感性理解视频异常检测

肖剑

前言

本文非正式地讲解视频异常检测（Video Anomaly Detection）任务，有些地方缺乏严谨性、逻辑性和关联性。为了理解上的充分性，文中穿插了一定量的“注”，有些地方高亮显示。

注：如今，深度学习技术用于解决视频异常检测取得了显著的成效，所以，本文主要介绍两种非常流行的方法，帧重构法和帧预测法。因为用到深度学习，所以有些最基本的概念一定要知道（学起来很简单）：

1. 有监督学习，无监督学习
2. 机器如何做图片分类？比如最简单的二分类。需要大量有标注的图片数据
3. 理解训练集、测试集[、验证集]。注：有些任务中没有验证集，比如视频异常检测一般就没有验证集。

对上述三点有简单理解即可，不要花太多时间。然后，就可以往下看了。

一、视频异常检测是什么？

从宏观角度介绍视频异常检测

视频异常检测是监控视频中的一项关键任务，具有重要的现实研究意义和应用价值。视频异常检测（Video Anomaly Detection）任务，是指模型（需要设计和训练）自动发现视频中存在的异常行为或异常实体。换句话说，给定一段视频作为模型的输入，要求模型可以判别出视频中的哪一部分有异常发生。

举个**例子**，给定一段视频，其时长为 10 秒，帧率为 30（所以该视频共有 10*30=300 帧），视频画面宽为 360 像素，高为 240 像素。假设是人行道场景，有一辆汽车从视频画面中驶过，其在时间维度上出现在 0.5 秒~3.5 秒，在空间维度上出现在**（因为汽车在移动，所以**每时每刻也在变化）。因为是人行道场景，所以汽车的出现就是异常事件。视频异常检测的任务，就是对于这个视频，要求模型可以判断出异常在时间维度上发生在第 15 帧~105 帧，在空间维度上发生在**。注：（不严谨地讲）有的很多方法仅仅做到判别帧级异常（时间维度），有的方法可以不仅可以判别帧级异常，还可以实现异常定位（空间维度，判断视频帧中那些像素位置发生异常）。

既然是检测异常，那么就要说明什么是异常。那么，异常有确切的定义吗？没有，我们没有办法给异常下确切的定义。例如，汽车行驶在高速公路上是一个正常事件，而行驶在人行道上是一个异常事件。由此可见，异常具有环境依赖性，异常事件的定义因环境的不同而发生变化。正常与异常是一个相对的概念，如果我们可以定义准确完备地定义正常（穷尽列举所有的正常事件），那么我们就相当于定义了异常，但不幸，我们无法准确完备地定义正常，我们只能定义部分正常。因此，异常事件通常被定义为不熟悉或在给定环境下不被期望出现的事件。简单理解，可见图 1。

图 1 三个被广泛使用的公开数据集中的正常与异常样例示意图。（注：最难的数据集为 Street Scene）第一行为正常事件，第二行为异常事件。可以看到，这三个数据集，正常事件基本上都是人在正常走路（Street Scene 不是）。Ped2、Avenue、ShanghaiTech 在图中的异常事件分别为：人行道上骑自行车、滑滑板；跑步；推婴儿车。

与正常事件相比，异常事件的发生频率更低，类型更多，甚至有些异常事件我们还没有见过，这导致异常事件数据难以收集，所以，我们无法收集大量的异常数据，继而利用传统的有监督学习范式予以解决视频异常检测问题。注：视频异常检测可以看做广义的二分类问题，一类是正常，一类是异常，所以，类比图片二分类，如果我们可以收集大量的有标注数据，那么我们就可以使用解决图片二分类的方法来解决视频异常检测问题。但不幸，我们无法收集大量的有标注数据，所以无法使用传统的有监督学习范式来训练模型。

但，考虑到监控视频中存在着大量的正常数据，所以视频异常检测有这样一种问题设置——半监督视频异常检测：训练集中仅包含正常样本数据；测试集中既有正常数据，也有异常数据。因此，许多文献将其视作单分类问题，训练阶段利用模型直接对正常样本进行建模，学习得到关于正常样本的分布表示，测试阶段将远离正常分布的输入数据判别为异常。示意图如图 2 所示。

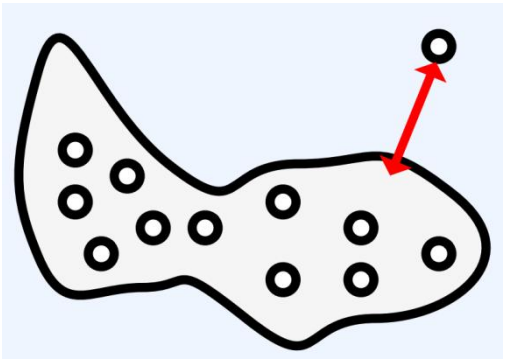

图 2 曲线所包裹的点为正常数据，曲线外的右上角的点为异常数据。在半监督问题设置下，训练阶段：训练集中仅有正常数据，模型利用这些数据学习得到一个分布表示（曲线所围住的区域）；测试阶段：给定一个样本点，看它是否在分布内，在分布内，那就判别它为正常，否则判别为异常（这个点也可以叫做离群点 outliers）。注：这是基本思想，具体工程实践（代码）中，不同方法会有差别。后面会详述。

在半监督视频异常检测中，有两类非常流行的方法：帧重构法（Frame-Reconstruction）和帧预测法（Future Frame-Prediction）。本文详述这两种方法，见第三节。

二、视频异常检测分类

从微观角度介绍视频异常检测：标注数据

视频异常检测从监督类型的角度可分为 4 种：1）有监督视频异常检测，训练集中提供帧级别的标注；2）半监督视频异常检测，训练集中提供正常样本数据；3）弱监督视频异常检测，训练集中提供视频级别的标注；4）无监督视频异常检测，模型学习过程中不提供任何有标签数据。如图 3 所示。

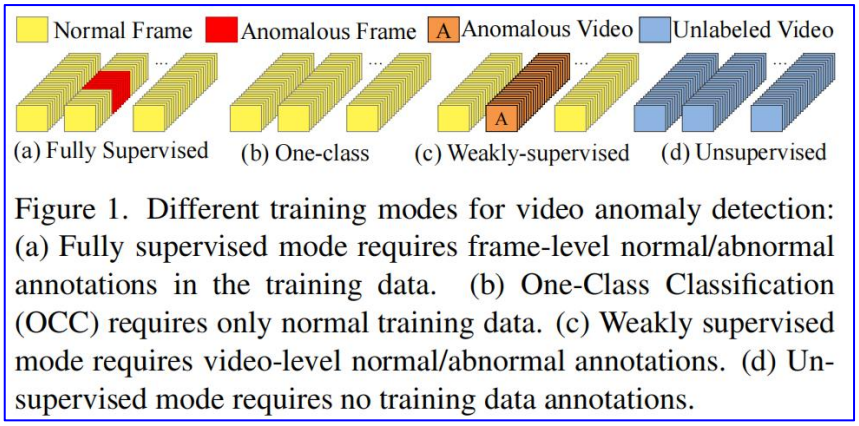

Figure 1. Different training modes for video anomaly detection: (a) Fully supervised mode requires frame-level normal/abnormal annotations in the training data. (b) One-Class Classification (OCC) requires only normal training data. (c) Weakly supervised mode requires video-level normal/abnormal annotations. (d) Unsupervised mode requires no training data annotations.

图 3 视频异常检测的四种问题设置。a）有监督；b）半监督<单类别>；c）弱监督；d）无监督；注 1，无监督中还可以细分，此处略去。注 2，这只是其中一种分类方法（从监督信息的角度），不同角度有不同分类。

接下来详述每类问题设置（训练集的标注有差别）：

a. 有监督。训练集中既有正常数据，也有异常数据。标注信息精确到帧，对每一个视频帧，若其中有异常发生，那么为该视频帧标记为 1，否则标记为 0。即，正常为 0，异常为 1。
b. 半监督。训练集中仅包含正常数据。所以无需像有监督设置那样标注帧，知道训练集中的都为正常即可。
c. 弱监督。训练集中既有正常数据，也有异常数据。标注信息精确到视频，对每一个视频，若其中有异常发生，那么为该视频标记为 1，否则标记为 0。即，对于一个标记为 1 的异常视频，我们仅仅知道其中有异常发生，但是对于异常所发生的位置（时空维度）我们一无所知。
d. 无监督。顾名思义，就是没有标注信息，我们不知道哪些是正常数据，哪些是异常数据。这里面还可以细分，但对于入门，可以略去。

注：以上所述，均为帧级别标注，这是现有的大多数方法在训练模型时所使用的标注；但应当指出的是，也有少量方法使用的是像素级别的标注。公开数据集中，有的只提供了视频级标注，有的只提供了帧级别标注，有的不仅有帧级别标注也有像素级标注。（有了像素级标注，当然就有了帧级别标注）

截止 2022 年，在顶会订刊中，有监督的文献绝对不超过 5 篇，半监督的文献最多，弱监督的文献次之（2018 年提出弱监督视频异常检测），无监督的文献有 6 篇（最早 2016 年）。

简要分析：为什么有监督的文献那么少呢？因为有监督的问题设置，不符合现实情况，我们根本无法收集大量的有标签的异常数据。为什么无监督的也不多呢？因为这个相对来说比较难，6 篇情况是，2016 年、2017 年、2018 年、2020 年，每年 1 篇，2022 年有 2 篇。而半监督和有监督，它们每年都有好多篇。

三、两种视频异常检测方法

两种流行且简单的半监督视频异常检测方法：帧重构法与帧预测法。

神经网络模型选择简单的自编码器（Auto-Encoder），当然也可以使用其它的模型，如生成对抗网络（GAN），变分自编码器（VAE）等等。

自编码器的相关知识，此处略去；简单提及，就是给定一张图片，喂进模型，模型输出一张图片，要求输出的这样图片尽量与输入的图片相似。不严谨地讲，这个输出的图片，也可以叫做重构出来的图片，重构图片。

帧重构方法与帧预测方法，它们假设正常数据可以被很好的重构或预测，而异常数据则不能。帧重构方法学习一个模型来重构正常的训练数据，并利用重构误差来识别异常。帧预测方法可看作是一种特殊的帧重构方法，它学习一个以一系列连续帧作为输入的模型，并预测下一帧，利用预测帧与真实帧的差异程度来区分异常。重构法与预测法的原理示意图如图 4 所示。

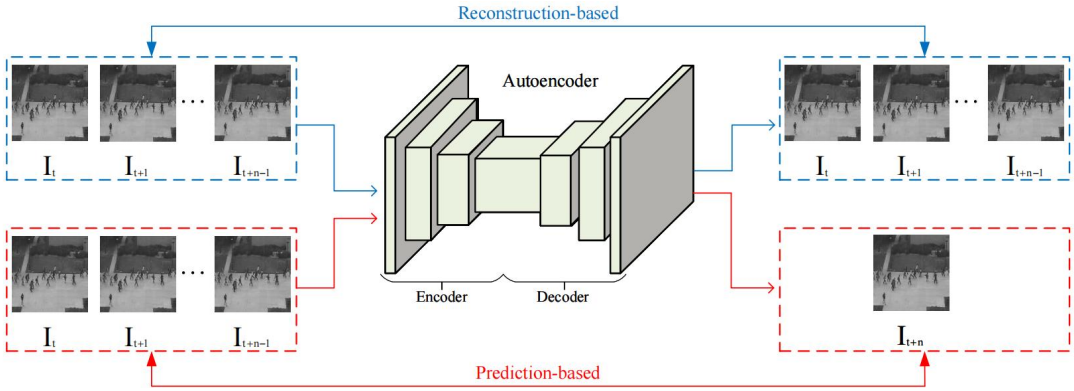

图 4 重构法与预测法的原理示意图

图 4 中，上面为重构法的示意图，下面为预测法的示意图。大白话说，在训练阶段要求模型只认识正常数据，而在测试阶段，如果模型认识，那么就将其视作正常，如果模型不认识，那么就把它视作异常。下面举例说明：1）于重构法，假设输入数据的时间维度为 5，那么连续输入 5 帧 ABCDE，重构也为连续的 5 帧 A′B′C′D′E′，如果 ABCDE 与 A′B′C′D′E′非常相似，那么就说明模型认识输入 ABCDE，否则，如果 ABCDE 与 A′B′C′D′E′差别很大，那么就说明模型不认识输入 ABCDE，那么就把 ABCDE 判别为异常。2）与预测法，它可以看做是一种特殊的重构。假设输入数据的时间维度为 4，那么连续输入 4 帧 ABCD，重构为第 5 帧 E′，如果重构帧 E′与真实的帧 E 相似，那么就说明模型认识输入 ABCD，将其判别为正常；否则，如果重构帧 E′与真实的帧 E 不相似，那么就说明模型不认识输入 ABCD，将其判别为异常。

对这两类方法的简单评价：

1）帧重构法可以很好地捕获表观信息，帧预测法可以有效地捕获运动信息

2）两种方法都是要重构出图片，或者说生成图片，所以都属于生成模型。（emmmm，说法不严谨，因为自编码器本身就属于生成模型）。重点是，它们都属于像素级别的生成，极其耗费计算资源，这或许并不必要

3）由于生成模型强大的泛化能力，导致模型对异常数据也可以生成的很好，这就导致模型性能的下降。怎么理解呢？以帧重构为例，训练时，我们要求模型喂进去什么，吐出来什么，那么如果模型泛化能力很强，它就是真的喂进去什么吐出来什么。训练时，喂进去正常数据吐出正常数据，但如果模型泛化能力太强，就真的是喂进去什么吐出来什么，测试时，喂进去异常，也可以吐出异常。但，这不是我们所希望的，我们希望的是，模型可以很好的重构正常，而不能够很好的重构异常。