
Inpainting Transformer for Anomaly Detection

School of Industrial and Management Engineering, Korea University

Yongwon Jo

Contents

- ❖ Research Purpose
- ❖ Inpainting Transformer (InTra)
- ❖ Experiments
- ❖ Conclusion

Research Purpose

❖ Inpainting Transformer for Anomaly Detection (arXiv, 2021)


- 독일 Fujitsu 기업에서 발표한 논문이며 2021년 7월 16일 기준으로 인용 횟수는 0회

Inpainting Transformer for Anomaly Detection

Jonathan Pirnay^a, Keng Chai^a

^a*Digital Incubation, Fujitsu Technology Solutions GmbH, Germany*

Abstract

Anomaly detection in computer vision is the task of identifying images which deviate from a set of normal images. A common approach is to train deep convolutional autoencoders to inpaint covered parts of an image and compare the output with the original image. By training on anomaly-free samples only, the model is assumed to not being able to reconstruct anomalous regions properly. For anomaly detection by inpainting we suggest it to be beneficial to incorporate information from potentially distant regions. In particular we pose anomaly detection as a patch-inpainting problem and propose to solve it with a purely self-attention based approach discarding convolutions. The proposed Inpainting Transformer (InTra) is trained to inpaint covered patches in a large sequence of image patches, thereby integrating information across large regions of the input image. When learning from scratch, InTra achieves better than state-of-the-art results on the MVTec AD  dataset for detection and localization.

Research Purpose

❖ Inpainting Transformer for Anomaly Detection (arXiv, 2021)

- Anomaly detection for the visual inspection task
 - 이미지 내 이상이 존재하는 지 여부(Anomaly detection)를 예측하는 문제
 - 이상이 존재한다면 이상 지역까지 동시에 탐지(Anomaly segmentation or localization)해야함
- 해당 문제의 어려움
 - 대부분 정상에 대한 이미지만 획득 가능하며 이상에 대한 이미지를 수집하기 어려움 → 지도학습 불가
 - 이상이라는 것의 형태가 정해져 있는 것이 아니기에 레이블링 하기 어렵다는 특징

Research Purpose

❖ Inpainting Transformer for Anomaly Detection (arXiv, 2021)

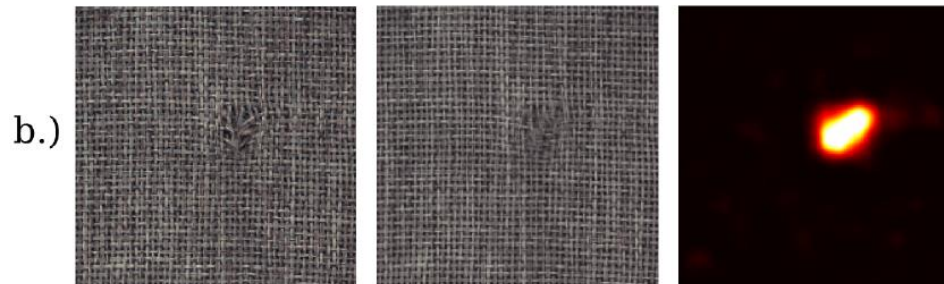
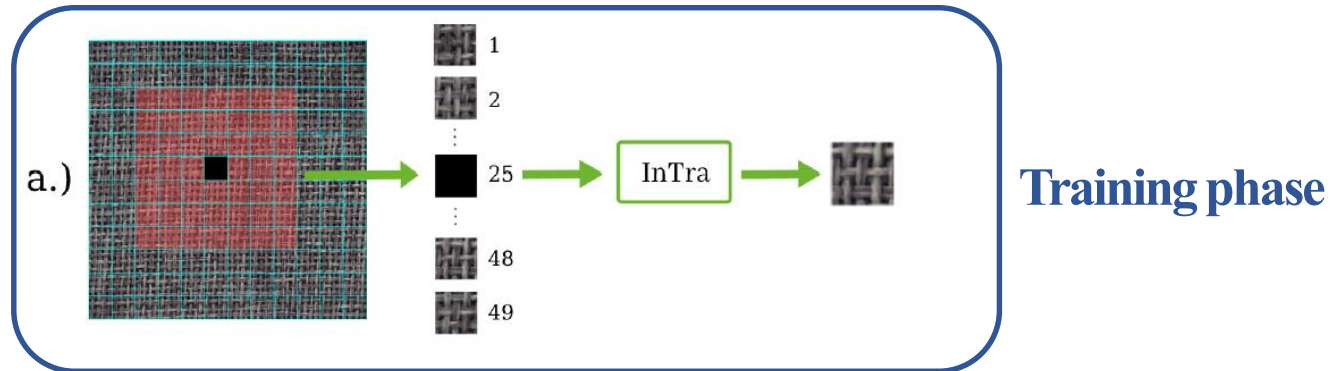
- Convolutional autoencoder (CAE) and Generative adversarial network (GAN)
 - 가정: 해당 CAE는 정상 데이터로만 학습하였기 때문에 이상 존재하는 이미지는 복원할 수 없을 것
 - 정상 데이터만으로 구성된 학습 데이터셋으로 입력 이미지를 복원하는 CAE와 GAN 학습
 - 하지만 CAE가 이상 데이터에 대해서도 잘 복원하기 때문에 정확히 정상만 모델링 불가(일반화 성능↑)
 - GAN의 경우, mode collapse 문제가 존재하며 학습하기 어려워 실제 상황 적용하기 어려움
- Inpainting Transformer (InTra)
 - InTra는 입력 이미지 내 특정 영역을 제거하여 이미지가 입력되어 제거된 영역을 복원
 - 합성곱 신경망이 아닌 Visual transformer 로 구성된 Encoder-deocder 구조를 제안

Inpainting transformer (InTra)

- Overview of InTra

❖ Inpainting transformer (InTra) – Training phase

- 입력 이미지 내 일부를 추출해 특정 정사각형 패치(Patch)로 분할하여 입력 데이터로 사용
 - 그림 a의 빨간색 영역 = 입력 이미지 내 일부 추출 결과
- 여러 Patch 중 하나를 제거하고 검은색 이미지로 대체
- 입력 데이터 패치 중 제거된 Patch는 출력 변수로 사용

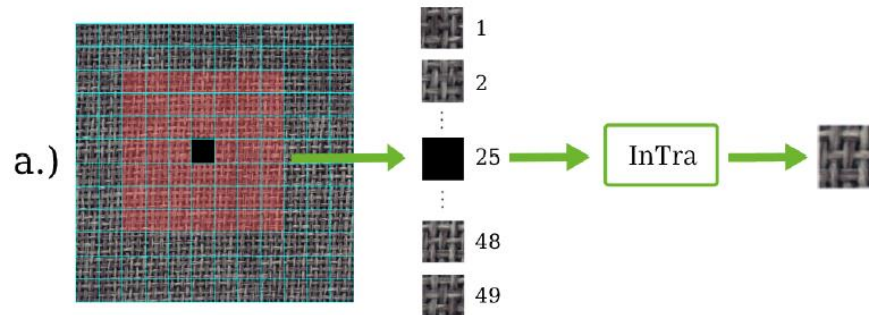


Inpainting transformer (InTra)

- Overview of InTra

❖ Inpainting transformer (InTra) – Inference phase

- 입력 이미지 내 일부를 추출해 특정 정사각형 패치(Patch)로 분할하여 입력 데이터로 사용
- 추론 단계에서는 b행 중간 이미지와 같이 이미지 복원 진행(b-middle, reconstructed image)
- 원본 데이터(b-left, real image)와 픽셀 단위 재구축 오차 계산하고 이를 이상치 맵으로 사용
 - 이상치 맵 (Anomaly map, b-right) – 노란색에 가까울수록 이상 지역이며 검은색에 가까울수록 정상



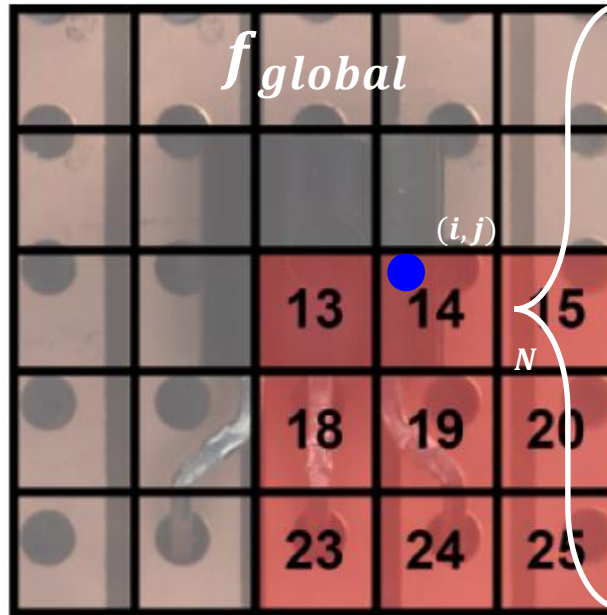
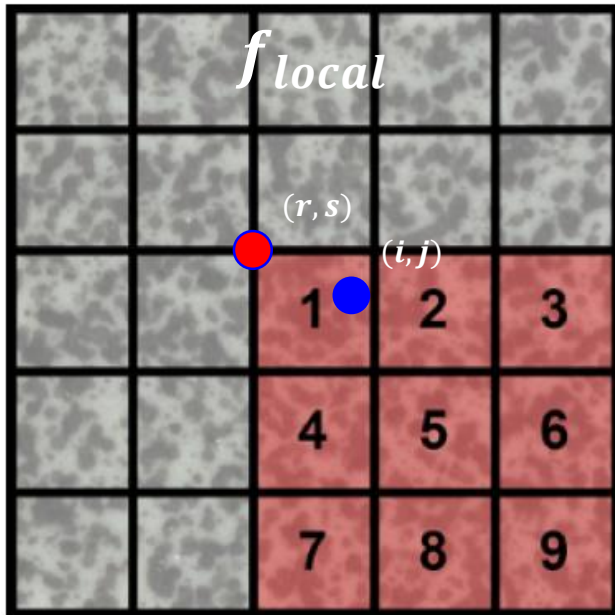
Inference phase

Inpainting transformer (InTra)

- Positional embedding for InTra

❖ Positional embedding for specific image domain

- 입력 데이터 전체가 중요한 경우(Texture)는 f_{local} 을 사용
- 배경과 관심있는 객체가 확실히 구별 가능한 경우(Object)는 f_{global} 을 사용
- 픽셀의 2차원 위치 (i, j) 를 Embedding 방법마다 오른쪽 수식과 같이 변경
 - i, j 는 패치의 위치 인덱스이며 L 은 특정 패치의 길이이고 r, s 는 특정 패치의 왼쪽 상단 좌표
 - N 은 Local의 경우 입력 이미지내 패치의 세로 길이 (6page, a의 빨간 상자의 세로 길이)



$$f_{local}$$
$$(i, j) \rightarrow (i - r) * L + j - s + 1$$

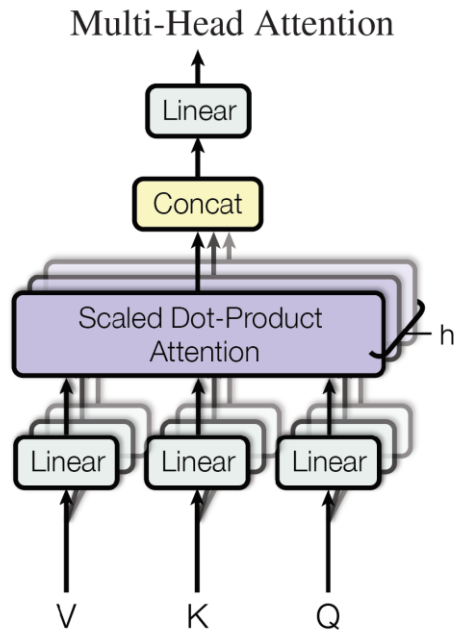
$$f_{global}$$
$$(i, j) \rightarrow (i - 1) * N + j$$

Inpainting transformer (InTra)

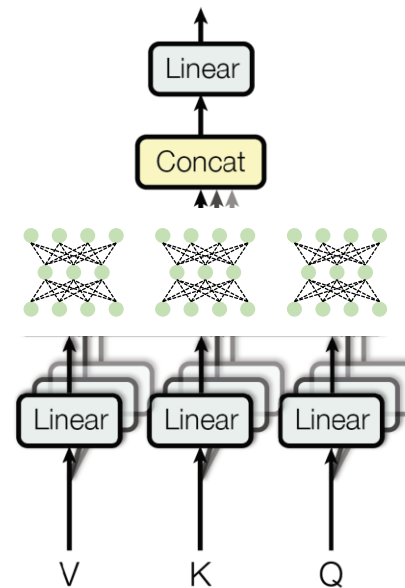
- Multi-head feature self-attention (MFSA)

❖ Multi-head feature self-attention (MFSA)

- 기존 Transformer 네트워크 self-attention은 Dot product와 Softmax 연산을 사용 (Multi-head attention)
 - key, query, value에 대한 weight와 feature vector 사이 dot product
- MFSA는 dot product연산이 아닌 Multi-layer perceptron을 사용해 self-attention 진행
- 선형 관계가 아닌 비선형 관계 추출을 위해 Multi-head attention에서 MFSA로 변경



Multi-head feature self-attention



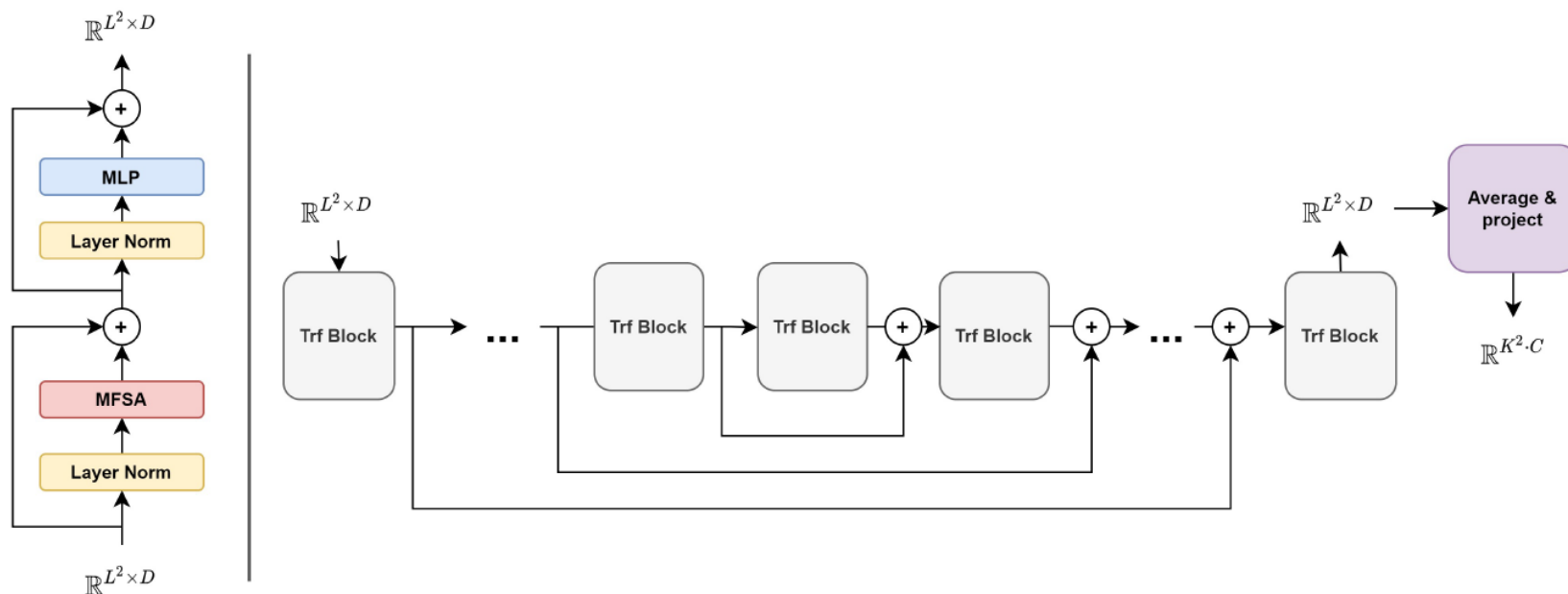
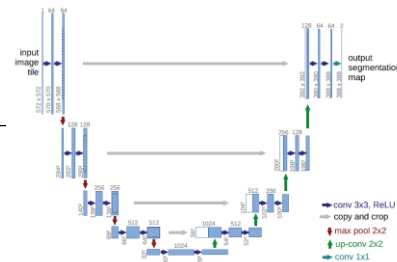
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Inpainting transformer (InTra)

- Multi-head feature self-attention (MFSA)

❖ Network architecture of InTra

- 각각의 Transformer 블록은 Layer normalization과 MFSA와 Multilayer perceptron (MLP)로 구성
- 잔차 학습을 위해 블록 내 잔차, 블록간 잔차를 학습에 사용
- 블록 간 잔차를 학습하여 U-Net과 같이 전반적 특징과 지역적 특징을 동시에 학습
 - U-Net의 Skip connection과 같은 형태로 블록 간 잔차 학습



- Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Inpainting transformer (InTra)

- Loss function

❖ Loss function of InTra

- 3개의 손실 함수 합으로 InTra 모델 학습 진행
 - Reconstruction loss (pixel-wise L2) + Gradient magnitude similarity (GMS) + Structure similarity (SSIM)
 - Pixel-wise L2는 색깔에 집중, GMS와 SSIM은 이미지 내 구조가 유사해지도록 학습
 - α 는 GMS의 가중치이고 β 는 SSIM의 가중치
 - x_p 는 제거된 패치이며 \hat{x}_p 는 제거된 패치를 복원한 패치

$$\begin{aligned} & Loss(x_p, \hat{x}_p) \\ = & \boxed{L_2(x_p, \hat{x}_p)} + \frac{\alpha}{K^2} \sum_{(i,j) \in K \times K} d_{GMS}(x_p, \hat{x}_p) + \frac{\beta}{K^2} \sum_{(i,j) \in K \times K} d_{SSIM}(x_p, \hat{x}_p) \end{aligned}$$

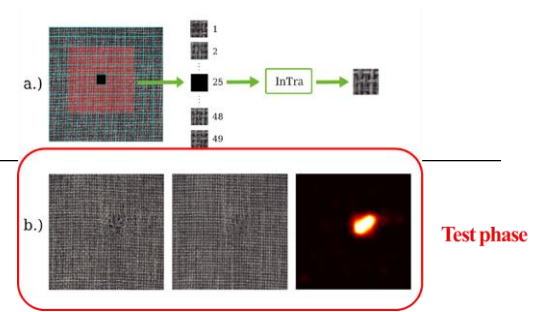
색깔 유사하도록 학습

객체 구조가 유사하도록 학습

- Xue, W., Zhang, L., Mou, X., & Bovik, A. C. (2013). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE Transactions on Image Processing, 23(2), 684-695.
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

Inpainting transformer (InTra)

- Loss function



❖ Anomaly score for each pixel and a image – Inference phase

- 테스트 할 이미지(x)를 정확히 복원하고 난 뒤 아래 이상치 점수를 산출
 - 6 page 그림 b의 중간 내 복원한 이미지를 우선적으로 생성
- 픽셀 당 이상치 점수는 RGB 채널별 재구축 오차의 평균과 Embedding 벡터간 차이의 합
- 이미지에 대한 이상치 점수는 픽셀 이상치 점수 중 최대 값 사용

$$Anomaly\ map(x) = \{diff(x, \hat{x}) - \frac{1}{|T|} \sum_{z \in T} diff(x, \hat{z})\}^2$$

$$Anomaly\ score(x) = \max_{(i,j) \in H \times W} anomaly\ map(x)$$

Experiments

- Experiment dataset

❖ MVTec dataset for Anomaly Detection

- 카메라 제조사인 MVTec사에서 만든 데이터 셋
- 총 15개의 산업 품목에 대한 이미지가 존재하며 5개는 Texture이며 나머지 10개는 Object
- 학습 데이터는 모두 정상만 존재하고 테스트 데이터에는 정상과 이상 제품 존재
- 이상에 대해서는 픽셀 단위 이상 여부에 대한 정답 존재

Texture

정상 이미지

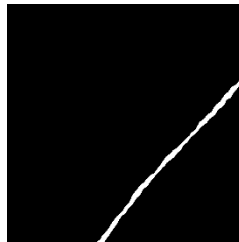
이상 이미지

이상에 대한 정답

나무



카펫



Object

정상 이미지

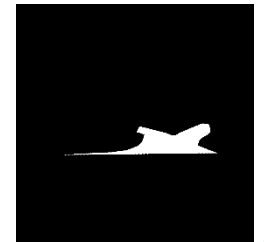
이상 이미지

이상에 대한 정답

헤이즐넛



캡슐



Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD—A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 9592-9600).

Experiments

- Experiment dataset

❖ 이미지 단위 이상 탐지 성능 비교

- 평가 지표로 ROCAUC 점수 사용
- 기존 방법론 대비 Transformer 기반 Encoder-decoder 구조를 사용하여 성능 향상 성공

Category	Patch-SVDD [17]	RIAD [11]	CutPaste [18]	PaDiM [20]	Ours
Carpet	92.9	84.2	93.1		98.8
Grid	94.6	99.6	99.9		100.0
Leather	90.9	100.0	100.0		100.0
Tile	97.8	93.4	93.4		98.2
Wood	96.5	93.0	98.6		98.0
avg. textures	94.54	95.1	97.0	99.0	99.0
Bottle	98.6	99.9	98.3		100.0
Cable	90.3	81.9	80.6		84.2
Capsule	76.7	88.4	96.2		86.5
Hazelnut	92.0	83.3	97.3		95.7
Metal Nut	94.0	88.5	99.3		96.9
Pill	86.1	83.8	92.4		90.2
Screw	81.3	84.5	86.3		95.7
Toothbrush	100.0	100.0	98.3		99.7
Transistor	91.5	90.9	95.5		95.8
Zipper	97.9	98.1	99.4		99.4
avg. objects	90.8	89.9	94.3	(97.2)	94.41
avg. all categories	92.1	91.7	95.2	(97.9)	95.94

Experiments

- Experiment dataset

❖ 픽셀 단위 이상 탐지 성능 비교

- 평가 지표로 ROCAUC 점수 사용
- 기존 방법론 대비 Transformer 기반 Encoder-decoder 구조를 사용하여 성능 향상 성공

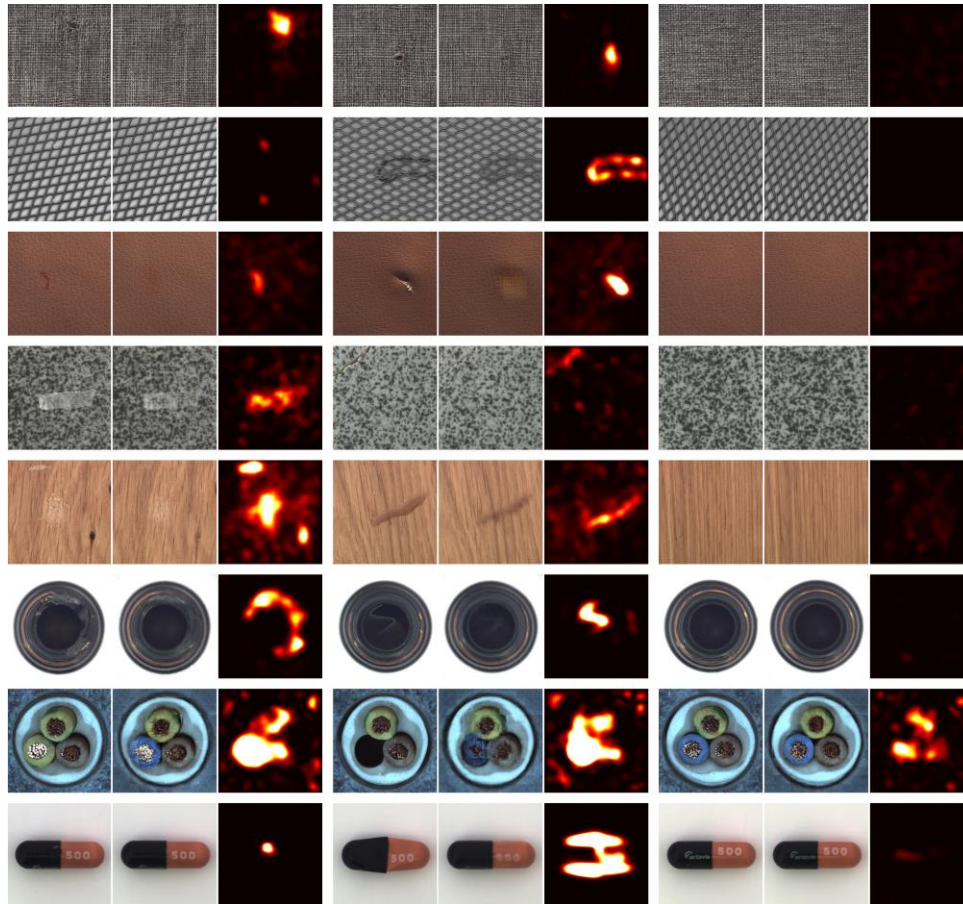
Category	Patch-SVDD [17]	RIAD [11]	CutPaste [18]	PaDiM [20]	Ours
Carpet	92.6	96.3	98.3	99.1	99.2
Grid	96.2	98.8	97.5	97.3	99.4
Leather	97.4	99.4	99.5	99.2	99.5
Tile	91.4	89.1	90.5	94.1	94.4
Wood	90.8	85.8	95.5	94.9	90.5
avg. textures	93.7	93.9	96.3	96.9	96.6
Bottle	98.1	98.4	97.6	98.3	97.1
Cable	96.8	84.2	90.0	96.7	93.2
Capsule	95.8	92.8	97.4	98.5	97.7
Hazelnut	97.5	96.1	97.3	98.2	98.3
Metal Nut	98.0	92.5	93.1	97.2	93.3
Pill	95.1	95.7	95.7	95.7	98.3
Screw	95.7	98.8	96.7	98.5	99.5
Toothbrush	98.1	98.9	98.1	98.8	99.0
Transistor	97.0	87.7	93.0	97.5	96.1
Zipper	95.1	97.8	99.3	98.5	99.2
avg. objects	96.7	94.3	95.8	(97.8)	97.17
avg. all categories	95.7	94.2	96.0	(97.5)	96.98

Experiments

- Experiment dataset

❖ 이상 지역 탐지 결과(Anomaly map) 시각화

- 왼쪽-원본 이미지/ 중간-복원된 이미지/ 오른쪽-Anomaly map

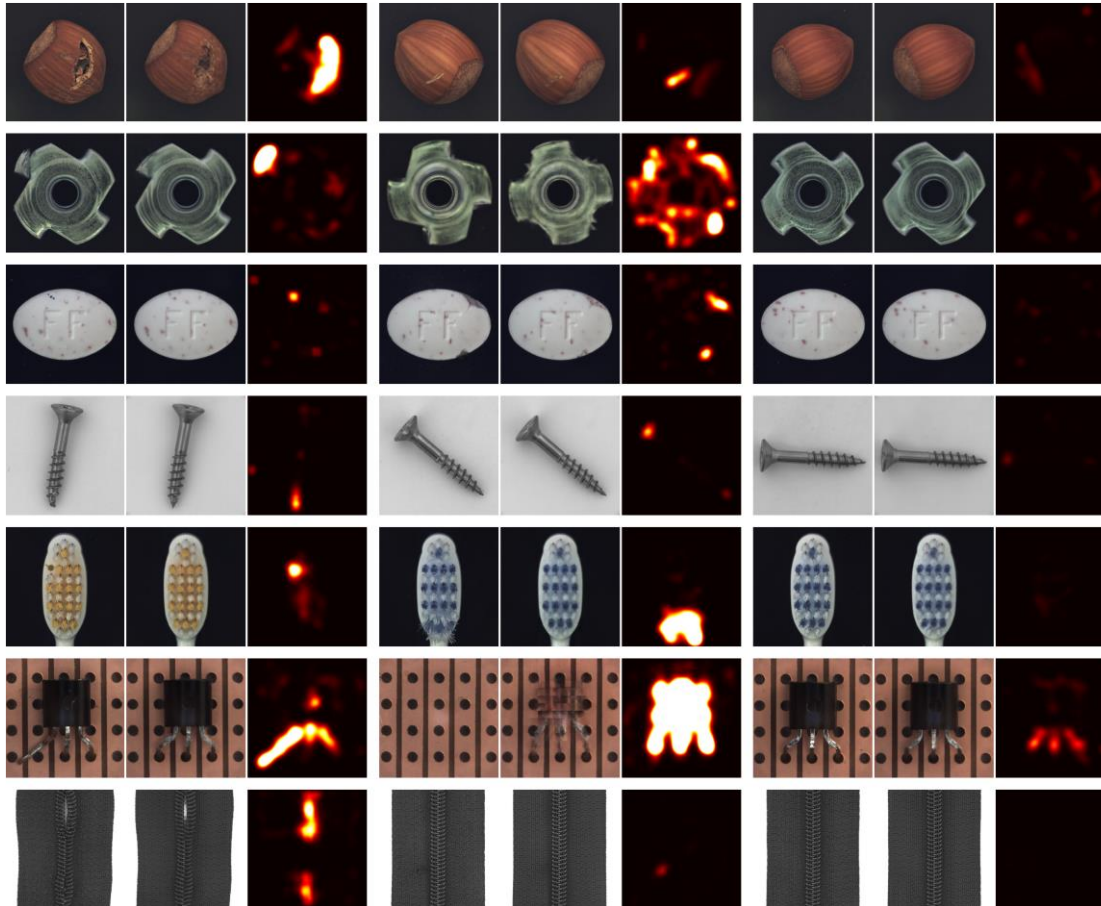


Experiments

- Experiment dataset

❖ 이상 지역 탐지 결과(Anomaly map) 시각화

- 왼쪽-원본 이미지/ 중간-복원된 이미지/ 오른쪽-Anomaly map



Experiments

- Ablation study for the residual connection

- ❖ Residual connection (잔차 학습) 여부에 따른 성능 비교
 - Texture 이미지 (3개)와 Object (2개)에 대한 성능 비교 진행
 - 입력 데이터 전반에 대한 관심이 있는 Texture에서 정상을 표현하는 벡터를 산출했다는 평가

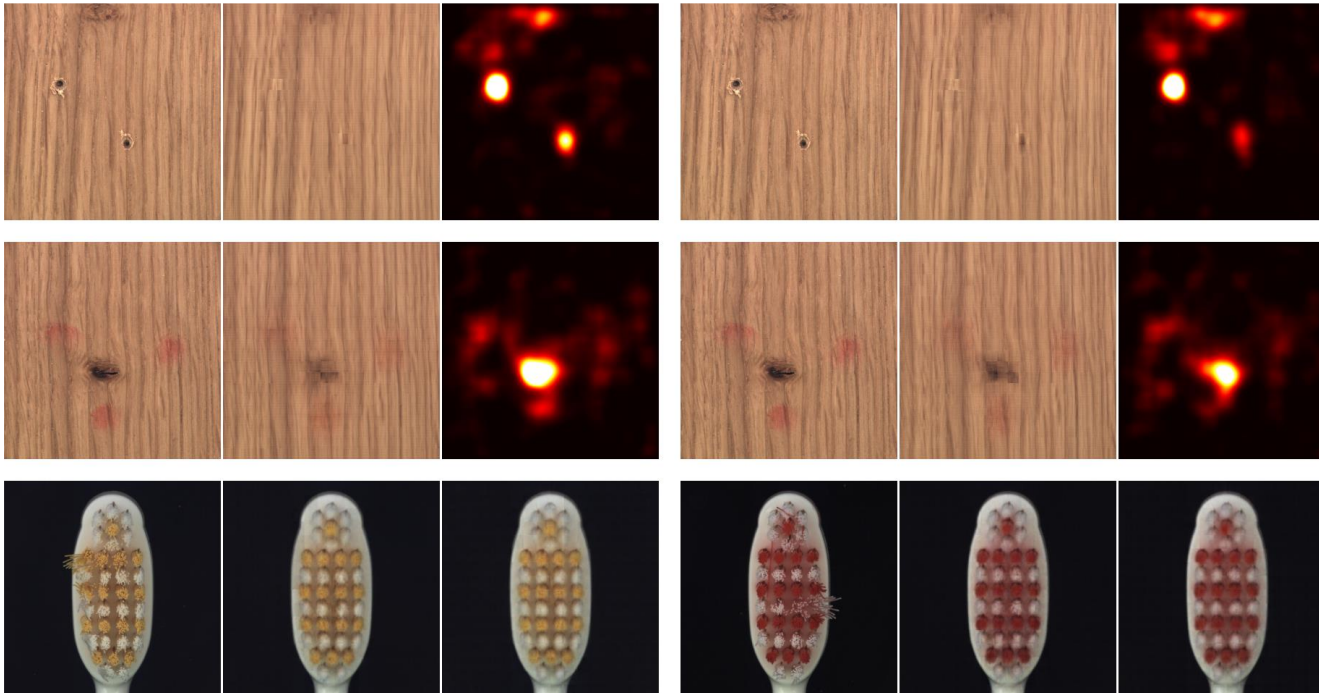
Category	Det. With	Det. Without	Seg. With	Seg. Without
Carpet	98.8	98.8	99.2	99.3
Leather	100.0	99.9	99.5	99.4
Wood	96.8	95.6	89.9	88.9
Toothbrush	99.7	100	99.0	98.9
Hazelnut	91.0	91.4	97.3	97.3

Experiments

- Ablation study for the residual connection

❖ Residual connection (잔차 학습) 여부에 따른 Anomaly map 비교

- 1,2행은 나무에 대한 입력 이미지 복원 및 Anomaly map
- 왼쪽 3장의 이미지는 Long residual connection을 사용한 경우이며 오른쪽 3장은 사용 ×
- 왼쪽의 경우가 Anomaly map 상 이상 영역을 더 정확히 인식



Experiments

- Ablation study between self-attention methods

❖ Self-attention 방법 변경에 따른 성능 변화

- Texture 이미지 (3개)와 Object (2개)에 대한 성능 비교 진행
- Dot product 연산에서 간단한 MLP를 사용해서 Object 자체에 집중하는 Self-attention

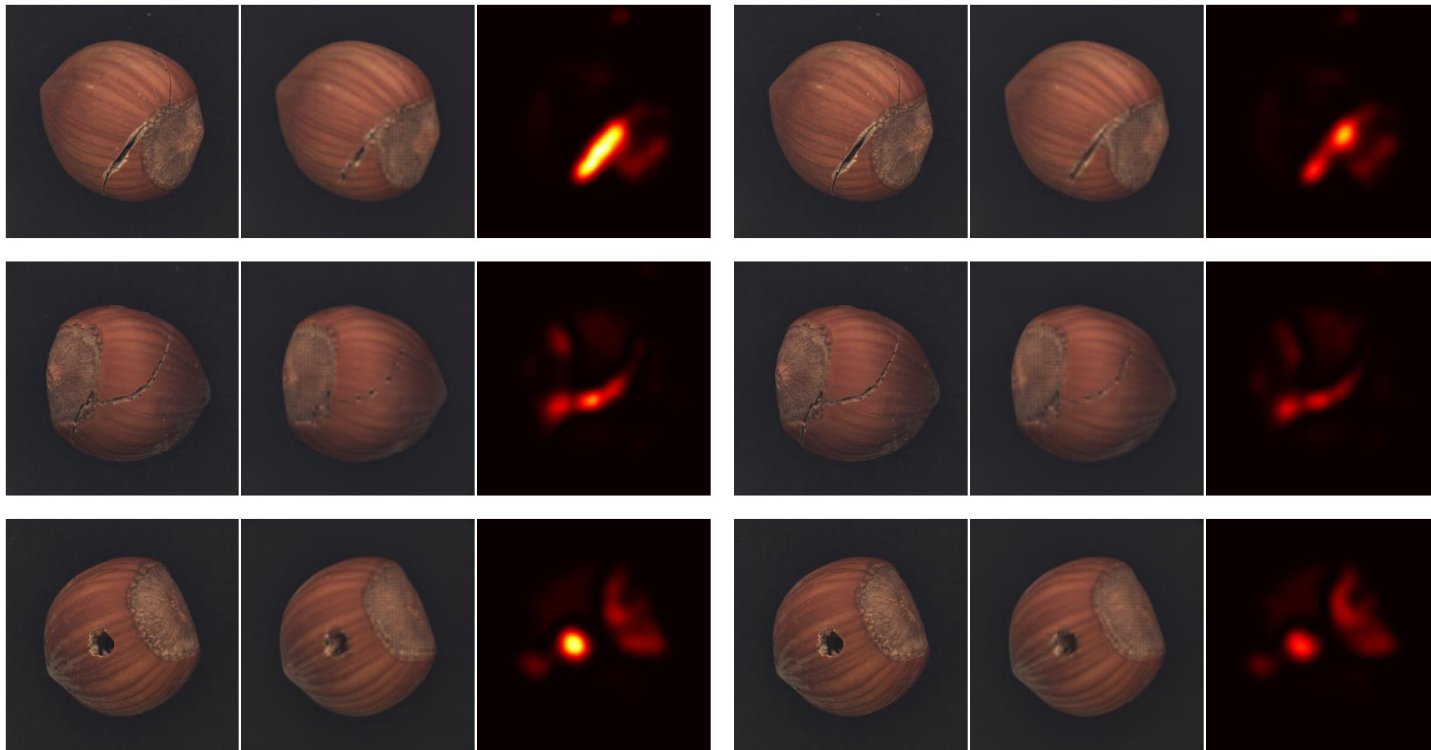
Category	Det. MFSA	Det. MSA	Seg. MFSA	Seg. MSA
Carpet	98.8	98.8	99.2	99.3
Leather	100.0	100.0	99.5	99.5
Wood	96.8	96.8	89.9	89.3
Toothbrush	99.7	98.9	99.0	98.8
Hazelnut	91.0	89.7	97.3	96.6

Experiments

- Ablation study between self-attention methods

❖ Self-attention 방법 변경에 따른 이상 지역 탐지 결과(Anomaly map) 비교

- Dot product 연산에서 간단한 MLP를 사용해서 Object 자체에 집중하는 Self-attention 으로 변경
- 왼쪽은 MFSA를 사용한 경우이며 오른쪽은 MSA를 사용한 경우
- Object 내 이상 영역을 상대적으로 더 높게 이상으로 탐지 (Anomaly map 내 색상이 노란색)



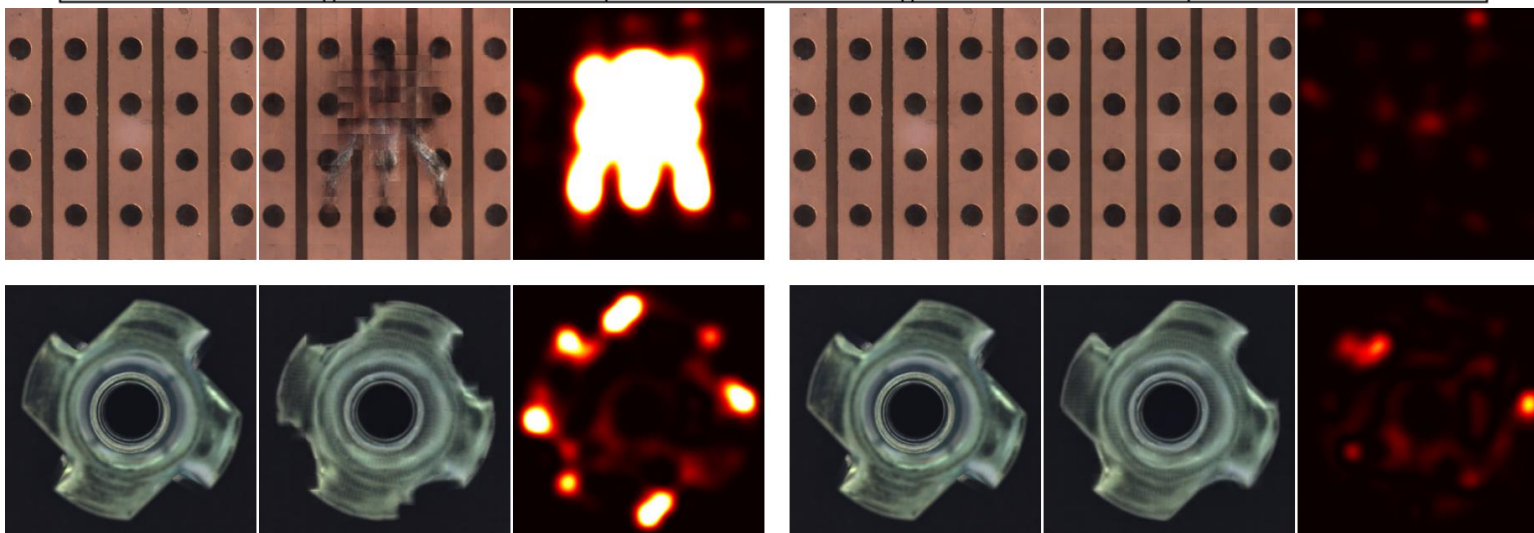
Experiments

- Ablation study for differences between the positional embedding methods

❖ Positional embedding 방법에 따른 성능 비교

- Object 이미지에 대해서는 Global embedding이 효과적
- Texture 이미지에 대한 성능 비교가 없어 정확한 판단을 내리기 어려움
- 왼쪽은 Global embedding을 사용한 경우이며 오른쪽은 local embedding을 사용한 경우

Category	Det. Local	Det. Global	Seg. Local	Seg. Global
Transistor	80.4	95.8	82.6	96.1
Metal Nut	86.4	92.5	77.4	80.5



Conclusion

❖ Conclusion

- 이상 탐지 및 이상 영역 탐지 문제에 Vision transformer 모델을 적용
- 특정 이미지를 Patch로 자른 뒤 특정 Patch를 제거(모든 값을 0)하고 입력 변수로 사용
- 제거된 Patch를 출력 변수로 하며 transformer block으로 구성된 Encoder-decoder 구축
- Transformer block 내 기존 Multi-head attention을 Multi-head Feature attention으로 변경해 성능↑
- 해당 블록 내 Residual connection을 사용해 성능 향상 및 이상 영역을 정확히 인식

❖ InTra를 읽은 뒤 나의 생각

- Ablation study가 Texture별, Object별로 더 추가된다면 제안하는 모듈 효율성에 대해 신뢰 가능
 - 일부에 대한 Ablation이 진행되어 신뢰하기는 어려움
 - 다만 Transformer를 이상 탐지 및 이상 지역 탐지에 최초로 적용한 것에 큰 의의가 있다는 생각
- InTra를 스마트 팩토리 내 삽입하여 실시간으로 확인하기는 어려울 것
 - Anomaly detection and segmentation(localization)이 삽입되기 위해선 실시간 예측이 필수적이라는 생각
 - 이상 탐지 및 이상 지역 탐지 성능은 뛰어나지만 Inference 과정을 가볍게 할 필요가 있다는 생각

References

- Pimay, J., & Chai, K. (2021). Inpainting Transformer for Anomaly Detection. arXiv preprint arXiv:2104.13897.
- Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9592-9600).
- Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
- Xue, W., Zhang, L., Mou, X., & Bovik, A. C. (2013). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE Transactions on Image Processing, 23(2), 684-695.
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

Thank You