VIDEO ANOMALY DETECTION VIA PREDICTION NETWORK WITH ENHANCED SPATIO-TEMPORAL MEMORY EXCHANGE

Guodong Shen, Yuqi Ouyang, and Victor Sanchez

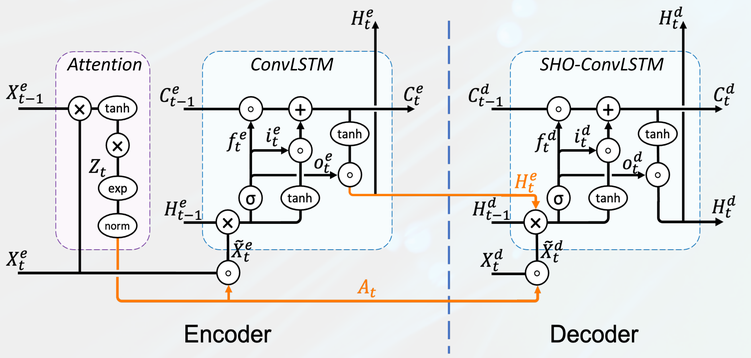
Signal and Information Processing Lab, Department of Computer Science, University of Warwick, Coventry, UK

Paper ID: 2569

Introduction

- Video anomaly detection aims to detect abnormal patterns in motion and appearance. The main challenges of this task come from the ***rarity*** and ***diversity*** of anomalies in the real world.
- This work presents a ***spatio-temporal memory-enhanced Convolutional LSTM Auto-Encoder (ConvLSTM-AE)*** framework to infer anomalies based on the dissimilarity between ***prediction*** and ground truth.
- A ***bi-directional*** mechanism lifts the restriction on temporal order, thus prompting the framework's responsiveness to the time dimension.
- A ***spatial higher-order ConvLSTM (SHO-ConvLSTM)*** mechanism enables the decoder to retrieve hidden states from the current encoder to boost spatial information exchange.
- An ***attention*** module dynamically highlights predictive features to cope with the inflexible receptive fields in ***ConvLSTM***s.

SHO-ConvLSTM and Attention



SHO-ConvLSTM:

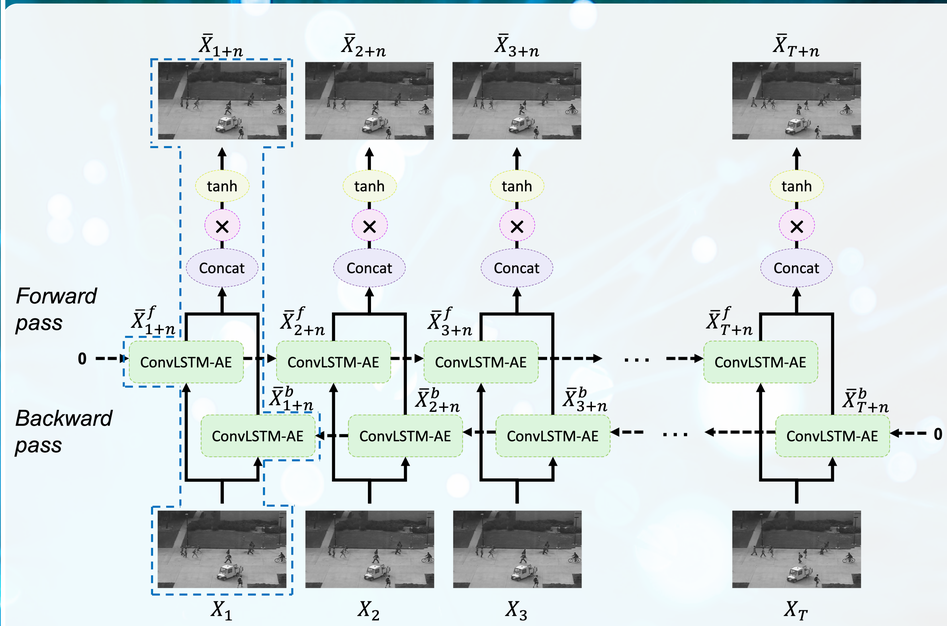
$$\bar{C}_t^d = \tanh\left(W_c \otimes \left[\tilde{X}_t^d, H_{t-1}^d, \boldsymbol{H}_t^e\right] + b_c\right)$$
$$i_t^d = \sigma\left(W_i \otimes \left[\tilde{X}_t^d, H_{t-1}^d, \boldsymbol{H}_t^e\right] + b_i\right)$$
$$f_t^d = \sigma\left(W_f \otimes \left[\tilde{X}_t^d, H_{t-1}^d, \boldsymbol{H}_t^e\right] + b_f\right)$$
$$o_t^d = \sigma\left(W_o \otimes \left[\tilde{X}_t^d, H_{t-1}^d, \boldsymbol{H}_t^e\right] + b_o\right)$$
$$C_t^d = f_t^d \circ C_{t-1}^d + i_t^d \circ \bar{C}_t^d$$
$$H_t^d = o_t^d \circ \tanh\left(C_t^d\right)$$

Attention:

$$Z_t = W_z^2 \otimes \tanh\left(W_z^1 \otimes \left[X_t^e, X_{t-1}^e\right] + b_z\right)$$
$$A_t^{ij} = \frac{\exp\left(Z_t^{ij}\right) - \min_{i,j}\left(\exp\left(Z_t^{ij}\right)\right)}{\max_{i,j}\left(\exp\left(Z_t^{ij}\right)\right) - \min_{i,j}\left(\exp\left(Z_t^{ij}\right)\right)}$$

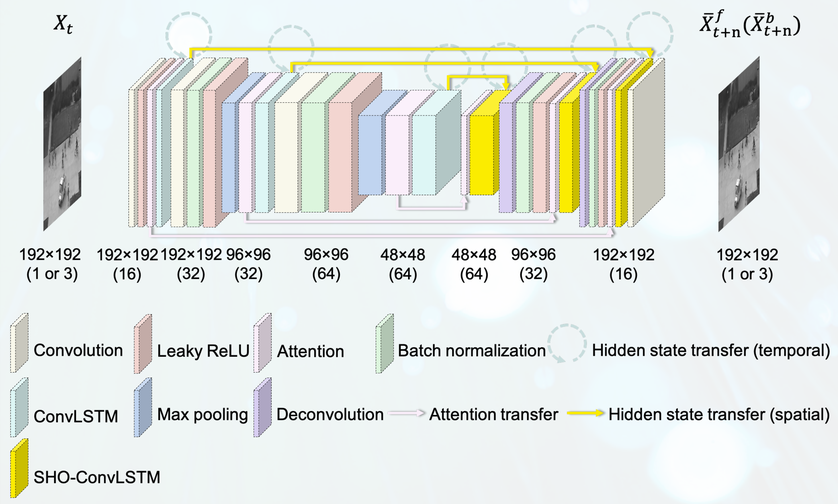
The ***SHO-ConvLSTM*** considers both the temporal and spatial dynamics. The ***attention*** is computed from the encoding feature maps and then imposed on both ***ConvLSTM***s and ***SHO-ConvLSTM***s.

Proposed Framework Overview



This framework leverages a ***Bi-LSTM*** as the backbone, where two enhanced ***ConvLSTM-AE***s are used to perform frame-to-frame prediction in forward and backward order.

ConvLSTM-AE Architecture



The ***ConvLSTM-AE*** consists primarily of Conv, Deconv, ***ConvLSTM***, ***SHO-ConvLSTM*** and ***attention*** layers.

SSIM + L1 Loss

$$\mathcal{L}^{ssim}(P,\hat{P}) = \frac{\left(2\mu_P\mu_{\hat{P}} + c_1\right)\left(2\sigma_{P\hat{P}} + c_2\right)}{\left(\mu_P^2 + \mu_{\hat{P}}^2 + c_1\right)\left(\sigma_P^2 + \sigma_{\hat{P}}^2 + c_2\right)} \qquad \mathcal{L}^{\ell_1}(P,\hat{P}) = \frac{1}{|P|}\sum_{i,j\in P}\left|p_{ij} - \hat{p}_{ij}\right|$$

$$\mathcal{L}^{mix} = \mathcal{L}^{ssim} + \lambda\left(W_{\ell_1} \otimes \mathcal{L}^{\ell_1} + b_{\ell_1}\right)$$

Anomaly Inference

The ***mean absolute error (MAE)*** is calculated for each frame and then normalized over each video to obtain anomaly score ***S(t)*** by using:

$$S(t) = \frac{MAE(t) - \min_t(MAE(t))}{\max_t(MAE(t)) - \min_t(MAE(t))}$$

Comparison with the SOTA

Authors	UCSD Ped2	CUHK Avenue	ShanghaiTech
Luo et al. [8]	88.1	77.0	-
Luo et al. [22]	92.2	81.7	68.0
Wang et al. [9]	88.9	**90.3**	-
Liu et al. [4]	95.4	85.1	72.8
Lee et al. [10]	**96.6**	90.0	**76.2**
Song et al. [11]	90.3	89.2	70.0
Chen et al. [6]	**96.6**	87.8	-
Lai et al. [5]	95.8	87.4	-
Our framework	**98.3**	**90.7**	**79.7**

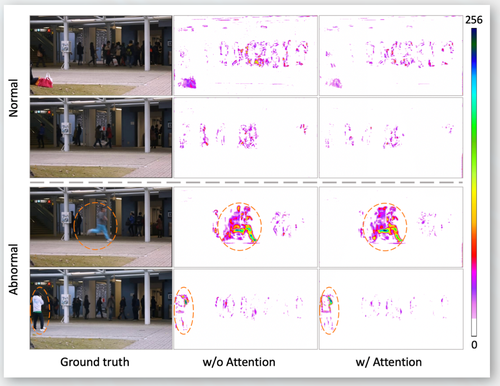
Please find the references in the paper.

AUC (%) of different models on three public benchmarks. A higher AUC value suggests a higher probability that the framework will correctly detect abnormal frames.

Ablation Study

Index	Framework design			Val loss ($\times 10^{-3}$)	AUC (%)
	Bi	SHO	Att		
𝔸	✗	✓	✓	17.58	88.1
𝔹	✓	✗	✓	10.49	88.6
ℂ	✓	✓	✗	4.62	89.8
𝔻	✓	✓	✓	4.54	90.7

Bi = Bi-directional. SHO = Spatial higher-order. Att = Attention. Val loss = Validation loss.



Evaluation of different components of our framework on the CUHK Avenue dataset. The results are assessed in terms of AUC and convergence ability (i.e., optimal validation loss).

Visualizations of the attention module's impact on normal and abnormal frames. The figure demonstrates that the attention module helps to diminish the prediction error around normal moving items, while preserving it in abnormal areas.