Google 发布 AVA：一个用于理解人类动作的精细标记视频数据集

原创：Google谷歌开发者2017-11-03

文 / Google 软件工程师 Chunhui Gu 和 David Ross

教机器理解视频中的人类动作是计算机视觉的一个基本研究课题，对于个人视频搜索和发现、运动分析和手势接口等应用必不可少。过去几年来，在图像中分类和查找对象取得了令人兴奋的突破，但识别人类动作仍然是一个巨大的挑战。原因在于，就其本性而言，人类动作的定义不如视频对象完善，因此，很难构建精细标记的动作视频数据集。尽管有许多基准数据集（如 UCF101、ActivityNet 和 DeepMind 的 Kinetics）采用图像分类标记模式，并为数据集中的每个视频或视频剪辑分配一个标签，但对于有多人执行不同动作的复杂场景，还没有相应的数据集。

为促进对人类动作识别的进一步研究，我们发布了 AVA，它诞生于“原子视觉动作”，是一个全新的数据集，为扩展视频序列中的每个人提供多个动作标签。AVA 由 YouTube 中公开视频的网址组成，注解了一组 80 种时空局部化的原子动作（如“走”、“踢（物体）”、“握手”等），产生了 5.76 万个视频片段、9.6 万个标记动作执行人以及总共 21 万个动作标签。

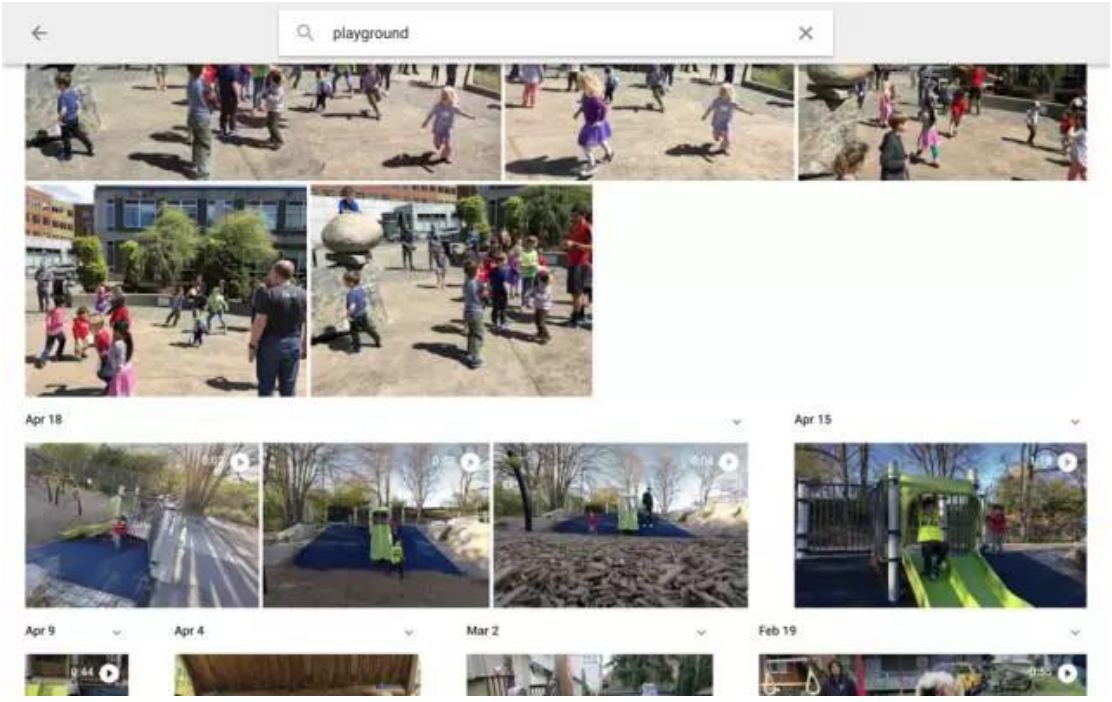
您可以浏览网站，了解数据集和下载注解：

https://research.google.com/ava

并阅读我们的 arXiv 论文，了解数据集的设计和开发：
https://arxiv.org/abs/1705.08421

与其他动作数据集相比，AVA 具有以下重要特征：

- 以人为中心的注解。每个动作标签与人相关，而不是与视频或剪辑相关。因此，我们可以将不同标签分配到同一场景中执行不同动作的多个人（这种情况很常见）。
- 原子视觉动作。我们将动作标签限于很小的时间尺度（3 秒），在此范围内，动作的性质是身体活动，具有清晰的视觉特征。
- 现实视频材料。我们使用电影作为 AVA 的来源，从很多不同的流派和原产国取材。因此，数据中包含广泛的人类行为。

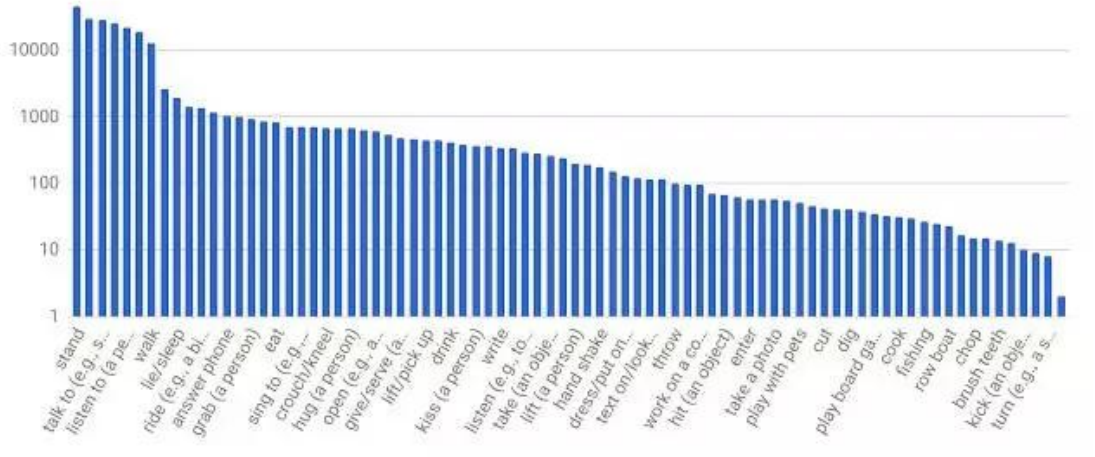


▲ 3 秒视频片段（来自视频来源）示例，其边界框注解在每个片段的中间帧中。（为清楚起见，每个示例只显示一个边界框）

为创建 AVA，我们先从 YouTube 收集了一组变化多的长形式内容，集中于“电影”和“电视”类别，有许多不同国籍的专业演员。我们对每个视频分析了 15

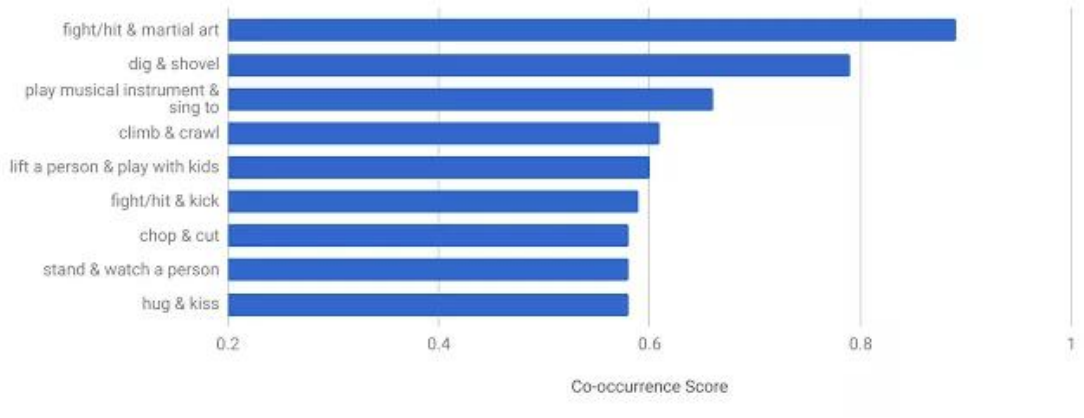
分钟的片段，将其统一分隔为 300 个不重叠的 3 秒片段。采样策略将动作序列保持在连贯的时间背景中。

然后，我们手动标识每个 3 秒片段中间帧中所有人的边界框。对于边界框中的每个人，注解人员从预定义的原子动作词汇（有 80 个类别）中选择不同数量的标签来描述个人在片段中的动作。这些动作分为三组：姿势/移动、人-物体互动以及人-人互动。因为我们详尽标记了执行全部动作的所有人，所以，AVA 标签的频率遵循长尾分布，下面进行了简要介绍。



▲ AVA 原子动作标签的分布。X 轴中显示的标签只是我们词汇的一部分。

AVA 的独特设计可让我们推导出其他现有数据集中没有的一些有趣统计信息。例如，如果很多人至少有两个标签，我们便可测量动作标签的共现模式。下图所示为 AVA 中最常见的共现动作对及其共现得分。我们确认预期模式，比如人们经常在唱歌时弹奏乐器，跟小孩玩时把人举起，以及在亲吻时拥抱，等等。



▲ AVA 中最常见的共现动作对。

为评估 AVA 数据集中人类动作识别系统的有效性，我们实现了现有基线深度学习模型，该模型可以从更小的 JHMDB 数据集获得更好的性能。由于缩放、背景杂波、摄影和外观变化等富有挑战性的变化，此模型在正确识别 AVA 中的动作时表现一般 (18.4% mAP)。这表明 AVA 是一个有用的试验台，可用于为未来几年开发和评估新的动作识别架构和算法。

我们希望，AVA 能帮助改进人类动作识别系统的开发，能基于精细时空粒度的标签在个人动作层级为复杂活动建模。我们将继续扩展和改进 AVA，也渴望听到社区的反馈意见，帮助我们指引未来的方向。

致谢

AVA 的核心团队包括 Chunhui Gu、Chen Sun、David Ross、Caroline Pantofaru、Yeqing Li、Sudheendra Vijayanarasimhan、George Toderici、Susanna Ricco、Rahul Sukthankar、Cordelia Schmid 和 Jitendra Malik。感谢许多 Google 同事和注解人员对此项目的全力支持。