Inpainting Transformer for Anomaly Detection

Introduction

作者为读者提供了异常检测领域的背景知识，以及他们提出的方法的动机和技术细节。

- **实际挑战：** 在实际工业应用中，异常是罕见事件，且由于异常可能具有不可预测的形状和纹理，这使得使用监督学习方法来处理异常检测问题变得困难。
- **无监督方法：** 鉴于监督方法的局限性，当前的研究通常采用无监督方法，这些方法只尝试模拟正常数据的分布。在测试阶段，每个图像会根据其与正常样本的偏差获得一个异常分数。
- **深度卷积自编码器：** 一种常见的无监督方法是使用深度卷积自编码器或生成模型（如对抗网络）来模拟正常训练数据的流形。然后，输入图像与重建图像之间的差异被用来计算异常分数。
- **自编码器的局限性：** 尽管这种方法在实践中常常有效，但它存在一个缺点，即卷积自编码器的泛化能力强，可能会错误地重建异常区域，导致漏检。
- **修复（Inpainting）方法：** 为了解决这个问题，最近的方法提出了将生成部分视为一个修复问题，即部分输入图像被覆盖，模型被训练以自监督的方式重建被覆盖的部分。这种方法可以有效地修复小的异常。
- **全局上下文的重要性：** 然而，由于全卷积神经网络（CNNs）的接受场有限，它们在模拟远距离上下文信息方面可能效果不佳，这使得移除较大的异常区域变得困难。
- **自注意力模型的启发：** 作者受到自注意力模型（如Transformers）在图像识别中取得成功的启发，提出了一种新的基于Transformer的方法，用于通过修复问题进行异常检测。

贡献：

- **Inpainting Transformer (InTra) 提出：** 作者提出了一种新的基于Transformer的模型，用于计算机视觉中的异常检测任务。该模型通过修复（inpainting）的方法来识别和定位图像中的异常区域。
- **自注意力机制的应用：** InTra模型摒弃了传统的卷积神经网络（CNNs）结构，完全采用自注意力机制来捕捉图像中不同区域之间的长距离依赖关系，这有助于更好地重建图像并识别异常区域。
- **全局上下文整合：** 通过将图像分割成补丁序列，并在Transformer模型中使用这些补丁序列，InTra能够在修复过程中整合来自图像大范围区域的信息，从而提高对异常区域的检测能力。
- **位置嵌入的使用：** 通过在补丁序列中加入位置嵌入，InTra能够在全局上下文中进行修复，即使补丁序列没有覆盖整个图像。

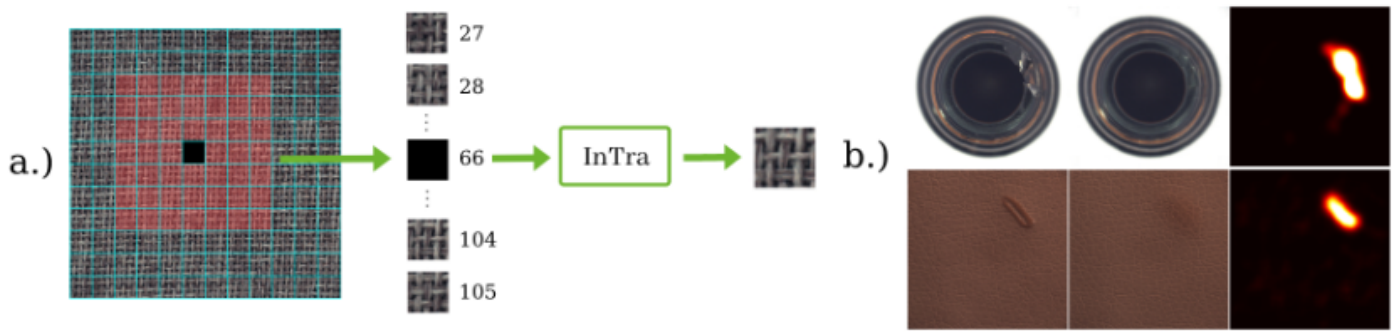
Related Work

作者探讨了与异常检测相关的现有工作，特别是基于重建的方法。

- **基于重建的异常检测方法：** 这些方法主要尝试对正常、无缺陷的样本进行建模。深度卷积神经网络（CNN）自编码器被广泛用于学习无缺陷图像的流形。在测试有缺陷的数据时，这些模型应该无法正确重建异常图像，因为它们只模拟了正常数据。
  局限性： 尽管自编码器在训练时仅使用无缺陷的样本，但它们通常能够很好地泛化到实际中的异常情况。可以使用修复（inpainting）方案来有效地隐藏异常区域，从而限制模型重建异常的能力。
- **U-Net架构：** Zavrtanik等人提出了一种使用U-Net架构的基于重建的方法，该架构利用了长残差连接。他们的方法随机选择图像的多个部分进行修复，为不同基准的异常检测通过修复提供了当前最佳的结果。
- **Transformer模型：** 在Transformer模型中，自注意力机制用于关联序列中的元素，能够全局处理长期依赖关系。Dosovitskiy等人提出了Vision Transformer，将图像数据分割成正方形非重叠的均匀补丁，并将每个补丁和位置嵌入到潜在空间中，将每张图像视为这些嵌入补丁的序列。应用Transformer架构在重构数据上，达到了与最先进的CNN相当甚至在某些任务上超越的结果，同时减少了模型偏差。
  注意力机制的挑战： 尽管理论上自注意力框架可以缓解全卷积自编码器的局限性，但在没有进一步简化的情况下，直接在整张图像上运行自注意力是不可行的。

Inpainting Transformer for Anomaly Detection

作者详细介绍了他们提出的Inpainting Transformer（InTra）模型，这是一种用于异常检测的基于Transformer的方法。模型如下图所示：



补丁嵌入和多头特征自注意力（Patch Embeddings and Multihead Feature Self-attention）：

将输入图像$X \in \mathbb{R}^{H\times W\times C}$裁剪为$N \times M$个小格子（补丁）

$$X_p \in \mathbb{R}^{(N\times M)\times(K^2\cdot C)}$$

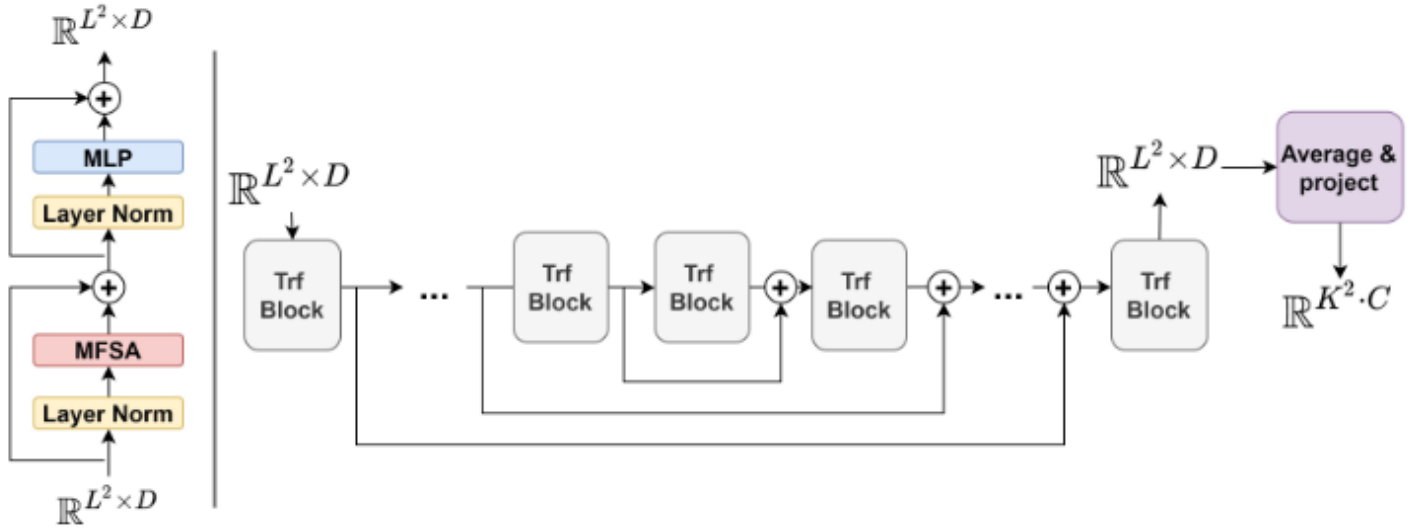
注：$X_p^{(i,j)} \in \mathbb{R}^{K^2\cdot C}$表示这个补丁在第i行和第j列。
目标：使用其他补丁的内容和位置去修复被覆盖的补丁。
位置信息使用$f(i,j) = (i-1)\cdot N + j$来表示
将补丁窗口以及位置信息映射到维度为D的某个潜在空间，对于每一个补丁$x_p^{(x,j)}$可以表示为：

$$y^{(i,j)} := x_p^{(i.j)}\mathbf{E} + posemb(f(i,j)) \in \mathbb{R}^D$$

注：$\mathbf{E}$表示需要学习的参数矩阵，并且$\mathbf{E} \in \mathbb{R}^{(k^2\cdot C)\times D}$,同时'posemb'表示需要学习的一维位置词向量（position embeddings）
为解决在训练图像的补丁非常相似但不明确的情况下，查询q和密钥K的点积非常接近，导致多头自注意力中几乎均匀的softmax加权和，使用多层感知机来代替标准的线性映射，以进行非线性降维。

网络架构和训练（Network Architecture and Training）：

作者说明了网络架构和训练过程，网络如下图所示：



1.**网络架构：** InTra的网络架构由多个Transformer块堆叠而成。每个Transformer块主要由两部分组成：多头特征自注意力（MFSA）和多层感知机（MLP）。这些块的结构遵循了标准的Transformer架构，但是在计算查询（queries）、键（keys）和值（values）时，作者使用了多层感知机来代替标准的线性映射，以进行非线性降维。

- **多头特征自注意力（MFSA）：** MFSA是标准的多头自注意力（MSA）的一个变体，它通过在计算查询和键时使用多层感知机来进行非线性降维，以提高模型对输入数据的处理能力。
- **多层感知机（MLP）：** 每个Transformer块中的MLP包含一个隐藏层，使用GELU激活函数，并将数据从维度D映射到4D，然后再映射回D。MLP用于在MFSA之后进一步处理数据。
- **长残差连接：** 为了将早期层的信息传递到网络的深层，作者在网络中引入了长残差连接，类似于U-Net的结构。这种设计有助于保留更多的空间上下文信息，并改善重建的结构细节。
- **层归一化：** 在MFSA之前应用了层归一化（Layer Normalization），以稳定训练过程。

2.**训练过程：** 网络通过随机采样正常图像数据中的补丁窗口进行训练。在每个窗口中，随机选择一个补丁位置进行修复。训练的目标是最小化原始补丁和重建补丁之间的差异，使用像素级L2损失函数，同时考虑结构相似性（SSIM）和梯度幅度相似性（GMS）

- **损失函数：** 损失函数由三部分组成：L2损失、结构相似性损失和梯度幅度相似性损失。这些损失的组合用于评估原始和重建补丁之间的差异。
- **模型选择：** 在训练过程中，作者使用验证集来监控模型的重建质量，并选择在验证损失上表现最佳的模型权重用于最终评估。

推理和异常检测（Inference and Anomaly Detection）：

推理分为两个步骤。首先，基于输入图像生成一个完整的修复（inpainted）图像。然后，使用原始图像和重建图像之间的差异来计算像素级的异常图。

- **生成修复图像：** 对于输入图像，首先将其分割成N×M的补丁网格。对于每个补丁位置（t, u），选择一个适当的补丁窗口（尽可能使补丁位于窗口中心），其边长为L，用于修复该位置的补丁。通过这种方式，可以重建整个图像。
- **计算异常图：** 使用原始图像和重建图像之间的差异来生成异常图。具体来说，对于原始图像和重建图像在不同尺度下的梯度幅度相似性（GMS）差异，应用一个平滑操作（包括平均滤波和高斯模糊）来增强对小的、重建不良的异常区域的鲁棒性。

Experiments

- 数据集
  选择MVTec AD数据集进行模型评估，该模型包含5中纹理和10种对象类别的高分辨率样本
- 指标
  使用ROC AUC作为检测的指标
  1.实验细节
  1）从正常图像随机选10%的图片（不超过200），在每一个epoch中，每个图像会随机采样600个

补丁窗口。为了扩充数据集，进行了数据增强操作（旋转和翻转）
2）因为patch长度k和patch窗口长度L对模型有比较大的影响，统一使用（k=16，L=7）
3）为图像重构质量和计算时间的平衡，使用图像尺度256×256、320×320、512×512在200epoch训练，以展示最优结果
4）使用消融实验，以评估长残差连接、多头特征自注意力（MFSA）的使用，以及补丁窗口大小对模型性能的影响。证明了MFSA（多头特征自注意力）的有效性，比一般的MSA（多头自注意力机）有效