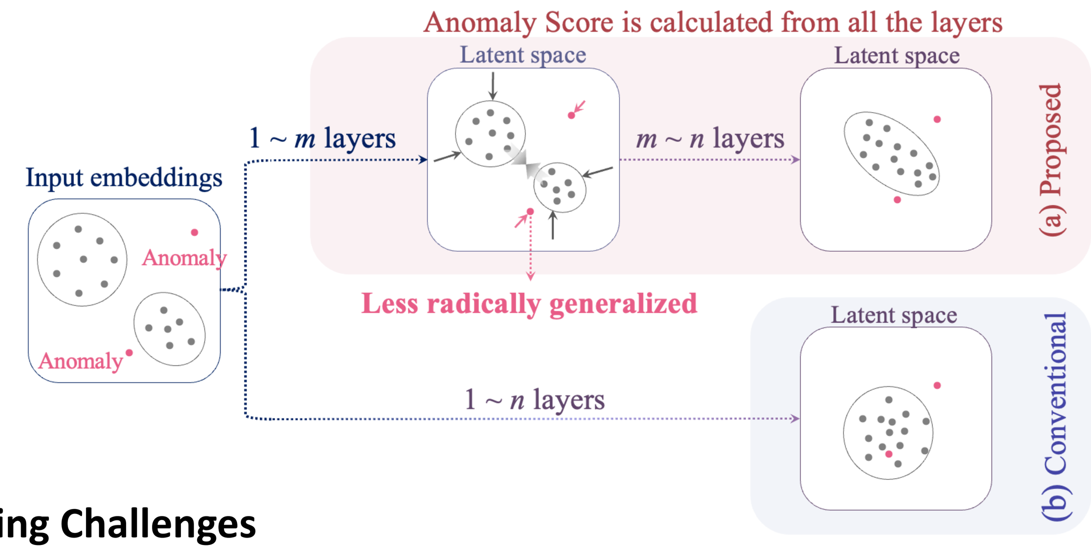
VATMAN Video Anomaly Transformer for Monitoring Accidents and Nefariousness

Harim Kim[1], Chang Ha Lee[2], Charmgil Hong[1]
[1]Handong Global University, [2]GMDSOFT

Introduction

- **Video Anomaly Detection**
  - Anomaly detection is identifying atypical patterns that diverge from the majority
  - Recent advances in deep learning have sparked increasing interest in utilizing video data for anomaly detection
  - Video anomaly detection methods [1,2] can identify abnormal behaviors in video footage, and reduce the workload of human operators
- **Main Challenges**
  - **Challenge 1: Extracting Spatio-temporal Representations from Video Footage**
    - The identification of specific situations often hinges on a complex mix of spatial and temporal data
  - **Challenge 2: Navigating the Complexity of Data in Anomaly Detection Models**
    - Video recordings capture a broad spectrum of real-world dynamics and result in complex and often non-uniform data distributions
    - Traditional anomaly detection methods, which typically generalize data into a singular distribution, are prone to high false positive and false negative rates



- **Overcoming Challenges**
  - We leverage a pre-trained 3D-CNN (Convolutional Neural Network) [3] to extract spatio-temporal representations
  - We propose a novel Transformer [4]-based Autoencoder (AE) to deal with complex data distributions
    - The encoder in our framework progressively generalizes the training data employing self-attention mechanism
    - The level of generalization for individual data points is indirectly assessed through attention weight at each self-attention layer
- **Remarks**
  - Discerning the context of human behavior from a single frame is challenging, so we segment continuous video data into uniform intervals along the temporal axis
  - We define a single data point as a segment and an anomaly data point as a segment that contains at least one anomaly frame

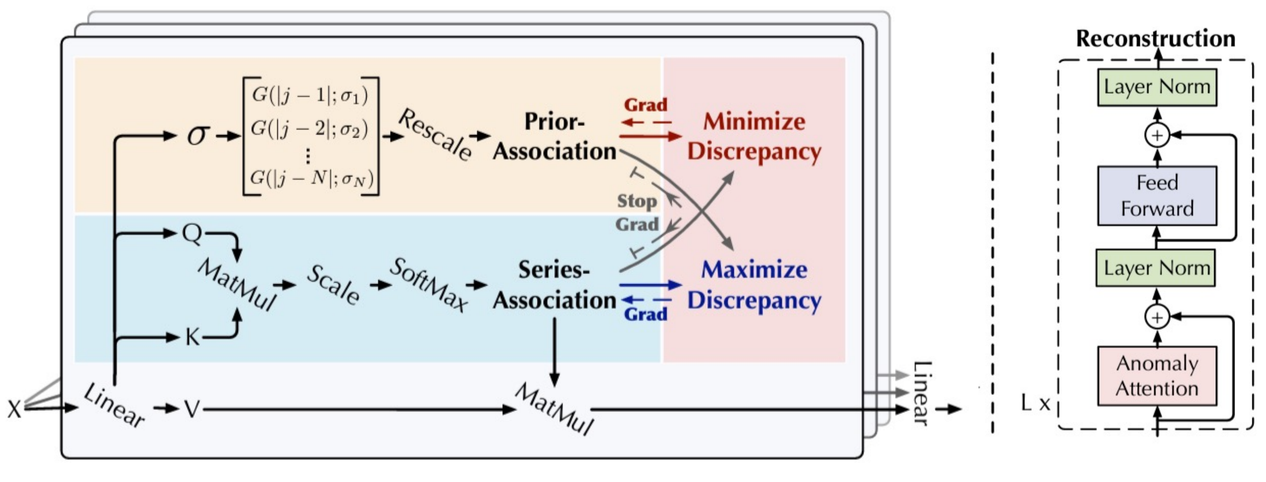
Related Work

- **Self-attention Mechanism**
  - By employing parallelizable self-attention mechanism, Transformer [4] can consider global dependencies within the data and reflect the complexity and context of input
  - Multi-head attention performs self-attention using multiple sets of weights to facilitate attention operations from various perspectives

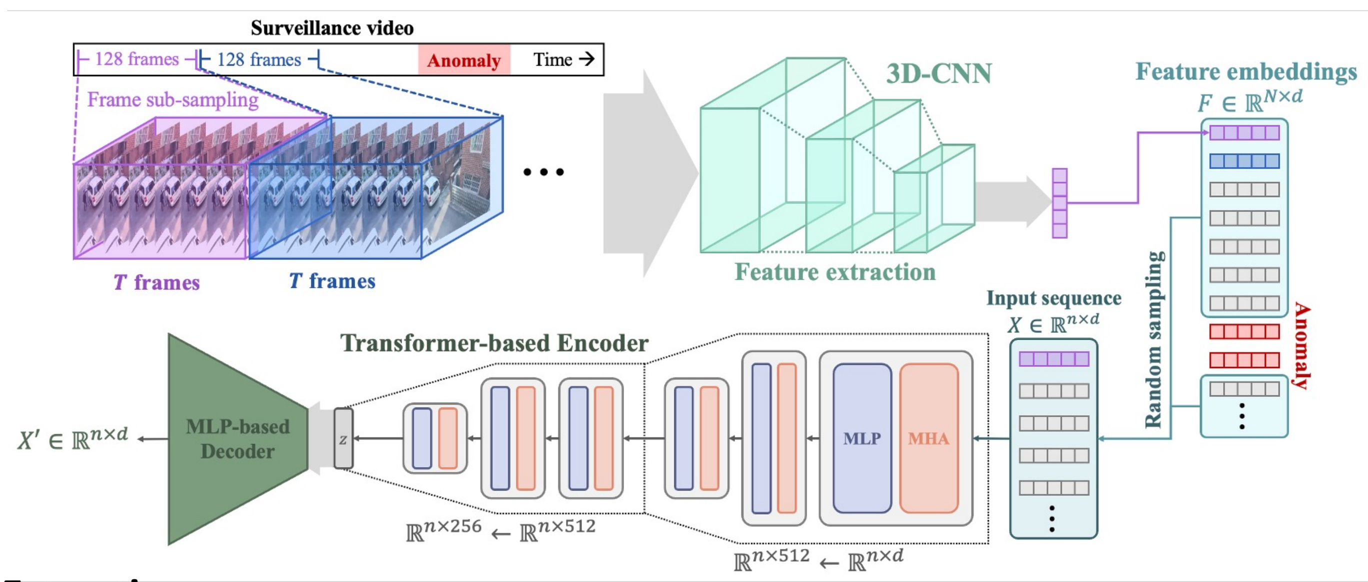
$$MHA(X) = Concat(head_1, \dots, head_h)W^O$$
$$head_i = Self\text{-}Attention_i(X)$$
$$Self\text{-}Attention_i(X) = softmax(QK^T/\sqrt{d_k})V$$
$$Q = XW_i^Q,\ K = XW_i^K,\ V = XW_i^V$$

- **Anomaly Transformer**
  - Anomaly Transformer [5] is anomaly detection approach in time-series data
  - This approach tends to make input data points appear more similar by applying the self-attention (weighted sum) operation across multiple windows
  - This may result in overly simplified generalization

$$\mathrm{AssDis}(\mathcal{P}, \mathcal{S}; \mathcal{X}) = \left[\frac{1}{L}\sum_{l=1}^{L}\left(\mathrm{KL}(\mathcal{P}^l_{i,:}\|\mathcal{S}^l_{i,:}) + \mathrm{KL}(\mathcal{S}^l_{i,:}\|\mathcal{P}^l_{i,:})\right)\right]_{i=1,\cdots,N}$$



Proposed Method

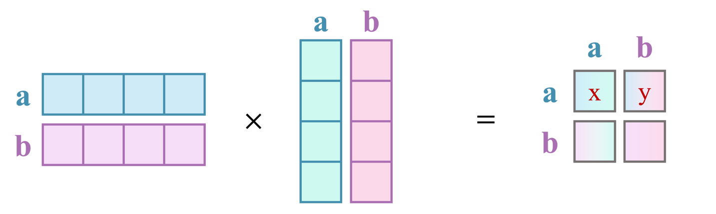


- **Feature Extraction**
  - To effectively capture spatio-temporal information from the segments, we employ a 3D-CNN pre-trained on large-scale action recognition datasets, such as Kinetics-400 [6] and Charades [7]
- **Transformer-based Autoencoder**
  - Key Features Distinguished from Conventional Transformer
    1) The MLP (Multi-layer Perceptron) layers within our encoder are designed to gradually reduce the dimensionality of the output space at the 3rd and 6th layers
    2) We employ random sampling from all input segments to generate input sequences
    3) The decoder is designed as an MLP structure to simply match the output dimensions
  - Loss Function

$$Reconstruction\ loss = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{x}'_i\|^2$$

- **Detecting Anomalies**
  - Self-attention is designed to allocate higher weights to input segments that exhibit greater similarities within the sequence
  - Due to their uniqueness and the scarcity of similar samples, anomalies tend to draw high attention among themselves
  - By exploiting this, we propose a novel method for determining the anomaly score, which involves aggregating the attention weights across all encoder layers

$$X = \{\mathbf{x}_0, \mathbf{x}_i, \dots, \mathbf{x}_{n-1}\}$$
$$Score = \frac{1}{S}\cdot\frac{1}{L}\cdot\frac{1}{H}\sum_{s=1}^{S}\sum_{l=1}^{L}\sum_{h=1}^{H} Attention(\mathbf{x}_0)^{(s,l,h)}$$



Experiments and Results

- **Dataset**
  - *Abnormal Behavior CCTV Video Dataset* [8] provided by the South Korean National Information Society Agency (NIA)
  - Three abnormal behavior types: trespassing, fighting and vandalism

Video specification of each behavior type

time of day	length (s)	frame rate (fps)
daytime	~ 3600	30

- **Data Preprocessing**
  - We segment the video data into intervals of 128 frames, with the number of data points (N) for each type of behavior
  - The input frames are resized and sampled at regular intervals to match the input shape required by each pre-trained 3D-CNN model used in the experiments

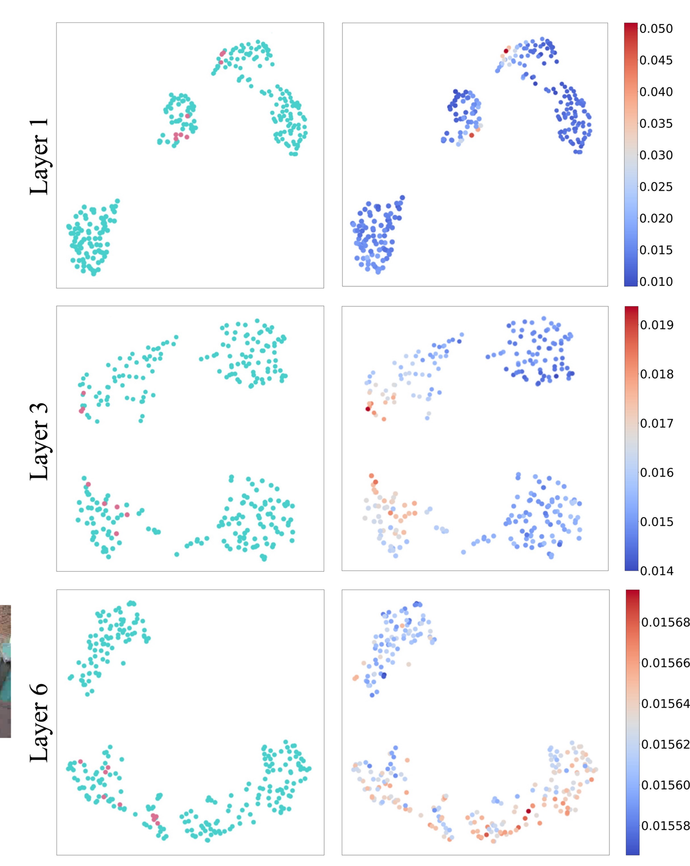
Dataset	Trespassing	Vandalism	Fighting
N	1,419	1,432	1,389
anomaly ratio	0.016	0.049	0.040
AE	0.907 / 0.343	0.839 / 0.224	0.895 / 0.330
VAE	0.907 / 0.300	0.834 / 0.219	0.892 / 0.321
AE-SVDD [17]	0.864 / 0.263	0.892 / 0.226	0.911 / 0.262
VAE-SVDD [22]	0.845 / **0.388**	0.802 / 0.197	0.731 / 0.119
Anomaly Transformer [21]	0.737 / 0.326	0.724 / 0.144	0.790 / 0.162
VATMAN (ours)	**0.913** / 0.278	**0.899 / 0.274**	**0.944 / 0.455**

Table 1. AUROC/AUPRC performances of compared methods (N: total number of data points)

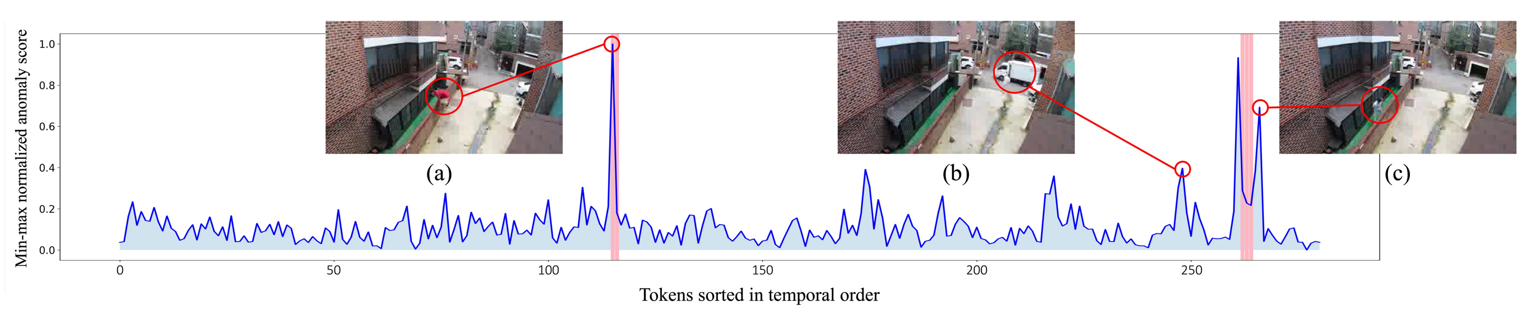
	Input shape	Dataset		
	(T×height×width)	trespassing	vandalism	fighting
P3D [16]	(16×160×160)	0.431 / 0.015	0.762 / 0.128	0.502 / 0.076
I3D [5]	(64×224×224)	0.842 / 0.324	0.878 / 0.305	0.826 / 0.245
S3D [20]	(64×224×224)	0.772 / 0.294	**0.919 / 0.357**	0.594 / 0.078
X3D [7]	(16×224×224)	**0.880 / 0.367**	0.897 / 0.236	**0.915 / 0.383**
TimeSformer [3]	(16×224×224)	0.690 / 0.172	0.755 / 0.173	0.684 / 0.142

Table 2. AUROC/AUPRC performances with difference features

- **Performance Comparison**
  - Our method outperforms the existing solutions
  - This indicates that our method not only precisely identifies the anomalies but also minimize false alarms and false negatives across different datasets
- **Visualization of Attention Weights**
  - **1-3rd layers**: suggests that our approach can extract meaningful anomaly scores at each stage of the generalization process
  - **6th layer**: while the data begin to show less varied distributions compared to earlier layers, higher attention weights continue to be assigned to anomalies and their vicinity



- **Case Analysis**
  - **(a)**: the precise moment when a person dressed in red jumps over the fence was successfully pinpointed
  - **(b)**: the detection was triggered by a white truck passing by, likely due to the scarcity of such events in the training data
  - **(c)**: despite being labeled as normal in the dataset, it can be considered as a potential anomaly



HANDONG ARTIFICIAL INTELLIGENCE LAB
GMDSOFT

Contact Info.
Presenter: Harim Kim (hrkim@handong.ac.kr)
Collaborators: Chang Ha Lee (yielding@gmdsoft.com)
Advisor: Charmgil Hong (charmgil@handong.ac.kr)

[1] Boekhoudt, K., & Talavera, E. (2022, November). Spatial-temporal transformer for crime recognition in surveillance videos. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-8). IEEE.
[2] Naji, Y., Setkov, A., Loesch, A., Gouiffès, M., & Audigier, R. (2022, November). Spatio-temporal predictive tasks for abnormal event detection in videos. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-8). IEEE.
[3] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
[5] Xu, J., Wu, H., Wang, J., & Long, M. (2021). Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*.
[6] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Zisserman, A. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.
[7] Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 510-526). Springer International Publishing.
[8] Aihub. https://aihub.or.kr/, 2019. Accessed: July 15, 2023.