A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction

Zhian Liu¹  Yongwei Nie¹*  Chengjiang Long²  Qing Zhang³  Guiqing Li¹

¹South China University of Technology, ²JD Finance America Corporation, ³Sun Yat-sen University
Video Anomaly Detection

• Motivation
  • Surveillance cameras are widely used.
  • VAD is an essential task to save human labor.
Video Anomaly Detection

• Goal: to identify unexpected behaviours in a video.

Ped2[1] test video #04

Avenue[2] test video #04

[1] http://www.svcl.ucsd.edu/projects/anomaly/dataset.html
[2] http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html
Video Anomaly Detection

• Goal: to identify unexpected behaviours in a video.

• Useful but challenging task.
Video Anomaly Detection

• Challenges
  • Anomaly rarely happens.
  • What is anomaly?

• Solution
Related work

• Reconstruction-based method
  • Train AE with L1 or L2 loss.

• Assume the anomalies lead to larger reconstruction errors.
Related work

- Reconstruction-based method
  - Memory-augmented AE to mitigate the "over-generalization" problem.

\[
\hat{z} = w M = \sum_{i=1}^{N} w_i m_i \quad \quad w_i = \frac{\exp(d(z,m_i))}{\sum_{j=1}^{N} \exp(d(z,m_j))}
\]
Related work

- Prediction-based method
  - Take the temporal information into consideration [Liu. et al, 2018].

\[
\mathcal{L} = \| \hat{I}_{t+1} - I_{t+1} \|^2_2
\]

[Future Frame Pred.] W. Liu et.al, CVPR, 2018
Our approach

• Insight
  • Previous work rarely exploits the **consistency between flows and frames**.
  • For an abnormal event, what if we manipulate the flows beforehand, and try to produce a poor prediction?
  • Propose to **reconstruct the flows first**, then using the reconstructed flows as condition to predict future frame.
A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction

HF$^2$-VAD pipeline

Previous $t$ frames

Future frame

Corresponding $t$ optical flows

ML-MemAE-SC

$y_{1:t}$

$y_{1:t}$

$\hat{y}_{1:t}$

$F_{\phi}$

$D_{\psi}$

$\hat{x}_{t+1}$

$E_{\theta}$

KL Loss

$p(z|\hat{y}_{1:t})$

$q(z|x_{1:t}, \hat{y}_{1:t})$

Features Concat.

Sampling

Memory Module

Weighted Sum

$M$

$C$

$\sum$
ML-MemAE-SC

- Observations
  - Memory only in bottleneck cannot remember all normal patterns.
  - AE with multi-level memories (ML-MemAE) leads to degradation.
  - Skip connection helps.

(a) MemAE  (b) ML-MemAE  (c) ML-MemAE-SC
ML-MemAE-SC

- Flow reconstruction objective

\[ \mathcal{L}_{\text{recon}} = \| y_{1:t} - \hat{y}_{1:t} \|_2^2 \]

\[ \mathcal{L}_{\text{ent}} = \sum_{i=1}^{M} \sum_{k=1}^{N} -\hat{w}_{i,k} \log(\hat{w}_{i,k}) \]

\[ \mathcal{L}_{\text{ML-MemAE-SC}} = \lambda_{\text{recon}} \mathcal{L}_{\text{recon}} + \lambda_{\text{ent}} \mathcal{L}_{\text{ent}} \]
CVE for prediction

• Formulation

  • Let $x_{1:t}$ and $x_{t+1}$ be the previous and future frame, $y_{1:t}$ be the reconstructed flows, $z$ be the hidden variables that control the content information:

    \[
    \log p(x_{t+1} \mid y_{1:t}) \geq \mathbb{E}_q \log \frac{p(x_{t+1} \mid z, y_{1:t})p(z \mid y_{1:t})}{q(z \mid x_{t+1}, y_{1:t})} \quad \text{(Evidence Lower Bound)}
    \]

    \[
    \approx \mathbb{E}_q \log \frac{p(x_{t+1} \mid z, y_{1:t})p(z \mid y_{1:t})}{q(z \mid x_{1:t}, y_{1:t})} \quad \text{(Short Duration Assumption)}
    \]

    \[
    = -KL[q(z \mid x_{1:t}, y_{1:t})\|p(z \mid y_{1:t})] + \mathbb{E}_q[\log p(x_{t+1} \mid z, y_{1:t})]
    \]

  • Resort the conditional Variational Autoencoder (CVAE).
CVE for prediction

- Frame prediction objective

\[ \mathcal{L}_{CVAE} = KL[q(z \mid x_{1:t}, y_{1:t}) \mid p(z \mid y_{1:t})] + \| x_{t+1} - \hat{x}_{t+1} \|_2^2 \]

\[ \mathcal{L}_{gd}(X, \hat{X}) = \sum_{i,j} \left| X_{i,j} - X_{i-1,j} - |\hat{X}_{i,j} - \hat{X}_{i-1,j}| \right| \]

\[ |X_{i,j} - X_{i,j-1} - |\hat{X}_{i,j} - \hat{X}_{i,j-1}| | \]

\[ \mathcal{L} = \lambda_{CVAE} \mathcal{L}_{CVAE} + \lambda_{gd} \mathcal{L}_{gd}(\hat{x}_{t+1}, x_{t+1}) \]
Anomaly detecting

• At test time, the anomaly score is composed of two parts:
  • Reconstruction error $S_r = \| \hat{y}_{1:t} - y_{1:t} \|_2^2$
  • Prediction error $S_p = \| \hat{x}_{t+1} - x_{t+1} \|_2^2$

• Frame-level anomaly score

$$S_{O_i} = w_r \cdot S_r + w_p \cdot S_p \quad S = max\{S_{O_1}, S_{O_2}, \ldots S_{O_N}\}$$
Anomaly detecting

• At test time, the anomaly score is composed of two parts:
  • Reconstruction error $S_r = \| \hat{y}_{1:t} - y_{1:t} \|^2_2$
  • Prediction error $S_p = \| \hat{x}_{t+1} - x_{t+1} \|^2_2$

• Frame-level anomaly score

$$S_{O_i} = w_r \cdot S_r + w_p \cdot S_p$$

$$S = \max\{S_{O_1}, S_{O_2}, \ldots S_{O_N}\}$$
Experimental results

• Datasets

a) UCSD Ped2

b) CUHK Avenue

c) ShanghaiTech

• Quantitative results

| Method               | UCSD Ped2 | CUHK Avenue | SHTech |
|----------------------|-----------|-------------|--------|
| Conv-AE [11]         | 90.0      | 70.2        | -      |
| ConvLSTM-AE [32]     | 88.1      | 77.0        | -      |
| GMFC-VAE [7]         | 92.2      | 83.4        | -      |
| MemAE [8]            | 94.1      | 83.3        | 71.2   |
| MNAD-R [39]          | 90.2      | 82.8        | 69.8   |
| Frame-Pred. [26]     | 95.4      | 85.1        | 72.8   |
| Conv-VRNN [31]       | 96.1      | 85.8        | -      |
| MNAD-P [39]          | 97.0      | 88.5        | 70.5   |
| VEC [50]             | 97.3      | 90.2        | 74.8   |
| ST-AE [53]           | 91.2      | 80.9        | -      |
| AMC [37]             | 96.2      | 86.9        | -      |
| AnoPCN [49]          | 96.8      | 86.2        | 73.6   |
| HF²-VAD w/o FP       | 98.8      | 86.8        | 73.1   |
| HF²-VAD w/o FR       | 94.5      | 90.2        | 76.0   |
| HF²-VAD              | **99.3**  | **91.1**    | **76.2** |
Experimental results

• Qualitative results

(a) Skateboarding and riding bicycle of Ped2.
(b) Kid running of Avenue.
Experimental results

• Visualization

|                | Normal                        | Abnormal                     |
|----------------|-------------------------------|------------------------------|
| GT             | ![Normal GT Image]            | ![Abnormal GT Image]        |
| Ours Pred.     | ![Normal Ours Pred. Image]   | ![Abnormal Ours Pred. Image]|
| Ours diff.     | ![Normal Ours diff. Image]   | ![Abnormal Ours diff. Image]|
| VEC            | ![Normal VEC Image]          | ![Abnormal VEC Image]       |
| MNAD-P         | ![Normal MNAD-P Image]       | ![Abnormal MNAD-P Image]    |

• Ablation study

Table 2. Ablation study results on UCSD Ped2 [35] dataset. The anomaly detection performance is reported in terms of AUROC ↑ (%). Number in bold indicates the best result.

| Memory-augmented Reconstruction Models | Prediction Models | AUROC |
|---------------------------------------|-------------------|-------|
|                                       | VAE               |       |
| Flow                                  | ✓                 | 96.27 |
|                                       | ✓                 | 97.75 |
|                                       | ✓                 | 98.81 |
| Frame                                 | ✓                 | 89.96 |
|                                       | ✓                 | 94.48 |
| Hybrid                                | ✓                 | 96.91 |
|                                       | ✓                 | 98.28 |
|                                       | ✓                 | 99.31 |

[VEC] G. Yu et al., ACM-MM, 2020
[MNAD-P] H Park et al., CVPR, 2020
Video anomaly detection demo

• On Ped2 dataset

Abnormal events: unusual lorry and bicycle.
Video anomaly detection demo

• On Avenue dataset

Abnormal event: kid running.
Video anomaly detection demo

- On ShanghaiTech dataset

ShanghaiTech Test Video 04_0001

Abnormal events: chasing and jumping.
Conclusion

- Design the Multi-Level Memory Autoencoder with Skip Connections (ML-MemAE-SC) for flow reconstruction.
- Propose to model the consistency between flows and frames by leveraging the conditional Variational Autoencoder (CVAE).
- Design a novel hybrid framework in a combination of flow reconstruction and flow-guided frame prediction, named as $HF^2$-VAD.
Project QR Code

https://github.com/LiUzHiAn/hf2vad

Thank you!