阿里文娱技术 阿里云 开发者社区

全景揭秘
阿里文娱智能算法

计算机视觉 | 搜索推荐 | 文娱智能

—— 阿里文娱技术精选系列——

文娱智能算法

关注我们

（阿里文娱技术公众号）

关注阿里技术

扫码关注「阿里技术」获取更多资讯

加入交流群

1）添加“文娱技术小助手”微信

2）注明您的手机号 / 公司 / 职位

3）小助手会拉您进群

By 阿里文娱技术品牌

更多电子书

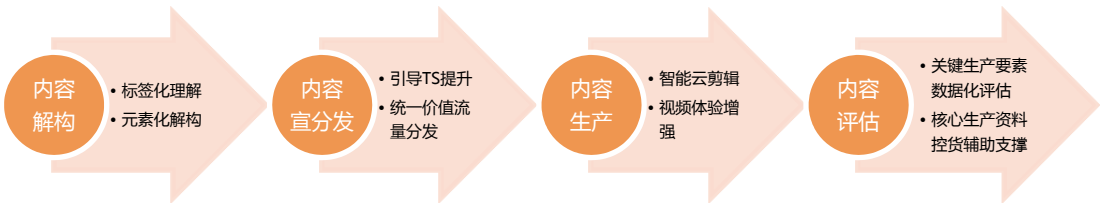
扫码获取更多技术电子书

目 录

序

阿里是一家坚信数据力量的公司，而文娱涉及的相关产业非常广泛，从线上到线下、从影剧综漫到现场娱乐以及文学小说等，其组成、形式、展现、分发的复杂性交织在一起为业务数据化带来了巨大的挑战。

近三年来，阿里文娱摩酷实验室始终以助力业务发展和增长为核心驱动，形成如下四个的技术方向：



内容理解是文娱相关算法技术的基石，IP、小说、剧本、视频、音乐等不同形态的内容对构建起领域知识图谱带来了很多困难，在这其中计算机视觉、自然语言处理、图谱&推理、图神经网络、多模态内容分析等技术被广泛应用于内容解构。以视频为例，影剧综视频的时长很难用一些低层级的标签来抽象表达其内容，基于多模态的分析技术在这类内容上也会碰壁，因此融合内容专家及机器学习系统的半自动化微标签体系成为一种可行的出路。与短视频快速的线上反馈闭环不同，即使制作周期最短的综艺节目也需要 3 个月以上，期间还面临内容监察审核的不确定，这就导致影剧综内容制作高度的不确定性，如何基于复杂的数据分析线索以及历史的成败规律来选择评估内容是各个综合视频平台所面临的核心挑战之一，而阿里文娱北斗星系统就是用来解决这一问题的。

搜索和推荐作为两种解决信息爆炸的重要手段被广泛应用于各个 APP 中，而影剧综内容的复杂性导致用户想精确描述一个内容非常困难，仅通过节目名、演员名去检索给用户也造成了很大的困扰。在文娱内容的分发体系中对搜索模式、推荐模式的融合成为新的用户需求，如何更为准确的通过类强化学习的用户意图理解过程来协助他们尽快找到喜爱的内容，成为文娱搜

推体系下一阶段的首要任务。

文娱作为产业互联网发展的重要行业，人工智能技术在这个领域中的应用空间广大，而我们也仅仅是迈出了一小步，期待工程师们能够创造出更大的奇迹，加速文娱产业数字工业化时代的到来。

阿里文娱摩酷实验室负责人 王晓博

2020.02.01

1

计算机视觉

分区域处理的图像和视频清晰化技术

作者| 阿里文娱高级算法专家 出林、阿里文娱高级算法工程师 文渊 苍华

一、UPGC 视频和图像质量面临的挑战

在优酷这种综合性的视频平台，用户的观看体验永远是第一位的，而画质是影响观看体验的重要因素。对于影视剧来说，画质和拍摄年代有较强相关性，也就是说随着拍摄设备和技术的提升，画质也在提升。用户一旦习惯了更高清的内容，就“回不去了”，进而对视频画质甚至显示设备提出更高要求。而对于目前大量增加的 UPGC 视频，画质情况却不容乐观，UPGC 视频来源主要包括两种：一种是由正片切条产生的短小视频经用户上传的，这种情况下，由于用户使用的片源清晰度无法保证，又经过多次的转码、压缩、缩放，会导致画质下降，导致压缩噪声、块效应等问题；另一种是用户拍摄上传的，虽然目前手机相机成像质量越来越好，分辨率越来越高，甚至出现了 1 亿像素、30 倍变焦等黑科技，但在不受控的拍摄环境中，普通用户终究无法控制拍摄质量，从而导致噪声、模糊、光线等问题。

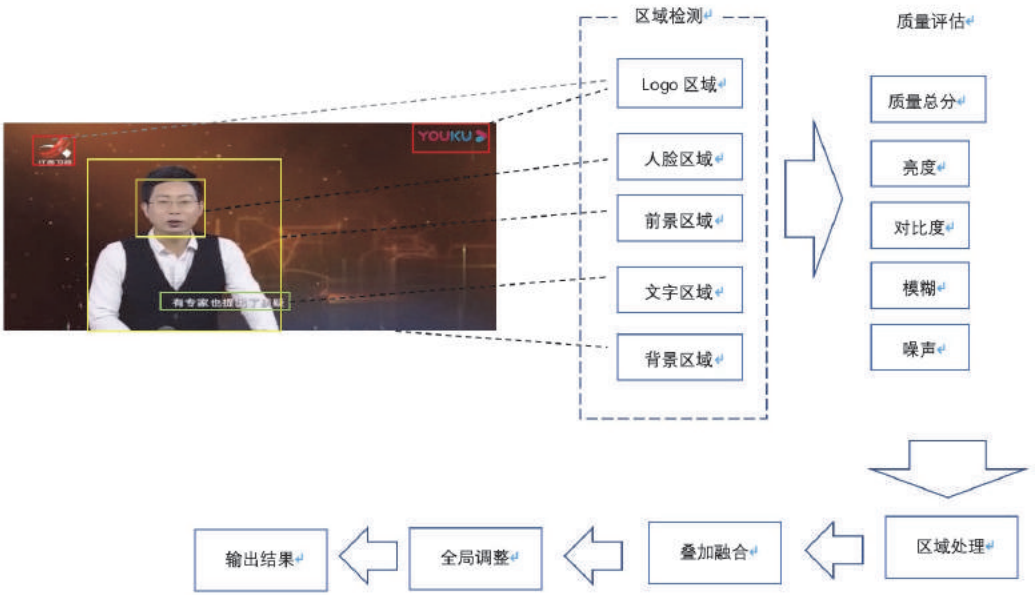
视频的封面图也是由原始视频截帧得到的，好的封面图会提升用户的观看欲望。如果原视频画质差，即使封面图经过人工和算法的精挑细选，也是“矬子里面拔将军”，提升空间不大。更坏的情况是，截帧之后选中的图片还要进行图片压缩，进一步降低了画质。

在所有画质问题中，“伪高清”问题最为突出，也就是说虽然表面上看视频分辨率很高，达到了 720p 甚至 1080p，但实际画质观感很差，甚至不如 540p。因为“伪高清”视频不能通过分辨率简单判别，所以想要解决“伪高清”问题，就要先识别它，然后再做针对性画质增强。

通过影视剧切条得到的 UPGC 视频，即前面提到的第一种来源，有非常显著的特点。这种视频有很强的背景虚化，原片中人脸等重点区域细节丰富，经常出现字幕或 logo。针对这些特点设计增强方案，会有事半功倍的效果。

二、图像和视频清晰化解决方案

基于对业务场景的深入理解和分析，我们设计出完整的视频和图像清晰化解决方案，该系统有几个明显的特点：画质评估和清晰化形成闭环，分区域清晰化后再进行融合。



图：图像清晰化方案

1. 区域检测

我们将区域划分为 logo 区域、人脸区域、前景区域、文字区域、背景区域等几个典型区域，分别利用文字和 logo 检测、人脸检测、saliency 区域检测等算法得到。后续的区域处理和融合模块对区域精度要求不高，所以出 saliency 区域有较精细的区域分割外，其余均使用检测框。

2. 分区域处理策略

划分前背景分别处理，是由于我们观察到超分辨率（super resolution）模型的一些特性，现有的 SR 模型会对“疑似”边缘做强烈的恢复。模型应用于背景虚化区域，某些轮廓会被增强成强边缘，而其他区域仍保持虚化的效果，这样就造成了“突兀”的效果，和人的主观认知不同。所以我们的模型对前景区域进行纹理恢复，背景区域只做简单的亮度对比度调整。

对于 logo 和文字区域，由于这类图像本身就是数字化内容，模式较单一，较容易通过简单

模型达到好的效果。顺便提一下，对于动画片的处理也是类似原理，相比复杂的真实场景图片，动画片总是更容易处理。

对于影视剧和短小视频，人脸是用户关心的重点，所以我们设计了人脸清晰化模型对人脸和头发等区域单独处理，通过大量高清人脸图片训练 SR 模型，并适当加入 GAN loss，可以恢复出人脸五官、毛发细节和皮肤纹理，达到分毫毕现的效果。

总结一下，我们多个模型对不同区域进行处理，分为 logo 和文字模型，人脸清晰化模型，和一般前景清晰化模型。其余还有一些通用的亮度对比度调整算法，对图像全局进行调整。

3. 质量评估模块

优酷摩酷实验室构建了大规模的 UPGC 图片质量数据集，并提出了 multi-level 特征融合的无参考质量评价框架（见我们的另一篇文章：基于人类视觉感知的视频体验评价体系），该方法不仅输出总体质量分，还可以输出失真类型。

得益于实验室良好的技术沉淀，我们的线上数据都可以打上质量分和失真类型，进而和清晰化模型结合形成评估+增强的业务闭环。

我们依据质量总分将数据划分为好、中、差三档，对于本来画质已经很好的图片不做处理，对于中和差的数据依据失真类型筛选出清晰化模型能处理的部分，并根据失真程度赋予清晰化模型不同的恢复参数。

4. 叠加融合模块

由于分区域处理模块只负责纹理和边缘的恢复，亮度和对比度后置到全局调整模块，我们的框架对分割和融合精度的要求较低，只需要简单的 alpha blending 就可以达到好的效果。

5. 视频清晰化

以上是面向图像的清晰化方案，对于视频场景我们做了适应性改进。为了保证前后帧效果的一致性，我们对增强参数做了时间平滑。将图像场景的 SISR（单帧超分辨率）模型替换为 VSR（视频超分辨率）模型，增强了对视频压缩问题的处理能力。同时，我们构建了 UPGC 视频质量评价数据集，并在此基础上训练了视频适量评价模型，将视频按质量分档，并针对失真类型进行处理。

三、重点算法原理介绍

1. 快速的融合模型

提升图像和视频清晰化的方法有超分辨率、锐化、以及将超分辨率和锐化结合的形式等。

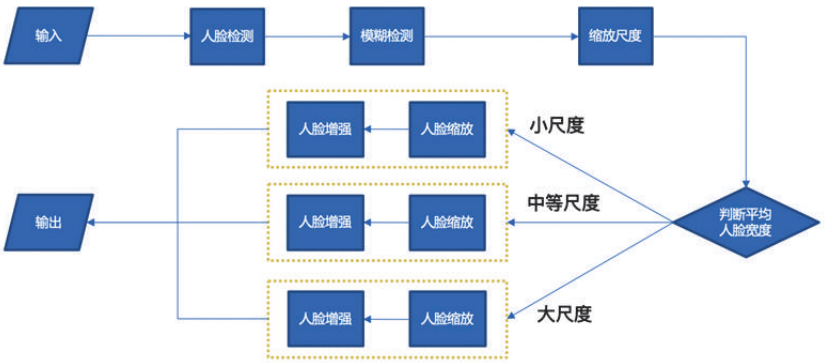
对于超分辨率，在学术界早年通常采用 bicubic 降采样的方式构造图像或视频数据对，这种方式构建的数据对的输入数据分布通常跟真实低分辨率图像或视频的分布相差很大，导致应用在真实低分辨率数据上，会出现各种各样的问题。比如在 bicubic 降采样方式构建的数据训练的模型应用在真实的低分辨率图像上后，会出现网格状的 artifacts。近几年，在构造数据和模型框架上，学术界做了一些新的尝试。比如阿里巴巴达摩院的研究人员在构造数据集时采用了 realSR 的方式，通过搜集同一场景下不同分辨率相机的图像，然后通过图像匹配的方式构建训练数据对，这种方式虽然一定程度上能够让获得的低分辨率图像更加接近真实的低分辨率图像，但也存在着对齐的问题。比如由于存在非严格对齐问题，造成光晕的现象。另外在模型框架下，近几年也涌现了一些采用非监督方式训练超分辨率。但非监督方式跟监督方式相比，在效果方面还有一定的差异，需要研究人员进一步提升模型的效果。

对于锐化而言，通常是采用传统算法，但传统算法也存在着一定的问题。比如传统的经典锐化算法 DOG，会存在噪声的放大和锐化过渡导致光晕的问题。另外一些锐化算法，采用经典的保边滤波算法，提取图像的低频，进而获得图像的高频信号，但这一类算法由于采用了较复杂的保边滤波算法，通常速度比较慢，很难达到工业界对于速度的要求。

另外一类锐化算法借鉴近几年大热的深度学习算法，将保边滤波提取低频这一步骤采用深度学习来做，一定程度上缓解了速度的瓶颈，但对噪声放大问题并没有得到很好的解决。另外一种是采用超分辨率和锐化相结合的方式，常见的做法是采用深度学习进行超分辨率，然后结合传统的 DOG 算法进行锐化。通常而言采用深度学习对低分辨率图像处理后的图像距离 GT 图像还有一定的距离，因此需要采用锐化进一步提升清晰度。但由于采用了先进行超分辨率，然后锐化的方式，而超分辨率采用深度学习算法，通常是在 GPU 上运行，而锐化通常是采用传统算法，是在 CPU 上运行，中间涉及到 GPU 和 CPU 的相互切换等，因此对于视频而言速度并不快，也很难达到实时处理的要求。为了解决这个问题，我们采用快速融合模型的方式，即采用单个深度学习网络，同时学习超分辨率和锐化，可以在基本不损失效果的基础上，速度得到很大程度的提升。

2. 人脸清晰化

线上的大量素材和长短视频大部分以人像为主体，人像的清晰程度是影响用户视觉体验的主要因素。针对人脸清晰化我们设计了如下算法流程：



流程：

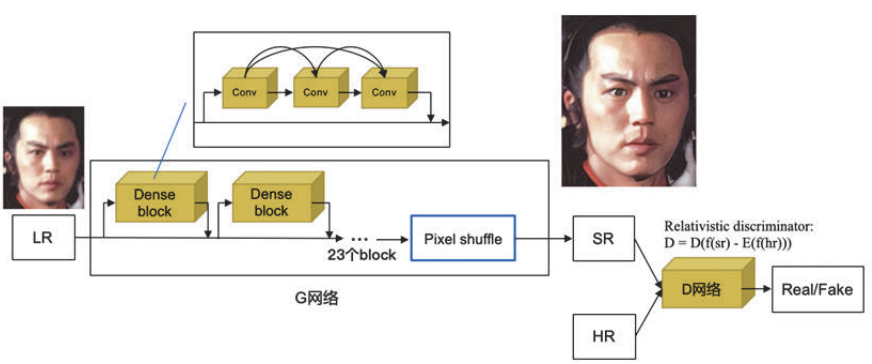
1）原始图像通过人脸检测器，检测平均人脸大小：为了提升不同尺度下人脸增强的效果，我们对比了单模型和多个尺度模型效果，多个尺度模型的还原效果优于单模型结果；

2）通过模糊检测预测降尺度系数，缩放图像以降低模糊程度：实际素材存在不同类型和程度的模糊退化问题，模糊程度较高时还原结果会存在较多失真纹理，因此单独训练了一个模糊检测器预测降尺度系数，通过图像降尺度，减小模糊因素产生的失真问题；

3）判断原图平均人脸尺度，缩放图像至三种尺度中心；

4）选取相应尺度增强模型，通过人脸增强模块，得到高清人脸。

我们的主要工作是针对人脸增强模块，设计了基于 gan-loss 的超分网络，结构如下：



LR 为低清图像，SR 为超分图像，HR 为高清图像。人脸增强模块的训练由生成器（Generator）和判别器（Discriminator）两部分组成，生成网络使用了稠密连接的 Residual-in-Residual Dense Block（RRDB）结构，有利于提取层级较深的图像特征，判别器参考 RaGAN 判别 SR 和当前批次 HR 图像特征差距来判别 SR 的真实度是否超过批次 HR。判别器为：

$$D_{ra}(sr, hr) = D(f(sr) - f(E(hr))$$

其中$f(sr)$为低清图像特征，$f(E(hr))$为当前 mini-batch 高清图像特征的期望

训练过程的损失函数包含三部分：

$$L = w_0 L_{pixel} + w_1 L_p + w_2 L_D$$

其中$L_{pixel} = L_2(sr, hr)$，L_p为感知损失函数，判别器损失函数L_D如下：

$$L_D = -\Sigma(log(D_{ra}(sr, hr)) + log(1 - D_{ra}(hr, sr)))$$

针对素材图场景，为了提升人脸细节清晰度，我们使用 Pixel shuffle 作为上采样方式，SSIM 指标提升。在线上测试过程中，发现增强结果中有 5%左右的样例存在失真现象，后通过实验对比，发现 gan-loss 权重过大、原图模糊程度较高是导致失真现象的主要因素，通过降低 gan-loss 权重，且仅在中间训练阶段采用 gan-loss，引入模糊检测模块对原图降尺度，最终有效减少了失真纹理的产生。为了解决大尺度人脸清晰度还原不足问题，使用特征金字塔结构融合多尺度信息以提升增强稳定性。针对短视频场景，使用轻量化模型提升模块速度，达到 50ms/帧，并对人脸区域边缘作平滑以减弱过渡纹理不自然的现象。

3. VSR 模型

深度学习视频超分辨率技术通常分为两种，一种是单帧超分辨率，另外一种是多帧超分辨率技术。

单帧超分辨率速度通常较快，但很难解决前后帧连续性的问题，从而导致画面的闪烁，导致人的主观感受较差。多帧超分辨率算法，一方面可以很好的解决前后帧连续性的问题，另外一方面由于利用了多帧的信息进行处理，在主观效果上要明显好于单帧算法。多帧超分辨率技术的主要问题是速度相对比较慢。目前 state of art 的算法是商汤的 EDVR，借鉴传统视频处理算法，包括帧对齐模块、帧间融合模块等。

优酷视频场景下，一方面面临着分辨率不足的问题，另外一方面面临着压缩、噪声等问题。

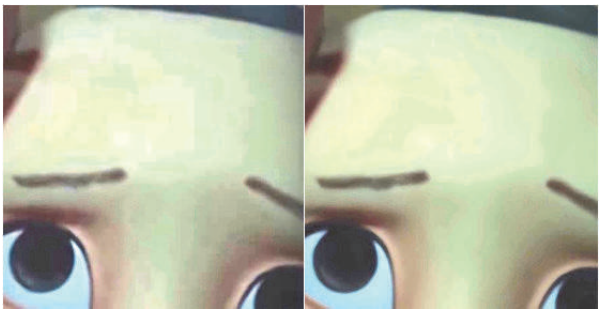
因此对于优酷场景，需要在对视频进行分辨率提升的同时，能够有效的解决压缩、噪声等视频画质退化问题。

为此，我们进行了大量的尝试和方案验证，从而找到了贴合优酷视频场景的相关解决方案。在数据处理方面，一方面，我们采用 GAN 网络等设计了视频降质工具包，可以一定程度上模拟优酷场景下的视频降质过程。另外一方面我们从优酷有版权视频库中获取同一视频的不同分辨率视频，并对视频进行匹配和清洗，从而构建贴合优酷业务场景的训练数据集。在模型结构设计上，为了解决分辨率不足的问题，我们借鉴了主流 VSR 模型的 PixelShuffle 模块，与此同时为了解决尺度连续性问题，我们采用了多尺度金字塔融合的方式。为了解决帧间连续性问题，我们借鉴传统视频多帧算法，引入了多帧对齐模块，并在此基础上融合了 attention 模块，对视频进行了分区域处理。为了解决噪声问题，我们借鉴传统的频谱分解方式，在网络结构中加入了小波分解和重建模块。为了解决去压缩问题，我们引入了 ResBlock 模块。最终融合了上述模块的网络结构，在优酷业务数据集上训练后，对优酷场景下视频面临的噪声、压缩、低分辨率等问题得到了很好的解决。

四、处理效果和业务收益

1. 去除压缩导致的噪声问题（建议放大观看）

左图为原图 右图为处理后图



原图　　处理后

为便于观察，局部做了提亮处理，可见处理后更细腻，条带/阶梯效应大幅减少。

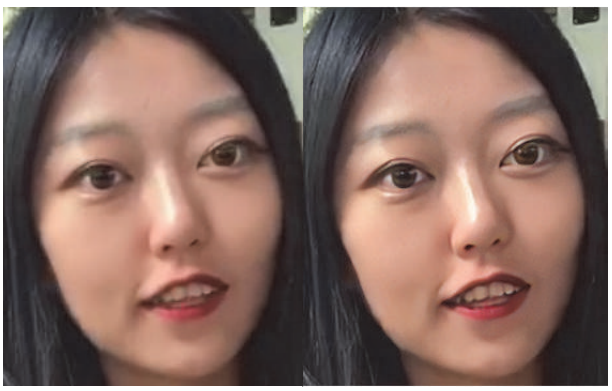
2. 算法采用分区处理，重点优化文字、人脸等区域，提升画面清晰度

原图

处理后

原图　　处理后

原图　　处理后



原图　　处理后

人脸部分可见明显提升，五官细节得到恢复。

原图　　处理后

Logo 和文字部分清晰度明显提升。

3. 用于素材海报图的清晰度提升

原图.　　处理后

人脸部分截图　　处理后图

五、总结

以上详细描述了优酷 UPGC 场景视频和图像清晰化解决方案，并介绍了关键算法的原理和改进思路。采取分区域处理的策略，对不同的区域分别处理，对文字、logo、人脸等区域效果尤其明显，人脸达到了分毫毕现，毛发和纹理细节得到了恢复。我们提出了自己的质量评价模型，结合视频和图像清晰化模型，构建了完整的质量评价+增强解决方案。

算法的进步是永无止境的，当前各种算法技术也是层出不穷，如何把算法融会贯通并加以改进，应用于我们的业务场景，让算法发挥最大价值，是我们长期研究的问题。

基于人类视觉感知的视频体验评价体系

作者| 阿里文娱资深算法专家 镜一

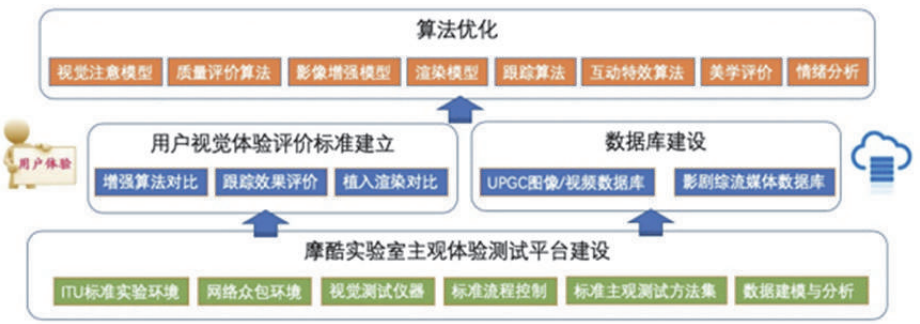
一、背景

视频质量评价技术是指基于视觉生理学心理学特性，例如人眼的多通道、多方向分解特性，视觉对比敏感度函数（Contrast Sensitive Function, CSF）和恰可失真门限（Just Noticeable Difference, JND），视觉注意（Visual Attention, VA）机制等对用户观看体验进行定量分析，包括主观评测以及客观建模。视频体验的终极受体是人眼，因此视频质量的评价可以与 4K/8K 极高清，HDR，AR/VR 等视频处理技术形成闭环，指导其优化最终达到增强用户观看体验的目的。

起初在大家还只是把电视/电脑显示器作为观看视频的主要手段的时候，由于人眼是视频的最终受体，视觉质量也因此称为 visual perceptual quality，即，只是视觉上的画面质量感受。随着多媒体和硬件技术的发展，3D 立体电视电影（需佩戴 3D 眼镜观看，裸眼 3D 技术尚不成熟效果极差这里不做讨论）的兴起导致行业内必须重新对视觉质量进行定义。在立体视频中，除了画面本身的质量这个维度外，又多了两个维度：深度（depth）和视觉疲劳（visual discomfort/fatigue）。在 2012 年，欧盟 Qualinet（European Network on Quality of Experience in Multimedia Systems and Services）发布了关于视频体验质量的白皮书，里面建议把这种多维度的感知体验用 Quality of Experience（QoE）来表示。其具体定义为“Quality of Experience is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and /or enjoyment of the application or service in the light of the user’s personality and current state”．也就是说，感知质量与具体应用和服务相关，基于用户对于设备或者服务在可用行上或享用性上是否达到期望的满足程度。期望因人而不同（受职业，性别，年龄，教育背景，个性等的影响），即便针对于特定的某个人，他/她的期望也会因他/她本人当前的状态（例如，情绪，生理状态）而有所改变。

随着4K电视，HDR技术，multi-view, free-viewpoint video, 360视频，虚拟现实Virtual Reality，增强现实 Augmented Reality 以及混合现实 Mixed Reality 的发展，Qualinet 定义的 QoE 的概念可以无差别的直接应用于这些多媒体载体上，所以在业界被广泛采用并认定其为标准定义。

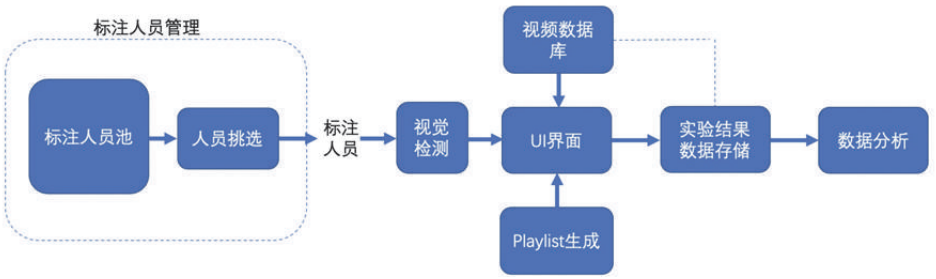
为什么要做质量评价？因为用户的观看体验永远是第一位。而在整个视频从获取，处理，压缩，传输到最后解码，增强，播放的 pipeline 中，每一个阶段视频质量的评估可以指导和优化相对应的算法实现，进而实现每一个阶段算法效果的提升，最终导致用户观看体验的提升。这是我们的终极目标。



图：摩酷实验室视频质量评价体系图

二、摩酷实验室主观测试平台

显而易见，QoE 是一个主观的感受。要去评价/测量（evaluate or measure）这个主观上的感受，需要让人去给视频打分。与 Computer vision 领域的数据标注不同，一段视频的质量不同的人给出的分不一样，在不同环境下看给的分不一样，放到电视上去看或者放到手机、平板上看质量不一样。甚至，离远了近了去看质量也不一样。为了解决这个多影响因素的问题，视频质量专家小组（VQEG）与国际电信联盟 ITU 联合致力于视频质量的标准化。在 ITU-R BT.500 等一系列的标准中，规定了测试视频质量的标准实验流程，包括人员筛选，实验环境，实验方法等（详情请参考 ITU-R BT500 文献）。摩酷实验室依据 ITU 国际标准，也搭建了自己的主观测试平台。



图：摩酷实验室主观测试流程

1. 标准测试环境

摩酷实验室搭建了符合 ITU-R BT.500 所规定的标准测试环境，如下所示：

测试环境亮度	低
背景色度 Chromaticity	D65
亮度峰值	70-250 cd/m2
显示器对比度	<=0.02
显示器背景亮度与图片亮度峰值的比	~0.15

2. 测试设备

在用户进行正式实验前，我们使用视力表，色盲检测书，立体视觉检测书等工具对用户的视觉能力进行检测并记录。对于显示设备，我们使用色彩分析仪/校准仪对显示器进行校准，利用与 imatest 类似的亮度/色彩显示范围等一系列检测工具例如 HDR targets, Arbitrary charts，xrite color chart 进行检测。从而确保实验中使用的显示设备符合规范要求。

正式实验中使用的显示设备包括手机，平板，PC 显示器，HDR 显示器，以及 OTT。以便于针对不同业务的质量评价需求进行测试。

3. 测试平台

摩酷实验室主观测试平台是基于 web 开发的多端（手机，平板，PC）可用，多用户并行可用，支持实验室标准环境以及 crowd-sourcing（网络众包）分发的测试平台。符合 ITU-R BT.500 以及 ITU-T P.910 的标准。

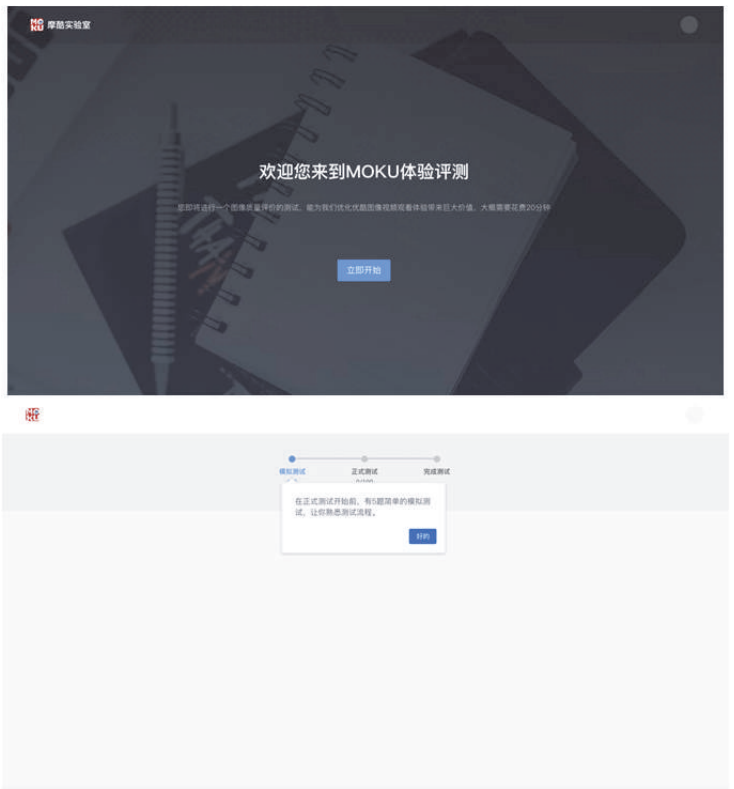


图 4: 摩酷实验室主观测试平台 PC 版界面

4. 标注人员管理

依据国际标准，标注（测试）人员全部为非视频处理领域的人员（即 naive observer)。标注人员数据，包括个人基本信息，以及参加实验的次数和实验类别全部通过数据库进行管理。保证每个标注人员在短时间内不重复参加类似的主观测试，以免产生 bias 效应。

5. 测试方法

对视频的质量进行有效的可靠的主观评价依旧是一个极具挑战的科研问题。对于不同的业务场景和实验目的，使用的评测方法需要仔细考虑才可得到较为有效的数据。ITU 在标准中针对不同需求提出了不同的测量方法 （详见 ITU-T P.910 国际标准）。最常见的为 ACR（Absolute Categorical Rating), 即给视频质量从 1-5 打分，1 代表极差，5 代表非常好。一个视频被多个观看者（一般大于 15 个）打分，最终将得到的平均分 MOS（Mean Opinion Score）作为 ground truth.

这种方法对于常见的 2D 视频来说结果比较稳定。然而试想一下，如果让一个从未感受过 VR 视频的观测者去看一段 VR 视频，让她/他依据感受从 1 到 5 打分，这个时候获得的数据可靠吗？答案是否定的。观测者很难对于多维度的视觉感受用一个绝对数值来打分。于是，配对比较法（Pair Comparison）被认为是一个针对多维度视频质量评价的可靠的方法。从生理心理学上来将，相比于给一个绝对的分数，从两个候选视频中挑出质量好的那个对我们来说更简单，因此获得的结果也就相对可靠。

本测试平台目前支持 ACR 和 Pair Comparison 两种测试方法，且 label，如：1 代表非常不舒服，或者 1 代表非常好等，可依据实验目的进行适配。

三、数据过滤

正确的主观评测方法可以有效减少数据的噪声，然而，噪声是一定存在的。因此，在业务落地中，直接拿已有的数据库去训练自己的算法模型时要认真考虑数据从何而来以及是否可以信任的问题。针对于主观测试数据去噪的问题，目前常用的两种方法为 ITU 定义的去除 outlier 的方法，以及 Li 提出的 MLE 模型。摩酷实验室主观测试平台也使用这两种方法进行数据过滤。下面对他们进行简单介绍。

1. ITU outlier 模型

ITU-R BT.500 提出了一种检测标注人员是否为 outlier 的方法，计算过程如下所示：

For each test presentation, calculate the mean, $\bar{u}_{jkr}$, standard deviation, S_{jkr}, and kurtosis coefficient, β_{2jkr}, where β_{2jkr} is given by:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \quad \text{with} \quad m_x = \frac{\sum_{i=1}^{N}(u_{ijkr} - \bar{u}_{ijkr})^x}{N} \tag{5}$$

For each observer, i, find P_i and Q_i, i.e.:

for $j, k, r = 1, 1, 1$ to J, K, R

if $2 \le \beta_{2jkr} \le 4$, then:

if $u_{ijkr} \ge \bar{u}_{jkr} + 2\,S_{jkr}$ then $P_i = P_i + 1$

if $u_{ijkr} \le \bar{u}_{jkr} - 2\,S_{jkr}$ then $Q_i = Q_i + 1$

else:

if $u_{ijkr} \ge \bar{u}_{jkr} + \sqrt{20}\,S_{jkr}$ then $P_i = P_i + 1$

$$\text{if } u_{ijkr} \le \bar{u}_{jkr} - \sqrt{20}\, S_{jkr} \quad \text{then } Q_i = Q_i + 1$$

$$\text{If} \quad \frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05 \quad \text{and} \quad \left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3 \quad \text{then reject observer } i$$

with:

N: number of observers

J: number of test conditions including the reference

K: number of test images or sequences

R: number of repetitions

L: number of test presentations (in most cases the number of presentations will be equal to $J \cdot K \cdot R$, however it is noted that some assessments may be conducted with unequal numbers of sequences for each test condition).

对于检测出来为 outlier 的标注人员，该标注人员的所有标注数据（分数）全部删除。

2. Li's MLE model

除了 ITU 规定的数据过滤方法，学术界普遍认为对于数据获取本身就很困难的情况下，将 outlier 的所有数据全部过滤掉是一种非常浪费的行为（over killing）。 Netflix 的 Li 提出了一种针对标注过程建模的方法，可以将真实分数，标注人员本身的 bias 和 inconsistency 恢复出来，即，最终只需要使用恢复出的真实分数作为该视频/图像的质量即可。标注人员的 bias 和 inconsistency 可以作为标注人员管理参考，即，对于 bias 或者 inconsistency 比较大的标注者，降低使用其标注的频率，甚至完全不再使用。

Li 的模型如下：

$$\begin{aligned} X_{e,s} &= x_e + B_{e,s} + A_{e,s}, \\ B_{e,s} &\sim \mathcal{N}(b_s, v_s^2), \\ A_{e,s} &\sim \mathcal{N}(0, a^2_{c:c(e)=c}) \end{aligned}$$

其中，$X_{e,s}$ 为标注者 s 对视频/图像 e 打的质量分。x_e 为视频/图像 e 的真实质量分数，$B_{e,s}$ 为标注者 s 对于视频/图像 e 所产生的偏差，$A_{e,s}$ 为视频/图像 e 本身产生的偏差，与 e 无关，与内容 c 有关。

$B_{e,s}$ 服从正态分布，均值为用户本身的 bias，即 b, 方差为用户的 inconsistency，即，v_s。

$A_{e,s}$ 是由视频/图像内容引起，不会改变分数的均值，只会影响方差，即，a_c

通过最大似然估计 MLE，即可将视频/图像的质量，标注人员的行为估计出来。

四、摩酷实验室客观质量评价模型

如上所述，使用主观实验方法来对视频进行质量评价是一件非常 expensive 的事情。利用 computational model 实现对视频质量的自动预测才是实际应用中可行的方法。

客观质量评价方法根据对参考视频（即，具有完美质量的视频）信息的利用程度来判断测试视频的质量而分全参考（Full Reference），部分参考（Reduced Reference）和无参考（No Reference）方法。目前效果比较好的视频质量评价模型大部分是基于全参考的，比如 NTIA 提出的 VQM，以及 Netflix 提出的 VMAF，以及较早前的 SSIM 等。无参考的评价方法一直是该领域的难点，然而随着近年来直播视频，小视频的火爆流行（属于无参考的范畴），无参考的视频质量评价将是未来该领域的重点研究方向。

客观质量评价方法根据它本身的实现方式也可以分为以下几类：1）早期的依据人类视觉系统（Human Visual System）特性建模的方法。该方法直接将人类视觉系统参与到质量评价过程的机制用数学建模的方式模拟出来，例如将人眼的多通道特性，Contrast Sensitivity Function, Luminance Adaptation, Masking Effect，Pooling 等按顺序连接最终实现对质量的评价。该方法的难点在于对视觉系统特性的建模，但是一旦建好，方法可以适用于不同类型的视频；2）基于人类视觉系统提取 hand-crafted feature，最后做 regression 的方法；3）基于 CNN 的方法，需要大量的数据训练（而视频质量数据库由于主观实验本身的限制决定了库不可能巨大），泛化性较差；4）其他方法。

摩酷实验室针对优酷视频业务，提出了自己的图片和视频质量评价模型。

1. 图片质量评价

从用户进入到优酷 app 的观看逻辑来看，首先进入用户视线的便是短小视频的封面图。封面图画质是否足够清晰是导致用户是否愿意点击观看的重要因素之一。其次，用户进入 app 所看到的封面图的质量也一定程度上影响着用户对于优酷 app 的视觉体验定位的判断。鉴于人工审核不可能覆盖到全站的所有图片量级，会被用户看到的封面图的质量甚至首页的封面图质量都无法做到高标准。因此我们针对 UPGC 封面图画面质量建立了一套评价体系来解决这个问题。

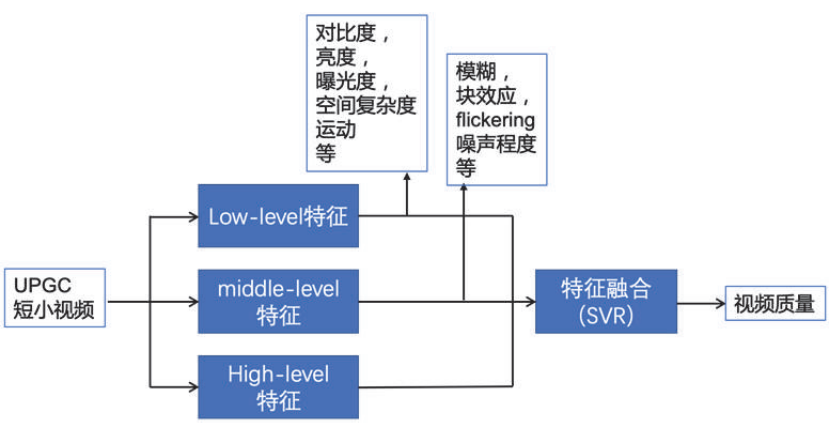
从优酷线上随机抽取 5000 余张视频封面图，经过系统抽样涵盖几十个视频种类，并考虑到不同分辨率，10 多种质量影响因子的均匀分布，以及总体质量分的均匀分布，最后共得到 750 张封面图作为训练数据。经过近 200 人在手机上进行主观评测实验得到平均质量分数作为模型的 ground truth。

由于线上封面图的快速生产要求，算法模型必须简单高效，因此我们采用基于 VGG 的网络训练模型，经过优化迭代最终达到预测分数与主观 ground truth 分数线上 PLCC=0.87，SROCC=0.86 的效果。目前算法模型已经应用于封面图筛选及封面图增强优化等业务。

2. 视频质量评价

随着近年来用户在社交媒体/短视频分享平台的重度参与，视频多媒体内容的消费主导从原来的 OGC 向 UPGC 转变。因此 UPGC 视频的质量评价成为了目前学术界重点关注的研究热点。由于其没有参考视频（即认为是完美的），研究课题本身即属于质量评价领域最具挑战的无参考质量评价。除此外，社交媒体/短视频分享平台等的视频来源十分复杂且多样，导致质量问题的根源可以追溯到整个视频从拍摄到传输到播放的整个 pipeline，由此带来的视频降质的因素为多种失真的叠加。这种对于失真类型的不可控（从失真类型到失真程度两个维度）导致 UPGC 视频的质量评价更具挑战性。

为解决上述问题，我们提出了一种基于 multi-level 特征融合的无参考质量评价框架。具体来讲，通过对 low-level（包括对比度，亮度等特征）, middle-level（包括模糊，块效应等失真）, high-level （各种画面质量评价算法得到的结果）的视频质量 features 的融合，可以更好的解决 UPGC 视频中失真来源复杂引发的质量评价难点。该方案的框架图如下所示（已申请专利）：



图：基于 multi-level 特征融合的无参考质量评价框架

该方法不仅输出总体质量分，还可以输出失真类型，因此针对于优酷视频业务，可以应用于视频质量监控，搜索推荐，以及视频增强算法的优化（针对失真类型进行有效处理）上。

五、5G 下未来多媒体质量评价的展望

5G 的到来势必颠覆目前用户的观看习惯和观看体验。目前已经出现的新型多媒体技术，比如 Light-field Imaging, AR, VR, 360 VR, MR, High Dynamic Range (HDR), Free-viewpoint video, 以及 Autostereoscopic 3D 将会是未来 5G 时代的主流。以提高用户多维度的感知体验为目的下一代视频内容生成，视频压缩，视频增强，depth estimation, view synthesis 等技术势必需要质量评价方法来做监控。同时，这其中有可能产生的会引发观众视觉疲劳等危害身体健康的视频更需要质量评价方法去做前期评估预警。

参考文献

[1] Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Patrick Le Callet, Sebastian Möller and Andrew Perkis, eds., Lausanne, Switzerland, Version 1.2, March 2013.

[2] ITU-R BT.500 : Methodology for the subjective assessment of the quality of television pictures

[3] ITU-T Rec. P.910, Subjective video quality assessment methods for multimedia applications.

[4] Margaret H. Pinson, Lark Kwon Choi, and Alan Conrad Bovik, “Temporal Video Quality Model Accounting for Variable Frame Delay Distortions (VQM-VFD)”, IEEE Trans. on Broadcasting, Vol. 60, No. 4, Decemebr 2014.

[5] Zhi Li, et.al, “Toward A Practical Perceptual Video Quality Metric” (VMAF), Netflix Tech Blog, 2016.

[6] Wang, Zhou, Bovik, A.C., Sheikh, H.R., Simoncelli, E.P. (2004-04-01). "Image quality assessment: from error visibility to structural similarity". *IEEE Transactions on Image Processing*. **13** (4): 600–612.

[7] 李静，王百超，周星光，一种基于 multi-level 特征融合的无参考图像质量评价框架， 已提交专利申请（提案号 101366206），2019

[8] Jing Li, Marcus Barkowsky, Patrick Le Callet, “Visual discomfort of stereoscopic videos: influence of motion”, Displays, vol.35, no.1, pp. 49-57, 2014.

[9] Z Li, CG Bampis, Recover subjective quality scores from noisy measurements, Data Compression Conference (DCC), 2017

端侧智能算法在优酷场景的应用——面向多种业务场景的统一端侧渲染 SDK

作者| 阿里文娱高级算法专家 王百超、阿里文娱算法专家 石海华 刘国友

一、业务背景

作为综合性的视频平台，优酷拥有完整且多样的视频内容形式，包括长视频、短视频、小视频，面向体育和秀场的直播平台、此外还有互动剧。总结起来我们主要面向视频的拍摄、编辑和播放开展业务。

放眼业界，对于拍摄，有大量的 APP 是围绕相机做文章的，美图秀秀、无他相机、轻颜相机等，且隔三差五会有新的爆款出来，比如 ZAO。对于视频编辑，抖音和快手都发布了自己的剪辑工具，剪映和快影，俨然成为短视频竞争的第二战场。在播放端，也可以结合 AI 技术，优化视频画质，或将内容升级为互动视频。

二、端侧渲染引擎功能和框架

1. 设计思路

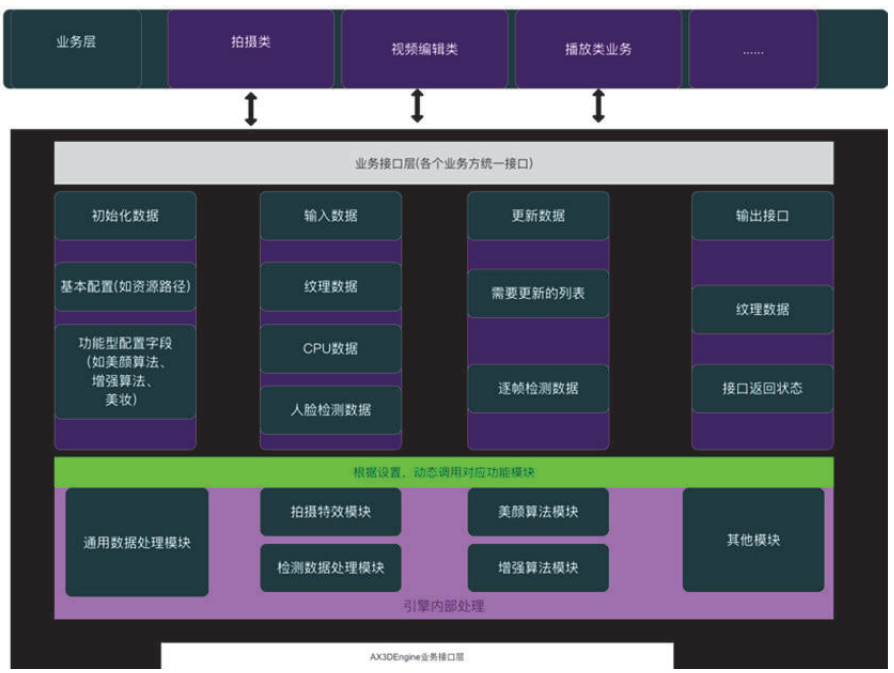
在长期对接拍摄、视频编辑、智能播放器等业务的过程中，摩酷实验室沉淀了可同时支撑多个业务场景的端侧渲染引擎——AX3D 引擎。

考虑到手机端的实际应用场景，在规划引擎的功能模块时我们采取了非常审慎的态度，时刻对焦业务主线，避免无意义的低频功能的开发。另外，在开发过程中避免引入开源引擎，虽然引入开源引擎可以加快开发进度，但更会导致引擎臃肿，bug 横飞，所以我们坚持独立开发。由于使用 C++开发，引擎做到了跨平台（PC、安卓、iOS），我们的渲染引擎还具有包小，速度快，稳定性好等优势。已经在优酷主客拍摄、云相册、播放器等场景得到应用，基本**覆盖了视频从生产，到编辑，再到播放的完整生命周期**。

在功能方面，得益于优酷丰富的内容形式，我们针对不同的内容形式研发了不同的功能点和特效。应用于 PUGC 短小视频、直播、社交等场景的互动类特效，包括美颜、滤镜、人脸贴纸等。应用于 O 周场景的编辑类特效，包括转场、动效、文字贴纸等。应用于播放器的真实感人脸美化和图像增强。

从架构角度，目前多种特效场景共用底层渲染引擎，能力得到最大程度的复用，同时能够快速灵活地支撑更多场景。端侧识别模型方面，阿里巴巴已经有许多团队研发了多种多样的端侧识别模型，还有非常成熟可靠的 MNN 等推理引擎，端侧有大量的模型可以复用。这使得我们可以站在巨人的肩膀上，聚焦在渲染引擎上，面向我们的业务，持续把渲染做深做透。接入多种 AI 识别能力，支持各种围绕人脸、人体的互动和编辑形式。

2. 业务结构图



从上面的业务结构图可以看到，接口层（Interface）是对外业务的输入/输出接口，通用的输入数据为 GPU 纹理数据和类型设置选项。识别信号是当作外部输入传入的，比如为了实现美颜、美妆等人脸检测相关功能，我们定义了 landmark 检测结果数据的通用结构体，这样即使是人脸检测算法结果输出方检测结果数据结构体不一致（如使用 MNN 或商汤的检测结果），但通过设置数据给通用结构体数据，可以实现数据的透传。

针对业务方变更、添加新的需求，引擎可简单、快速的增加注册新功能模块，快速将算法实现应用到业务中。例如添加端侧增强算法、美颜算法功能，只需要编写相对独立的功能模块，然后以配置的形式注册到引擎中，就可以实现相应的功能了。

3. 应用

1）美颜

2）滤镜

3）云相册视频编辑

模版中包含转场、帧动画、叠加特效等

三、重点算法

1. 真实感人脸美化

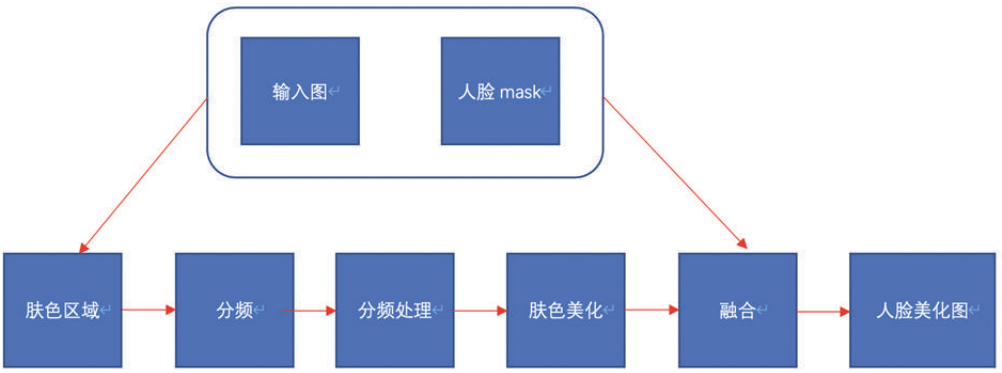
围绕人脸的美化和增强是业界关注的问题，我们将美颜和人脸的去噪、细节恢复统称为人脸美化。美颜技术行业应用非常普遍，在快手抖音等短小视频场景，直播场景，拍摄工具等都有大量应用。一般在拍摄或上传入口都需要美颜，且对算法实时性要求很高。而播放端涉及到大量的视频噪声、压缩等降质问题，会破坏人脸的细节和纹理，针对人脸图像做恢复和增强也很有必要。基于此我们提出统一的人脸美化工具包。

优酷 UPGC 场景对美颜和画质增强有特别的需求，强调人脸的真实感和肤色保持。对于播放器场景，如果人脸图像存在降质问题，需要前置使用人脸 SR 模型（人脸 SR 模型见我们的另一篇文章“分区域处理的图像和视频清晰化技术”），进行人脸图像基础画质的恢复和增强，先恢复已被破坏的纹理，并去除噪声，然后使用美颜技术进行修饰。对于这样复杂的业务需求，传统的美颜技术很难支持。因为传统的美颜技术往往把皮肤磨的非常光滑，并加入大程度的美白，极易产生失真和肤色变化的问题。事实上，我们对算法的要求已经超越了美颜的范畴，既要恢复细节，又不能有 artifacts，要控制和调整纹理的粗细，保持真实感和肤色，我们称之为真实感人脸美化。

我们根据业务需求，运用分区域和分频技术的来达到一种无磨皮感美颜的效果。 算法的主

要思路是：对图片进行分频处理，在不同频段采集不同处理方式，同时结合人脸肤色的 mask 图，细腻的控制人脸肤色的各个区域，达到去除痘渍，光滑皮肤，同时保持皮肤真实感的效果。

美颜处理流程：



1）分区域 mask 图的获取算法

分区域算法主要根据人脸检测算法先检测出人脸的区域，然后在根据肤色检测检测出人脸区域的皮肤区域，进而为后续的分频处理提供了 mask 处理区域。

2）分频处理算法

无磨皮美颜的核心算法包括分频技术和对特定频率段（皮肤痘渍）的处理。分频技术需要能从高层次分离正常皮肤和痘渍。对特定频率段的特殊处理要能保证处理后的效果不过渡均匀，不引起异常跳变，频率反转的问题。

分频技术是图像处理非常常用的一项技术，在各种图像增强的应用场景里面都得到了很多应用。分频的方法也有很多，比如小波技术，保边滤波算法等方式。考虑到性能的问题，我们利用 fast guided filter 来实现分频的效果。

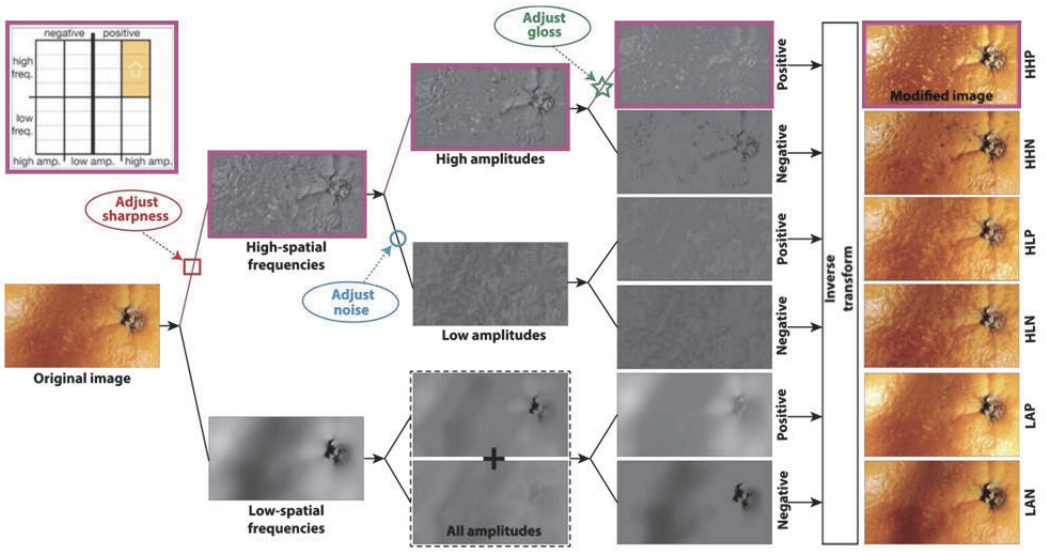
频率处理技术，对特定频率的处理部分我们主要三种操作：

scale：整体放大信号，可达到过滤信号的作用；

Amplitude：调整振幅；

Sign：改变符号。

如下图所示，有了上面的分频工具和分频处理操作，我们可以达到对图片纹理细节控制。



通过对大量人脸图像的实验，我们发现脸部皮肤的痘渍之类的不干净的东西基本处于高频区域信号里且频率符号为负。利用实验结论，我们确定了美颜算法需要处理的频率段和相应的处理系数，最终达到了最大肤质保持情况下美颜效果。

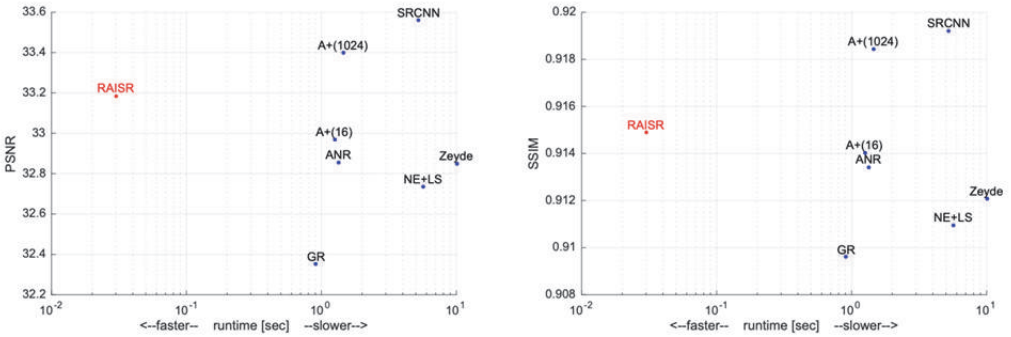
3）融合技术

最后我们需要人脸区域处理完结果和原图进行融合，这里面涉及的一问题就是要处理好边缘的过渡，我们通过 fast guided fitlter 羽化过渡的边缘区域，完美的把人脸区域图和原图就行融合。

2. 端侧视频增强优化

超分算法目主要算法都是基于深度学习的算法，但是深度学习虽然是目前学术的趋势，但是深度学习算法一般网络参数繁多，模型比较大。且需要大规模的数据才能得到较好的数据。这样的特点让深度学习算法增强算法具有迭代行差和性能差的弱点。如果想用深度学习还要集合模型压缩的技术，可想而知这是漫长的。

谷歌的 RAISR（Rapid and Accurate Image Super-Resolution）算法，利用机器学习将低分辨率图像转化为高分辨率图像。这项技术能够在节省带宽 75%的情况下分辨率效果达到甚至超过原图，同时速度能够提升大约 10 到 100 倍。该算法速度超快，比目前流行的深度学习网络算法速度提升了很多，从下图可以看出 RIASR 算法和一些经典的超分算法对比。



(a) PSNR, 2× upscaling　　(b) SSIM, 2× upscaling

由于 RAISR 算法优秀性能和效果特点，我们在端侧超清化处理的技术上选择 RIASR 算法作为基础。

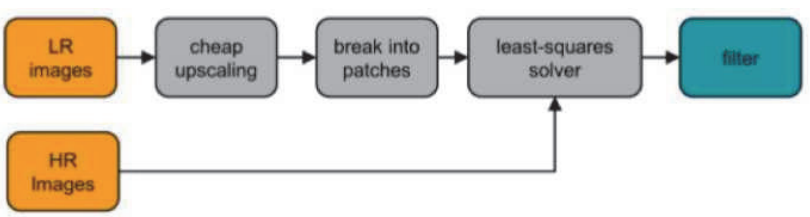
1）算法基本原理

a）学习滤波器

学习滤波器的过程就是学习一种高清映射的关系。给定一些图像对，用最小化恢复出来的图像和高清图像质检 的误差的方法，学习预设的滤波器。常用的 least-square 损失函数可写为[2]：

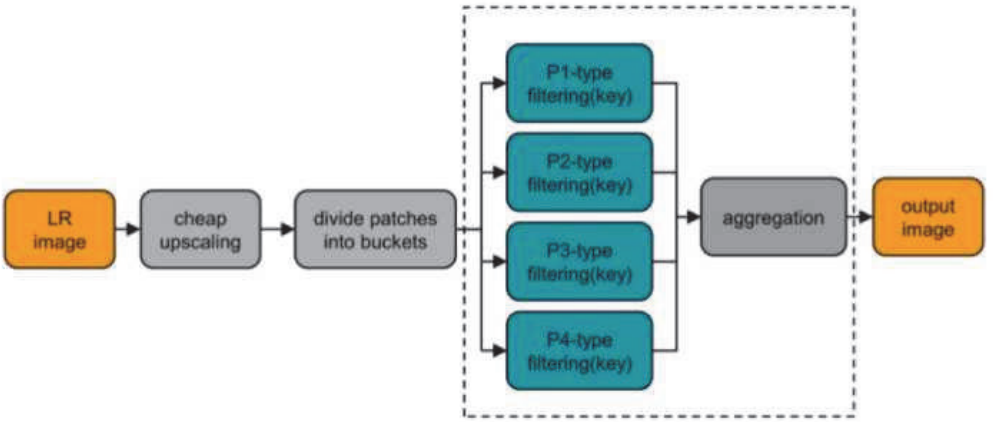
$$\min_{\mathbf{h}} \sum_{i=1}^{L} \|\mathbf{A}_i\mathbf{h} - \mathbf{b}_i\|_2^2$$

其中，h 是我们要求的滤波器，A 是从高清图像中扣取的小图像块，b 是这个小图像块对用的低清像素块。这个流程可以用下图[2]来直观表示。



b）低分辨到高分辨图像预测流程

综合以上步骤，RAISR 的流程可以概括为下图[2]：



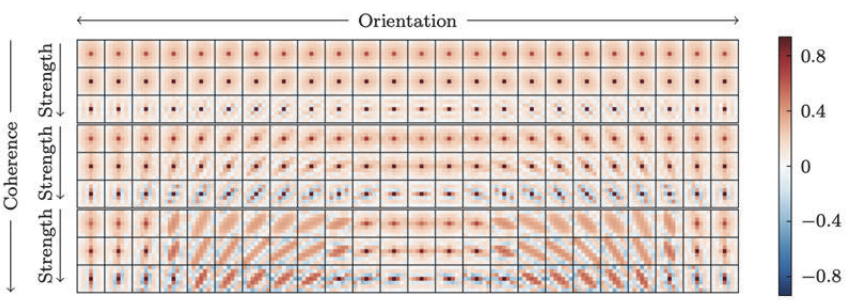
首先，对低清图像进行简单的双线性差值；然后，使用哈希算法快速将图像块分到不同的类别（bucket）中；对于每个类别，分别使用四个预先训练好的滤波器进行线性滤波；将不同的图像块的结果融合起来，得到最终的恢复结果。

2）性能优化

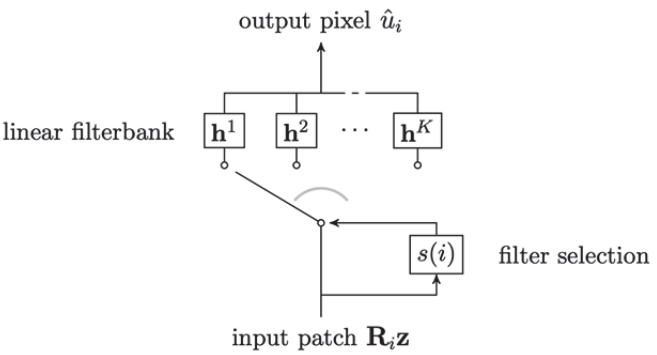
虽然 RIASR 算法本身性能超高，但是要想发挥算法的性能优势，也需要结合相应的并行优化技术。目前大多数的中高端手机都配备 GPU，且有相应的开发语言。如 IOS metal 语言，Android 支持 openCL 或 valkan 语言。因此我们结合手机特定的并行处理技术设计 RIASR 的算法计算流程，最大化并行化各个处理模块。

3）算法可调性优化

RIASR 通过一组训练数据来可以到一个滤波器，达到一种固定的增强效果。但这个算法天生的缺陷就是不具备可调整性，也就是学习完一组参数只能得到一种程度的增强效果。下图[2]RIASR 一组滤波的可视化展示：



但在实际业务场景中，往往需要快速调整算法的强弱程度。解决这个问题最简单的办法学习多组滤波器，根据需求按一定的策略动态选择不同的滤波器，如下图[2]所示：



但这样的做的缺点也显而易见。大大增加我们的工作，降低了算法的快速响应业务需求的能力，显然不可取。我们通过分析分析滤波器的特点，利用权重插值的办法来解决这个问题。利用一个最大程度，中间程度，最小程度的 filer 权重组，插值出中间态的效果。既减少了学习多组滤波器的工作量，又满足了算法可调性的需求。

4）训练数据

RIASR 算法不像深度学习方法需要那么多数据，且训练时间较短，一般几个小时就可以完成。但想要取得比较好的效果就需要数据特别贴合业务场景，不同业务场景需要不同的数据。因此我们设计一套数据模拟的方案。利用 Gan 方法快速学习生成各种降质数据，且接近真实场景。这样就使 RIASR 具备了快速适应各种场景的能力。

参考文献

[1] Pascal Getreuer, Ignacio Garcia-Dorado, John Isidoro, Sungjoon Choi, Frank Ong, and Peyman Milanfar, "BLADE: Filter Learning for General Purpose Computational Photography", *2018 International Conference on Computational Photography (Oral)*

[2] Y. Romano, J. Isidoro, and P. Milanfar “ RAISR: Rapid and Accurate Image Super-Resolution ” *IEEE Transactions on Computational Imaging*, vol. 3, no.1, Jan. 2017, pp. 110-125

[3] I. Boyadzhiev, K. Bala, S. Paris, and E. Adelson. Bandsifting decomposition for image-based material editing. ACM TOG, 34(5):163:1–163:16, Nov. 2015

大千 XR-Video 技术概述

作者| 阿里文娱高级算法专家 方如、
阿里文娱高级算法工程师 时镇、阿里文娱算法专家 姜岑

大千世界无奇不有，大千意味着创意无限，其二有汉语拼音 DaQian，谐音大钱，讨个好彩头，其三缩写 DQ 也很酷。而 XR，即 X Reality，俗称各种现实。长久以来，以 VR（Virtual Reality：虚拟现实）、AR（Augmented Reality：增强现实）和 MR（Mixed Reality：混合现实）为代表的"XR"技术备受世人瞩目。加入优酷以来，我们一直琢磨如何把 XR 的理念和方法应用到优酷视频中，打造具有特色的特效视频植入技术 — "大千 XR-Video"。大千 XR Video 特色在于(1)虚拟信息与视频在后期富有创意地植入；(2)通过 3D 视觉技术实现时空多维度地准确合成；(3)具有实时可交互性。本文结合优酷在植入特效广告场景上的应用来解开 XR-Video 技术背后神秘的面纱。

一、虚拟信息与视频后期植入

视频植入技术通过复用已有的普通视频素材，在视频制作后期植入新的内容，目的是在原视频中加载和扩充需要传递的视觉信息。XR-Video 已支持了高光时刻、跃享时刻、移花接木、动态混合现实、拍照特效、心动时刻、爆石特效和背景氛围等十几种特效，用户看到我们的特效的时候总会发出“喔”的感叹。

1. 植入内容

植入新的内容包括用户提供的和 AI 算法自动产生的两大类。用户提供的内容除了第三方提供的文本、动图、视频和 3D 模型等多媒体素材，还有 XR-Video 系统自带的丰富的特效库。AI 算法利用视频理解和视频分割等技术从已有的视频中通过算法模型计算生成相关的文本和图像。图（a）中 logo 是广告主提供的，粒子效果是特效库产生的，通过人体目标检测和人像分割算法确定主体轮廓，然后特效粒子对人像进行 3D 环绕，形成酷炫的效果。图（b）人体复刻效果通过人体目标检测和人像分割算法把人像抠出来再复制合成。图（c）是大千云渲染引擎与阿里体育合作的一个子弹时刻特效 Demo，展现了 XR-Video 在体育赛事上的应用场景，未来摩酷实验室将进一步提升自动化程度，通过一些 CV、AI 算法技术结合渲染在准实时的体育赛事上展现更多新颖的数据特效表达形式。

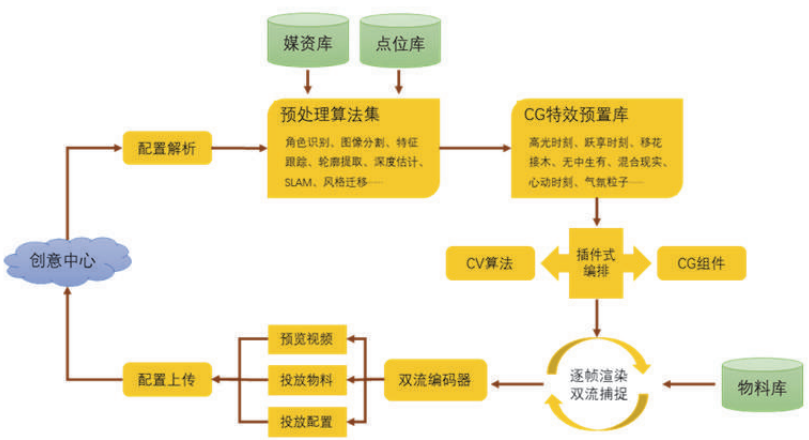
图(a) 高光时刻 demo

图(b) 人物复刻互动 demo

图(c) 体育比赛 360°视频

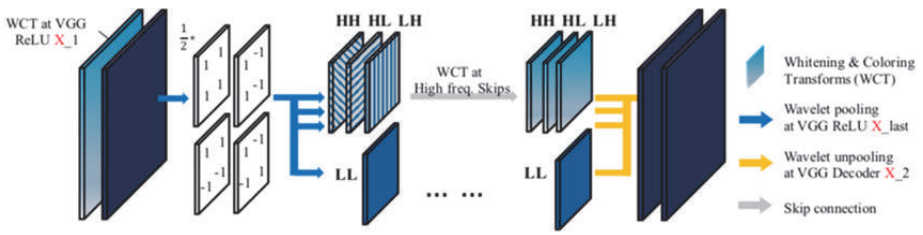
2. 云渲染

支撑特效视频制作的是大千云端渲染引擎。云端渲染引擎旨在解决目前特效制作的规模化和自动化，侧重解决效率和品质难点。以植入特效广告为例，云端渲染引擎从创意中心下单到自动化制作输出投放配置，为特效广告快速上线提供保障，丰富的预置模板和插件式编排能力为广告客户提供优质视觉曝光方案。传统特效广告制作存在渲染时间长、人工介入流程多、场景变动需要重新制作等耗时问题，我们采用实时渲染+双流捕捉技术实现了快速特效渲染，同时结合 CV 算法使视觉参数自动适配视频场景，达到了模板编排、一键生成、快速微调的高效制作链路。另外常规的影视包装技术比较匮乏，也难以与 CV 算法结合进行创新，因此我们采用开放式 CG 方案，将物理计算、粒子系统、光影渲染等 CG 技术进行插件式配置，从而可以灵活地与 CV 算法结合创造令人耳目一新的特效新形态。



云渲染架构

植入渲染是植入虚拟信息和视频内容融合的关键，直接影响到用户的观感。为了使得植入无违和感，需要在把植入位的图像风格迁移到待植入的素材图像，使得植入后的素材区域和原始视频的整体图像风格一致。以植入特效广告为例，目前方案大都是使用颜色迁移、亮度迁移、模糊迁移等方法对广告素材进行处理，传统的图像处理方法很难自动的准确估计出原始视频的图像风格，需要人工介入调整大量的参数，且操作人员要求有较高的经验才能调整得到较好的效果。我们引入深度学习方法结合 Wavelet Transforms，提取植入区域图像的风格纹理特征，并对素材图像风格纹理进行迁移，达到了植入后的广告位自然无违和感。



网络框图

植入效果如下图所示：

原片　　素材植入后

二、时空多维度合成

大千 XR-Video 另一个特色是从时空多维度解析视频，近自动地搜寻和定位植入区域，数据生产、标注项目生成、点位结果入库和特效制作都能在线上完成。时空多维度中的“时”指的是对视频进行人物、动作、场景、物体、精彩时刻、BGM 和平面等多方面检测，确定特效开始和结束的时间点，精确度到帧级别。时空多维度中的“空”指的是在一帧或多帧图像中连续的植入位置。

在时的基础上，融入“空”的感知与理解，由于视频包含的信息不同，对植入空间确定的方法也不同。如果是 360 度视频，数据采集由多个摄像机同时拍摄完成，每个摄像机内参已知，通过传统 SFM 方法/深度学习的方法进行深度估计、三维平面检测和重建，图像和点云结合利用 PointNet 系列的目标检测模型实现场景理解；如果是已有的影剧综，视频拍摄时的相机内参我们是不知道的，需结合物体检测、几何形状检测、深度估计和基于图的平面追踪等多种方法实现平面检测和追踪；如果是优酷自制的影剧综，我们能介入前期的拍摄，例如优酷自制的互动剧，我们可以利用标定板铺助进行相机位姿估计，实现平面检测和重建，叠加虚拟场景改变画面效果，推动剧情发展。

移花接木、无中生有和动态混合现实等特效制作就是利用了时空多维度合成。下面以移花接木为例阐述一下相机内参未知的情况下的平面植入位置检测和平面追踪。

1. 显式平面检测

平面植入位置检测是规模化自动化的关键。镜头分割在线抽帧后，对序列帧图片进行显式/隐式的平面检测。显式平面检测的步骤包括边缘检测、直线拟合、轮廓提取、筛选和精确查找起止帧、以及生成植入点位信息等。边缘检测有 Hough 变换/Canny / EDLines 等，但采用 sobel，canny 等算子与霍夫变换直线检测，需要较多的参数调节， 且计算代价较大，方法相对复杂，有些场景仍然不能达到要求。 EDLines 作为一个不需要后处理，无调节参数的直线检测方法，从实际效果来看其鲁棒性强于 sobel，canny 等算子。我们采取基于改进的 EDLines 做边缘检测。

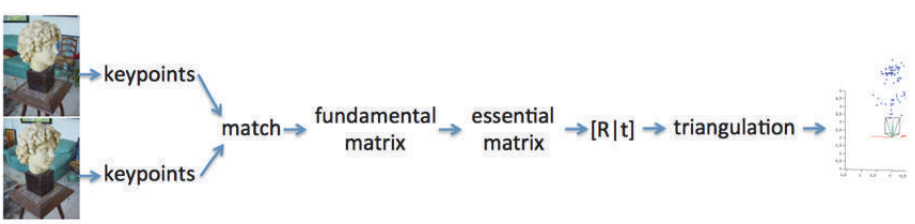
原图　　Canny　　EDLines

轮廓提取筛选是对边缘检测结果进行轮廓拟合，按照周长、面积、顶点数目、长宽比和内角等几何条件筛选出符合一定要求的轮廓，再按照相邻帧四边形出现位置、符合要求的四边形连续出现次数来合并间断检测片段。接下来，在一个镜头内，对检测到的视频片段，向前向后查找，弥补漏检的图像帧。初步检测结果是：在视频中显著的平面四边形区域检测准确，对运动和遮挡有一定的适应能力，漏检主要是边缘检测后寻找轮廓的错误所致。有了显式平面检测初版的结果，我们通过边缘检测和轮廓提取辅助人工标注，收集了一批平面检测数据，利用深

度学习进一步提升平面检测的准召率。

2. 隐式平面检测

隐式平面检测是为了进一步挖掘点位信息。如果视频场景内无运动的目标且仅有平移旋转的视频序列，我们尝试了传统点云重建的方法。点云是某个坐标系下的点的数据集，点包含了丰富的信息，包括三维坐标 X，Y，Z、颜色等，可以说万物皆点云。通过高精度的点云数据可以还原现实世界，当然这也是我们的梦想。这里我们简单描述一下利用摄像机成像原理从图像序列获取物体三维点云模型的方法。首先对 2 幅或多幅图片序列进行特征点提取与匹配，利用外极几何原理通过直接线性变换法（8 点法+最小二乘法）对基础矩阵 F 进行估计进而得到本征矩阵 E，在计算过程中采用 RANSAC 算法消除错误匹配的的点。本征矩阵 E 进行 SVD 分解获得 R 和 T 矩阵。知道了两个相机之间的变换矩阵(R 和 T)，还有每一对匹配点的坐标，通过这些已知信息还原匹配点在空间当中的坐标，将三维点三角化并重映射到摄像机得到二维点，同时算出这两帧图像所对应的相对相机姿态，通过相机位置就可以恢复物体稀疏三维点云，再通过 CMVS/PMVS 重建稠密三维点云。针对原始的 PMVS 算法在扩展面片时由于检测到的特征点检测稠密分布不均，容易导致重建结果出现部分区域空洞的问题，提出了一种自适应优化特征点检测的三维重建改进。先将 Harris 算子分 patch 计算响应值，根据 patch 中的特征点个数自适应的选择 Harris 算子的阀值，保证每个 patch 中有数量近似的特征点个数，减少了特征点的集群现象，改善了三维重建的结果容易出现大量空洞的缺陷。



从二维图片（序列）恢复稀疏点云

如果视频中特征点较少，存在运动目标的场景，传统方法效果比较差。针对这种情况我们推出了利用图像深度估平面检测的方案：

1）使用 CNN 估计图像深度信息，重建 3D 坐标。

LossFunction：

$$D(y,y^*) = \frac{1}{2n^2}\sum_{i,j}\left(logy_i - logy_j - \left(logy_i^* - logy_j^*\right)\right)^2$$

2）SLIC 图像超像素分割得到 cluster。

SLIC 即 simple linear iterative clustering。分簇的依据是像素之间的颜色相似性与邻近性。其中颜色相似性的度量因子是 lab 颜色空间的 L2 范数，颜色邻近性的度量因子是图像二维坐标空间 xy。因而综合的度量因子是[labxy]五维空间。下面所述的距离度量因子由下式计算得到：（其中 Ns 与 Nc 分别是距离与颜色的权重）

$$d_c = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2}$$

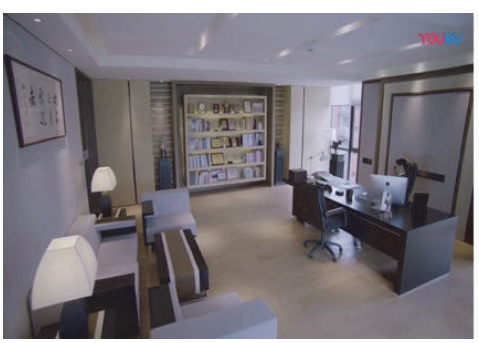
$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

$$D' = \sqrt{\left(\frac{d_c}{N_c}\right)^2 + \left(\frac{d_s}{N_s}\right)^2}$$

3）对于每一个 cluster，根据 cluster 中的 3D 点拟合出一个平面并求出一个法向量。

4）根据超像素分割的结果建立一个邻接矩阵判断 BlockIndex 之间是否相邻。

5）根据 BFS 算法遍历邻接矩阵，通过每个 cluster 拟合平面法向量的夹角余弦来判断相邻 cluster 之间是否共面。



原片

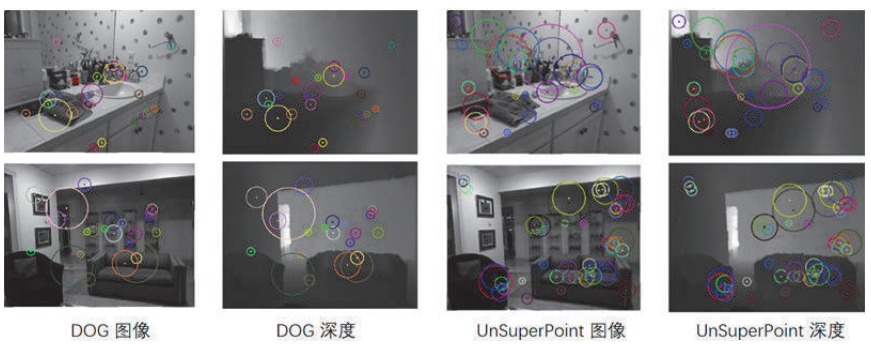
深度图

3. 平面追踪

平面追踪是移花接木植入的关键环节之一。追踪待植入区域，使植入区域在视频运动中仍

然可以保持与画面运动的同步。平面追踪大体有基于特征点的、基于区域的和 generic object（KCF）几种。我们采用基于特征点的多融合跟踪方案，引入图模型和图匹配，结合 H 矩阵平滑来提高计算单应性矩阵的准确性。特征点计算传统方法有 SIFT、SURF、KAZE、AKAZE、BRISK 和 ORB 等，Learning-based 方法有 D2-Net、DELF、LF-Net 和 SuperPoint 等。Learning-based 方法取代传统基于 SIFT 的匹配是一个大趋势。

整体来讲，UnSuperPoint 框架从效果和数据训练方面更符合我们的需求。UnSuperPoint 延续了 SuperPoint 的框架，使用 self-supervised 方法，多任务网络同时估计关键点和描述子，利用 homography transformation 建立对应关系用于训练。在多融合方案中，我们将推进利用类似 UnSuperPoint 深度学习方法进行关键点和描述子联合的自适应学习，保证特征点和描述子的稳定性（reliability）和重复性（repeatability），提升在植入平面运动幅度大和植入平面旋转的情况下的植入稳定性。



移花接木效果对比

三、实时交互

交互从简单的人面对屏幕观看视频发展到将 2D/3D 信息融合于周围的空间与对象中，不再与视频内容脱离，而是和人们的当前视频自然而然地成为一体。交互的动作除了以往的按键或

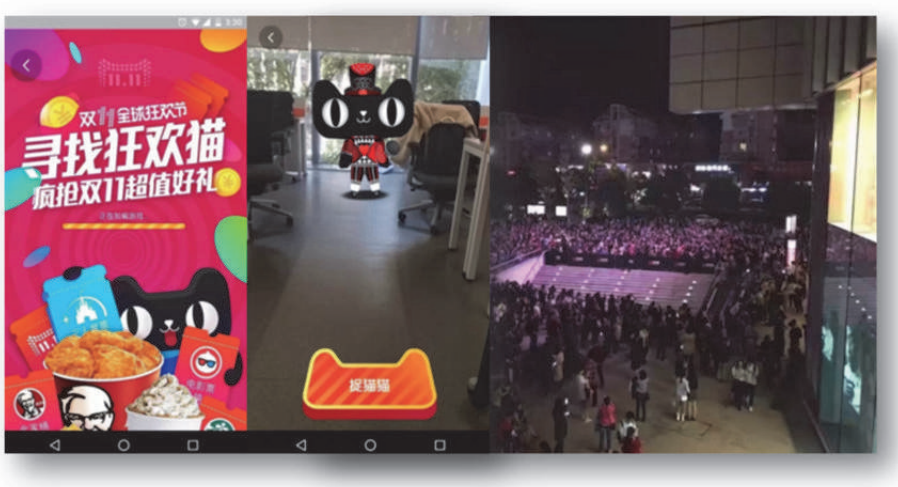
者触屏，可以扩展到头部、眼部、表情、手势和语音等，从位置扩展到原有视频某个空间。下面分享一下我们在实时交互中的体会。

1. 点哪儿 活哪儿

汽车 3D 互动广告 demo

首先，让我们认识一下 3D 形式互动。例如在视频广告中，我们可以在出现保时捷品牌汽车的点位进行预埋点，通过特效触发召唤出汽车模型，用户可以与汽车模型进行三维触控互动，模型可动态展现品牌汽车的各个角度以及开关门、开关灯等各种行车效果，这种 3D 互动式广告可以大幅增强广告的品牌感知度和认可度。

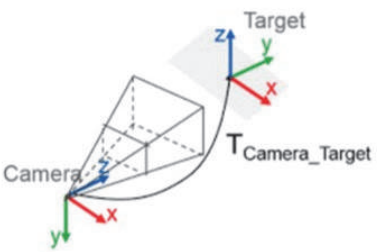
2. 转哪儿 看哪儿



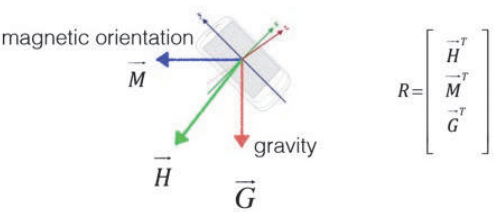
AR 捉猫猫手机截图和活动现场的人群

2016 年双 11 AR 捉猫猫游戏上线，是 LBS+AR 技术的一种成功运用。游戏活动期间总 PV16 亿，日均 PV 3 亿多，UV3100 万，支持星巴克、KFC、苏宁易购等 60 多款品牌猫，是那年最火的双 11 预热互动活动。其主要的交互方式就是转动手机，这种方式将应用在手机观赏体育比赛和综艺节目等场景上。

我们要解决的技术主要问题就是利用智能手机传感器和 GPS 获取信息，实时计算出 3D 模型在屏幕上的显示位置，给用户一种该 3D 模型（例如星巴克猫）就在其真实世界周围的某个方位上的“错觉”。这个“错觉”的视线方向通常表示为一个旋转矩阵。



视锥体



手机惯性测量单元（IMU）

智能手机中跟手机位恣相关的传感器有加速度计、磁力计、陀螺仪。加速度计可以感知加速度大小，磁力计感知磁场的方向和大小，陀螺仪能够计算角速度，即转动速度。智能手机在上述三种基本传感器之上，进一步计算 orientation（即欧拉角）和 rotation vector（即四元数）等。以四元数形式输出的结果就是利用卡尔曼滤波算法综合使用加速度、磁力计、陀螺仪得到的。利用四元数和 Rodrigues' rotation formula 推导出以下公式便可计算旋转矩阵：

$$R(\mathbf{q}) = \begin{bmatrix} 1-2(y^2+z^2) & 2(xy-zw) & 2(xz+yw) \\ 2(xy+zw) & 1-2(x^2+z^2) & 2(yz-xw) \\ 2(xz-yw) & 2(yz+xw) & 1-2(x^2+y^2) \end{bmatrix}$$

虽然四元数可以求出旋转矩阵，在手机上计算出准确的旋转矩阵没有那么简单。廉价的 Android 机器没有陀螺仪，有部分机器虽然检测到有陀螺仪，但实际上陀螺仪输出的数据有问题。针对这个问题，我们采用分级策略来解决：默认使用加速度和磁力计的直接计算方案，使用低通滤波算法降低抖动；对测试过证明陀螺仪正常的机器或者市面上的高端旗舰机，使用稳定的四元数方案；如果没有加速度计或者磁力计，直接返回一个默认旋转矩阵（3D 模型默认显示在屏幕中央）。

3. 看哪儿 买哪儿

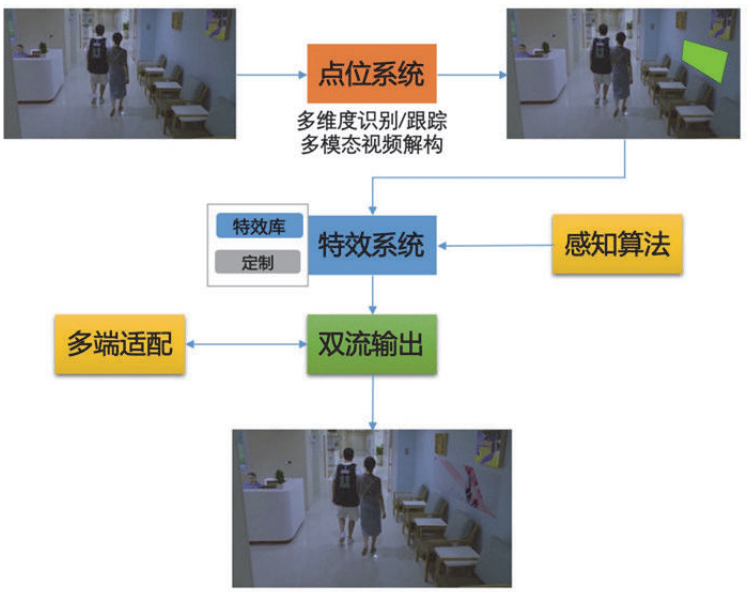
Buy+ 沉浸式购物（全景视频版）

在 VR/AR 中通过空间定位，人置身其中，参与其中的互动，犹如身临其境一般。在阿里推广 VR/AR 技术时，淘宝和优酷身边都有一批志同道合的同行者，Buy+ 就是大家一起打造的品牌。VR Buy+是世界上首次大规模沉浸式用户购物体验，除了 HTC Vive VR 版外，2016 年双 11 还发布了手机版。因为成本、时间和受众用户的考虑，采用了全景视频+手淘+Cardboard 的方案。在 3 个国家取了 7 个场景，这 7 个场景和 Buy+产品有机地融合。用 VR 手机盒子体验的购物应用，带你穿越到世界各地的商场购物，遇到喜欢的还可以直接线上下单。

交互方式主要只有一种，就是盯住触发按钮。虽然 Cardboard 上有点击按键，更建议用自然的人机交互方式。有一点需要大家知道，在全景视频中，用户停下来时商品总是能标定得非常准确是非常不容易的。经过多个方案对比，最终采用了空间移动方案，把一个视频拍完以后，转成一个倒播的视频。商品方面，每个商品环拍一圈，然后每隔一定度数取一张照片，结合绿抠手段把物体抠出来，把它形成一个连播的文件。有了这些准备工作，只需要在正向走动的时候播正向的视频，用户每时每刻就都知道物品在哪里了。由于安全距离的问题，对小商品的细节描述会遇到问题，比如货柜上的手表会看不清楚，可以通过场景交互手段解决这个问题。走到一个区域时，VR 应用让你进入另外一个场景，这个场景是全景图片，这个时候商品细节表现会好一些。不断推陈出新，Detail AR、Detail 3D、高清照片质量与空间定位方案等技术也在持续研发中。

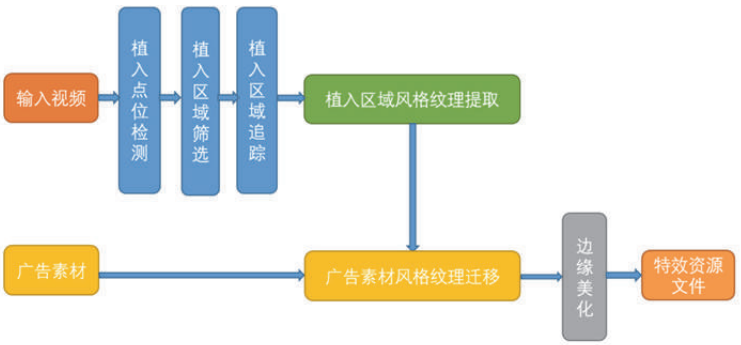
四、应用案例

视频特效广告植入是 XR-Video 技术的应用场景之一，它解决了广告与会员的二元矛盾，让广告也能给用户较好的体验，以及降低新内容排播不确定性对广告的影响。人工智能平台实现了点位生产、特效制作、投放和播放等全链路的打通，银鹭、良品铺子、OPPO、携程、曼秀雷敦和御泥坊等多家广告客户上线尝试了这种新型的广告形式。



视频特效植入系统示意图

我们以移花接木为例解释一下特效植入的大致流程。首先输入视频，点位系统根据物体识别、品牌识别、场景识别、动作识别、BGM 识别、OCR、情绪识别和明星识别等识别手段，找到合适的场景，并进行平面点位检测和筛选，然后进行植入区域的跟踪，生成植入区域指示数据；广告素材和特效蒙版（通过分割和粒子效果等手段生成）进行特效合成生成特效素材，对特效进行风格迁移和光照估计，生成的植入效果做边缘美化后，根据植入区域指示数据进行合成，生成特效资源文件。



移花接木整体流程

在保证播放原视频的同时，还要保证动态渲染广告的视觉效果，要求严格的帧同步，精准和轻量级渲染，另外 Android、iPhone 和 OTT 播放设备差异大，技术方案普适性和兼容性挑战大。我们通过大千制作平台巧妙地进行资源分离，集成了多种特效资源生成和灵活地对接，特效层与播放层高性能的通讯机制，独特的特效视频掩码设计，实现了轻量渲染和快速同步，在安卓千元机（Android5.0 以上），iPhone 6 以上均可流畅播放，OTT 2 款投放，后续继续推进摸底测试通过的 12 款魔盒，联盟盒子和一体机等。帧同步双流渲染技术经过了多次迭代，直接叠加带透明通道的视频会存在兼容性问题，因此我们提出了滤色 Key 方案（性能消耗较大）、WebP 渲染方案（内存占用较大），逐步演变到了双流掩码方案（性能、资源占用情况均较佳），最后通过 pts 基准合流渲染的方法达到了严格的帧同步，至此特效广告与视频资源达到了解耦+同步的两全其美效果。

帧同步动态渲染技术演进

五、结束语

大千世界无奇不有，大千 XR-Video 技术在植入特效广告和互动剧等应用场景上有了初步的成效，后续拓展到 360 度视频以及智能影像生产服务中。它们需要更沉浸、更准确、更有趣地互动式植入，我们将从以下两个大方向努力：

1）2D 与 3D 结合的三维感知技术是研发重点之一，这项技术处理摄像机的运动轨迹、景

深、遮挡关系，满足影视级多场景的视频生产要求。一方面，我们利用估算摄像机的运动并通过算法从 2D 点中生成 3D 点，从而实现 3D 重建与人物遮罩，为后续妥善处理视频中元素的遮挡问题进行技术铺垫。另一方面，结合语义信息用深度学习进行 2D 语义分割，将分割结果与 3D 重建的结果结合即可得到了一个含有语义信息的深度图或点云，在此基础上进行的平面检测，通过大量影剧视频训练的神经网络提高准召率；

2）为了达到真实感的效果，光照估计是我们另一个研发重点。利用深度学习对光照进行估计，研发基于深度学习的场景光照特征识别算法，智能感知视频画面中的光源方向及光源照度分布，为增强现实 3D 模型渲染提供数据支撑，保证虚实场景视效的一致性。

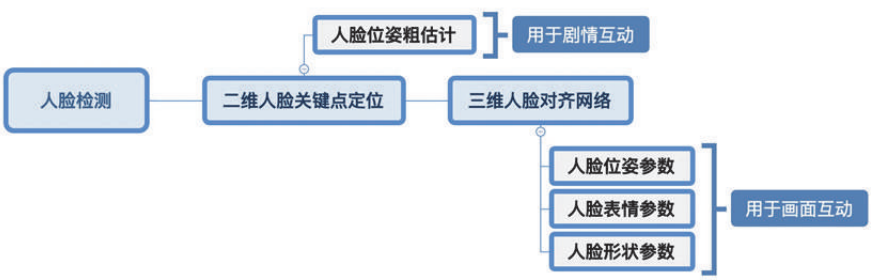
XR 技术，是人工智能技术非常重要的一部分，而 XR 更符合人类接受信息的方式，能大幅提升信息获取的效率。大千世界无奇不有，让我们一起发挥想象力，一不小心，也许未来的生活方式将因您而改变。

大千 XR-Video 技术在互动剧上的应用

作者| 阿里巴巴文娱高级算法专家 方如、高级算法工程师 黑仔

2019 年的流媒体关键词又多了一个：互动视频。它是一种用户能“玩”的交互式网络视频，是一种游戏化的视频，或者说视频化的游戏。2019 年 6 月 20 日，爱奇艺推出的《他的微笑》，是平台建立了自己的互动视频标准后真正意义上的第一部互动剧。优酷也在布局互动剧，一场新赛道上的战役已经打响。用户在观看的互动剧时候，每触发一个情节点，都需要通过选择操作来确定剧情的走向。

传统的互动剧因为有很多分支，所以情节复杂，但是互动性却并不算高，因为所有的情节依然是编剧事先设置好的，并非观众的创造，因此用户的参与成程度并不高。我们在传统的选择事先拍摄好的分支剧情（AB 分支）互动基础之上，利用技术手段，增加了用户与画面内容的交互。例如传统的互动剧根据点击选择剧情的基础上，我们增加了各种体感玩法，其中人脸互动玩法中，我们采用了 3d face alignment，通过用户的人脸图像，进行人脸识别，定位，对齐，估计用户的人脸形状、位姿、表情等，并利用这些信息影响剧情的走向以及剧集画面互动。



人脸互动算法流程如下：

1）人脸检测、二维关键点定位：为满足实时互动，采用快速人脸检测算法，如 mtcnn、Face R-CNN 等，粗定位出人脸区域；

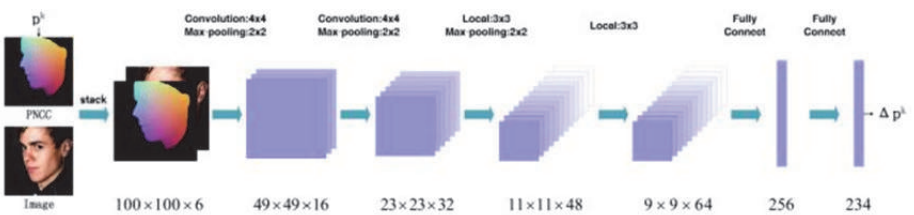
2）人脸二维姿态粗估计：利用二维关键点，进行 PnP 方法，与现有三维模型上的关键点进行反算，可以得到人脸的姿态，由于二维向三维映射的误差，仅可以得到粗略估计；

3）三维人脸对齐：在 3DDFA 算法的基础上，我们进行了很多新的尝试。

（1）不同于 3DDFA 输入为 100x100 的 RGB 图像和 PNCC（Projected Normalized Coordinate Code）特征，经过实验，PNCC 特征对于人脸重建的精度并没有明显的提升，因此我们去掉了这一特征；并将网络的输入调整至 128x128，重建效果提升较大，且网络运行时间上并没有明显增长；

（2）此外，卷积神经网络的损失函数进行了调整。在引入权重，让网络优先拟合重要的形状参数，如尺度、旋转和平移的基础上，增大了表情特征的权重，让人脸表情更加贴近真实；

（3）为了更好的重建人脸模型，在重建过程中采用 uv 自适应的方法，对于重建效果，在肉眼观察下明显提升。



3ddfa 原始网络框图

获取了人脸的位姿、形状、表情参数后，即可进行 3d 建模以及渲染，我们实现了一些有意思的互动：

人脸画面互动

当用户在观看视频时，如果喜欢一个演员，可以采用“点赞”、“送花”等形式，画面中演员会反馈一张笑脸，剧情走向可以走向一个开心的分支；反之，不喜欢一个演员，采用“丢砖”等

形式，画面中演员反馈一张生气的脸，剧情走向另一分支。除此之外，人脸的位姿参数和其他表情参数也可作为剧情走向互动的输入条件，例如点头、摇头、哭、笑等。用户在观看视频时候，不再只可以通过 AB 选项的方式机械化的引导视频走向，而是更加鲜活的影响剧情中的人物或物品，达到更加“身临其境”的互动。这种互动玩法涉及的技术专利已被受理。

除此之外，模型植入与互动这个创新点值得后续继续跟进。目前优酷互动剧基本上是自制的，这就给在视频内容中融入互动内容留下很大空间。在视频内容的拍摄过程中，预留一些已知模式的标记物，即标定板，可以在后期制作中获取相机的各种参数信息，利用这些参数信息，我们可以在视频内容的基础上，“玩出”各种各样的花式玩法。

求解相机参数（内参、外参和畸变参数）的过程就称之为相机标定（或摄像机标定）。利用标定板（例如二维码）辅助进行 3D 模型植入。在拍摄现场布置多个标定板，一次拍摄搞定多相机自动化标定，提高三维空间定位与跟踪的精度。图中使用标定板，利用 PNP 的方法实现相机的定位，更精确地计算出相机参数。植入 3D 模型会让互动剧更有趣，例如：

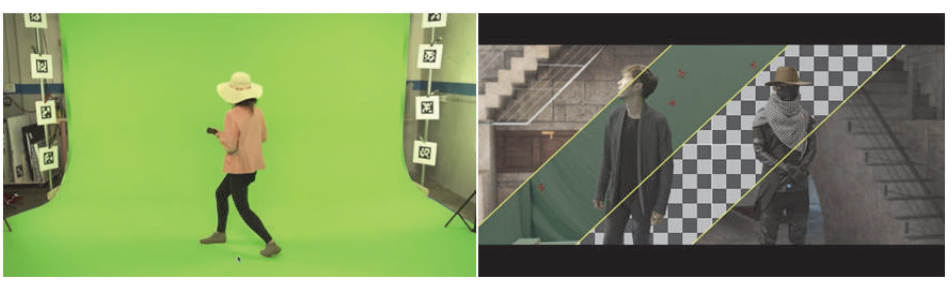
（1）植入模型可交互，可以根据不同操作有不同的互动效果，不需要预先拍摄；

（2）植入模型可推动剧情发展，利用植入模型参考和选择线索，进入不同的剧情；

（3）植入内容可多次配置，针对相同的视频内容，可以多次植入不同物体；

（4）植入手段更贴合内容，让视频观众更加容易接受植入；

（5）植入物体可作为广告，增加用户对品牌的认知度。



标定板和绿幕辅助拍摄标记现场以及合成后画面

优酷致力于互动剧的发展，先后颁布了内容制作和技术标准，涉及到的体感互动算法也不断完善。期待互动剧这个新形式被越来越多人接受，期待有更多新颖的玩法提供给观众。

优酷视频换脸技术实践

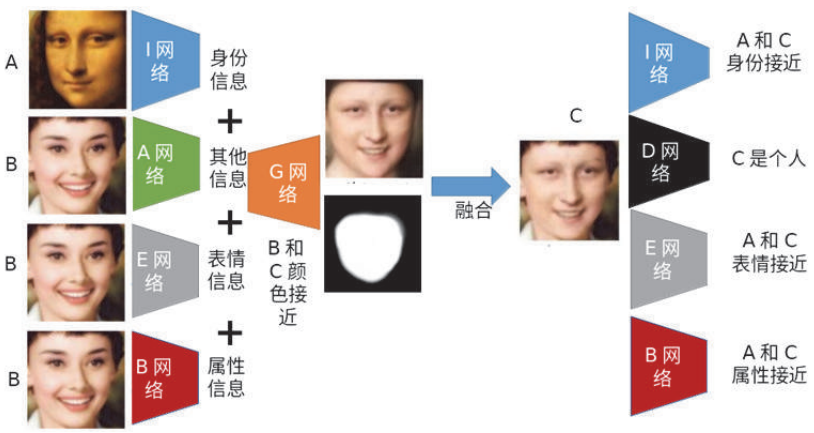
作者| 阿里文娱算法专家 崇伯

一、背景

优酷在 2018 年双 11 期间上线了一款换脸黑科技的活动。在这个活动中，用户仅需上传一张照片，无论是自己的还是别人的，即可将指定视频中明星的脸换成自己的脸，能看到样貌和自己一样的角色按剧中情节进行表演和对话，好像自己穿越到剧中，成为自己喜爱的明星，或者和自己喜爱的明星搭戏。也可以让自己的朋友或领导在剧中反串，可能会产生非常鬼畜的效果。针对我们选定的这些十几二十秒的小片段，换脸耗时仅需 5 秒。对，跟 ZAO 的功能绝对雷同，只是比 ZAO 早上线了将近一年。

二、核心算法

我们最终选用的算法架构的核心是在[1]的基础上改进的一个多分支的 CVAE+GAN 损失的结构，使用多个编码器来对不同的人脸特征（身份、表情、属性、环境）进行去耦合，然后通过鉴别器加上多个条件回归网络来确保生成的人脸的逼真程度以及各个特征的还原程度。

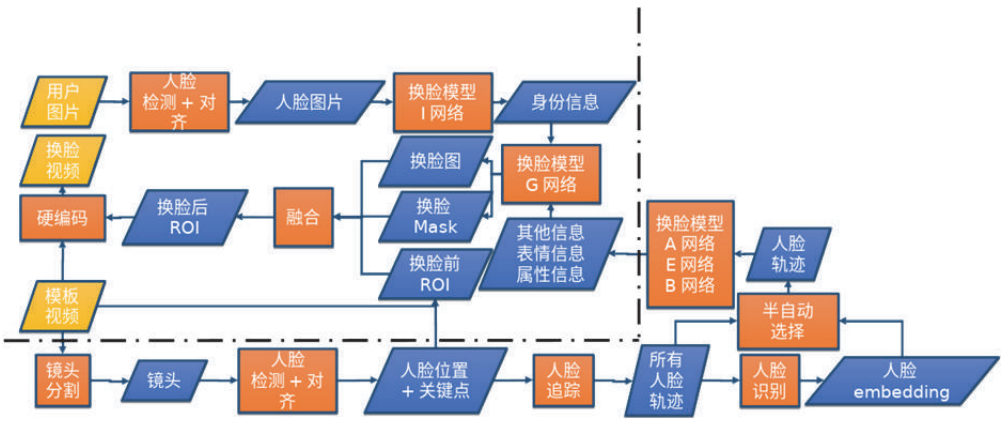


整体的网络架构如上图所示，常规的模型使用了 I 网络和 A 网络来学习被换脸人的身份信息和其他信息。但我们为了我们使强化学习到被换脸人的实时表情，比如眼珠和嘴角的细微动作，增加了 E 网络来强化表情的相似程度，并用表情数据集进行额外训练；另外增加 B 网络来学习人物的属性信息，使画面更干净、清晰度更高。

除了脸部本身的生成之外，将生成的脸部与完整画面融合也是一大难点，尤其是面部有阴

影，或者头发遮挡较多的情况。使用分割-融合的方法可以解决这个问题，但一是计算效率比较慢，二是 badcase 也多，所以我们为了解决生成的人脸使用 seamless clone 算法融合到原视频中速度慢的问题，提出了在使用 CVAE 生成人脸的 RGB 图像的同时也生成一个 alpha 通道的遮罩的魔心结构，在解决融合性能问题的同时也大大减少了融合后的瑕疵。

我们的算法流程如下图所示。算法流程分为两大部分，第一部分为对待换脸模板视频进行的离线预分析，第二部分为线上用户换脸请求时的实时处理。



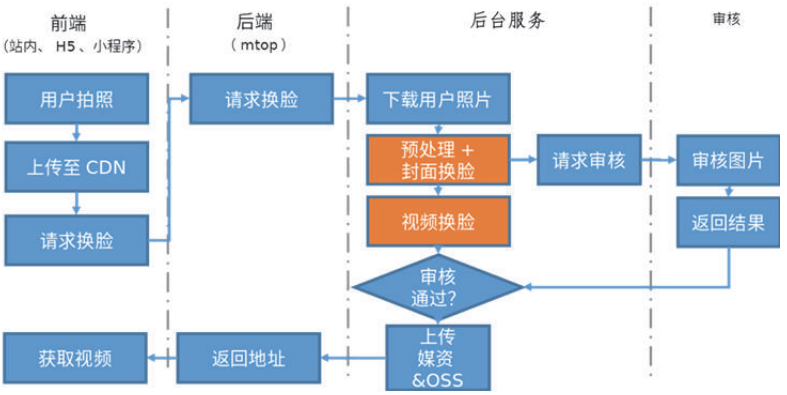
对换脸视频进行预分析包括几个主要部分，首先是对整段视频进行镜头分割，然后对每一段镜头中所有的人脸进行检测和对齐，得到人脸的位置和关键点位置。然后，标定将要被换脸的对象是哪条脸的轨迹，并视情况对转头、人物间遮挡等特殊情况进行修补。得到了需要换脸的轨迹之后，通过换脸模型的 A 网络、E 网络和 B 网络求得待换脸任务的其他信息、表情信息和属性信息，准备工作就完成了，将该可换脸视频模板上线。

用户上传图片后，首先对用户照片进行人脸检测和对齐，将人脸图片通过 I 网络得到用户的特征信息。将用户特征与用户选择的要将自己换入的视频的各种信息一起通过 G 网络，得到每一帧中用户生成的换脸图以及其对应的位置遮罩，将其融合后逐帧贴回原视频，之后进行编码即可生成完整的换脸视频了。

三、工程实现

由于考虑到视频换脸存在较大的法律风险，我们最终的产品形式定义为由运营和审核的同学来挑选优酷版权的片子片段，然后用户上传的照片也会通过优酷的安全审核系统来防止用户违规上传非法图片。整个线上的工程链路在能实时给到上传用户以反馈的同时有完整的安全保

障链路。



我们的工程链路如上图所示，我们开放的所有入口包括优酷 app、H5 活动页、小程序活动页，用户上传自己的照片后，后台服务会先进行封面图的换脸，我们会挑选一张典型的正面高清图作为封面图，并将这张图和原图提交机审及人工审的审核。这样审核团队仅需审核图片即可确认原始上传图片及换出来的效果是否含有风险，大大提升了处理的人工效率和机器效率。在审核的同时，后台服务会并行地进行整个视频的换脸，逐帧替换并生成新的完整视频，当审核结果为通过时，第一时间将视频放入可播的媒资库并显示在用户界面上。

为了保证用户的体验，即在上传照片后不用等待太久即可看到换脸后的效果，我们定了从上传照片到播放出换脸后视频之间时长不能超过 5 秒的技术目标。我们选取的被换脸的视频长度在 10-25 秒之间，其中需要换脸的画面时长在 10-20 秒之间，所以我们的算法效率要达到 1：4 的时效比。

为了达到这个目标，我们从工程上也进行了优化。首先，将从人脸检测到视频编码在内所有可能的运算都放到 GPU 上，避免了在计算链路中内存与显存之间的拷贝。其次，实现了一个多模型的管道机制，确保每个网络被调用的时候显卡计算尽可能处于独占状态，并结合 GPU 和硬编解码的能力，实现了单任务独占条件下 15 秒视频 3 秒内处理完成，并发条件下 5 秒完成，单卡高并发下处理极限效率为 1 视频/秒。

四、展望

由于各种客观条件的限制加上目前技术的效果，视频换脸算法的应用领域还是比较有限。但随着深度学习研究的不断深入，视频换脸技术本身还有更多的优化空间。如今年不断更新的

高分辨率人脸生成算法以及基于 3D 的人脸生成算法，为生成模型点亮了新的明灯。相信不久的将来，更逼真的视频生成算法会带来新一波的娱乐革命。

引用

[1] Bao, Jianmin, et al. “Towards Open-Set Identity Preserving Face Synthesis.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6713–6722.

基于多模态内容理解的视频智能裁剪

作者| 阿里文娱算法专家 崇伯

一、背景

随着大家观看视频的设备形态越来越丰富，在不同宽高比例的显示屏幕上观看同一个视频的需求变得越来越旺盛。从 4：3 到 16：9 的电视宽屏化过程中，大家已经经历过一个令人印象深刻的电视上所有人物都很胖的过渡时期，现在又要面临一个很严重的竖屏观看的问题，以及其他更多的剪辑加工需求。尤其对于竖版视频的生产而言，原生的竖版内容和由横版内容转换而来的竖版内容都很重要，但目前由横版内容转换而来的竖版视频的体验和产能都不尽如人意，在人工裁剪的过程中一方面费时费力，另一方面也很难处理视频中大幅运动的显著目标。

下图就是一帧视频中人眼注意力在水平方向上的大致分布示意。

二、算法目标

针对围绕从横版视频转竖版视频的一系列视频裁剪需求，我们在技术上将问题重新定义成：分析画面中内容的重要程度，并能够在任意裁剪尺寸的约束下，尽可能多地保留其最重要的部

分，并在时域上位置相对连续使人眼感受最好。

目前业界在算法领域领域的研究没有非常成熟的解决方案，最接近的研究领域是图像重定向（image retargeting）算法，即将单张图片的内容重新组织以适应变化的宽高比的问题。但由于视频的重定向需要保证图像内容的时域连续性。单帧内容重组织的方法几乎不可能做到时域的平滑自然，对视频内容可用性还不强。

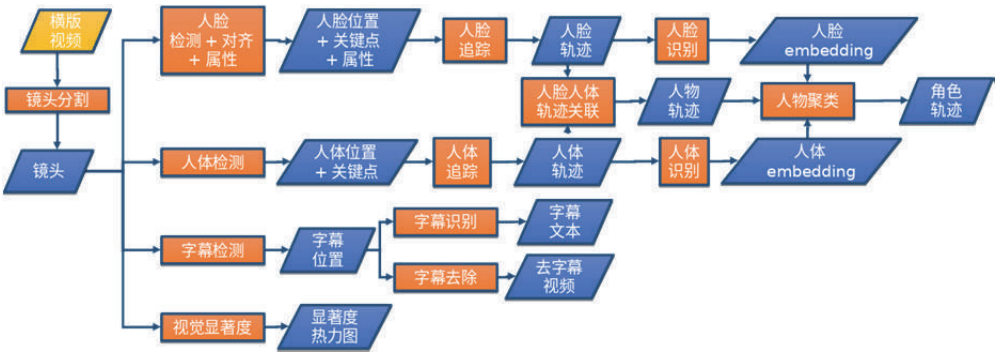
视频智能裁剪算法的基本目标是尽量保证裁剪区域内的内容完整性。理论上内容完整性指的是一个镜头中导演希望观众看到的元素尽量完整并位于显著位置，但这个目标很难真正实现。所幸的是绝大部分视频中导演希望观众看到的元素都会设法做得对人眼显著，这也是运用构图、景深、灯光等拍摄技法的一个重要目标。这样就可以将这个目标转化为人眼视觉显著区域完整性这个指标，这样就可以充分利用到较大规模的显著度标注数据。

另外一个需要权衡和算法目标是用户观看的舒适程度。如果仅仅考虑最大化内容完整性这一指标，得到的竖版视频会产生抖动和晃动等问题，很大地降低用户的观看体验。因此视频裁剪算法另一的重要的约束条件是尽量避免影响用户观感的镜头晃动。为了做到这一点，我们实验了不同的轨迹平滑策略，发现人眼对裁剪区域运动对时间的一阶和二阶导数变化更加敏感。据此使用了一阶+二阶全变分（total variation）最小化的算法来在保证抖动对观感影响最小。以下两幅图说明了这种优化的重要性。画面的横轴为逐帧时间 t，纵轴为实际画面的横坐标 x，橙色细线和深蓝色细线为两个人的实际位置变化，所以可以看出这是一个典型的剧集内容中包含 2 个人互动的镜头，内容是两个人一开始在画面左右两端，之后逐渐靠近最后又稍稍分开的过程。如果我们直接采用真实位置作为画面中心来裁剪，裁剪后的画面特征如左图所示，虽然人在画面中间但背景及交互物始终小幅晃动，人眼的顿挫感很强，经过我们优化后的画面截取轨迹如右图所示，在最大化地将任务主体包含在轨迹中的同时，实现了裁剪位置的尽可能平滑稳定。

三、算法流程

虽然端到端的算法一般来说更加精简高效，但对于智能裁剪来说，由于端到端的算法可以

利用的数据集比较小，很难达到比较理想的效果。反之，分步的方法能够有效利用现有的成熟的人脸、人体目标检测等大规模的数据集和模型，能够大大提高算法的准确性。另外分步的方法也能够提高整个算法体系的灵活性，使得算法能够快速根据业务需求进行灵活调整。因此我们采用了多步的算法来实现智能裁剪，整体的流程如下图所示。



第一步为镜头分割，由于相邻镜头间的裁剪位置是不相关的，因此可以以镜头为单位独立进行裁剪，避免了镜头边界造成的相关问题。镜头边界检测算法的准确率直接影响到智能裁剪错误率。我们在 TransNet[1]的基础上改进了一个高速高精度的模型将镜头分割导致的错误率降低到 0.2%以下。

第二步为特征抽取，这一步通过一系列算法群对每一个镜头生成了多个和视觉显著度密切相关的特征、包括人脸和人体的检测、朝向估计、清晰度评估、追踪和识别，光流的估计，视频视觉显著度的预测。

最后一步为特征的融合和裁剪位置的估计，这一步是整个算法流程中的关键部分。我们使用了对每一个镜头内目标能量函数最大化的方法，同时保证视频内容的完整性和裁剪区域运动的平稳。其中能量函数的定义如下所示。

$$E_{saliency} = \frac{1}{N}\sum_{t=0} D_{KL}(S, \mathcal{N}(c, \beta r^2)) = \frac{1}{N}\sum_{t=0}\sum_{x}^{N} S_x \log\left(\frac{S_x}{f_{\mathcal{N}(c,\beta r^2)}(x)}\right)$$

$$E_{objects} = \frac{1}{N}\sum_{t=1}^{N}\left(E_{base} + \sum_{o} w_o\left(E^o_{offset} + E^o_{inter}\right)\right)$$

$$\text{minimize } \alpha E_{object} + \beta E_{saliency} + \gamma E_{mv} + \delta\frac{dx}{dt} + \epsilon\frac{d^2x}{dt^2}$$

四、业务实践

我们目前已经和正在将这套智能裁剪的算法能力用于竖版视频的生产以外的多个应用场景，包括智能视频平铺播放、智能竖版直播、动感旋转视频、以及算法辅助的人工视频编辑等。

五、展望

目前优酷的智能裁剪算法已经能够自动为竖版视频提供海量的内容，也能够催生更多更新颖的播放形式。我们下一步的重点包括将智能裁剪算法适用于包括动画、体育等特殊的视频场景，以及用更加自然放来的重新组织复杂的综艺类的画面内容。

引用

[1] Soucek, Tomás, et al. “TransNet: A Deep Network for Fast Detection of Common Shot Transitions.” ArXiv Preprint ArXiv:1906.03363, 2019.

阿里文娱视频智能裁剪技术实践

作者| 阿里文娱算法专家 闵公

全球文娱视频市场存在海量统一横屏制作的大剧热综和 UPGC 短视频等，同时优酷中也存在着大量横屏播放的长短视频，随着近两年来竖版视频的流行和较高的播放转化效率，用户对竖版视频的消费需求越来越旺盛。

阿里文娱优酷首次将基于机器视觉的视频裁剪技术应用于视频二次生产和智能封面图生成业务中，该技术已经覆盖优酷的 OGC，海量 UPGC 竖版短小视频以及静态封面图，同时我们将该技术能力输出给阿里云，服务于中国的文娱企业客户。

智能裁剪技术主要应用于以多人或者单人为主体的场景，我们将目标检测，跟踪，识别等技术进行创新和结合，开发了完整的视频智能裁剪技术链路，面对实际业务中的主体标定，视频帧间抖动，视频黑边填充等问题针对性的研发了算法解决方案，可以根据不同的业务场景将各算法可插拔的配置进主裁剪 pipeline 中，阿里文娱视频智能裁剪技术的研发给内容行业的素材自动化制作，剪辑作品的视觉效果和制作成本降低等方面都带来了大幅度的提升。

在视频智能裁剪技术链路（如图 1 所示）中，我们研发了前处理模块（包含镜头切分，画面尺寸判定，黑边检测裁剪等），主体标定模块，主体追踪模块和后处理模块（包含画质增强，字幕/logo 检测，画面内容修补等），下面分别介绍四个模块。

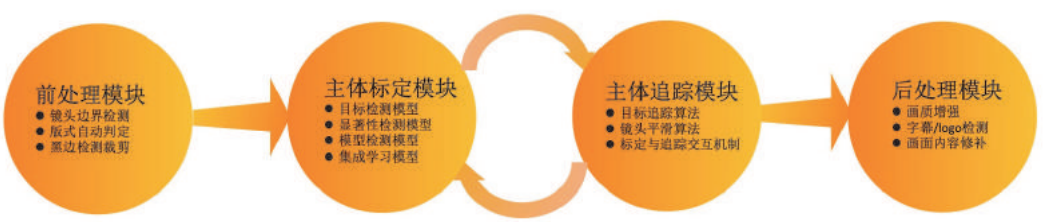


图 1

一、前处理模块

1）前处理模块包括分镜边界检测模型，画面尺寸判定算法，黑边检测与剪裁算法等三个模块，其中分镜边界检测模型根据视频画面将视频分成多个镜头片段（如图 2 所示）；

图 2

2）画面尺寸判定算法使得裁剪可以在不同的画面尺寸中进行自动选择，包括（宽：高）16:9, 4:3, 1:1, 3:4, 9:16 等任意尺寸，通过对视频帧进行抽样后根据目标的显著性和运动特性来计算得出显著区域的大小进行剪裁尺寸适配；

3）由于大量竖版视频存在上下黑边填充现象（如图 3 所示），但上下黑边在自动裁剪后会严重影响用户体验。因此我们使用边缘检测算子和霍夫变换等来解决黑边检测与剪裁的问题。

图 3

二、主体标定模块

主体自动标定模块（如图 4 所示）包含目标检测模型，显著性检测模型，模糊检测模型和

集成学习模型，我们通过模型集联结构来进行视频帧中主体目标的自动标定。

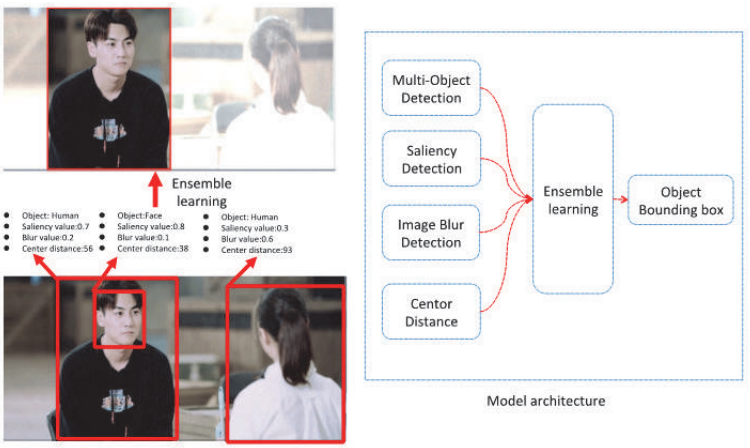


图 4

根据目标检测模型对视频中的人脸和人体进行检测后，将包含人脸或人体的 bounding box 作为候选主体集合，为了避免前景虚化现象,在自动标定模块中引入了显著性检测模型，通过显著性获取画面中不同位置为显著区域的概率；由于不同视频存在不同的降质现象，我们研发了模糊检测算法，通过模糊检测模型提供图像清晰度比较结果，从而实现选择更为清晰主体的目的，除了上述子模型的输出结果外，我们还设计了主体检测框离画面中心的欧式距离，基于相邻帧的预测位置等特征，通过集成学习模型将上述输出结果进行拟合，根据标定的 110 部大剧主体标定的结果来训练主体判定模型使得主体标定的 Accuracy Rate 达到 99%。

三、主体追踪模块

主体追踪模块包括目标追踪算法，镜头平滑算法，主体标定和主体追踪交互机制。通过对多个物体运行多次 SOT 追踪得到关键帧后续相邻帧中主体目标对应的位置，形成连续视频帧的镜头标定结果。由于目标追踪算法得到的镜头剪裁位置并不是渐变的，这导致了画面抖动，引起用户观看眩晕等较差体验，因此通过时间序列离群点检测和 Kalman filter 等技术，将异常定位点 t 进行平滑，解决了裁剪后视频帧间抖动问题，抖动幅度 Jitter Degree 由 1.93 优化至 0.07，人工评估视频帧后观感流畅。同时通过主体标定和主体追踪交互机制(如图 5 所示),保证了主体目标在镜头切换情况下的镜头内容连续性。

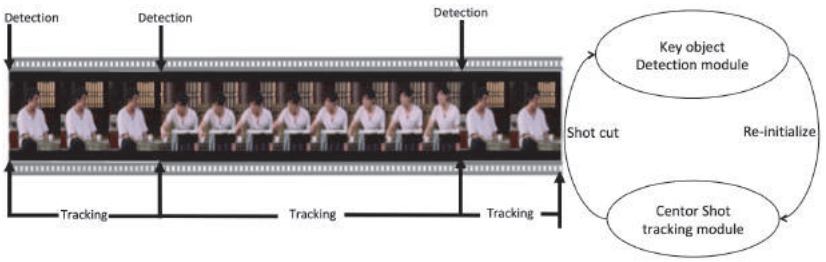


图 5

四、后处理模块

针对视频剪裁后的视频画质问题，我们开发了后处理模块（包含画质增强，字幕/logo 检测，画面内容修补等），主要解决剪裁边界可能的 logo/字幕截断问题和分辨率降低的问题。首先对原视频帧进行字幕检测和 logo 检测，得到分割 Mask，并将其作为输入进行图像修补（Inpainting）。最后我们使用去噪、去模糊、和超分辨率模型，对裁剪后的视频进行画质提升。

五、结束语

视频智能裁剪技术生产的视频和封面图广泛应用于优酷的各个场景，并得到了业务方和阿里云客户的一致认可，我们对视频智能裁剪算法栈进行了整体性能优化，达到处理时间仅 1:2 视频时长，目前该技术累计对优酷综艺：演技派、这就是街舞、这就是灌篮；优酷剧集：陆战之王、天雷一部之春花秋月、微微一笑很倾城等百部 OGC 进行裁剪服务，裁剪后的竖版视频用于抖音，微博等外渠宣发和站内投放，同时主体标定算法服务于搜索双列封面图转竖项目，镜头平滑算法服务于弹幕人脸项目，目前视频裁剪算法已经部署在阿里云上，由于目前行业内竞品尚无成熟技术方案，已经通过申报《基于主体目标标定与追踪的视频智能剪裁技术》，《基于智能画面分析和多层级主体目标标定的图像智能剪裁技术》专利的方式来保障该产品技术的竞争优势，期待阿里文娱视频裁剪技术为中国的视频娱乐行业创造更大价值。

技术实践-精准的视频物体分割算法以及应用

作者| 阿里巴巴资深算法专家 任海兵

视频物体分割（Video Object Segmentation，简称 VOS），顾名思义就是从视频所有图像中把感兴趣的物体区域完整的分割出来。为了方便大家的理解，先给出一个我们自己的视频物体分割的结果：

视频 1：视频物体分割示例，选中的人体区域用红色高亮显示

视频物体分割是进行内容二次创作的重要素材。例如目前火爆的“裸眼 3D 视频”，基于视频中主要物体与观众之间的距离，利用蒙皮遮挡的变化产生 3D 效果。其核心点是将前景物体从视频中分割出来，这部分会花费创作者 99%以上的时间。

因此，对于优酷这样的视频类网站，视频物体分割是非常有价值的算法，能够赋能内容生产者，提升内容生产效率。特别是交互式视频物体分割算法，能利用用户少量交互，逐步提高视频物体分割正确率，提升用户观感体验。这是任何无监督视频物体分割算法所不能达到的。

目前，CV 学术界在视频物体分割方面的研究主要分为三个方向：

- 半监督视频物体分割（Semi-supervised video object segmentation）；
- 交互式视频物体分割（Interactive video object segmentation）；
- 无监督视频物体分割（Un-supervised video object segmentation）。

这三个研究方向对应于 Davis Challenge 2019 on Video Object Segmentation[1]中的三个赛道。其中，学术界更倾向于研究半监督视频物体分割，因为这是视频物体分割的最基础算法，也是比较纯粹的一个研究点。接下来，我首选介绍视频物体分割的三个研究方向，然后结合优酷认知实验室的探索，分享在视频领域的最新应用。

一、半监督视频物体分割

半监督视频物体分割，又称为单一样本视频物体分割（one-shot video object segmentation，简称 OSVOS）。在半监督视频物体分割中，给定用户感兴趣物体在视频第一帧图片上的分割区域，算法来获取在后续帧上的物体分割区域。物体可以是一个，也可以是多个。在视频中，存在物体和背景运动变化、光照变化、物体旋转变化、遮挡等，因此半监督视频物体分割算法研究的重点是算法如何自适应获取变化的物体表观信息。一个示例如下图所示：

（a） Frame No. = 1　（b） Frame No. = 20　(c) Frame No. = 40　(d) Frame No. = 60

图 1. 半监督视频物体分割示例

在图 1 中，第一行为序列的 RGB 图片，第二行为感兴趣物体区域。其中（a）为视频第一帧图像，骆驼区域是给定物体的 ground-truth。（b）(c)和 (d)是后续的第 20、40 和 60 帧，后续的图像只有 RGB 图片，需要算法去估计物体的区域。该示例的难点是：(1) 前景背景颜色非常相似；（2）随着目标骆驼的运动，背景中出现一个新的骆驼，需要分割出这两个不同的骆驼区域。

目前半监督视频物体分割算法分为两大类：有在线学习、无在线学习。

基于在线学习的算法根据第一帧物体的 ground-truth，利用 one-shot learning 的策略来 fine-tune 分割模型。经典的在线学习算法包括 Lucid data dreaming[2]，OSVOS[3]，PreMVOS[4]等。在线学习算法针对每个物体单独训练模型，可以达到很高的分割正确率。但是在线学习本身是深度学习模型的 fine-tuning，需要耗费大量的计算时间。在 2019 年之前，在线学习算法是主流。今年出现了不少无在线学习的算法，它的模型是事先训练好的，不需要针对样本进行 fine-tune，具有更好的时效性，例如 CVPR2019 的 FEELVOS[5]，Space-time memory network[6]等。

半监督视频物体分割的最主要的结果评估标准是平均 Jaccard 和 F-measurement.平均 Jaccard 值是所有物体在所有帧上分割精度 Jaccard 的均值。F-measurement 为分割区域边缘的准确度。半监督视频物体分割由于其需要第一帧物体区域的 ground-truth，因此无法直接应用于实际应用。但它是交互式和无监督视频物体分割算法的核心组成部分。

二、交互式视频物体分割

交互式视频物体分割是从去年开始兴起的、更贴近实用的视频物体分割方法。在交互式视频物体分割中，输入不是第一帧物体的 ground-truth，而是视频任意一帧中物体的用户交互信息。交互信息可以是物体 bounding box、物体区域的划线（scribble)、外边缘的极值点等。

基本流程如下图所示：

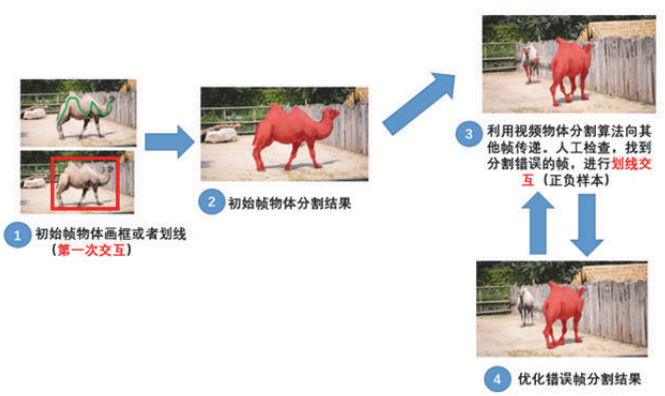


图 2. 交互式视频物体分割流程

交互式视频物体分割通常包括以下 5 个步骤：

1）用户输入交互信息，标记感兴趣物体，例如物体的 bounding box，scribble 信息、边缘点等；

2）根据用户输入的交互信息，利用交互式图像物体分割算法分割出物体在该帧图像上的物体区域；

3）根据前一帧物体区域，利用半监督视频物体分割算法向视频其他帧图像逐帧传递，进行物体分割，得到所有帧图像上物体区域。然后，用户检查分割结果，在分割较差帧上，给出新的交互信息；

4）算法根据新的交互信息，修改该帧图像上的分割结果；

5）重复步骤 3 和 4，直到视频物体分割结果让用户满意。

交互式视频物体分割不是一个单一算法，而且多种算法有机融合的解决方案，包括交互式图像物体分割、半监督视频物体分割、交互式视频物体区域传递算法等。其主要评估方法为 Davis Challenge on Video Object Segmentation 中提出的 Jaccard&F-measurement@60s(简称 J&F@60s) 和 AUC（Area Under Curve，简称 AUC)。Davis 竞赛提出限定 8 次用户交互，建立准确度随时间的变化曲线图，曲线下方区域的面积就是 AUC，t=60s 时刻曲线插值就是 J&F@60s。下图为一个 J&F 随时间变化曲线图。

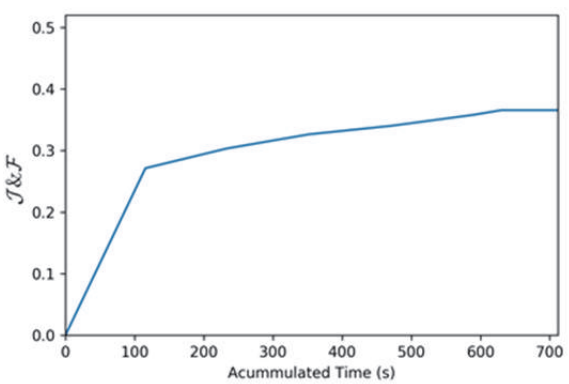


图 3. 交互式分割结果 J&F 曲线示例

从评估指标可以看出，交互式视频物体分割强调分割算法的时效性，不能让用户长时间等待。所以，在交互式视频物体分割中一般不采用基于在线学习方法的半监督视频物体分割算法。

目前还没有交互式视频物体分割的开源代码。但是交互式视频物体分割算法对工业界有非常重要的意义，其原因是：

1）半监督视频物体分割需要物体第一帧的 ground-truth，实用中获取比较麻烦。而交互式视频物体分割只需要用户的简单交互，非常容易达到；

2）交互式视频物体分割可以通过多次交互，达到非常高的分割正确率。高精度的分割结果能够提供更好的用户体验，才是用户需要的结果。

三、无监督视频物体分割

无监督视频物体分割是全自动的视频物体，除了 RGB 视频，没有其他任何输入。其目的是分割出视频中显著性的物体区域。在上述三个方向中，无监督视频物体分割是最新的研究方向。

Davis 和 Youtube VOS 竞赛今年第一次出现无监督赛道。从算法层面上说，无监督视频物体分割需要增加显著性物体检测模块，其他核心算法没有变化。

半监督和交互式视频物体分割中，物体是事先指定的，不存在任何歧义。而在无监督视频物体分割中，物体显著性是主观概念，不同人之间存在一定的歧义。因此，在 Davis VOS 中，要求参赛者总共提供 N 个物体的视频分割结果（在 Davis Unsupervised VOS 2019 中，N=20），与数据集 ground-truth 标记的 L 个显著物体序列计算对应关系。对应上的物体和遗漏的物体参与计算 J&F 的均值。N 个物体中多余的物体不做惩罚。

四、优酷认知实验室的研究现状

目前很多半监督视频物体分割算法在学术上有很好的创新，但是实用中效果不佳。我们统计了今年 CVPR 的论文，在 Davis 2017 val 数据集上，没有一篇正会论文 J&F>0.76。FEELVOS[5]、siamMask[7]等算法理论上有很好，实用中却存在多种问题。交互式视频物体分割更是没有开源代码。

所以，优酷认知实验室从 2019 年 3 月底开始从事半监督和交互式视频物体分割算法的研究。

2019 年 5 月，我们完成一版基础的半监督视频物体分割算法和交互式视频物体分割解决方案，并以此参加了 DAVIS Challenge on Video Object Segmentation 2019，在交互式视频物体分割赛道获得第四名。

我们提出的 VOS with robust tracking 策略[8]，可以较大幅度的提高基础算法的鲁棒性。在

Davis 2017 验证集上，我们交互式视频物体分割算法 J&F@60s 准确率从 3 月底的 0.353 提高到 5 月初的 0.761。现在，我们的半监督视频物体分割算法也达到了 J&F=0.763。可以说，在这个集合上我们的结果已经接近业界一流水准。一些分割结果示例如下：

视频 2. 我们的交互式视频物体分割结果示例

五、优酷认知实验室的后续计划

目前，我们在继续探索复杂场景下的算法应用，这些复杂场景包括小物体、前景背景高度相似、物体运动速度很快或表观变化很快、物体遮挡严重等。后续，我们计划在 online learning、space-time network、region proposal and verification 等策略上发力，以提高视频物体分割算法在复杂场景下的分割精度。

另外，图像物体分割算法、多目标物体跟踪算法也是视频物体分割算法的重要基础，我们也将在这些方面持续提升精度。

Reference

[1] The 2019 DAVIS Challenge on VOS: Unsupervised Multi-Object Segmentation. S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool .arXiv:1905.00737, 2019

[2] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. In arXiv preprint arXiv: 1703.09554, 2017. 2

[3] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taix´e,D. Cremers, and L. Van Gool. One-shot video object segmentation. CVPR, 2017

[4] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. arXiv preprint arXiv:1807.09190, 2018.

[5] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, Liang-Chieh Chen. FEELVOS: Fast End-to-End Embedding Learning for Video Object Segmentation. CVPR 2019

[6] Seoung Wug Oh, Joon-Young Lee, Ning Xu, Seon Joo Kim.Fast User-Guided Video Object Segmentation by Interaction-and-Propagation Networks. CVPR2019

[7] Wang, Qiang，Zhang, Li，Luca Bertinetto, Weiming Hu, Philip H.S. Torr.Fast Online Object Tracking and Segmentation: A Unifying Approach. CVPR2019

[8] H. Ren, Y. Yang, X. Liu. Robust Multiple Object Mask Propagation with Efficient Object Tracking. The 2019 DAVIS Challenge on Video Object Segmentation - CVPR Workshops, 2019

2

媒体智能引擎 SmartAI

媒体智能平台之推理服务

作者| 阿里文娱开发专家 欢朋

一、背景

随着人工智能算法领域的快速发展，机器学习在智能内容生产、安全审核、体育直播分析、视频内容结构化等领域的应用需求越来越多。算法开发工程师们面临以下挑战：

- 算法迭代频繁——业务发展快速，业务需求多变且变更频繁；
- 需要快速交付——业务驱动，需要快速给出结果；
- 系统环境复杂——依赖不同的计算底层，例如 GPU 或 CPU 等，同时也要保证算法服务的整体稳定性。

二、行业对比

目前业界有很多视频推理平台，如国外的 Deep Video Analytics (deepvideoanalytics.com)，实现了从视频标注到推理服务的链路；阿里云的视频云平台提供了具有很多能力的推理服务；优酷 smart 平台基于业务需求，整合了链路上的所有节点，串联了从标注到模型，再到推理，最后沉淀数据反哺标注的完整系统，实现了对模型迭代提升的一个正向循环；优酷业务复杂多变，算法开发模型也随着业务不断迭代，新需求新算法不断部署，smart 的产生就是为了解决这些问题，提供一个稳定又能促进算法提升的目的而生。参考了国内外平台，在此基础上，smart 实现了以下特性：

- 完整系统：实现从标注->数据->模型->推理->标注的循环；
- 智能标注：实现了以算法能力为基础的智能标注；
- 存储：实现了灵活高扩展的海量数据存储；
- 调度：根据算法能力自适应调度，多维负载均衡；
- DAG：算法能力实现图形化串联。

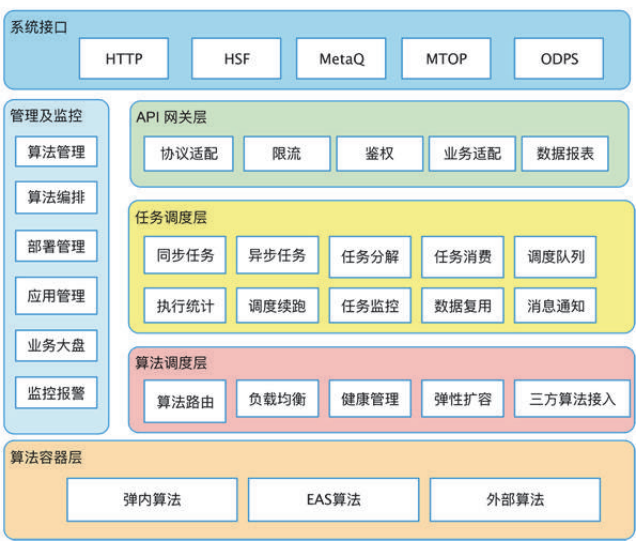
三、架构设计

1. smart 致力打造一个正向循环的数据算法服务平台

一站式的算法开发服务平台，集成了 Tron 算法开发平台、Smart 算法在线服务平台、标注系统、数据集等多个子系统来解决实际算法开发、生产发布与在线服务的各种痛点。

通过 smart 平台，能够赋能算法开发与业务应用，算法能够快速响应业务变化，驱动业务创新应用。

2. smart 逻辑架构



smart 整体由上到下分为 5 个部分：

1）API 网关层：实现统一外部接口，包括 qps 限流、请求参数签名验证、防止重放验证。并统计算法执行情况：每个业务方的算法调用量、当日总调用量等；

2）任务系统：监控报表、任务报表、qps 报表、任务执行统计；

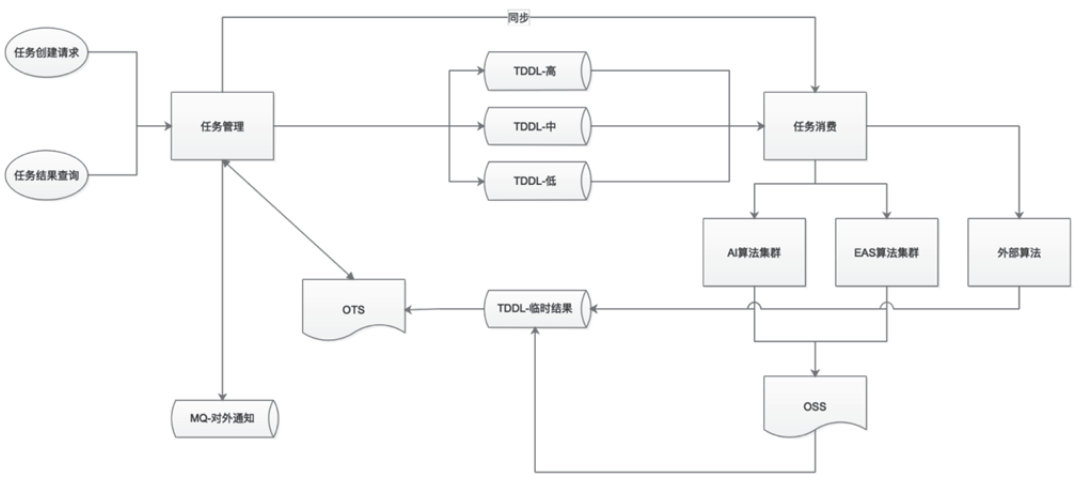
3）算法调度层：算法的蓝绿部署与流量分配、算法的负责均衡、算法机器的健康管理、算法机器发布管理、以及第三方算法的接入与适配；

4）算法计算层：包括弹内的 CPU，GPU 计算容器；以及弹外的 EAS 算法容器；

5）管理及监控：算法的配置、算法模板的配置、业务调用方的配置、限流配置、业务大盘、监控报警等内容。

四、技术细节

1. 任务调度策略



Smart 任务调度使用 MySQL 数据库作为任务数据的存储。Smart 任务调度引擎可以随时调整处于队列中的任务优先级，来灵活干预队列的执行，调度计算资源的分配。

1）smart 的任务调度支持优先级调度，可以根据不同的业务来源方设置不同的优先级。优先级 priority 的值越高代表任务的优先级越高；

2）开始执行的任务先进先出：进入到执行中的任务也会优先完成，避免被后续优先级高的

任务占领导致已触发的任务一直无法完成；

3）qps_limit 的任务优先重试：由于算法执行引擎繁忙导致 qps_limit 失败的任务，也会优先进行重试，保证已经开始执行的任务尽快完成；

4）通过数据库乐观锁确保每个任务同一时刻只被某个 task-consumer 处理，但处于运行状态的任务经过指定时间没有返回成功，会被放置回任务队列进行重试；

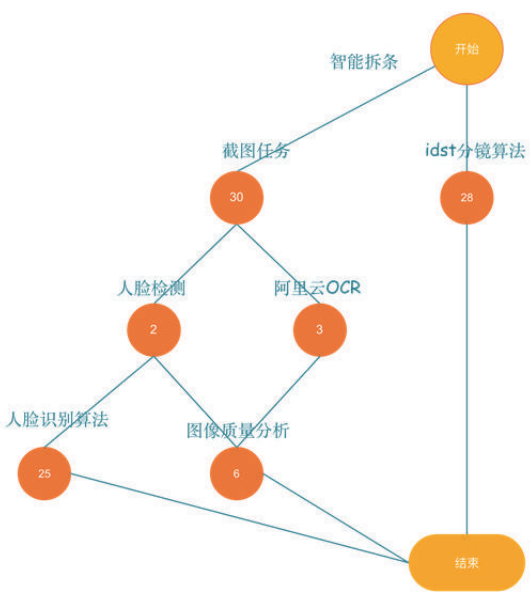
5）不同算法间调度的负载均衡：任务调度系统会根据不同算法在队列中等待个数以及相应算法执行引擎的饱和情况，来动态调整算法的权重，进行不同算法间的负载均衡。

2. 算法能力编排

算法模板是在单个算法能力的基础上，根据业务需求把一系列算法组装成一个完整的业务处理流程。

通过算法模板，能够将灵活的进行算法能力编排定制，快速响应业务需求，而不需要手工重新编码开发。

算法编排能力在满足特定业务需求的同时，也沉淀了优酷素材内容的各种解决方案。



图像质量分的算法模板

```
[

{"nodeId":30,"parentNodes":[],"childNodes":[2,3],"level":0,
"taskType":"video_recognition","nodeName":"snapshotAnalysis"},

{"nodeId":28,"parentNodes":[],"childNodes":[],"level":0,
"taskType":"full_video","nodeName":"idstSceneAnalysis"},

{"nodeId":2,"parentNodes":[30],"childNodes":[6,25],"level":1,
"taskType":"video_recognition","nodeName":"faceDetectAnalysis"},

{"nodeId":3,"parentNodes":[30],"childNodes":[6],"level":1,
"taskType":"video_recognition","nodeName":"aliyunOcr"},

{"nodeId":6,"parentNodes":[2,3],"childNodes":[],"level":2,
"taskType":"video_recognition","nodeName":"newImageQualityAnalysis"},

{"nodeId":25,"parentNodes":[2],"childNodes":[],"level":2,
"taskType":"video_recognition","nodeName":"faceRecognitionAnalysis"}

]
```

图像质量模板对应的 json 配置

算法模板的内容包括：

1）算法节点的任务处理内容：包括算法的名称、算法节点 id；

2）算法节点的依赖关系：一个算法节点可能依赖多个上一层级的算法节点的任务完成，并把上一层级的算法节点的输出结果作为下一层级算法的输入参数；

3）整个模板的最终输出节点：通过配置算法输出节点，来灵活定义整个处理流程的返回结果，可以定义为多个算法节点的返回结果；

4）算法节点的预置元数据：通过预置元素材实现调用算法时的参数干预；

5）算法节点的结果保存方式：是否复用 smart 系统中已有算法处理结果。

3. 灵活拓展的海量数据存储

为了满足不断日益增长的算法分析需求、与视频内容结构化算法结果复用，需要针对视频图像的每一秒一帧的图像算法分析结果进行存储。存储的数据量级达到了 70 亿+。

基于上述需求，选用了阿里云的表格存储（Table Store）作为 smart 的算法结果存储。

表格存储（Table Store）是阿里云自研的 NoSQL 多模型数据库，提供海量结构化数据存储

以及快速的查询和分析服务。表格存储的分布式存储和强大的索引引擎能够提供 PB 级存储、千万 TPS 以及毫秒级延迟的服务能力。

4. 通过列拓展满足动态算法存储需求

主键	type	videoId	url	startTime	endTime	算法结果1	算法2	....	算法n

如上图所示，这是某个视频图片的算法结果存储行。

Table Store 支持多列拓展，一行中除主键列外，其余都是属性列。属性列会对应多个值，不同值对应不同的版本，一行可存储不限个数个属性列。通过灵活的拓展属性列，来保存不同算法的算法结果。

在每个列的值可以对应不同的版本，版本的值是一个时间戳，可以用来保存算法不同版本的处理结果。

5. ots 主键的生成规则

主键使用 a.b.c.d 的规则

a 位，b 位的 hash 前 5 位，用于随机分布

b 位，最常用的查询条件。比如 site_videoId，taskId 等

c 位，任务类型

d 位，范围，比如 startTime_endTime，或者随机 uuid 的前 5 位 hash

例如：md5（key）#videoId#site#task_type#begin#end

以 md5(key)的前 5 位作为主键的第一部分，可以把数据散列，让数据存储整体负载均衡，避免热点问题。

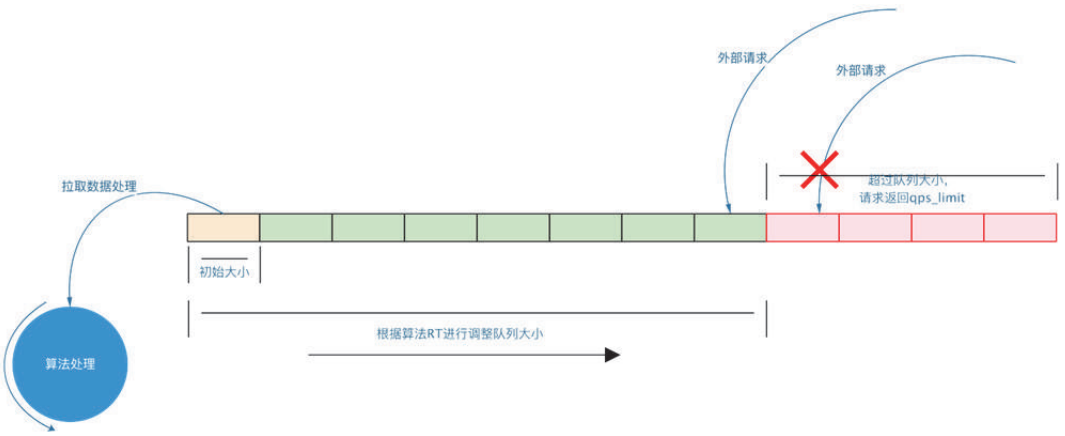
6. 算法的行级结果复用能力

假设一个视频以每秒 1 张的频率截图，总共有 1000 张图片，在算法执行分析的过程中有 999 张都分析成功了，剩下 1 张由于意外原因导致分析失败。在下一次进行任务失败重跑的时候，还可以复用已有的 999 张，只需要再增量去跑失败的那 1 张图片，减少了不必须的重复计算损耗。

如上图所示，随着同步算法 qps 的提升，系统会优先分配更多的资源给同步算法请求，同

时也会给异步算法保留一台机器。当同步请求减少时，异步机器可以获取更多的计算资源。统一同异步计算资源，有效地提升了系统的资源利用率，也优先了保障了在线算法服务请求的响应时间。

7. 动态自适应的算法队列处理策略



挑战：由于机器学习算法很多都需要独占 GPU 进行运算，在每个 GPU 上同时只能处理一个任务。如何保证 GPU 算法能够达到最大的吞吐量，并且每个算法的执行 RT 也不能太久。不同算法模型的处理耗时也不相同，无法设置统一的队列长度或等待时间。

针对每个算法运行时的 RT 来动态计算 1 秒内所能处理的请求 qps(1 秒/ 最近 100 次平均的算法耗时)，初始的队列长度为 1，根据算法的 RT 耗时进行自动拓展，超过所能承受 qps 长度的算法请求将直接被拒绝，返回 qps_limit。如果算法本身支持批量处理，那么在请求算法时会以相应批次的大小组装成 1 个请求处理。

五、问题与展望

1）机器资源总是有限的，如何最大化的提供资源利用率是后续优化的重点；

2）智能化调度也是后续的研究方向，实现自动动态扩缩容和流量调度。

海量视频解构数据全生命周期流转

作者| 阿里文娱高级开发工程师 见羽

一、背景

优酷 Smart 是一个算法服务平台，致力于提供一个完整的从标注到训练、生成服务的流程。

本文重点讲解优酷 Smart 平台上数据的使用与流转。

二、行业对比

Netflix（美国最大的 PGC 视频内容商）在 2018 年下半年陆续发了几篇文章来讲述他们内部的 NMDB 系统的设计和实现，NMDB 的全称是 Netflix Media Database，用于解决 Netflix 内部视频结构化数据的统一存储和分析问题。

优酷 Smart 数据层在设计的时候参考了 NMDB，并在 NMDB 的基础上增加了很多特性来适应优酷的业务场景，具体有以下特点：

- 为结构化数据服务（Affinity to structured data）：可定义结构化数据的 schema，对数据进行存储和索引，灵活支持查询、搜索和分析等不同需求；
- 时间线模型（Efficient media timeline modeling）：支持对媒体的 Timeline 类数据进行建模，例如视频截帧、字幕等拥有时间线属性的数据；
- 时间和空间查询（Spatio-temporal query-ability）：支持时间（截帧、字幕等数据）和空间（视频截帧部分区域数据）维度的查询；
- 多租户（Multi-tenancy）：支持不同视频源的内容存储；
- 高可扩展（Scalability）：支持模型的横向扩展，无需手动扩列。

三、整体架构

从使用人工标注数据、三方数据等数据集开始训练，到生成模型、发布上线，服务于业务之后，数据贯穿算法的始终。所以 Smart 需要、必要打造一个基于数据的闭环。

架构思路如下：

- 数据集：来自于人工标注、三方数据集的数据构成最原始的数据集
- 在开发平台上做数据的清洗、整理
- 使用清洗后的数据做算法开发
- 算法开发产生的模型发布到 Smart 服务上，供业务使用
- 算法服务产生的结果，可以供其他算法服务使用(比如图片质量分可以使用 OCR 的结果)，也可以在处理后形成新的数据集

当然，实际的数据流转情况会比上图复杂，比如开发平台不止承担了数据的清洗整理，同时还提供部分算法的在线开发、部署等功能。下面会详细介绍数据流转中各个环节的功能。

1. 数据集

数据集主要来自于三个地方：标注、三方数据集、算法服务产生的结果

1）标注平台（http://smart.youku.com）

对图片、视频进行人工打标，除提供基本的标注功能之外，还提供算法服务的使用。

- 智能标注演示
- 人体轮廓标注

2）三方数据集

youtube-8M（https://research.google.com/youtube8m/）

2. 开发平台（https://smart.youku.com）

开发平台提供一个从数据到模型、从模型到服务的线上操作环境。开发平台上共有 4 个主要概念：

- 项目：项目负责一个独立的业务模块，负责项目下数据、模型、发布的管理、隔离等
- 数据：人工标注、三方数据集、算法服务结果等数据管理，供算法训练、测试使用
- 模型：在线的模型训练、测试

- 发布：流式发布

1）项目空间；

2）流式发布。

3. 算法开发

在线、离线的算法开发过程，将在其他文章做详细讲解。

4. 算法服务

高效、稳定的算法服务，将在其他文章做详细讲解。

四、数据存储

在存储算法数据之前，先对算法数据做一个简单的分析

1. 算法数据特点

- 大：数据规模很大，百万只是起步，亿是基本操作，十亿、百亿也不罕见
- 高：数据质量很高，因为数据的质量直接决定了算法的上限
- 多：数据格式繁多，算法快速发展以来，产生了许多数据格式
- 贵：高质量的算法数据多来自于人工甚至专家标注，一份优质数据集的成本相当高

算法数据的以上特点，对存储提出了很高的要求：不流失、不蒸发。经过调研，我们找到了自己想要的金坷垃

2. TableStore-表格存储（https://help.aliyun.com/document_detail/27280.html）

TableStore 很多特点都契合了算法数据的存储要求：

- 行扩展

一行代表一个处理对象：一个视频、一张图片、一段文字或者其他复合数据

通过数据分片、负载均衡等技术，实现了数据无缝扩展。简单粗暴来说：无限容量

- 列扩展

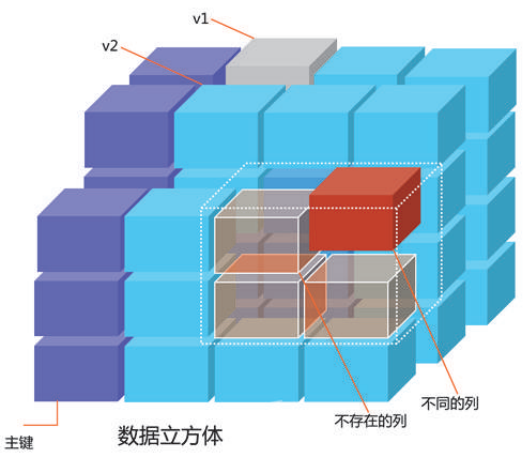
一列代表对处理对象做某个算法处理，比如对一张图片做质量分、OCR、人脸识别等算法

当要对处理对象进行新的算法处理时，需要扩展一个列。常见的数据库扩展一列相当繁琐痛苦，TableStore 是 NoSql 数据库，它的数据更像一个 Json

- 良好设计的主键：用于数据分布式存储、查询
- 其他列任意扩展
- 每行的列不要求相同

- 版本扩展

假如行代表一张图片，列代表一个算法，那么同一张图片的同一个算法还可以做多版本的存储将行、列、版本视为长、宽、高，则 TableStore 的存储方式形成了一个近乎无限扩展的数据立方体。



当然，仅仅是这样还不够，TableStore 还有其他特性：

- 高可靠：10 个 9 的可靠性，99.99999999%
- 千万 TPS，毫秒级查询
- 多级索引

那么有人要问了，这么好的数据库，我要如何使用呢，这个问题放到下篇详细讲解。

五、统一数据结构

前文讲述了数据闭环、特点、存储等，但这些做的再好，都是一个外壳。

如果每一条算法结果存储毫无章法，则该算法结果没有任何意义，因为除了算法开发者，其他人都不能理解它的意义、用法。

所以，要对算法结果做统一的数据结构处理。

上图即为优酷 Smart 设计的统一数据结构，主要包含以下几个字段

- summary: 概括结果，处理对象只会产出一个，比如一个图片的宽高
- details：可并列的算法结果，比如一个图片包含多张人脸
- shape：形状，包含各种几何形状以及 bitmap。与 coordinates 共同决定一个几何形状
- coordinates：坐标集，与 shape 功能决定一个几何形状
- innerShape：内部形状，用法同 shape。辅助 shape 使用，比如表达一个同心圆
- innerCoordinates：内部形状坐标集，用法同 coordinates
- tags：标签
- description：描述
- metaData：自定义字段

这个结构看起来很复杂，这里举几个例子帮助了解：

1. 图像质量分

很简单的算法，为图像打一个分，可以这样存储

```
{
   "summary":{
      "score":0.81
   }
}
Copy
```

2. 人脸检测

检测一张图片中的人脸，标矩形框

```
{
   "details":[
      {
         "shape":"rectangle",
         "coordinates":[
```

```
            [
                10,
                10
            ],
            [
                400,
                400
            ]
        ],
        "metaData":{
            "personName":"刘德华",
            "personId":100
        }
    }
  ]
}
```

Copy

关于数据格式有个段子：“我数了数，业界居然有 18 种格式，太难用了，我们来把它统一一下吧！”于是，第 19 种格式产生了。

要做好统一数据结构，又不增加算法同学的开发成本，最好的方式就是数据的兼容与可转换，这个也是我们后面的一个发展方向。

六、数据流转

最后对优酷 Smart 的数据流转做一个总结：

- 虚线箭头：用户看到的数据流转

从人工标注、三方数据集和其他方式获取原始数据，提供给算法开发，算法模型部署成为算法服务

- 实线箭头：实际的数据流转
  - 人工标注、三方数据集和其他方式获取的数据经过统一结构存储于 TableStore。同时已存储的数据也会辅助人工标注等
  - 算法同学使用数据做开发

- 算法服务产生数据，同时也可能使用其他算法结果来辅助自己的算法

- 最后，Smart 给自己制定了一个小目标：生成一个优酷的数据集

七、问题与展望

优酷 Smart 将不同方式产生的数据以统一结构存储于 TableStore 做数据流转，并希望搭建流畅的算法开发流程，我们遇到了或者正在、即将遇到许多问题，比如：

- 统一数据结构难以理解，结构层次太深
- 开发流程因为对算法了解不充分，做出一些不足的、过度设计的步骤

数据是算法的起点，也是算法的终点，希望优酷 Smart 在音视频方面能为大家提供一份好的数据。

3

内容智能

内容全生命周期里的文娱大脑

作者| 阿里文娱资深算法专家 牧已

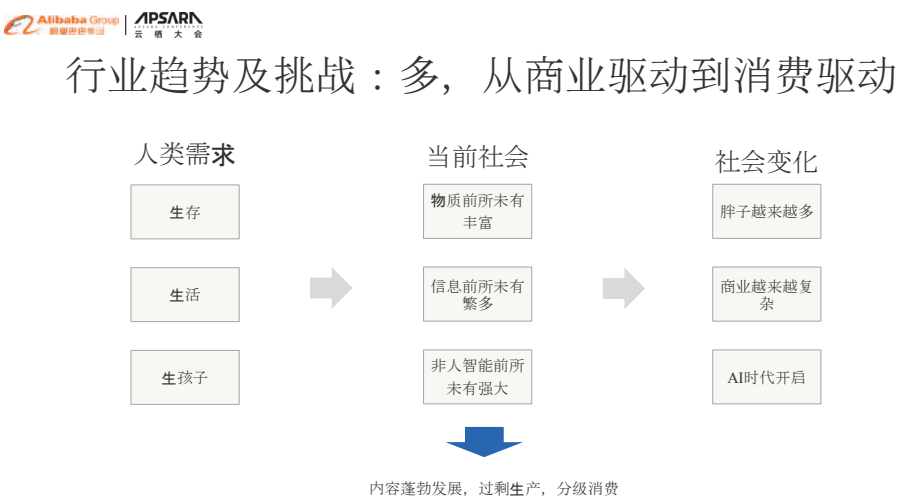
内容这个产业，2000 多年前就有人开始研究了，电影产业也发展有 100 年了，但这个产业一直面临行业的问题没有很好的解决过，内容不像商品有非常完整的量化指标体系，它是一个复杂的实体，它跟意识形态以及用户体验强相关，对内容进行量化评估和衡量是非常困难的，比如，在内容早期进行节目选角儿，如何通过数据手段来衡量我们选择角色的有效性这件事情，我们不能通过单一的指标去衡量一个演员好还是不好，我们可能需要思考这个演员的演技好不好？这个演员本身的气质和角色气质是否符合？演员的颜值是否匹配角色要求（越来越多的用户对演员的颜值有要求）？演员后续的潜力？

另外，选择的导演、主演组盘是否是最优的组合，是否能够成为爆款，这个是选择模式的问题，这件事情更加复杂困难，我们今天面临的技术挑战是我们如何进行知识的抽取，知识的

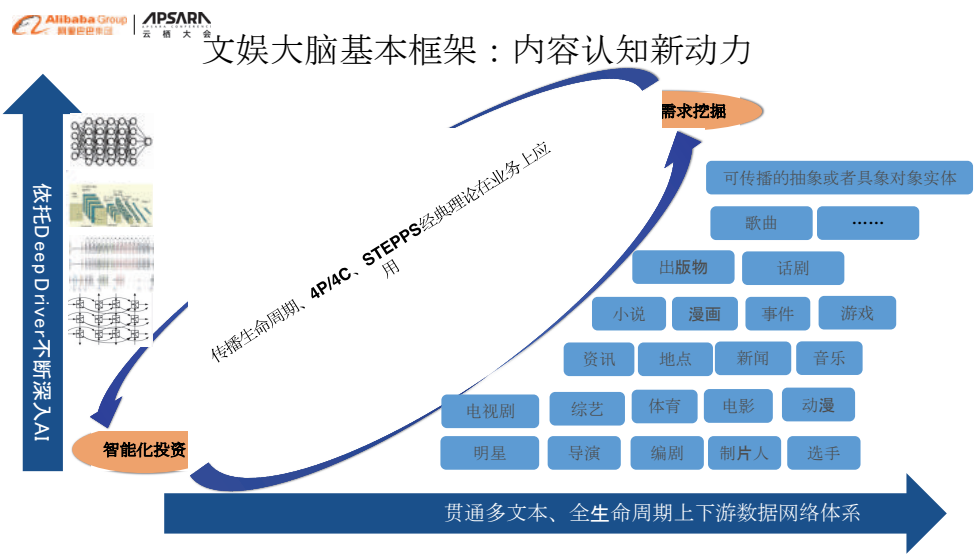
挖掘以及知识的推理，推理出来什么样的盘子是最好的，什么样的盘子的组合是最优的。

除了上述两个问题外，一部好的内容它的拍摄过程是一个庞大的系统工程，同时它也是一部艺术作品，我们以《长安十二时辰》为例，该片非群演有八百到一千人，群演有 300 到 1500 人，这里面包括服装、摄影、摄像、美术等等，历时 7 个月拍摄 217 天，如何让这个过程变成特别好的系统工程这件事情也是特别特别的困难。我们可以参考软件工程这个行业，软件工程发展了 70 年，主要研究三个层面的东西，方法论、过程以及工具，然后就是怎么将三者组合，近些年软件行业的敏捷开发对于软件工程的质量和效率都有非常好的提升，我们怎么把这些理论应用到内容制作这个产业，让内容制作敏捷起来，

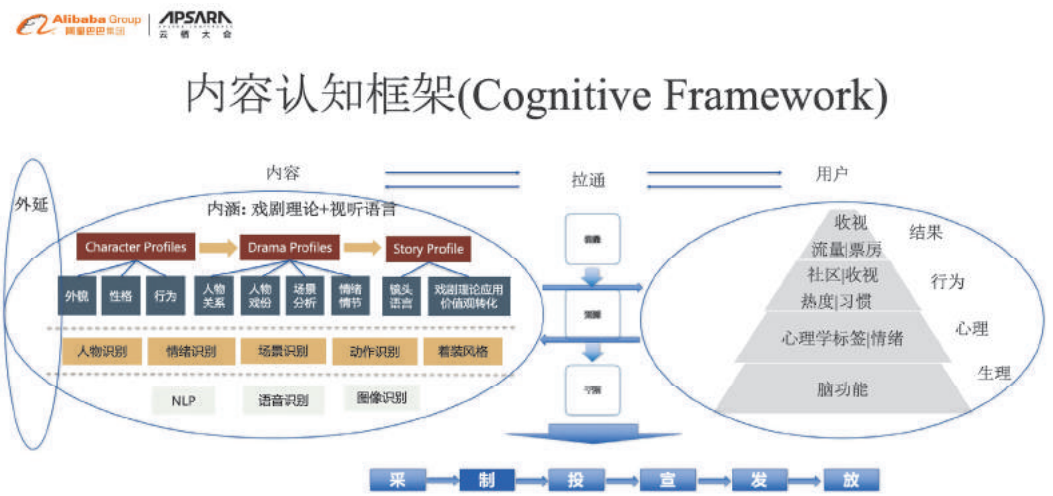
我们希望内容敏捷在过程里面会知道，过程对结果造成的影响是什么，我们可以快速的调整内容创作的过程，让它更敏捷，但是内容行业面临的独有的特点“延迟满足”，用户在内容的某一分钟特别嗨，可能来自于前面的 30 分钟铺垫在那一分钟爆发了，针对内容的这个特点，我们除了要做基本的知识图谱语义的理解之外，还要考虑如何去做有效的对应分析，如何去做对应的知识抽取等问题。



上面这些是我们在过去 2000 年当中一直存在的问题，那么今天这个问题加剧了，比过去还要复杂，全世界的 GDP 在过去的十年涨了 30%，人类有钱了，物质丰富了，人类的手机的持有量涨了 25 倍，在这种情况下促进了内容的消费和内容的需求量，在过去的 5 到 10 年里面，UPGC 加上整个内容的生产量极大的发展，用户的消费分层化，多样化，从前的全民爆款越来越少，换句话说，也许你喜欢的内容只有你那一小类人喜欢，用户对内容的需求更加个性化。相应于内容生产端，就需要考虑不同用户群的个性化需求。

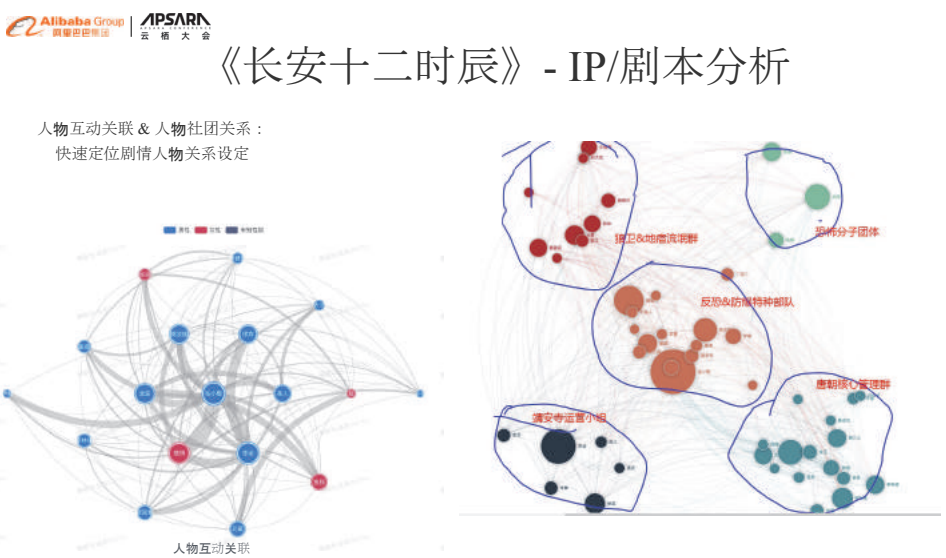


针对上面几大困难，我们今天在做文娱大脑，优酷北斗星这样的系统来解决。我们把所有的内容形式和用户消费的数据都采集下来，然后整合人工智能的技术手段，同时我们把业务里面的细分理论整合进去，然后来提供今天我们对内容上面的理解能力，基本思路如上图所示，具体怎么做我们提出了内容认知的框架。

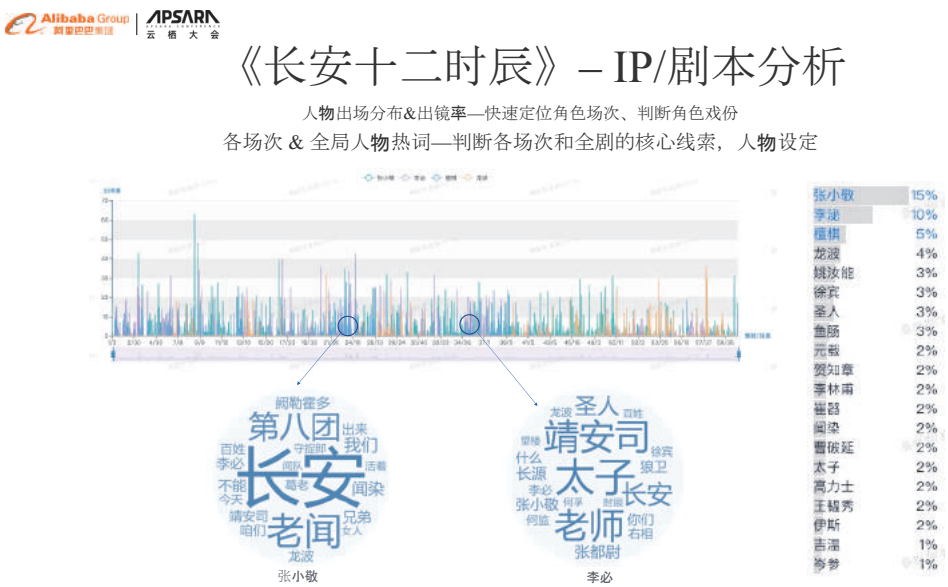


分两部分，左面是内容，右面是用户，基本的思路就是近些年心理学发展的基本的思路。内容侧，我们对内容进行理解，包括外延和内涵，外延就是内容的各种基本属性，比如主创阵容，比如题材类型等等，内涵主要研究内容的戏剧理论和视听语言，围绕制作内容的支撑要素，我们用传统的人工智能的机器学习的方式对内容进行理解，理解了之后基于戏剧理论和视听语言构造了内容的衡量要素，用户侧，分析用户的观看行为，用户的行为来自于用户的心理偏好、心理情绪，用户的心理偏好、心理情绪来自于生理构造，基于心理学的五大人格理论和用户的观看行为，构建模型建立左面和右边的连接，从而就知道创造什么样的内容，用户会有什么样的感受，基本上的思路就是这样。

基于我们的内容认知框架，落到内容的生命周期中，我们在内容生命周期的每个阶段做了一些具体的工作，已开播时间为切分点，开播前提供内容评估、艺人挖掘和内容情绪挖掘等能力，在早期为内容评估提供有效的数据支撑，在制作阶段提供了现场解决方案比之前更敏捷的反馈机制，同样在播出后也提供一些数据的支持能力去帮助更好的宣发。

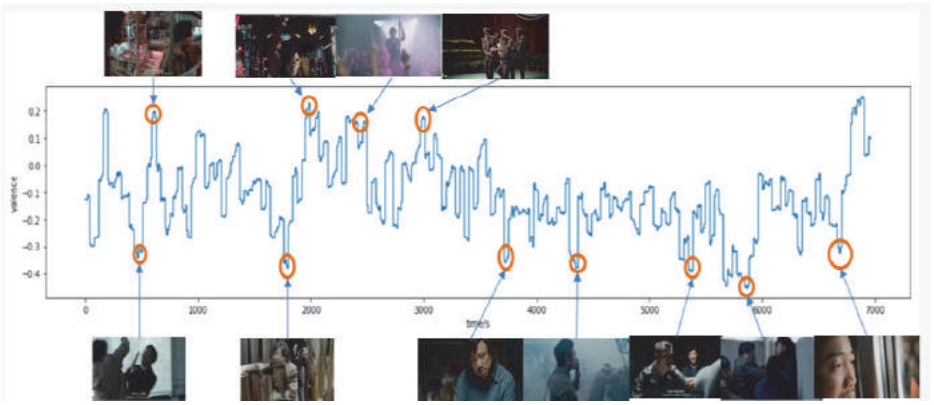


接下来我们展现一些我们在实际业务中的一些能力尝试，上图是《长安十二时辰》的剧本分析的例子，我们把已有的一些剧本作为样本，让机器去学习，识别出剧本的所有角色，把角色直接交互的对白、行为识别出来，然后再进一步，根据交互进行社团的划分，长安的剧本最终划分出来几个群体，如中间的这个群体就是反恐、特爆的小分队以张小敬为中心，下面这个是唐朝核心管理团队，就是皇上，基本上通过这样的方式能够快速的定位整个剧本的人物和人物关系的展开。

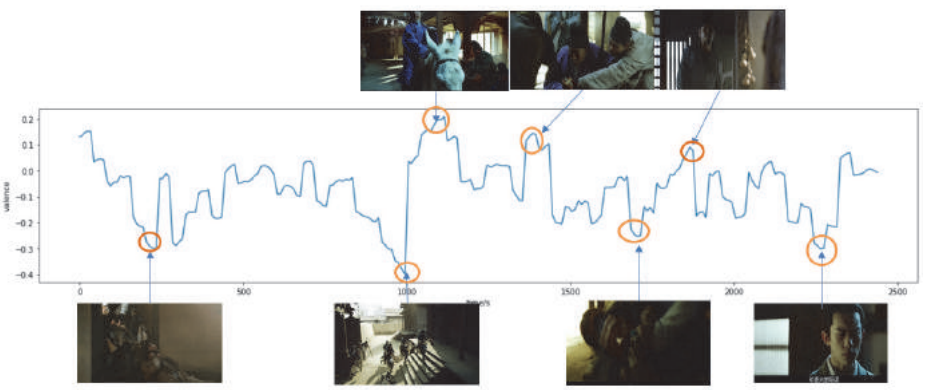


围绕上面角色的关系，我们进一步展开，可以把刚才整个剧本剧情里面的角色情绪情况也识别出来，构造成上面的曲线，然后通过分析很多的剧本，分析曲线中各个指标（出镜率、戏份、情绪值等）形成 benchmark，然后对于后面的每一个过来的剧本进行衡量，相当于是对剧本进行一个“体检”，帮助让剧本更有效。

《药神》用户情绪VA

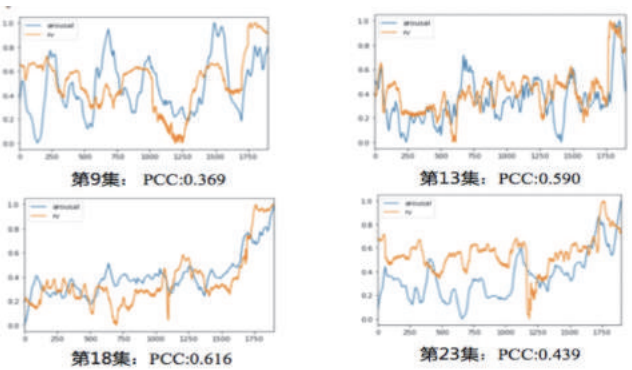


《长安十二时辰》成片情绪挖掘



同样是“体检”的方法，对于《药神》和《长安十二时辰》这两部电影我们做了一个用户情绪的识别，然后做了体检的扫描，参考零线的位置，我们看到《药神》这部片子差不多都是正向和负向级的，直到最后有一个正向区间，基本上后面以眼泪为主，伤感为主，而《长安十二时辰》这个片子的情绪状态还是比较稳定，比较沉稳的一个片子。对照情绪高低点的具体情节，我们发现，曲线表达的情绪和具体的故事情节还是很相符，很合理的。

《长安十二时辰》情绪强度预测 VS 网络收视率



然后我们拿更多的方式去验证它的合理性，上图中抽取《长安》的几集来看，每集有两条曲线，蓝线是刚才预测的情绪曲线，黄线是播放指数（表示每一秒钟有多少用户在看），通过两条曲线对比，我们可以发现，两条曲线的相关性比较高的将近 60%，情绪的高峰、低谷和最后的用户的观看行为状态是吻合的，由此我们就提供了一种能力，拿这个能力对剧本或是片子做一个情绪的扫描，我就知道你这个片子里面是不是用户真有一些喜欢的点，再对比下 benchmark，从而帮助他们更好的更高效的完成制作。

用户情绪强度预测 VS 网络收视率：综艺/电影

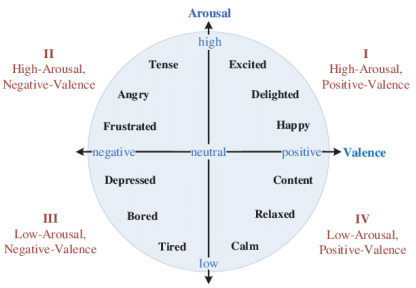
这是优酷的一部综艺，叫做《一起乐队吧》，同样做了一下体检，大家感兴趣的可以去看一下，这个片子。

基于VA的情感模型

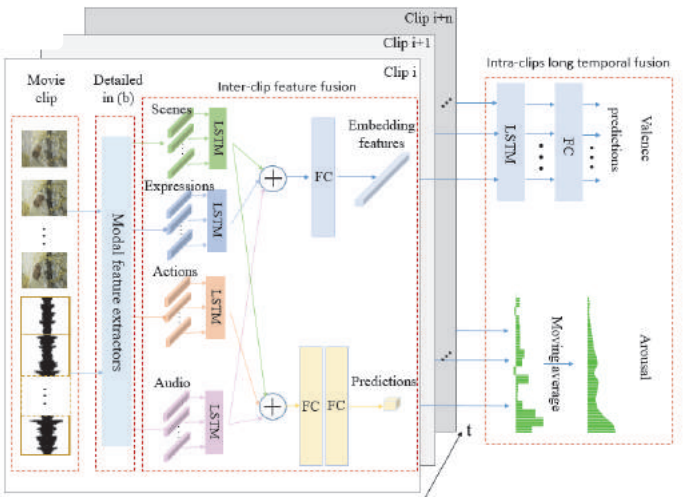
业内通用的情感模型:

Valence: 情绪正负向。-1 到 +1 之间，-1 表示负向情感，如悲伤，+1表示正向情感，如高兴

Arousal: 情绪的强烈程度。-1表示情绪最不强烈，如困乏平静，+1表示最强烈，如激动兴奋。

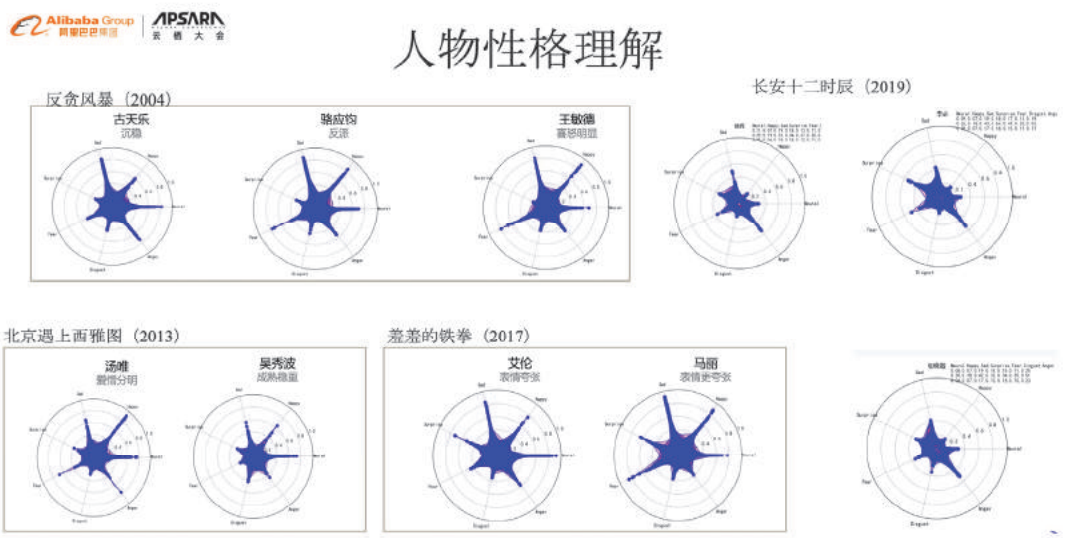


多模态的VA识别模型

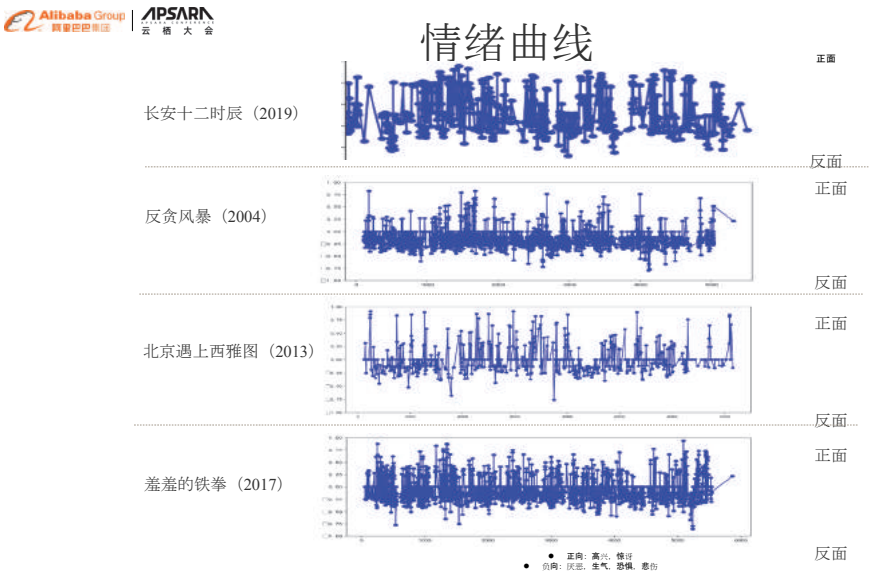


下面我们简单介绍一下用户情感曲线技术上怎么做的，首先，我们把用户观影情绪的表述，映射到认知计算中常用的二维空间表示，也就是 Valence 和 Arousal，Valence 表示情绪正负极性，Arousal 表示情感激烈程度。然后基于情绪极性跟强度提供了一个预测，这个是我们今年的

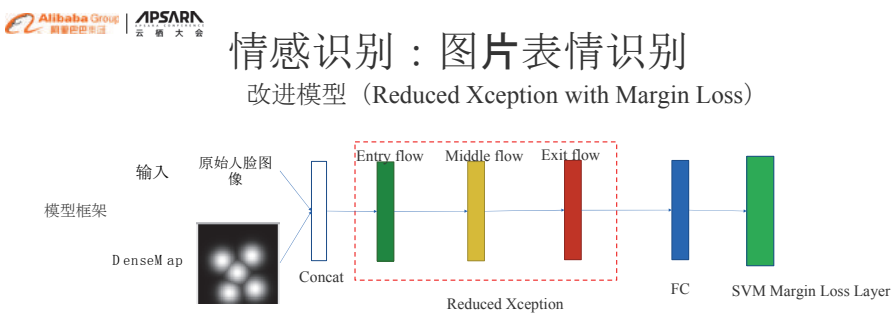
一个产出，这个论文我们已经开放出来了（论文地址：https://arxiv.org/abs/1909.01763），大家感兴趣的话可以上去看看，因为心理学最近这两年研究的核心观点是为什么用户会感同身受，这来自于前两年的一个理论叫做静向神经元，所以我们选择了场景、表情、动作以及声音作为基本的模型的输入，对模型参数进行学习。



如刚刚所讲，内容这个产业，它有强延迟满足的问题，我们目前通过两层分析来解决之前长短期满足的问题，除了上面讲的用户情绪分析外，我们也做了内容角色的情绪识别，通过图片表情识别模型，我们识别了不同题材类型的片子，可以看出来不同题材类型的片子中不同角色刻画的人物性格，港剧《反贪风暴》这么多年，主创人物形象的脸谱还是比较正的，图中显示负面角色的情绪是开心、害怕、为主的，正面形象是以悲伤、生气为主，与负面反派的开心正好相对，正面的人一直很沮丧，是一个有些压抑角色形象。



同样我们分析每秒角色的情绪，形成角色的正负情绪曲线，部分片子的分析结果曲线如上图，不同题材类型的节目会有不同的情绪密度，所以你想放松的时候，要看的不一定是喜剧，喜剧其实不一定会放松，因为角色的正负向情绪不停交替，由于延迟满足，你的大脑还要负荷的非常大，你要做长短记忆，反而很多爱情片常常对你大脑的占用确是相对低的。



输入：引入人脸关键点densemap

原理：精确判断人脸表情需重点关注五官如眼睛，鼻子，嘴的区域信息

关键点检测模型（MTCNN）Densemap 计算

$$M(i,j) = \max_{k=1}^{K} e^{\frac{(i-x_k)^2+(j-y_k)^2}{2*\sigma^2}}$$

$M(i,j)$为densmap 在像素i,j 处的强度，(x_k, y_k)关键点k的坐标，$k \in \{1,2,\cdots,K\}$

情感识别：图片表情识别

模型提升（Reduced Xception with Margin Loss）

Reduced Xception*

使用可分离卷积（deepwise 卷积 + pointwise 卷积）

Entry flow, middle flow, exit flow 各缩减至2层卷积

SVM Margin Loss Layer**

$$\min_{w} \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max(1 - w^T x_n y_n, 0)$$

$$\min_{w,\xi_n} \frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n$$

$$\text{s.t. } w^T x_n y_n \geq 1 - \xi_n \quad \forall n$$

$$\xi_n \geq 0 \quad \forall n$$

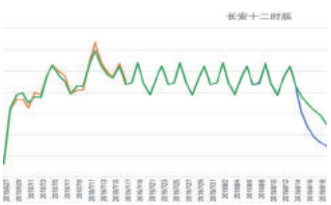
$x_n \in R^p$ 表示上一层输出，$y_n \in \{-1,1\}$ 为分类标签，w 表示本层参数

*Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions[J]. 2016:1800-1807

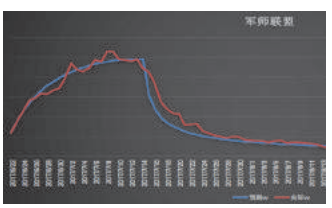
**Yichuan Tang, Deep Learning using Linear Support Vector Machines, ICML 2013

角色情绪检测是一个分类问题，所以我们用的是人脸 landmark 对于原来的初始的图象做了识别，进而生成 densemap 作为输入的一个通道，结合原始图片会使区域会更明显；合成的输入送入到 Reduced Xception 网络进行特征提取；在 loss 方面，我们引入了基于 SVM 的 marge loss，来提升各个情绪类别的类间差距，这样对于情绪的识别效果会更好，具体参考上图。

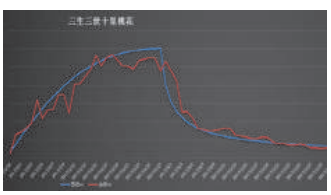
采制阶段：预测能力建设



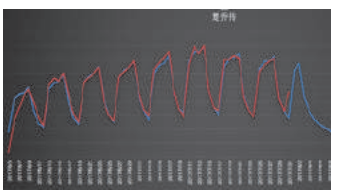
准确率92%



准确率90%+



准确率90%



准确率92%

基于前面对内容的各种理解产生的各种纬度的内容的量化纬度，我们构建了一个预测模型，可以提前预测出节目的流量情况，如前面内容认知框架中说讲，首先对内容进行量化，然后对内容相应的量化纬度进行提前的预测，进而可以更好的为业务的动作提供辅助决策。

今天整个的分享，介绍了产业当中严重的问题以及技术挑战，介绍我们通过建立文娱大脑和内容认知的框架，来尝试去解决内容行业的这些问题，最后展望一下，未来我们会花一些时间，去把人工的经验通过推理以及心理学的一些研究整合到我们的人工智能的框架下，帮助我们更好的产生内容的制高点。

《长安十二时辰》背后的文娱大脑：如何提升爆款的确定性？

作者| 阿里大文娱资深算法专家 蔡龙军（牧己）

据优酷北斗星数据显示，《长安》的“北斗星日指数”高达到 100W+，普通热门剧的“北斗星日指数”为 50-60W，是普通热剧的 2 倍。

爆款稀有，所以可贵。长视频爆款的复杂和挑战主要来源于不确定性，并且这种不确定性渗透在内容的采集、宣发和投放的所有环节中。以《长安》为例，拍摄 217 天，从定剧本、选角色、搭场景、道服化、到拍摄、后期处理，以及宣发和投放等等，每一个环节都可能影响最后效果的呈现。

一、长视频爆款的复杂与挑战：较高的不确定性

长视频内容的三大不确定性：

第一个不确定叫做延迟满足和信息不完备。长视频通过组织多个有效的事件序列，形成价值转换，刻画出不同人物，最终体现一个或多个价值观，整个过程需要很多剧集逐渐被用户感知。每个用户对于内容的偏好点和关注点不同，获取的只是内容片面的信息，信息的不完备性，导致对于内容理解的偏差。

优酷主要通过 NLP/CV/语音的语义理解&KG 等技术，进行“内容外延的解构”和“内容内核的创作理解”，获取到内容从外到内的各维度数据，保证相对的确定性。

第二个不确定是涉及工种极多的复杂系统工程。需要对复杂过程中的关键点数字化、模式化，对过程进行量化衡量&计算。

第三个不确定来自于内容本身的专业技能。如何与流量商业价值相结合，内容人在内容创作过程中会加入各种专业的技术，如在大场景的还原上，镜头语言的处理上，服化道的配置上，画面的构图上等等。但是这些技术，哪些是用户关注的？哪些是用户不在乎的？这很重要，涉及到最终的流量商业价值。所以，优酷要在技术上解决用户理解和用户心理学的问题，洞察用户偏好，将用户和内容做关联。

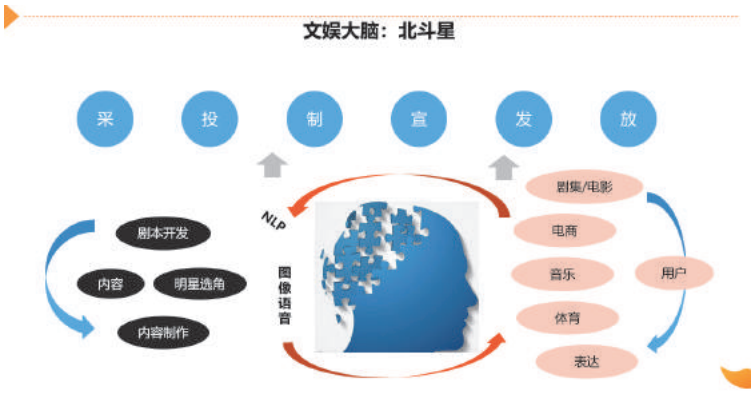
1. 内容产业发展背后的趋势思考

商业需要确定性，而内容具有极强的不确定性，如何依靠技术达到平衡？这是内容产业发展所引发的思考。

在崇尚个性化的当下，爆款也从“全民爆款”演进成“圈层爆款”，非圈层受众对某些内容完全没有感知，与之前万人空巷的气势完全不同。

二、爆款《长安》可复制吗？向算法和数据榨取确定性

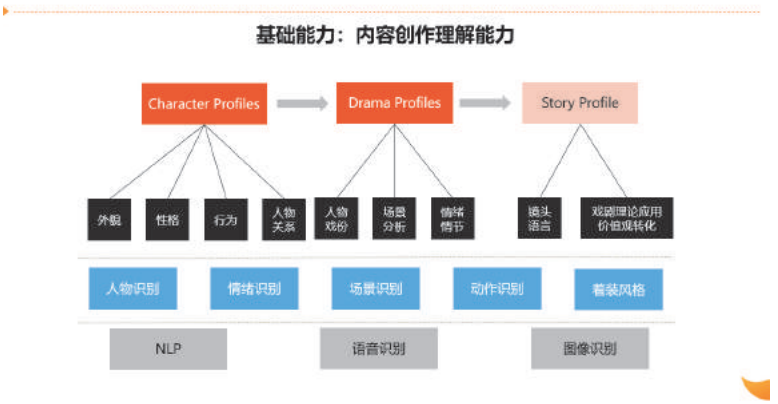
从内容的不确定性出发，优酷搭建了人机结合的智能系统即“北斗星”，它是一个具有思考能力的 AI 大脑。在采、投、制、宣、发、放的内容全生命周期中，都融入了 AI 能力，目的就是向算法和数据榨取确定性。



1. 基础能力：内容创作理解能力

处理庞大信息对于人工智能来说是“小菜一碟”，难的是提升内容创作中的理解、预测和挖掘能力。内容创作理解能力，是对剧本进行智能化的分析和挖掘。

内容主创班子是一个极强的系统化工程，在上图左侧会基于内容理解做分析和挖掘，而右侧会基于数据对左侧内容理解做“量化”，从而在内容创造阶段提供类似大脑的思考和决策能力，提高这部分的确定性。



内容创作有自身规律，内容创作理解就是围绕基于镜头语言和“两千多年的戏剧理论应用价值观”转化为技术能力，即对剧本和视频的智能理解。

以《长安》剧本分析为例，全剧本共有 120 多个人物，主创戏份评估如下：

1）张小敬的戏份占 15%，李必占 10%，檀棋、龙波、姚汝能分别占 5%、4%、3%；

2）张小敬和李必在全剧分别贡献了 90%以上的人物关系；

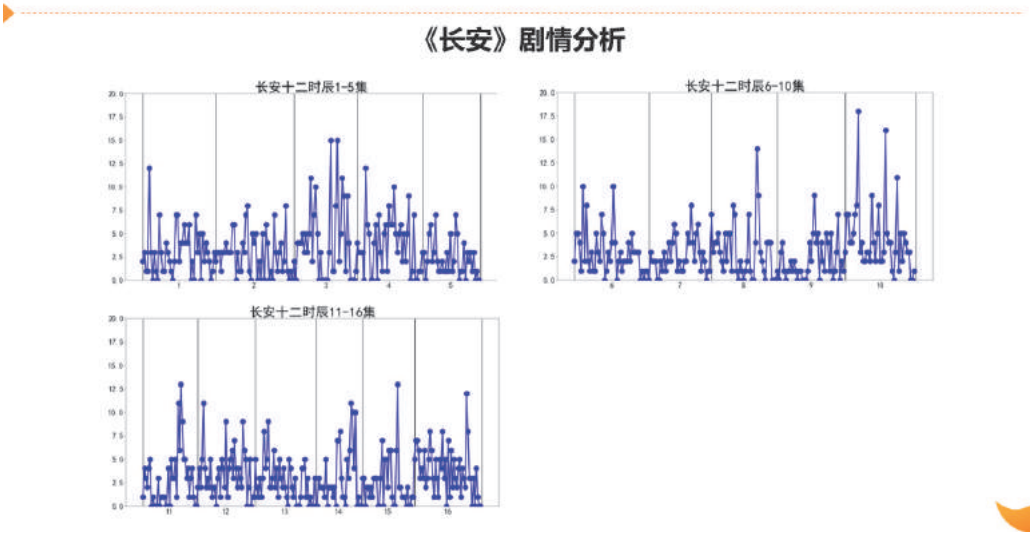
3）檀棋贡献了 80%以上的人物关系，在剧中作为功能性人物推动剧情发展。

对《长安十二时辰》剧本的角色交互分析如下：

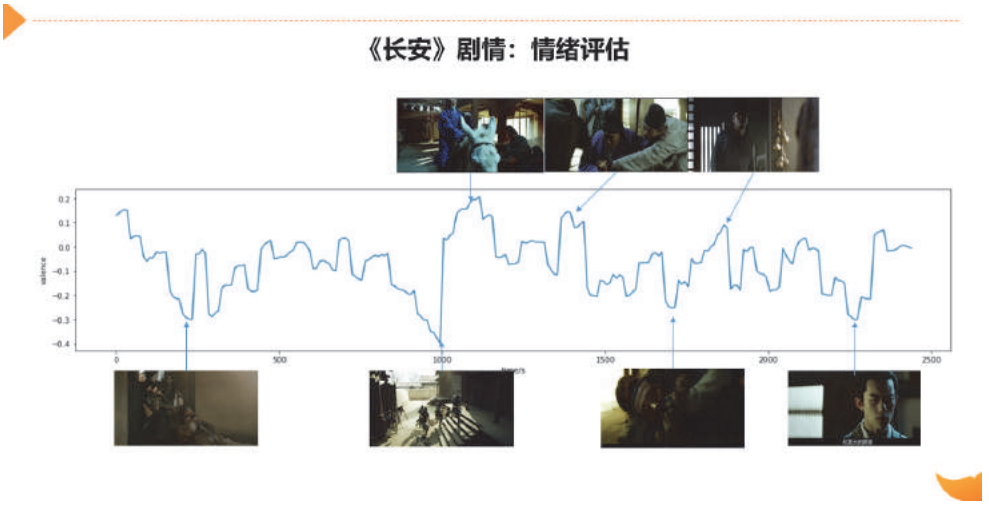
1）张小敬与檀棋的交互最多；

2）李必与檀棋、徐宾交互较多；

3）相比 IP 剧本减少了张小敬和李必的交互。

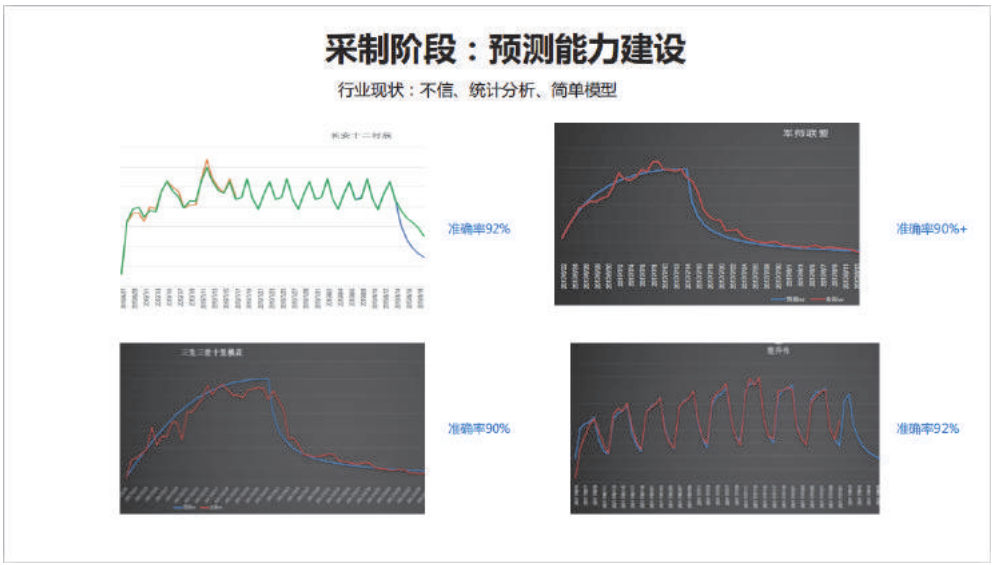


对《长安》中人物情绪进行分析发现：在前 16 集中，第 3 集和第 10 集出现了情绪表达的高峰，为剧情创造紧张情节。

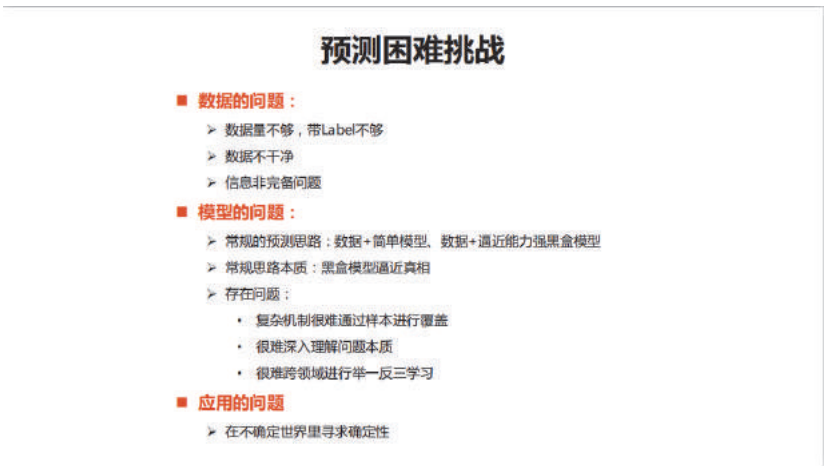


对于《长安》第一集的成片进行多模态，包括声音与图像。图像综合了演员表情、场景、动作等分析，预测出一条“用户观影情绪曲线”，后续结合用户真实观看情况对数据进行升级优化。

2. 采制阶段：预测能力建设



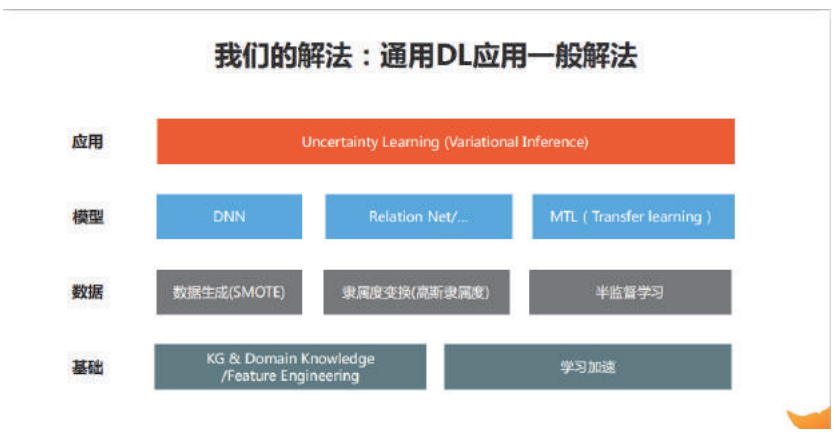
对于不确定的事情，如果可以计算出不确定性有多强，便可有效提升商业决策效率，提高决策结果的确定性。基于此，建设识别和理解不确定性的预测模型。



预测中会面临数据、模型和应用三方面问题。数据问题分为数据量不够，数据不干净和信息不完备。模型的问题包括复杂机制很难通过样本进行覆盖、很难深入理解问题本质和很难跨领域进行举一反三学习。从优酷的经验出发，是正确识别应用上的不确定性可以在应用上有很

好的改观。

常规解法也分为数据、模型和应用三方面解法。数据量由数据采样和数据生成解决，数据不干净由数据清洗解决，数据不完备由 Domain Knowledge&KG 解决。应用解法中不确定分析模型有 Belief Network 等解法。



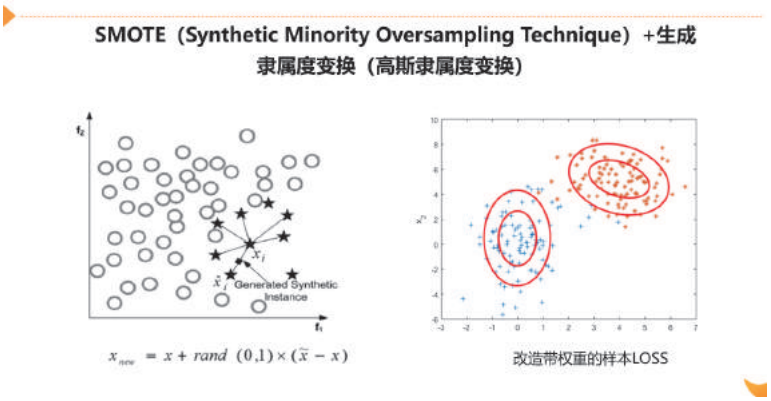
根据之前解决的问题，解法可以分为四层：

第一层是基础层。分为 KG&Domain Knowledge/Feature Engineering 和学习加速；

第二层是数据层。分为数据生成（SMOTE），隶属度变换（高斯隶属度）和半监督学习；

第三层是模型层。通过 DNN 和 Relation Net 以及 MTL 相结合，降低过拟合，提高模型的学习能力；

第四层是 Uncertainty Learning，基于变分推断的框架进行内容不确定性的预测。

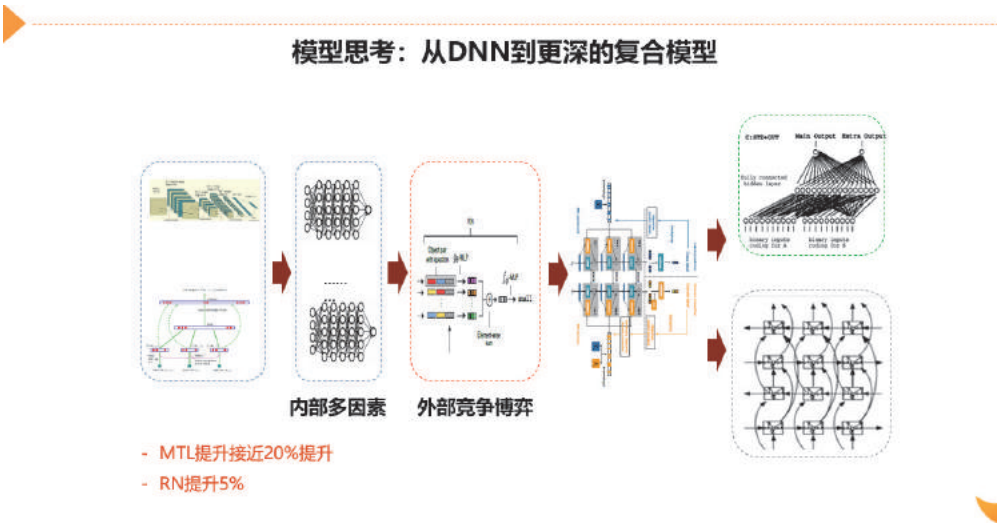


SMOTE（Synthetic Minority Oversampling Technique），合成少数类过采样技术。

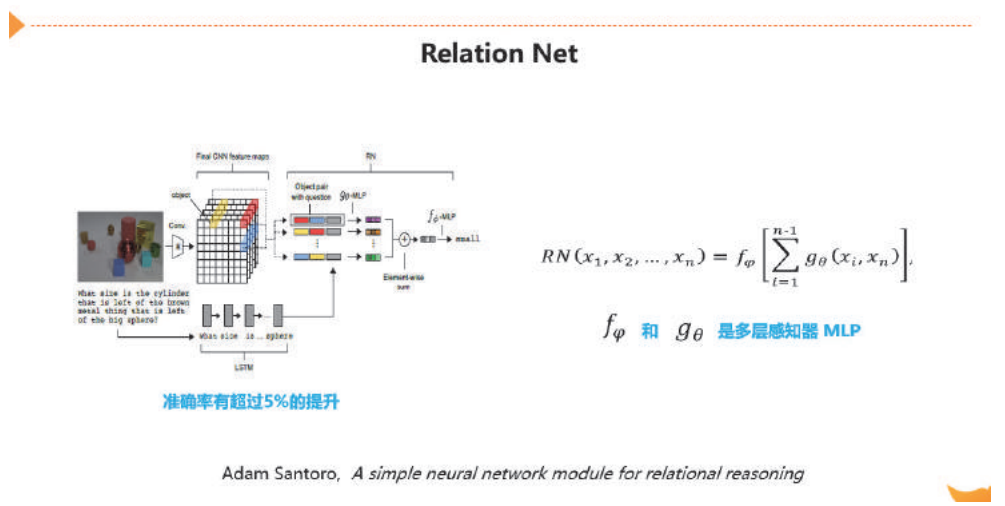
它是基于随机过采样算法的一种改进方案。由于随机过采样，采取简单复制样本的策略来增加少数类样本，这样容易产生模型过拟合的问题，使得模型学习到的信息过于特别(Specific)而不够泛化(General)。SMOTE 算法是对少数类样本进行分析，并人工合成新样本添加到数据集中，新样本的公式为 xnew=x+rand(0,1)*|x−xn|，生成的样本可直接应用到项目中，但提升效果不稳定。

优酷得到的结论是：在生成新样本后引入隶属度变换，来计算新样本与真实样本的接近程度。经验证，加入隶属度变换后，效率提升约 5%。

所有模型都会面临过拟合问题，优酷的基本思路是分析预测事件的基本特点，对于不同的特点建立不同的模型，分别有生命周期模型、竞争博弈模型和复杂影响因子。



对于复杂模型的逻辑：把前面的模型各部分的因素都拆开，复杂因素用 DNN 去拟合，外部竞争的关系去 Relation Net 做推理去解决，最后用 MTL 整合模型，根据实际情况也会加入其它模型。



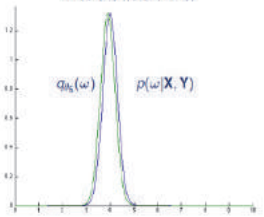
Relation Net 是 2016 年发表的 CNN 模型。基本思路是将包含各种圆柱、椭圆等形状的图片，经由 CNN 网络输出生成 feature maps，把图中涉及到形状的 object 通过通道取出，每两个 object 配对形成一个对比串，然后与 LSTM 编码 question 的 embedding 向量叠加到一起，输入到一个深度网络中进行学习，最后 softmax 分类到某个答案词上面，进行正确与否的判断。

Uncertainty Learning 这块，从 2016 年开始它逐步热起来，我们也用变分去做了一些事情。

Variational Inference

- Approximate $p(\omega|\mathbf{X}, \mathbf{Y})$ with simple dist. $q_\theta(\omega)$
- Minimise divergence from posterior w.r.t. θ

$$\mathrm{KL}(q_\theta(\omega) \,||\, p(\omega|\mathbf{X}, \mathbf{Y}))$$



Yarin Gal, Uncertainty in DL 2015

- Def $q_\theta(\omega)$ to approximate posterior $p(\omega|\mathbf{X}, \mathbf{Y})$
- KL divergence to minimise:

$$\mathrm{KL}(q_\theta(\omega) \,||\, p(\omega|\mathbf{X}, \mathbf{Y}))$$
$$\propto \boxed{-\int q_\theta(\omega) \log p(\mathbf{Y}|\mathbf{X}, \omega) d\omega} + \mathrm{KL}(q_\theta(\omega) \,||\, p(\omega))$$
$$=: \mathcal{L}(\theta)$$

- Approximate the integral with MC integration $\hat{\omega} \sim q_\theta(\omega)$:

$$\tilde{\mathcal{L}}(\theta) := -\log p(\mathbf{Y}|\mathbf{X}, \hat{\omega}) + \mathrm{KL}(q_\theta(\omega) \,||\, p(\omega))$$

Minimise divergence between $q_\theta(\omega)$ and $p(\omega|\mathbf{X}, \mathbf{Y})$:

- Repeat:
  - Sample $\hat{\mathbf{z}}_{i,j} \sim \text{Bernoulli}(p_i)$ and set

$$\hat{\mathbf{W}}_i = \mathbf{M}_i \cdot \mathrm{diag}([\hat{\mathbf{z}}_{i,j}]_{j=1}^{K_i})$$
$$\hat{\omega} = \{\hat{\mathbf{W}}_i\}_{i=1}^{L}$$

  - Minimise (one step)

$$\tilde{\mathcal{L}}(\theta) = -\log p(\mathbf{Y}|\mathbf{X}, \hat{\omega}) + \mathrm{KL}(q_\theta(\omega) \,||\, p(\omega))$$

  w.r.t. $\theta = \{\mathbf{M}_i\}_{i=1}^{L}$ (set of matrices).

Variational Inference (last page)

Instead, **predictive mean**, approx. with MC integration:

$$\mathbb{E}_{q_\theta(\mathbf{y}^*|\mathbf{x}^*)}(\mathbf{y}^*) = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{y}}(\mathbf{x}^*, \hat{\omega}_t).$$

with $\hat{\omega}_t \sim q_\theta(\omega)$.

- In practice, **average stochastic forward passes through the network** (referred to as "MC dropout").[5]
- Dropout after convolutions and averaging forward passes = **approximate inference in Bayesian convnets.**[6]

$$\boxed{\mathrm{Var}(\mathbf{y}^*)} = \tau^{-1}\mathbf{I} + \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{y}}(\mathbf{x}^*, \hat{\omega}_t)^T \hat{\mathbf{y}}(\mathbf{x}^*, \hat{\omega}_t) - \mathbb{E}(\mathbf{y}^*)^T \mathbb{E}(\mathbf{y}^*)$$

结论：

- 实验效果，VI更好
- Laplace逼近的方式假设条件更多，而VI更加严谨
- 另外逻辑上讲，逼近Laplace的函数是选择最优解，而VI更像Ensemble

这一部分可以在网上参考“贝叶斯 Network”，重点看它如何利用“变分”得到最后结果。

3. 宣发阶段：挖掘能力建设

挖掘能力更多应用于已经发生的事件，使其更具有确定性。上图是《长安》播出后，每一分钟用户的收视状况、复看状况和弹幕状况，再结合每一时间段的剧情内容对用户喜好做更精准的分析，以此来做更好的宣推和挖掘。

内容产业是个不确定性非常高的产业，越是爆款就越有不确定性。互联网下半场我们积累了特别多的数据，AI 能力也得到了前所未有的发展，我们建立了“文娱大脑”北斗星、AI 剧本等内容形式的挖掘能力，和采买不确定性预测的评估能力，以及对于宣发挖掘的能力，都在业务应用上取得了不错的成绩。

传统的内容制作领域，依然依赖人的经验，在强人工智能尚遥远的情形下，如何结合机器 AI 和人工经验是个永恒的主题，例如结合符号主义（计算机学派）和连接主义（仿生学派）的智能。因此，我们将在两个方向发力：1）决策引擎的建设，包括结合人工逻辑规则和机器学习，不确定性分析框架和经久不衰的贝叶斯因果决策，以及神经元化的混合智能计算框架。2）量化的心理学研究也越来越重要，我们也会推进这部分探索。

4

搜索

智能多轮对话式搜索技术实践

作者| 阿里文娱资深算法专家 刘尚堃

一、问题定义和分析

用户在应用大屏收看视频的时候常常面临一个困难就是“不知道看什么，并且不知道如何搜索”的问题，针对用户这个痛点，优酷人工智能平台提出了基于多轮对话式搜索系统。

交互式搜索系统采用模块化的设计思路，按照分层逻辑结构，分为应用技术层、核心技术层和基础数据层共三个部分。应用技术层主要包括是自然语言理解（NLU）和对话技术，其中NLU 包括意图理解（Intent Understanding）技术和成分分析（Slot Filling）技术；对话技术包括对话管理（DM）以及对话生成（NLG）。核心技术层包括知识图谱（Knowledge Graph）的构建和推理应用。基础数据层是基于视觉技术的智能媒资库。

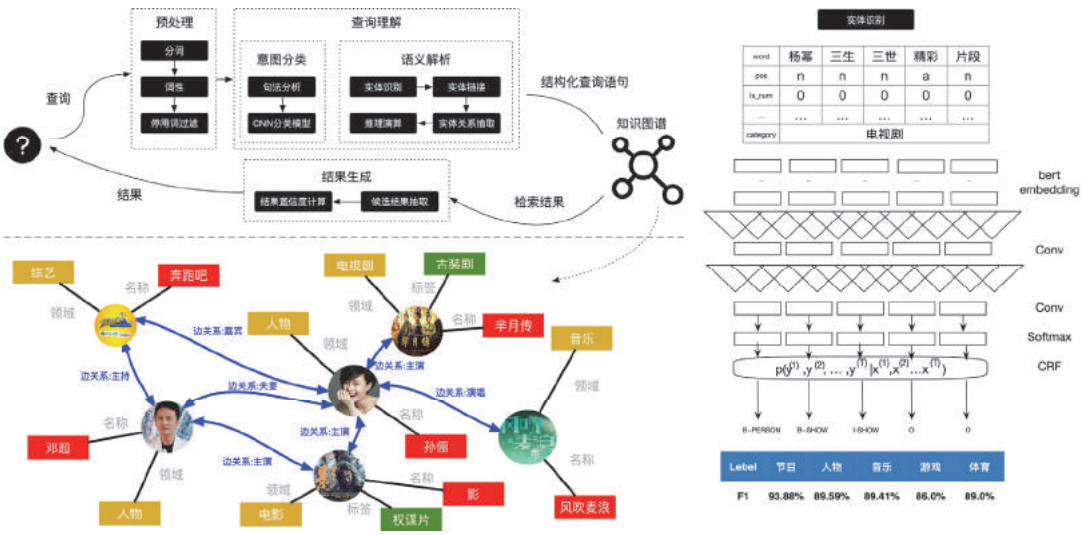
1. 自然语言理解（NLU）

在 NLU 中，意图代表用户想要达到的目的，就是在语言表达中所体现出的“用户想干什么”，解决的是人与人，人与机器之间的交互问题。自然语言理解的结果，就是要获得一个语义表示（semantic representation），常用的是框架语义表示（frame semantics）的一种变形：采用领域

（domain）、意图（intent）和属性槽（slots）来表示语义结果，其中，领域是指同一类型的数据或者资源，以及围绕这些提供的服务，比如“视频”，“歌曲”，“购物”等；意图是指对于领域数据的操作，一般以动宾短语来命名，比如视频领域中，有“检索”，“播放”、“快进”等意图；属性槽用来存放领域的属性，比如视频领域有“节目名”，“演员”，“角色”等；槽位填充为了让用户意图转化为用户明确的指令而补全信息的过程。

query 成分分析技术，依托实体知识图谱提供的节目/人物/角色等文娱数据，实现用户需求各维度的成分理解，满足用户对视频内容多维度的检索需求，是属性槽提取的基础技术。如“我想看易烊千玺跳地板舞的视频”，成分分析的结果是“易烊千玺:人物”，“跳:动作“，“地板舞:舞种”，在此结果上构建召回策略实现结构化检索。

意图理解技术，基于成分分析技术的全面理解以及完善的意图分类体系，精准识别用户查询意图。意图分类体系的构建是意图识别的基础，在此基础上我们构建基于 CNN 的意图分类模型去实现意图的基础理解。



2. 对话技术

对话技术包括对话管理（DM）以及对话生成（NLG），DM 是对话系统的大脑，它主要干两件事情：

1）维护和更新对话的状态。对话状态是一种机器能够处理的数据表征，包含所有可能会影响到接下来决策的信息，如 NLU 模块的输出、用户的特征等；

2）基于当前的对话状态，选择接下来合适的动作。举一个具体的例子，用户说“我要看香港电影”，此时对话状态包括 NLU 模块的输出、这次搜索的结果信息以及用户历史等特征。在这个状态下，系统接下来的动作可能有几种：1）向用户确认筛选指令，如“请问您想看谁主演的电影”等；2）向用户询问播放详细节目，如“请问您想播放第几个影片”；3）根据用户兴趣推荐具体节目或者更换搜索词，如“没有找到 xxx，为您推荐 xxx”。

从技术实现维度，常见的 DM 实现方法有如下几种：

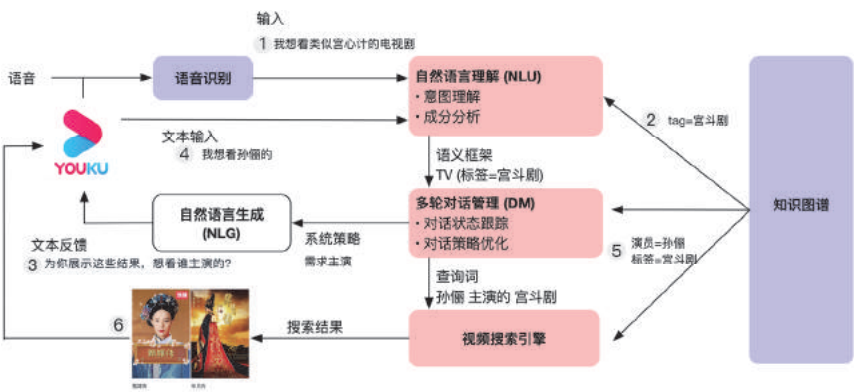
第一种是基于有限状态机（FSM），显示的定义出对话系统应有的状态,基于 FSM 的 DM，优点是简单易用，缺点是状态的定义以及每个状态下对应的动作都要靠人工设计，因此不适合复杂的场景；

第二种采用基于统计的方法,可以是马尔可夫决策过程或强化学习；

第三种是基于神经网络的，它的基本思路是直接使用神经网络去学习动作选择的策略，即将 NLU 的输出等其他特征都作为神经网络的输入，将动作选择作为神经网络的输出，不再需要人工去显式的定义对话状态。

用户在视频这个垂直领域，意图和属性槽相对比较明确，整体以有限状态机的方法为基础，基础动作迁移状态以人工设计动作为主；模型的方法作为泛化能力，解决不确定场景的理解。

系统对话流程如下，用户说“我想看类似宫心计的电视剧”，系统先后通过语音识别（ASR）和自然语言理解（NLU）技术理解分析用户想看‘宫斗剧’，通过检索反馈给用户‘宫斗剧’相关电视剧，并通过自然语言生成（NLG）技术主动和用户作进一步的交互，得到用户想看‘孙俪’主演的需求后，系统基于多轮对话管理（DM）技术将前后两轮的用户综合理解，向搜索引擎发起再次检索实现多轮交互。

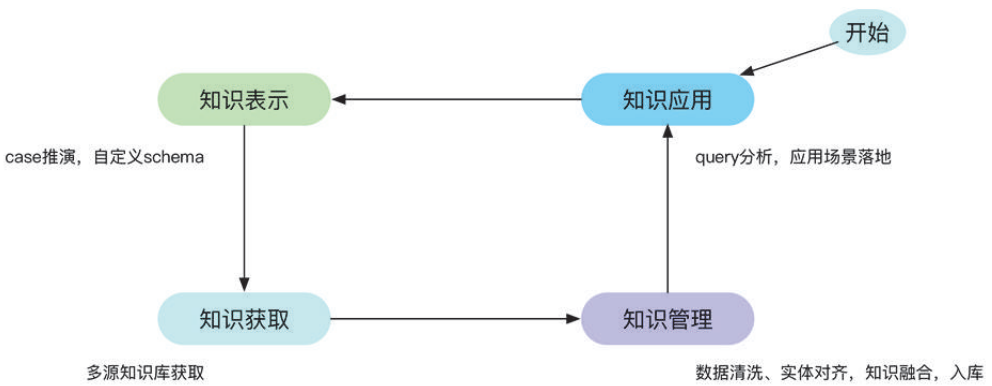


3. 知识图谱构建及应用

知识图谱（Knowledge Graph）本质上是一种大规模语义网络，是一种基于图的数据结构，由节点（Point）和边（Edge）组成。在知识图谱里，每个节点表示现实世界中存在的“实体”（Entity），每条边为实体与实体之间的“关系”（Relation）。当知识图谱聚焦在特定领域，就可以认为是领域知识图谱（Domain-specific KnowledgeGraph：DKG）。比如“文娱知识图谱”，里面大多都是跟文娱相关的实体和概念。领域知识图谱虽然在广度上不及通用知识图谱，但在深度和粒度上，DKG 通常表现更为优秀。比如在文娱领域，追星族们可能更关心“内地 90 后演员获金鸡奖的电影”，DKG 就可以在更深更细的层面上为垂直搜索赋能。

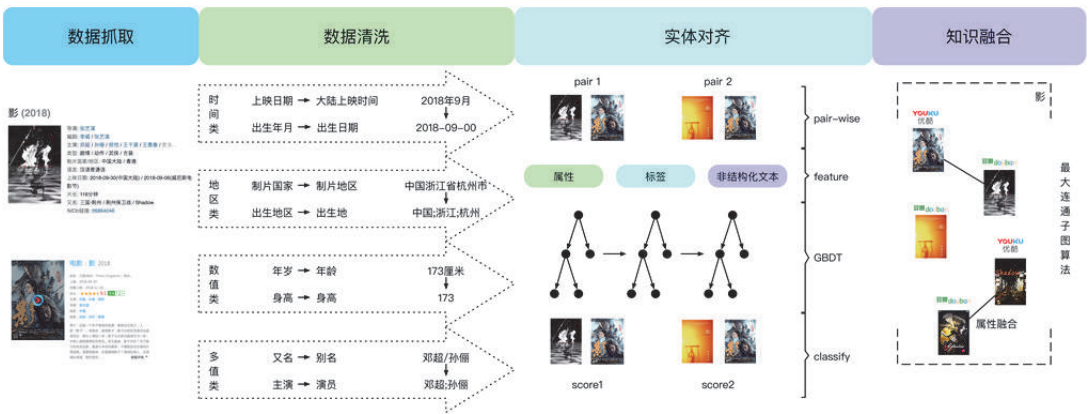
在行业智能化的实现进程中，通过领域知识图谱对数据进行提炼、萃取、关联、整合，形成专业的领域知识，让机器形成对于行业工作的认知能力，从而实现一个专业领域的知识引擎，实现专业领域知识的工作自动化，将对提高机器生产力有着巨大的价值。

领域知识图谱系统的生命周期包含四个重要环节：知识表示、知识获取、知识管理与知识应用。这四个环节循环迭代。优酷知识图谱的初期构建和领域知识图谱的构建流程严格一致。从最开始明确知识的应用场景，通过客观评估场景收益、人力资源消耗、技术与应用的适配程度等一系列问题，最终选择落地场景；再通过应用来进行 case 推演，反推出知识图谱需要怎样的知识表示，明确 schema 和知识边界，完成模式层的构建；进而根据知识边界选取特定源的知识库进行知识获取并根据获取的知识进行整合、管理。最终完成了优酷在文娱领域的知识图谱构建



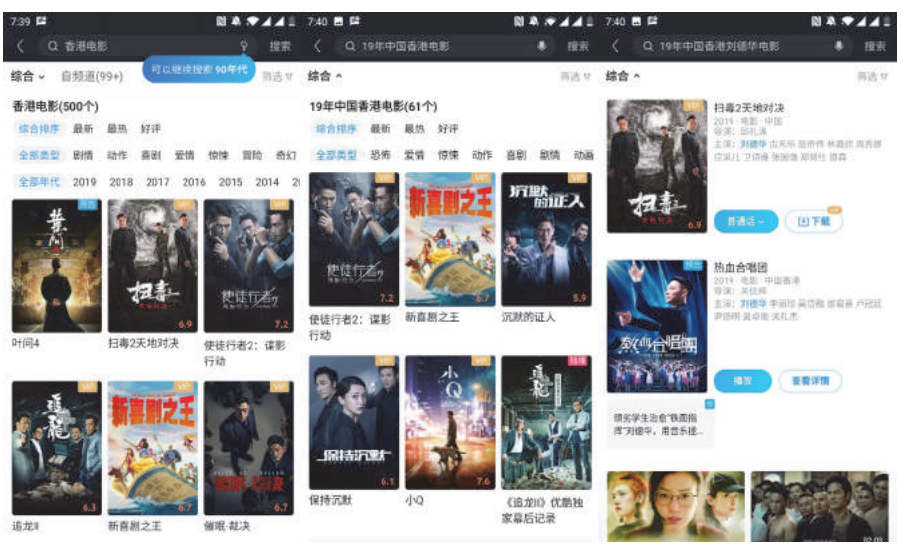
在明确模式层和知识边界后，圈定了知识库的来源，比如内部底层数据的转换、来自开放互联网的领域百科抓取、来自兄弟部门的数据拉通，通过优先选择数据结构化程度相对较优的

数据源，来降低知识获取、知识清洗带来的人力成本。根据这些数据，就可以开始领域知识图谱的数据层构建。数据层构建大致分为 4 个步骤：数据抓取、数据清洗、实体对齐、知识融合。总体流程及相关细节如下图所示：



二、多轮对话式搜索效果案例

多轮对话式搜索首先在优酷 OTT 端落地实现，随后通过将这一服务集成到 AI SDK 扩展到优酷主 APP 等各个应用服务上，整体效果提升非常明显，用户使用率提升 30+%，语音搜索量也有很大的涨幅。



优酷语义模态匹配模型设计与实现

作者| 阿里文娱算法专家 惟秋

一、业务问题

在优酷搜索中，查询词（query）与视频的相关性一般包括文本模态、图像模态以及视频模态三个方面。文本模态中，我们主要解决用户的查询词和视频标题、视频上传用户以及视频其他文本信息的相关性。然而从 query 维度来看，优酷搜索有 90%多的 query 是长尾词，相关性无法获取很多有效的统计特征；从视频维度来看，大量普通用户上传的视频（UGC）命名比较随意，标题中经常出现口语化的表达以及无用信息等。这就对文本相关性提出了极高的要求。因此，我们需要引入语义匹配来对基础文本匹配（词语匹配）进行补充，从而提升最终的搜索效果。

二、技术定义

语义匹配是用来解决传统字面匹配无法解决的用户意图（查询词）与文档表述中语义鸿沟问题的技术。传统语义匹配模型主要是通过各种方法将 query 和文档转化到同一空间进行计算，典型的方法如 pLSA，LDA，SMT 等。深度文本语义匹配模型根据 query 和文档交互时间和内容主要分为三种：表示型、交互型、表示&交互融合。表示型深度语义匹配模型主要代表有 DSSM，ARC-I，CNTN 等。交互型语义匹配的模型代表有 ARC-II，MatchPyramid 等。BERT 以强大的知识迁移能力以及学习能力让语义匹配效果上又上升了新的台阶。

我们将从样本构建以及模型演化两方面简单介绍下优酷搜索在语义匹配方面的工作。

三、样本构建

一方面数据和特征决定了机器学习的上限，另一方面深度学习需要大量的训练数据。虽然对搜索、推荐场景来说拥有丰富的用户行为数据。但是在相关性任务上直接利用用户反馈进行

样本构建是行不通的。首先相关性模块离最终给用户展示的文档排序较远，后续有多个模块影响排序结果，不能单纯利用用户行为来判定文档相关性；其次相关性模块需要处理的文档是底层引擎返回的全部文档，而用户看到的文档则是经过相关性截断以及排序后的结果，因此二者在分布上是不一致的。再次在视频搜索领域用户猎奇、浏览心智较强，用户的行为不能完全反应查询词和文档的相关性。

但是所有数据都交由人工标注也是不切实际的。因此在数据集的构建方面，我们采用主动学习来提升迭代速度。首先我们利用开源数据集以及平时积攒的高质量样本集合训练基础模型。考虑到准确率要求和超参引入，我们没有采用主动学习力的不确定采样来挑选后续标注样本而是利用人工标注前一版模型在待标注集上预测结果的抽样，对准确率很高的区间内样本直接进入训练集合，其他样本交由人工标注从而形成第二轮迭代。在后续的代过程中，我们会根据当前模型预测错误的样本分布不断调整样本采样策略，确保特征的覆盖率以及样本分布的均衡。

四、模型设计

一般而言，语义匹配 中有四大难题即一词多义、多词同义、结构组合、表达多样（表 1）。传统方法解决语义问题往往需要人工进行大量繁琐的工作进行特征挖掘。例如，构造主题模型解决一词多义的问题，挖掘同义词解决多词同义问题。然而在数据清洗、特征构造、线上应用的过程中误差会不断累积放大，这就势必造成最终应用效果有所打折。

表 1. 语义匹配中的难题及传统解决方法

挑战	例子	传统解决方法
一词多义	笑傲江湖（综艺，电影，电视剧）	意图识别、主题挖掘
多词同义	德罗巴（非洲刘德华）	同义词挖掘
结构组合	志明与春娇（春娇与志明）	term 紧密度、匹配顺序
表达多样	英语书写（英语怎么写才好看）	核心词，term 赋权

深度学习的出现让我们能端到端的解决问题成为可能。CNN 利用 N-Gram 多通道方式对句子进行特征提取更多获取的是局部信息，擅长建模短距离的依赖关系；RNN 通过逐步递归（前向或者后向）抽取句子特征能直接建模前后顺序、长距离依赖的关系。因此，我们就尝试引入 RNN 来解决视频搜索中的语义匹配问题。

1. 基于 BiLSTM + Attention 的语义相关性模型

考虑到下游任务的通用性，我们是参照 DSSM 设计了第一版语义相关性模型。除了得到

query 和视频标题的相关性得分，它的副产物（文档表达）也可以应用于其他业务。

在选取合适的训练样本后，语义相关性模型的最大挑战来自于 OOV （Out-of-vocabulary）。DSSM 通过 letter trigram 来增加词典的覆盖率。然而不同于英文单词词干就带有单词的语义信息，词组或者单字是中文中包含语义信息的最细粒度单元。

在视频搜索领域，我们还面临额外的挑战。一是新词出现的频率很高包括新出现的影视节目以及网络用语，二是词语本身也会经常出现转义。例如“镇魂”这个词，前一段时间可能是“镇魂街”或者“镇魂曲”中的子词，但是下一刻可能是独立的一个实体，指代《镇魂》这部电视剧。另外，优酷站内有相当一部分的视频标题包含数字（表示日期、期数等）或者英文（人名、音乐名等），还有很多 UGC 英文标题不规范造成分词困难。

因此我们利用不同的方式对英文、数字和中文分别进行预处理。对英文和数字，我们参考 DSSM 进行 letter trigram 处理。虽然在一定程度上这会损失部分语义信息，但是可以极大的缩小词表从而减少 OOV 情况。对中文我们则尽可能利用最大粒度分词确保语义的无损。当大粒度分词不在词表里时，我们转向中小粒度分词甚至单字以提升词典的覆盖率。

模型层面，为了形成更准确的文档语义表达，我们对 DSSM 中全连接层进行了改造。虽然 CNN 通过 N-Gram 能学习到部分匹配顺序和短距离的信息，但是更长距离的依赖关系则需要通过 RNN 来建模。在输入层后面，我们引入用 BiLSTM 来获取每个单词的上下文信息，于是每个单词的表征可以表示为前向和后向序列隐藏状的拼接 $[h_{fi}, h_{bi}]$。

得到输入每个词的向量之后，大家通常利用池化（pooling）来得到句子的表示。常用的池化方法如均值池化，最大值池化，最后池化都无法区分文档中的主次关系。在搜索这个场景下，人们对一个文档的兴趣点通常来自几个核心词。因此我们引入全连接网络作为注意力网络（attention）来学习每个单词的权重：

$$\mathrm{weight_i} = \mathrm{FFN}([\mathrm{h_{fi}}, \mathrm{h_{bi}}]).$$

传统 IDF 得到的权重是全局信息，而此处学习到的权重包含了词的位置及上下文信息，能对当前单词在文档中的关系的进行更精细的刻画。当然我们也可以在注意力网络中引入位置、词性等其他信息。于是，文档的最终表示可以由每个单词表达的加权求和生成。

$$\mathrm{Rep_{sent}} = \sum_{i \in sent} a_i [h_{fi}, h_{bi}],$$

$$a_i = \frac{\exp(weight_i)}{\sum_{j \in sent} \exp(weight_j)}.$$

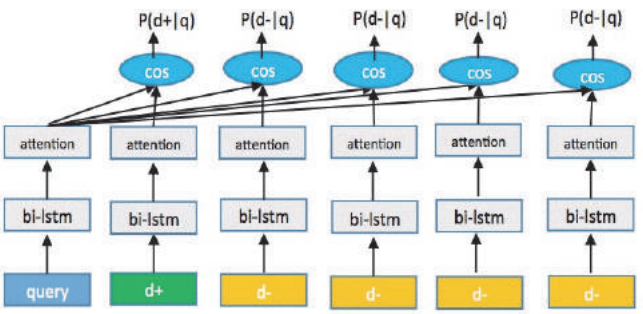


图 1. BiLSTM+Attention 网络结构

得到 query 和文档的表达后，二者的距离可以由 cosine 函数表示。最终的损失函数定义为：

$$J(\theta) = -\sum \log \frac{\exp(\gamma \cos(q, d^+))}{\exp(\gamma cos(q, d^+)) + \sum \exp(\gamma \cos(q, d^-))}.$$

为了提升模型的泛化能力，在训练样本里我们加入了神马的搜索日志，最终训练数据达到了 8000 万。在测试集上，人工评测 g:b=2.2：1；模型上线一周后，长尾流量上平均跳出率下降 0.58%。图 2 是我们给出的两个例子。从图中可以发现，加入了 BiLSTM 后，模型能够捕获到 query 中蕴含的主谓关系以及主题信息，基准桶里“法国”不是核心词甚至只是被提及语义跟 query 意图大相径庭的问题得到了有效解决；而表征的引入也能解决 term 没有匹配时的相关性问题。

图 2. 左边两张图对应“法国输球”；右边两张图对应着“小腿酸胀怎么回事”

2. 基于 bert 的语义相关性模型

在搜索中特别是长尾流量，要获得 query 的意图以及 query 与文档的相关性需要大量的先验知识。例如在优酷搜索中大部分流量是影剧综相关的，对长尾”强化学习”这样相关的文档则无法很好建模。单纯的 Word Embedding 虽然能进行大规模语料学习，但是它学到表示的是静态的，无法很好的表达不同语境下的不同意思。RNN 的串行特点限制了模型的可扩展性和规模。

Bert 的双向 transformer 特征提取器具有良好的并行能力和可扩展性。在 bert 第一阶段利用大规模语料可以获取的丰富的先验知识能够大幅提升第二阶段微调任务的效果。此外 bert 中自注意力模块能够直接学到句子中的 term 权重信息，无需添加新的注意力模块就能产生更准确的文档表达为下游任务服务。

利用 Bert 进行语义相关性计算，通常有两种方法：一是交互型 bert，query 与待计算文档用分隔符 SEP 拼接起来送入 bert，最后输出的 CLS 向量送入前向网络进行特定任务微调（图 3）。二是表征性 bert，query 与 doc 分别送入 bert 得到二者的表示，然后再把他们送入特定的深度网络进行相关性计算。大量文献和实践表明，在相关性或者相似性任务上，交互型 bert 的效果要远远超过表征型 bert。

另一方面，在落地 bert 的过程中大家一般从两方面着手提升具体任务的指标。一是预训练，以一般性语料（如 wiki 或者百科）为基础结合行业垂直领域知识以及实体信息进行预训练是提高模型效果最直接也是最明显的手段，但是它受限于训练成本和样本准备难度。二是多任务学习，一般而言，只要多个任务的关系不是无关的或者同解的，那么多任务学习出来的模型效果要比单任务好，而且具有更强的泛化能力。通常我们会挑选跟主任务强相关且学习难度低的任务作为我们的辅助任务。

回到语义相关性这个问题上来，同一个 query 下两个文档相关性比较的 pairwise 任务更加稳定、简单而且和文档相关性任务强相关。因此我们选择 pairwise 分类和 query-doc pointwise 相关性分类的多任务模型（图 2）。在标注数据足够的情况下命名实体识别、文档分类、文本蕴含等都是相关性较相关的辅助任务。

如图 2 所示 query-doc 拼接串经过 bert 后，我们将 CLS 变量送入三层全连接网络。全连接网络的输出会流向两个 loss 节点。一个是标准的 query-doc 多分类模型，几图中的 Softmax Loss。另一个是和另外一个 doc 比较的 Pairwise Loss。Pairwise 部分标签生成方式参考了 Ranknet 的做法：

$$\bar{P}_{12} = \begin{cases} 1 & l_1 > l_2 \\ 0 & l_1 < l_2 \\ 0.5 & l_1 = l_2 \end{cases},$$

偏序对模型预测结果为:

$$\mathrm{P}_{12} = \frac{1}{1 + e^{-(s_1 - s_2)}}$$

其中 s_i 为图 2 中的相关性得分。最终整个模型的损失函数定义为三部分之和：

$$\mathcal{L}^{tot} = \mathcal{L}^1 + \mathcal{L}^2 + \mathcal{L}^{pair}$$

其中每部分损失函数军采用交叉熵损失：

$$\mathcal{L} = -\bar{P}\log(\mathrm{P(s)}) - (1 - \bar{P})\log(1 - \mathrm{P(s)})$$

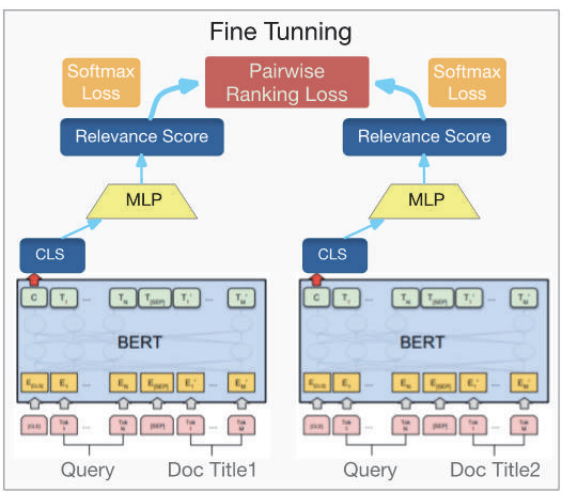


图 3. 相关性微调多任务模型

作为 bert 的尝试和落地，在第一版中我们没有直接训练我们自己的模型。经过调研，我们采用了目前各项任务上表现比较好的 24 层的 RoBERTa-zh-Large 的中文预训练模型。我们利用一期人工标注的 23 万语义相关性样本对模型进行微调。在特意挑选的比较困难的测试集合上，bert 预测结果的 AUC 达到 0.79，与目前在线模型相比，测试集上 bert 的 F1 值提升了约 0.3。

在实验的过程中，我们发现当用户的搜索词为短语或者几个关键词时，添加了 bert 相关性的模型与基准模型差不多；而当用户输入倾向于口语表达时，bert 的优势就凸显出来了。下面给出线上的两个例子（图 4），可以看出由于 bert 在交互层直接对 query 和 doc 进行 attention 交

互，能更准确的把握 query 的意图信息以及 query 中的关键词，如在第二个例子中，bert 能发现“9x”是整个词的核心。

图 4. 左边两张图对应“什么是反式脂肪酸”右边两张图对应“华为荣耀 9x 如何截屏”

五、总结

对深度学习模型来说，最大的挑战还是数据。这不仅包括数据的数量，还包括数据的质量、分布、正负样本比例等。它们都会对最终结果产生巨大的影响。然而现实中却没有统一的理论指导我们应该如何选择，只能靠平时不断试验去选择适合自己业务的模式。

除此之外，搜索是个复杂的系统，语义匹配只是其中的一个模块。语义匹配如何更好的与其他模块协作，包括如何输出特征，输出哪些特征，如何保证打分一致性等等，真正解决用户的痛点也是需要我们不断去尝试完善的。

优酷多模态搜索设计与实现

作者| 阿里文娱高级算法专家 若仁

一、多模态搜索问题定义和分析

基于标题和描述作为被检索文本视频搜索引擎会遇到如下困难：

单模态信息缺失：用户在上传 UGC 内容时，标题比较简单，很难将丰富的视频内容表达清楚，并且有些文字信息与视频内容不相关；

搜索输入多元化：用户查询词意图越来越多元化，即使是版权视频的搜索也不再集中于节目名字的搜索，社交与互动的需求逐渐增长；

TO B 侧需求增长：内容二次创作型的的用户对于视频内容语义检索的诉求显著增加，独立检索词数量近两年增长迅速；

基于多模态技术的搜索能将语言、语音、文字、图像等各种模态集成来进行搜索，综合各模态信息来理解视频，方便用户更好的找到所需的视频内容，做出很多新的搜索体验。

最近几年视频搜索和多模态技术研究在学术界非常火热，比如 Ad-Hoc 视频搜索（TRECVID AVS 任务）很多方案是基于将 query 和视频映射到中间维‘concept’空间，然后在这个中间维度做相似度排序；此外，也有 VQA/GQA 等各类基于视频问答的数据集去推动自然语言和图像的关系推理，最近 video/visual bert 的端到端的解决方案也有很大的突破。

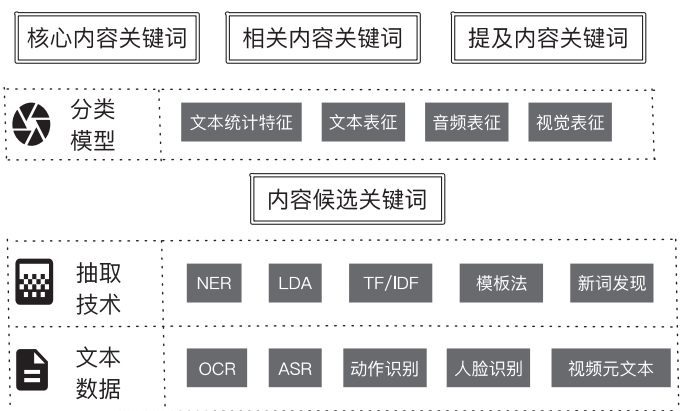
学术界比较喜欢端到端的解决方案，对短小视频的理解可能是不错的方案，但是针对长视频这些方案很难真正做到准确的理解；此外在工业界搜索引擎需要可解释性和可控性，很少采用单一的端到端的解决方案，优酷多模态搜索采用的技术方案：首先利用先进的 CV 算法将其它模态的信息降维转换到文本模态，然后通过“多模态内容检索”技术实现召回，最后综合“多模态内容相关性排序”算法技术达成多模态内容搜索的用户需求，具体技术方案如下：

二、多模态内容检索

整体来说，从一个完整的视频可以切分出不同的片段，每个片段可以拆解到镜头、关键帧、关键元素等不同粒度；对视频内容做细颗粒度拆解，将图像、视频动作、人物、声音、背景音乐等信息通过检测和识别等手段做标签化，通过上述手段完成对视频内容降维到文字模态的转换。

1. 内容关键词检索

由于视频内容涵盖社会各个领域，我们将需要分析的视频文本源不断扩大，不仅包含内容本身的文本元信息（内容的标题，IP 相关元知识），还从内容各维度解构的方法抽取文本信息，如采用 OCR/ASR 等技术将视频对话信息降维到文本域；其核心的内容关键词抽取技术框架如下图所示：

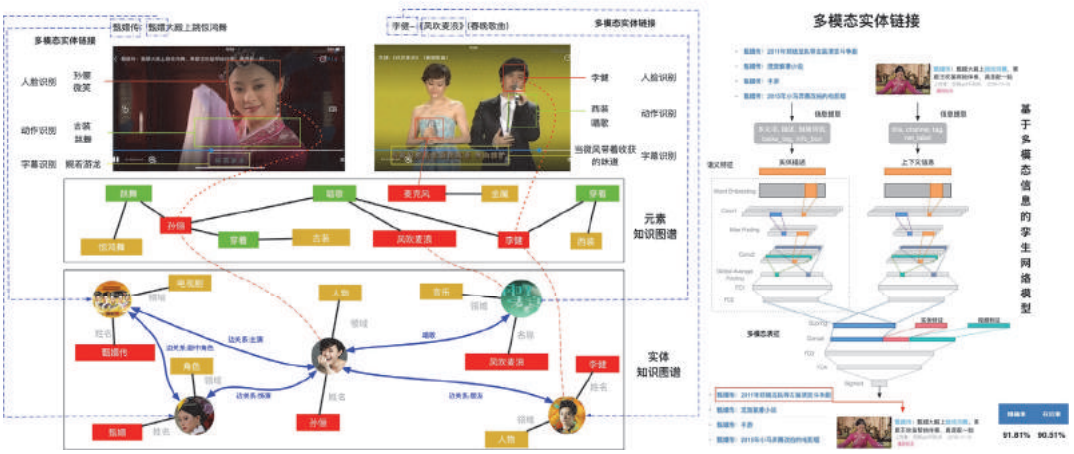


从上图可以看到，通内容关键词的词库体量非常大，且内容和关键词属于多对多关系，我们通过各种关键词抽取技术去抽取内容候选关键词，扩大候选词来源的多样性，比如基于“NER”的方法能确保抽取的内容关键词是百科类实体名称，有较广泛的知识内涵；“新词发现”方法会综合 Ngram 以及语言模型（LM）等多种基础能力扩大对未知知识领域的挖掘。候选关键词是一个不断扩充的过程，随着我们在视频内容理解的维度扩大，候选关键词的来源会越来越丰富。

在丰富的内容候选关键词基础上，根据内容候选关键词和视频内容相关程度我们构建分类模型预测 3 档，分别为核心内容关键词，相关内容关键词以及提及内容关键词，预测模型的核心特征除了文本统计特征外，还会采用文本/音频/视觉表征网络生成的多模态表征特征共同训练，提升关键词分类预测的准确率。

2. 视觉元素级多模态检索

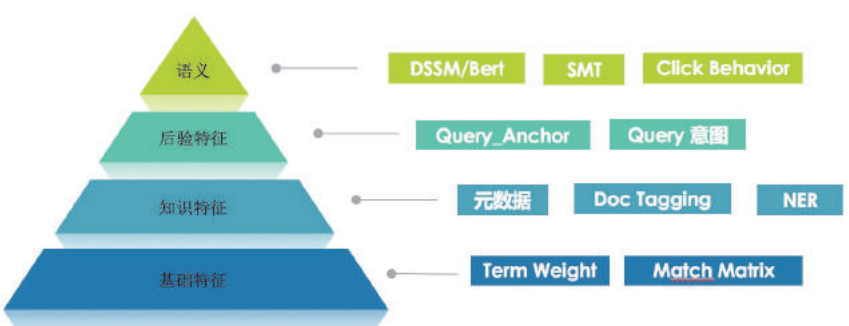
在实体知识图谱基础上，我们基于视觉技术，实现视觉元素级知识图谱的构建，如人物/动作/表情/场景等细粒度视频元素进行内容解构。在线检索阶段，基于自然语言理解技术，QP识别用户对视频元素级别的检索需求，并通过搜索引擎实现在实体知识图谱和视觉元素级知识图谱的结构化检索，如“刘德华打戏“，NLU 向搜索引擎传递”刘德华:人物“,”打戏:动作“，结合视频视觉技术的打戏标签，采用结构化检索，真正实现视频内容理解层面的检索。



三、多模态内容相关性排序

1. 内容相关性匹配

内容相关性是视频搜索的基础，是搜索体验的关键保障。如下图所示，内容相关性整体分四个维度展开：



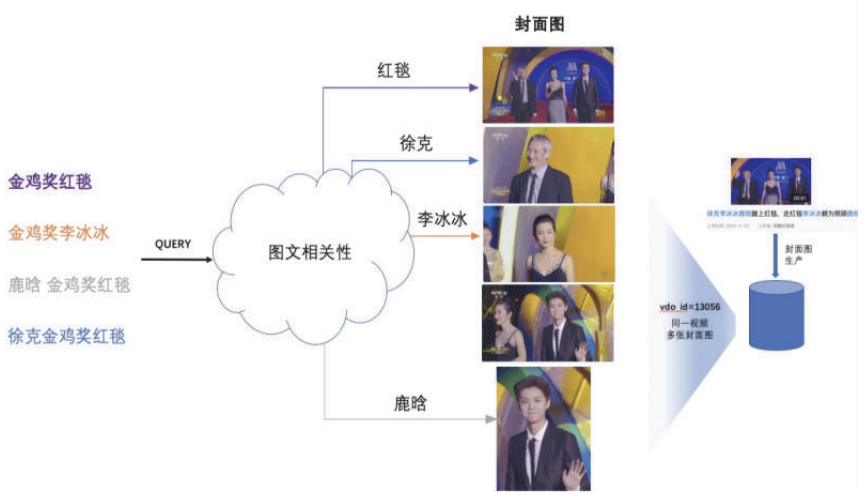
“基础文本匹配”和通用搜索的基础文本匹配有所区别，优酷内容相关性基础文本匹配会综合考虑多模态内容关键词挖掘的的内容关键词类别以及其响应权重值，视觉内容元素级解构扩充的文本元信息，以及字幕等各维度的复合匹配的特征加入到基础匹配模型中；

“知识特征匹配”层面不仅会利用视频内容理解的 IP 指纹信息去扩充视频元信息，比如 IP 名，内容包含演员人物，角色等基础信息，同时还会综合多模态特征去提升 Doc 端标题实体识别和链接的准确性，以提升知识特征匹配的效果；

2. 封面图相关性匹配

用户在搜索视频衡量结果相关性时，是通过每个搜索结果的标题文本和封面图内容进行感知的。多模态相关性除了需要考虑内容相关外，还需要考虑感知相关性，即让“**封面相关性更高、美观度更强**”的内容排序更靠前。

我们采用离线和在线结合的整体框架，离线阶段对视频从用户常见的需求维度算法生产出多张封面图，扩大对视频内容的覆盖；在线阶段，根据用户的搜索词，根据图文相关性选择最符合搜索意图的封面图展示。整体方案如下图所示:



四、方案设计

当前的封面图生产状况，每个 ugc 视频（记为 U)产出 8 张横版封面图，部分 ugc 同时产出多张竖版的封面图。其中，竖版图为搜索结果的双列展示生成。这两个集合分别记为 H 和 V。

输入：视频 U、U 的封面图集合 V 和 H，以及 query q，

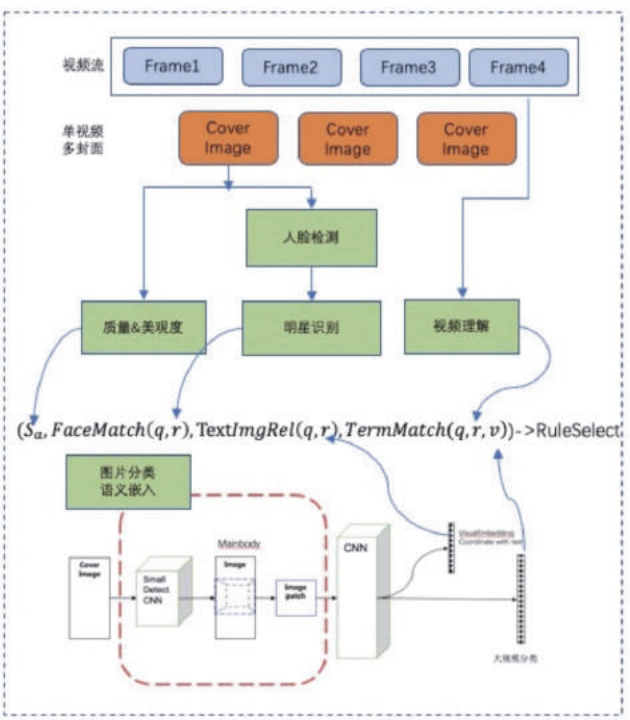
输出：封面图 C，满足相关性 R(C, q) = Max R(c, q) for c in V or c in H.

且美观度 Sa(C) > thresh；

是方案引入视觉内容相关性这个模块，在排序阶段输出视觉相关性得分，从封面美观度，人脸信息，以及在通用语义表达三个维度进行相关性度量。记 query 为 q，检索召回集合为 R，考察的视频目标为 r

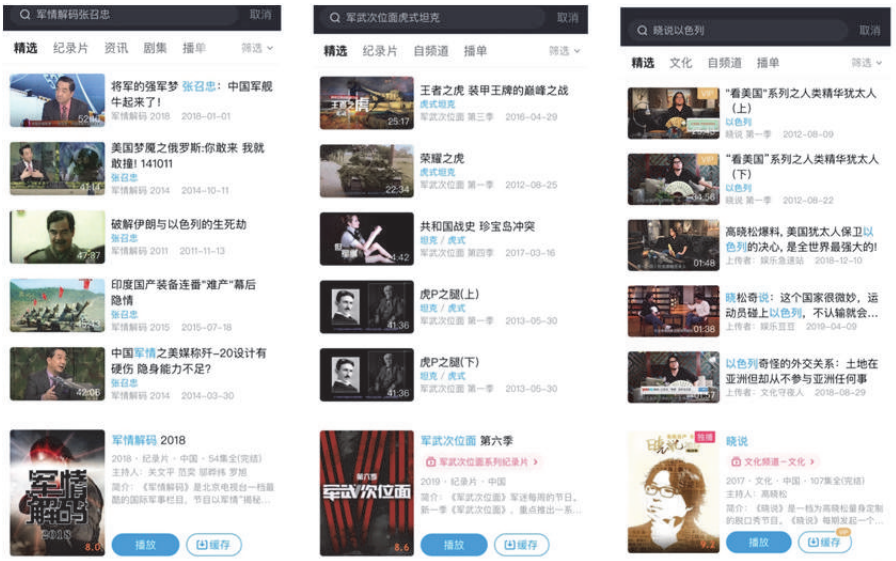
```
R(c, q)  =  SelectRule(Sa,  FaceMatch, TermMatch, TextImageRel)
```

Sa 是封面质量得分，FaceMatch 是 query 中主体人物和通过人脸识别得到的人物的匹配分数。TermMatch 是在视觉通用语义空间中 tag 与 query 中 NE 的匹配结果。TextImageRel 为图文嵌入空间中的距离度量。方案如下图所示



五、多模态搜索效果案例

优酷主 APP 线上案例效果如下图，基于内容关键词的多模态搜索大大弥补了标题等单模态文本信息的缺失的影响，同时也很好的满足了用户查询词意图多元化的需求。上线实验在这类触发 query 上跳出率有 25%的下降，点击率的能提升 130%，这也从数据维度验证了用户的搜索意图得到了较好的满足。



5

推荐

基于 Bi-LSTM 深度学习模型的 Term Weight 算法

作者| 阿里文娱算法专家 苇凌

一、摘要

Term weight (TW) 是计算 Query 与 Title 词匹配程度时，决定词重要性的模块，在优酷的搜索相关性计算中具有较为重要的作用。为了改善优酷搜索的词匹配的相关性计算，我们与优酷搜索团队合作，对 Term Weight 进行优化。我们遇到的主要问题是词重要性的标注语料不足，导致模型缺乏泛化能力。针对这一问题，我们引入无监督学习，在大量的优酷、神马、豆瓣语料上训练词向量，提高模型泛化能力，使用深度学习模型融合词语的预训练词向量、POS Tag 向量和词统计信息，并应用 Pairwise Ranking 训练词重要性的偏序关系。我们设计了一种结合局部信息、全局信息与上下文信息的 Bi-LSTM 模型，在测试集上取得了 91.3% 的准确率，相比基线 SVM 算法（准确率 84.4%）提升 6.9%。针对 LSTM 模型计算延时过长的问题，我们将模型简化为 MLP 结构，将计算延时从 10ms 降低到 2ms，而准确率仅降低到 89.8%，满足在线预测要求。基于深度学习的 Query 端 TW 模块在搜索相关性模型上线后，在 AB Test 中「单次搜索有效点击比率」提高 0.37%，「用户跳出率」降低 4.2%，效果显著，现已全流量上线。

二、背景

在搜索的 Query 与 Title 中，有些词在句子中占有中心地位，如果没有匹配会导致意思完全不同；有些词则不那么重要，即使没有匹配掉也不影响 QT 之间的相似度。自动确定句子中词语权重的模型被称为 Term Weight 模型。

1. Term Weight 在优酷搜索中的使用

Term Weight 模型被应用在相关性计算和倒排查找中。由于权重较高的词语具有更重要的作用，可以将 TW 应用在结合词语权重的词匹配算法中，使得 Query 与 Title 的匹配得分与实际的语义相关性更加一致。同时，当 Query 中词语较多时，倒排查找求交集时可能会返回空

集合，此时可以丢弃 Query 中权重较低的词再重新进行倒排查找，这样可以在保证召回数量的同时提高召回结果的相关性。

2. Term Weight 的标注数据

Term Weight 的标注数据采用偏序关系来表达词语之间的重要性关系，中心词的权重大于普通词，单纯修饰性词的权重则更小一些。标注数据由人工标注，共 2700 条。

样例

句子 经典歌曲《求佛》,太好听了
标注 求佛>歌曲>好听>经典>太>了>《=》=,

3. 其他工作：使用搜索点击来构造样例

构造 Term Weight 标注样例的另一种方法是通过搜索的点击来确定哪个词是重要的，如果一个 Query 中出现了词 A 与词 B，但是 Title X 中只出现了词 A，而 Title Y 中只出现了词 B，如果用户点击 Title X 的几率显著高于 Title Y，那么可以认为词 A 比词 B 更重要，这种方式广泛使用在 Term Weight 的优化中。

然而，在优酷搜索中，绝大多数 Query 是剧名与其他短 Query，使得绝大多数搜索结果都能匹配到 Query 中的所有词。而不满足这一条件的搜索词出现机会又很少，进一步使得我们很难通过搜索点击来构造出高质量的样例。

基于以上原因，我们最终采用人工标注者对句子进行偏序关系标注的方式来生成 Term Weight 的训练/测试集。

三、基线 Term Weight 模型

基线 Term Weight 模型是一个线性 Rank SVM 分类器[1]，它的输入特征有： 词性特征 / query idf 特征 / title idf 特征 / 作为 query 出现次数 / jieba idf 特征。

这个模型中，词性特征贡献最大，几乎主导了分类的结果，统计特征对分类的贡献很小，这导致我们难以区分两个相同词性的词之间的重要性。

基线模型有两个不合理之处，一是对于新词缺乏泛化能力，二是句子上下文不影响词语的重要性。针对这两个问题，我提出使用大量语料来预训练词向量，并使用深度学习模型来融合统计特征、词性特征和词向量特征。

四、深度学习 Term Weight 模型

相比传统线性分类器，深度学习模型的建模方式更加灵活，可以方便地结合预训练的词向量、在模型中引入非线性，以及在单模型中引入从浅层到深层模型的混合。这些优势是我们将模型优化的重点放在深度学习模型的主要原因。我们使用 Tensorflow 来实现文中的深度学习模型。

1. 多特征词表示

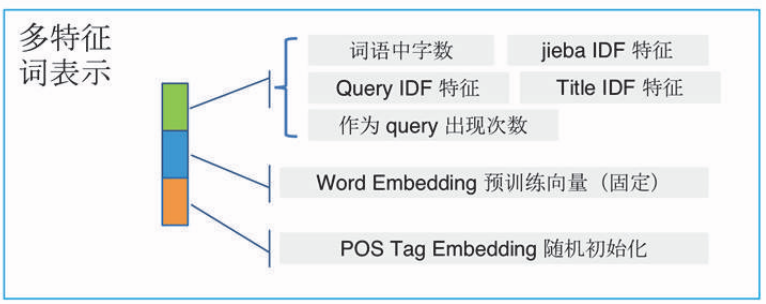
为了充分利用分布式词表示、统计特征和语言学特征，我们设计了结合三种特征的「多特征词表示」方法。

预训练词表示 使用优酷、神马、豆瓣等十亿级语料训练 Skip-gram 词向量作为 Embedding 层的初值，由于 TW 的数据集较小，我们在模型中保持这些词表示固定不参与训练。

统计特征 与 SVM 模型相同的统计特征，即 词性特征、query idf 特征、title idf 特征、作为 query 出现次数和 jieba idf 特征，为保证值域在一个量级内，统计特征被归一化到 [0,1] 区间内。

语言学特征 采用 AliNLP 的 POS Tagger 的词性分析结果作为语言学特征。由于 POS Tag 是枚举值，我们用 4 维的 Embedding 来表示它，该 Embedding 矩阵参与模型训练。

对于输入句子中的每个词，这三种特征被拼接在一起，作为词的表示，如下图所示。



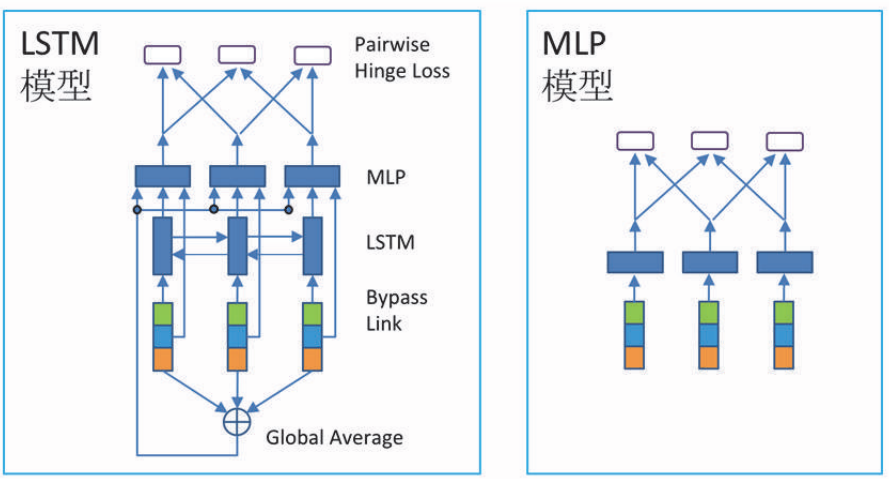
2. 深度学习模型

我们提出了两种模型：上下文有关的双向 LSTM 模型，可以建模上下文对词权重的影响，具有较强的建模能力；上下文无关的 MLP 模型，无法建模上下文对词权重的影响，但是运算开销小，计算延时低。

双向 LSTM 模型是一种序列模型，可以建模上下文的依赖关系。这里的 LSTM 模型有两个主要的改动，一是在 LSTM 的输出层后又添加了一个隐层，这主要是为了融合其他来源的信息，比如全局平均词特征；二是将每个位置的词特征与 LSTM 输出层拼接起来，形成一个 Bypass 连接。这使得该 LSTM 模型实际上成为序列模型和浅层模型的一个融合。

MLP 模型是一种多层前馈神经网络模型，开发 MLP 模型主要是为了优化模型效率。在网络结构调优中，有两个参数非常重要，一是词向量的维数，词向量的维数越高，它能承载的参数量就更大，而神经网络的性能在一定程度上是与参数数量成正相关的；二是隐层的维数，隐层的维数越高，参数量就越大，辅以适当的 dropout 规范化，宽隐层会起到 ensemble 的效果。

模型输出一个与句子等长的权重序列，使用 Sigmoid 作为输出层激励来使得权重在 0 到 1 之间，并根据权重序列和标注样本构造 Pairwise Hinge Loss 的损失函数，句子中每两个权重不同的词之间都可以构造出一个样例。



五、实验结果

1. 算法迭代

我们在偏序关系测试集、QT 匹配的相关性测试集上对模型进行了测试。原基线模型的准确率为 84.4%。我们发现，使用原统计特征在特征交叉后达到了词序准确率的上限，对模型进行进一步的优化还会导致过拟合。单添加词和 POS Tag Embedding 特征的 MLP 模型可以将模型的效果至 88.4%。而使用带有上下文信息的 Bi-LSTM 模型则能够将模型的效果进一步提升

至 90.8%。

2. 模型简化

由于 Embedding 矩阵是模型参数的主要部分，为了节约存储、网络传输和加载时间开销，我们对词表进行了压缩。我们发现，即使将词表缩减到 50 万，Bi-LSTM 模型也没有出现性能下降，而 MLP 模型只出现了轻微的性能下降。最终我们采用 50 万词表的模型进行在线预测。

3. 上线效果

有效减低了「用户跳出率」并提高了「搜索命中率」

4. MLP 模型与 Bi-LSTM 模型的比较

由于 MLP 模型是上下文无关模型，一个词的权重只与词语本身有关。多数情况下，训练集中的人名被标注为中心词，因此测试时 MLP 模型倾向于将人名预测为最高权重；Bi-LSTM 模型是上下文相关模型，因而不存在这样的问题。另外，Bi-LSTM 模型也比 MLP 模型更倾向于将书名号内的一个和多个词预测为更高的权重，这也说明 Bi-LSTM 模型能够更好地利用上下文信息。

样例 1：人名的权重

```
MLP 三生三世十里桃花:0.9999 杨幂:0.9992
Bi-LSTM 三生三世十里桃花:0.9350 杨幂:0.3244
```

样例 2：书名号的影响

```
MLP 《:0.000 古代:0.914 岩画:0.841 》:0.000 _:0.000 王金成:1.000 新课标:0.834 小学:0.983 美术:0.992 优质课:0.868 展示:0.611
Bi-LSTM《:0.000 古代:0.917 岩画:0.889 》:0.000 _:0.000 王金成:1.000 新课标:0.597 小学:0.892 美术:0.732 优质课:0.510 展示:0.434
```

六、现有问题与未来方向

本任务的排序优化目标与实际使用中搜索的优化目标仍有不匹配问题：排序优化只是按照词的重要性排序，但是搜索中则需要对 Title 按照相关性排序。一种可能的解决方案是使用点击 / 未点击行为构造样本对直接对构造的相关性特征进行训练，我们将在后续工作中对这种方式进行探索。

多模态视频多标签分类在优酷推荐算法中的实践

作者| 阿里文娱算法专家 苇凌

一、背景

在个性化视频推荐中，为了对视频的内容进行表征，我们需要准确理解视频的主题元素。一种常用的方法是给视频打上多个标签，每个标签代表了一个视频中的主要元素。优酷过去的标签算法主要依赖于文本分析，当视频的文本元信息（标题、描述、评论等）对主题的描述不明确时，我们常常无法分析视频内容。为了解决这一问题，我们采用文本、封面图、音频、视频多种模态信息对视频进行多标签分类，大大提高了建模的准确率。

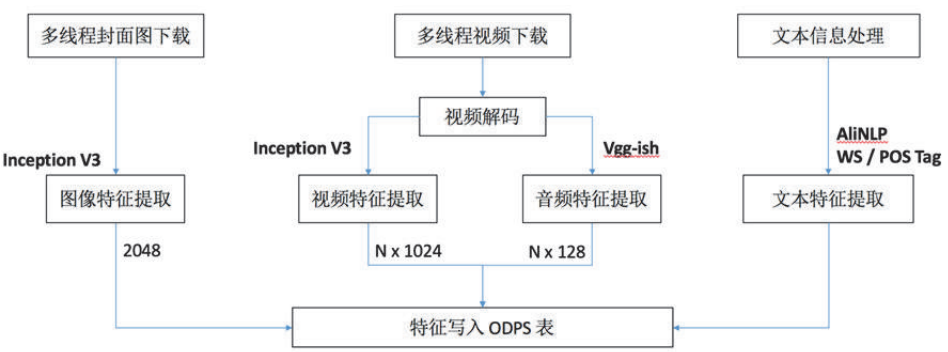
多模态视频多标签分类结果示例

二、数据采集

我们建设了一个拥有百万标注视频，1000+ 个标签的数据集，涵盖音乐、舞蹈、游戏、亲子、电影、电视剧等多个领域，其中平均每个视频有 2.4 个标签。我们分别对视频特征和音频特征进行了采样抽取，并抽取了封面图的图像特征。

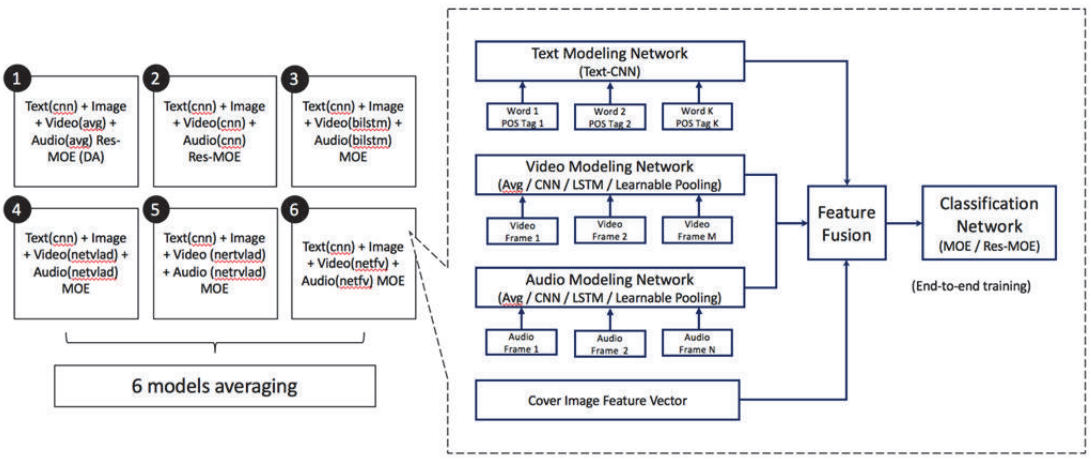
三、特征抽取

为了减少存储和处理的开销，我们对视频画面进行抽帧处理，每 2s 视频抽取一帧画面并提取 Inception V3 特征。对于音频，我们对每 1s 视频抽取了 Vggish 网络特征。我们对文本与封面图也做了相应的处理，见下图。



四、系统设计

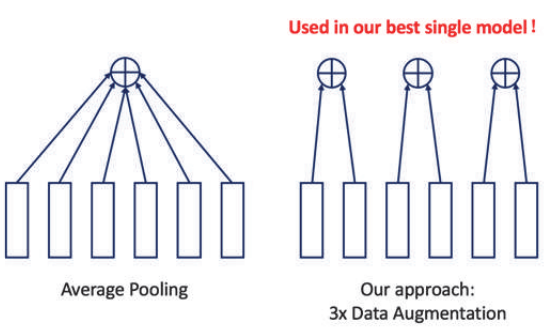
我们采用 Late-fusion 的方式对多个模态进行融合，所谓 Late-fusion 就是将每个模态单独建模映射为向量，将代表不同模态的向量再进行融合的方法。通过使用不同的音视频建模方法，我们构建了六种不同的网络，最终我们将多个网络的预测结果进行融合。



1. 特征聚合网络

我们采用了 Average Pooling / CNN / Bi-LSTM / Learnable Pooling 四种算法进行音视频的特征聚合。

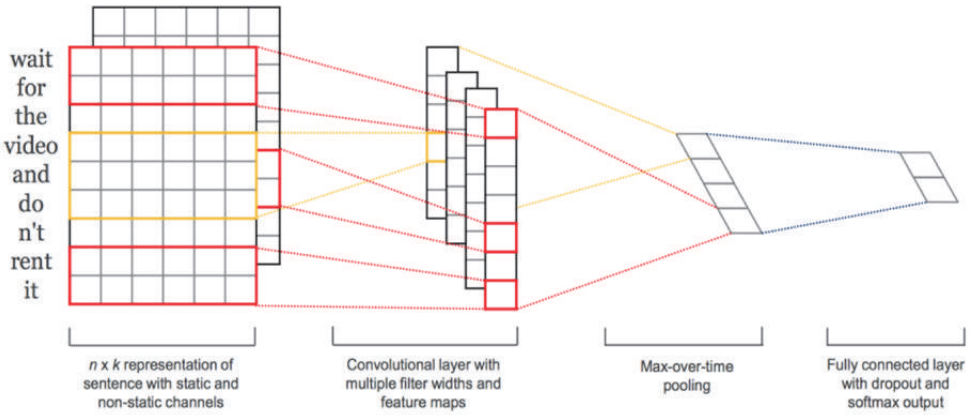
Average Pooling 将各帧特征直接平均起来，是最简单的特征聚合方法。因为计算量少，也容易进行数据扩增，我们采用将音频/视频的特征平均分为三份的方法进行数据扩增，扩增后模型效果均有一定程度的提高。



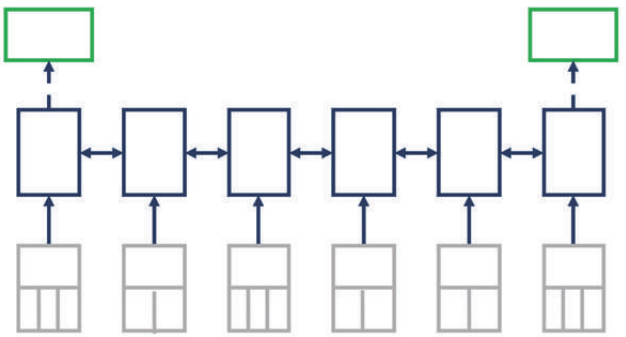
GAP@5（调试集）

Models	Original	Data Augmentation
Multi-modal MOE	60.00%	60.96% (+0.96%)
Multi-modal Res-MOE	58.53%	61.49% (+2.96%)

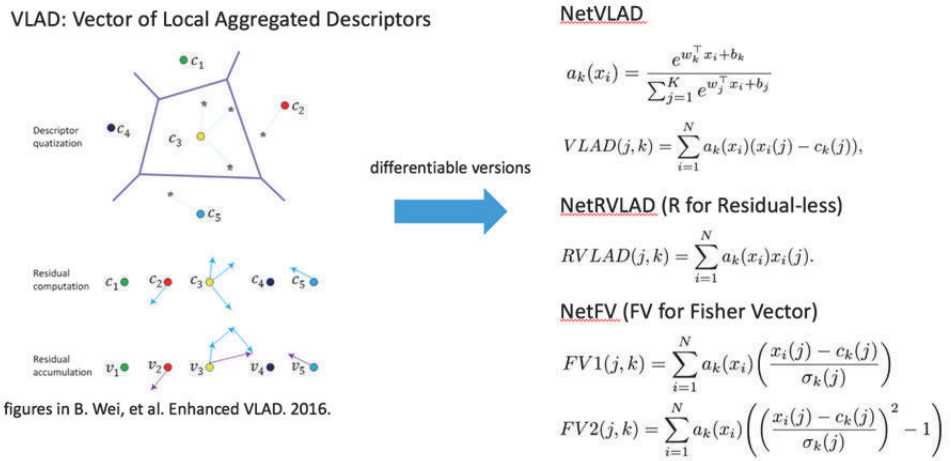
CNN 网络可捕捉短时内的局部 pattern，由于我们已经将每一帧提取为特征，这里我们采用的是 1D-CNN，结构与 Text-CNN 相同。



Bi-LSTM 是一种捕捉双向序列依赖关系的 RNN 网络，我们取得第一帧和最后一帧的 memory 信息作为视频的表示。



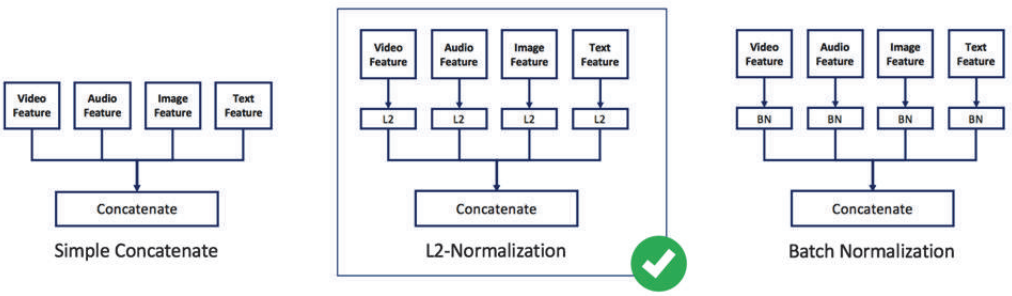
Learnable pooling 是一类基于聚类的 Pooling 方法，包括 NetVLAD / NetRVLAD / NetFV。NetVLAD 是一种 VLAD 的可微版本，而 NetRVLAD 和 NetFV 则是它的变形。VLAD 将空间分成若干个区域，并在每个区域内分别计算落在该区域内的向量的残差。它是一种不考虑帧间顺序的 Pooling 方法，相比于 Average Pooling 而言，它的优势在于可以精细地刻画向量集合在空间中的局部分布。



figures in B. Wei, et al. Enhanced VLAD. 2016.

2. 多模态融合

在多模态特征进行融合时，简单地进行向量拼接是最常用的方式，我们发现，预先对向量进行适当的正则化可能会取得更好的效果。例如，在 Average Pooling 中，预先 L2 正则会明显好于简单拼接，也好于进行 Batch Normalization 的结果。这可能是因为预先正则化对各个模态的数据分布进行了一定预处理，从而一定程度上防止了过拟合的发生。



Modality	GAP@20
Image	35.54%
Audio	37.13%
Video	51.24%
Text	61.89%

GAP@20 Multi-modal Fusion

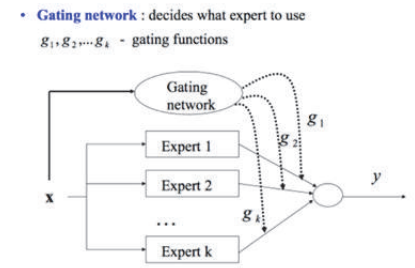
Classification Model	Simple Concatenate	L2-Normalization	Batch Normalization
MOE	64.20%	**66.17%**	65.90%
Residual MOE	60.03%	**67.12%**	65.67%

3. 分类模型

Mixture-of-Experts 是多标签分类中常用的末端分类器单元，我们在使用了 MOE 的同时，

还测试了其改进版 Residual MOE 的效果。我们发现，Residual MOE 通过加深网络层数，层层拟合残差的方式可以提高分类的效果。同时，由于它是一个浅层与深层网络的结合，不容易产生严重的过拟合问题。

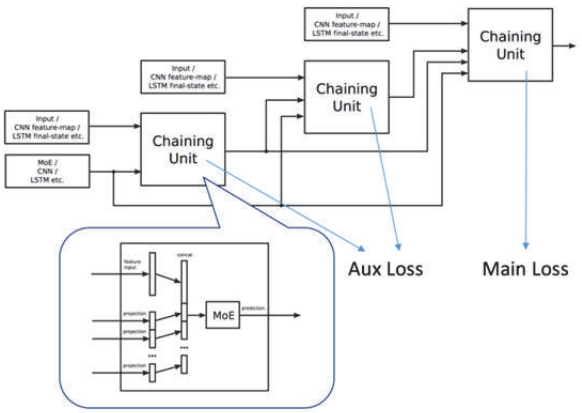
Mixture of experts model (aka. MOE)



GAP@20

Models	MOE	Res-MOE
Multi-modal (avg v/a)	66.17%	67.12% (+0.95%)
Multi-modal (cnn v/a)	64.50%	66.72% (+2.22%)

Residual MOE model



五、结果分析

1. 评测结果

我们使用 41.5 万数据用于训练，5 万数据用于验证，10 万数据用于测试。在测试集上得到如下的单模型与融合模型的对比结果，其中，Text(cnn) 是文本单模型单独的分类效果，加入了音视频模态信息后，最好的单模型相比其好 5% 左右，而融合模型则比其好 9% 左右。

Models	GAP@20
Text(cnn)	61.89%
Text(cnn) + Image + Video(avg) + Audio(avg) Res-MOE (DA)	67.12%
Text(cnn) + Image + Video(cnn) + Audio(cnn) Res-MOE	66.72%
Text(cnn) + Image + Video(bilstm) + Audio(bilstm) MOE	65.54%
Text(cnn) + Image + Video(netvlad) + Audio(netvlad) MOE	65.75%
Text(cnn) + Image + Video (nertvlad) + Audio (netrvlad) MOE	64.87%
Text(cnn) + Image + Video(netfv) + Audio(netfv) MOE	66.18%
6 models ensemble (simple averaging)	70.85%

2. 定性比较

从定量指标上，我们肯定了多模态融合的效果。但是多模态融合到底解决了什么样的问题呢？我们用几个 show case 来说明不同模态的融合的意义。

下面这个案例说明了一个典型的文本模态失效问题，即当分词出现问题时，文本模态彻底失效，而音视频模态还可以做到准确的推断。当「莫文蔚演唱会」被分为同一个词时，文本模型就无法看到具有较好可分性的词语了，而音视频模态依然可以判断这是一首歌曲。

- Bad word segementation
  - 莫文蔚演唱会 上的这一段视频，网友表示值回票价了
- 莫文蔚演唱会 在训练集中仅出现一次
- From text model's perspective:
  - XXX 上的这一段视频，网友表示值回票价了
  - nearly no informative words

Top6 scores:

VideoAudio	音乐 0.8634	歌曲 0.2179	华语男歌手 0.0976	翻唱 0.0695	经典歌曲 0.0614	越南歌曲 0.0608
TextOnly	资讯 0.4371	纪录片 0.3347	娱乐 0.0840	生活 0.0413	历史 0.0352	教育 0.0240
SingleModel	音乐 0.4489	综艺 0.2483	我是歌手 0.1991	演唱会 0.1494	娱乐 0.1079	华语女歌手 0.0920
Ensemble	音乐 0.6306	综艺 0.2530	华语女歌手 0.0760	翻唱 0.0593	歌曲 0.0515	娱乐 0.0438

即使分词没有问题，文本模态依然面临的一个问题是文本本身表意的能力有限，下面这个 case 说明了这个问题。对于文本模型来说，这里的大部分信息来自「奥特曼」与相关词语，因此「动漫」获得了较高的关联权重，而从音视频的角度来看可以排除掉动漫、游戏的可能，最终多模态的结果也可以较好地提炼「奥特曼」「玩具」「亲子」的主题。

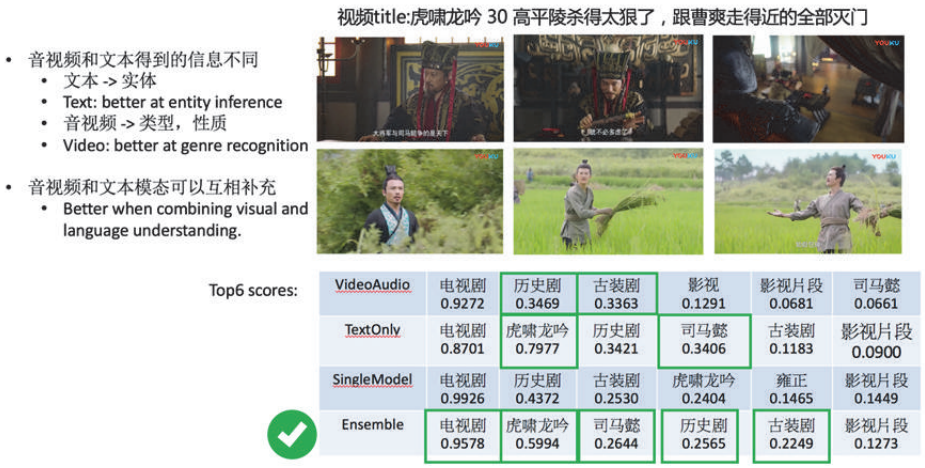
视频title：
【菠萝上传】初代奥特曼科学特搜队光线枪 奥特曼&巴尔坦星人套装

Top6 scores:

VideoAudio	玩具 0.7572	奥特曼 0.5311	生活 0.3290	动漫 0.1823	模型 0.1795	体育 0.1747
TextOnly	奥特曼 0.4832	动漫 0.4205	游戏 0.2294	迪迦奥特曼 0.1957	亲子 0.1008	玩具 0.0663
SingleModel	玩具 0.7393	奥特曼 0.4895	模型 0.3495	动漫 0.2590	亲子 0.1854	生活 0.1730
Ensemble	玩具 0.7262	奥特曼 0.6186	亲子 0.2783	模型 0.2304	生活 0.1972	体育 0.1796

Text model can only **guess** that Altman is about animation or game.
But the video is about toys.

当然，音视频模态面临的一个严重问题是对于知识的提取能力有限，文本模态对于实体有更好的提取与推断能力。在下面的例子中，音视频模态会以更高的权重推断「古装剧」和「历史剧」，而文本模态则会推断「虎啸龙吟」与「司马懿」，最终的融合模型则可以融合两者的优势获得更完整的推断结果。



六、未来工作

我们希望从以下几个方向开展工作：

利用与 Collaborative Deep Metric Learning 相似的训练框架实现多模态视频相关性算法，在 CDML 的基础上还可以加入文本，使得视频向量的表达能力更佳；

目前我们对于特征的处理停留在抽帧的层面，由于采样率较低，两帧之间一般不带有强依赖关系，也失去了动作等短时信息。我们还计划通过 C3D 等算法对短时视频信息进行表示，以获得更细粒度的视频特征。

多模态是个 Multi-view 问题：Multi-view 要求各个模态一一对应的样本，当样本少的时候，如何利用 Single-view 的样本帮助 Multi-view 训练，如果采用 Multi-view 的思路，是否可以将各个 view 投影到公共空间来进行分类，充分利用各个 view 自身的样本。

Label Correlation 的利用：利用 Label 本身的知识以及之间的各种关系，如何让少样本的类预测得更好。

Collective Learning：视频很重要的信息是用户的行为信息，包括用户搜索点击和推荐点击

行为，从这个角度看视频是图上的节点，图表示对分类是有正向作用的。

Active Learning：特征分布跟训练集差别较大的样本，可以提人工标注，增加关键的标注样本，这个方向与 outlier detection 又有着一定的联系。

参考文献

[1] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

[2] Li, Fu, et al. "Temporal modeling approaches for large-scale youtube-8m video understanding." arXiv preprint arXiv:1707.04555 (2017).

[3] Miech, Antoine, Ivan Laptev, and Josef Sivic. "Learnable pooling with Context Gating for video classification." arXiv preprint arXiv:1706.06905 (2017).

[4] Arandjelovic, Relja, et al. "NetVLAD: CNN architecture for weakly supervised place recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[5] Jégou, Hervé, et al. "Aggregating local descriptors into a compact image representation." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[6] Wang, He-Da, Teng Zhang, and Ji Wu. "The monkeytyping solution to the youtube-8m video understanding challenge." arXiv preprint arXiv:1706.05150 (2017).

[7] Lee, Joonseok, et al. "Collaborative Deep Metric Learning for Video Understanding." (2018).

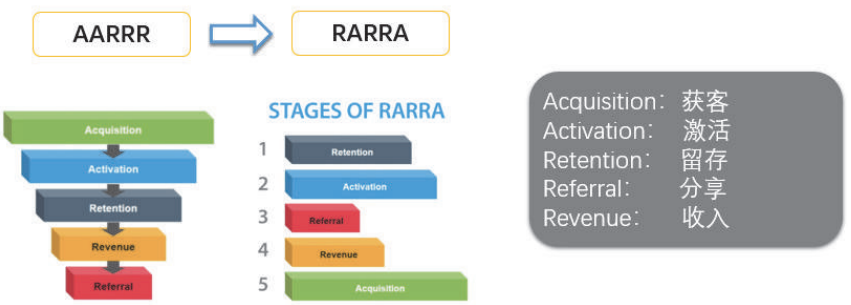
6

增长与营销算法

本节摘要

作者| 阿里文娱高级算法专家 天师

如何实现产品的用户增长？显然，这是各家移动互联网应用的头等大事，也是悬在各家业务负责人头上的“天问”。在移动互联网进入下半场的大趋势下，过去粗放式的买量、厂商合作等模式越来越会受到掣肘，将更加依赖精细化的用户增长策略和产品用户体验的细致打磨；经典的 AARRR 模式会逐步转向 RARRA 模式，提升产品留存、拉活、分享传播等方式是构建增长的主要战场。而在此之中，对于一个内容型产品，个性化算法对于用户留存、拉活将起到决定性的作用。



考察与优酷类似的应用，在内容领域，增长的成功案例有：

1）“头条快手”模式：内容分发类产品，代表是“今日头条”、“抖音”、“快手”等。这类产品构建了完善的内容生产和消费生态，旨在通过推荐系统同时刺激生产和消费，实现两端的同时增长；

2）“趣头条”模式：该产品同属内容分发类产品，但较早地参考了网络游戏模式，从各个环节设计用户里程碑和激励，不断引导新用户一步步完成点击、下刷、完整阅读、分享、关注等目标里程碑，并给予虚拟货币和真实货币的激励，在短时间内获取了大量下沉用户；

3）“爱奇艺”、“腾讯视频”模式：这类产品利用大量资金和精准的内容采买眼光，利用头

部内容的流量聚集效应，在前几年迅速圈定大批用户，并形成长视频 app 特有用户心智。由于内容头部化，个性化算法在其中发挥的空间和作用较小，产品、模式趋于同质化，内容采买的巨大资金投入使得长视频网站的盈利遥遥无期。

会员增长是长视频产品体系下用户增长的特有子问题。优酷作为国内顶尖的视频内容提供商，上述三种增长模式都是需要进行借鉴的。用户增长问题需要从内容供给、内容分发、权益设计、产品设计等多环节进行联合优化，从算法的角度，其目标可以拆解为两大部分：

1）用户状态建模：深度建模用户状态和行为，从大数据集中找到使用户从低阶状态到高阶状态转化的干预因子；

2）个性化分发的升级：将用户行为建模后，在多个场景将这些干预动作落地为个性化推荐算法和营销算法，满足用户的视频内容消费需求。

阿里大文娱是阿里集团双 H 战略（Happiness & Health）中最为重要的践行者，在不断为广大网民提供优质内容与良好体验的同时，我们也面临着用户规模化增长以及营收有计划提升的压力。我们已经逐步形成以消息推送（push）、站外引导（dsp）以及新用户承接推荐等场景组成的用 户增长业务体系，也已经逐步形成了以权益发放（营销）以及商业化（广告）等抓手组成的收入增长业务体系。在这一章介绍的基于因果推断的推荐算法、基于双 pid 的动态报价算法以及基于 uplift model 的营销增益模型正是应用在这两大业务体系中的，我们已经在多个业务场景中取得了较为显著的效果提升，我们相信其中的一些技术必将对整个互联网业内在增长算法体系带来一些崭新的视角、思考和实践经验。

因果推断在用户增长中的应用

作者| 阿里文娱高级算法专家 天师

一、用户增长和智能营销算法的目标

在本章节的序中已经介绍了优酷用户增长的业务打法和构思，其中已经提到，个性化的分发算法是实现用户增长的主战场。其中有两大目标：

1）用户状态建模：深度建模用户状态和行为，从大数据集中找到使用户从低阶状态到高阶状态转化的干预因子；

2）个性化算法的升级：将用户行为建模后，在多个场景将这些干预动作落地为个性化推荐算法和营销算法，满足和刺激用户的视频内容消费需求。

针对目标 1，传统数据分析主要是建模变量之间的相关性而非因果关系，不能从真正的因果关系来设计干预手段。

针对目标 2，传统的推荐算法主要进行短期的点击、时长等多目标预估，未能从用户状态的跃迁去设计个性化的目标机制；其次目前大量应用的深度学习类算法同样属于统计学习派别的延展，其模型可解释性差，不能从中推断用户兴趣与内容的因果关系，而该类技术方向的演化会导致用户画像的算法较为单薄，不能满足优酷会员营销核心业务的需求。

基于因果推断的推荐算法我们已经成功应用在消息推送(push）以及 dsp 外投买量算法等业务中，而在营销场景中应用的 uplift 模型本质上也是因果推断思想的一个典型应用。因此，我们在整个用户增长以及智能营销的业务场景中逐步推广地应用了因果推断的思想，在某些实验中取得了非常好的业务结果，比如我们在 push 和 dsp 业务中的沉默用户召回这个场景下就取得了点击量和点击率的显著提升。

二、用户状态表示

1. 用户画像与状态表示法

传统的用户画像表示技术要么服务于运营可解释性，要么服务于推荐或广告系统的模型预估，通常建模成向量（离散高维或低维稠密）。而我们在深入研究在线视频和付费会员业务后，发现状态转移图是更有力地建模该业务下用户画像的数据结构，原因如下：

1)用户从非会员到购买会员并逐步进入高阶会员的阶段，本质属于一种强规则定义的状态；

2）在线视频，尤其是长视频领域具备长时间、连续型消费（追剧、追网红）等特点，对比传统的图文推荐系统、电商推荐系统和广告系统，用户的消费行为可以在连续的时间上进行切分，状态表示法是对向量表示法的有力补充；

3）新用户的承接和推荐策略是用户增长中“促留存”，建立心智的重要阶段。借鉴网络游戏和趣头条的思路，将难度较大的“促留存”问题拆分为“目标达成”问题，产品通过策略不断使得用户完成高阶里程碑，是业内目前已证明成功的用户增长方法。

序中已经提到，会员模式是长视频业务的核心付费模式，在用户的整个生命周期内，其大体的会员状态转移图如下：

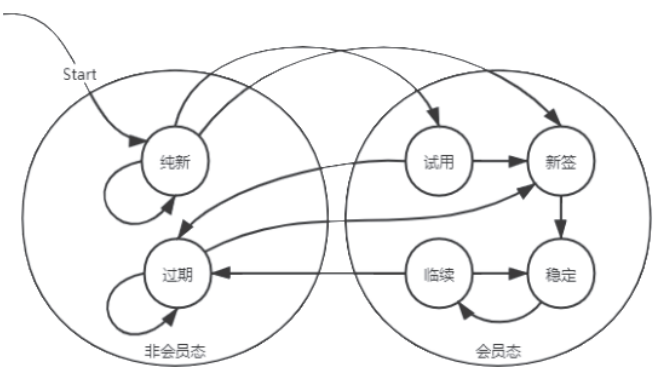


Figure1：会员转化状态

新用户阶段是产品对用户建立信任感的最重要时期，新用户在优酷 app 中的里程碑可以大致描述如下：

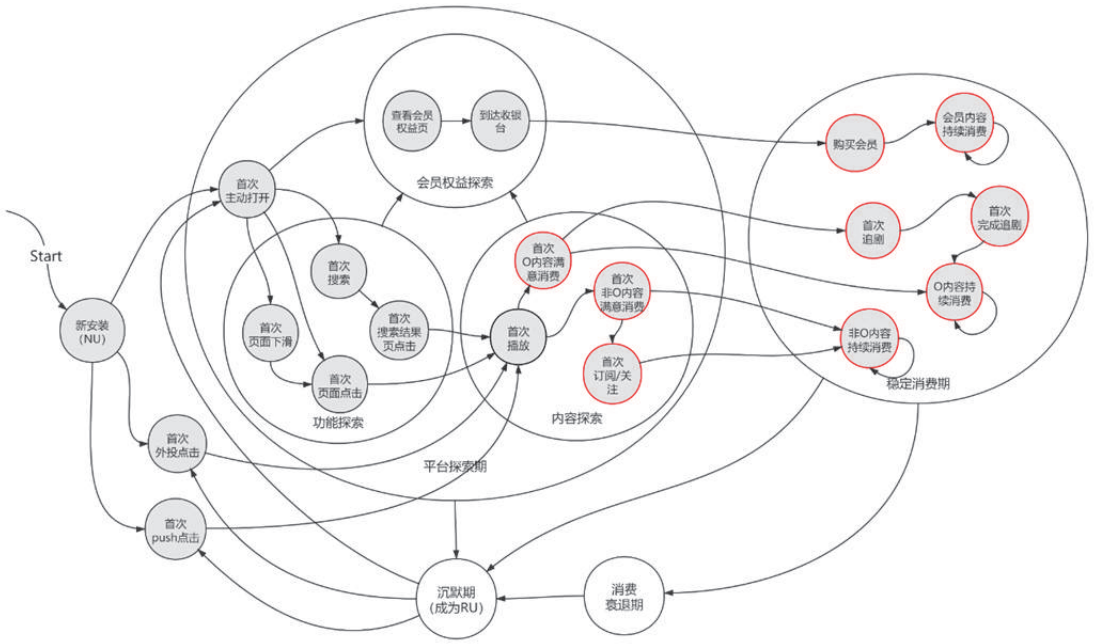


Figure2：新用户状态里程碑

可以看到，用户在不同的状态下，我们期望他们能完成状态的“跃迁”，也就是从低阶状态不断往高阶状态：“持续消费”，“订阅/关注”，“追剧”，“会员稳定期”等转化。

可以预见的是，这种用户画像的表示方法，将会对业内长久以来已经趋于成熟的个性化推荐算法步向新的发展阶段：即为用户增长这个核心业务问题更好的服务。首先是多目标的排序机制，对于在不同状态下的用户，个性化算法的机制目标会不同（跃迁至目标态）；其次启发我们从更前沿的算法高度来研究状态跃迁的干预手段问题，进而解决推荐系统中长期难解的“可解释性”、“幸存者偏差”、“兴趣探索”等问题。

针对干预手段的研究，在 2019 年用户增长&智能营销团队组建之后，对因果推断（Causal Inference）算法率先进行了研究和落地，目前在个性化推送、外投 DSP 应用了基于 matching 的无偏 user-cf 算法，智能红包发放场景应用了 uplift model，取得了显著的核心业务指标提升，并得到了业务方和兄弟团队的一致认可。现将无偏 user-cf 算法介绍如下，uplift model 将在本章第二部分进行阐述。

三、基于因果推断的无偏 user-cf 设计

1. 因果推断（Causal Inference）简介

因果推断（Causal Inference）作为新兴的人工智能技术方向，旨在突破传统数据分析和机器学习方法的瓶颈，建模大规模数据集中的因果关系，为干预手段的设计提供指导，为构建下一代面向用户增长的全域分发系统提供理论基石。

因果推断的核心研究课题：

1）从众多观测到/未观测到的变量中找出致因（causes）；

2）预估某个行为/因素的影响力/效益（causal effect）。

对于个体，来说，核心是寻找反事实(counterfactual)镜像。在个性化推荐中，一个难题就是消除“幸存者”偏差，即如何将低活用户通过良好的路径推荐，逐步变成产品的高活用户。我们定义问题如下：

2. 目标

消去推荐系统的偏差。用户增长需要消去高活用户带来的行为偏置，提升低活用户推荐效果。

3. 假设

用户变成低活、沉默的原因主要是因为对之前推荐不满意（负例）。

4. 方法

1）构建 Counterfactual 镜像人：

利用无偏信息构造相似度量，构造低活 user 到高活 user 的 matching：

a）基础人口属性、安装的长尾 app 信息等；

b）主动搜索行为（非被动推荐），尤其是长尾 query。

2）去除低活用户的 leave causes，推荐相似高活用户的 stay causes。对于推荐系统来说，这些 causes 包括：

a）item 本身：但缺少泛化容易推出老内容；

b）item 的泛化特征：标签、时效性、质量。

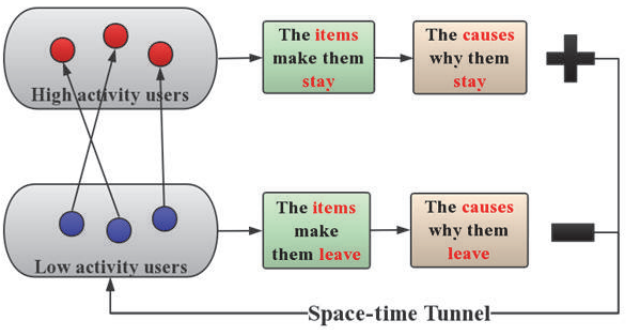


Figure 3：无偏 usercf 设计

注意，由于使用了 matching 方法，这里的算法非常类似传统的 user-cf 类算法，但是和传统 user-cf 核心的区别在于：

1）matching 不使被动推荐数据：个性化推送、站内推荐、运营推荐的内容都不使用；

2）只匹配低活到高活，活跃度相同的用户之间不进行匹配。

5. 业务收益

该算法落地后，在两个 baseline 相对较高的算法场景中取得了较大的收益：其中个性化推送（push），在沉默用户中获得了 50%+ ctr 和 50%+ click 的双增长，在外投 dsp 业务中，拉活量对比峰值接近翻倍。

四、总结&展望

目前算法的应用，只是对应了两个用户状态（低活→高活）之间的推断，如 Figure1 和 Figure2 所描述的，用户增长的目标是将细分的低阶状态往高阶目标态上进行跃迁，那么该类算法很显然将会在数据分析、产品设计、分发优化等各个环节发挥巨大作用。整个 2019 年团队的实践虽然取得了很大的业务效果，但只是对该算法方向相对较浅显的应用，且对于优酷整体的增长问题来说，应用的场景还不够多，未来期望在其基础理论和实践都投入更多的资源。

可以预见的是：对于整个业内用户增长的方法论，该方法在未来必将成为核心的理论基石。对于个性化推荐这一经典领域，该方法为解决经典难题“幸存者偏差”，“可解释性”，“用户表示”，“兴趣探索”等提供了漂亮的解法。

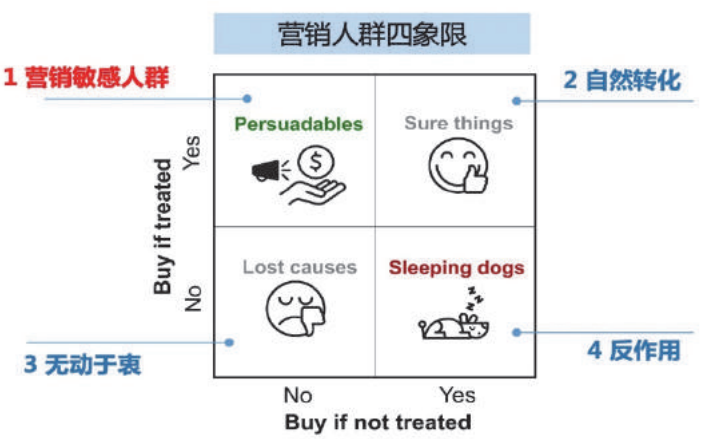
基于 Uplift Model 的营销增益模型

作者| 阿里文娱高级算法专家 毕达

一、智能营销的核心任务与主要挑战

在移动互联网普及的今天，基于大数据的营销已经成为用户运营中的常规动作。然而如何找出“营销敏感“的人群，而不把预算浪费在“自然转化”的那部分人身上，成为智能营销算法最重要的挑战。

从营销人群的四象限可以看到，不做任何营销干预（treatment）就会“自然转化”的人群和对于权益刺激“无动于衷”甚至起到“反作用”的人群均不是营销动作的最佳受众，而只有营销刺激能够促使用户从“不买”到“买”转化的这部分“营销敏感人群”才是各类营销动作最该触达的人群。



针对此问题，我们构建了基于营销激励的增量效果的增益模型，也称为 Uplift Model，并且应用在淘票票智能红包等阿里文娱的营销实践中。基于此模型，实现了淘票票个性化的红包类型和面额决策，显著提升了营销预算对于购票转化撬动效率。后续我们还将探索模型在大文娱

更多业务场景中的应用，比如在优酷会员营销场景中提升会员转化率。

二、基于营销增益的建模方法（Uplift Model）

营销和广告中常见的 CTR、CVR 预估模型，其关注的正样本并不是点击、购买等用户行为反馈（Response），我们称之为 Response Model。而营销增益模型人群的关键在于准确衡量和预测各种营销动作带来的“转化增量”，而不仅仅关注用户“是否转化”，我们称之为 Uplift Model。可以形式化的表达为：

$$\text{Uplift} = p(Y_i|X_i, T=1) - p(Y_i|X_i, T=0)$$

其中，X_i代表用户特征，Y_i代表该用户的实际反馈，而T则代表平台是否针对此用户施加营销干预（Treatment）。

Uplift Model 本质上是一种因果推断模型，可以用于预测/估计某种干预对个体状态/行为的因果效应（ITE, individual treatment effect）。同时也与大多数因果推断任务一样，Uplift 建模遇到的主要挑战是**反事实**，即同一个用户不能既被干预又不被干预，上述计算公式中$p(T=1)$和$p(T=0)$至少有一项是未知的，需要通过模型预估的方式得到。

营销实践中，针对 Uplift 的建模需要解决以下两个问题：

1）如何构建模型，使得可以实时预测 Uplift？与传统的 Response Model 有何差异？

2）针对反事实问题，如何构建合理的训练&评估样本集，以及采用合适的效果评估方法？

1. 建模方法

常见的建模方法有两大类：一类为沿着 uplift 定义的思路，首先建模 Response Model，然后计算干预前后的 Response 差别，得到最终 uplift，我们称之为差分模型（Differential Model）；另一类，我们可以直接根据输入的用户特征预测最终 Uplift，我们称之为直接模型（Directly Uplift-Model）。

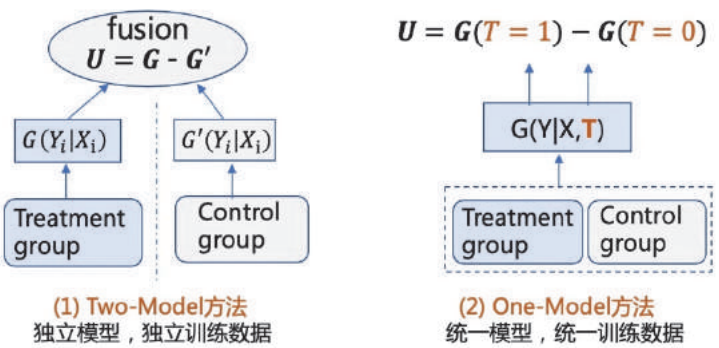
1）差分模型（Differential Model）

差分模型的思路非常直接，即利用有干预的 treatment 组样本和无干预的 control 组样本分别训练两个 Response Model，干预模型 $G_{T=1}(Y_i|X_i)$ 和非干预模型 $G'_{T=0}(Y_i|X_i)$ ，然后二者的差值记为预估的 Uplift：

$$U = G' - G$$

这种建模方式优点是原理简单，可以复用常见的 Response Model 进行建模，包括常见的 LR/GBDT/DNN 等模型。

实际应用中，G和G′可以分别训练出两个不同模型，我们称之为 Two-Model 方法；也可以用同一个模型来预测G和G′，把二者的差异通过 Treatment 特征的不同（T = 1 vs. T = 0）来表达，我们称之为 One-Model 方法。二者的训练过程和数据区别如下图所示：



Two-Model 方法建模更为直接，样本约束也更为简单，但实际应用中容易出现两个模型的 Bias 方向不一致，形成误差累积，使用时需要针对两个模型打分分布做一定校准。

One-Model 方法使用一个模型得到一组参数，避免了误差累积的情况，而且由于 treatment 组样本和 control 组样本贡献，模型训练数据相对更丰富，模型训练也相对更充分。同时，One-Model 方法针对T = i这维特征稍加改造就可以建模 multiple-treamtment 的情况，比如营销干预是不同面额的红包，甚至是不同类型的红包，大大拓宽了 uplift model 可应用的场景。

但是，无论是 Two-Model 还是 One-Model 方法，其本质都是训练和预估传统 Response Model，然后在此基础上做简单算术运算。那么在建模过程中模型更专注于 Response 的准确性，而由于缺少 uplift 相关 loss，而有可能对 treatment 特征不同带来的 uplift 微小差别并不敏感。

2）直接模型（Model Uplift Directly）

增强模型对 uplift 建模敏感度的一种方式是直接针对 uplift 建模，即模型的预估目标或 Loss Function 直接基于 uplift 来计算。这里需要针对原始模型做一定改造，具体改造方法因模型而异，这里仅简单介绍一种基于 GBDT 等树模型的构造方法。

GBDT 模型中，子树的分裂规则是关键，传统树模型的分裂规则为：

$$\Delta_{\text{gain}} = \text{info}_{after-split}(\text{D}) - \text{info}_{before-split}(\text{D})$$

而改造后的 uplift model 分裂规则可以表示为：

$$\Delta_{\text{gain}} = \text{D}_{after-split}(\text{P}^{\text{T}}, \text{P}^{\text{C}}) - \text{D}_{before-split}(\text{P}^{\text{T}}, \text{P}^{\text{C}})$$

其中，P^{T}与P^{C}分别代表 Treatment 组与 Control 组的转化概率分布，D其分布差异的度量规则，可以采用 KL 散度、欧氏距离、Chi-平方距离等常见度量方法。

可以看到，分裂规则由信息熵或者基尼等的增益，变成了 T 组和 C 组转化概率分布的差异，即模型更关注 Treatment 人群和 Control 人群的转化差异，即 Uplift。

这类建模方法对与 Uplift Loss 的量化更灵敏更精准，但实际应用中需要对机器学习模型进行大量的改造和优化，实现成本较高。而且该方法对于训练样本的要求较高，需要严格构建同质的 Treatment 组样本和 Control 组样本集，基于此才能准确计算 Uplift 值，这在实际应用中样本获取难度更大。所以在我们的营销实践中，以 One-Model 的差分模型应用为主。

2. 样本构建

Uplift Model 训练过程中，无论采用以上哪种建模方法，都会遇到前文提到的反事实问题，即样本集合中存在同一个用户 u 分别在有干预（$\text{T} = 1$）和无干预（$\text{T} = 0$）情况下的反馈时，我们才能准确衡量用户u 的真实 Uplift，而这在实际应用中几乎是无法做到的。

好在大部分的模型其实并不严格要求单用户同时存在干预前后的两种样本，即使是对样本要求最高的直接模型（Model Uplift Directly），也只需要训练集符合 CIA 假设（Conditional Independence Assumption）就好，即样本特征 X 与 T 独立（$\text{X}_{\text{i}} \perp \text{T}$），直观点的表达就是对用户是否施加干预T与用户特征是独立的。

构建符合 CIA 假设的数据集最简单的方法是随机实验，即随机选取一部分人进行干预，保证无论是 Control 组中还是 Treatment 组中的人群都符合天然分布。此外也可以采用因果推断中常见的人群 Matching、propensity-score 等方法构建无偏数据集，这里不做过多展开。

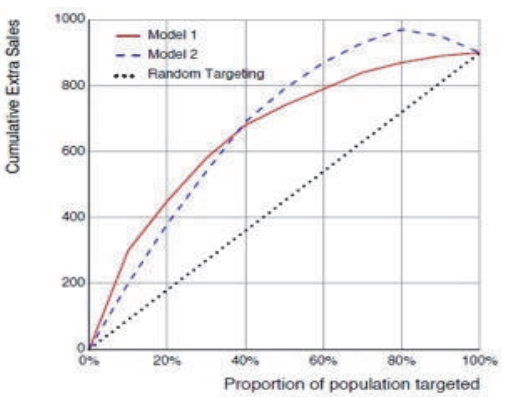
3. 效果评估

对于模型效果的离线评估中，Response Model 中常见的 AUC，ROC，Fscore 等常见的指标也无法完全胜任，原因主要在于这些指标主要客户模型对于 Response Groud-truth 的拟合程度，

而并不关注 Uplift。

但是类似的，有了满足 CIA 假设的无偏数据集，我们可以构建基于 Uplift Ground-truth 的评估指标：

- 当 control/treatment 分组符合 CIA 假设时，可按照两组用户的预估 uplift score 进行降序排列来对齐构造“镜像人群”；
- 针对每组镜像人群对，可以计算其 Response 表现差异即为 Uplift Ground-truth，在图形中表达出来记为图中的 Uplift Curve；
- 与 AUC 类似，曲线之下的面积，即为 AUUC 指标（Area Under Uplift Curve）。



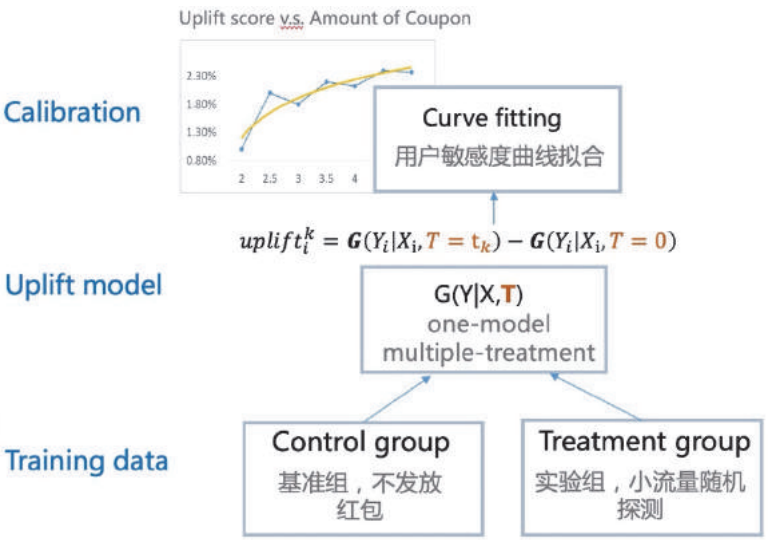
除了 AUUC 之外，还有其他一些评估指标可以衡量 Uplift Model 效果，这里不做展开，只是简单罗列几个：

- Valid-Gini：越高的基尼指数，意味着模型表现越好
- Monotonicity-of-Incremental-gains：模型打分越高，uplift 越高，随着模型打分的降低，uplift 依次递减
- Maximum impact：关注可以达到的最高的 uplift

三、Uplift Model 在智能红包中的应用简介

在淘票票等智能红包业务中，关键要解决的问题是“发给谁、发什么、发多少”。在预算有限和 ROI 等约束之下，只有准确识别对红包敏感的人群，并且根据实际转化增益发放精准面额的红包，才能提升红包资金的利用效率，以及最终提升平台购票转化率。

我们基于 Uplift Model 方法，直接建模每次红包发放对于每个用户购票转化增量的期望。这里，由于我们构建了实时预估模型，用户“敏感度”不再仅仅是对存量用户的一个静态描述，而可以针对“用户 x 场景 x 影片”做到 PV 级的增量预估，其基本流程见下图。



在 Uplift Model 的建模选择中，我们采用了 One-Model 的差分模型方法，基于以下考虑：

- 传统的 CTR/CVR 预估模型比较成熟，用户、场景、内容等基础特征的表示和深度模型可以直接复用，并且 PV 级的大规模训练和实时预估基础设施已比较完善。我们只需要关注无偏样本构建，以及模型中与 Treatment 特征和 Uplift 相关的模型改造；
- 基于 One-Model 方法，可以平滑支持多面额的 Multiple-treatment 建模，建模粒度更精细。当然在实践中，对于多面额的曲线拟合需要更高精度和更多数据，可以结合校准等方法做些后处理，使曲线更平滑更符合先验假设。

实际红包发放过程中，在实时预估出精细量化的“红包敏感度”之后，还需要解决“预算&ROI 约束下的最优面额决策问题”。这本质上是一个带约束的组合优化问题，可以用分组背包等方法求解，因为这超出了本文的范畴，这里不做过多阐述。

四、总结和展望

本文简要介绍了营销增益模型的定义和建模方法，给出了从关注 Response 到关注 Uplift 的营销建模思路，使得在红包等智能营销实践中，算法可以更“直接”的去优化最终目标，模型

效果和数据利用效率都得到提高。

沿着增量建模的思路走下去，还有更多的优化值得去探索。比如，无偏样本的构造是增量建模的关键问题，在随机实验之外，通过人群 Observation Study 的方法更充分利用历史存在的有偏样本，可以有效提高样本利用效率；在模型层面上，也可以通过 Multi-Task Learning 的方法将多场景的任务，甚至 Uplift 任务和 Response 任务进行联合建模，提升模型学习效率，降低对无偏样本规模的依赖。

智能营销实践中，营销玩法越来越复杂，场景越来越多样，也给算法建模提出了更多的挑战。比如，智能红包业务中，券的类型和面额越来越精细，对应的 Treatment 组合也维度剧增；同时，多种营销工具共存带来了多模型间的互相影响，用户对于营销玩法的不断适应存在用户与平台的长期博弈。这些挑战都需要在营销建模方法上有更多突破，路漫漫其修远兮，吾将上下而求索。

外投 DSP 自动报价算法实践

作者| 阿里文娱高级算法工程师 晨翊

一、报价算法对于 DSP 的意义

在广告竞价业务中，自动化报价是不可或缺的一环。业务要求成本和预算是可控的，需要在这些约束下实现最大化量的目标。在最开始，没有自动化报价算法，我们采用了手动调节，固定出价的办法，这种方案一方面需要大量的人力，另一方面经常要么成本过低，没能最大化利用流量，要么成本过高，买的用户不具备性价比。基于这些原因，我们探索并开发了多种自动化报价算法，解决了满足约束条件下的最大化 DAU 量问题，不仅减少了人力耗费，且有效利用了流量。

1. 累积 PID 报价

在解决 DSP 冲量投放场景的报价问题时，我们存在着单一约束，即用户点击成本或转化成本约束，而我们的冲量目标在于最大化首活 DAU 量。根据业务情况我们产生以下形式化定义：

$$\max_{x_i} \quad \sum_{i=1}^{n} x_i * CVR_i \tag{1}$$

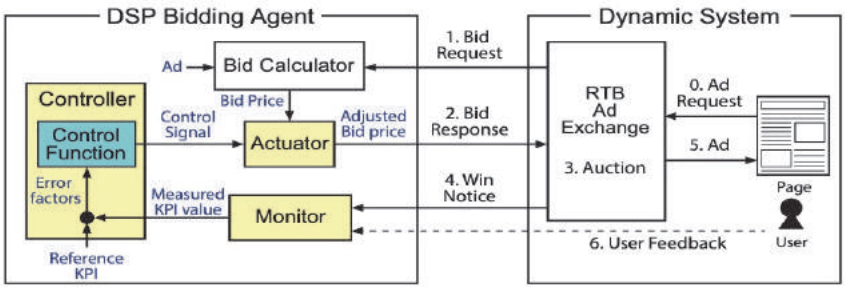
$$s.t. \quad \frac{\sum_{i=1}^{n} x_i * wp_i}{\sum_{i=1}^{n} x_i * CVR_i} \leq C \tag{2}$$

xi表示第 i 次竞价是成功，1 表示成功否则 0，$CVRi$表示第 i 次转化率预估值，wpi 是第 i 次竞价成功的扣费值。我们需要最大化转化数量，但是存在着点击成本或转化成本的约束，即(2)式所表示的。根据计算，我们能得到以下类似出价公式

$$bid_i = \alpha * CVR_i \tag{3}$$

从以上出价公式中得知，通过控制 a 变量，我们能实现满足(2)公式的约束。那么 a 的值该如何计算？我们通过引入 PID 控制器来解决该问题。同时，由于(2)约束属于最终约束，我们借

鉴《Bid Optimization by Multivariable Control in Display Advertising》（KDD 2018）中的办法对其进行修正，从而实现稳定满足最终约束的出价方式。



如上图所示，DSP 内部的核心控价组件由控制函数与执行函数组成，监视器则提供对竞价 Feedback 数据的处理和统计，统计后的数据由控制器进行计算输出控制信号，最终由执行器根据信号对出价进行放大或者缩小，从而满足(2)中的约束。

常用控制器函数的计算方法：

$$e(t) = r(t) - y(t)$$

$$u(t) = k_p e(t) + k_i \sum_{i=1\ldots t} e(k) + k_d(e(t) - e(t-1))$$

e(t)表示 t 时间步的误差，及期望 KPI 即 r(t)与实际值 y(t)之间的差值，这里的 KPI 可以是点击成本、转化成本，甚至竞价成功率等。然后通过第二个公式计算出信号 u(t)。在我们这里面，由于是希望最终成本满足约束，于是对误差计算和信号计算进行修正

$$e_q(t) = click(t) \cdot (r(t) - y(t))$$

$$u_q(t) = \frac{1}{\sum_{i=1\ldots t} click(t)} \cdot u(t)$$

如图所示，误差计算我们改为原有的误差乘以 t 时间步内的点击数或转化数(转化数对应着转化成本约束)，然后按照常用控制公式计算出 u(t)，最后用第二个公式将信号修正，即用信号值除以从开始到 t 时间所产生的点击数。这样每个时间步内的点击数量对信号也产生了影响，

同时点击数累加的分母有助于使信号的变化满足最终的点击约束，同时还增强了稳定性。

控制器产生了控制信号后，然后通过执行器对原始出价进行放大或者缩小，我们采用指数形式的执行器：

$$x(t+1) = x(0) \cdot exp(-u(t))$$

x(t+1)表示 t+1 时间步的出价，其值为初始出价 x(0)与后面的指数信号变换器的乘积。原始出价值 x(0)的选择需要一定的技巧性，可以利用历史投放的数据计算或者搜索出来，能有小减少控价波动。同时，针对 PID 较为难调的问题，可以将中间变量可视化，能提供一定的帮助。

2. 双 PID 报价

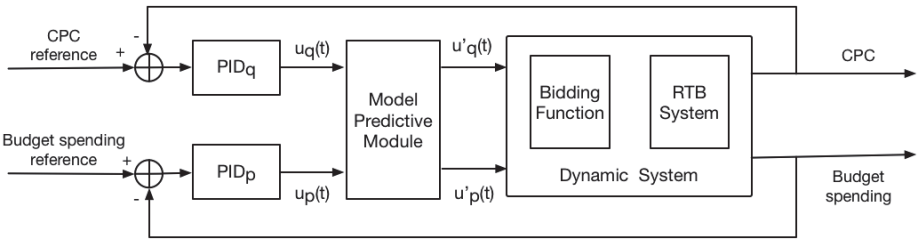
上一部分主要是针对单约束场景，但在实际业务中，我们面临着更多的约束。例如我们的日常投放场景中，经常会遇到预算问题。即在预算不足和存在成本约束的条件下，最大化首活 DAU 量。将问题形式化如下：

$$\begin{aligned} \max_{x_i} \quad & \sum_{i=1\ldots N} x_i \cdot CTR_i \cdot CVR_i \\ \text{s.t.} \quad & \sum_{i=1\ldots N} x_i \cdot wp_i \le B \\ & \frac{\sum_{i=1\ldots N} x_i \cdot wp_i}{\sum_{i=1\ldots N} x_i \cdot CTR_i} \le C \\ \text{where} \quad & 0 \le x_i \le 1, \forall i \end{aligned}$$

相对于最早的形式化，我们增加了一个新的约束，即所有竞价成功花费的钱不能大于 B（budget）。根据以上公式，我们能够解出类似以下的出价公式：

$$bid_i = \frac{1}{p+q} \cdot CTR_i \cdot CVR_i + \frac{q}{p+q} \cdot CTR_i \cdot C$$

公式引入了两个参数 p 和 q，通过 p 来控制预算约束，而通过 q 控制成本约束。然后我们通过双 PID 控制器来实现这些控制目标：



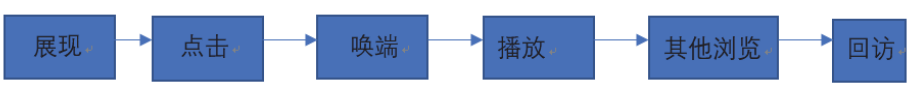
PIDq 和 PIDp 分别控制点击成本和预算约束。当通过控制器计算出各自的信号之后，需要建模两个 PID 信号之间的相互影响作用，采用以下线性插值的方式实现，也就是图中的 Model Predictive Module 组件：

$$\begin{bmatrix} u'_p(t) \\ u'_q(t) \end{bmatrix} = \begin{bmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{bmatrix} \begin{bmatrix} u_p(t) \\ u_q(t) \end{bmatrix}$$

通过引入参数α和β两个参数对原有信号进行修正，从而建模两个信号的补偿关系。最后分别将两个信号送入各自的执行器，再由出价公式和预估的 CTR/CVR 计算出出价值。

二、用户价值预估

在以上报价技术介绍中，提到了 CTR、CVR 预估，这是整个以竞价实现用户增长业务中最重要的一环。以下是用户在竞价成功之后的行为路径：



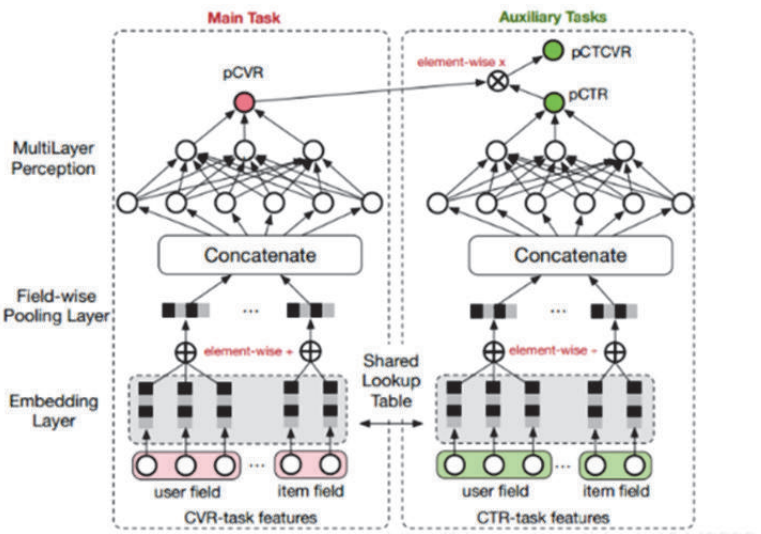
在整个过程中，涉及多方面的预估，如点击率(CTR)、唤端率(CVR)、播放率、效用预估、留存率等。其中效用是指用户在进站之后带来的广告消费、会员购买、视频消费甚至站内互动、多天回访等带来的一系列收益，我们可以用效用代表用户的最终价值。

理想是美好的，但现实是残酷的。实际上效用涉及到很多方面，很难做一个具体的预估。通常实践中，我们采取一些简单的办法来代替这部分，如点击率、唤端率、播放率等，用这些目标代表用户价值。

常用的 CTR 预估模型有：LR、GBDT、FM、DNN 各种种类。除去 DNN 类，其他的基本上都是对单一目标进行预估。LR 模型的优点在于训练和打分速度都很快，耗时少，且模型可以控制在较小的规模，缺点在于需要大量的特征工程引入非线性，且模型精度上限较低。GBDT

的优势也在于快速训练和打分，且能做自动特征筛选和组合，可达到较为不错的精度。DNN 类模型预估精度高，在未来可融合 CV 和 NLP 类的内容理解特征，且可通过网络结构的变化实现特征的自动组合，但也存在巨大的缺点：一方面很耗资源，训练慢，另一方面打分慢，RT 高。由于我们业务的延时要求高，在 100ms 内必须返回，现有工程技术条件下，DNN 类模型由于过高的延时而没有被采用，LR 和 GBDT 则因为资源消耗小，RT 基本满足要求成为了我们业务的主力模型。LR、GBDT 只对单一业务目标进行预估，而根据我们以上用户行为路径分析，则存在多个任务及目标的预估问题。所以用户价值预估方面还是要继续解决工程上的问题，走 DNN 模型方式，以实现多目标预估，实现对最终用户价值的逼近。

多目标最开始是通过多个模型实现，由于 DNN 的引入，使得通过网络结构的变化及 Loss 函数的创新，能够实现多目标预估。例如阿里巴巴公司提出的 ESMM 模型：

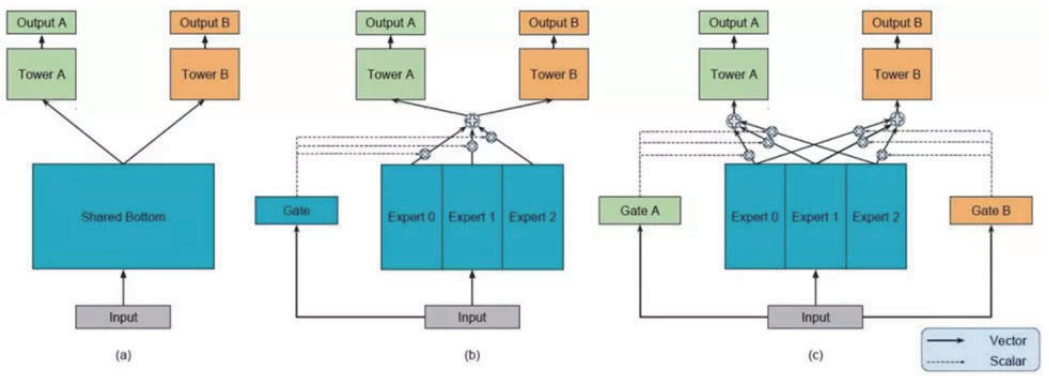


该模型的创新点在于将 CTR 和 CVR 任务统一在同一个空间内建模，虽然是两个网络结构，但在底层共享 embedding 特征结构，也就是所谓的 Share-Bottom MTL 结构。在上层分别预估 pCTR 和 pCVR，并将其相乘得到 pCTCVR。Loss 函数定义为：

$$L(\theta_{cvr}, \theta_{ctr}) = \sum_{i=1}^{N} l(y_i, f(\boldsymbol{x}_i; \theta_{ctr})) + \sum_{i=1}^{N} l(y_i \& z_i, f(\boldsymbol{x}_i; \theta_{ctr}) \times f(\boldsymbol{x}_i; \theta_{cvr}))$$

即由两个子任务的 Loss 函数之和构成。在 CTR 任务中，有点击行为的曝光事件标记为正样本，没有点击行为发生的曝光事件标记为负样本。在 CTCVR 任务中，同时有点击和购买行为的曝光事件标记为正样本，否则标记为负样本。

而 Google 公司在 2018 年 KDD 会议上发表了一个新的多任务模型，称之为 MMoE。针对 Share-Bottom MTL 结构，在一些任务中，底层表示差异过大，噪声过多，差异冲突等问题，MMoE 显式建模任务与任务之间的关系。



以上图中(a)为常见的多任务实现方式，即共享隐层。(b)、(c)中将共享的底层表示层分为多个 expert，同时设置了 gate，使得不同的任务可以多样化的使用共享层。(b)是加入单门（one gate）的 MoE layer 的多任务学习模型，(c)是 Google 提出的 MMoE 模型。

通过从单任务预估到多任务多目标预估，我们可以更好的实现用户价值逼近，从而在引流过程中获得更多有价值的用户，使其在我们的产品上产生更多的活跃性、更多的消费、更长的使用时长、更多的效用。

7

搜推统一分发系统

本节摘要

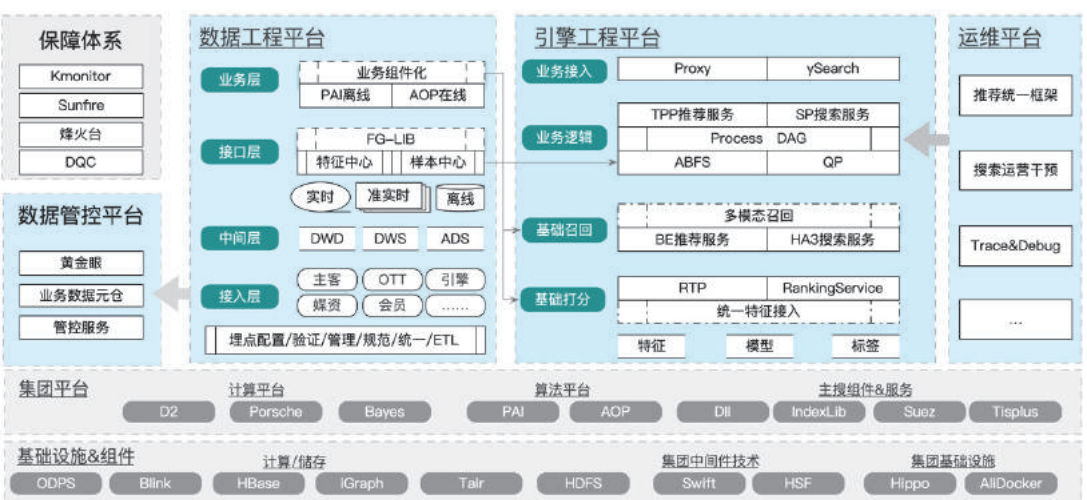
作者| 阿里文娱高级专家 治庸

搜索和推荐是优酷内流量最大的两类算法分发场景，优酷投放引擎基于集团中台技术，构建起弹性高扩展性的基础架构，稳定索引数亿视频内容，高性能支撑数十亿 PV 请求；在此基础上，优酷引擎架构立足于视频平台，快速支撑产品发展和算法迭代，并针对过程中的业务域痛点和特点，对整体架构进行了系统化的思考和探索：

1）推荐场景，针对倍数级场景接入和算法迭代要求，我们基于图执行引擎封装了一套算法服务框架，很好提升了场景迭代效率；

2）视频作为高维多模态的数据载体，我们开创性的研发了多级多模态召回引擎，为视频理解后的结构化检索提供了新的架构支撑；

3）将图谱体系的思想引入到内容数据更新的架构设计中，构建起面向视频域的实时特征和索引更新平台。



基于图执行引擎的算法服务框架

作者| 阿里文娱高级专家 随方、阿里文娱开发专家 轩成

一、背景

在阿里的业务中，有广泛的算法应用场景，也沉淀了相关的算法应用平台和工具：基础的算法引擎部分，有成熟的召回和打分预估引擎、在线实时特征服务；推荐算法应用领域，有算法实验平台 TPP（源于淘宝个性化平台），提供 Serverless 形式的算法实验平台，包括资源弹性伸缩，实验能力（代码在线发布、AB 分流、动态配置），监控管理（完善的监控报警、流控、降级）等能力，是算法在线应用的基石。

但在实际的算法应用业务中，比如优酷推荐业务，算法应用场景众多（100+活跃场景），需求灵活多变，如果没有一套通用业务框架，用于抽象出通用和定制化的部分来提高算法组件的复用度；会严重拖慢算法实验的节奏。基于图引擎的算法服务框架就是为了封装一套框架，抽象算法在线服务的通用算子，支持运行时的算法流程的装配，提升算法服务场景搭建的效率。

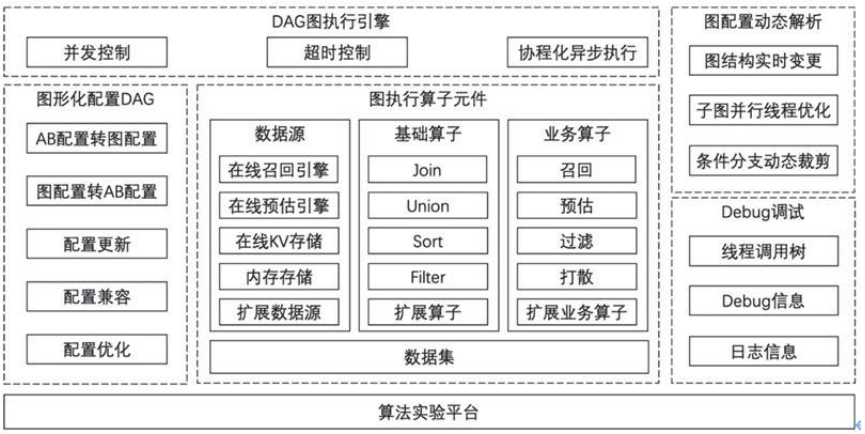
二、设计概览

算法推荐典型在线处理执行流程：多路粗排召回，合并，预估，打散策略。推荐服务根据用户的设备 ID 等其他必要信息进行多路并行召回，在召回引擎中进行粗排后，经过必要的过滤处理，截取一定数量的内容调用 Rank 引擎进行精排预估，预估结果经过一系列算法策略处理后输出最终结果。

整个过程中召回，合并，预估，打散等业务处理有并行处理，有串行处理，根据业务需要能够灵活配置。基于图的推荐业务执行引擎是运行在算法实验平台上的执行引擎，它的典型处理流程是：在 AB 实验分桶上，通过图形化交互页面配置数据源、业务算子的执行依赖关系，并配置每个算子的运行时动态参数。

系统总体结构如下图所示：分成五个主要的模块（DAG 图执行引擎、图执行算子元件、图

形化配置 DAG、图配置动态解析、Debug 调试）。



图：系统总体架构

当推荐请求到达时，引擎读取 AB 参数，根据参数上配置的算子信息通过反射机制创建算子实例，动态组装成可运行的 DAG。根据条件分支配置，动态裁剪运行时的 DAG 实例，根据图运行占用最大线程数配置，动态调整线程复用。算子通过算法实验平台的底层协程池并行运行。

三、关键模块

1. 图执行算子元件

1）数据集

在 DAG 图中流转的数据统一封装为 DataSet 数据集，数据集是结构化多行二维数据的封装，在数据集上封装便利的基础算子操作。

数据集上一系列处理操作基于 Java 的 Stream API 来进行处理，以此来达到集合处理的最好性能，将非 Action 操作延迟到最后数据处理时运行。

2）数据源

将能够返回数据或者数据交互的二方服务封装为通用数据源，所有业务算子围绕数据源的数据进行业务开发，通用数据源包括召回数据集、在线算法需要的辅助数据集（如存放在 KV 内存存储的旁路召回数据、特征等数据）、打分预估结果集、内存数据源等等。

数据源的封装通过动态参数配置方式实现通用性和可扩展性。数据查询只需要修改配置即可实现数据获取，不需要开发代码。

3）基础算子

在 DataSet 数据集上封装的基本操作作为基础算子，比如 Join、Union、Filter、Sort、Map、Collect 等流式操作。在 DataSet 上重新封装 Stream 相关 API，便于对 DataSet 进行流式处理。

4）业务算子

召回、预估、合并、打散、过滤等业务操作封装为业务算子，在业务算子中可以查询数据源，返回数据集后通过基础算子计算得到结果。

2. 图形化配置 DAG

3. 配置动态解析和优化

1）根据 AB 配置实时变更图执行结构

图引擎在运行时为了减少解析图结构的耗时，将图结构进行了缓存，在 AB 配置更新时需

要实时反映到图引擎中，所以根据图配置的哈希值校验的方式检测图配置是否更新，图结构变更后会重新创建引擎实例。

2）子图并行线程优化

在 DAG 执行时，所有算子都交给线程池异步运行，但是在大多数情况下子图可能是一个顺序执行图，不需要并行，不应该占用其他线程，所以在图执行时，动态根据依赖关系识别节点是否需要占用新线程运行。

3）条件分支动态裁剪

如果图结构中存在条件节点，会根据条件节点的动态结果裁剪后续图节点的运行。如果一个图节点的执行条件为否，后续单独依赖它的节点都不会运行，条件节点具备传递性。如果后续节点不单独依赖不运行的节点，则当前节点可运行。

4）子图并行线程优化

在 DAG 图中流转的数据统一封装为 DataSet 数据集，数据集是结构化多行二维数据的封装，在数据集上封装便利的基础算子操作。

数据集上一系列处理操作基于 Java 的 Stream API 来进行处理，以此来达到集合处理的最好性能，将非 Action 操作延迟到最后数据处理时运行。

4. DAG 图执行引擎

1）并发控制

通过图中依赖关系动态解析节点需要通过并行还是串行执行，在图中配置最大并发线程数来控制图的最大并发度。最大程度复用线程，减少线程切换带来的开销。

2）超时控制

通过整个图上配置超时时间来控制图的超时，根据业务粒度，会将子业务配置为子图，从而通过控制子图的超时时间来控制子业务的超时时间。

3）通过协程优化异步执行

DAG 运行依赖线程池运行，算法实验平台提供了基础线程池，并同时将线程池在 JVM 层面优化为协程，通过压测比对，普通线程池的性能要低于协程池的性能。

四、总结&展望

基于图引擎的算法服务框架建设，通过抽象算法业务的通用组件，提供图形化流程编排工具和图执行引擎，实现了 0 代码、配置化支持算法业务需求。为快速的算法应用，不断提升用户的个性化服务打下了坚实基础。对推荐、搜索、广告等算法应用业务有参考价值。

接下来，为了进一步提升引擎性能，我们将在构图优化和引擎执行性能上做优化，在保持业务表达灵活简洁的同时，追求更优的执行性能。

面向多级多模态场景的召回引擎

作者| 阿里文娱开发专家 崇懿、阿里文娱开发专家 慧善

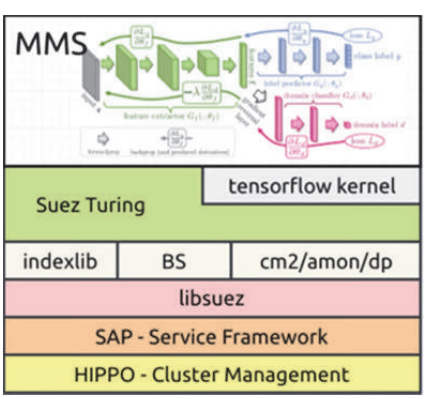
一、背景

随着智能手机及移动互联网的快速发展，人们接触到的多模态数据在数量和类别上都在飞速增长。计算、存储能力不断突破，人工智能技术也得以发展，在云、端测中 CV 技术、AR/VR 技术都为人们提供更便利、智能的体验。

优酷作为视频平台拥有海量 OGC、UGC 视频内容，视频内容数据是一个高维度多模态的数据，有标题、简介、评论等文本信息，有视频帧的图像信息，有声音，也有连贯的动作视频片段。传统的基于倒排索引的搜索引擎只适合检索文本信息，对于多媒体内容检索能力不足。

为了让用户更便捷的找到多媒体内容，增加多模态搜索能力，开创性设计与研发多级多模态搜索引擎（MMS）。提供分布式大规模多层级多模态索引能力，低延时跨模态级联检索能力，多层级检索、融合、排序能力。

二、系统概况

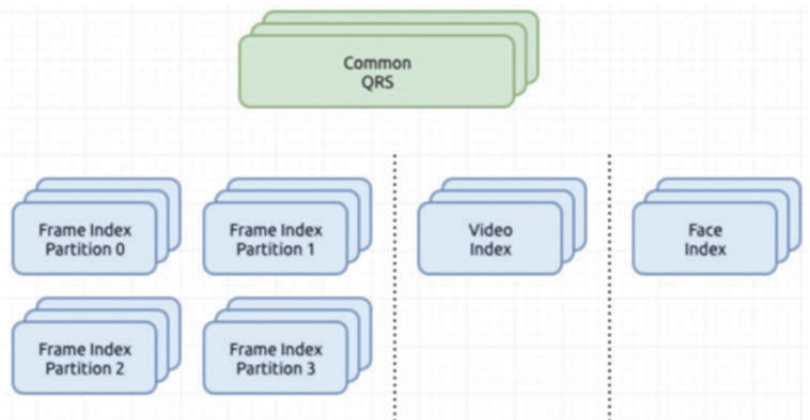


基于阿里中台的 Hippo(在线服务调度)、SAP(应用服务框架)等基础设施开发部署。MMS 主要是在索引结构、检索控制、执行框架、部署等方面做了系统设计和选型。

三、关键技术

1. 分布式多级多模态索引结构设计

每层级独立构建分布式索引，索引类型包括倒排及向量索引。以视频、帧（图片）、人脸举例，索引结构如下：

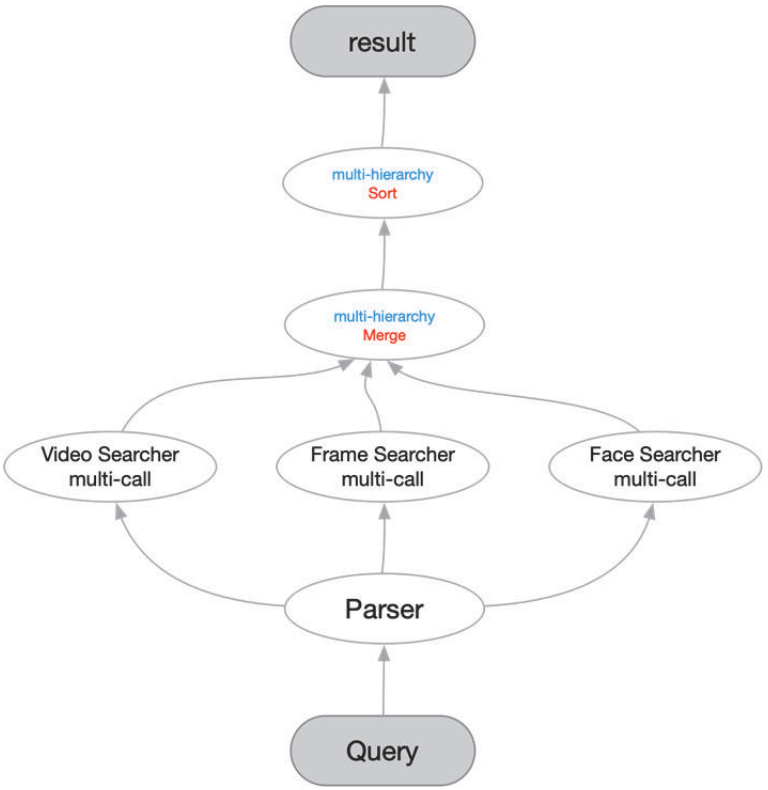


视频、帧、人脸存在层级关系；同时帧图片及人脸都有表征向量建的向量索引，支持亿级别规模，分 10 个分片，Top10 的召回率 90%以上。

2. 检索调度

MMS 的复杂性在于其在线检索逻辑，在支持层级及多模态 query 输入的基础上，如何控制跨层级、跨模态的检索。通过定义标准的跨层级和跨模态准则，根据用户的输入形成在线检索逻辑。

基本的检索流程如下图：



会由 multi-call 进行多层级、多模态扩展查询逻辑，其中关键是多级、跨模态的扩展查询逻辑。

1）跨层级

跨层级的检索由用户输入的层级作为起始检索点，用户想要的输出作为终点作为扩展，系统具有自适应推理能力。

2）跨模态

跨模态检索会有两种形式的解决方案，使用不同场景，索引构建前，不同模态数据做统一表征，映射到统一空间，在线进行向量检索，此处内容表征的占主要，MMS 主要解决是在文本、向量间进行跨模态检索。

3. 图化执行引擎

复杂检索逻辑及低延迟服务能力要求，需要有高效的执行框架，图化执行引擎具备最大限

度并行能力。同时对算子进行抽象，可以更自由编排及复用。

MMS 选择 Suez 图化执行引擎，采用 DAG 执行引擎+业务逻辑算子的实现方式。

4. 通用性算子实现

搜索逻辑中会设计 query parser、merge、sort 通用逻辑，基于图化执行引擎接口实现通用算子：

1）query parser 算子负责解析请求，查询串使用简单文本方式，相对于 pb/binary 方式，可视化的查询串更加直观，同时查询语法简单且强大。查询串支持查询文本查询，向量查询，或者同时有两者，支持高级语法，可以控制的查询参数粗排精排等；

2）merge 融合多层级 doc，补全所需要的正排、summary 信息；

3）sort 是搜索排序逻辑，排序后选取 Top N 返回；

4）result 是结果返回和处理逻辑，基于查询使用文本方式，我们希望结果也是直接可视化，所以在构建结果的时候支持了 json/xml，同时为了兼顾性能，我们也支持 protobuf 的返回格式，同时还加入 snappy/lz4 方式压缩，使返回结果集更小，传输效率更高；为了方便调试，我们加入了调试参数，可以保存聚合调试参数，输出引擎内部的调试内容。

四、产品应用

1. 优酷智能搜索

采用 MMS 对视频、帧、元素（人物、动作）等多级内容进行索引，召回出视频解构后的信息，可以实现定帧播放，支持用户对于精准视频内容片断的需求。

2. 以图搜剧

用户可以通过拍照、上传图片搜索人物及节目、相似画面的视频。输入态丰富为图像，召回系统采用 MMS，既具备传统的通过人脸识别后用人名召回节目，又可直接通过图片向量进行召回。

五、总结&展望

多媒体信息不断丰富，直播、小视频等相关应用增长迅猛，5G 移动通信技术的进一步普及，多媒体信息的生产、传播将会持续爆发式增长。人工智能技术日臻成熟，对于多模态内容理解、表征会进一步加强。多模态的人机交互体现会渗透到生活、生产各个环节。多级多模态的检索能力是必须要面临的核心问题。

优酷的多级多模态搜索引擎（MMS）提供了低延迟的跨模态、跨层级搜索能力，支持大规模多模态的索引。在视频分发、视频创作中都有着十分关键的应用场景。MMS 技术在更多的智能交互场景也将发挥更广泛的应用场景。

基于内容图谱体系的特征与索引更新平台

作者| 阿里文娱开发专家 遨翔、阿里文娱开发专家 玄甫

一、背景

搜索推荐系统作为在线服务，为满足在线查询的性能要求，需要将预查询的数据构建为索引数据，推送到异构储存引擎中，供在线查询。这个阶段主要通过 Offline/Nearline 把实时实体、离线预处理、算法数据进行处理更新。这里包含了算法对这些数据在离线和在线的处理，不同业务域之间的协同（召回、排序、相关性等），在平台能力方面采用传统的数仓模式即围绕有共性的资源、有共性的能力方面建设，形成分层策略，将面向业务上层的数据独立出来，在实现业务敏捷迭代、知识化、服务化特征方面已不能很好满足需求。

知识图谱作为对数据进行结构化组织与体系化管理的核心技术，在实际面向业务侧应用过程中能很好的满足知识化、业务化、服务化方面的诉求，基于内容图谱体系的特征平台建设，以内容（视频、节目、用户、人物、元素等）为中心，构建一个知识融合特征和实时数据更新平台。

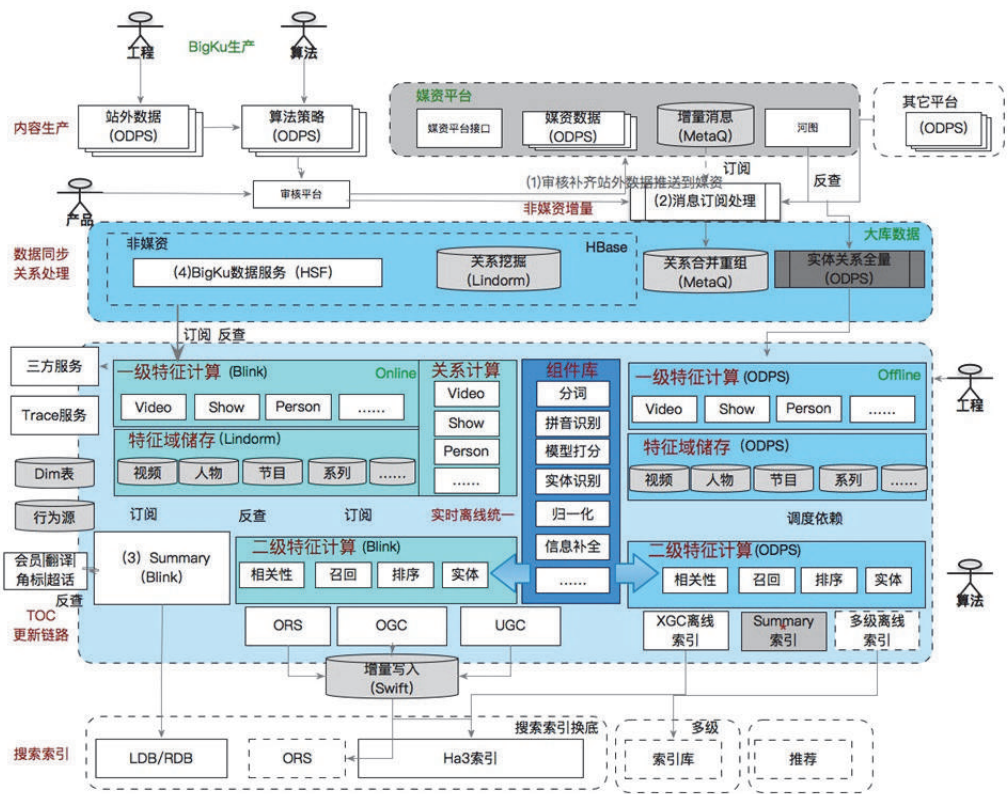
二、设计概要

基于搜索推荐系统数据处理链路一般包括以下几个步骤：从内容生产端（媒资、互动、内容智能、包罗、粮仓、琳琅等）接收 dump 出来的全量数据和业务侧增量数据，然后业务侧拿到这些数据按业务域进行一层一层加工，最终通过 build 构建索引进入到引擎端。不同于其它业务场景，在优酷场景中我们接收的内容生产端并不是源头生产端，中间掺杂了很多半加工的异源异构数据，数据的一致性（逻辑一致性、功能一致性）是摆在用户侧实际性面临问题，特别实时和全量产出的数据需要保持结构一致，同时搜索引擎的字段结构保持一致。我们从数据结构化组织与业务体系化管理方面进行索引平台更新设计。

数据结构化组织：设计文娱大脑面向应用侧的中间层，将知识图谱引入中间层，实现了面

向业务领域的数据组织方式。将知识图谱融入在中间层数据模型这一层，用包含实体、关系、事件、标签、指标的知识图谱统一视图来定义面向领域的数据模型。将视频领域知识图谱作为中间层数据组织的基础，实现面向业务领域数据组织方式的转变。

业务体系化管理：将算法的逻辑以组件化的模式进行封装，实现了业务方只需要维护一套逻辑，实时和全量一套代码，采用统一 UDF 来实现。利用 Blink 的流批一体化的架构，实现全增量架构模式，如在数据清洗订正逻辑时进行全量（实时引擎中做到了消息不丢的机制保证，不需要每天实现全量），让全量数据走一遍逻辑。



三、关键模块

1. 特征库

1）特征库包含两层，第一层是全增量的一级特征计算，主要对接不同的数据源（包括实时

和离线），在特征域计算中不存在离线全量，对于冷数据或修正数据采用存量的全集重新走一次流处理。整体采用顶点和边关系的储存，在实时更新过程中为了减少对上游的反查导致的性能压力，不同实体属性的变更直接走内部图查询，统一封装了 DataAPI 进行对这些数据进行操作，不同的顶点都边采用独立的一个 blinkjob 进行计算；

2）在离线能力方面，由于搜索引擎在线服务的机器并不持久化数据。 当新的在线机器加入集群时需要从某个地方拉取全量文件，加载到内存里，由于全量文件只是某一个时间戳的快照，全量文件的时间戳越早需要追的实时消息就越多，故障的恢复速度就会越慢，需要有一个机制尽量及时的产出最新的全量文件，减少实时和增量消息带来的性能压力；

3）二级特征计算，面向算法的接入，包含了搜索的相关性、排序、召回这层直接面向业务域，它直接消费一级特征库中的数据，业务主要逻辑集中到这层进行计算，此时实时离线逻辑主要通过组件库来完成。

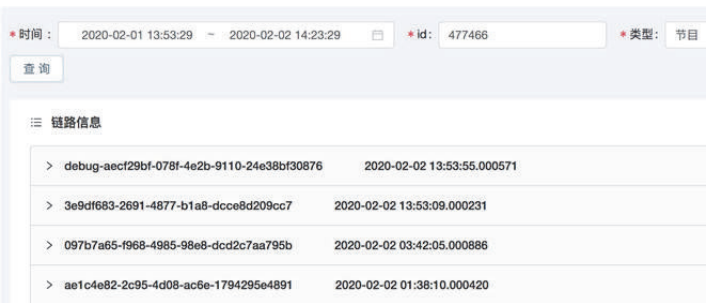
2. 组件库

各个业务线算法都是按各自的业务从同一份数据中获取自身需要的数据进行处理，这无形中就导致了代码的重复。组件库建立主要应开发适配的接口，让相同功能的代码得以复用，避免重复开发。

组件库将业务逻辑抽象成简单的基于 UDF 的算术表达式来组织，简单、简洁，并且更易维护，特征使用方，只需关注特征粒度，不需要关注整体。

3. Trace&Debug 模块

每一个消息有唯一的一个签名，从源头数据一致会在各个流程中流转，同时在流程处理过程中为了便于业务更好理解处理流程，将不同的数据按 uuid 和实体 id 进行聚合，通过 Trace&Debug 服务能较好的理解业务处理过程中的业务信息和系统处理的信息。



四、技术细节

整体计算框架采用新一代的实时计算引擎 Blink，主要优势在于流批一体化，业务模块通过 job 切分，不同的计算模块可以随意组合；消费位点自动保存，消息不丢失，进程 failover 自动恢复机制；分布式的计算可消除单点消费源和写入性能瓶颈问题。储存引擎采用 Lindorm 进行实体数据储存，主要利用 Lindorm 二级索引来储存 KV 和 KKV 数据结构，用于构建知识图谱的底层数据。

1. 知识图谱储存和组织

采用标签属性图（Labeled Property Graph，LPG）建模，Lindorm 作为主储存，实体表（视频、节目、人物等）作为顶点表，实体间关系利用 lindorm 的二级索引能力作为边表。

数据访问方面，实现数据驱动层，提供给外部使用的接口 API，开发人员通过本地 API 来操纵 Lindorm。接口层一接收到调用请求就会调用数据处理层来完成具体的数据处理，屏蔽了 java 代码属性和 lindorm 列值的转换以及结果查询的取值映射，使用注解用于配置和原始映射，解决 java 对象直接序列化到 lindorm 的行列储存问题。

2. 计算和更新策略

采用 Blink 计算平台实现特征计算和索引更新，由于采用了全增量架构，在全量更新过程减少上游服务反查的压力，采用列更新策略。在不同实体属性或边表属性（边表属性为了减少图查询过程中顶点查询的压力开销）更新采用级联更新策略，即属性更新后生成新的消息推送到总线链路端，不同实体或关系订阅消息后按需进行自身属性更新。

更新一个业务核心诉求就是一致性，其本质就是不丢消息和保序，我们采用 MetaQ 作为主消息管道，本身具备不丢消息，更多是在外部服务、储存、处理链路层面上失败情况。

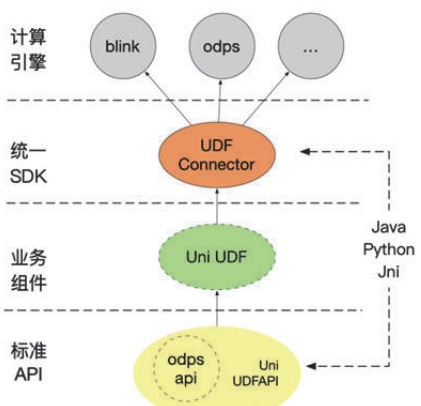
对于一个实体数据或关系数据（通常一个 job），采用原子操作，内部有一定的重试机制，如访问外部服务，自身会有重试机制，这种重试为了不影响整体的链路性能我们称作 Fast try，一般应对网络抖动如超时等情况，如果失败会保留一定现场，将数据写入重试队列中，抛出异常由最外层捕获异常，丢弃本次更新接受下一个消息，失败的消息会在 5 分钟、10 分钟，20 分钟去重试 3 次如果依然失败则发出通知人为干预。

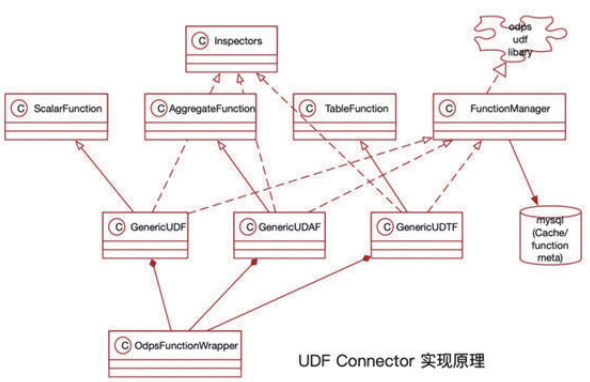
3. 计算和更新策略

采用 Blink 计算平台实现特征计算和索引更新，由于采用了全增量架构，在全量更新过程减少上游服务反查的压力，采用列更新策略。在不同实体属性或边表属性（边表属性为了减少图查询过程中顶点查询的压力开销）更新采用级联更新策略，即属性更新后生成新的消息推送到总线链路端，不同实体或关系订阅消息后按需进行自身属性更新。

4. 统一 UDF

采用核心解决 UDF 的业务逻辑，在各个系统之间的可移植，通过技术手段保证只维护一套业务逻辑，各个计算平台（离线/实时）可复用，解决 UDF 业务逻辑的一致性和可移植性问题。





五、总结&展望

基于内容图谱结构化特征与索引更新平台，在结构化方面打破传统的数仓建模方式，以知

识化、业务化、服务化为视角进行数据平台化建设，来沉淀内容、行为、关系图谱，目前在优酷搜索、票票、大麦等场景开始进行应用。

随着用图神经网络、表征学习方面不断的发展，进一步在图存储和图计算在面向 OLTP 和 OLAP 进行着重深度优化，通过深度算法策略来补充实时融合和实时推理方面的建设。

在索引更新平台建设方面，随着多方业务的接入、搜推融合带来的挑战，索引更新朝向全增量化的进行推进，在业务自助方面，进一步探索抽象 DSL，提升业务整体接入效率。

关注我们

（阿里文娱技术公众号）

关注阿里技术

扫码关注「阿里技术」获取更多资讯

加入交流群

1）添加“文娱技术小助手”微信

2）注明您的手机号 / 公司 / 职位

3）小助手会拉您进群

By 阿里文娱技术品牌

更多电子书

扫码获取更多技术电子书