**YouTube推出基于深度神经网络的推荐系统**

YouTube的推荐系统是是世界上规模最大、最复杂的推荐系统之一。最近Google的研究人员公布了他们投到今年ACM会议的一篇文章，详细介绍了他们最近利用深度神经网络实现YouTube推荐系统的技术细节。相关会议也会将于本月15号至19号在美国波士顿召开。

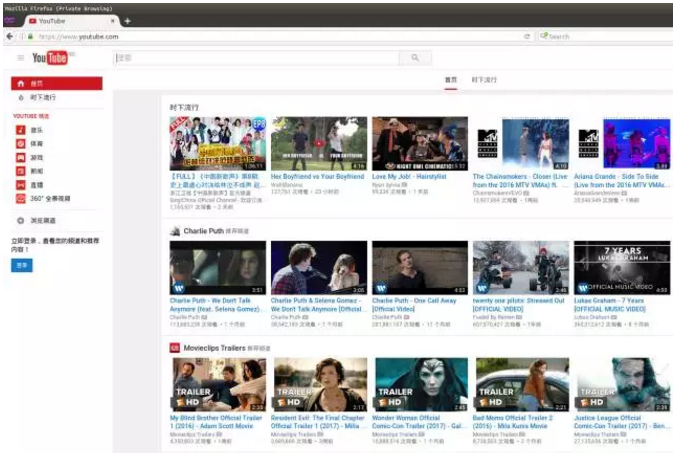
YouTube的全球用户已经超过十亿，每秒上传的视频长度以小时计。视频“语料”库存日益增长，就需要一个推荐系统及时、准确地将用户感兴趣的视频不断推荐给用户。相比其他商业推荐系统，Youtube推荐系统面临三个主要的挑战：

1. **规模**。现有绝大多数可行的推荐算法，在处理YouTube级别的海量视频就力不从心了。
2. **新鲜度**。YouTube视频“语料”库不仅仅是储量巨大，每时每刻上传的新增视频也是源源不断。推荐系统要及时针对用户上传的内容进行分析建模，同时要兼顾已有视频和新上传视频的平衡。
3. **噪声**。由于用户行为的稀疏性和不可观测的影响因素，用户的历史记录本质上难以预测。

因此YouTube推荐引擎很难获得用户满意度的真实信息，取而代之的是使用带有噪声的反馈信号建模分析。

此外，与视频内容相关的元数据通常结构不良、没有明确定义的本体，这也对推荐算法的鲁棒性很有挑战。

YouTube网页推荐效果示意图：



和其他Google产品一样，YouTube推荐系统也将研究重点转移到深度学习，它是基于Google Brain开发的。而Google Brain已经开放源代码，也就是现在风头正旺的TensorFlow。TensorFlow给YouTube推荐系统带来了开发测试上的灵活性，在会议 文章中也有体现。

整个模型的深度网络配置大概需要学习十亿个参数，并且使用了数千亿的数据进行训练。

**系统概况**

YouTube推荐系统主要由两个深度神经网络组成：

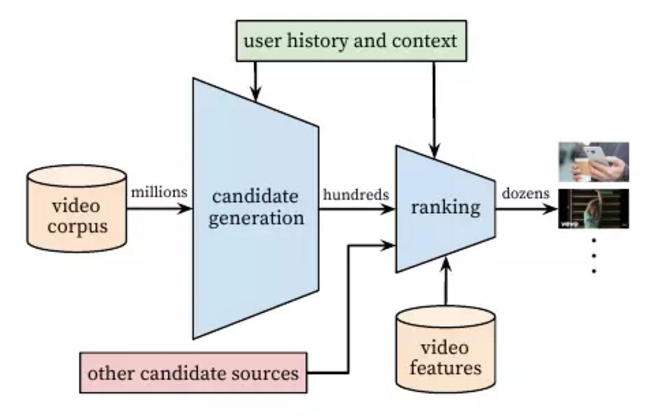
第一个神经网络用来生成候选视频列表；

第二个神经网络用来对输入视频列表打分排名，以便将排名靠前的视频推荐给用户。

候选视频生成是依靠协同过滤算法产生一个宽泛的针对用户的个性化推荐候选名单。排名神经网络是基于第一个候选生成网络的列表，提供更精细的区分细 化，从来达到较高的推荐命中率。通过定义目标函数来提供一系列描述视频和用户的特征，排名网络则根据目标函数来给每一个视频打分。分数最高的一组视频就被 推荐给用户。

两级推荐系统的好处是可以处理海量（百万量级）视频，且保证推荐给用户的视频是个性化的或者有吸引力的。

此外，这种分级设计可以使用其他来源产生的候选视频，而不必完全依赖于来自第一个神经网络的输出。



推荐系统在开发和训练阶段也和其他深度学习技术一样，采用了各种量化指标，如准确率、覆盖率、排名损耗等等。但是为了最终确定算法或者模型的有效 性，Google团队还采用了A/B测试。通过A/B测试，可以衡量例如鼠标点击率、观看时间以及其他量化指标的微妙变化和联系。这一点相当重要，因为很 多时候A/B测试跟离线实验并不相关。

**候选推荐生成**

在候选生成阶段，针对每个用户，上百个相关的视频从YouTube的视频语料库中被选出。推荐过程可以看作是基于排名损失的矩阵分解。Youtube推荐系统早期的迭代是使用浅神经网路，根据用户观看记录，模仿矩阵分解的过程。

现在的手段则可以看作是矩阵分解的非线性实现。

实际上目前很常用的基于协同过滤的推荐算法，本质上就是针对高维度的输入数据进行矩阵分解。而YouTube推荐系统中的候选推荐生成网络，恰好高效准确的模拟出矩阵分解的非线性实现了。这也恰好体现了深度学习在实际工程应用里的普适性。

**以分类方式进行推荐：**

YouTube推荐系统中的推荐过程被当作极多种类的分类来处理。模型中含有几百万个种类，要从中选出N个视频，整个过程仅耗时几十微秒。

**模型架构：**

受NLP中连续词包的启发，推荐系统在固定词汇表里获取每个视频的高维嵌入，再将这些嵌入输入进一个前向传输神经网络。嵌入是通过其他网络参数共同学习的，通过正常的梯度下降后向传输来更新模型参数。

**异质信号：**

使用深度神经网路来实现矩阵分解一般功能的关键好处之一是：任意连续的特征和分类特征都可以随时加进模型中。搜索历史与观看历史同样处理——每一个查询用单字组和双字组标记，每一个标记都被嵌入。

嵌入后的用户查询代表了总结过的密集搜索历史记录。人口分布信息则为新用户推荐提供先验信息。用户的地理位置和设备信息都嵌入并链接在一起。

**标签与上下文选择：**

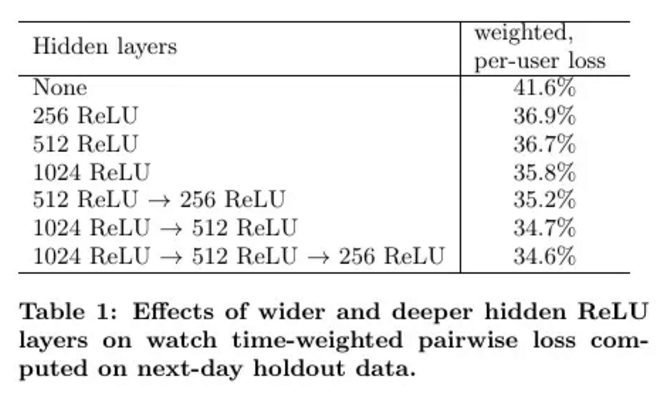
自然的视频消费模式一般会导致非对称的共同观看概率。

**针对特征与网络深度的实验：**

增加特征和网络深度会显著改善推荐系统性能。

**排名系统神经网络**

排名神经网络的主要作用就是根据印象数据，针对特定用户定制和校准推荐。采用了预期观看时间建模测试推荐效果，正对隐藏层的实验表明：更深更宽的ReLU层显著**减小了错误预测的比例**。



YouTube推荐系统被分成两部分来处理两个截然不同的问题：候选生成和排名。这种深度协同过滤可以有效的吸收多种信号，并且通过深层网络对信号间的相互作用建模学习。

候选生成网络的性能优化可以从离线量化测试和在线A/B测试看出。排名网络的性能提升，相比之前线性或树状方法，对观看时长的预测更为准确。深度网络能供有效针对上百种特征间的非线性相互作用建模分析。

对数回归被稍作修改：用于训练的正面例子以观看时间加权，反面例子被设为单元值。这种方法相比于直接预测点击率能更准确的基于观看时间加权排名评估指标。