天津大学

《高级计算机视觉》课程报告

姓　　名　刘坚

学　　号　2020244252

2021 年　1 月　9 日

目录

摘　要

监控视频在社会民生方面都有广泛的应用。本文对监控视频异常检测概念、常用的或者经典的监控视频异常检测算法进行了概括总结。同时对 CVPR2020“MNAD”[1]的工作进行了研究以及复现，修改完善了这篇论文开放源码中存在的问题，尝试性地将 ECCV 2020 “CDDA”[2]中提出的使用 RGB 差异来模拟光流[3]的方法应用在原工作中，提高了异常检测的效果。最后对本工作未来的方向进行了探讨和展望。

Abstract

Surveillance video has a wide range of applications in all aspects of society and people's livelihood. In this paper, we summarize the concept of surveillance video anomaly detection, common or classical surveillance video anomaly detection algorithms. It also studies and reproduces the work of CVPR-2020 "MNAD", modifies and improves the problems in the open source code of this paper, and tries to apply the method proposed in ECCV-2020 "CDDA", which uses RGB difference to simulate optical flow, to the original work. We also try to apply the method of using RGB difference to simulate the optical flow proposed in ECCV-2020 "CDDA" to the original work to improve the anomaly detection effect. Finally, the future direction of this work is discussed and prospected.

一、引言

随着监控摄像机应用的普及，监控视频的数量越来越庞大。实现对监控视频的内容进行分析，检测监控视频中存在的异常现象，对于安防、智能家居、病人监护乃至自动驾驶等领域都有重要的意义。

视频中的异常检测问题类似于：检测出驶入人行道的汽车、检测出地铁口与人流反向的行人、检测出校园内漫步学生中奔跑的学生等。视频异常检测是一个非常具有挑战性的任务[4]，因为：①视频的正常样本和异常样本之间没有明确的边界[5]，同一个事件或者行为在不同的场景下可能被归类于不同的异常性[6]；②在数据样本中，异常事件往往是稀少、多样且不可穷举的；③训练样本存在着噪声；④数据隐私性。由于异常样本的稀少、不可穷举的特点，所以往往使用半监督、无监督的算法。比较经典的思路是采用仅包含正常样本的训练数据集训练一个正常模型，然后将测试集中不符合正常模型的样本判断为异常。因此，视频异常检测任务的一般步骤是：首先进行特征提取，之后进行模型训练，最后对异常进行判断。实验结果的评估标准通常使用接收器操作特性曲线 ROC 及其对应的曲线下面积 AUC。

二、相关工作

异常检测：异常点检测(Outlier detection)，又称为离群点检测，是找出与预期对象的行为差异较大的对象的一个检测过程。这些被检测出的对象被称为异常点或者离群点。异常点检测在生产生活中有着广泛应用，比如信用卡反欺诈、工业损毁检测、广告点击反作弊等。根据异常检测任务的特点，它常被划分为是一类无监督的学习问题。当异常检测的输入被当做点的形式时，它常使用以下三种方法来处理：①聚类判别[7]：根据特征空间的分布判断异常，将远离聚类中心的点、属于小聚类的点或分布概率密度低的点判断为异常；②重构判别[8]：用低维子空间/流形拟合正常样本特征空间的分布，通过将测试样本向正常 样本子空间/流形投影计算重构误差，进而根据重构误差的大小判断测试样本是否服从正常样本的分布并判断异常。③共发判别[9]：根据测试样本与正常样本共同出现的概率（共发概率）判断异常，将共发概率低的样本判断为异常。

记忆网络：有许多工作尝试捕捉时序数据中的长期依赖性。长短期记忆(LSTM)利用局部记忆单元解决了这个问题，网络的隐藏状态部分记录了过去的信息。然而，由于单元的大小通常较小，而且隐藏状态中的知识被压缩，因此记忆性能受到限制。为了克服这一局限性，最近有人提出了记忆网络[10]。它使用了一个可以读写的全局存储器，比传统的方法更好地执行记忆任务。然而，记忆网络需要层层监督来学习模型，因此很难使用标准的反向传播来训练它们。最近的工作使用连续的记忆表示或键值对来读/写记忆，允许端到端

地训练记忆网络。同时也有一些工作将记忆网络用于计算机视觉任务，包括视觉问题回答、One-shot 学习、图像生成[11]和视频摘要[12]。

MNAD：即“Learning Memory-guided Normality for Anomaly Detection”，是 CVPR2020 提出的一个工作。该项工作的核心是对视频异常检测任务重构判别方法的完善。作者认为由于现在的视频异常检测工作，在它的第一个阶段大多是采用 CNN 进行提取特征，CNN 强大的提取图像/视频表示的能力，使得传统的重构判别的方法，即使实在重构异常事件时也有比较好的能力，因为一个异常事件内容的组成部分往往包含有很多的正常样本的部分，或者说，异常帧里的特征都可以在正常样本中找到。这样就会导致重构判别的准确性大大下降。此外，作者认为传统的重构判别也没有考虑到正常样本的多样性，这样就会导致由正常样本训练出的正常模型受限于训练的数据量。对此，作者提出了两个解决方案，一是引入记忆网络（Memory Network），提出了新颖的特征压缩和分类损失来训练训练该网路，使得记忆项和深度学习的特征对正常数据有较好的分辨力，一定程度上弱化 CNN 强大特征提取的能力。二是对记忆模块使用一个新的更新策略，以此来记录正常样本的典型的模式。

CDDA：即“Clustering Driven Deep Autoencoder for Video Anomaly Detection”，是 ECCV2020 提出的一个工作，核心方法是结合了重构判别和聚类判别一起进行视频异常检测。作者认为异常事件通常可以从事件的外观表现或者运动变化检测到。外观表现大多

时间在时序上是静止的，运动变化是贯穿于一个视频段的，因此采用了双过程流的方法，一方面对事件的外观表现进行聚类判别，另一方面使用特定的方法捕获事件的运动信息，然后对运动的特征表示进行聚类得到聚类后的运动聚类中心特征。有了表示事件外观表现以及事件运动变化的特征，再对原始的输入片段进行重构判别，即先进行编码，再进行解码，以此来实现对异常事件的检测。作者在这项工作中的一个主要贡献就是，提出了使用 RGB 差异来模拟光流的效果，以捕获事件的运动信息。因为，传统的光流虽然可以捕获运动轨迹，但是计算量比较大难以训练。所以使用 RGB 差异可以有效的解决这一问题。

三、个人工作

个人工作主要是在 MNAD 工作的基础上，复现了 MNAD 原有的工作，并且将 CDDA 中提出的以 RGB 差异来实现光流效果的思想实现，将其添加到了 MNAD 编码的环节。

MNAD 提出的网络结构以及工作的流程如下图(1)所示：

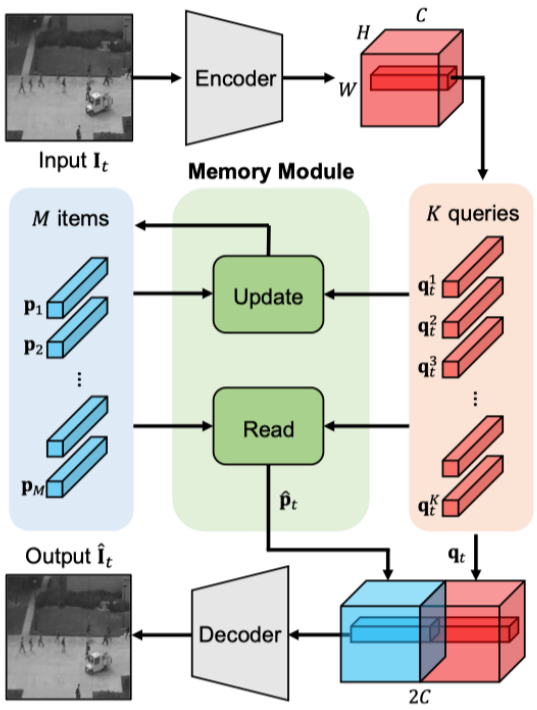

图 1

可以看到，MNAD 是以每个视频帧为单位，直接送入卷积层提取特征。之后使用提取的特征对记忆项进行更新，记录正常样本的典型特征。本文的工作没有改变记忆模块和后续的检测模块。使用 CDDA 中提出的基于方差的注意力模块，这里之所以使用注意力模块，是因为作者认为在视频样本中，监控视频中的绝大部分都是静态的，异常样本更可能有大幅度的动作变化，于是作者提出了使用一个基于方差的注意力模块来动态的分配视频片段运动部分的权重。本文将基于方差的注意力模块以及使用 RGB 差异获取运动特征的模块添加到 MNAD 编码过程中，建立了外观表现特征结合运动表现的特征的记忆项。下图(2)为原 CDDA 中使用 RGB difference 获取运动特征的结构，本文去除了对运动特征聚类的过程，使用获得的多个运动特征的均值进行表示。

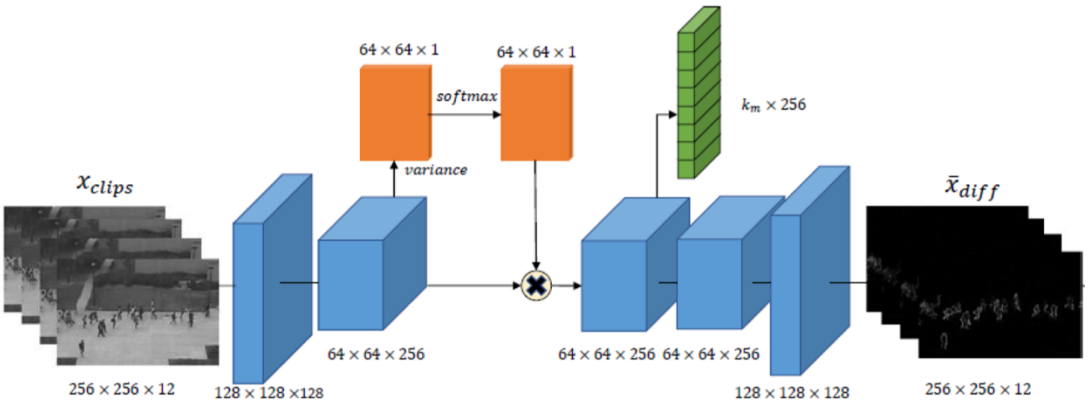

图 2

具体来讲，以E_m、D_m分别表示运动的编码器和解码器，x_{clips}表示连续的视频帧，z_m表示动作表示，x_{diff}表示输入的连续帧以及最后一帧的 RGB 差异，x_a为输入的连续视频帧的第一帧，也就是说x_{diff}=x_{clips}−x_a，则运动自编码器的公式表达为：z_m=E_m(x_{clips};θ_e^m)$\bar{x}_{diff}$=D_{m}(z_m;θ_d^m)，其中$\bar{x}_{diff}$即为重构的 RGB 差异。通过最小化重构误差得到上图中绿色的表示动作的特征。将其与原 MNAD 方法得到的外观表示结合在一起得到新的输入的表示来更新记忆项。

因为对原结构进行了修改，所以采用了新的损失函数如下：

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_c\mathcal{L}_{compact} + \lambda_s\mathcal{L}_{separate} + \mathcal{L}_m(x_{diff}, \bar{x}_{diff})$$

即在原 MNAD 损失函数的基础上添加了动作自编码的重构损失。实验结果如下：（原方法和修改后的方法在不同数据集上的 AUC）

Algorithm	UCSD Ped2	Avenue	ShanghaiTech
MNAD	93.82%	84.27%	70.11%
Our method	92.77%	84.43%	\

四、总结

本文对视频异常检测的任务描述、挑战问题、常用方法、工作流程以及评估标准做了介绍，并且在相关工作部分对本工作基于的两个工作进行了分析，最后通过结合两个工作的优点完成了视频异常检测的任务。实验表明，通过 RGB 差异模拟光流的效果，捕获异常事件的动作变化确实可以起到很好的效果，由于本工作复现时并没有达到原文提及的最有效果，因此在修改之后方法性能提升也不是很明显。此外，本工作在将动作变化的特征和外观表现的特征结合时，是采用简单的对应的 pixel 相加，是否合理还需要进一步的实验。

代码：https://github.com/XiaoXueHou/Anomaly-Detection

五、参考文献

[1] Park H , Noh J , Ham B . Learning Memory-guided Normality for Anomaly Detection[J]. 2020.

[2] Chang Y , Tu Z , Xie W , et al. Clustering Driven Deep Autoencoder for Video Anomaly Detection[M]. Springer, Cham, 2020.

[3] Hetherington R . The perception of the visual world[M]. Houghton Mifflin, 1950.

[4] 王志国, 章毓晋. 监控视频异常检测:综述[J]. 清华大学学报(自然科学版), 2020, v.60(06):73-84.

[5] Chandola V , Banerjee A , Kumar V . Anomaly Detection: A Survey[J]. ACM Computing Surveys, 2009, 41(3).

[6] Chong Y S , Tay Y H . Modeling Representation of Videos for Anomaly Detection using Deep Learning: A Review[J]. Computer Science, 2015.

[7] 黄鑫, 肖世德, 宋波. 监控视频中的车辆异常行为检测[J]. 计算机系统应用, 2018.

[8] B X Z A , B J L , B J W , et al. Sparse representation for robust abnormality detection in crowded scenes[J]. Pattern Recognition, 2014, 47(5):1791-1799.

[9] Hu D H , Zhang X X , Yin J , et al. Abnormal Activity Recognition Based on HDP-HMM Models.[C]// International Jont Conference on Artifical Intelligence. Morgan Kaufmann Publishers Inc. 2009.

[10] Weston J , Chopra S , Bordes A . Memory Networks[J]. Eprint Arxiv, 2014.

[11] Zhu M , Pan P , Chen W , et al. DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis[J]. 2019.

[12] Lee S , Sung J , Yu Y , et al. A Memory Network Approach for Story-based Temporal Summarization of 360{\deg Videos[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.