- Reverse Distillation[1)]
  - Abstract
    - Knowledge distillation based anomaly detection
      - Training : normal data에 대해 student가 teacher의 representation을 모방하도록 학습
      - Test : teacher와 student가 출력하는 feature의 차이가 크다면 anomaly로 분류
      - 문제점
        - ✓The identical or similar architectures of the teacher and student networks
        - ✓The same data flow in the T-S model
    - Contributions
      - Teacher-encoder, student-decoder 구조의 새로운 KD 모델 제안
      - One-class bottleneck module

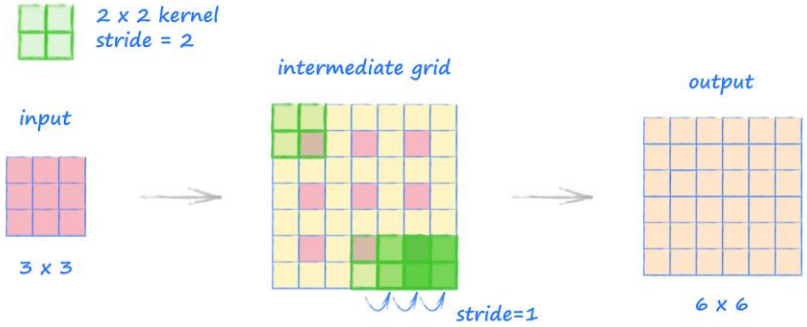


(a) 기존의 KD 구조도 (b) 제안된 KD 구조도

Papers

- Reverse Distillation[1)]
  - Architecture
    - Teacher Encoder E
      - ImageNet pre-trained ResNet
    - Bottleneck Module
      - Basic Res-block
    - Student Decoder D
      - Teacher encoder와 mirror-symmetric





	Encoder (ResNet50)		Decoder De-ResNet50	
Block3	1X1, 64 3X3, 64 1X1, 256	X1	1X1, 256 2X2, 256 1X1, 1024	X1
	1X1, 64 3X3, 64 1X1, 256	X2	1X1, 256 3X3, 256 1X1, 1024	X5
Output shape	16X16		16X16	
Block2	1X1, 128 3X3, 128 1X1, 512	X1	1X1, 128 2X2, 128 1X1, 512	X1
	1X1, 128 3X3, 128 1X1, 512	X3	1X1, 128 3X3, 128 1X1, 512	X3
Output shape	32X32		32X32	
Block1	1X1, 256 3X3, 256 1X1, 1024	X1	1X1, 64 2X2, 64 1X1, 256	X1
	1X1, 256 3X3, 256 1X1, 1024	X5	1X1, 64 3X3, 64 1X1, 256	X2
Output shape	64X64		64X64	

Papers

- Reverse Distillation[1)]
  - Training
    - 학습 중에 teacher의 모든 weight를 고정
    - Student가 teacher의 representation을 모방하도록 학습
      - Cosine similarity map M^k

$$M^k(h,w) = 1 - \frac{(f_E^k(h,w))^T \cdot f_D^k(h,w)}{\|f_E^k(h,w)\| \, \|f_D^k(h,w)\|}$$

$$a \cdot b = |a||b|cos\theta \qquad cos\theta = \frac{a \cdot b}{|a||b|}$$

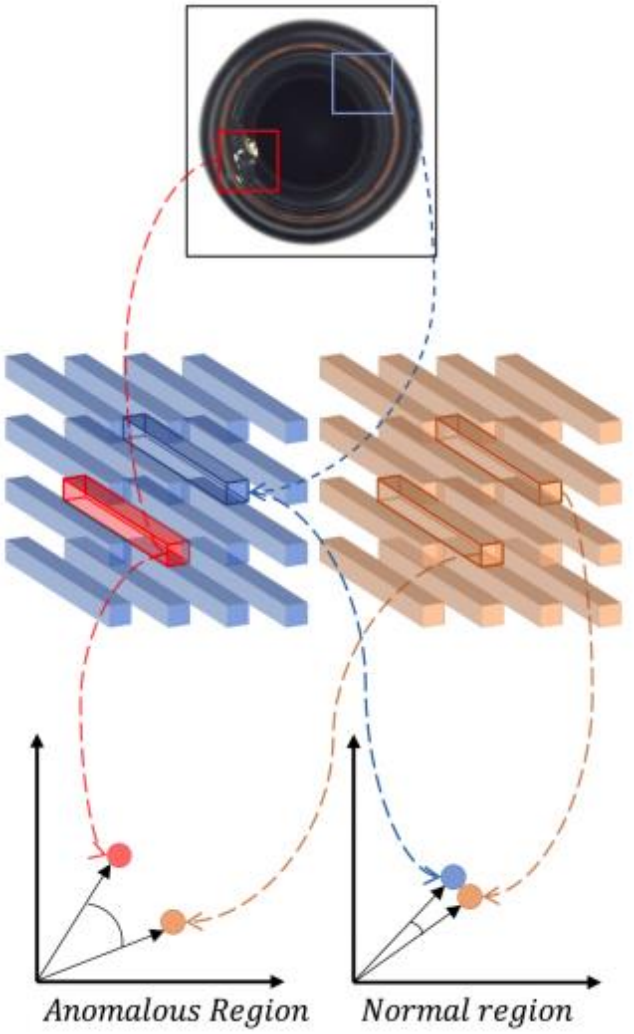
M^k : Cosine similarity map
f_E^k :k번째 encoder의 출력 feature
f_D^k : k번째 decoder의 출력 feature
h, w : pixel 좌표

      - Loss function

$$\mathcal{L}_{KD} = \sum_{k=1}^{K} \left\{ \frac{1}{H_k W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} M^k(h,w) \right\}$$

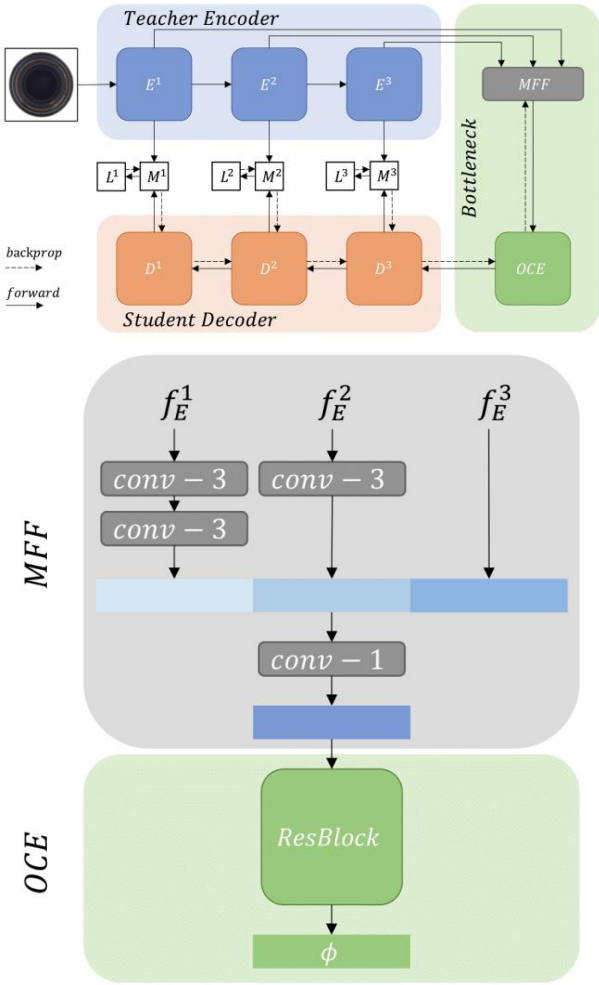
L_{KD}: loss
W_k:k번째 similarity map의 width
H_k: k번째 similarity map의 height
h, w : pixel 좌표

    - Bottleneck module은 student가 normal data에 대해서만 teacher를 모방할 수 있도록 제한



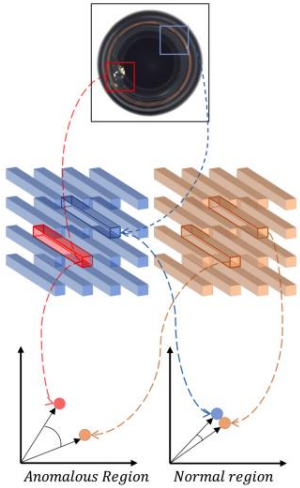
Papers

- Reverse Distillation[1)]
  - One-Class Bottleneck Embedding
    - Teacher encoder가 출력하는 high level feature는 anomaly-free information뿐만 아니라 너무 많은 정보를 포함함
      - OCE
        - ✓Project high-dimensional representation into a low-dimensional space
        - ✓Compact embedding acts as an information bottleneck
        - ✓Prohibit the propagation of anomaly-information to the student
    - Reverse order of knowledge distillation
      - MFF
        - ✓Concatenate multi-scale representations



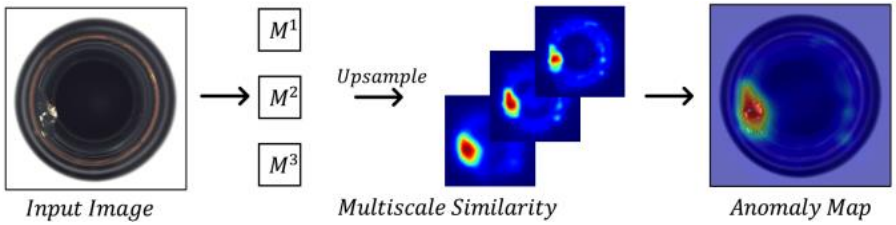
Papers

- Reverse Distillation[1)]
  - Anomaly scoring
    - Anomaly Segmentation
      - Student는 normal sample에 대해서만 teacher를 모방하도록 학습됨
      - 따라서 anomalous region에 대해서 similarity map M^k가 큰 값을 가짐
      - Anomaly map



$$S_{AL} = \sum_{i=1}^{L} \Psi(M^i)$$

ψ : bilinear up-sampling operation
S_{AL} : Anomaly map
L : number of blocks



- Anomaly detection
  - Normal sample에 대해서는 S_{AL}이 큰 값을 가지는 픽셀이 없을 것이라는 idea에 착안하여 S_{AL}의 최대값을 anomaly score로 사용

$$S_{AD} = Max(S_{AL}(H, W))$$

Papers

- Reverse Distillation[1)]
  - Experiments
    - MVTec

Image Size		128		256								
Category/Method		MKD [33]	Ours	GT [10]	GN [2]	US [4]	PSVDD [43]	DAAD [16]	MF [40]	PaDiM [8]	CutPaste [23]	Ours
Textures	Carpet	79.3	99.2	43.7	69.9	91.6	92.9	86.6	94.0	**99.8**	93.9	**98.9**
	Grid	78.0	95.7	61.9	70.8	81.0	94.6	95.7	85.9	96.7	**100**	**100**
	Leather	95.1	100	84.1	84.2	88.2	90.9	86.2	**99.2**	**100**	**100**	**100**
	Tile	91.6	99.4	41.7	79.4	**99.1**	97.8	88.2	99.0	98.1	94.6	**99.3**
	Wood	94.3	98.8	61.1	83.4	97.7	96.5	98.2	**99.2**	**99.2**	**99.1**	**99.2**
	Average	*87.7*	*98.6*	*58.5*	*77.5*	*91.5*	*94.5*	*91.0*	*95.5*	*98.8*	*97.5*	***99.5***
Objects	Bottle	99.4	100	74.4	89.2	99.0	98.6	97.6	99.1	**99.9**	98.2	**100**
	Cable	89.2	97.1	78.3	75.7	86.2	90.3	84.4	**97.1**	92.7	81.2	**95.0**
	Capsule	80.5	89.5	67.0	73.2	86.1	76.7	76.7	87.5	91.3	**98.2**	**96.3**
	Hazelnut	98.4	99.8	35.9	78.5	93.1	92.0	92.1	**99.4**	92.0	98.3	**99.9**
	Metal Nut	73.6	99.2	81.3	70.0	82.0	94.0	75.8	96.2	98.7	**99.9**	**100**
	Pill	82.7	93.3	63.0	74.3	87.9	86.1	90.0	90.1	93.3	**94.9**	**96.6**
	Screw	83.3	91.1	50.0	74.6	54.9	81.3	**98.7**	**97.5**	85.8	88.7	97.0
	Toothbrush	92.2	90.3	97.2	65.3	95.3	**100**	99.2	**100**	96.1	99.4	**99.5**
	Transistor	85.6	99.5	86.9	79.2	81.8	91.5	87.6	94.4	**97.4**	96.1	**96.7**
	Zipper	93.2	94.3	82.0	74.5	91.9	97.9	85.9	**98.6**	90.3	**99.9**	98.5
	Average	*87.8*	*95.4*	*71.6*	*75.5*	*85.8*	*90.8*	*88.8*	*96.0*	*93.8*	*95.5*	***98.0***
Total Average		*87.8*	*96.5*	*67.2*	*76.2*	*87.7*	*92.1*	*89.5*	*95.8*	*95.5*	*96.1*	***98.5***

Sample AUROC (%)

Image Size		128		256						
Category/Method		MKD [33]	Ours	US [4]	MF [40]	SPADE [7]	PaDiM [8]	RIAD [46]	CutPaste [23]	Ours
Textures	Carpet	95.6/-	98.1/95.3	-/87.9	-/87.8	97.5/94.7	**99.1**/96.2	96.3/-	98.3/-	98.9/**97.0**
	Grid	91.8/-	97.3/92.6	-/95.2	-/86.5	93.7/86.7	97.3/94.6	98.8/-	97.5/-	**99.3/97.6**
	Leather	98.1/-	99.0/98.6	-/94.5	-/95.9	97.6/97.2	99.2/97.8	99.4/-	**99.5**/-	99.4/**99.1**
	Tile	82.8/-	92.6/84.8	**-/94.6**	-/88.1	87.4/75.9	94.1/86.0	89.1/-	90.5/-	**95.6**/90.6
	Wood	84.8/-	92.1/82.3	**-/91.1**	-/84.8	88.5/87.4	94.9/**91.1**	85.8/-	**95.5**/-	95.3/90.9
	Average	*90.6/-*	*95.8/90.7*	*-/92.7*	*-/88.6*	*92.9/88.4*	*96.9/93.2*	93.9/-	*96.3/-*	***97.7/95.0***
Objects	Bottle	96.3/-	98.2/94.7	-/93.1	-/88.8	98.4/95.5	98.3/94.8	98.4/-	97.6/-	**98.7/96.6**
	Cable	82.4/-	97.8/90.5	-/81.8	**-/93.7**	97.2/90.9	96.7/88.8	84.2/-	90.0/-	**97.4**/91.0
	Capsule	95.9/-	96.5/87.2	**-/96.8**	-/87.9	**99.0**/93.7	98.5/93.5	92.8/-	97.4/-	98.7/95.8
	Hazelnut	94.6/-	98.8/89.2	**-/96.5**	-/88.6	**99.1**/95.4	98.2/92.6	96.1/-	97.3/-	98.9/95.5
	Metal Nut	86.4/-	96.6/84.1	-/94.2	-/86.9	**98.1/94.4**	97.2/85.6	92.5/-	93.1/-	97.3/92.3
	Pill	89.6/-	97.0/90.0	-/96.1	-/93.0	96.5/94.6	95.7/92.7	95.7/-	95.7/-	**98.2/96.4**
	Screw	96.0/-	98.3/94.4	-/94.2	-/95.4	98.9/96.0	98.5/94.4	98.8/-	96.7/-	**99.6/98.2**
	Toothbrush	96.1/-	98.2/86.7	-/93.3	-/87.7	97.9/93.5	98.8/93.1	98.9/-	98.1/-	**99.1/94.5**
	Transistor	76.5/-	97.6/85.2	-/66.6	**-/92.6**	94.1/87.4	**97.5**/84.5	87.7/-	93.0/-	92.5/78.0
	Zipper	93.9/-	97.0/92.3	-/95.1	-/93.6	96.5/92.6	**98.5/95.9**	97.8/-	99.3/-	98.2/95.4
	Average	*90.8/-*	*97.6/89.4*	*-/90.8*	*-/90.8*	*97.6/**93.4***	*97.8/91.6*	94.3/-	*95.8/-*	***97.9/93.4***
Total Average		*90.7/-*	*97.0/89.9*	*-/91.4*	*-/90.1*	*96.5/91.7*	*97.5/92.1*	94.2/-	*96.0-*	***97.8/93.9***

Pixel AUROC, AUPRO (%)

- One-Class Novelty Detection

Method	MNIST	F-MNIST	CIFAR10	Caltech-256
LSA [1]	97.5	92.2	64.1	-
OCGAN [27]	97.3	87.8	65.7	-
HRN [17]	97.6	92.8	71.3	-
DAAD [16]	99.0	-	75.3	-
MKD [33]	98.7	94.5	84.5	-
G2D [28]	-	-	-	95.7
OiG [45]	-	-	-	98.2
Ours	**99.3**	**95.0**	**86.5**	**99.9**

AUROC (%)

감사합니다.