新智元
来自雪球 发布于08-04 22:20

YouTube-8M 数据集中前 18 个高级类别里的视频示例。

1 新智元原创

作者：王鹤达，清华大学电子系多媒体信号与信息处理实验室

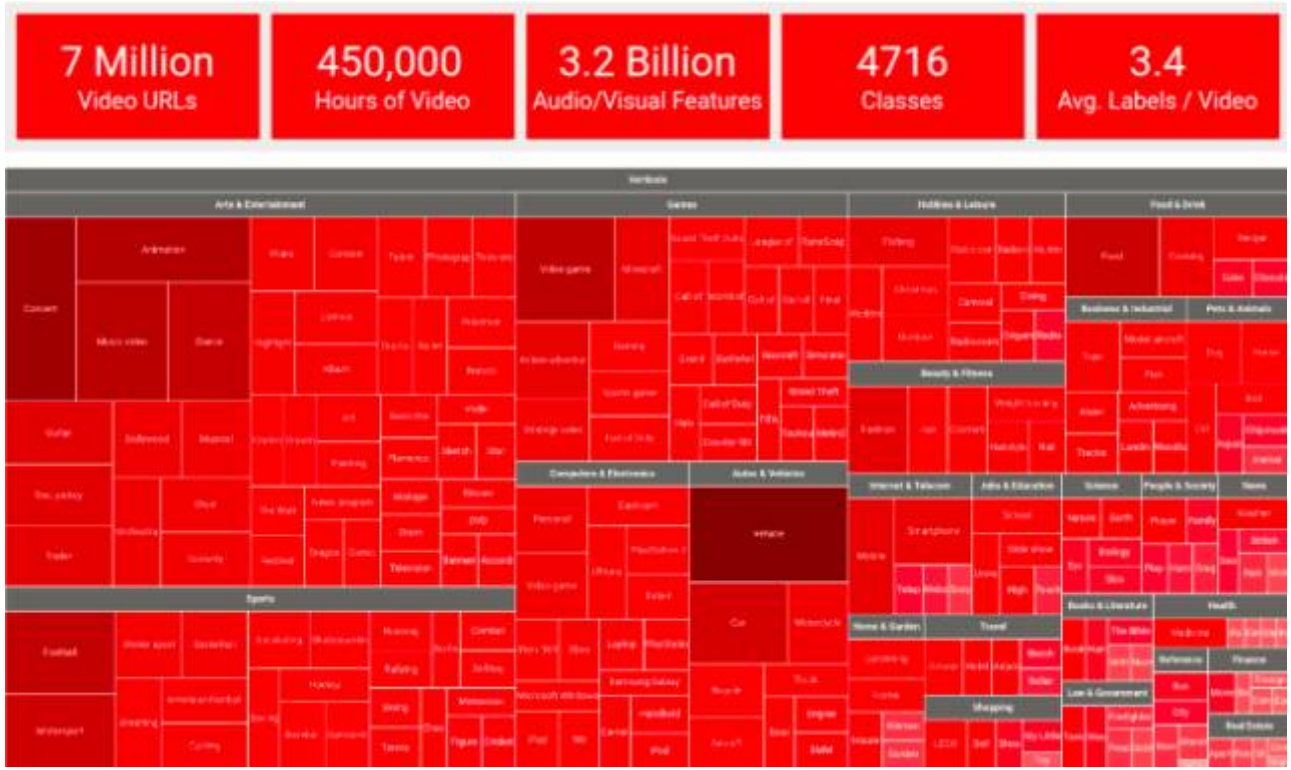
【新智元导读】谷歌云和 Kaggle 共同主办的 YouTube-8M 大规模视频理解竞赛，来自清华大学电子系的团队主要从三个方面对视频进行建模：标签相关性、视频的多层次信息，以及时间上的注意力模型。最终，他们的方法在 600 多支参赛队伍中获得第二。来看他们的实战技术分享。

论文地址：https://arxiv.org/abs/1706.05150

代码地址：https://github.com/wangheda/youtube-8m

理解和识别视频内容是计算机视觉中的一大主要挑战。理解视频也有很多的应用，包括安防监控、智能家居、自动驾驶，还有影视素材搜索和体育视频分析。今年 2 月，谷歌更新了此

谷歌还同时宣布了与 Kaggle 平台联合举办视频理解竞赛，邀请参与者使用 Youtube-8M 作为训练数据，利用谷歌云机器学习，构建视听内容分类模型。表现最佳的参赛队伍将获得 10 万美元的奖金。



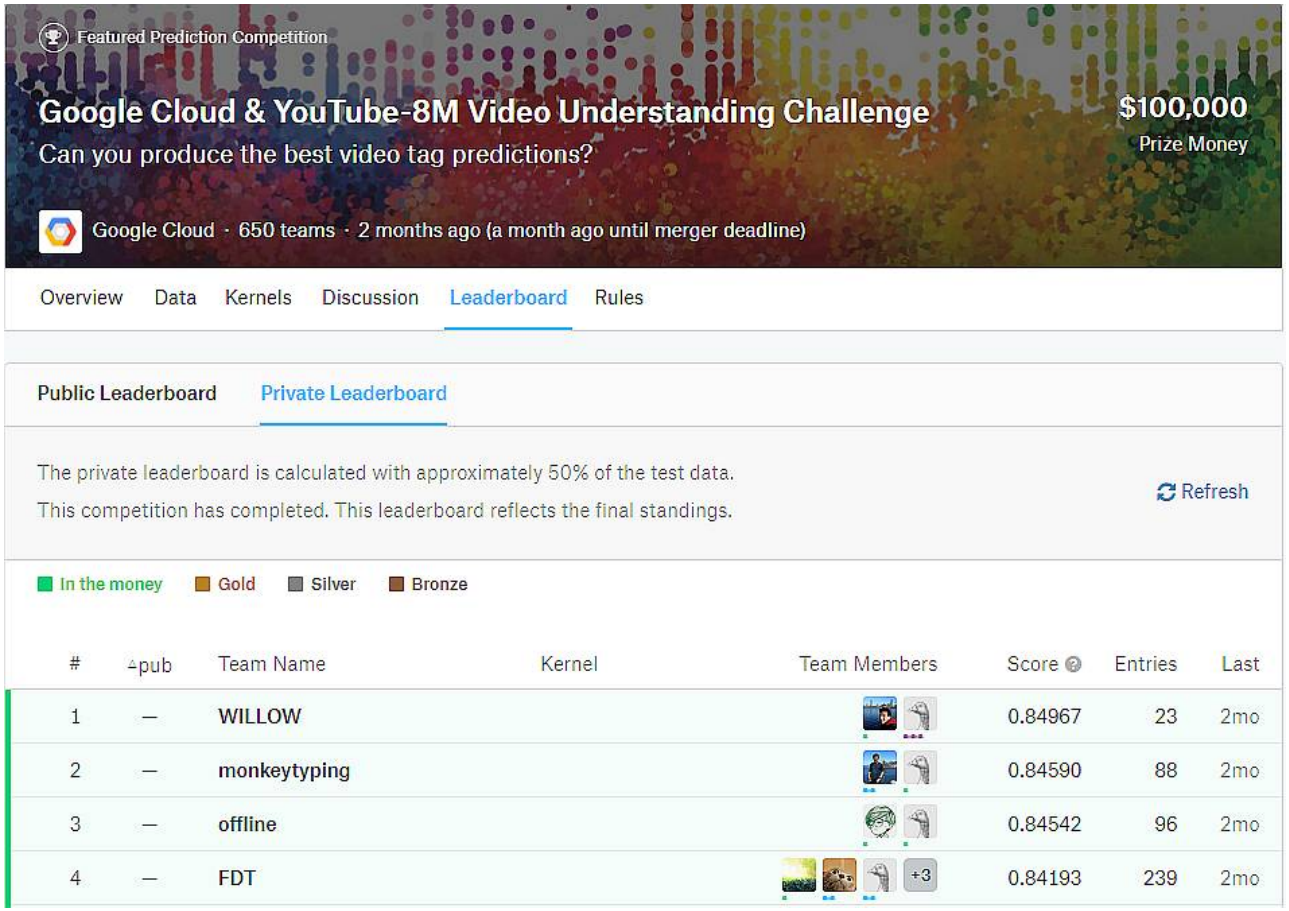
更新后的 YouTube-8M 数据集的 tree-map可视化，分为 24 个高级垂直类别，包括前 200 个最常见的实体，以及每个类别的前5个实体。

YouTube-8M 数据集中前 18 个高级类别里的视频示例。

6 月 30 日，比赛公布了结果。在刚刚结束的 CVPR 2017 YouTube-8M 大规模视频理解挑战赛 Workshop 上，主办方与各参赛团队就各自使用的方法进行了交流探讨。

下文是获得第二名的 monkeytyping 团队所做的赛后总结。团队成员王鹤达与张腾均来自清华大学电子系多媒体信号与信息处理实验室，导师为吴及副教授。张腾目前正在攻读博士学位，研究方向为多媒体事件检测；王鹤达于今年 7 月硕士毕业，他的研究兴趣为推荐系统、自然语言处理与计算机视觉。

Youtube-8M 大规模视频理解挑战赛由 Google Cloud 与数据科学竞赛网站 Kaggle 共同主办，从今年 2 月开始，到 6 月初结束，在四个月的时间里吸引了超过 600 个团队参加比赛。最终，来自法国国立计算机及自动化研究院（INRIA）的 WILLOW 团队夺得第一名，第二名



谷歌 YouTube-8M 大规模视频理解竞赛结果：来自法国国立计算机及自动化研究院（INRIA）的 WILLOW 团队夺得第一名，第二名的 monkeytyping 团队来自于清华大学电子系，第三名的 offline 团队来自于百度深度学习实验室和清华大学，第四名的 FDT 团队来自于复旦大学、中山大学和武汉大学。

ActivityNet 和 UCF101 数据集。另外，这个新的数据集的领域也更加多样，共有 4716 个不同的分类标签，平均每个视频的标签数量为 3-4 个。这些标签取自 Knowledge Graph 中的实体，是由标注程序根据视频的文本和视觉信息进行标注，并经人工检验和过滤得到的。

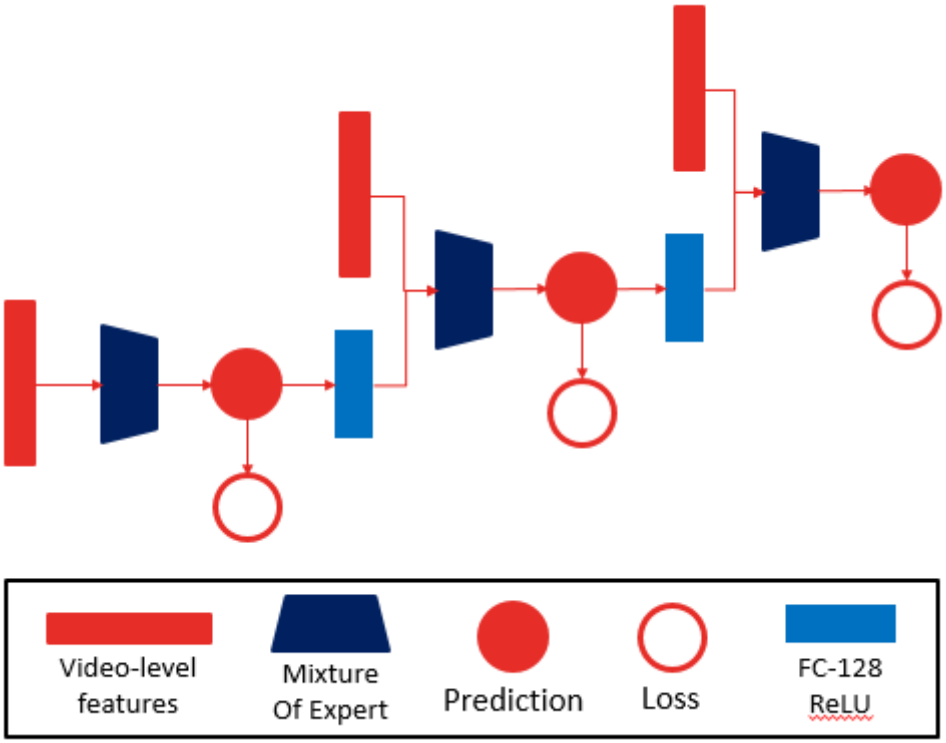
尽管在多样性和数据规模上具有优势，Youtube-8M 数据集也存在着一些限制。首先，为了减少计算上的门槛，Google 对视频数据进行了每秒 1 帧的采样，并使用在 ImageNet 上预训练的网络对每帧图像提取特征。由于数据集中仅包括预提取的特征，这使得参赛者所能使用的手段变得较为有限。其次，数据集中仅包括视频级别的类别标注，没有细粒度的其他种类标注，这限制了数据集应用的场景。第三，数据集中的标签是由机器生成的，在召回率方面具有一定的缺陷。

我们最终提交的结果是 74 个模型组成的 Ensemble，在最终测试集上取得了 0.8459 的全局平均准确率。我们主要从三个方面对视频进行建模：标签相关性，视频的多层次信息，以及时间上的注意力模型。在标签相关性建模中，我们采取一种不断对分类结果进行降维并用于后续预测的网络结构，这种结构可以有效提升多种模型的分类性能。我们使用一种深层卷积神经网络结合循环神经网络的结构，在多个时间尺度上对视频的帧特征序列进行建模。另外，我们采取注意力模型对序列模型的输出进行 Attention pooling 取得了较好的效果。

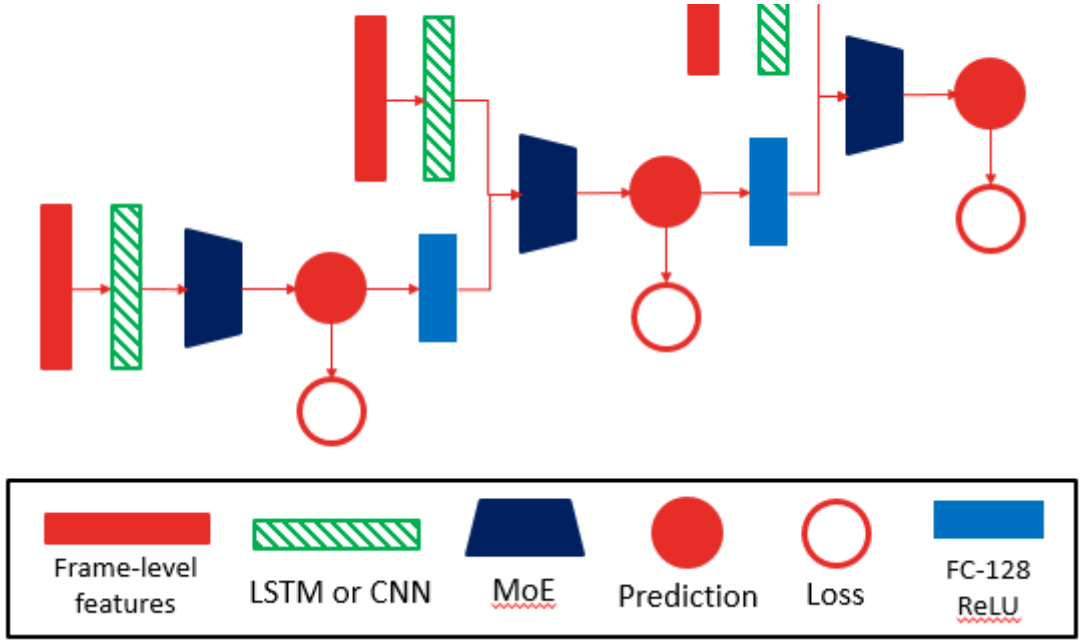
1、标签相关性

我们提出一种链式神经网络结构来建模多标签分类时的标签相关性。如下图所示，当输入是视频级别特征时，该结构将单个网络的预测输出进行降维，并将降维结果与视频表示层合并成一个表示并再经过一个网络进行预测。网络中最后一级的预测结果为最终分类结果，中间

越好。

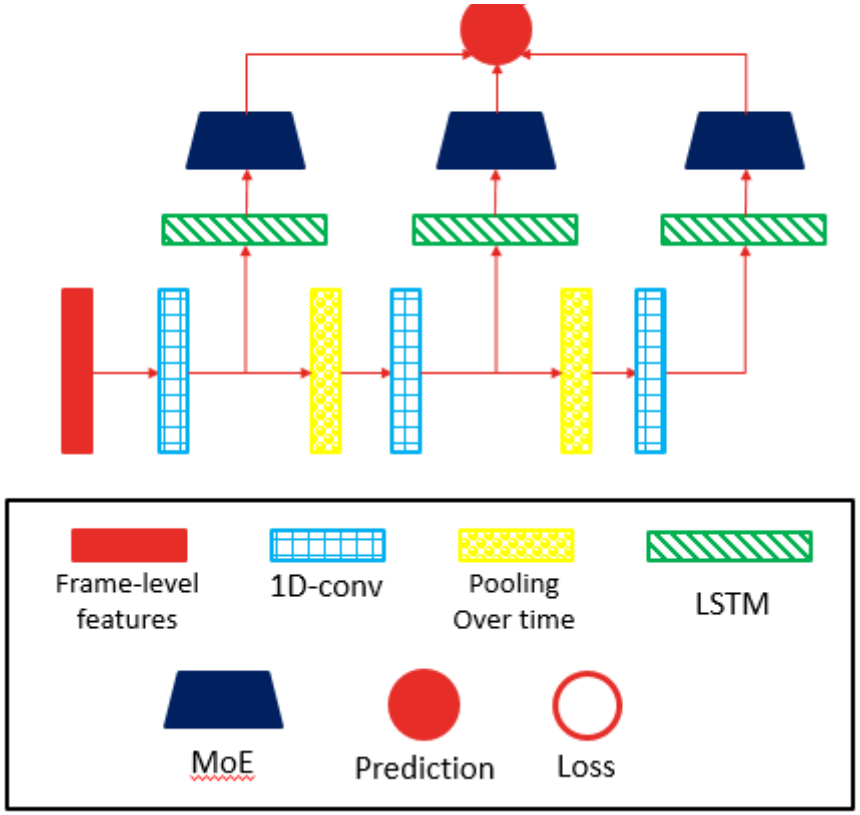


不仅视频级别特征可以使用链式结构，通过如 LSTM、CNN 和注意力网络等视频表示网络，同样也可以对帧级别特征使用链式结构网络。在对该网络进行实验时，我们发现，对其中不同层级的视频表示网络使用不共享的权重，可以获得更好的性能。



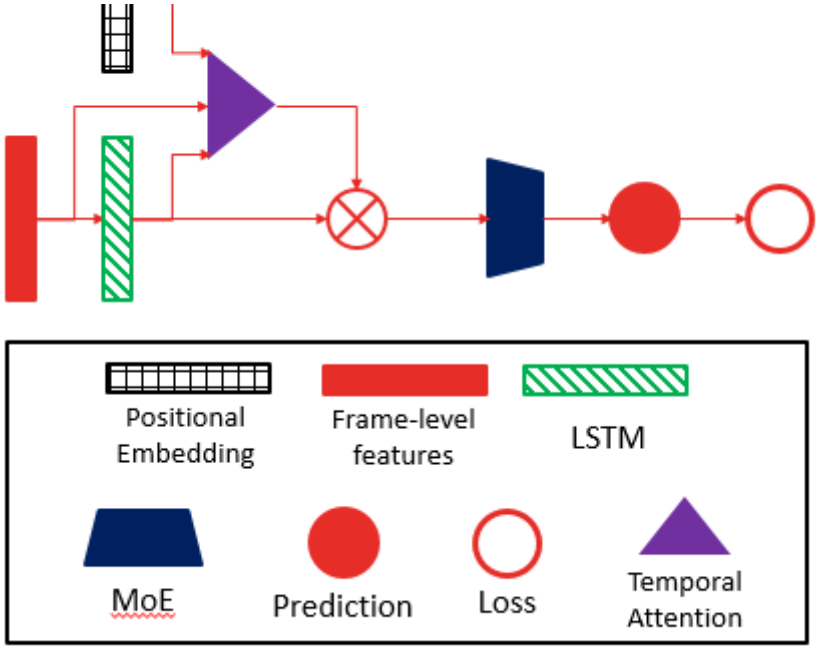
2、利用时间多尺度信息

由于不同的语义信息在视频中所占据的时长不同，在一个时间尺度上进行建模可能会对某些分类较为不利。因此，我们采取一种在时间上进行 pooling 的方式来利用在更大的时间尺度上的语义信息。我们采用 1D-CNN 对帧序列提取特征，通过时间上 pooling 来降低特征序列的长度，再通过 1D-CNN 再次提取特征，如此反复得到多个不同长度的特征序列，对每个特征序列，我们采用一个 LSTM 模型进行建模，将最终得到的预测结果进行合并。通过这种方式，我们利用了多个不同时间尺度上的信息，该模型也是我们性能最好的单模型。



3、注意力模型

我们使用的另一模型是对帧序列的表示采用 Attention Pooling 的方式进行聚合，由于原始序列只反映每帧的局部信息，而我们希望聚合具有一定的序列语义的信息，因此我们对 LSTM 模型的输出序列进行 Attention Pooling。实验表明，这种 Attention Pooling 的方式可以提高模型的预测效果。另外，在注意力网络中使用位置 Embedding 可以进一步改善模型性能。



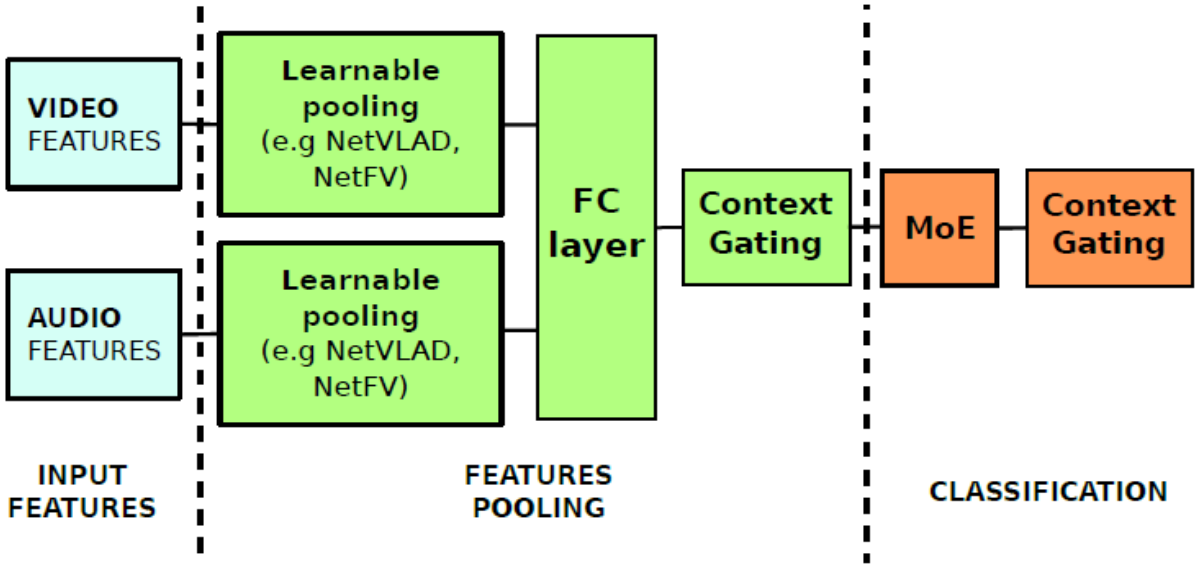
我们对注意力网络输出的权重进行了可视化，我们发现，注意力网络倾向于给予呈现完整的、可视的物体的画面更高的权重，而对于没有明显前景的、较暗的或呈现字幕的画面更低的权重。

在本次比赛中我们感受非常深的一点是大规模深度学习中硬件架构的重要性。由于我们的服务器之间仅以千兆网相连，我们无法做到高效的梯度同步，因而无法利用多机集群来加速运算。我们全部的算法都是在单卡上运行的，其结果是验证性实验的迭代周期变长了，并进行了很多目的性不明确的探索。而在工业界的深度学习集群中，万兆以上的机房网络已经是主流，架构的落后给我们带来了很多困难。

另外，我们也认识到视频分析中算法效率的重要性，在 Youtube-8M 数据集中，预处理阶段需要数千小时的 GPU 时间，而各队所提交的方案又各需要一千至数千小时的 GPU 时间来训练。在实际应用中，这样的运行效率常常是无法接受的，这也是为什么我们认为 Attention Pooling 相关的方法会更加流行的原因。

1. WILLOW 团队：可学习的Pooling + Context Gating

表示，并称之为 NetFV。这两个网络的优点在于计算量小，可以使用帧采样，易于并行。他们对门控线性单元 GLU 进行了简化，将简化的模块称为 Context Gating，通过这个门控单元捕获特征之间的相关性。Gated NetVLAD 也是本次比赛中单模型性能最佳的帧级别分类网络。

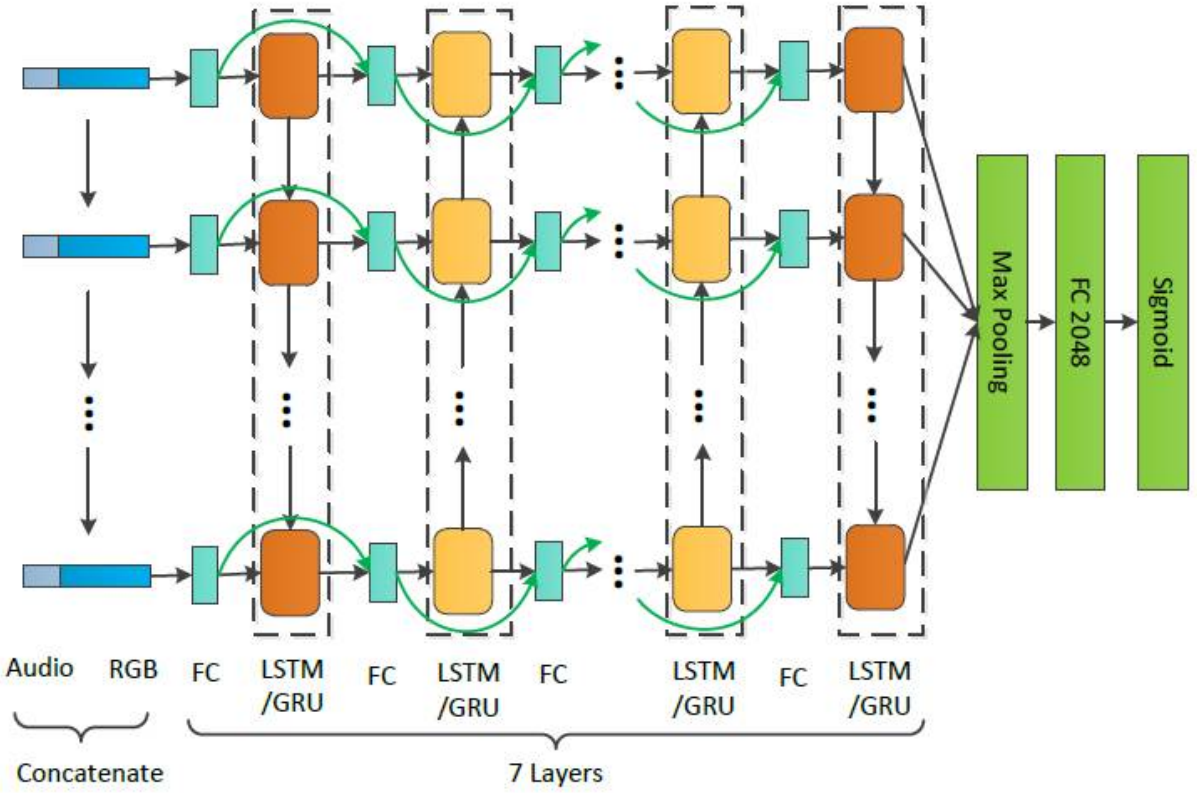


报告：https://arxiv.org/abs/1706.06905

代码：https://github.com/antoine77340/Youtube-8M-WILLOW

2. offline 团队：Fast-forward 序列模型

入了 Fast Forward 连接，有效缓解了训练的困难。该模型是本次比赛中单模型性能最佳的时间序列模型。



报告：https://arxiv.org/abs/1707.04555

代码：https://github.com/baidu/Youtube-8M

点击阅读原文可查看职位详情，期待你的加入~

打赏