Google 通过视频着色进行自监督跟踪

原创：Google谷歌开发者7 月 11 日
文 / 机器感知研究员 Carl Vondrick

跟踪视频中的物体是计算机视觉领域的一个基本问题，对活动识别、物体交互或视频风格化等应用而言至关重要。不过，教会机器对物体进行视觉跟踪十分困难，这在一定程度上是因为此过程需要使用大量带标记的跟踪数据集进行训练，而大规模地标注在实际中并不可行。

在 “Tracking Emerges by Colorizing Videos” 一文中，我们介绍了一种卷积网络，这种网络可以对灰度视频着色，但被限定为仅从一个参考帧复制颜色。通过这种方式，网络可在没有监督的情况下自动学习对物体进行视觉跟踪。重要的是，尽管模型从未明确进行过跟踪训练，它仍然可以跟踪多个物体，跟踪被遮挡的物体并在物体发生变形时保持稳定，而不需要任何带标记的训练数据。

基于公开学术数据集 DAVIS 2017 的跟踪预测示例
在学习对视频着色之后
在没有监督的情况下出现了一种自动跟踪机制

我们在第一帧中指定感兴趣的区域（用不同颜色表示）
模型无需任何额外学习或监督即自动跟踪

学习对视频重新着色

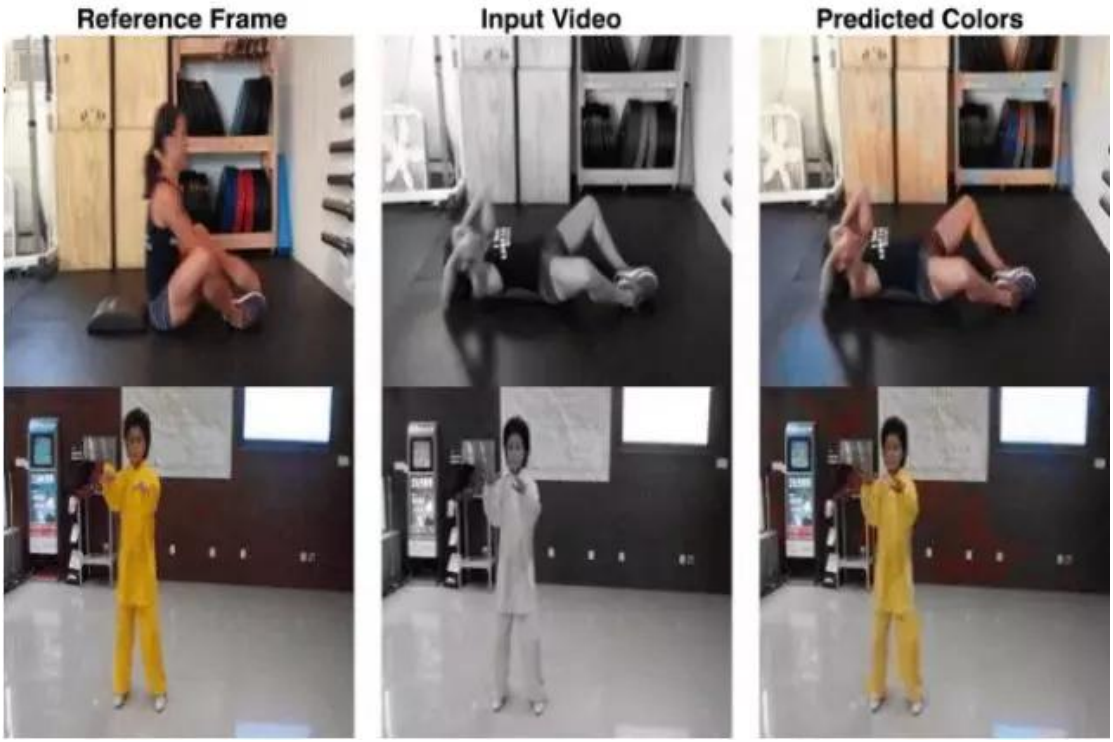
我们假设颜色的时间一致性为教机器跟踪视频中的区域提供了大规模的优秀训练数据。显然，总有一些例外情况，即颜色不具备时间一致性（如突然开灯），但一般而言，颜色不会随着时间而变化。并且，大部分视频都包含颜色，这就提供了可扩展的自监督学习信号。我们先去掉视频的颜色，然后再添加着色步骤，因为视频中可能有多个物体颜色相同，而通过着色我们可以教机器跟踪特定的物体或区域。

为了训练系统，我们使用了大型公开数据集 Kinetics 中的视频，此数据集汇总了大量描述日常活动的视频。我们将除了第一帧以外的所有视频帧都转换为灰度图像，并训练一个卷积网络来预测后续帧中的原始颜色。我们期望模型学会跟踪区域，以准确恢复原始颜色。我们的主要观察结果是，跟踪物体着色这一需求使得自动学习物体跟踪模型成为可能。

我们使用 DAVIS 2017 数据集中的视频
来展示视频重新着色任务模型
接收一个彩色帧和一个灰度视频作为输入

然后预测视频其他帧的颜色
它学习从参考帧中复制颜色
这使得无需人工监督即可学习跟踪机制

学习复制单个参考帧的颜色要求模型学会内在地指向正确的区域以复制正确的颜色。这迫使模型学习一种可用于跟踪的明确机制。为了展示视频着色模型的工作原理，我们在下面显示了一些对 Kinetics 数据集中的视频进行着色预测的示例。

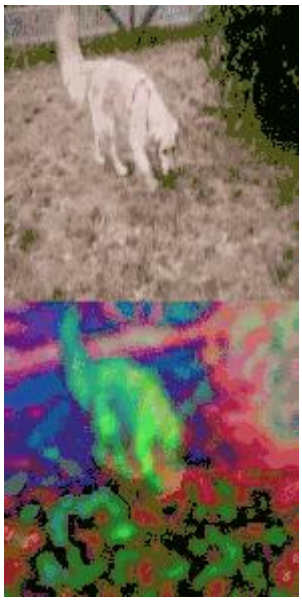


使用公开数据集 Kinetics 将着色参考帧
应用到输入视频后的预测颜色示例

尽管网络未使用真实标识进行训练，我们的模型还是能学会跟踪视频第一帧中指定的任何视觉区域。我们可以跟踪视频中的物体轮廓或单个点。唯一做出的改变是在视频中传播表示感兴趣区域的标签，而不是传播颜色。

分析跟踪器

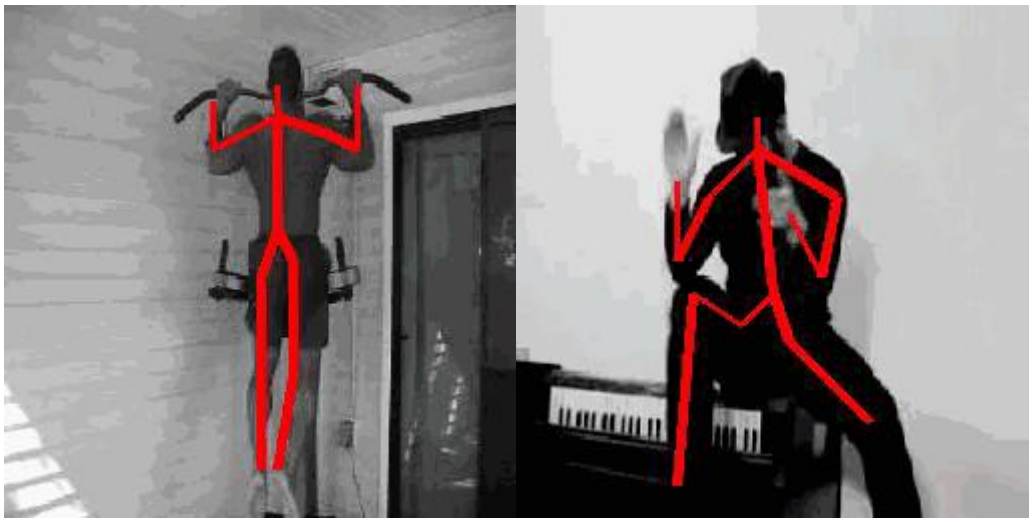
由于模型是基于大量未标记的视频进行训练的，因此我们希望深入了解它学习的内容。下面的视频展示了一个标准的跟踪过程：通过主成分分析 (PCA) 将模型学到的嵌入投影到三维空间进行可视化，并做成 RGB 影片的形式。结果表明，学到的嵌入空间的最近邻往往会对应物体标识，即使经过变形或视角改变也是如此。



上面一行：DAVIS 2017 数据集中的视频
下面一行：对着色模型的内部嵌入进行可视化
在这个可视化中，相似的嵌入具有相似的颜色
这表明学到的嵌入按物体标识将像素分组

跟踪姿态

我们发现，在给定初始帧关键点的条件下，模型还可以跟踪人类姿态。下面所示为基于公开学术数据集 JHMDB 的结果，其中模型跟踪的是人类关节骨架。



使用模型跟踪人类骨架运动的示例
在本例中，第一帧的输入是人类姿态，
后续运动由模型自动跟踪
即使模型从未明确进行过此项任务的训练
它依然能够跟踪人类姿态

虽然着色模型并没有超越强监督模型，但它可以学习跟踪视频分割和人类姿态，且超越了基于光流的最新方法。按运动类型细分性能的结果表明，我们的跟踪器在许多自然复杂场景（例如动态背景、快速运动和遮挡）下比光流方法更加强大。

未来工作

我们的研究结果表明，视频着色提供的信号可以用于学习跟踪视频中的物体，且无需监督。此外，我们发现系统中出现的失败与视频着色失败有关，这表明进一步优化视频着色模型可以改善自监督跟踪。