基于改进 VGG16 的自编码器视频异常检测算法

杨大为,刘志权
(沈阳理工大学 信息科学与工程学院,辽宁 沈阳 110159)

摘 要:在使用自编码器结构的神经网络处理视频异常检测任务时,U-Net 风格的自编码器由于编码器层数深度过浅,导致在面对复杂的数据集时,不能充分抽取更多有用的特征信息。同时,在训练模型时使用 MSE(均方误差),仅考虑了预测帧与真实帧之间的像素级相似性,对于复杂场景,像素级相似性可能无法准确判断预测帧与真实帧之间的相似性。针对以上问题,对基于 U-Net 风格的自编码器进行改进,提出了一种使用改进的 VGG16 作为编码器的视频异常检测算法,同时在均方误差的基础上添加结构相似性(SSIM)损失函数。改进的 VGG16 去掉了全连接层,并加入了残差连接防止特征退化,添加 SSIM 在计算像素级相似性的同时计算图像的亮度、对比度和结构等方面的相似性来优化网络。实验结果表明,改进后的算法,在 Ped2 数据集上检测效果达到 95.91%,在 Avenue 数据集上检测效果达到 84.89%,与改进前的方法相比分别提高了 0.80% 和 0.19%,验证了所提方法的有效性。

关键词:自编码器;U-Net;特征提取;VGG16;残差连接;结构相似性
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1673-629X(2024)04-0095-06
doi:10.20165/j.cnki.ISSN1673-629X.2024.0015

Auto-encoder Video Anomaly Detection Algorithm Based on Improved VGG16

YANG Da-wei, LIU Zhi-quan
(School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China)

Abstract: When using the auto-encoder structure neural network to process video anomaly detection tasks, the U-Net style auto-encoder cannot fully extract more useful feature information when facing complex data sets due to the shallow depth of the encoder layer. At the same time, when training the model, MSE is used, only considering the pixel level similarity between the predicted frame and the real frame. For complex scenes, pixel level similarity may not accurately determine the similarity between the predicted frame and the real frame. To solve the above problems, the U-Net style auto-encoder is improved, and a video anomaly detection algorithm using the improved VGG16 as the encoder is proposed. At the same time, the structure similarity (SSIM) loss function is added on the basis of MSE. The improved VGG16 removes the fully connected layer and adds residual connections to prevent feature degradation. SSIM is added to optimize the network by calculating pixel level similarity while also calculating image brightness, contrast, and structural similarity. The experimental results show that the improved algorithm achieves a detection performance of 95.91% on the Ped2 dataset and 84.89% on the Avenue dataset, which is 0.80% and 0.19% higher than that of the previous method, respectively, verifying the effectiveness of the proposed method.

Key words: auto-encoder; U-Net; feature extraction; VGG16; residual connection; structure similarity

0 引 言

视频异常检测[1-2]是对视频流进行实时监控,自动检测视频中是否出现非正常行为,以保障公共安全和财产安全。视频异常检测在众多领域中得到广泛的应用,如城市监控、交通管理、工业制造等。

随着深度学习的发展,基于深度学习的方法在视频异常检测中得到了广泛应用。文献[3]提出了 2 种基于自编码器的方法,首先在手工制作的特征上训练一个完全连接的自动编码器。其次建立学习局部特征和分类器的卷积前馈自动编码器作为一个端到端学习框架。但由于神经网络有强大的表征能力,导致异常样本也会较好重构。针对这一问题,文献[4]使用记

收稿日期:2023-06-14 **修回日期**:2023-10-18
基金项目:辽宁省教育科学研究经费项目(LG201915)
作者简介:杨大为(1976-),男,教授,博士,研究方向为机器学习;通信作者:刘志权(1999-),男,硕士研究生,研究方向为大数据与智能信息处理技术。

忆模块来增强自编码器,并开发一种改进的自编码器,称为记忆增强自编码器,即 MemAE(Memory-augmented Auto-Encoder)。对于给定的输入,MemAE 首先从编码器获取编码,然后将其作为查询来检索与重构最相关的内存项,以增强异常的重构误差,用于异常检测。在此基础上,文献[5]使用了新更新方案的记忆模块,其中记忆模块的项记录正常数据的原型模式,降低了神经网络的表示能力。同时提出了新的特征紧凑性和分离损失来训练记忆,提高了特征的辨别能力。文献[6]引入了一个原型学习模块来显式地建模视频序列中的正常动态,并以端到端的方式进行训练,有效降低了内存的训练消耗。这些方法仅在记忆模块上进行优化,没考虑对网络模型的特征提取能力进行优化。文献[7]提出基于预测未来帧的异常事件检测框架,使用 U-Net 生成未来帧,比较真实帧与生成帧的差距来判断事件是否异常,但该方法仅考虑了真实帧与生成帧之间的像素级差距,没有考虑亮度、对比度等方面。文献[8]将长短期记忆(Long Short-Term Memory, LSTM)与现有模型结合,提出 ConvLSTM 网络,对输入的视频序列进行编码重构,并预测当前帧。但使用长短期记忆会大量增加模型参数,导致难以训练。

该文将对视频异常检测网络进行改进,使用 VGG16 作为编码器,并使用残差连接防止特征退化,以此加强模型编码器对输入数据的特征提取能力。在训练模型时,在考虑真实帧与生成帧之间像素级差距的同时对亮度、对比度等方面进行考虑,从而更准确地判断真实帧与生成帧的相似性。实验结果表明,改进后的方法与改进前的方法相比性能得到了提升,验证了所提方法的有效性。

1 相关研究

1.1 自编码器结构

自编码器[9]结构是一种适用于无监督的对称神经网络结构,分为编码器(Encoder)和解码器(Decoder)。编码器会对输入数据 X 进行特征信息提取得到特征 Z,这一过程称为编码。解码器会使用特征信息 Z 上采样重构还原出原始输入数据得到 X^R,这一过程称为解码,如图 1 所示。

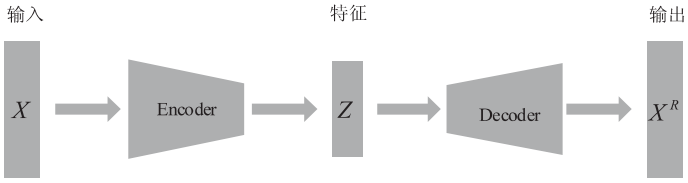

图 1 Auto-Encoder 结构

编码器的作用是将输入数据 X 进行编码得到低维的特征信息变量 Z。从 X 到 Z 的编码过程为:

$$Z = \sigma(W_1 x + b_1) \tag{1}$$

解码器的作用是将编码器得到的特征 Z 还原到原始的维度。从 Z 到 X^R 的编码过程为:

$$X^R = \sigma(W_2 z + b_2) \tag{2}$$

最优的自编码器是输出的 X^R 可以近似等于 X,即 $X \approx X^R$。优化目标函数为:

$$\text{MinimizeLoss} = \text{dist}(X, X^R) \tag{3}$$

1.2 U-Net 模型

为了解决生物医学图像的问题,提出了 U-Net[10]网络模型,其良好的效果在其他领域也非常适用,在无监督的计算机视觉任务上也有着不错的效果。U-Net 是自编码器结构,这种结构不仅简单而且很有效,如图 2 所示。

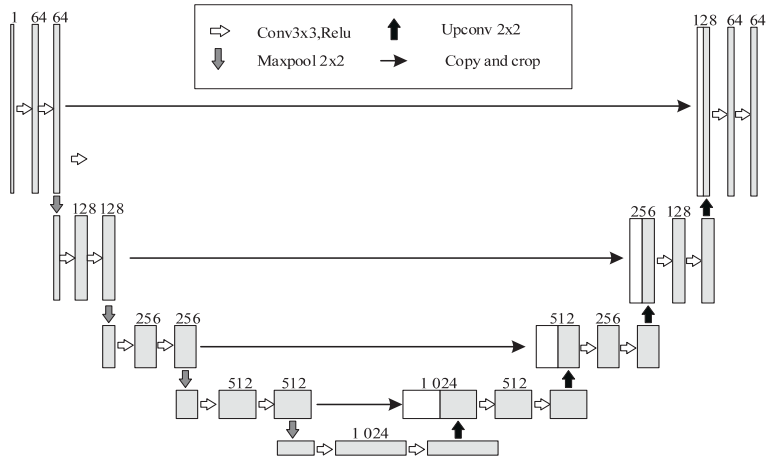

图 2 U-Net 网络结构

编码器部分由一系列的卷积、批量归一化和 ReLU 激活函数构成,逐渐降低图像的分辨率,并提取出高层次的特征信息。在编码器部分,网络主要运用了步长为 2 的卷积操作实现下采样,每次卷积后得到的特征图的尺寸将缩小一半。

解码器部分是编码器的逆过程。在此过程中,网络使用反卷积进行上采样,将分辨率提高到原始输入图像的尺寸,并结合编码器中相同深度对应的特征图进行特征融合。在特征融合时,U-Net 采用跳跃连接,将编码器中的高分辨率特征图与解码器中的低分辨率特征图相连,以保留更细节的特征信息。

2　基于 VGG16 的改进方法

2.1　残差连接

在处理计算机视觉相关任务时通常会加深网络层数从而得到较好的效果。但网络层数的增加会带来很多问题,如梯度消失、梯度爆炸、模型过拟合、特征退化等。该文使用了残差连接[11]试图解决这些问题。

图 3 是残差连接的示意图。用非线性变化函数描述网络的输入和输出,也就是输入 X ,输出 $F(x)$ 。当在函数输出中加入一个输入时, $G(x)$ 可用于描述输入与输出之间的关系,因此, $G(x)$ 可拆成 $F(x)$ 与 X 线性叠加。即:

$$G(X) = F(X) + X \tag{4}$$

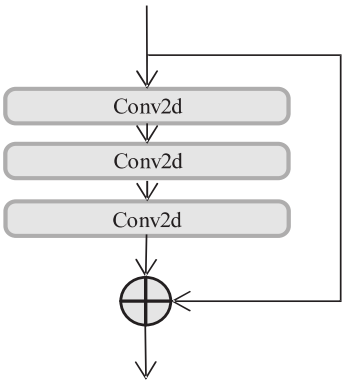

图 3　残差连接

实验发现,残差连接在其他神经网络上同样有优秀的表现。该文在 VGG16 上添加了两个残差连接来解决网络层数太深导致特征退化的问题。

2.2　基于 VGG16 的编码器

VGG16[12]是应用在分类任务的神经网络模型,拥有 13 个卷积层,3 个全连接层。卷积层的卷积核大小均为 3×3,步长为 1。该文只采用 VGG16 的特征提取部分作为编码器,也就是 13 个卷积层。在视频异常检测任务中,通常会将一段视频转换成视频帧,使用前 4 帧图像去预测第 5 帧图像,然后将预测的第 5 帧图像与真实的第 5 帧图像进行误差计算,使预测出来的第 5 帧图像尽可能逼近真实的第 5 帧图像。因此输入数据的通道为 12,输入数据第一次经过 64 个卷积核的 2 次卷积后,再经过一个池化层,第二次经过 2 次 128 个卷积核卷积后,经过池化层,再经过 3 次 256 个卷积核的卷积后,以此类推最终得到 1 024 通道的特征图,如图 4 所示。

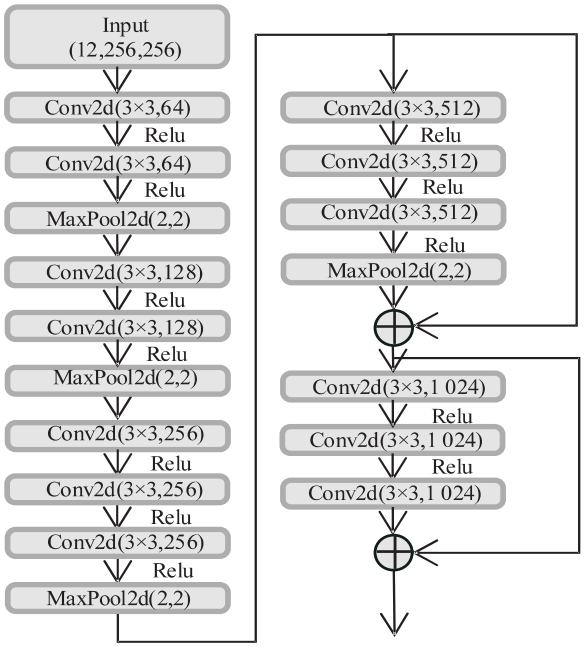

图 4　基于 VGG16 的编码器

2.3　整体网络结构

基于改进 VGG16 的自编码器视频异常检测算法整体网络模型如图 5 所示。网络整体是编码器-解码器结构,风格与 U-net 网络相似。编码器是改进的 VGG16 网络,添加了两个残差连接防止特征退化。解码器有四个上采样,解码时将编码器输入的特征进行上采样,并与浅层的编码器特征进行拼接作为输出。Outhead 是将解码器输出的特征图映射为任务需要的输出结果,即(256,256,3)的图片。

2.4　损失函数

一个好的损失函数可以更好地优化模型。该文使用预测帧与真实帧之间的相似程度作为模型的损失函数,损失函数 L 由两部分组成,分别为帧像素级预测损失 L_{MSE} 和帧结构性预测损失 L_{SSIM} ,如公式 5,其中 λ 为超参数。

$$L = L_{\mathrm{MSE}} + \lambda L_{\mathrm{SSIM}} \tag{5}$$

MSE(均方误差)是预测值与真实值之间的差异的平方的平均值,通过对每个样本的预测误差平方进行求和并除以样本数量求得,具体过程为:

$$L_{\mathrm{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{6}$$

其中, n 表示样本数量, Y_i 表示真实的目标变量值, $\hat{Y}_i$ 表示模型对第 i 个输入样本的预测值。

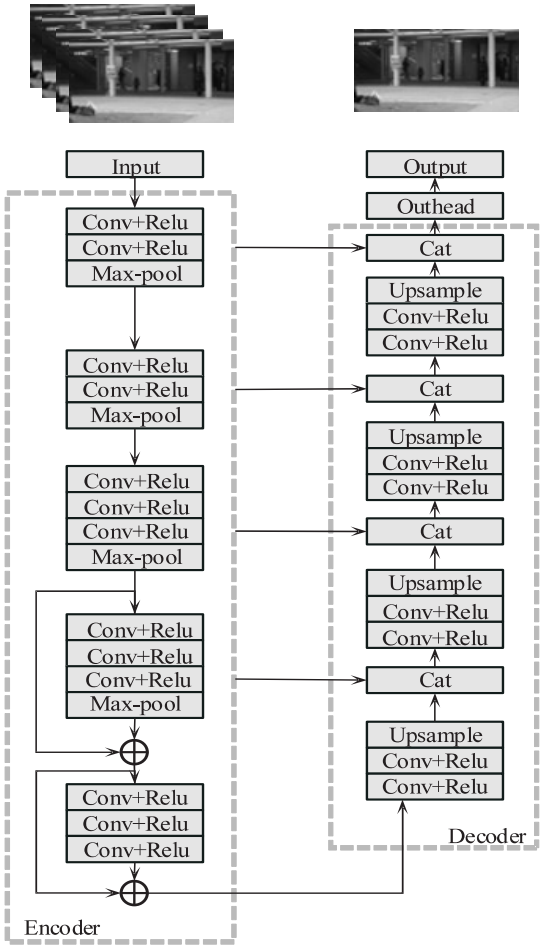

图 5 整体网络结构

SSIM[13]是一种图像质量评价指标,用于度量一段图像与另一段图像之间的结构相似度。SSIM 指标在比较两幅图像时,会分别对它们进行亮度、对比度和结构信息的分析,然后通过计算它们之间的相似性指数得出最终的相似度得分。亮度计算过程如公式 7 所示:

$$l(Y_i,\hat{Y}_i)=\frac{2\mu_{Y}\mu_{\hat{Y}_i}+c_1}{\mu_{Y_i}^2+\mu_{\hat{Y}_i}^2+c_1} \tag{7}$$

其中,μ_{Y_i} 和 $\mu_{\hat{Y}_i}$ 分别表示图像 Y_i 和 $\hat{Y}_i$ 的平均值,c_1 是小的正常数,用于防止分母为 0。

对比度计算过程如公式 8 所示:

$$c(Y_i,\hat{Y}_i)=\frac{2\sigma_{Y_i}\sigma_{\hat{Y}_i}+c_2}{\sigma_{Y_i}^2+\sigma_{\hat{Y}_i}^2+c_2} \tag{8}$$

其中,σ_{Y_i} 和 $\sigma_{\hat{Y}_i}$ 分别表示图像 Y_i 和 $\hat{Y}_i$ 的标准差,c_2 是小的正常数,用于防止分母为 0。

结构计算过程如公式 9 所示:

$$s(Y_i,\hat{Y}_i)=\frac{\sigma_{Y_i\hat{Y}_i}+c_3}{\sigma_{Y_i}\sigma_{\hat{Y}_i}^2+c_3} \tag{9}$$

其中,$\sigma_{Y_i\hat{Y}_i}$ 表示 σ_{Y_i} 和 $\sigma_{\hat{Y}_i}$ 的标准差,c_3 是小的正常数,用于防止分母为 0。

SSIM 计算如公式 10 所示:

$$L_{\mathrm{SSIM}}(Y_i,\hat{Y}_i)=[l(Y_i,\hat{Y}_i)]^{\alpha}[c(Y_i,\hat{Y}_i)]^{\beta}[s(Y_i,\hat{Y}_i)]^{\gamma} \tag{10}$$

其中,α,β,γ 是权重系数,它们都是 0~1 之间的值,一般取 $\alpha=\beta=\gamma=1$。

SSIM 的得分一般在 0 到 1 之间。分数越高意味着两幅图像的相似度越高。通常,当 SSIM 得分在 0.95以上时,代表两幅图像结构基本相似,质量较高。当得分在 0.8 到 0.95 之间时,代表两幅图像基本相似,质量一般。当得分在 0.8 以下时,代表两幅图像相差较大,质量较差。

3 实验结果与分析

3.1 实验数据集

UCSD Ped2[14]是由固定摄像机俯视采集的行人移动的异常检测数据集,视频片段主要包括行人在街道步行,以及骑自行车和观光车等异常事件。UCSD Ped2 包含 16 个训练视频样本(共 1 578 帧图像)和 12 个测试视频样本(共 1 962 帧图像),空间分辨率为 360×240 像素。

Avenue[15]数据集由一台固定视角的摄像机拍摄的短视频片段组成,视频片段主要包括行人进出建筑物,建筑物中的柱子会使人流造成严重的堵塞。Avenue 包含 16 个训练视频样本(共 15 328 帧图像)和 21 个测试视频样本(共 30 652 帧图像),它的空间分辨率为 640×360 像素。

3.2 实验设置

为了方便实验,统一对数据集进行预处理,将所有数据集图像缩放到 256×256 像素,并归一化到[-1,1]。使用训练批量大小为 4,学习率设定为 0.000 1,训练时间为 1 000 轮。所有实验数据均在单块 RTX-3060 GPU,CPU 为 Ryzen 5600@4.4 Hz,16 GB 内存,pytorch1.7 深度学习框架,Windows11 操作系统的环境下得到。

3.3 评估指标

在视频异常检测中常使用 AUC(ROC 曲线下面积)来评估模型,ROC 是以真正率(True Positive Rate,TPR)为纵坐标,假正率(False Positive Rate,FPR)为横坐标的曲线。AUC 不会受到类不平衡问题的影响,具有较好的鲁棒性。AUC 的取值范围是[0,1],AUC 值越大表示模型性能越好,最大值为 1 表示模型预测完全准确。如果 AUC 等于 0.5,表示模型检测效果为随机判定,无实用价值。AUC 计算过程如下:

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{11}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{12}$$

其中,TP 是真阳性指实际阳性的样本被正确地判断为

阳性;FP 是假阳性指实际阴性的样本被错误地判断为阳性。同样,TN 是真阴性表示实际阴性的样本被正确地判断为阴性,而 FN 是假阴性表示实际阳性的样本被错误地判断为阴性。

3.4　实验结果

为了证明改进方法的有效性,该文在同一实验平台下将改进方法与现有典型方法进行对比,其中粗体为最优结果,下划线为次优结果,实验结果如表 1 所示。

表 1　不同数据集下的 AUC 对比　%

方法	Ped2	Avenue
ConvAE-2D[3]	85.04	80.02
ConvAE-3D[16]	91.44	78.23
MemAE[4]	94.10	83.36
MNAD[5]	95.65	85.06
ST-CaAE[17]	94.94	84.12
Ada-Net[18]	92.71	**88.72**
LGN-Net[19]	**96.27**	85.85
Baseline[6]	95.11	84.70
Ours	95.91	84.89

从表 1 可以看到,改进后的方法比改进前的方法在 Ped2 数据集和 Avenue 数据集上的检测效果分别提升了 0.80% 和 0.19%。与 ConvAE-2D 和 ConvAE-2D 相比检测效果提升较大,更深的卷积网络提取到了更多信息,残差网络也很好地降低了特征信息退化。MemAE 和 MNAD 与该文使用了相似的网络风格,但该文在训练时使用了结构相似性损失来优化网络,对比使用像素级差距损失,文中方法更具优势。其中 Ada-Net 和 LGN-Net 使用了长短期记忆(LSTM),对时间信息进行处理,从而准确地捕捉视频中的时间相关性和空间相关性。虽然提高了性能,但使用长短期记忆会大量增加模型参数。文中方法与使用了长短期记忆的典型方法相比,由于 Ped2 数据集分辨率较小,且单个视频帧数较少,所以性能差距不大,但 Avenue 数据集分辨率更大,单个视频帧数也更多,使用长短期记忆的方法将更具优势。从结果来看,文中方法更适用于小型数据集。

3.5　消融实验

为验证提出的改进 VGG16、残差连接和 SSIM 损失函数的必要性,在 Ped2 数据集上进行了消融实验,实验结果如表 2 所示,粗体为最优结果,下划线为次优结果。

通过消融实验发现,在使用 VGG16 不添加残差连接时由于加深了网络层数,所以导致特征退化,从而使检测效果下降。在加入残差连接后,缓解了网络层数过深导致的特征退化,检测效果也得到了提升。在使用 SSIM 损失函数后,模型网络会对输入数据的亮度、对比度和结构进行计算优化,对比不使用 SSIM 的时候,检测效果会更好。

表 2　不同模块对 AUC 的影响　%

VGG16	残差连接	L	Ped2
×	×	×	95.35
√	×	×	93.44
√	√	×	95.69
×	×	√	95.31
√	×	√	95.26
√	√	√	**95.91**

为探究超参数 λ 对模型的影响,在 Ped2 数据集上进行了实验,实验结果如表 3 所示,其中粗体为最优结果,下划线为次优结果。

表 3　超参数 λ 对 AUC 的影响　%

λ	AUC
0.2	94.47
0.4	94.47
0.6	**95.91**
0.8	95.39
1	94.46

通过实验结果可知,超参数 λ 取值为 0.6 时,模型的检测效果最好,过大或过小的 λ 值反而会导致 AUC 降低。

3.6　检测结果可视化

将 Avenue 数据集的异常检测结果进行可视化,可视化结果如图 6 所示,可以看到模型可以生成相似的图像。

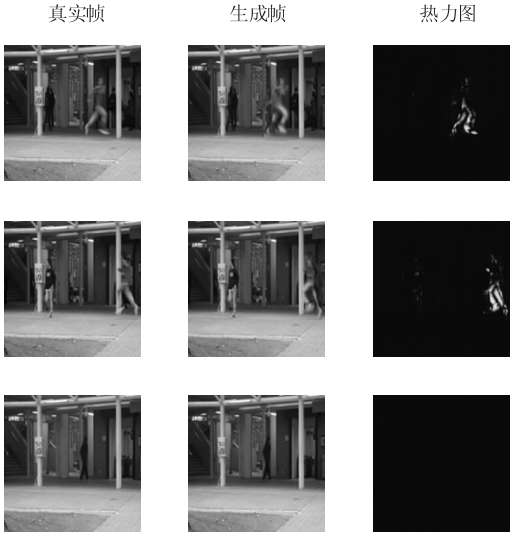

图 6　视频异常检测结果可视化

利用真实帧和生成帧制作热力图,其中高亮的区

域表示两张图像存在误差,即出现异常。模型在训练时只使用正常数据集,所以在测试时当有行人奔跑的数据输入,模型生成的图像会和真实图像之间产生误差,从而在热力图上产生高亮区域。当行人步行时,模型生成的图像会和真实图像之间误差很小,热力图上不产生高亮区域。

4 结束语

对基于 U-Net 风格的自编码器进行改进,该文提出了一种使用改进的 VGG16 作为编码器的视频异常检测算法,并添加结构相似性(SSIM)损失函数。改进的 VGG16 去掉了全连接层,能够让自编码器从编码阶段中抽取更多的特征信息,加入了残差连接防止特征退化,使深层次的网络减少丢失特征信息,加入 SSIM 在像素级相似性的基础上,同时计算图像的亮度、对比度和结构等方面的相似性来优化网络。实验结果显示改进后的方法与改进前的方法相比检测效果得到了提升,证明了所提方法的有效性。

参考文献:

[1] 邬开俊,黄 涛,王迪聪,等. 视频异常检测技术研究进展[J]. 计算机科学与探索,2022,16(3):529-540.

[2] 吕承侃,沈 飞,张正涛,等. 图像异常检测研究现状综述[J]. 自动化学报,2022,48(6):1402-1428.

[3] HASAN M,CHOI J,NEUMANN J,et al. Learning temporal regularity in video sequences[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:733-742.

[4] GONG D,LIU L,LE V,et al. Memorizing normality to detect anomaly:memory-augmented deep autoencoder for unsupervised anomaly detection [C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul:IEEE,2019:1705-1714.

[5] PARK H,NOH J,HAM B. Learning memory-guided normality for anomaly detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle:IEEE,2020:14372-14381.

[6] LV H,CHEN C,CUI Z,et al. Learning normal dynamics in videos with meta prototype network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville:IEEE,2021:15425-15434.

[7] LIU W,LUO W,LIAN D,et al. Future frame prediction for anomaly detection - a new baseline[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City:IEEE,2018:6536-6545.

[8] LU Y,KUMAR K M,NABAVI S S,et al. Future frame prediction using convolutional VRNN for anomaly detection [C]//2019 16th IEEE International conference on advanced video and signal based surveillance (AVSS). Taiwan,China:IEEE,2019:1-8.

[9] 来 杰,王晓丹,向 前,等. 自编码器及其应用综述[J]. 通信学报,2021,42(9):218-230.

[10] RONNEBERGER O,FISCHER P,BROX T. U-net:convolutional networks for biomedical image segmentation[C]//Proc of 18th medical image computing and computer-assisted intervention. Munich:Springer,2015:234-241.

[11] 谢小红,李文韬,孙晓燕. 深度残差网络综述[J]. 信息与电脑,2021,33(16):85-87.

[12] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556,2014.

[13] WANG Z,BOVIK A C,SHEIKH H R,et al. Image quality assessment:from error visibility to structural similarity[J]. IEEE Transactions on Image Processing,2004,13(4):600-612.

[14] MAHADEVAN V,LI W,BHALODIA V,et al. Anomaly detection in crowded scenes[C]//2010 IEEE computer society conference on computer vision and pattern recognition. San Francisco:IEEE,2010:1975-1981.

[15] LU C,SHI J,JIA J. Abnormal event detection at 150 fps in matlab[C]//Proceedings of the IEEE international conference on computer vision. Sydney:IEEE,2013:2720-2727.

[16] ZHAO Y,DENG B,SHEN C,et al. Spatio-temporal autoencoder for video anomaly detection [C]//Proceedings of the25th ACM international conference on multimedia. New York:ACM,2017:1933-1941.

[17] LI N,CHANG F,HAN Y. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes[J]. IEEE Transactions on Multimedia,2021,23:203-215.

[18] SONG H,SUN C,WU X,et al. Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos[J]. IEEE Transactions on Multimedia,2020,22(8):2138-2148.

[19] ZHAO M,LIU Y,LIU J,et al. LGN-Net:local-global normality network for video anomaly detection[J]. arXiv:2211.07454,2022.