20th Century Fox — 利用机器学习来预测电影观众

原创：Google谷歌开发者 11 月 21 日

文 / Miguel Campo-Rembado，SVP of Data Science, 20th Century Fox

和 Sona Oakley，Solutions Architect, Google Cloud

电影业的成功仰赖电影公司吸引观众的能力，但有时说起来容易做起来难。观影爱好者是一个多元化群体，他们对电影的爱好多样、兴趣广泛。从历史角度来看，电影公司在决定是否投资某个剧本时严重依赖经验，而这可能会招致巨大风险，在投资新的原创故事时尤其如此。将故事与观众匹配是一个反复迭代的复杂过程，因此 20th Century Fox 总裁、首席数据战略分析师兼媒体主管 Julie Rieger 与数据科学高级副总裁 Miguel Campo-Rembado，以及该公司的数据科学家团队决定使用数据理清这个过程。

适合使用机器学习解决的数据挑战

了解电影观众的市场细分是电影公司的核心职能。多年来，电影公司在高层次数据处理方面进行投资，尝试理清客户细分情况，以及对未来电影的观众作出预测。然而迄今为止，由于技术和制度障碍，细分市场层面的精细预测仍然难以实现，更不必说客户层面。

通过与 Google Cloud 这样的合作伙伴一起努力，Miguel 和他的团队已经能够消除其中一些障碍。我们一同建立起一个隐私保护健全的数据合作伙伴关系，以便更好地了解观影爱好者，还开发出内部深度

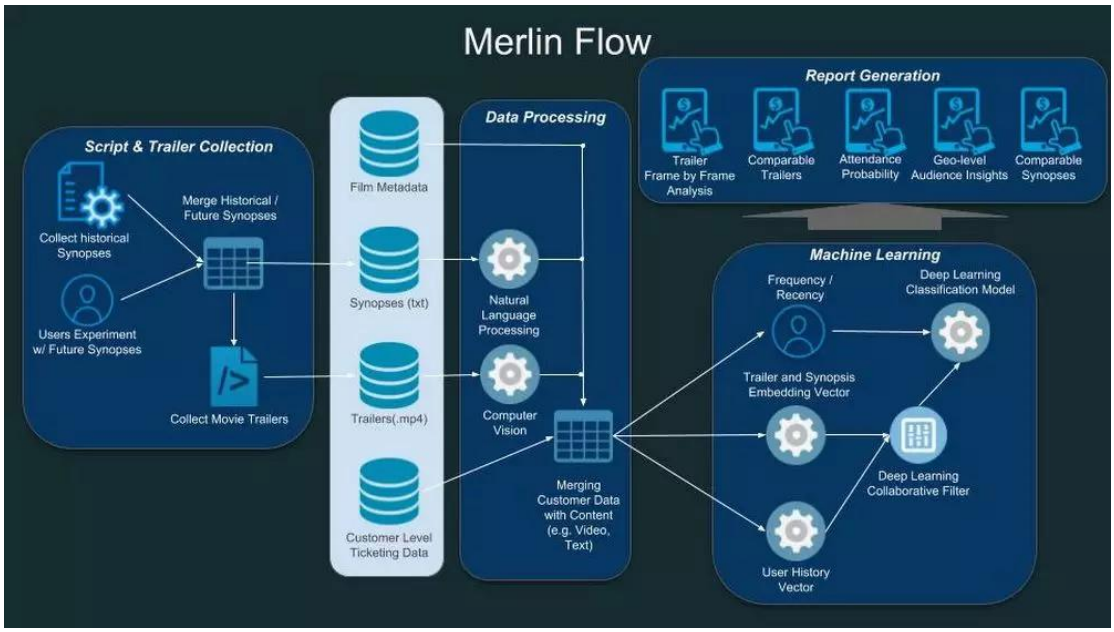
学习模型，并使用精细客户数据和电影剧本来训练该模型，以使其能够识别出观众对不同类型电影偏好的基本模式。在 18 个月的时间里，这些模型已经成为重要商业决策的常规考虑因素，并为评估电影的基调、核心观众与弹性观众的相似性，以及可能的财务业绩提供一份最客观、最讲求数据，也最高效的参考信息。

下面我们来更详细地讨论这些方法。对于电影，只分析从剧本中选取的文本远远不够，因为这些文本只能提供故事的框架，并不具备能够吸引观众观影的任何额外推动力。公司团队想知道是否有某种办法可以使用先进的现代计算机视觉技术来研究电影预告片，因为电影预告片仍然是整个电影营销活动中最核心的元素。新电影的预告片发布是一项备受期待的活动，有助于预测未来的成功，因此公司理应确保预告片能够打动观影爱好者。为了实现这一目标，20th Century Fox 的数据科学团队与 Google 的 Advanced Solutions Lab 合作创造出 Merlin Video。这是一个计算机视觉工具，可用于通过学习电影预告片中的密集表征来帮助预测特定预告片未来电影观众。

设计数据管道

团队的第一步工作是确定应选择哪项技术为工具提供支持。当然是要选择 Cloud Machine Learning Engine (Cloud ML Engine)，并将其与 TensorFlow 的深度学习框架一同使用。由于是托管式服务，因此 Cloud ML Engine 自动配置和监控所有资源，这样团队就可以集中精力为 Merlin 构建深度学习模型，而无需配置基础架构。Cloud ML

Engine 与 Cloud Dataflow 的集成还使其可以在 Data Studio 中无缝生成报告，从而使团队能够更深入地理解流程的工作原理。对该系统的日常维护（主要是数据获取）既简单又轻松，而且可以完全由数据科学家处理，无需其他业务部门的工程师介入。



Merlin 的结构流程图

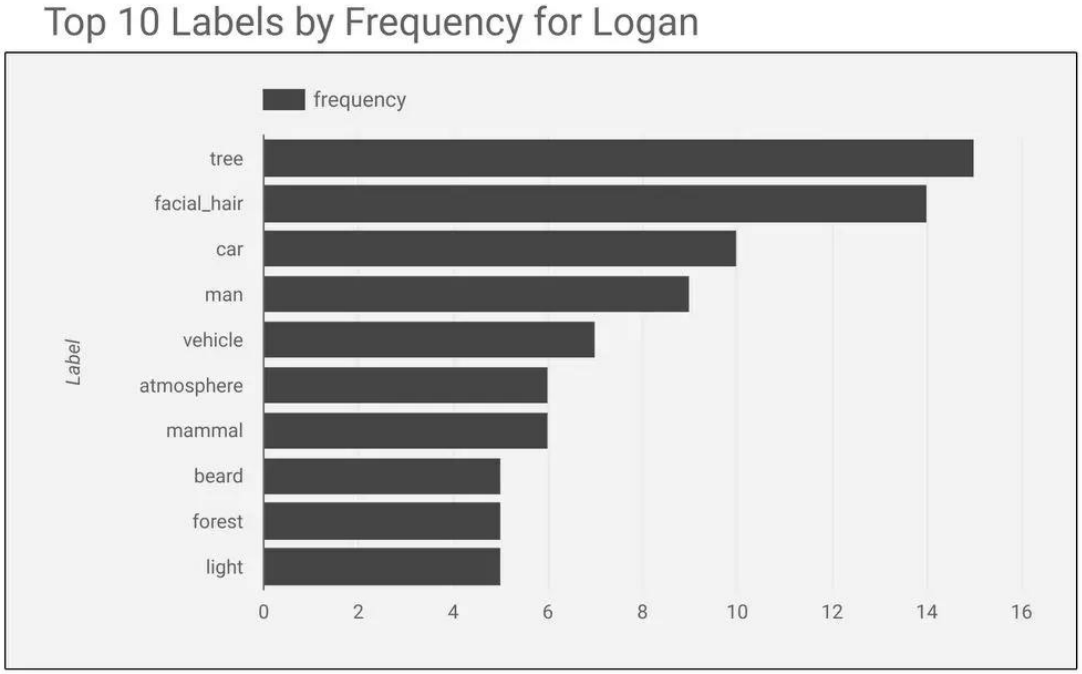
合适的基础架构就位后，团队开始在 YouTube 视频的公开可用数据集（YouTube 8M https://research.google.com/youtube8m/ ）上进行分析。此数据集包含一个 Google 的预训练模型，能够用于分析特定的视频特点，例如颜色、照度、多种脸型、几千种物体，以及一些景观。如上表所示，Merlin 架构中的第一步是解析这些预定义特征，并将其用作前兆，以确定预告片的哪些元素最能预测观影爱好者的偏好。

例如，如果某些观众以前主要观看主演为男性角色的动作电影，那么他们是否很有可能会看另一部主演为男性角色的动作片？下面我们以

《金刚狼》("Logan") 为例，进行深入探讨。该片是一部 20th Century Fox 发行的动作电影，金刚狼一角由 Hugh Jackman 饰演。下图是官方预告片中，第 12 秒画面的快照。

《金刚狼》官方预告片，第 12 秒画面

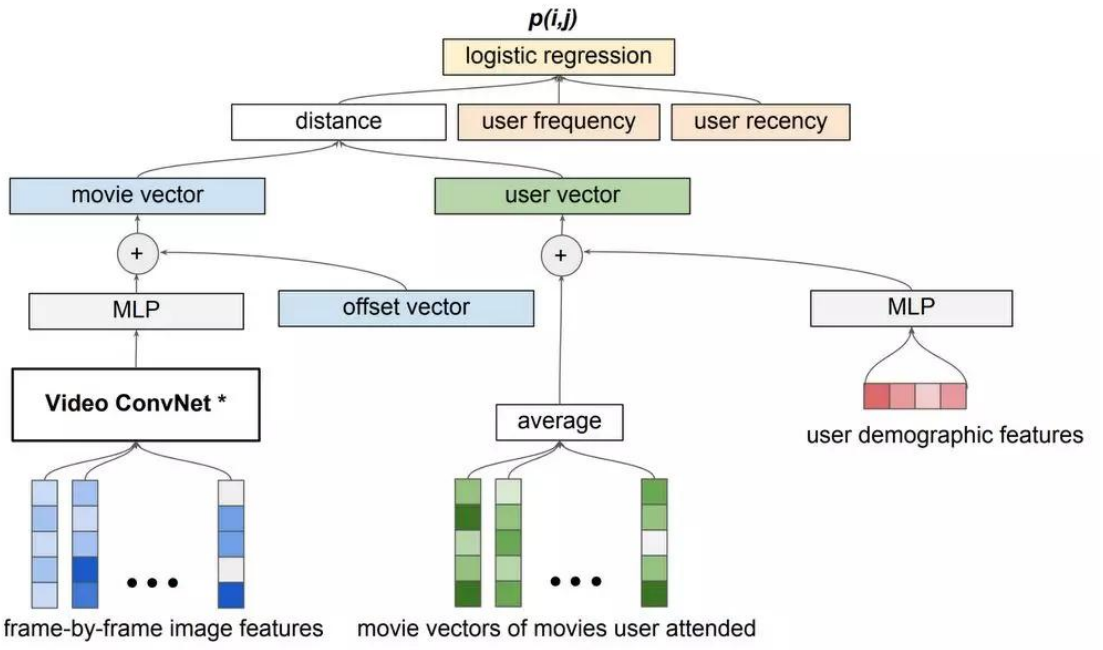
对于这张快照，Merlin 返回以下标签：facial_hair（面部毛发）、beard（胡须）、screenshot（屏幕截图）、chin（下巴）、human（人类）、film（电影）。在完成对整个预告片的逐秒分析后，Merlin 展示出《金刚狼》中最常出现的标签，如下图所示：



Fox 的屏幕截图工具 Merlin 标记的标签，按频率降序排列

在分配对《金刚狼》的标签分析任务后，20th Century Fox 团队想要将这种新的分析与之前从其他电影预告片中生成的标签进行比较，以识别类似的电影。我们可以假定，《金刚狼》的观众和其他动作电影的观众存在重叠，但要验证假设，需要应对双重挑战。第一个挑战是预告片中标签的时间位置：标签在预告片中的出现时间很重要。第二个挑战是这些数据的高维度。对于任何给定的电影，其预告片中都可能有大量能够预测观众兴趣的元素，Merlin 的目标正是同时分析这些元素。Cloud ML Engine 的弹性使数据科学团队能够快速进行迭代和测试，而不会有损深度学习模型的完整性。如此一来，Merlin 便在数日之内就成为可随时用于电影制作的工具，而不必花上数月或数年。

具体来说，分析管道将这些单个组件（标签）送入数据科学团队开发的自定义神经网络中。这个自定义模型会学习电影预告片中标签的时间序列。时间序列（例如，针对某个物体的长镜头与间歇短镜头）可以传达一些相关信息，例如电影类型、电影情节、主要人物角色，以及电影制作人的摄影选择。当与历史客户数据结合后，便可以使用序列分析来创建对客户行为的预测。该管道还包括一个基于距离的“协同过滤”(CF) 模型和一个逻辑回归层，可以将所有模型输出组合到一起，以生成电影上座概率。该模型经过端到端训练，可以将逻辑回归的损失反向传播到所有可训练的组件（权重）。Merlin 的数据管道每周更新，以便分析新发布的预告片。管道结构如下图所示：



最后一步，团队使用 BigQuery 和 BigQueryML 将数百万的客户预测与其他数据资源合并，以创建实用报告，并为营销活动快速设计媒体计划原型。

验证模型

我们回到《金刚狼》这个示例，看看得出的数据能否印证我们的直觉，即之前看过由“粗犷”男性主演的动作片的观影爱好者可能也会看《金刚狼》。我们可以在电影上映后处理有关该观众之前看过哪些电影的数据。下表展示了前 20 名真正的影迷观众（Comp ACT）与前 20 名预测观众 (Comp PRED) 的比较情况。我们重点分析前 5 部真正的电影（下表中以绿色标示），看看它们是否也出现在我们的预测栏中，结果是前 5 部电影都出现在预测栏中。

COMP - ACTUAL	WAS PREDICTED?		COMP - PREDICTED
x-men__apocalypse	TRUE		the_magnificent_seven
john_wick__chapter_2	TRUE		jason_bourne
doctor_strange	TRUE		john_wick__chapter_2
batman_v_superman__dawn_of_justic	TRUE		terminator_genisys
suicide_squad	TRUE		the_legend_of_tarzan
deadpool	FALSE		mad_max__fury_road
terminator_genisys	TRUE		the_revenant
mad_max__fury_road	TRUE		independence_day__resurgence
ant-man	FALSE		spectre
captain_america__civil_war	FALSE		rogue_one__a_star_wars_story
star_trek_beyond	TRUE		the_hunger_games__mockingjay_part_
independence_day__resurgence	TRUE		the_accountant
the_magnificent_seven	TRUE		star_trek_beyond
avengers__age_of_ultron	FALSE		suicide_squad
kingsman__the_secret_service	FALSE		the_martian
arrival	FALSE		x-men__apocalypse
split	FALSE		batman_v_superman__dawn_of_justic
rogue_one__a_star_wars_story	TRUE		san_andreas
fantastic_beasts_and_where_to_find_t	FALSE		doctor_strange
furious_7	FALSE		mission__impossible_-_rogue_nation

Merlin Video 的结果输出：真正的观众与预测观众的对比

从表面上看，我们的直觉没错。《金刚狼》的主要观众实际上是喜爱超级英雄（这点我们已经知道）和“粗犷男主角动作片”（这一点我们尚不能完全确定）观众的组合。观察一下对“粗犷男主角动作片”这一关键词的预测，例如《豪勇七蛟龙》(The Magnificent Seven)（上图中以蓝色标示）、《疾速追杀》(John Wick)（上图中以绿色标示）和《终结者：创世纪》(Terminator Genisys)（上图中以蓝色标示）也出现在前 20 名真实观众的列表中，我们就能更清楚地看到这一点。这是双赢的结果，因为新观众“加入”了超级英雄的核心观众，而且我们或许可以通过他们将该电影的覆盖人群扩展到核心观众之外。

这些工具对 20th Century Fox 的营销和数据团队有非常重大的影响。该团队现在可以部署更多精密工具来确定客户意向，而不只是依赖粗略的观众调查结果。这些数据分析比电影公司之前一直依赖的分析数据集至少详细两个数量级。自 2017 年《马戏之王》(The Greatest Showman) 上映以来，20th Century Fox 一直在使用此工具，并会继续将其用于分析他们最新上映的电影。现在，他们还开始整合从家庭娱乐来源购买和租用的数据，以确定观众成员与他们观看过的电影之间更大的关联性。

最后，由于数据更加精细，该团队可以查看实际票房表现与内部预测的比较情况，以了解哪些细分等级的预测与实际相符。现在，Miguel 的数据科学团队会在每个星期一的早上创建记分卡，然后通过电子邮件将其发送给公司的其他人员。

Movie	Comp PRED	Comp ACT	Was Comp ACT predicted?
the_greatest_showman	la_la_land	wonder	true
the_greatest_showman	wonder	murder_on_the_orient...	true
the_greatest_showman	spy	la_la_land	true
the_greatest_showman	cinderella	cinderella	true
the_greatest_showman	the_intern	the_intern	true
the_greatest_showman	trainwreck	three_billboards_outsi...	true
the_greatest_showman	hidden_figures	a_bad_moms_christm...	true
the_greatest_showman	tomorrowland	lion	true
the_greatest_showman	daddy's_home_2	the_shack	true
the_greatest_showman	fifty_shades_of_grey	sully	true
the_greatest_showman	the_shack	daddy's_home_2	true
the_greatest_showman	sully	the_imitation_game	false
the_greatest_showman	a_bad_moms_christm...	hidden_figures	true
the_greatest_showman	bad_moms	bridge_of_spies	false
the_greatest_showman	pitch_perfect_2	going_in_style	false
the_greatest_showman	murder_on_the_orient...	pitch_perfect_2	true
the_greatest_showman	gone_girl	jumanji:_welcome_to...	false
the_greatest_showman	lion	tomorrowland	true
the_greatest_showman	three_billboards_outsi...	beauty_and_the_beast	false
the_greatest_showman	fifty_shades_darker	miss_peregrine's_ho...	false

Movie	Comp PRED	Comp ACT	Was Comp ACT predicted?
love_simon	bad_moms	game_night	true
love_simon	a_bad_moms_christm...	pitch_perfect_3	true
love_simon	the_greatest_showman	three_billboards_outsi...	false
love_simon	trainwreck	the_greatest_showman	true
love_simon	game_night	red_sparrow	true
love_simon	fifty_shades_freed	the_post	false
love_simon	wonder	a_wrinkle_in_time	true
love_simon	pitch_perfect_3	wonder	true
love_simon	la_la_land	a_bad_moms_christm...	true
love_simon	pitch_perfect_2	fifty_shades_freed	true
love_simon	girls_trip	pitch_perfect_2	true
love_simon	ghostbusters	tomb_raider	false
love_simon	a_wrinkle_in_time	trainwreck	true
love_simon	cinderella	lion	false
love_simon	red_sparrow	daddy's_home_2	true
love_simon	spy	murder_on_the_orient...	true
love_simon	murder_on_the_orient...	bad_moms	true
love_simon	daddy's_home_2	miss_peregrine's_ho...	false
love_simon	fifty_shades_of_grey	la_la_land	true
love_simon	the_intern	baby_driver	false

Actual scorecards for *The Greatest Showman* (left) and *Love, Simon* (right)