DIPARTIMENTO DI INGEGNERIA INFORMATICA, MODELLISTICA,
ELETTRONICA E SISTEMISTICA

Corso di Laurea Magistrale in Ingegneria Informatica

Machine & Deep Learning

Elaborato Finale

Re-Engineering AE-XAD with Vision Transformers for Explainable Anomaly Detection

Docenti:

Prof. **Fabrizio Angiulli**
Ing. **Francesco De Luca**

Studente:
Presta Vincenzo
matr. 252290

ANNO ACCADEMICO 2024/2025

Indice generale

1 Introduzione

L'Anomaly Detection in ambito industriale rappresenta un problema di grande rilevanza applicativa, poiché consente l'individuazione automatica di difetti e anomalie su superfici e componenti prodotti in serie. In molti scenari reali, le anomalie risultano rare, eterogenee e difficilmente annotabili in modo esaustivo, rendendo complessa l'applicazione di approcci supervisionati tradizionali. Per questo motivo, si sono affermati metodi basati su Autoencoder, capaci di apprendere una rappresentazione delle sole istanze normali e di identificare le anomalie come deviazioni rispetto al comportamento appreso.

All'interno di questo paradigma, i metodi di anomaly detection basati su ricostruzione hanno dimostrato particolare efficacia in contesti industriali, soprattutto quando è richiesta una localizzazione spaziale dei difetti. In tale direzione si colloca il framework AE-XAD, che introduce una pipeline strutturata per la rilevazione e la localizzazione delle anomalie attraverso l'analisi dell'errore di ricostruzione [1]. Il metodo combina un encoder convoluzionale, un decoder progettato per enfatizzare le discrepanze rispetto alla normalità e un meccanismo di decisione basato su statistiche globali pixel-wise, ottenendo risultati competitivi sia a livello di immagine che a livello di localizzazione.

Negli ultimi anni, l'evoluzione delle architetture di visione ha portato all'emergere di modelli basati su meccanismi di attenzione globale, come i Vision Transformer, che hanno mostrato elevate capacità di rappresentazione in numerosi compiti di visione artificiale. Questo progresso solleva naturalmente l'interrogativo se tali architetture possano sostituire efficacemente le reti convoluzionali anche all'interno di pipeline di anomaly detection basate su ricostruzione.

Tuttavia, l'integrazione di un Vision Transformer all'interno del framework AE-XAD non è immediata. AE-XAD non è un autoencoder generico, ma un metodo che fa affidamento su specifiche assunzioni strutturali riguardanti la distribuzione spaziale dell'errore di ricostruzione e sulla sua separabilità statistica dal rumore di fondo. In questo contesto, la sostituzione dell'encoder convoluzionale con un'architettura caratterizzata da un diverso tipo di rappresentazione solleva interrogativi fondamentali sulla compatibilità tra il modello di features apprese e il meccanismo decisionale adottato.

L'obiettivo di questa tesi è quindi analizzare in modo sistematico se, e in quali condizioni, un Vision Transformer possa sostituire l'encoder convoluzionale originale di AE-XAD mantenendo inalterati il decoder, la funzione di loss, le metriche di valutazione e l'intera pipeline di test, al fine di garantire un confronto equo e scientificamente rigoroso.

1.1 Definizione del Problema e Ipotesi di Ricerca

Il framework AE-XAD assume implicitamente che le anomalie producano errori di ricostruzione *spazialmente localizzati e compatti*, tali da poter essere distinti dal rumore di fondo mediante una soglia statistica globale basata su media e deviazione standard ($\mu + \sigma$). Questa assunzione risulta naturalmente coerente

con le proprietà delle architetture convoluzionali, che favoriscono una rappresentazione gerarchica e localmente strutturata delle informazioni spaziali.

Le architetture basate su attenzione globale, come i Vision Transformer, presentano invece un comportamento rappresentazionale differente, orientato alla modellazione di relazioni globali tra regioni dell'immagine. Sebbene tale caratteristica possa risultare vantaggiosa in compiti di natura semantica, non è immediatamente evidente se essa sia compatibile con un paradigma di anomaly detection basato su errori di ricostruzione pixel-wise e su una sogliatura statistica globale.

La domanda di ricerca che guida questo lavoro può pertanto essere formulata come segue:

> *Un Vision Transformer può sostituire efficacemente l'encoder convoluzionale di AE-XAD, mantenendo inalterata la pipeline decisionale, in un contesto di anomaly detection few-shot supervisionato?*

L'ipotesi investigata in questa tesi è che, pur essendo in grado di apprendere rappresentazioni utili per il ranking delle anomalie, i Vision Transformer tendano a produrre errori di ricostruzione più diffusi e meno localizzati rispetto alle architetture convoluzionali. Di conseguenza, il meccanismo di binarizzazione basato su $\mu + \sigma$ risulterebbe intrinsecamente meno efficace, portando a un degrado delle prestazioni di localizzazione in specifiche classi del dataset considerato.

2 Fondamenti teorici

In questa sezione vengono introdotti i concetti teorici necessari a inquadrare il contesto metodologico di questo lavoro e a chiarire le assunzioni implicite alla base del framework utilizzato. In particolare, viene discusso il paradigma di anomaly detection basato su ricostruzione, viene descritto il framework AE-XAD dal punto di vista concettuale e viene analizzato il ruolo dell'inductive bias nelle architetture di visione, in relazione al meccanismo di decisione adottato.

2.1 Anomaly Detection basata su ricostruzione

Gli approcci di anomaly detection basati su ricostruzione si fondano sull'idea di apprendere un modello delle sole istanze normali, in modo tale che le anomalie possano essere identificate come deviazioni rispetto al comportamento appreso. In questo paradigma, un autoencoder viene addestrato a ricostruire immagini prive di difetti, minimizzando un errore di ricostruzione calcolato a livello pixel-wise.

In fase di test, la presenza di anomalie si traduce tipicamente in un aumento dell'errore di ricostruzione nelle regioni difettose, rendendo possibile l'individuazione delle anomalie sia a livello di immagine sia a livello pixel-wise. L'efficacia di tali approcci dipende pertanto non solo dalla capacità del modello

di rappresentare correttamente la normalità, ma anche dalla struttura spaziale dell'errore di ricostruzione prodotto.

2.2 Il framework AE-XAD

Il framework AE-XAD (AutoEncoder for eXplainable Anomaly Detection) rappresenta un'evoluzione degli approcci classici basati su autoencoder, introducendo una pipeline progettata specificamente per la localizzazione delle anomalie in contesti industriali [1]. Il metodo è composto da un encoder convoluzionale, un decoder con una struttura a rami e un meccanismo di decisione basato su statistiche globali dell'errore di ricostruzione.

Un elemento centrale di AE-XAD è la costruzione di una *reconstruction error map*, ottenuta confrontando l'immagine di input con la ricostruzione prodotta dal decoder. Tale mappa rappresenta la distribuzione spaziale dell'errore di ricostruzione ed è utilizzata come base per la localizzazione delle anomalie. Per ridurre il rumore ad alta frequenza, la mappa viene sottoposta a un'operazione di filtraggio, seguita da una binarizzazione mediante una soglia statistica globale definita come $\mu+\sigma$, dove μ e σ indicano rispettivamente la media e la deviazione standard dei valori di errore.

Questo meccanismo di sogliatura implica che le anomalie siano caratterizzate da errori di ricostruzione significativamente superiori al rumore di fondo e concentrati in regioni spazialmente limitate. Di conseguenza, l'efficacia della pipeline AE-XAD dipende in modo critico dalla distribuzione spaziale dell'errore di ricostruzione prodotto dall'encoder–decoder.

2.3 Inductive bias nelle architetture di visione

Con il termine *inductive bias* si intende l'insieme di assunzioni strutturali che un modello incorpora a priori, influenzando il modo in cui generalizza a partire da un numero limitato di esempi. Nel contesto della visione artificiale, l'inductive bias riveste un ruolo particolarmente rilevante in scenari few-shot, dove la quantità di dati disponibili non è sufficiente a guidare completamente l'apprendimento.

Le architetture convoluzionali, come quelle impiegate in AE-XAD, incorporano un forte inductive bias locale, che favorisce la modellazione di pattern spaziali e la produzione di rappresentazioni gerarchiche sensibili alla localizzazione. Tale caratteristica risulta naturalmente coerente con un paradigma di anomaly detection basato su errori di ricostruzione pixel-wise e su una sogliatura statistica globale.

Architetture caratterizzate da un diverso tipo di rappresentazione, orientate alla modellazione di relazioni globali tra regioni dell'immagine, possono invece produrre distribuzioni dell'errore di ricostruzione più diffuse e meno concentrate. In un framework come AE-XAD, in cui la decisione finale dipende dall'applicazione di una soglia globale a una mappa di errore spaziale, tale differenza rappresentazionale può tradursi in un disallineamento tra le assunzioni della pipeline di decisione e le proprietà dell'encoder utilizzato.

3 Setup sperimentale

In questa sezione vengono descritti il dataset utilizzato, l'architettura del modello, la funzione di loss, le metriche di valutazione e il protocollo sperimentale adottato. Tutti gli esperimenti sono stati condotti mantenendo invariata la pipeline di AE-XAD, al fine di analizzare in modo isolato l'impatto della sostituzione dell'encoder convoluzionale con un Vision Transformer.

3.1 Dataset e preprocessing

Gli esperimenti sono stati condotti sul dataset MVTec AD, ampiamente utilizzato per la valutazione di metodi di anomaly detection in ambito industriale. Il dataset include immagini di componenti e superfici industriali, accompagnate da maschere pixel-wise che annotano le regioni anomale nelle immagini di test.

Seguendo il protocollo adottato nel framework AE-XAD, il dataset viene utilizzato in regime few-shot supervisionato, in cui un numero limitato di campioni anomali è reso disponibile durante la fase di addestramento. Le immagini sono ridimensionate a una risoluzione uniforme e sottoposte a un preprocessing coerente con quello previsto dal metodo originale, senza introdurre trasformazioni aggiuntive che possano alterare la distribuzione spaziale dei difetti.

3.2 Architettura del modello e funzione di loss

Il framework AE-XAD adotta una pipeline di anomaly detection basata su ricostruzione, composta da un encoder $E(\cdot)$, un decoder $D(\cdot)$ e un modulo di decisione a valle [1]. Dato un input $x \in \mathbb{R}^{H\times W\times C}$, l'encoder produce una rappresentazione latente $z = E(x)$, che viene utilizzata dal decoder per ottenere la ricostruzione $\hat{x} = D(z)$.

L'addestramento del modello è guidato da una funzione di loss di ricostruzione definita a livello pixel-wise, che misura la discrepanza tra l'immagine originale e quella ricostruita. In forma generale, la loss può essere espressa come:

$$\mathcal{L}_{rec}(x, \hat{x}) = \frac{1}{HW} \sum_{i,j} \|x_{i,j} - \hat{x}_{i,j}\|,$$

dove (i, j) indicano le coordinate spaziali dei pixel.

Nel regime few-shot supervisionato considerato in AE-XAD, un numero limitato di campioni anomali è reso disponibile durante l'addestramento. Tali campioni contribuiscono alla funzione obiettivo secondo la formulazione proposta nel framework originale, consentendo di guidare l'apprendimento senza alterare la natura reconstruction-based del metodo.

In fase di inferenza, la differenza pixel-wise tra input e ricostruzione produce una *reconstruction error map* $M \in \mathbb{R}^{H\times W}$, definita come:

$$M_{i,j} = \|x_{i,j} - \hat{x}_{i,j}\|.$$

La mappa di errore viene successivamente filtrata per ridurre il rumore ad alta frequenza e binarizzata tramite una soglia statistica globale definita come:

$$T = \mu + \sigma,$$

dove μ e σ rappresentano rispettivamente la media e la deviazione standard dei valori di M.

La localizzazione delle anomalie è infine ottenuta confrontando la mappa binarizzata con le maschere di ground truth fornite dal dataset. In questo lavoro, l'intera pipeline descritta viene mantenuta invariata, sostituendo esclusivamente l'encoder convoluzionale originale con un Vision Transformer, al fine di analizzare l'impatto del diverso inductive bias sulla distribuzione spaziale dell'errore di ricostruzione.

3.3 Metriche di valutazione

Le prestazioni del modello sono valutate utilizzando le metriche previste dal framework AE-XAD. In particolare, vengono considerate metriche di rilevazione a livello di immagine e metriche di localizzazione a livello pixel-wise.

La localizzazione delle anomalie è ottenuta applicando una soglia statistica globale basata su media e deviazione standard ($\mu + \sigma$) alla mappa di errore di ricostruzione filtrata. Le prestazioni di localizzazione sono quindi misurate confrontando la mappa binarizzata con le maschere di ground truth fornite dal dataset.

3.4 Protocollo sperimentale

Per valutare in modo rigoroso l'impatto della sostituzione dell'encoder, gli esperimenti sono condotti modificando un solo componente alla volta. In particolare, decoder, funzione di loss, metriche di valutazione e pipeline di test sono mantenuti invariati rispetto al framework AE-XAD originale.

Questo protocollo consente di attribuire eventuali variazioni nelle prestazioni osservate esclusivamente al tipo di encoder utilizzato, permettendo un'analisi controllata e scientificamente fondata del ruolo dell'inductive bias nelle prestazioni di anomaly detection.

References

[1] F. Angiulli, F. Fassetti, L. Ferragina, and S. Nisticò, "Explaining anomalies through semi-supervised autoencoders," *Array*, 2025.