自己注意機構を用いた長期的分類による現実世界での異常検出

渡邉　祐大[†]　岡部　誠[†]　　原田　泰典[‡]　鹿島　直二[‡]
静岡大学[†]　　中部電力株式会社[‡]

1　はじめに

街中に設置されている防犯カメラの台数は世界中で増加を続けている．また工場や発電所などの大規模施設でも複数台の監視カメラを用いて安全確認が行われている．しかし，これらの映像全てを人間が目で見て確認することは難しいため，人間に代わって人工知能が映像を解析し，自動的に異常事象を検出できる技術の開発が急務である．異常事象はめったに観測されないため，正常状態のみを学習データとして用いる手法が多く存在する．推論時には入力動画が学習済みの正常状態からどれくらい乖離しているかという基準で正常か異常か判断する．しかし，こういった手法は動画の見た目や速度の違いなどの低レベルな特徴量に基づいた異常検出しかできない．そこで近年，正常な動画と異常な動画の両方を含む弱い教師データセットを用いて異常検出器を学習する手法が提案されている [1, 3, 4]．

弱い教師データセットでは動画単位で正常/異常のラベルが付いている．即ち，正常とラベル付けされた動画では全フレームを通して正常状態のみが記録されている．一方で，異常とラベル付けされた動画には正常状態と異常状態のフレームが混在している．このようなデータセットを用いることで，動画中のフレーム毎に正常/異常のラベル付けをする必要がなくなり，ラベル付けの労力を削減できる．

既存手法では連続する複数のフレームを 1 つの短期的なセグメントとして扱い，多重インスタンス学習 (MIL) を用いてセグメント毎の正常/異常を判定している．MIL では各動画を 1 つのバッグとして扱い，バッグの中には対応する動画から得られた全セグメントが入っている．異常度を表すスコアが最も高いセグメントをバッグから取り出し，そのセグメントが正常な動画から来たものであれば 0 を出力するように，あるいは，そのセグメントが異常な動画から来たものであれば 1 を出力するように異常検出器を学習する [1]．

しかし，近年成功している既存手法を観察すると，単一のセグメントだけを用いるのではなく，動画の長期的な情報を効率良く扱うことで高い精度を達成していることが分かった．例えば，Tian ら [4] らは，時間方向の畳み込み演算を組み合わせることでセグメントの長期的な時間変化を考慮し，その結果として高精度な異常検出器を実現した．Tian ら [4] の手法に触発され，我々は動画から得られる全セグメントを入力とし，その動画が正常か異常かを判定するモデルを提案する．学習段階において，提案手法はセグメント単位でなく動画単位の学習を行うため MIL が不要となる．その結果，学習手法の実装が簡単となる．

2　提案手法

我々は動画の長期的な情報を効率良く扱う異常検出器 *Video Classifier* を提案する．

2.1　特徴量抽出

データセットを $\mathcal{D} = \{(\mathbf{V}_i, y_i)\}$ とする．ただし，$\mathbf{V}_i$ は学習データセット中の i 番目の動画，y_i は $\mathbf{V}_i$ に付けられたラベルである．$y_i = \{0, 1\}$ で 0 ならば正常，1 ならば異常を表す．正常とラベル付けされた動画では全フレームを通して正常状態のみが記録されている．一方で，異常とラベル付けされた動画には正常状態と異常状態のフレームが混在している．$\mathbf{V}_i$ は T 個のセグメントに分割され，各セグメントは特徴抽出器によって D 次元の特徴ベクトル $\mathbf{F}_{i,j}$ に変換されている．ただし，$\mathbf{F}_{i,j}$ は $\mathbf{V}_i$ の j 番目の特徴ベクトルを表す．特徴抽出器は全ての実験を通して，Kineticts データセットによって学習済みの I3D[2] を用いた．

2.2　Video Classifier

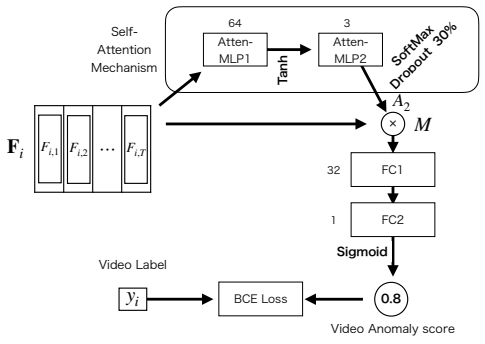

図 1　Video Classifier の概要図

Real-world Anomaly Detection by Long-Term Classification using Self-attention Mechanism

† Yudai Watanabe, Makoto Okabe, Graduate School of Engineering, Shizuoka University
‡ Yasunori Harada, Naoji Kashima, Chubu Electric Power Co., Inc.

モデルの概要図を図 1 に示す．*Video Classifier* は自己注意機構（Self-Attention Mechanism）と 2 層の全結合層からなるシンプルなモデルである．自己注意機構では，入力された $\mathbf{F}_i$ は Atten-MLP1 の多層パーセプトロンによって $64 \times T$ の行列 $\mathbf{A}_1$ に変換される．この時，活性化には Tanh 関数を用いる．$\mathbf{A}_1$ は Atten-MLP2 の多層パーセプトロンによって $3 \times T$ の行列 $\mathbf{A}_2$ に変換される．この時，活性化には softmax 関数を用いる．更に 30% のドロップアウト正則化も行う．

自己注意機構から得られた重み行列 $\mathbf{A}_2$ を用いて $\mathbf{M} = \mathbf{F}_i \mathbf{A}_2^t$ を計算する．$\mathbf{M}$ の形を $D \times 3$ 次元のベクトルに変形した後，2 層の全結合層 FC1(32 ユニット) と FC2(1 ユニット）を経て動画の異常性を表すスコアに変換される．FC1 の活性化関数は恒等関数，FC2 の活性化関数は sigmoid 関数である．損失関数にはバイナリクロスエントロピー (BCE) 関数を用いた．

2.3 Video Classifier による推論

Video Classifier を適用して異常検出したい動画を $\mathbf{V}^e$ とする．まず，連続する 16 フレームを 1 つのセグメントとし，$\mathbf{V}^e$ を N 個のセグメント $\{\mathbf{V}_1^e, \mathbf{V}_2^e, \cdots, \mathbf{V}_N^e\}$ に分割する．各セグメント $\mathbf{V}_i^e$ は I3D[2] によって D 次元の特徴ベクトル $\mathbf{F}_i^e$ に変換される．特徴ベクトルの集合を $\mathbf{F}^e = \{\mathbf{F}_1^e, \mathbf{F}_2^e, \cdots, \mathbf{F}_N^e\}$ とする．次に分割サイズを l とし，$\mathbf{F}^e$ を l 個のセグメント毎に $m = N/l$ 個のバッグに分割する．即ち，

$$\mathbf{F}^e = \{\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_m\}$$
$$= \{\{\mathbf{F}_1^e, \cdots, \mathbf{F}_l^e\}, \{\mathbf{F}_{l+1}^e, \cdots, \mathbf{F}_{2l}^e\} \cdots \{\mathbf{F}_{N-l}^e, \cdots, \mathbf{F}_N^e\}\}$$

となる．次に，$\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_m$ を 1 つずつ *Video Classifier* に入力とすることで推論結果 $\mathbf{S}^v = \{s_1^v, s_2^v, \cdots, s_m^v\}$ を得る．s_i^v は $\{\mathbf{V}_{(i-1)l+1}^e, \cdots, \mathbf{V}_{il}^e\}$ の各セグメントに対する異常スコアとして用いられる．フレームレベルの異常検出の結果を作る必要がある際は，得られたスコアを各セグメントの全フレームに割り当てることとする．

3 実験

提案手法の評価を行うために，UCF-Crime データセット [1] を用いて，実験を行い，いくつかの既存手法との検出精度比較を行った．評価指標は，既存研究 [1, 3, 4] と同様に，フレーム単位の異常検出精度に基づいて計算された Receiver Operating Characteristic (ROC) 曲線における Area Under the Curve (AUC) を評価指標として使用した．AUC の値が大きいほど，高精度な異常検出器であることを表している．また，最適化アルゴリズムには Radam を用い，学習率は 0.001 とした．バッチサイズは 64 とした．ハイパーパラメータについては，T を 64，推論時の分割サイズ l を 27 とした．

3.1 UCF-Crime データセットの実験結果

表 1 に UCF-Crime データセットにおけるフレーム単位での AUC 性能と各手法のモデルの学習可能なパラメータ数を示す．提案手法である *Video Classifier* モデルは MIL に基づく既存手法と同等か，もしくはそれよりも高い検出精度を達成できている．また，*Video Classifier* モデルは既存手法の中で最もパラメータ数の少ない Wu ら [3] の手法と比較して，半分以下のパラメータ数で高い検出精度が達成できている．シンプルで軽量な手法にも関わらず，*Video Classifier* は MIL に基づく手法と遜色ない検出精度が達成可能であることが分かる．

手法	特徴量	パラメータ数	AUC(%)
SVM Baseline	-	-	50.00
Sultani et al. [1]	C3D RGB	2114113	75.41
Wu et al. [3]	I3D RGB	769155	82.44
RTFM [4]	I3D RGB	24718849	84.30
Video Classifier	I3D RGB	328004	84.30

表 1　UCF-Crime データセットにおけるフレーム単位の AUC 性能比較.

4 結論

我々は，自己注意機構を用いて動画の長期的な特徴を解析して学習する *Video Classifier* モデルを提案した．*Video Classifier* は MIL に基づく手法ではないため，実装が簡単であり，既存手法と比べて学習可能なパラメータが最も少ないシンプルなモデルにも関わらず，最先端の手法と比較して遜色ない検出精度が達成できることを示した．

参考文献

[1] Waqas Sultani, Chen Chen, Mubarak Shah, "Real-world anomaly detection in surveillance videos", *CVPR*, 2018

[2] Joao Carreira, Andrew Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", *CVPR*, 2017

[3] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, Zhiwei Yang,"Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision", *ECCV*, 2020

[4] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, Gustavo Carneiro,"Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning", *ICCV*, 2021