ここで、i はチャネルのインデックスを表し、u は空間位置のインデックス（簡素化のため高さおよび幅を意味する）です。異常局所化は異常領域を特定し、各ピクセル u に対して異常スコアを割り当てる異常スコアマップ s(u) を生成します。s(u) は特徴量差分ベクトル d(:, u) の L2 ノルムとして計算されます。

$$\boldsymbol{s}(u) = ||\boldsymbol{d}(:, u)||_2. \tag{3}$$

異常検出は、画像に異常な領域が含まれているかどうかを検出することを目的とします。直感的に、平均プールされたs(u)の最大値を画像全体の異常スコアとして採用します。

3.2 トランスフォーマーによる「同一マッピング」の防止

CNNと比較して、注意層のクエリ埋め込みがTransformerが「同一マッピング」を学習しにくくすると推測されます。正常な画像の特性を$x^+ \in_R^{K\times C}$ と表し、ここでKは特性数、Cはチャネル次元です。異常画像の特性を$x^- \in_R^{K\times C}$と表します。再構築ネットワークとして1層のネットワークを採用し、x^+に対してMSE損失で訓練し、x^-の異常領域を検出するためにテストします。
CNNの畳み込み層。まず、全接続層を訪れます。その重みとバイアスはそれぞれw $\in_R^{C\times C}$、b $\in_R^C$で表されます。この層を正常サンプルの再構築モデルとして使用する場合、次のように表せます。

$$\hat{\boldsymbol{x}} = \boldsymbol{x}^+\boldsymbol{w} + \boldsymbol{b} \in \mathbb{R}^{K\times C}. \tag{4}$$

MSE損失関数がx^をx⁺に押し上げるため、モデルはw → I（単位行列）、b → 0への近道を取る可能性があります。最終的に、このモデルはx^-を適切に再構築できるため、異常検出に失敗する可能性があります。1×1カーネルを持つ畳み込み層は、全接続層と等価です。さらに、n × n (n > 1) のカーネルはより多くのパラメーターと大きな容量を持ち、1×1カーネルが可能なことはすべて実行できます。したがって、畳み込み層もショートカットを学習する可能性があります。
クエリ埋め込みを含むトランスフォーマーは、学習可能なクエリ埋め込みq $\in_R^{K\times C}$ を持つ注意層を含みます。この注意層は次のように表せます：

$$\hat{\boldsymbol{x}} = \mathtt{softmax}(\boldsymbol{q}(\boldsymbol{x}^+)^T/\sqrt{C})\boldsymbol{x}^+ \in \mathbb{R}^{K\times C}. \tag{5}$$

x^をx⁺に押し上げるため、アテンションマップ（softmax(q$(x^+)^T$ ／ C）はI（単位行列）に近似する必要があり、qはx⁺と密接に関連している必要があります。訓練されたモデルでqが正常なサンプルに関連しているため、モデルはx^-を適切に再構築できません。第4.4節の消去実験では、注意層またはクエリ埋め込みを削除すると、トランスフォーマーの性能がそれぞれ2.4%または3%低下し、これはCNNとほぼ同じです。これは、注意層のクエリ埋め込みがトランスフォーマーが「同一のショートカット」を学習するのを防ぐのに役立つことを示しています。

3.3 異常データ利用可能ケースへの適応

実践では、生産ラインの稼働に伴い異常が徐々に増加するため、これらの増加する異常に対応する互換性が求められます。そのため、ADTRをADTR+に拡張し、異常データが利用可能なケースに対応できるように適応させます。

where i represents the index of channel, u is the index of spatial position (height together with width for simplicity). *Anomaly localization* aims to localize anomalous regions, producing an anomaly score map, $\boldsymbol{s}(u)$, which assigns an anomaly score for each pixel, u. $\boldsymbol{s}(u)$ is calculated as the $L2$ norm of the feature difference vector, $\boldsymbol{d}(:,u)$.

$$\boldsymbol{s}(u) = ||\boldsymbol{d}(:,u)||_2. \tag{3}$$

Anomaly detection aims to detect whether an image contains anomalous regions. We intuitively take the maximum value of the averagely pooled $\boldsymbol{s}(u)$ as the anomaly score of the whole image.

3.2 Preventing "Identical Mapping" with Transformer

We suspect that, compared with CNN, the query embedding in attention layer makes transformer difficult to learn an "identical mapping". We denote the features in a normal image as $\boldsymbol{x}^+ \in \mathbb{R}^{K\times C}$, where K is the feature number, C is the channel dimension. The features in an anomalous image are denoted as $\boldsymbol{x}^- \in \mathbb{R}^{K\times C}$. We take a 1-layer network as the reconstruction net, which is trained on $\boldsymbol{x}^+$ with the MSE loss and tested to detect anomalous regions in $\boldsymbol{x}^-$.

Convolutional layer in CNN. We first visit a fully-connected layer, whose weights and bias are denoted as $\boldsymbol{w} \in \mathbb{R}^{C\times C}, \boldsymbol{b} \in \mathbb{R}^C$, respectively. When using this layer as the reconstruction model of normal samples, it can be written as,

$$\hat{\boldsymbol{x}} = \boldsymbol{x}^+\boldsymbol{w} + \boldsymbol{b} \in \mathbb{R}^{K\times C}. \tag{4}$$

With the MSE loss pushing $\hat{\boldsymbol{x}}$ to $\boldsymbol{x}^+$, the model may take shortcut to regress $\boldsymbol{w} \to \boldsymbol{I}$ (identity matrix), $\boldsymbol{b} \to \boldsymbol{0}$. Ultimately, this model could also reconstruct $\boldsymbol{x}^-$ well, failing in anomaly detection. A convolutional layer with 1×1 kernel is equivalent to a fully-connected layer. Besides, An $n \times n$ ($n > 1$) kernel has more parameters and larger capacity, and can complete whatever 1×1 kernel can. Thus, the convolutional layer also has the chance to learn a shortcut.

Transformer with query embedding contains an attention layer with a learnable query embedding, $\boldsymbol{q} \in \mathbb{R}^{K\times C}$. This attention layer can be denoted as,

$$\hat{\boldsymbol{x}} = \texttt{softmax}(\boldsymbol{q}(\boldsymbol{x}^+)^T/\sqrt{C})\boldsymbol{x}^+ \in \mathbb{R}^{K\times C}. \tag{5}$$

To push $\hat{\boldsymbol{x}}$ to $\boldsymbol{x}^+$, the attention map, $\texttt{softmax}(\boldsymbol{q}(\boldsymbol{x}^+)^T/\sqrt{C})$, should approximate $\boldsymbol{I}$ (identity matrix), so $\boldsymbol{q}$ must be highly related to $\boldsymbol{x}^+$. Considering that $\boldsymbol{q}$ in the trained model is relevant to normal samples, the model could not reconstruct $\boldsymbol{x}^-$ well. The ablation study in Sec. 4.4 shows that without the attention layer or the query embedding, the performance of transformer respectively drops by 2.4% or 3%, which is almost the same as CNN. This reflects that the query embedding in attention layer helps prevent transformer from learning an "identical shortcut".

3.3 Adaptation with Anomaly-available Case

In practice, anomalies gradually increase with the runs of production lines, which brings the demands of compatibility with these increasing anomalies. Thus we adapt ADTR to ADTR+ for compatibility with the anomaly-available case.