Weakly Supervised Video Anomaly Detection using Deep Learning

Gaetano Magazzù

Febbraio 2024

1 Introduzione

Negli ultimi anni l'anomaly detection ha assunto sempre maggiore importanza e interesse in diverse applicazioni, come diagnosi mediche, controllo qualità, rilevamento di guasti o difetti. Il rilevamento di anomalie è uno dei *task* più complessi e più studiati nel settore della *computer vision*. La difficoltà risiede principalmente nel distinguere immagini normali da immagini anormali, poiché le anomalie possono essere ambigue e avvenire in modo imprevedibile e presentarsi in diversi contesti. Un'anomalia può essere definita come quegli esempi o eventi rari che si discostano significativamente dalla maggior parte dei dati, ossia dai casi normali [2]. Nel contesto della *video anomaly detection* si tratta principalmente di individuare attività o comportamenti inusuali, da parte dei soggetti ripresi in video di sorveglianza. Esempi di anomalie includono veicoli che attraversano luoghi non autorizzati, persone che corrono in zone vietate, abusi, atti vandalici, furti, incidenti stradali, incendi o esplosioni.

Recentemente, visto l'aumentare del numero di videocamere impiegate oggigiorno in strade, negozi, banche ed altri luoghi pubblici, al fine di incrementare la pubblica sicurezza è emersa la necessità di sviluppare metodi per il rilevamento di anomalie anche su video. Tipicamente questo *task* veniva trattato come problema di classificazione ad una classe in modo non supervisionato [2]. Tuttavia, data la mancanza e la difficoltà di ottenere dati etichettati, recentemente sempre più metodi stanno ottenendo successo utilizzando un approccio debolmente supervisionato (anche detto *weakly supervised*). Il *task* di *weakly supervised video anomaly detection* può essere definito quindi come un problema di regressione, dove a ogni frame di un video viene assegnato uno *score* di anomalia. Uno *score* alto indica un'anomalia, mentre uno score basso la normalità. L'unica informazione disponibile durante l'addestramento è sapere se il video contenga o meno anomalie, senza conoscere la loro posizione temporale.

I dati trattati sono solitamente ottenuti da telecamere di video sorveglianza che presentano scene reali in ambienti non controllati, cosiddetti *in-the-wild*. Ovvero scene con condizioni molto variabili e imprevedibili come diverse angolature, rotazione, scala, risoluzione, illuminazione e occlusioni, le quali complicano il riconoscimento delle anomalie. In particolare, rappresenta una grossa sfida gestire i cosiddetti video *untrimmed*, i quali sono video senza tagli o modifiche temporali, che mantengono la loro durata originale. Questo comporta avere video di lunga durata senza una chiara segmentazione temporale delle anomalie, con numerosi frame normali e pochi frame anomali. Il che rende molto costoso reperire dati etichettati e causa uno sbilanciamento dei dati. Ulteriori sfide dipendono dal tipo di anomalia e dalla durata. Ad esempio, riconoscere anomalie di breve durata può risultare problematico. Inoltre, all'interno di un video possono essere contenute diverse anomalie e avvenire più volte. Un'ulteriore aspetto da considerare nel *task* di rilevamento di anomalie è il contesto sia spaziale che temporale. Ad esempio, azioni normali come correre, camminare o guidare un auto, possono essere anomalie in base alle circostanze.

L'obiettivo della tesi è quello di effettuare un attenta analisi dello stato dell'arte, proporre e sviluppare una soluzione per il rilevamento delle anomalie nei video con approccio debolmente supervisionato a grana fine, considerando l'importanza delle singole *clip* del video. La valutazione del metodo proposto sarà effettuata considerando i tre principali *dataset* di *benchmark*.

2 Stato dell'Arte

Come stabilito da Joo et al. [1], i metodi che affrontano il riconoscimento di anomalie nei video, possono essere categorizzati in base al metodo di apprendimento: non supervisionato, supervisionato e debolmente supervisionato. Per i metodi non supervisionati un approccio molto comune è impostare il problema come una classificazione ad una classe, sfruttando solo video normali durante l'addestramento e definendo le anomalie in base alla differenza rispetto agli esempi normali [1]. Uno dei primi approcci impiega *feature hand-crafted* e modelli *SVM one-class*. Successivamente, sono stati proposti metodi di *Sparse Dictionary Learning*, che apprendono un dizionario che cattura i *pattern* normali attraverso la ricostruzione con vettori sparsi. Tuttavia, ottenere questi vettori risulta molto oneroso. Altri approcci usano un'architettura *AutoEncoder*, ma spesso generalizzano così bene da ricostruire le anomalie con un basso errore, causando potenziali errori di classificazione [7]. In seguito, per ottenere ricostruzioni più raffinate sono impiegate delle GAN(*Generative Adversarial Network*). Infine, gli approcci recenti integrano moduli di memoria negli *AutoEncoder*, per catturare meglio i *pattern* normali, concentrandosi esclusivamente sui quelli più rilevanti attraverso meccanismi di *attention*. Nonostante i vari progressi, gli approcci non supervisionati sono ormai superati dai metodi debolmente supervisionati [5].

Sebbene i metodi supervisionati riescano a ottenere risultati migliori, la loro praticità è limitata dalla scarsità di dati etichettati con grana fine [1]. Per questo motivo sono presenti pochi metodi di questo tipo, e quelli esistenti etichettano manualmente i dati.

I metodi debolmente supervisionati rappresentano l'approccio più pratico per il rilevamento delle anomalie nei video, in quanto a differenza dei metodi supervisionati, vengono considerate esclusivamente etichette *video-level* che non sono molto costose da reperire, e rispetto ai metodi non supervisionati [1] le performance ottenute sono migliori.

Uno dei lavori più influenti per il rilevamento delle anomalie con approccio debolmente supervisionato è quello di Sultani et al. [8], il quale introduce il *deep* MIL*(Multiple Instance Learning) ranking framework*. L'addestramento prevede l'utilizzo di coppie di video normali ed anormali. I video sono suddivisi in frammenti, i quali formano una *bag* positiva e una *bag* negativa, per rispettivamente i video anomali e normali. L'obiettivo è massimizzare la differenza tra gli *score* di anomalia della la *bag* positiva e della *bag* negativa, assumendo che lo *score* più alto nella *bag* positiva corrispondi ad un frammento anomalo. I metodi *MIL* soffrono di alcuni problemi come la scelta delle istanze anomale nella *bag* positiva e la classificazione della cosiddetta *hard instance*. Nello stato dell'arte sono stati introdotti accorgimenti e vincoli per cercare di ridurre tali problemi. Un altro filone molto diffuso nello stato dell'arte dei metodi debolmente supervisionati è quello di adottare strategie di *self-training*, generando delle *pseudo-etichette* per rifinire gli *score* di anomalia, oppure supervisionando interamente l'apprendimento. Altri propongono metodi ibridi reintroducendo tecniche usate nei metodi non supervisionati, al fine di cercare e correggere le predizioni rumorose ottenute dai metodi debolmente supervisionati.

Un problema affrontato in vari studi, è il *domain gap*, causato dall'uso di *feature* pre-addestrate per il riconoscimento di attività. Per gestire questo problema, alcuni metodi addestrano sia il classificatore che le *feature*, anche noti come *Encoder-Based*, per ottenere *feature* specifiche per il *task*, al contrario dei metodi *Encoder-Agnostic*, che non effettuano il *fine-tuning* delle *feature* [4]. Un'altra opzione è l'uso di *feature* pre-addestrate più sofisticate come

CLIP(*Contrastive Language Image Pre-Training*), che essendo addestrato su oltre 400 milioni di coppie immagine-testo, è capace di generalizzare su diversi contesti. Notevoli sono inoltre gli avanzamenti dei vari lavori nel gestire il contesto, dalle convoluzioni dilatate alle GCN(*Graph Convolutional Network*), fino ai moduli di *self-attention* e ai metodi più recenti basati su *Trasformer*, che cercando di catturare in modo sempre più preciso il contesto spazio-temporale, sia locale che globale. Una delle principali caratteristiche dei metodi *weakly supervised* è che essi effettuano esclusivamente il rilevamento temporale delle anomalie, e prevedono solitamente un approccio *Encoder-Agnostic* basato su MIL, il quale cattura il contesto spazio-temporale tra i frammenti. Inoltre, tipicamente il rilevamento avviene offline, considerando l'intero video a disposizione. Sebbene alcuni metodi sono in grado di processare una *clip* in *real-time*, non sono pensati per gestire uno *streaming* video. Un'altra importante osservazione riguarda il livello di granularità utilizzato per predire gli *score* di anomalia, ovvero la scelta tra frammenti o *clip*. Wan et al. [6] evidenziano che considerare frammenti rende difficile l'identificazione di anomalie brevi, poiché le istanze anormali nell'intero frammento potrebbero essere sopraffatte dalla predominanza di elementi normali. Nonostante ciò, la maggior parte dei lavori, anche recenti, suddivide i video considerando i frammenti.

3 Lavoro Svolto

Come punto di partenza si è scelto il lavoro Tian et al. [5] e analizzando le sue debolezze sono state effettuate delle modifiche al fine di definire una nuova soluzione. Tian et al. [5] propongono un modello *Encoder-Agnostic* basato su MIL. Questo impara delle feature temporali tramite una MTN(*Multi-Scale Temporal Network*), catturando sia dipendenze locali e globali tra i frammenti. Durante l'addestramento, un classificatore assegna ai *top-k* frammenti, scelti in base alla magnitudine media delle *feature* più alta, degli *score* di anomalia, mentre nella fase di *test* invece, viene predetto uno *score* per ogni *clip*. Infine, si utilizza una *Feature Magnitude Loss* per aumentare la separazione tra video normali e anormali, massimizzando le magnitudini dei *top-k* frammenti anomali e minimizzando quelli normali. Questo modello divide i video in 32 frammenti di dimensioni uguali, indipendentemente dalla loro lunghezza. Questo può causare la perdita di informazione, specialmente nei video lunghi e sopraffare le *clip* anomale, in quanto le *clip* di ogni frammento vengono mediate senza considerare l'importanza delle singole *clip*.

La prima modifica apportata al modello consiste nel considerare le singole *clip* anche durante la fase di addestramento. Questo permette di riconoscere dettagli più fini, individuare anomalie più brevi e prevenire che le *clip* anomale siano sopraffatte da quelle normali [6]. Per far ciò, durante la fase di addestramento si introduce un campionamento uniforme di un numero fisso di *clip* per ogni video. Un'ulteriore modifica consiste nell'introdurre un *task* ausiliare di video classification, al fine di aiutare il più complesso *task* di rilevamento di anomalie. Infine, ispirati dalla *speaker recognition*, viene introdotto un *layer* SAP(*Self Attentive Pooling*) per aggregare l'informazione temporale considerando l'importanza delle singole *clip*. Il metodo così definito rappresenta una soluzione innovativa per il *task* trattato.

Seguendo lo stato dell'arte, vengono considerate le feature C3D e I3D e i principali *dataset* di *benchmark* per il riconoscimento delle anomalie con le loro rispettive metriche, ovvero AUC per Shanghai-Tech e UCF-Crime, e AP su XD-Violence. L'addestramento segue l'approccio comune con *train set* basato su etichette *video-level* e *test set* con etichette *frame-level*. Il modello migliore è scelto in base al valore più alto della metrica di valutazione sul *test set*. Come eseguito da altri lavori allo stato dell'arte [6, 5, 2], vengono effettuate diverse prove per scegliere gli iperparametri migliori sul *test set*, utilizzando una una *grid search*.

4 Risultati

Il modello proposto supera il modello base su tre *dataset* di *benchmark*. Sul *dataset* Shanghai-tech, migliora di +5,22% con C3D e +0,77% con I3D, ottenendo AUC rispettivamente di 97,03 e 97,98. Sul *dataset* UCF-Crime, le performance migliorano di +1,77% e +1,26% con AUC di 85,05 e 85,56 per rispettivamente C3D e I3D. Sul *dataset* XD-Violence invece, migliora di +2,95% raggiungendo un AP di 80,76 con I3D, mentre peggiora di -0,41%, ottenendo un AP di 75,48 con C3D. Per quanto riguarda il *task* di *video classification*, si è dimostrato essere più semplice con performance del modello quasi sempre superiori rispetto alla controparte del *task* di *video anomaly detection* con valori di AUC e AP sempre superiori al 90%.

Confrontando con lo stato dell'arte, il modello proposto raggiunge il quarto posto nel *dataset* Shanghai-Tech con *feature* I3D, con uno scarto di -0,68% dal primo in classifica con *feature* I3D+CLIP, il quale ottiene un AUC di 98,66. Considerando le singole *feature* C3D e I3D, il modello si colloca al secondo posto rispettivamente con scarti di -0,14% e -0,16% dai modelli migliori con 97,19 e 98,14 di AUC. Nel *dataset* UCF-Crime invece, il modello proposto ottiene il suo risultato migliore con la *feature* I3D con uno scarto di -3,41% rispetto al modello migliore, con *feature* I3D+CLIP ottenendo 88,97 di AUC. Considerando le singole *feature*, con C3D il modello si colloca al secondo posto con uno scarto di di -0,73% rispetto al modello migliore che ottiene 85,78 di AUC, mentre con I3D uno scarto di -1,42, rispetto al modello migliore che raggiunge un AUC di 86,98. Nel *dataset* XD-Violence, il modello proposto si colloca al terzo posto e diventa secondo considerando solo la *feature* I3D. Risulta uno scarto di -4,83 rispetto al modello migliore, con *feature* I3D, che ha un AP di 85,59.

Inoltre si è dimostrata l'efficacia delle modifiche apportate al modello su I3D, in particolare riguardo l'utilizzo di *clip* e del *layer* SAP. L'introduzione di SAP mostra un vantaggio rispetto al più semplice *layer* GAP(*Global Average Pooling*), con un miglioramento di +0,1% su Shanghai-Tech, +0,32% su UCF-Crime e +1,63% su XD-Violence. Inoltre, il modello proposto mostra miglioramenti considerando le *clip* rispetto ai frammenti, con un aumento del +0,2% su Shanghai-Tech, +0,09% su UCF-Crime e +2,41% su XD-Violence. Infine, si è visto che il modello proposto è in grado di raggiungere con I3D tempi di inferenza *real-time*, processando 10,42 *clip* al secondo e 166,70 FPS su una GPU.

5 Conclusione

Nonostante i risultati ottenuti, il metodo proposto presenta delle limitazioni e presenta spazio per diversi possibili sviluppi futuri. Il campionamento delle *clip* è una strategia rischiosa, poiché nel caso di video lunghi si rischia di non catturare *clip* anomale, e quindi confondere il modello durante l'addestramento. Inoltre, campionare in modo uniforme genera dei salti temporali tra le *clip*. Un possibile miglioramento di questa strategia potrebbe essere effettuare un campionamento di insiemi di *clip* consecutive, il che potrebbe contribuire a mitigare dipendenze temporali brevi e lunghe. Sebbene il modello sia in grado di elaborare *clip* in tempo reale, non è pensato per il rilevamento online di anomalie in uno *streaming* video, poiché il metodo non sfrutta completamente il modulo MTN e il layer SAP, i quali sono progettati per catturare il contesto temporale tra tutte le *clip*, incluse quelle future.

Un'ulteriore possibilità è quella di effettuare un *task* ausiliario più complesso che consiste nella classificazione multi-classe delle anomalie di diverso tipo. In questo modo il modello potrebbe apprendere ulteriori informazioni specifiche per le singole anomalie migliorando le performance del *task* principale.

Un'altra limitazione riguarda le *feature* pre-addestrate per il *task* di riconoscimento di attività, poiché non sono specifiche per i dati trattati, e non permettono di catturare in modo preciso il contesto. In alternativa, possono essere considerate altri tipi di *feature* per rimpiazzare C3D

e I3D, oppure per essere combinate ad esse. Per esempio, si potrebbero aggiungere *feature* per l'audio se disponibile, oppure le *feature* più recenti che sono basate su *Transformer*. In particolare CLIP, che nonostante sia pre addestrata, è capace di generalizzare in diversi contesti. Oppure estrarre le *feature* considerando delle *tublet* (patch di *clip*) rispetto alle *clip*, permetterebbe di catturare il contesto in modo più fine [3]. Infine, si potrebbe pensare di addestrare una nuova *feature* basata su *Transformer* effettuando *fine-tuning* sui dataset di *benchmark*, effettuando il *task* di riconoscimento di attività considerando le anomalie come classi. Ottenendo così delle *feature* spazio-temporali specifiche per il *task* trattato catturando in modo migliore il contesto, per poi essere usate come *feature* pre-addestrate.

In aggiunta alle limitazioni del metodo attualmente proposto, allo stato dell'arte rimangono molte sfide aperte da affrontare per il *task* di rilevamento di anomalie nei video. I dataset di *benchmark* di solito consistono di video a bassa risoluzione, il che rende difficile identificare dettagli e quindi il riconoscimento di anomalie. Tuttavia considerare risoluzioni maggiori rende complicato il processo di estrazione delle *feature* e l'addestramento. Ottenere dati etichettati a *frame-level* è molto oneroso, il che rimane un problema che i metodi debolmente supervisionati non hanno risolto pienamente. Infatti, si possono presentare errori di selezione delle istanze positive durante l'addestramento, che spesso comporta strategie di addestramento complesse (e.g. strategie ibride, *multi-task* e *self-training*) e rende difficile formulare un opportuno schema di valutazione. Altre sfide aperte sono il rilevamento spaziale delle anomalie e il rilevamento delle anomalie online, il che non si limita a modelli in grado di gestire *streaming* video *real-time*, ma anche capaci di aggiornare online il concetto di normalità vista la natura *open-set* delle anomalie.

Riferimenti bibliografici

[1] Hyekang Kevin Joo et al. "Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection". In: *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2023, pp. 3230–3234.

[2] Seongheon Park et al. "Normality Guided Multiple Instance Learning for Weakly Supervised Video Anomaly Detection". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 2665–2674.

[3] Shengyang Sun e Xiaojin Gong. "Long-Short Temporal Co-Teaching for Weakly Supervised Video Anomaly Detection". In: *arXiv preprint arXiv:2303.18044* (2023).

[4] J Feng, F Hong e W Zheng. "Multiple instance self-training framework for video anomaly detection". In: *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. 2021.

[5] Yu Tian et al. "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 4975–4986.

[6] Boyang Wan et al. "Weakly supervised video anomaly detection via center-guided discriminative learning". In: *2020 IEEE international conference on multimedia and expo (ICME)*. IEEE. 2020, pp. 1–6.

[7] Dong Gong et al. "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1705–1714.

[8] Waqas Sultani, Chen Chen e Mubarak Shah. "Real-world anomaly detection in surveillance videos". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6479–6488.