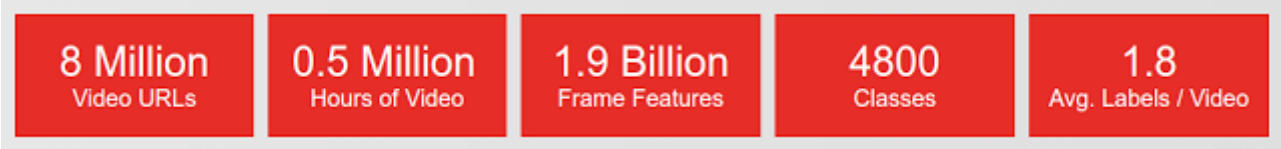
谷歌发布YouTube-8M：有史以来最大最多样化的视频数据集

By 吴攀2016年9月29日 13:00

机器学习和机器感知领域近来的许多突破都可归功于大型有标注数据集的可用性，例如 ImageNet，其包含了分成了数千个类别的数百万张有标签的图像。它们的可用性显著加速了图像理解领域的研究，例如检测和分类静态图像中的物体。

视频分析能为检测和识别物体、理解人类行为和与世界的交互提供更多的信息。改善视频理解能带来更好的视频搜索和发现，这类似于图像理解帮助重新想象照片中的经历的方式。但是，这一领域进一步发展的一个关键瓶颈是缺乏与图像数据集同等规模和多样性的真实世界视频数据集。

今天，我们很高兴宣布发布 YouTube-8M——该数据集包含了 800 万个 YouTube 视频 URL（代表着 500,000 小时的视频）以及它们的视频层面的标签（video-level labels），这些标签来自一个多样化的包含 4800 个知识图谱实体（Knowledge Graph entity）的集合。相比于之前已有的视频数据集，这个数据集的规模和多样性都实现了显著的增长。比如说，我们所知的之前最大的视频数据集 Sports-1M 包含了大约 100 万段 YouTube 视频和 500 个体育领域的分类——YouTube-8M 在视频数量和分类数量上都差不多比它高一个数量级。



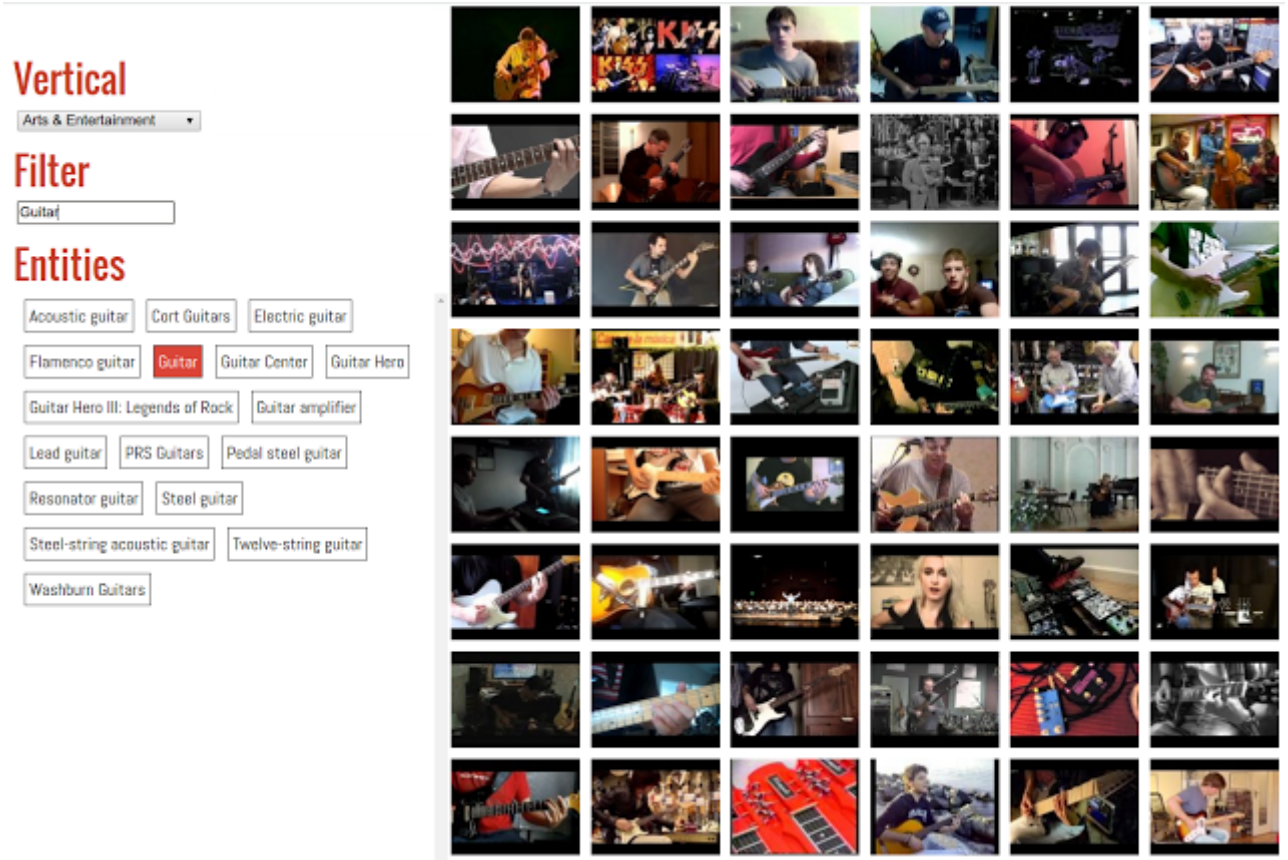
为了构建一个这样大规模的有标签视频数据集，我们需要解决两个关键难题：

1. 用人工标注的话，视频标注比图像标注所需的时间远远更多；
2. 视频的处理和存储的计算成本非常高。

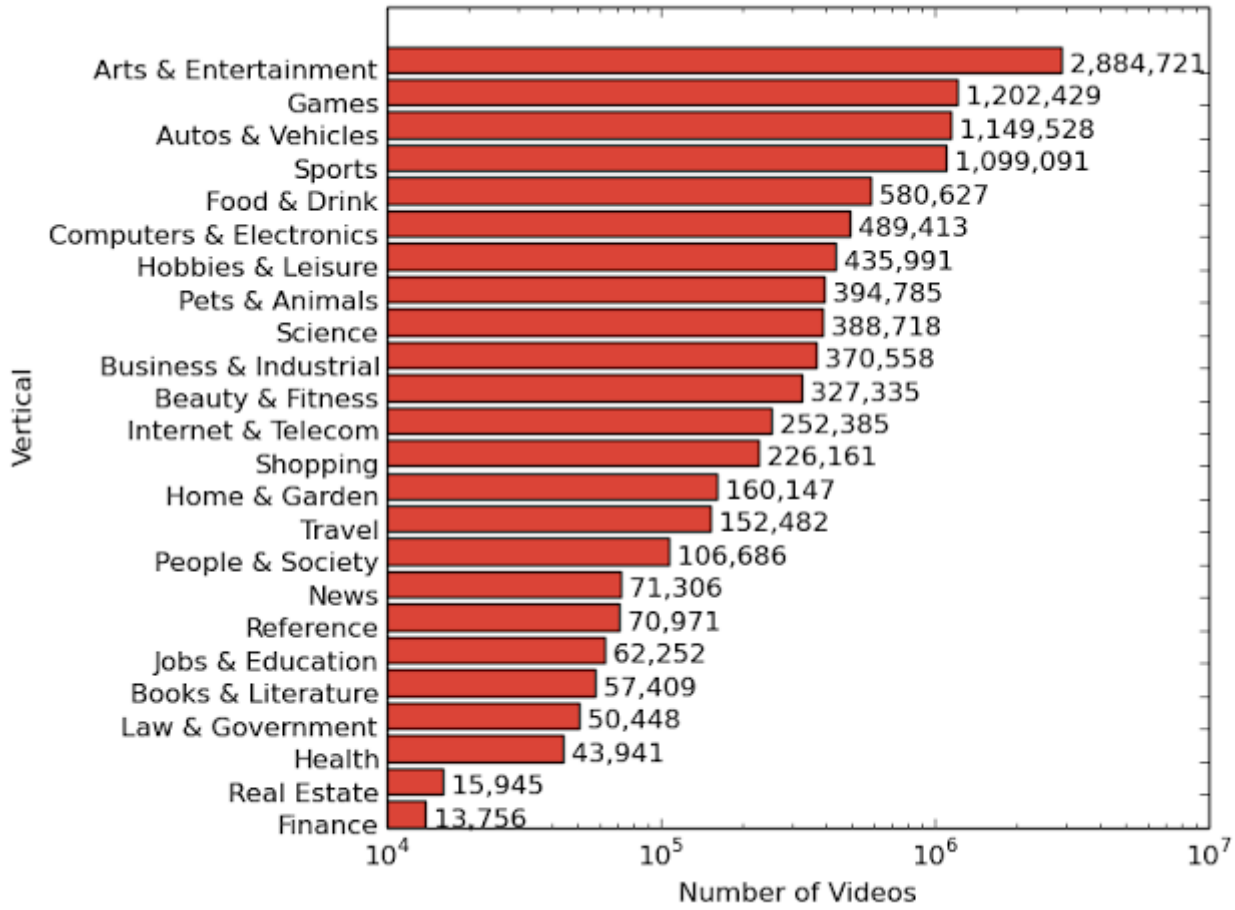
为了克服难题 1，我们使用了 YouTube 及其视频标注系统（video annotation system）。该系统可为所有公开的 YouTube 视频确定相关的知识图谱主题。尽管这些标注是机器生成的，但它们整合了来自数百万用户的强大的用户参与信号（user engagement signals）以及视频元数据和内容分析。由

此，这些标注的质量是足够高的，可用于视频理解研究和制定标准的目的。

为了确保该有标注视频数据集的稳定性和质量，我们仅使用了包含超过 1000 条评论的公开视频，而且我们创建了一个多样化的实体词汇集（vocabulary of entities）——这些实体都是视觉上可见的，而且出现的频率也足够高。该词汇集的创建结合了频率分析、自动过滤、人类评估者验证该实体是视觉上可见的、以及分组成 24 个顶层的垂直类别（更多详情参见我们的技术报告）。下图描述了其数据浏览器和在顶层垂直类别的视频分布，同时也说明了该数据集的规模和多样性。



数据浏览器允许浏览和搜索整个知识图谱实体词汇集，它们被分成了包含了对应视频的 24 个顶层的垂直类别。这张截图描述了一个标注了实体「Guitar」的数据集视频的子集。



该顶层垂直类别的视频分布说明了该数据集的规模和多样性，同时也反映了流行的 YouTube 视频的自然分布。

为了解决难题 2，我们必须克服研究者处理这些视频时所面临的存储和计算资源的瓶颈。在 YouTube-8M 的规模上进行视频理解通常应该需要 PB 级存储以及相当于 CPU 工作几十年的处理能力，为了让计算资源有限的研究者和学生也能用上这个数据集，我们对视频进行了预处理并使用了在 ImageNet 上训练的公开可用的 Inception-V3 图像标注模型（一种目前最佳的深度学习模型）提取出了帧层面的特征（frame-level features）。这些特征是按每秒 1 帧的时间分辨率从 19 亿个视频帧中提取的，然后它们还被进一步压缩到了可装入单个商品级硬盘的大小（少于 1.5 TB）。这使得我们可以在单个 GPU 上只用低于一天的时间就能全规模地下载该数据集并完成基准的 TensorFlow 模型的训练。

我们相信该数据集能极大地加速在视频理解上的研究，因为它能让研究者和学生无需使用大数据和大机器就能进行之前前所未有的规模的研究。我们希望这个数据集将能激励在视频建模架构和表征学习上的激动人心的新研究，尤其是能有效处理噪声或不完整标签的方法、迁移学习（transfer

learning）和领域适应（domain adaptation）方面的研究。事实上，我们的实验表明：在该数据集上对模型进行预训练并在其它外部数据集上进行应用/微调，可以在这些外部数据集（如 ActivityNet、Sports-1M）上实现当前最佳的表现。关于我们使用该数据集进行的所有实验以及我们构建它的更多细节，请参阅我们的技术报告论文。

下面是对该技术报告论文的摘要翻译。

- **论文：YouTube-8M：一个大型视频分类基准（YouTube-8M: A Large-Scale Video Classification Benchmark）**

摘要：计算机视觉领域近来的许多进步都可归功于大型数据集。机器学习的开源软件包和不再昂贵的商品级硬件大幅降低了探索新方法的进入壁垒。我们可以在几天时间内就在数百万个样本上完成模型的训练。但是大型数据集（如 ImageNet）都是为图像理解存在的，还没有规模能与之媲美的视频分类数据集。

在这篇论文中，我们介绍 YouTube-8M——这是目前最大的多标签视频分类数据集，包含了约 800 万段视频（约 50 万小时），这些视频用一个包含了 4800 个视觉实体（visual entity）的词汇集进行了标注。为了获取这些视频和它们的（多）标签，我们使用了一个 YouTube 视频标注系统（video annotation system），该系统可以给其中的视频标注上主要的主题。尽管这些标签是机器生成的，但它们的准确度非常高并且是衍生自各种基于人类的信号，其中包括元数据（metadata）、查询点击信号，所以可以说它们是基于内容的标注方法的一个非常好的目标。我们使用了自动和人工兼用的调制（curation）策略对视频标签（知识图谱实体）进行了过滤，其中包括询问人类评估者标签是否可以通过视觉识别。然后我们以每秒一帧的速度对每个视频进行了解码，然后使用了一个在 ImageNet 上预训练过的 Deep CNN 来提取刚好在分类层之前的隐藏表征（hidden representation）。最后，我们对帧特征（frame features）进行了压缩，使帧层面和视频层面的标签都可供下载。该数据集包含了超过 19 亿个视频帧和 800 万段视频的帧层面的特征，所以它是最大的公开的多标签视频数据集。

我们在该数据集上训练了多种（中等的）分类模型，并使用了流行的评估标准对它们进行了评估，然后将它们报告作为了基准。尽管这个数据集很大，但我们的一些使用了公开公用的 TensorFlow 框架的模型在单台机器上只用不到一天的时间就训练到了收敛（convergence）的程度。我们计划发布用于训练基本 TensorFlow 模型和用于计算标准的代码。

我们的实验表明：在大型数据集上的预训练可以泛化到其它数据集上，比如 Sports-1M 和 ActivityNet。我们在 ActivityNet 上实现了当前最佳的表现，将 mAP 从 53.8% 提升到了 77.6%。我们希望 YouTube-8M 的前所未有的规模和多样性可以带来在视频理解和表征学习上的进步。

谷歌开源论文工程数据集YouTube-8M视频

提交评论

登录后参与评论去登录

关于我们寻求报道商务合作服务条款

京 ICP 备 12027496

全球人工智能信息服务

友情链接

Synced Global机器之心 Medium 博客PaperWeekly网易智能动脉网硬蛋网

联系电话：+86 010-57150141

联系邮箱：contact@jiqizhixin.com