Divide-and-Assemble: Learning Block-wise Memory for Unsupervised Anomaly Detection

Jinlei Hou[1,2†], Yingying Zhang[1], Qiaoyong Zhong[1], Di Xie[1], Shiliang Pu[1⋆], Hong Zhou[2]
[1]Hikvision Research Institute [2]Zhejiang University
{houjinlei,zhangyingying7,zhongqiaoyong,xiedi,pushiliang.hri}@hikvision.com
zhouh@mail.bme.zju.edu.cn

Supplementary Material

Impact of Memory Size. Figure 1 shows the impact of the memory size, i.e. the number of items N in the memory bank to the detection performance. As the memory size increases, AUROC gets steadily improved, and saturates after 500. Notably, increasing the memory size further does not lead to an obvious performance degradation. It indicates that the proposed block-wise memory module enjoys the benefit of good reconstruction on normal samples without worrying about learning an identity mapping, which would not be possible by simply increasing the model size.

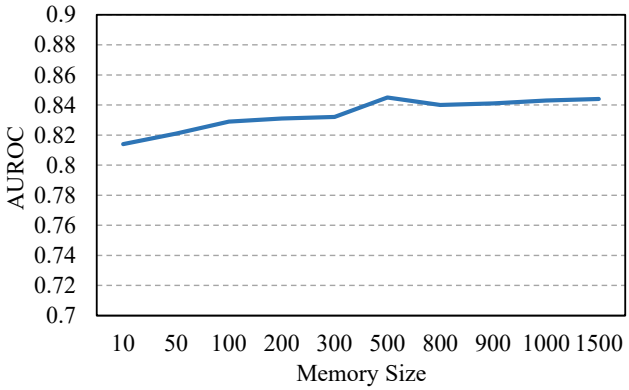

Figure 1. The AUROC score of DAAD over different memory sizes on the MVTec AD dataset.

Experiments on Video Datasets Our model is designed for anomaly detection in images, but to validate the generalization of our method, we conduct experiments on two real-world video anomaly detection datasets, i.e. UCSD-Ped2 [3] and CUHK Avenue [4]. Following the experimental setting in [6], we resize each frame of the video into the size of 256×256, and normalize the pixel values to the range of $[-1, 1]$. r_h, r_w, r_c and the memory bank size N are empirically set to 16, 16, 1 and 2000. The model is optimized via the Adam optimizer with a fixed learning rate of $2e^{-4}$, a weight decay of 0, momentums of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size of 4 for 60 epochs on both UCSD Ped2 and CUHK Avenue. As shown in Table 1, compared with other reconstruction-based methods, our model can still achieve competitive results. It is worth noting that our model has not been tailored for video data as MemAE does, where 3D convolution is utilized to exploit the temporal information in videos. We leave such improvement as future work.

Method\Dataset	UCSD-Ped2	Avenue
AE-Conv2D [2]	0.850	0.800
AE-Conv3D [9]	0.912	0.771
TSC [5]	0.910	0.806
StackRNN [5]	0.922	0.817
AbnormalGAN [8]	0.935	-
AE [1]	0.917	0.810
MemAE [1]	0.941	0.833
AE [6]	0.864	0.806
MNAD [6]	0.902	0.828
AE	0.901	0.805
Ours	0.941	0.829

Table 1. Comparison with existing methods on two video datasets (UCSD-Ped2 and CUHK Avenue) in terms of AUROC.

Structure of DAAD Table 2 shows the structure of DAAD. We use the skip connection to improve the reconstruction ability of our model and furthermore utilize the block-wise memomry module to balance the reconstruction of the normality and anomaly.

Structure of the Discriminator Table 3 shows the structure of the discriminator used in DAAD+. Its architecture is designed by following the discriminator of DCGAN [7]. The dimension of the flattened output representation from the feature extractor is 100 if the size of the input image is 256×256.

Layer	Filter	Channels	Input	Output
ConvBNReLU	$3 \times 3, s1$	64	$input$	enc_{1-1} $(H \times W)$
ConvBNReLU	$3 \times 3, s1$	64	enc_{1-1}	enc_{1-2} $(H \times W)$
Maxpooling	$2 \times 2, s2$	64	enc_{1-2}	enc_{1-3} $(H/2 \times W/2)$
BW Memory1	-	-	enc_{1-3}	m_1 $(H/2 \times W/2)$
ConvBNReLU	$3 \times 3, s1$	128	enc_{1-3}	enc_{2-1} $(H/2 \times W/2)$
ConvBNReLU	$3 \times 3, s1$	128	enc_{2-1}	enc_{2-2} $(H/2 \times W/2)$
Maxpooling	$2 \times 2, s2$	128	enc_{2-2}	enc_{2-3} $(H/4 \times W/4)$
BW Memory2	-	-	enc_{2-3}	m_2 $(H/4 \times W/4)$
ConvBNReLU	$3 \times 3, s1$	256	enc_{2-3}	enc_{3-1} $(H/4 \times W/4)$
ConvBNReLU	$3 \times 3, s1$	256	enc_{3-1}	enc_{3-2} $(H/4 \times W/4)$
Maxpooling	$2 \times 2, s2$	256	enc_{3-2}	enc_{3-3} $(H/8 \times W/8)$
BW Memory3	-	-	enc_{3-3}	m_3 $(H/8 \times W/8)$
ConvBNReLU	$3 \times 3, s1$	512	enc_{3-3}	enc_{4-1} $(H/8 \times W/8)$
ConvBNReLU	$3 \times 3, s1$	512	enc_{4-1}	enc_{4-2} $(H/8 \times W/8)$
Maxpooling	$2 \times 2, s2$	512	enc_{4-2}	enc_{4-3} $(H/16 \times W/16)$
BW Memory4	-	-	enc_{4-3}	m_4 $(H/16 \times W/16)$
ConvBNReLU	$3 \times 3, s1$	1024	m_4	dec_{4-1} $(H/16 \times W/16)$
ConvBNReLU	$3 \times 3, s1$	1024	dec_{4-1}	dec_{4-2} $(H/16 \times W/16)$
ConvTranspose	$2 \times 2, s2$	512	dec_{4-2}	dec_{4-3} $(H/8 \times W/8)$
ConvBNReLU	$3 \times 3, s1$	512	$[dec_{4-3}, m_3]$	dec_{3-1} $(H/8 \times W/8)$
ConvBNReLU	$3 \times 3, s1$	512	dec_{3-1}	dec_{3-2} $(H/8 \times W/8)$
ConvTranspose	$2 \times 2, s2$	256	dec_{3-2}	dec_{3-3} $(H/4 \times W/4)$
ConvBNReLU	$3 \times 3, s1$	256	$[dec_{3-3}, m_2]$	dec_{2-1} $(H/4 \times W/4)$
ConvBNReLU	$3 \times 3, s1$	256	dec_{2-1}	dec_{2-2} $(H/4 \times W/4)$
ConvTranspose	$2 \times 2, s2$	128	dec_{2-2}	dec_{2-3} $(H/2 \times W/2)$
ConvBNReLU	$3 \times 3, s1$	128	$[dec_{2-3}, m_1]$	dec_{1-1} $(H/2 \times W/2)$
ConvBNReLU	$3 \times 3, s1$	128	dec_{1-1}	dec_{1-2} $(H/2 \times W/2)$
ConvTranspose	$2 \times 2, s2$	64	dec_{1-2}	dec_{1-3} $(H \times W)$
ConvBNReLU	$3 \times 3, s1$	64	dec_{1-3}	$head_{1-1}$ $(H \times W)$
ConvBNReLU	$3 \times 3, s1$	64	$head_{1-1}$	$head_{1-2}$ $(H \times W)$
Conv	$1 \times 1, s1$	3	$head_{1-2}$	$output$ $(H \times W)$

Table 2. Structure of DAAD.

Name	Layer	Filter	Channels	Stride
Feature Extractor	Conv	4×4	64	2
	LeakyReLU	$negative\_slope = 0.2$		
	Conv	4×4	128	2
	BatchNorm	-		
	LeakyReLU	$negative\_slope = 0.2$		
	Conv	4×4	256	2
	BatchNorm	-		
	LeakyReLU	$negative\_slope = 0.2$		
	Conv	4×4	512	2
	BatchNorm	-		
	LeakyReLU	$negative\_slope = 0.2$		
	Conv	4×4	1024	2
	BatchNorm	-		
	LeakyReLU	$negative\_slope = 0.2$		
	Conv	4×4	2048	2
	BatchNorm	-		
	LeakyReLU	$negative\_slope = 0.2$		
	Conv	4×4	100	1
Classifier	Conv	3×3	1	1
	Sigmoid			

Table 3. Structure of the discriminator used in DAAD+.

References

[1] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1705–1714, 2019. 1

[2] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 1

[3] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013. 1

[4] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 1

[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. 1

[6] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020. 1

[7] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[8] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017. 1

[9] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941, 2017. 1