Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning - Supplementary Material

Yu Tian[1,3] Guansong Pang[1] Yuanhong Chen[1] Rajvinder Singh[3]
Johan W. Verjans[1,2,3] Gustavo Carneiro[1]
[1] Australian Institute for Machine Learning, University of Adelaide
[2] Faculty of Health and Medical Sciences, University of Adelaide
[3] South Australian Health and Medical Research Institute

1. Theoretical Motivation of RTFM

Theorem 1.1 (Expected Separability Between Abnormal and Normal Videos)**.** *Assuming that* $\mathbb{E}[\|\mathbf{x}^+\|_2] \geq \mathbb{E}[\|\mathbf{x}^-\|_2]$*, where* $\mathbf{X}^+$ *has* μ *abnormal samples and* $(T-\mu)$ *normal samples, where* $\mu \in [1, T]$*, and* $\mathbf{X}^-$ *has* T *normal samples. Let* $D_{\theta,k}(.)$ *be the random variable from which the separability scores* $d_{\theta,k}(.)$ *of Eq.3 in the main paper are drawn [2].*

1. *If* $0 < k < \mu$*, then*

$$0 \leq \mathbb{E}[D_{\theta,k}(\mathbf{X}^+, \mathbf{X}^-)] \leq \mathbb{E}[D_{\theta,k+1}(\mathbf{X}^+, \mathbf{X}^-)].$$

2. *For a finite* μ*, then*

$$\lim_{k\to\infty} \mathbb{E}[D_{\theta,k}(\mathbf{X}^+, \mathbf{X}^-)] = 0.$$

Proof.

$$\begin{aligned}\mathbb{E}[D_{\theta,k}(\mathbf{X}^+, \mathbf{X}^-)] &= \mathbb{E}[g_{\theta,k}(\mathbf{X}^+)] - \mathbb{E}[g_{\theta,k}(\mathbf{X}^-)] \\ &= p_k^+(\mathbf{X}^+)\mathbb{E}[\|\mathbf{x}^+\|_2] + p_k^-(\mathbf{X}^+)\mathbb{E}[\|\mathbf{x}^-\|_2] - \mathbb{E}[\|\mathbf{x}^-\|_2]\end{aligned} \tag{1}$$

1. Trivial given that $\mathbb{E}[\|\mathbf{x}^+\|_2] \geq \mathbb{E}[\|\mathbf{x}^-\|_2]$ and that $p_{k+1}^+(\mathbf{X}^+) > p_k^+(\mathbf{X}^+)$ for $0 < k < \mu$
2. Trivial given that as μ is finite, $\lim_{k\to\infty} p_k^+(\mathbf{X}^+) = 0$.

□

Intuition of feature magnitude: Assuming the expected magnitude of abnormal samples is larger than of normal samples, we can derive Thm. 3.1 that proves that the expected feature magnitude-based separability score between normal and abnormal videos grows for $0 < k < \mu$ and reduces to zero for $k \to \infty$. Hence, to use Thm. 3.1, we need to enforce larger magnitude for abnormal features using our proposed RTFM. The similarity between the theoretical and empirical curves in Fig.2(left) is evidence of the soundness of Thm. 3.1.

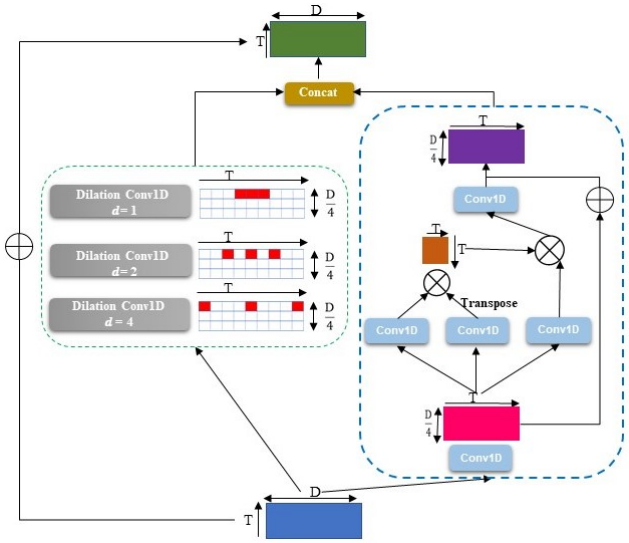

Figure 1. Our proposed MTN consists of two modules. The module on the left uses the pyramid dilated convolutions to capture the local consecutive snippets dependency over different temporal scales. The module on the right relies on a self-attention network to compute the global temporal correlations. The features from the two modules are concatenated to produce the MTN output.

2. Multi-scale Temporal Feature Learning

Our proposed multi-scale temporal network (MTN) captures the multi-resolution local temporal dependencies and the global temporal dependencies between video snippets, as displayed in Fig. 1.

3. Computational Efficiency

We investigate if our system can run in real time. During inference, our method processes a 16-frame clip in 0.76 seconds on a Nvidia 2080Ti–this time includes the I3D extraction time. This indicates that our system can achieve

good real-time detection in real-world applications.

4. Temporal Dependency

Temporal Dependency has been explored in [1, 3–5, 7, 8, 10]. In anomaly detection, traditional methods [1, 8] convert consecutive frames into handcrafted motion trajectories to capture the local consistency between neighbouring frames. Diverse temporal dependency modelling methods have been used in deep anomaly detection approaches, such as stacked RNN [5], temporal consistency in future frame prediction [4], and convolution LSTM [3]. However, these methods capture short-range fixed-order temporal correlations only with single temporal scale, ignoring the long-range dependency from all possible temporal locations and the events with varying temporal length. GCN-based methods are explored in [7, 10] to capture the long-range dependency from snippets features, but they are inefficient and hard to train. By contrast, our proposed module combines PDC [9] and TSA [6] on the temporal dimension to seamlessly and efficiently incorporate both the long and short-range temporal dependencies into our temporal feature ranking loss.

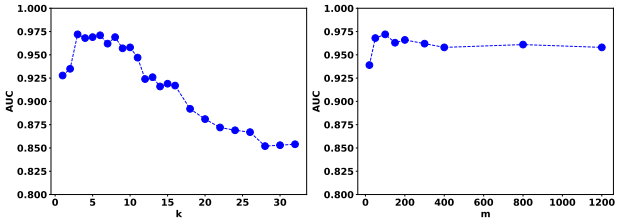

Figure 2. AUC w.r.t. top-k (Left) and the margin m (Right).

5. Ablations for k and m

We show the AUC results as a function of top-k and margin m values on ShanghaiTech in Fig.2. Consistent to our theoretical analysis, the performance of our model peaks at a sufficiently large k, flattens at around $k \approx \mu$ and then drops with increasing k (Fig.2(left)). It is also robust to a large range of $m \in [50, 1200]$ with a stable AUC in $[93\%, 96\%]$ (Fig.2(right)).

References

[1] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1446–1453. IEEE, 2009. 2

[2] Weixin Li and Nuno Vasconcelos. Multiple instance learning for soft bags via top instances. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4277–4285, 2015. 1

[3] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, Shenghua Gao, et al. Margin learning embedded prediction for video anomaly detection with a few anomalies. 2

[4] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018. 2

[5] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017. 2

[6] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[7] Peng Wu, jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[8] Dan Xu, Rui Song, Xinyu Wu, Nannan Li, Wei Feng, and Huihuan Qian. Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. *Neurocomputing*, 143:144–152, 2014. 2

[9] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[10] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019. 2