Appearance-Motion Co-memory Network for Video Anomaly Detection

Supervisor:
Dr. Phan Duy Hung
Presenter:
Le Duc Anh
Nguyen Ba Duong

FPT University, Hanoi, Vietnam

Our team

Name: Nguyen Ba Duong
Email: duongnbhe130658@fpt.edu.vn
Computer science student
FPT University, Hanoi campus

Name: Le Duc Anh
Email: anhldhe130082@fpt.edu.vn
Computer science student
FPT University, Hanoi campus

Table of content

Introduction

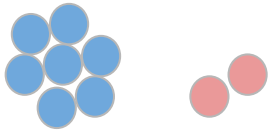
What is Video Anomaly Detection?

Anomalies are deviations from normal activities.

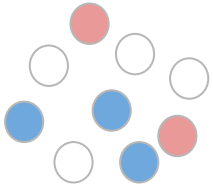
Applications

- Intrusion Detection Systems
- Fraud Detection Systems
- Surveillance Systems
- Defect Detection in Images
- Event Detection

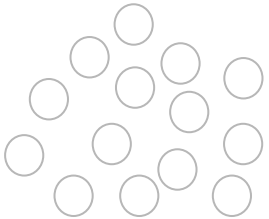
Challenges



High Imbalance
normal > anomalies



Scarce Labels
expensive, time

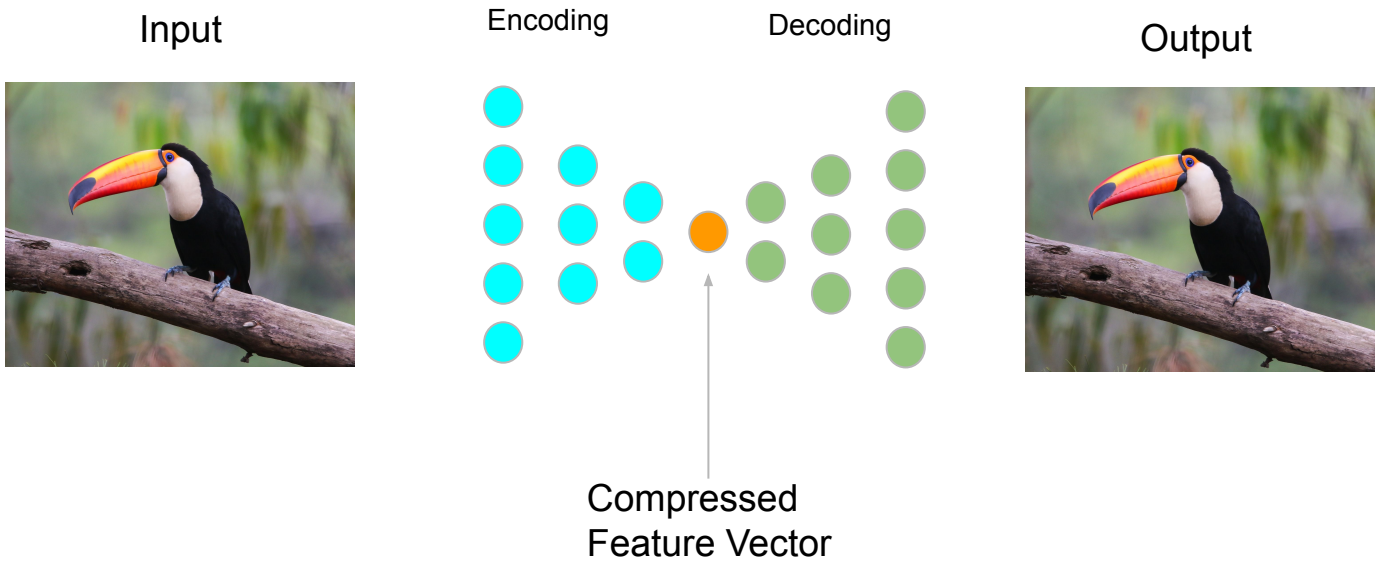


Ambiguity
anomalies depend on context

Related works

Autoencoder (AE)

Reconstruct input from unsupervised pre-training



Base models

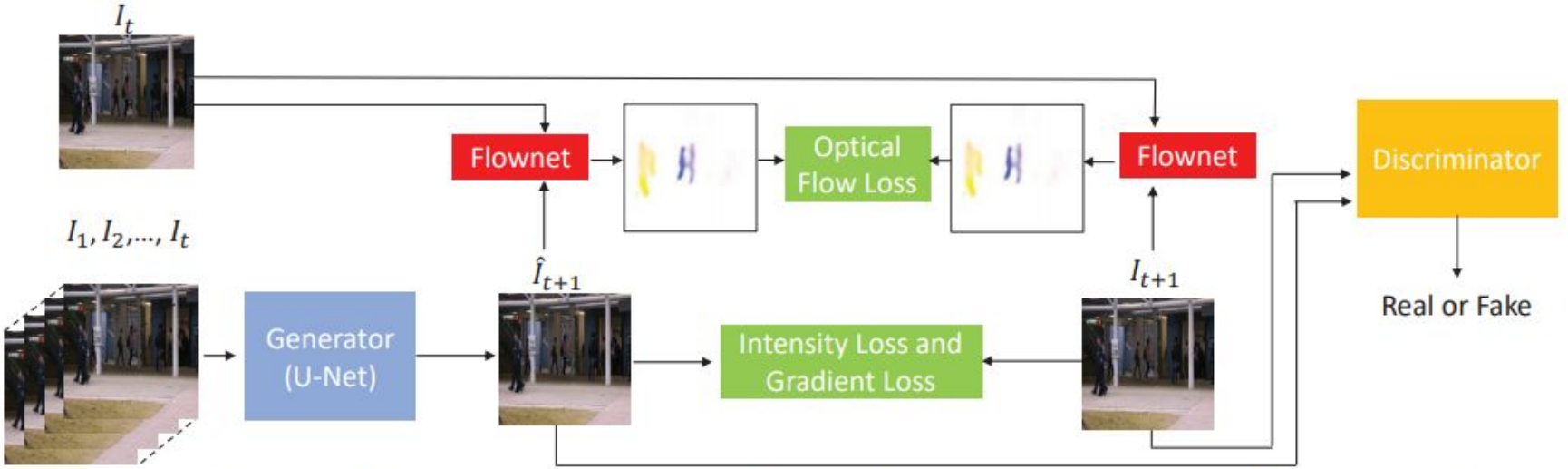


Figure 2. The pipeline of our video frame prediction network. Here we adopt U-Net as generator to predict next frame. To generate high quality image, we adopt the constraints in terms of appearance (intensity loss and gradient loss) and motion (optical flow loss). Here Flownet is a pretrained network used to calculate optical flow. We also leverage the adversarial training to discriminate whether the prediction is real or fake.

Image by: [Future Frame Prediction for Anomaly Detection – A New Baseline]

Base models

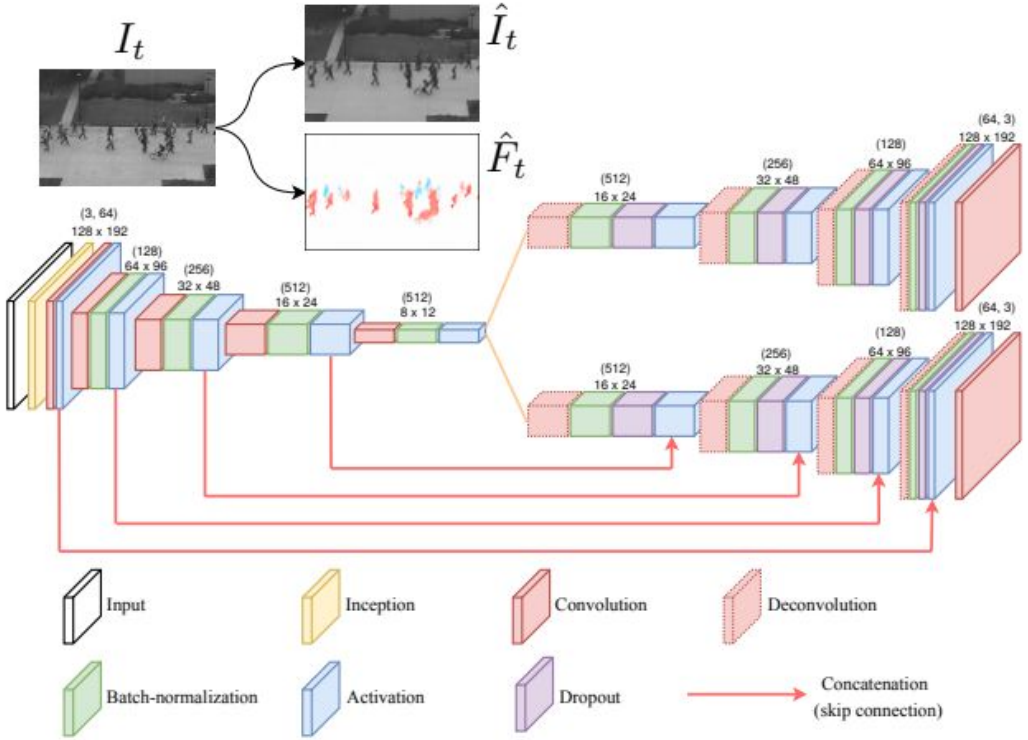


Figure 1. Overview of our model structure together with the spatial resolution of feature maps in each block (*i.e.* a sequence of layers with the same output shape). The number of channels corresponding to each layer in each block is also presented (in parentheses). The input and two output layers have the same size of $128 \times 192 \times 3$. There are three clusters of layers: common encoder (left), appearance decoder (top right) and motion decoder (bottom right). Each concatenation is performed along the channel axis right before operating the next deconvolution. The model input is a single video frame I_t and the outputs from the two decoders are a reconstructed frame $\hat{I}_t$ and an optical flow $\hat{F}_t$ predicting the motion between I_t and I_{t+1}. Best viewed in color.

Image by: [Anomaly Detection in Video Sequence with Appearance-Motion Correspondence]

Base models

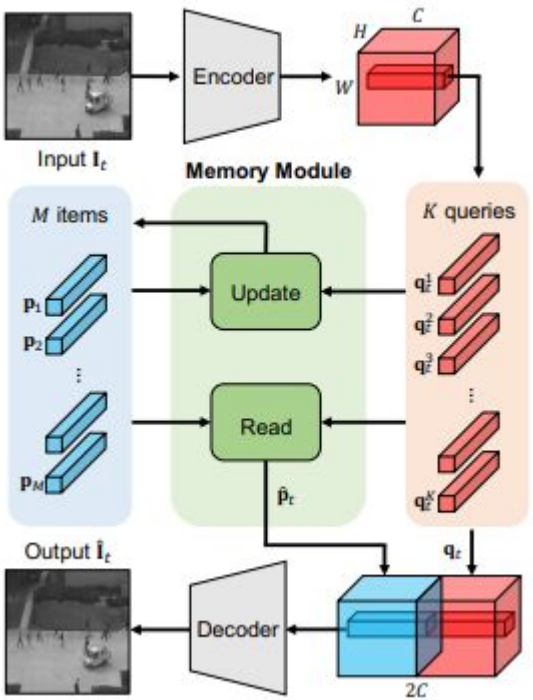


Figure 2: Overview of our framework for reconstructing a video frame. Our model mainly consists of three parts: an encoder, a memory module, and a decoder. The encoder extracts a query map $\mathbf{q}_t$ of size $H \times W \times C$ from an input video frame $\mathbf{I}_t$ at time t. The memory module performs reading and updating items $\mathbf{p}_m$ of size $1 \times 1 \times C$ using queries $\mathbf{q}_t^k$ of size $1 \times 1 \times C$, where the numbers of items and queries are M and K, respectively, and $K = H \times W$. The query map $\mathbf{q}_t$ is concatenated with the aggregated (*i.e.*, read) items $\hat{\mathbf{p}}_t$. The decoder then inputs them to reconstruct the video frame $\hat{\mathbf{I}}_t$. For the prediction task, we input four successive video frames to predict the fifth one. (Best viewed in color.)

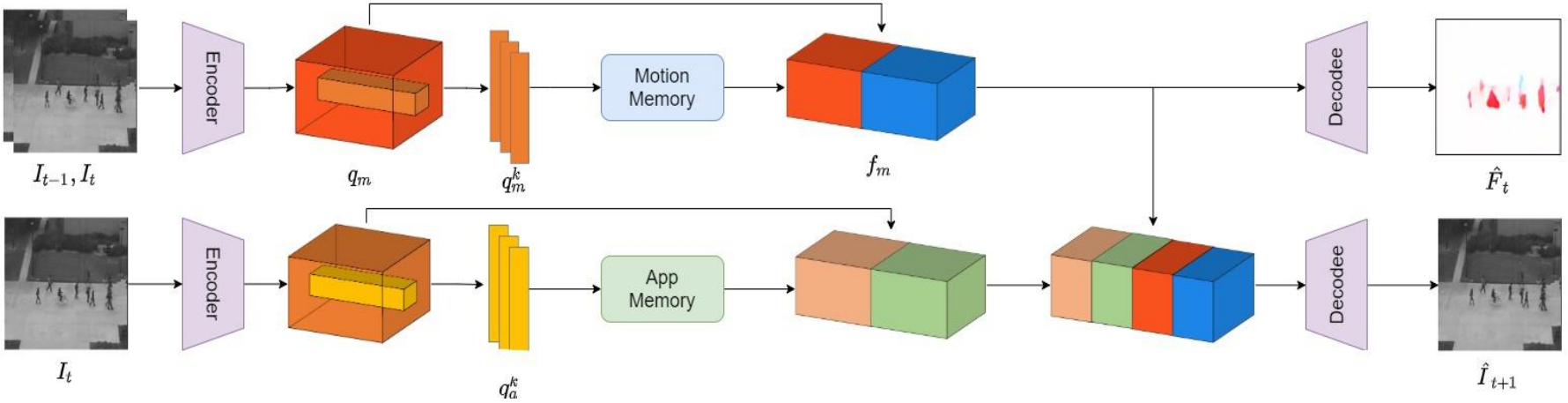
Image by: [Learning Memory-guided Normality for Anomaly Detection]

Contributions

- We proposed a model that includes a motion branch with a motion representation extraction task combined with an appearance branch to increase prediction efficiency. Also integrates the memory module into the motionbranch to store information about the normal movement of objects.
- We proposed a loss function synthesized from memory, prediction, optical flow losses to take advantage of these loss functions to the training model.
- Our model was trained on 2 different datasets and resulted in 2.7% (UCSD Ped2 dataset), 0.9% (CUHK Avenue dataset) increase compared to the original MNAD model. Hence, we have achieved state-of-the-art with both datasets Ped2 and Avenue for abnormal detection in video frames.
- We also provide a breakdown of memory size - an important factor in memory modules.

Proposal Method

Architecture Overview



- Our model consists of two branches motion (upper branch) and appearance (lower branch) with each branch is composed of 3 main components: **encoder**, **memory**, and **decoder** modules.

Motion branch

- The motion-encoder converts the consecutive frames I_{t-1} ,I_t to motion query representations q_m^k.

$$p_m^k = M_{motion}(q_m^k)$$

- Applying the advantage of the skip connection in using the concatenate operation to fuse the feature query map q_m with obtained normal motion feature $\hat{p}_m$.

$$f_m = Concat(q_m, \hat{p}_m)$$

Motion branch

- The motion-decoder D_{motion} inputs the feature f_m and gives an optical flow F_t.

$$\hat{F}_t = D_{motion}(f_m)$$

Appearance branch

- The process of extracting feature representations and reading normal patterns is similar to the motion branch it inherits.

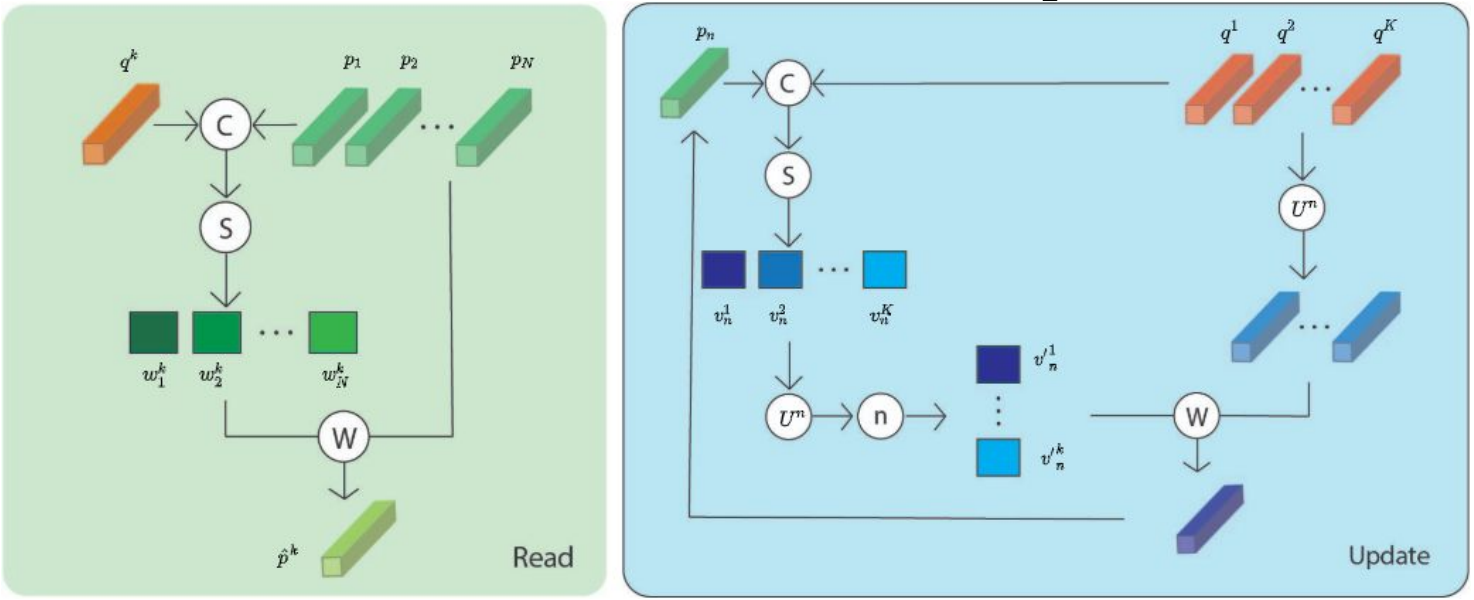
$$f_a = Concat(q_a, \hat{q}_a)$$

- where qa is appearance feature query map and q_a is normal appeara $\hat{p}_a$ e features .
- To complete the information required for prediction, motion features are exploited from motion memory f_m is combined with f_a in order to predict the future frame $\hat{I}_{t+1}$:

$$f = Concat(f_a, f_m)$$

Co-memory module

- The architecture of the memory module:



Co-memory module

- **Read**:
  - First, we calculate the matching probabilities w_n^k between each feature query q^k and all normal patterns p_n in the memory:

$$w_n^k = \frac{\exp((p_n)^T q^k)}{\sum_{\acute{n}=1}^{N} \exp((p_{\acute{n}})^T p^k)}$$

  - Second, we compute the feature $\hat{p}^k$:

$$\hat{p}^k = \sum_{\acute{n}=1}^{N} w_n^k P_{\acute{n}}$$

Co-memory module

- **Update**:
  - For updating normal patterns in memory, we first also use cosine similarities and softmax functions to achieve matching probabilities v_n^k :

$$v_n^k = \frac{\exp((p_n)^T q^k)}{\sum_{\acute{k}=1}^{K} \exp((p_n)^T q^{\acute{k}})}$$

Co-memory module

- **Update**:
  - We choose K's highest matching probability in v_N^K, which means each query will be assigned to a normal pattern in the memory. These $\acute{v}_n^k$ are renormalized through a class of max normalization:

$$\acute{v}_n^k = \frac{v_n^k}{\max_{\acute{k} \in U^n} v_n^{\acute{k}}}$$

Co-memory module

- **Update**:
  - To the end, memory update normal patterns are based on matching index queries in U^n through g(.) is L2 norm:

$$P_n \longleftarrow g(p_n + \sum_{k \in U^n} \acute{v}_n^k q^k)$$

Loss functions

- **Optical flow loss**:

$$L_{flow}(\hat{F}_t, I_{t-1}, I_t) = ||\hat{F}_t - f(I_{t-1}, I_t)||_2$$

- **Prediction loss:**

$$L_{pred}(\hat{I}_{t+1}, I_{t+1}) = ||\hat{I}_{t+1} - I_{t+1}||_2$$

- **Compactness loss:**

$$L_{compact} = \sum_{k}^{K} ||q^k - p_i||_2 \qquad i = \arg\max_{n \in N} w_n^k$$

Loss functions

- **Separate loss**:

$$L_{separate} = \sum_{k}^{K} [||q^k - p_i||_2 - ||q^k - p_j||_2 + \alpha] \quad j = \arg \max_{n \in N, n \neq i} w_n^k$$

- **Total loss:**

$$Loss = \lambda_f L_{flow} + \lambda_p L_{pred} + \lambda_{ac} L_{a-compact} + \lambda_{as} L_{a-separate} + \lambda_{mc} L_{m-compact} + \lambda_{ms} L_{m-separate}$$

Abnormality Score

- The estimated weighted combination score of motion and appearance branch defines as:

$$S = \log[w_F S_F(\tilde{\mathrm{P}})] + \lambda_S[w_I S_I(\tilde{\mathrm{P}})]$$

- where

$$\begin{cases} S_I(P) = \frac{1}{|P|}\sum_{i,j\in P}(I_{i,j} - \hat{I}_{i,j})^2 \\ S_F(P) = \frac{1}{|P|}\sum_{i,j\in P}(F_{i,j} - \hat{F}_{i,j})^2 \end{cases} \quad \begin{cases} w_F = [\frac{1}{n}\sum_{i=1}^{n} S_{F_i}(\tilde{\mathrm{P}}_\mathrm{i})]^{-1} \\ w_I = [\frac{1}{n}\sum_{i=1}^{n} S_{I_i}(\tilde{\mathrm{P}}_\mathrm{i})]^{-1} \end{cases}$$

$$\tilde{\mathrm{P}} \longleftarrow \underset{P \text{ slide on } frame}{\arg\max} \; S_F(P).$$

Abnormality Score

- Finally, the abnormal score of each frame t in a video with m frames can be computed as:

$$\hat{S}_t = \frac{S_t}{\max(S1..m)}$$

Experimental Results

Datasets

For evaluation the effectiveness of our model, we use the two most common benchmark dataset:

1. UCSD Ped2 dataset contains 4560 frames which separated into 2550 use for training and 2010 using for testing, and the rare event is riding a bike and driving the vehicle.
2. CUHK Avenue dataset consists of 30652 frames that split into 16 clips for training and 21 abnormally event clips for testing. The irregular action contains 47 abnormal events such as anomaly actions (e.g., running, throwing), wrong moving direction, anomaly object (e.g., bicycle).

Datasets

- Example of anomaly events in Ped2 dataset:

Datasets

- Example of anomaly events in Avenue dataset:

Strange action	Wrong direction	Abnormal object
		

Results analysis

Method	Ped2 dataset	Avenue dataset
AMCorrespondence [24]	0.962	0.869
Frame Prediction [15]	0.954	0.851
TSC [31]	0.910	0.806
Stacked RNN [31]	0.922	0.817
ConvLSTM-AE [55]	0.881	0.770
Abnormal GAN [14]	0.935	-
Any-Shot [56]	0.978	0.864
CDAE [57]	0.965	0.860
AMCM [22]	0.966	0.866
MemAE [23]	0.941	0.833
CAC [58]	-	0.870
MNAD [16]	0.970	0.885
Our Method w/o motion memory	0.970	0.890
Our Method w motion memory	**0.997**	**0.894**

Experimental Results

Dataset	M-memory size	A-memory size	AUC
Ped2	0	10	0.970
	10	10	**0.997**
	15	15	0.981
Avenue	0	10	0.890
	10	10	0.891
	15	15	**0.894**

Conclusion

Future Works

Many improvements may be applied to the method proposed by us to solve the limitations stated above. Moreover, we introduce some directions for our future works.

- First, limiting the number of input frames can make the model design easier, making our model more challenging to deal with long-term dependencies. The solution is to adopt LSTM or Conv3D models to represent long-term motion factors instead of convolution in the encoder of the motion branch.
- Secondly, integration of the update operation during the testing process is affected by the calculation of W_I, W_F parameters during the update, which can change with each update normal pattern in the memory module. Therefore, it is necessary to design a new abnormality score to reduce the effect of update operation on the coefficients while ensuring the properties of the consideration score.
- Third, for the improvement of inferences times purpose, the replacement of the optical flow extractor to RGB difference maybe apply in our model to reduce the complexation of our model, which inspired by

Close remark

This thesis has introduced frameworks where deep convolutional networks learn Spatio-temporal dynamics on optical flow fields and predict a future frame with the additional memory module. Our approach has stepped toward archiving the exploiting motion and appearance feature based on quantitative and qualitative results while extract normal-activation features more efficiently. Moreover, our proposal approach archives state-of-the-art on two benchmark datasets. In conclusion, we hope our proposal method would substantially improve where a normal manifold is described effectively to detect anomaly events advantageously.

Thank you for your attention!
Feel free to ask any questions.

Contacts:
Le Duc Anh
anhldhe130082@fpt.edu.vn
Student of Computer Science
FPT University
Hanoi, Viet Nam

Website
http://fpt.edu.vn/

Contacts:
Nguyen Ba Duong
duongnbhe130658@fpt.edu.vn
Student of Computer Science
FPT University
Hanoi, Viet Nam

Website
http://fpt.edu.vn/