

2017年10月刊

A I - F R O N T



关注落地技术，探寻AI应用场景



# 卷首语

## AI 前线伴你踏上人工智能征途

刘志勇

人工智能方面的科幻作品层见迭出，从 *The Matrix* 到 *Neuromancer*，不断引发广泛热议。如果要问未来什么最重要，很多人都会回答：科技。如果要问科技界什么最重要，所有人都会回答：人工智能。对于人工智能，人们有着无尽的期望与想象，同时也难免恐惧与担忧。伴随着这些复杂的情愫，人工智能时代不知不觉已经降临。在这种大背景下，极客帮科技确定了“ All Around AI ”的内容战略，推出了 AI 前线专栏和《AI 生态》期刊。面对人工智能，我们无力改变科技的进程，但我们可以改变自己，以及我们下一代的知识结构。这正是我们推出《AI 生态》期刊的初衷。

1956 年，达特茅斯会议（Dartmouth Conference）奠定了人工智能的基础，经过 60 年厚积薄发的准备，人工智能终于可以奔跑了。是的，我们现在已经真真切切生活在人工智能时代了。这是最好的时代，也是最坏的时代。这个时代的伟大之处在于，它永远在更新，永远在前进；但是这个时代的悲哀之处在于，跟不上时代的人，可能永远就跟不上了。

鉴于此，2017 年 7 月 20 日，国务院正式印发《新一代人工智能发展规划》，集举国之力，将新一代人工智能提到了国家级的战略高度，首次提出“三步走”的战略规划。预计 2030 年中国将成为世界主要人工智能创新中心，人工智能核心产业规模超过 1 万亿元，带动相关产业规模超过 10 万亿元。

随着人工智能提升到国家战略高度，国家安全、经济繁荣、人口健康、生活质量和生态环境，都比以往任何时候更需要人工智能。人工智能将成为人类认知自然与社会，扩展智力，走向智慧生活的重要伴侣，引发了人人联网、物联网的崭新形态，也改变着人类的生产活动、经济活动和社会生活。这些，霍泰稳（极客邦科技创始人兼 CEO）在 2017 年 9 月 17 日第一期的刊首语都已经说得很清楚了，在此我不再赘述。

人工智能的每一点进步，是成千上万的科学家、程序员背后的推动，没有这推动力，人工智能就无法继续发展，比如，光纤宽带、移动宽带、移动互联网、云计算、大数据及物联网等，都推动了人工智能的进步。如果不是联网通信、数据量爆发及算力的极大提升，人工智能也许还在实验室“养在深闺人未识”。

最近几年来，人工智能的发展趋势是开始走出实验室，步入人类生活的方方面面，它们变得能用了，过去的许多技术承诺终于得以实现了。

不管怎么说，这是一个诱人的时代，我们正向它走去。AI 时代，星辰大海！



AI 前线

ID: ai-front

# 关注落地技术 探寻 AI 应用场景



关注AI前线公众号

InfoQ

# 助力人工智能落地

2018.01.13 – 01.14 · 北京国际会议中心

近年来，获得投资界助力的AI市场发展迅猛，人工智能正在渗透到各行各业。过了“尝鲜期”，大多数企业开始发力AI落地：

- (1) 如何用机器学习实现2亿月活跃用户?
- (2) 如何基于人工智能为用户精准推荐他们最喜欢的商品?
- (3) 如何通过深度学习网络结构使准确度提升 11% ?
- (4) 如何优化风控模型来解决智能金融带来的安全问题?
- (5) 如何通过机器学习等 AI 技术来提高运营效率?
- (6) 如何解决VR直播中高码率带来的“三高”问题?

为了帮助企业摆脱“落地难”的困扰，InfoQ中国为大家梳理了整个AI产业生态链，并瞄准全球顶尖AI落地案例策划了AICon全球人工智能技术大会。大会将精选30+国内外AI技术专家共享他们的落地痛点及填坑经验。

## 演讲嘉宾



颜水成  
360人工智能研究院  
院长及首席科学家



洪亮劫  
Etsy  
数据科学主管



张浩  
饿了么  
技术副总裁



尹大朏  
摩拜单车  
首席科学家



裴少芳  
iTutorGroup  
大数据部总监



张瑞  
知乎  
机器学习团队负责人



杨骥  
国美在线  
大数据中心副总监



胡时伟  
第四范式  
首席架构师



胡南炜  
微博  
机器学习计算和  
服务平台负责人



吴甘沙  
驭势科技  
联合创始人&CEO



陈伟  
搜狗  
语音交互技术中心  
研发总监



张重阳  
微信小程序  
商业技术高级研究员

8折

限时购票，立减720元

团 购 享 受 更 多 优 惠



# AI 前线

InfoQ 中文站 AI 月刊 2017 年 10 月

## 生态评论

6 视觉 AI 到底发展到了什么地步？

## 重磅访谈

16 技术大 V 老师木：软件平台是深度学习计算力突破的关键

## 落地实践

32 LinkedIn 的机器学习实践

43 爱奇艺视频场景下的自然语言处理应用

53 小红书里的秘密：机器学习如何帮助十人算法团队快速达成目标

70 分析海量视频中的违规内容，七牛如何构建弹性深度学习计算平台

83 YouTube 整合 Google Brain 推荐算法，视频播放量提升 20 倍

## 企业机器学习平台

89 Uber 的机器学习平台：从打车到外卖，一个平台如何服务数十个团队？

# 视觉 AI 到底发展到了什么地步？

作者 贾佳亚



AI 这个词从进入大家的视野到变得巷闻皆知才用了两年时间，所以 AI 在这个时间发展过程中有点像突然发现的东西，用什么来比喻它？我会用哈利波特的隐形斗篷来比喻它，这个隐星斗篷就是当你穿上它的时候，你会发现空无一人，但是你把隐形斗篷取下来，你发现原来里面躲了一个庞然大物。

其实 AI 视觉技术就这样一个过程，AI 的发展从最开始到现在经历了几十年的发展，所以到今天这个规模绝对不是一朝一夕能够形成的。所以我从隐性斗篷的例子来从头看看到底怎么理解 AI 技术。

我们在很多的小说、电影、科幻读物里都有很多拟人化的机器人或者



产物，其中有四个最重要的功能，第一是看，第二是听，第三是说，第四是动。当然不是所有东西都会动，但如果这是一个超级的智能产物一定会控制其他东西在动，自己不用动。当我今天想跟大家介绍 AI 的时候，我会专注在一个方向上，那就是看。为什么我们要去讲看这件事？我觉得还要从自然智能理解起。



自然智能不是 AI 智能的对立面，但是却是反方面的词。人工智能是人创造的，自然智能是从远古时代演化到现在的，我们从自然智能里学到很多东西，比如说看到自然智能的时候，我会想，我们有非常多的视网膜神经细胞，有柱状和椎装细胞但是我们有超过 40 亿以上的神经元会处理我们的视觉信息，相比之下，我们的触觉和听觉可能只有 8% 和 3% 的比例，这说明什么？说明我们这个世界太复杂了，当我们从第一天人类开始去理解这个世界的时候，我们就有足够多的神经元或者处理单元去理解这个世界，所以“看”是我们理解这个世界最重要的部分。

## 我们做到了什么？

视觉的 AI 可以运用在很多的游戏里面，比如说体感游戏或者是增强现实游戏，满大街去找小精灵的游戏就是重要的体现。除了视觉娱乐之外还有很大用处，比如解决在监控、安防或者需要大量人手去观察视频和图

像的分之内，我们从几百人减成几个人，这也是视觉 AI 发挥的作用。



腾讯是一个非常大的社交网络公司，里面有各种各样的媒体或者软件帮大家做交流，比如我有一个好的照片想给大家看看，是不是能够达到把人年轻十岁的效果呢？这个事情是可以做到了，甚至于如果想把自己变一个性别，从男生变成女生，那也很容易，甚至不用去医院了。这是在相册上或者是在手机端产生的变化，除此之外还有两块非常大的部分，一个是智能医疗，如何能够让一个机器智能读懂所有医疗的片，比如说 CT 片、MRI 片，这是非常重要的部分。还有自动驾驶，我们能不能辅助驾驶、自动驾驶的功能加入在视觉 AI 里面。

这些 AI 的技术代表在这个领域飞速发展的进程，但是与此同时，在不同的途径、不同的视频或者不同的专家给大家介绍各种方法的时候会说，我们的技术已经做到多么强、多么好，我在这里更希望通过科学家的角度跟大家介绍，我们的视觉 AI 角度到底发展到什么地步。

首先可以超过 1000 个类别的上亿张图像的分类理解。

当我有一张图像的时候，人和机器都可以告诉你这张图像是什么，这是一头牛还是一朵花，有的时候你可以想象机器甚至做得比人更出色，我三岁的女儿经常跟我说，爸爸，我看到那边有非常漂亮的蝴蝶。我就纠正她，宝贝，那不是蝴蝶，那是蛾子。但是我的宝贝说，这个蛾子比蝴蝶还漂亮，肯定是蝴蝶。

说明我们在图像理解上有一个过程，我需要理解它的含义得到一个结果，但是在机器学习的时候，甚至可以达到比成年人更高的境界，我们可以细分到山丘、山陵的区别，而超越人的理解。科学家已经不满足于这个问题，这个问题被认为已经在这个领域解决，下一个要解决的是检测问题。

当我们有一张图，我希望不但知道这个图的整体表达是什么，还要知道这个图里哪个地方是车，哪个地方是路面，哪个地方是人，这是检测过程。由于现在有强大的计算资源和计算能力，我们可以超过五亿个品种的检测，这是视觉 AI 的另外一个可以达到的目标。



除此之外，科学家们想，当我们能检测到一些物体的时候，能不能把细致度做得更深？比如说颗粒度更深的每个像素、每个点，我是不是能知道这个点是属于马路的，属于人还是属于车的，这是远远超越于之前问题的更加进一步的推广问题。所以我们管它叫做语义分割，现在可以超过总数四千亿像素级别的多图图像分割，这是这几年整个领域产生的巨大推进作用和研究成果，能够达到的效果。除此之外更加熟悉的是对人脸的匹配



查询，可以超过一亿张人脸匹配查询，找到你想要的人，你问问自己，能不能认识一亿个人？认识一百个人，我就很开心了，这在电脑上是远远超越了人。

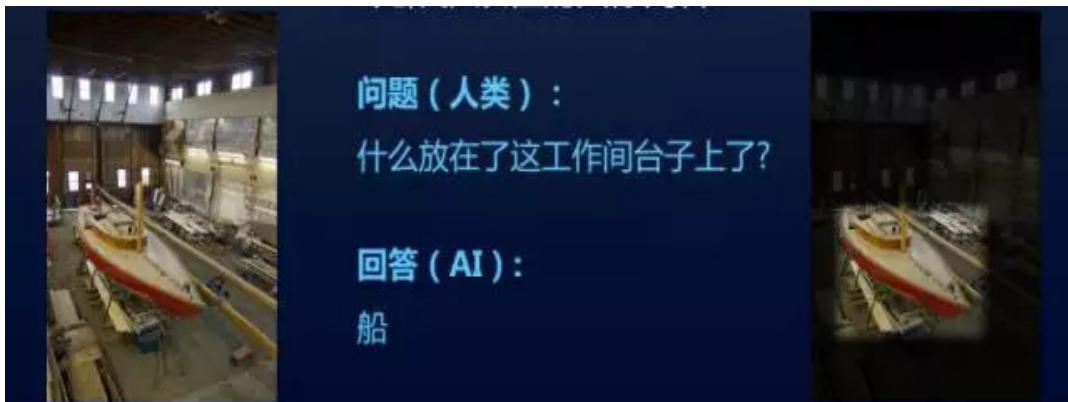
在之前五到十年的时候，我在学校的团队还会做一些有意思的研究：我们当时想，如果看到这样一张模糊的图片你会做什么事情？你看到一张模糊图像会做什么事情？可能大家要做的就是把它删除，为什么？因为这张图片模糊了，已经没有用了，但是对我们科学家而言，是非常珍贵的资源，因为通过这张图像我们发掘出一些人类看不到的东西。



为大家举一个有趣的例子，比如在这张图像里，我们是在一个高速行驶的车上，周围的环境是容易被模糊的，当我们理解环境的时候你发现，车牌或者路标已经被模糊掉，在这张图上，在这个图标上看着公路的信息，但是左边小的是什么东西？左边路牌上的数字是什么东西呢？之前大学里的团队经过五到十年的研究，把这些信息充分理解出来，最后通过我们的技术手段，能够看到最后这是在美国 101 公路上的场景，我们可以超越人类的图像模糊。

还有一个东西，我们希望十年之后出现家具机器人，什么叫家具机器人？就是你希望他能帮你洗衣服、洗碗、做饭，甚至带孩子，但是要达到这个智能机器人，怎么样才能做到这一步？其中重要的就是如何把自然语言和自然图像结合起来，也就是两者的充分结合。所以我们在之前有一系列的研究，是当你看到一张图像的时候，我的人来问一个问题，是什么放在了这个工作间的台子上，电脑看到通过你这句话，分析是什么意思，然后再去寻找在这张图像上是问了什么问题，最后把图像上重要的位置找到

之后反馈回来，得到一个结果，这个结果就是船，这就是说这个答案是对的。



这个说明现在电脑可以结合自然语言，我说的话和看到的场景可以结合起来，这是一个非常了不起的进步。正是因为有这个进步，我相信令到家具机器人的理想在十年之内能够变为现实。

## 我们还可以做到什么？

可能有讲座介绍说，幸亏 AI 达到的程度还没有人那么高，因为人会创造，AI 不会创造。我想跟大家说，其实这句话是不对的，电脑也会创造，而且创造出来的种类和试样，很多时候是让我们惊叹的，在以往知识库里是找不到这些模组的，这就是创造的功能，视觉 AI 已经可以实现创造

比如看这两幅图，看在座各位认为右边这张图是电脑画的？真理永远



掌握在少数人手里，右边这张图确实是电脑画的，电脑用铅笔画出了整体轮廓的表达、阴影的表达，达到了非常高的层次，人类需要长期训练才能画出，但是我们画出这张图只用了 0.1 秒，就是它的创造过程，这是非常有趣的事情。

我大学的团队在去年为了去理解这个非常复杂的场景，创造了全世界最像素级的分割技术，做场景理解分析。



这个例子可以看到车是一个颜色，因为我识别出来这是车，所以是蓝色，旁边的树，我识别出是树，所以标成绿色。我们在去年实现了在大规模场景的多复杂环境下的内容分析，这是去年做的事情。今年我们不满足这样一个结果，我说我们还能做什么？然后我的团队开始在今年做了另外一件让大家激动的事情，我们实现了到迄今为止最准确的道路上的像素级语义分割技术，在已知的论文里面，我们这个技术远远高出第二名，直接到每秒 30 倍的运算速度，没有改变任何的硬件资源，我们加速了一百倍。

## AI 视觉技术的社会价值

除了我刚才跟大家讲的商业价值或者学术价值之外，我今天还有另外一个内容。我想跟大家讲讲优图 AI 所产生的社会价值。有一个优图跟腾讯的公益部门以及腾讯云、腾讯互联网 + 的部门合作，我们开创了一个新的活动，这个活动叫做天眼。

大家有没有看过这部《亲爱的》？讲述的就是现在有很多很多的家庭，



家里的孩子被拐卖走失，这些父母组成了一个团体，他们希望通过这个团体找寻自己的孩子，这就是 2014 年这部电影上映反映的严重现实，孩子的丢失也许是社会的一小部分现象，但是这个现象确实是存在的。所以可以想像，在这样一个环境下，可以有一个大概的估计，但是现在走失的现象在统计意义上而言非常巨大，每个城市这样的现象很少，但是因为中国人口多，在这么大的中国人口的基数上，我们能够把这个比例一点点的上升，这也是我们能贡献的力量。虽然我们有了各种各样的途径，有微博打拐，有大家贡献的力量来找寻，但是贡献率依然是 0。

2015 年，优图团队跟腾讯公益部门和腾讯云和腾讯互联网 + 的单位一起开始加入了“天眼”计划，希望通过优图的技术积累，帮助社会实现社会价值，而不仅仅是商业价值，因为这对我们而言是一件相对比较容易的事情，但是社会价值如何体现在 AI 上？

于是我们加入了这个公益计划，当时在上海的小伙伴们拿到这个计划的时候非常兴奋，他们觉得终于有一天，他们坐在电脑前面也可以像蜘蛛侠一样出去救人。他们做了很多评测，发现我们在人脸识别率上只有 40%，这个数字让我们的小伙伴非常惊讶，发现原来问题这么难，不是我们拿到一张图做一个寻人就可以把人找回来。主要有以下三点困难：

## 第一个是场景

我们有非常复杂的场景，有城市，有农村，有山林，有不同区域，甚至当我找回这样一个失踪人口的时候，他的发型、衣着、轮廓改变都是复杂因素。

## 第二个是年龄

很多的失踪人口找回来的时候，他们可能在外面已经漂流了几年的时间甚至十年时间，这是一个非常长的时间段，所以从我们的面容上看，他们改变了很多，从轮廓、皱纹、皮肤的粗糙程度，这都是对我们实用算法是非常大的挑战。

### 第三需要具备有亿级人脸的检索能力

当我拿到一张检索照片的时候，是不是能够通过实时寻找对比，找到这个人出来。

这三大挑战是我当时面对的，但是好在我们优图的小伙伴们并没有放弃，他们觉得这件事情既然做了就要做到底，而且要做好。所以我们在经历一系列的，超过一年多的研究，把 Megaface 的准确率从 40% 提升到 83.29%，这代表了我们可以在亿级人脸检索上达到毫秒的速度，把成功率从之前的不到 50% 提高到 99%。我们立项之后，在短短三个月时间之内，就开始用在福建省公安一起合作，做了网上在线系统找回人群。

2017 年 3 月份的时候，福建省公安厅接到一个群众电话，他们在小学边上找到一个老奶奶，神智不清语言也不通，把他接到公安局以后，通过我们的线上人脸比对系统，发现可能是这个奶奶失踪了，最后我们发现，家人为了防止她走失，在公安系统已经把她挂上号，最后通过这个系统把这个老人找了回来。



我们上线这个系统短短三个月时间，在整个福建省公安部门的帮助下，实现了找回人数超过 120 人，才三个月时间。这样的成效是高过以往通过群众电话，再去通过大海捞针式的访问拿到结果的过程，所以整个福建系统“牵挂你”是有一个过程的，我们发上名单照片，通过群众找到某一个个人群的时候，拍张照片，最后在数据库里做比对，然后把这个找回来，这样成功的案例已经超过一百起，这是非常振奋人心的，而且也是很有意义的。

除此之外，我们希望技术不仅仅是帮到这样一些走失的人，我们甚至可以走得更广一点，所以我们这几年做了一个“万象鉴黄”的全球儿童网络保护行动，整个优图团队开始贡献对成人图片的检测，我们发现准确率在大部分上线系统上可以超过 99%，也就是可以实现对儿童在网络上的保护，防止这些儿童受到欺凌欺骗，这样的事情是我们团队的小伙伴最愿意做的事情之一。

## 结语

AI 就是一种工具，AI 的出现可能会令一些人失去自己的工作，但是 AI 确实便利了我们这个社会，使得我们这个社会更加容易和谐，做得更好。当我们发现一些不好的事实的时候，AI 这个系统能够准确判别，打击犯罪，这是一种工具，就像是一把刀一样，你切菜是好的工具，但是伤害人的时候是一个坏的工具。AI 无所谓好和坏，但是好的部分需要我们去弘扬，发光广大。最后我想说，每一个技术人员虽然都坐在电脑前面，大家都认为我们是电脑高手，但是我们每个人都有一颗蜘蛛侠的心，我们希望自己有一天不上街也能够帮助人们，打击犯罪。

# 技术大 V 老师木：软件平台是深度学习计算力突破的关键

作者 吴少杰 Tina 陈思



在知乎上，“老师木的机器学习水平怎么样”的问题，被浏览了 3.7 万次。虽然关注者众，却不少评论他线下实际“为人低调”。

同时有人称他是“微博大 V 老师木”，没错，作为一个技术人，他的微博粉丝有 5.8 万。老师木说这个影响力“还太小”。

有人关心他为什么做得好好的要从微软亚洲研究院离职，更有人关心老师木为啥要创业。

老师木，真名袁进辉。读书时成绩一直优异，本科后保送清华大学直博生，师从人工智能领域张钹院士。期间多篇论文在国际顶级会议上发表，在竞争激烈的国际技术评测（TRECVID）中连续多年名列第一。博士后出

站后，于 2011 年入职网易有道。2012 年作为早期成员加入 360 搜索创业团队，一年之后，产品上线成为国内市场份额第二的搜索引擎。2013 年，加入微软亚洲研究院（MSRA），主要从事大规模机器学习平台的研发工作。

从博士后到第一次创业，是从学术研究人员转型成为工程师；进微软亚洲研究院，则又重回到了学术道路。在 MSRA 期间，专注于研发大规模机器学习平台，以出色的科研和工程综合能力，发明了世界上最快主题模型算法 LightLDA 及分布式训练系统：只用几十台服务器就能完成之前需要数千服务器才能完成的训练任务。“LightLDA 的确是迄今为止，我做出来最有影响力的工作。人常说，评价一个学者水平高低不是看成果多少，而是看他能到达的最高水平，可以说这项研究让我跻身于世界一流研究人员的行列”。

MSRA 被称作中国 IT 届的黄埔军校，精英荟萃，并且老师木的成就也开始受到各界的认可，但是他却出人意料的放弃了 MSRA 的优厚工作，走上了创业的路途，更是参与到深度学习框架这种战略级产品竞争中。众所周知，很多大公司都出有自己的深度学习框架，Google 的 TensorFlow，微软 CNTK，Amazon 的 MxNet，Facebook 的 Caffe2 等，并且都在努力的建立生态。以老师木的视角，他是如何看待这些框架？我们从深度学习技术和框架、LightLDA 两大方向和老师木进行了一次深度访谈。

## 关于创业

### InfoQ：为什么你放弃 MSRA 的优厚工作去创业？

**老师木：**创业者一定有一个或大或小的愿景，或者说使命感，未来的世界应该是什么样的，怎么努力促使愿景实现。我的愿景是：人工智能技术赋能各行各业，推动人们工作效率和生活质量更高，把人类从机器擅长的工作中解放出来，让人类去做更需要创造力的事。在这种使命驱动下，首先选择做什么事最有利于这个愿景实现，其次选择做事的形式。要选能突破自我，能最大化创造价值的事和形式。

### InfoQ：怎么看待人工智能的市场潜力？

**老师木：**首先，互联网行业已经充分验证了数据驱动的业务模式。其次，互联网行业之外的存量业务有显著的人工智能技术红利可吃，或者刚刚尝到人工智能技术的甜头，或者是尚未开垦的处女地，仅仅把人工智能技术引入已有业务，就能获得竞争优势，甚至带来质的飞跃。最后，人工智能技术革命会催生一些新的产业，譬如自动驾驶，精准医疗等。据此，有人认为这次由深度学习引发的大潮可能是第三次工业革命。

### InfoQ：深度学习在业界有哪些靠谱应用？

**老师木：**每个高商业价值的互联网应用背后都有深度学习的身影，搜索引擎，广告，推荐引擎，用户画像，社交媒体，共享经济等等。人类智能可概括为感知，决策和控制三方面，有监督深度学习方法最先在感知类型的任务（图像视频，语音，语言的理解）中取得成功，譬如安防，医学影像，色情信息过滤，语音助手，机器翻译等都已经商用落地。强化学习在决策和控制方面也取得很多成果，主要是机器人自动控制，自动驾驶，处在快速发展中。

### InfoQ：深度学习在技术上存在什么瓶颈？最可能在哪里获得突破？

**老师木：**先分别说有哪些关键问题。在算法和理论方面，目前有监督学习应用最成功，各行各业积累了大量的无标签的数据，怎么利用上无标签或弱标签的数据？深度学习在感知（Perception）类型的任务上非常成功，怎么与认知（Cognition）方法（符号推理）结合形成最终决策？在理论上如何理解深度学习这么惊人的效果，怎么在理论指导下设计模型，而不是靠 ad-hoc 经验试？在计算效率方面，服务器端主要考虑扩展性，怎么能让一批高吞吐协处理器协同解决一个大型任务时总体利用率最高，在终端上则主要是考虑低功耗实现，能否同时实现易用性和高效性。在应用方面，主要是在一些高商业价值的问题上能否从技术上打通达到可用程度，AlphaGo 非常成功，但商业价值还不明确，在杀手级应用如自动驾驶，精准医疗，自动化交易等方向上取得成功，更值得期待。

理论和算法研究上的突破通常可遇不可求，更难预测，而且是否真的

突破最终也要落实到实际应用中去评判。在计算力和应用上的突破确定性更高一些。我们是瞄准了计算力这个方向的商机，一会儿可以深入探讨下这方面的问题。某些垂直应用如自动驾驶方向聚集了大量资金和人才，这方面的突破希望也很大。

### InfoQ：为什么计算力会成为深度学习的一个突破方向？

**老师木：**首先，计算力是极其关键的一项支撑技术。最近发生的人工智能革命通常被认为是三驾马车驱动，数据，算法和计算力。与上世纪九十年代相比，深度学习在算法原理上并无二致，在数据和计算力方面进步更大，各行各业积累了大量的优质数据，GPU 作为新的计算手段引爆了此次深度学习的热潮。

其次，计算力方面还有现成的红利可吃，相同的算法，如果能用上更多的数据，或者用更大规模的模型，通常能带来效果的显著提升，能不能做的更大取决于计算力的水平。

再次，算法和原理的研究进展依赖于计算能力，好的计算力平台可以提高算法和原理研究的迭代速度，一天能实验一个新想法就比一星期才能实验一个新想法快的多。有些理论问题本身是一个大规模计算问题，譬如神经网络结构的自动学习等价于在一个超大规模假设空间的搜索问题，没有强大计算力的支持就只能停留在玩具数据上。深度学习是受生物神经网络启发而设计出来的，现在人工神经网络的规模还远远小于人脑神经网络的规模，人脑有上千亿神经元细胞，每个神经元平均有成千上万的连接。

最后，如何在低功耗约束下完成高通量的计算也是制约了深度学习在更多终端上应用的一大因素。

### InfoQ：计算力具有什么样的商业价值？

**老师木：**一方面，计算力的商业价值体现在它是数据驱动型公司的大部头营业支出（硬件采购，人力成本等）。数据驱动型业务的完整链条包括数据收集，预处理，深度分析和在线预测，无论是私有部署还是上公有云，建设高扩展性的基础设施等支撑技术，都是一笔不可忽视的开销。另一方面，计算力也是数据驱动型公司获得竞争优势的关键，人工智能可提

高公司业务效率，而计算力又可提高人工智能的效率。目前，围绕着计算力已经出现了诸多成功的商业模式，譬如公有云，面向私有部署的商业技术服务，深度学习加速器（GPU, DPU）等。

### InfoQ：计算力在技术上有哪些瓶颈？

**老师木：**从硬件看，我们现在使用的都是冯诺依曼结构的计算机，它的主要特点是计算单元和存储单元分离，主要瓶颈表现在摩尔定律（Moore's law）的失效和内存墙（Memory wall）问题上。克服摩尔定律的主要途径是增加中央处理器上集成的核心（core）数量，从单核，多核发展到现在众核架构（GPU, Intel Xeon Phi），但芯片的面积及功耗限制了人们不可能在一个处理器上集成无穷无尽个核心。内存墙的问题是指内存性能的提升速度还赶不上CPU性能的提升速度，访存带宽常常限制了CPU性能的发挥。纯从硬件角度解决这些瓶颈问题，一方面要靠硬件制造工艺本身的发展，另一方面可能要靠新型的计算机体系结构来解决，譬如计算和存储一体化的非冯诺依曼结构计算机。除了高通量的计算，在电池技术没有大的突破的前提下，终端应用场景（物联网，边缘计算）里低功耗也是计算力的一项重要指标。当前，深度学习专用硬件创业如火如荼，有可能会被忽视的一点是：对突破计算力瓶颈，软件至少和硬件一样关键。

### InfoQ：为什么软件会成为计算力突破的关键？

**老师木：**计算力的基础设施要满足上层用户对易用性，高效率，扩展性的综合需求，仅有硬件是不够的。一方面，数据科学家和算法研究员不像系统研发工程师那样深刻立刻硬件的工作机理，不擅长开发释放硬件计算潜能的软件，对数据科学家最友好的界面是声明式编程，他们只需要告诉计算力平台他们想做什么，具体怎样算的快要由软件工具链来解决。另一方面，尽管单个众核架构的协处理设备（如GPU）吞吐率已远超CPU，但出于芯片面积 / 功耗等物理限制，任何一个单独的设备都无法足够大到处理工业级规模的数据和模型，仍需由多个高速互联的设备协同才能完成大规模任务。出于灵活性需求，设备之间的依赖必定由软件定义和管理，软件怎样协调硬件才能提高硬件利用率和释放硬件潜能极具挑战，至关重

要。在相关领域，软件定义硬件已是大势所趋：上层软件决定底层硬件的发展方向，底层硬件要取得成功离不开完善的上层软件生态。

### InfoQ：业界已经有很多软件平台，为什么要再打造一个？

**老师木：**用户选择众多，但仍有重要需求未被满足，深度学习框架技术演化仍未收敛。深度学习框架一定会出现 Hadoop 那样具有市场支配地位的产品，也就是所谓的事事实工业标准，而现在还没有任何一个软件平台达到这种地位。工业标准级的平台不仅要解决眼前的需求，更要面向未来。现在的确有一些知名的软件平台，但业界还有相当一部分重要需求没有被满足。比如，现有技术方案对于单设备或多设备数据并行这种简单场景的支持已经非常优秀，但在模型更大或者神经网络拓扑更复杂时，通用框架的易用性和效率都大打折扣，有这种需求的工业级应用只好去用定制的 HPC 方案（譬如百度的 DeepSpeech）。问题的根源是，设备之间互联带宽远低于设备内访存带宽，这是和传统 CPU 上内存墙（Memory Wall）类似的难题。我们团队经过艰苦卓绝的努力，探索一条走向通用解决方案的技术路径。沿这个思路开发的软件平台，有望既享受软件的灵活和便利，又享有专用硬件的高效性。我们坚信，通用的解决方案是深度学习平台技术收敛的方向，只有这种通用的解决方案才是深度学习平台的最终形态。

### InfoQ：能说说你们产品的主要技术特点是什么吗？

**老师木：**深度神经网络和人脑信息处理本质数据流计算，信号的传播即计算，然而当前主流的底层硬件都是冯诺依曼结构。纯硬件实现的数据流计算机还不现实，现在必须依赖深度学习软件平台来完成这样一个翻译或者映射的过程：从数据流表达式到冯诺依曼结构上的指令序列。软件平台最终价值体现在易用性和高效性。易用性，要支持用户能够使用最自然的表达方式来描述各种神经网络计算的需求；高效性，对所支持的任何一种上层需求，都能基于通用硬件资源表现出专用硬件的那种效率。我们的产品开创了一种和现有深度学习框架截然不同的技术路线，细节上表现出来静态编译，全链路异步，去中心化，流式计算等特点，我们认为这是深度学习基础架构实现易用和高效的必由之路，是深度学习框架技术收敛的方

向。

### InfoQ：长江后浪推前浪，这样一个先进的技术架构生命力会有多久？

**老师木：**首先，我们可以探讨一下深度学习的范式还有多久生命力，毕竟技术架构应需求而生。可以从这几方面看：从数据流计算模型是生物体采用的信息处理机制，是人工智能的效仿对象；人工神经网络已经在多个领域取得成功，而且深度学习本质上还是统计学习理论，利用算法在数据中挖掘统计规律性，这种学习机制的本质不会变化；深度学习算法便于利用并行硬件的威力，算法和硬件的天作之合，还看不出取代它的必要。其次，从计算机体系结构及硬件演化方向上看，软硬件结合的数据流计算机代表着突破摩尔定律和内存墙限制的方向。

### InfoQ：是不是只有大公司才需要这样的基础设施？

**老师木：**并不是。目力所及，这样的基础设施已经不是大公司的独享的专利，拥有数十台服务器的中小企业，大学研究院所比比皆是。数据驱动是一种先进的生产力，所有行业最终都会变成数据驱动，每个行业的每个公司的数据都在积累，每个公司对数据分析的需求都在进化，从浅层的分析到深度分析，这个大趋势呼之欲出不可逆转。十年前，会有多少公司需要 Hadoop，现今几乎所有的公司都要用到 Hadoop。历史一再证明，无论计算能力发展到多强大，应用总能把它用满。多年以前，有人还觉得 640K 内存对于任何人来说都足够了，今天 64G 的内存都开始捉襟见肘，一辆自动驾驶测试车每天收集的数据达数 TB 之多。从来不是强大的计算力有没有用的问题，而是计算力够不够用的问题。

### InfoQ：深度学习框架竞争很激烈，而且看上去都是业界巨头在玩。

**老师木：**是的。一个深度学习框架一旦像 Hadoop 那样成为事实工业标准，就占据了人工智能各种关键应用的入口，对各类垂直应用，基于私有部署的技术服务，公有云上的 AI 即服务业务，甚至底层专用硬件市场都有举足轻重的影响。它的角色就像互联网时代的浏览器，移动互联网时代的安卓操作系统一样，是战略级产品，业界巨头谁都不想让给他人也就不奇怪了。目前，大公司出品的比较知名的框架有 Google

的 TensorFlow, 微软 CNTK, Amazon 的 MxNet, Facebook 的 Caffe2, PyTorch, 国内百度的 PaddlePaddle 等。

### InfoQ：为什么用创业的方式做这样一件事？

**老师木：**这种事既有技术攻关上的挑战，也有资源组织上的挑战。这就需要科研院所那种人才密集度，又需要公司的组织支持。我既有在大公司工作的经历，也有两次创业的经历，个人理解，创业是社会资源组织和分配的一种优秀机制，能最大化这项事业的成功率。首先，创业是社会鼓励创新和承担风险的一种资源分配形式，有潜力的创业团队能得到所需要的资源（资金和人才），同时有高度灵活的机制，在大公司，未必是最适合做这项事业的人来承担这样的项目。其次，一项充满挑战的事业需要具有聪明才智的人以持久的热情投入其中，创业公司那种公平合理的利益分配机制才能最大激发成员的主观能动性，为业界做出实质贡献的人也应该得到回报。

### InfoQ：创业公司做这样一件事看上去很不可思议。

**老师木：**有很多大公司加入这场竞争，说明存在真实的需求，而且市场容量足够大，看上去创业公司做这样的产品非常难，实际上大公司做也是同样地难。深度学习框架的用户是开发者 (developer)，也就是说的 To developer，要把这样一件产品做成功，被业界广为采用，关键看两点：

首先，这种深度学习框架是技术密集型产品，一定要做到最广泛的满足实际需求，而且在某些方面要有不可替代的优势，有突出的长板。

其次，要形成生态，具有完善的社区支持，做到没有明显的短板。一个组织只要具备实现这两点目标的要素，就有机会，而不在于那是小公司，还是大公司。

事实上，在开源软件范围竞争还是非常公平的，原来名不见经传的人开发出的软件的确好用就能火，大公司开发出的软件质量不行也没人用，最终靠产品质量说话。现在，创业公司聚集了业界最优秀的一批人，聪明，更重要的是有野心（进取心）。当然，对创业公司来说，不仅要取得产品的成功，还要取得商业上的成功，让所有参与这项事业的人拿到现实的回

报，公司自身也获得更充足的资金支持投入再生产，做出更优秀的产品。大公司在开源产品的商业化上更从容一些。个人观点，很多大公司与你竞争不可怕，更可怕的是面对很多创业公司的竞争。最终结果取决于产品质量。

**InfoQ：如何取得商业上的成功？只有好的技术也可能赚不了钱。**

**老师木：**取得商业上的成功是创业公司的最终追求，我们也一样。我的理解，这涉及两个“价值”问题。

第一，我们在做的事是否为用户创造了价值，我坚决信奉 `create value, money follows;`

第二个是回归商业价值，在为用户创造价值的前提下，我们需要探索出一条双赢的利益分配机制，把用户转化成客户。

现阶段，我们聚焦在解决第一个问题，打造出解决用户需求和痛点的产品：深度学习平台，不贪大求全，只追求把整个链条中的那最关键一环打造到极致。这是我们这个团队在人工智能大潮中参与顶端竞争的切入点，在我们眼里是那个撬动地球的杠杆支点。从为用户创造价值这个角度切入能最大化实现商业目标的成功率，而且有可能把我们推举到比其它选项要高的多的高度。微软，谷歌，英伟达，甲骨文，华为这样伟大的公司都是因为有了创新的产品才形成了伟大的商业公司。我们对商业模式的各种选项都持 `open` 态度，不排斥和高商业价值的垂直场景结合。

**InfoQ：您们的深度学习平台第一版预计什么时候公测？需要从哪些方面准备？**

**老师木：**系统主体开发已经完成，目前处在内测阶段，计划年底时开源。开源之前需要从以下方面做充分准备：第一，产品功能完整性，要支持主流的深度学习模型，譬如 CNN/RNN/LSTM，支持图像，语音和语言经典应用；第二，验证高效性，在业界公认的大规模评测中表现出效率优势，给出具体技术指标，如在多大规模上跑到什么水平的加速比，设备利用率等等；第三，打磨易用性，和上下游工具，和已有深度学习框架的兼容性，以及文档建设等等。我们团队先从技术方面打好一个底子，当用户想为这个项

目做贡献时，可以更容易加入进来。

**InfoQ：您们研发深度学习平台会兼容哪些芯片？支持什么操作系统，支持 Linux, Windows, Android 和 iOS 吗？**

**老师木：**目前我们聚焦在服务端的训练场景，在这种场景下，GPU 是最经济的选择，所以目前只支持纯 CPU 或 CPU+GPU 的异构集群，如果未来硬件市场发生变化，我们也可以支持其它芯片。服务器上主要操作系统是 Linux 和 Windows，所以目前只支持这两种。终端的应用场景主要是在线推理（inference），我们团队目前没有投入。

## 关于 LightLDA

**InfoQ：LightLDA 是您的代表作之一么？能给大家介绍下这个项目的一些情况么？**

**老师木：**LightLDA 的确是迄今为止我做出来最有影响力的工作。人常说，评价一个学者水平高低不是看成果多少，而是看他能到达的最高水平，可以说这项研究让我跻身于世界一流研究人员的行列。首先，算法结果是一流的，LightLDA 是当时业界最快的训练 Latent Dirichlet Allocation (LDA) 主题模型的算法，它把单个词采样降低到  $O(1)$  复杂度。其次，系统实现是一流的，我们仅用数十台服务器，完成之前成千上万台服务器才能做的事。LightLDA 和许多其它优秀科研成果一样，是集体努力的结晶。那个时候，CMU 的邢波教授 (Eric Xing) 在 MSRA 任顾问，微软团队和他领衔的 Petuum 团队合作达成此项成果，论文发表在 WWW 2015，系统代码在 Github 开源，也成功应用于微软搜索广告和情景广告产品中。

主题模型特别是 LDA 是广告系统和推荐系统中的关键组件，据说“Google AdSense 背后广告相关性计算的头号秘密武器 Google Rephile”就是一个巨大规模的主题模型。大约三四年前，微软很多产品想用类似的技术，然而并没有大规模主题模型的训练系统。有一天，主管这个领域的副院长马维英（现今日头条副总裁）和我讨论时，说起这件事，产品

部门经常问他的团队有没有这样的解决方案，问我愿不愿意干。恰好那时邢波教授也开始做 MSRA 的顾问，邢教授的团队在这方面有很积累，微软正好可以和他在 CMU 的团队合作研发大规模主题模型训练技术，双方一拍即合。当时，从公开渠道能了解到，为解决工业级需求，训练数据可能涵盖数亿个文档，每个文档包含十几到数百个词，为了覆盖长尾词和长尾语义，词典可能包含数十万到百万个单词，主题个数远超业界发表论文的数字（仅数百个主题），达到万，十万，甚至百万，最先进的解决方案需要数千台服务器运行数天才能得到结果。我们当时立下的 flag 是，相对于业界最好解决方案，做到各个维度上都有数量级的超越（服务器数量必须是数十台，我们那时拿不到数千台这么奢侈的硬件支持，数据规模做到数十亿 Bing 索引的主流网页，词典和主题数至少做到十万级别）。稍微推算一下，就可以知道，即使是当时最先进的算法 SparseLDA，在给定的硬件环境中训练这样规模的模型需要半年到一年的时间。再加上身处研究部门，一没有可供使用的集群，二没有工程师团队的支持，微软这边全时投入的只有我和实习生高飞，这个目标看上去是 mission impossible。我当时的想法是，最低目标要做出来一个能满足产品部门需求可用的主题模型，能不能做出打破纪录，就看运气了。

### InfoQ：请问大规模训练 LDA 模型的瓶颈是什么？

**老师木：**训练 LDA 的算法可以分成两类，一类是变分贝叶斯法，一类是 Gibbs 采样算法。前者计算过程和中间表示都是稠密的，分布式实现时通信量较大，后者是稀疏计算，通信量小，一般大规模主题模型都基于 Gibbs 采样算法实现。使用 Gibbs 采样算法时，算法复杂度和系统实现两方面都有困难。假设有 100 亿文档，平均每个文档有 100 个词，一共有 10000 亿个词，训练过程迭代 100 次，那就需要对 10000 亿个词扫描 100 遍。标准的 Collapsed Gibbs 采样算法处理一个词的计算复杂度与模型的主题数量有关，假设要训练包含 10 万个主题的模型，那么每个词就包含 10 万次计算，主频为 2GHz 的 CPU 核心每秒能处理 1000 个词，这样估算一下下来，假设使用一个单线程程序来做这件事，共需要 1000 亿秒，

也就是 100 万天。使用 10000 个 CPU 核心的分布式集群去训练，假设线性扩展性，也需要 100 天之久。假如每个词的采样效率能提高 100 倍，那么使用 10000 个 CPU 核心的集群去训练这个模型就只需要 1 天。前人已经提出了 Gibbs 采样算法的多种改进，譬如 SparseLDA, AliasLDA，但这些算法的单个词的计算复杂度仍与模型的主题数量相关，与“创造奇迹”仍有距离。另外，实践上，算法中总有一些步骤是无法并行化，受制于阿姆达尔法则，分布式系统很难做到线性加速比，所需要的时间会比上述预估的时间更长。

### InfoQ：LightLDA 设计之处，面临了哪些挑战？

**老师木：**我们 LightLDA 团队资源匮乏（计算资源，工程师资源），同时在算法和系统实现上都挑战极大。我个人认为最大挑战在信心方面：我们能不能做到？在此之前，有多位知名科学家和资深工程师在训练大规模 LDA 的问题上耕耘已久，他们已经把算法和系统实现推进到相当的高度，即使采用当时最先进的技术，仍不可能实现我们的目标。必须做出显著超越前人的奇迹技术突破才有可能实现目标。我和学生都是第一次从事大规模机器学习的项目，名不见经传，何德何能，能比另外一些特别牛逼的人物做的还要好？

首先是算法上的突破。我在重现和把玩 SparseLDA 和 AliasLDA 时，被可遇不可求的灵感眷顾：解耦 Gibbs 采样中与词自身相关的因素和词所在文档上下文的因素这两个因子，能做到单个词采样复杂度与主题个数无关。马维英院长第一次听我介绍完这个想法和初步实现结果后说 *too good to be true*，的确，谁能想到这样一个小小的 insight，竟然能把单个词采样复杂度降到  $O(1)$ ，理论上使得达成那个宏伟的目标成为可能。这个灵感来的偶然又必然，机遇偏爱有准备的人。我动手能力比较突出，很快就重现了 SparseLDA 和当时刚刚在 KDD 上发表并获得最佳论文奖的 AliasLDA 算法，同时理论功底又比较扎实，很快就深刻理解了它们的关键所在。我不断把玩这两个算法，在直觉和理论分析指引下做一些改动，然后观察是否有效，终于在一次改动后发现计算效率陡升，让人怀疑是不

是出现了有益处的 bug，再三推敲后终于确认，这是一个有深刻内涵的新发现。这又一次印证了我从清华数学系林元烈老师那学到的一个诀窍：熟能生巧。他的随机课程巨难无比，我刚开始怎么都入不了门，和很多自认佼佼者的同学一样竟然期中考试不及格。林老师说了一番这样的道理：他认识很多大牛数学家，即使是像他们那么聪明的人，在掌握一些艰深的数学科目时，也是通过做特别多习题才能悟道。我就硬着头皮做了很多习题，有的证明看不懂，甚至都背下来了，也是突然一瞬就知道了随机过程怎么回事。每次遇到困难，在说放弃之前再坚持一会儿结果就会不同。

找到理论上性质很好的算法，只是万里长征第一步。怎么高效地用程序实现，特别是在分布式环境下接近线性加速，包含了一系列的技术挑战，任何一个环节掉链子，所有努力都会化成泡影。做这类事的特点就是，兵来将挡，水来土掩，在你不知道前人这些技巧时，你要自己发明出来，但在系统领域极大概率是这个发明已经在经典文献中被提出过了。我们解决了两个突出的难题，超大规模模型的内存瓶颈和通信瓶颈。100 万的词典和 100 万个主题，模型之大，前所未有，意味着需要若干 TB 的内存，如何存储和支持快速访问也极其严峻。在分布式环境下，如何有效掩盖通信开销又不损失模型精度，也是当时面临的一个主要难题。我的学生高飞在工程实现方面特别给力，交给他的事情总能又快又好的做完。事后回顾这段经历，他说，这段日子是他最愉快的经历之一，偶尔会感到绝望，总发现我在前面仍激情满满的坚持，他深感佩服。我的领导马维英和刘铁岩研究员则克服重重困难，为这个项目提供资源支持和高屋建瓴的指导。同时，我们和 CMU Petuum 团队，Eric Xing, David Dai, Jinliang Wei, Qirong Ho，尽管身处太平洋两岸，但几乎每天都有邮件讨论，每周都有好几次电话会议，遇到技术难题大家凑在一次分析，提出不成熟的好点子又立刻能得到挑战，共鸣和支持，缺少任何一个人，结果都不是大家看到的样子，这就是一个优秀团队的魅力所在。

没有前面技术突破，绝不可能达到目标。仅仅有前面的算法突破，没有执行成功，这项研究也就是一个微不足道的 trick，绝不可能产生后来

的影响。

### InfoQ：LightLDA 如何借助 DMTK 框架做并行化？LightLDA 有哪些优点？

**老师木：**这里可能有一个小小的误解。在 Github 上发布时，LightLDA 是作为 Distributed Machine Learning Toolkit (DMTK) 的一个组件发布的，但实际上 LightLDA 最初是使用 Petuum 的参数服务器实现并行化。在 LightLDA 论文发表后，微软酝酿和发布了 DMTK 项目，这时候把 LightLDA 作为 DMTK 的一个主要应用集成进去了。LightLDA 的优点就不多说了，主要是快，扩展性好，用少得多的硬件资源就可以解决规模大的多的问题。我来说一下开源版本的缺憾吧。首先，理论上单个词采样复杂度是  $O(1)$ ，在工程实现上，因为随机访存造成 cache miss 太多的原因，没有完全发挥算法的优势，不久以后，清华大学朱军和陈文光教授的课题组做了一些新的创新，提出了 WarpLDA，重排训练数据的访问顺序，大大减少 cache miss，才真正发挥了这类  $O(1)$  复杂度算法的威力；其次，LightLDA 开源的代码并没有包含数据预处理和在线预测这一整套工具链，使得用户必须自己去开发和踩坑；最后，有一些较高级的特性虽然在内部版本实现了，却并未在开源代码中发布，譬如能搞定长尾语义的非对称先验的 LDA 等。我们也没有把单线程版本发布出来，方便同行做纯粹地算法比较。

### InfoQ：通过 LightLDA 项目，得到了什么启发？

**老师木：**第一，的确存在不可替代的技术，平凡的创新和破坏式创新的效果不可同日而语，后者往往有四两拨千斤的效果。

第二，要敢于迎接挑战，承担风险，个人理解，这相对于平凡而稳妥的道路更划算，做一件挑战但有风险的事，可能需要付出于平常事 3 倍的努力，但可能获得做 10 件平常事才会有的回报。

第三，无论是科学研究，产品研发，还是商业竞争是智商，意志，情商等综合素质的全面比拼，不仅要有不可替代的优势，在其它任何方面还不能有短板。

第四，机会总是眷顾有准备的人，有所准备才能抓住稍纵即逝的机会。总结，LightLDA 让我体验了做成一件有影响的事所需要的所有困难，我好像对看上去很难的事不会感到畏惧。

## 关于人工智能从业

**InfoQ：人工智能前景良好，那么从业者能发挥什么角色？**

**老师木：**有三种类型的技术可做：

1. 研究机器学习算法或原理，解答怎么做（How）或为什么这么做（Why）的问题，譬如研究怎么训练深度学习模型，什么样的神经网络结构效果最好，为什么深度学习要比其它机器学习方法效果好等等，简略称为原理问题；
2. 机器学习的基础设施，什么样的软硬件设计能使得机器学习算法计算更快，能用上更多数据，或者使模型规模更大，例如研发深度学习软件框架，或深度学习专用硬件等等，可归结为计算力问题；
3. 如何应用机器学习技术（算法和计算力）解决工作和生活中的实际问题，譬如互联网广告系统设计，推荐系统，游戏博弈（如AlphaGo），自动驾驶等等，可归类为应用问题。

**InfoQ：从事哪种类型的工作更有竞争优势？**

**老师木：**这三种类型的工作我恰好都做过，应该说哪个都很有用武之地，哪一个方向能做到顶尖水平都不易，做好了都能赢者通吃，全栈则更有优势。当然，这些工作也存在一些具体的差别。理论问题，进入门槛较高，工作岗位不太多，一般是兴趣驱动，看天赋和运气，这方面的突破，影响范围广，惠及全行业，从创业看，难以形成独立的商业模式，一般是在大学或企业研究院开展。计算力问题，影响力能到达全行业，通常是业界巨头和精干的创业团队的强项，岗位不太多，门槛也比较高，但主要看后天努力，一般是努力总有结果，创业上有可能形成独立的商业模式。应用类型的问题，业界需求最大，进入门槛低一些，确定性高，离商业近，

周期短，见效快，影响力一般受限于特定领域。统计上，少数人从事理论和计算力类型的工作，大部分人从事应用驱动的工作。现在的开源软件和公开课非常普及，为有志于在此方向上有所造诣的同行提供了前所未有的良好条件。

最后的话：“遍地黄金的日子过去了，低垂的果子已经没了”，技术创新主导的时代必将来临，让我们以“像鹰一样的眼光，像狼一样的精神，像熊一样的胆量，像豹一样的速度”，去抓住属于技术人的机遇。

# LinkedIn 的机器学习实践

作者 Divye Kapoor 译者 核子可乐



作为服务于全球超过 5 亿用户的专业社交网络，领英已经成为专业交流的首选平台。这里提供多种多样的职位选项，吸引到大量会员参与，其中部分文章更是拥有极高人气。但面对大量的受众与评论内容，人们往往发现有价值信息被快速淹没在干扰性内容当中。

为了为领英会员提供切实有效的评论内容，领英的研发团队构建起一套具备可扩展能力的评论排名系统。该系统利用机器学习技术为访问领英内容生态系统的每一位会员提供个性化的会话体系。在今天的文章当中，来自领英的技术专家将详尽介绍自己的设计思路、面临的可扩展性挑战与解决办法，以及系统运行中必须承受的有限延迟空间。

## 发展历史

领英馈送内容丰富多样，可供各类会员根据需求随意使用。其中一部分由会员有机生成（例如由领英公司创始人 Reid Hoffman 撰写的新文章等等），也有一部分来自第三方网站。研发团队需要确保能够定期发布高质量内容。这些文章将根据观点与偏好倾向吸引到广泛而热烈的会员参与，但就在不久之前，领英还没有能力将这种参与行为转化为平台之上真正具备现实意义的会员间对话。

领英馈送评论排名机制采取的默认模式为按新鲜度排序：如果您是最 后一位在某热门主题之下发表评论的用户，则您的评论将被显示在顶端。由于无法理解评论的具体内容，因此无法建立起个性化概念，更遑论解析评论内容与参与度之间的实际关联。

领英的技术团队从 2016 年开始对这个问题加以关注。他们建立起一 款简单的最低可行产品（简称 MVP），尝试利用其根据获得的“赞”数对 评论进行排名（即简单利用这一数字作为评论内容质量的判断依据）。这 款 MVP 获得了一定程度的左转，因为其确实能够将有价值评论优先显示 在顶端。然而，其也暴露出依据单一且缺少个性化因素的弱点：只有在得 到足够的交互评判之后，评论才会得到较高评价。在理想情况下，技术团 队希望能够提前找出那些值得一读的评论内容。另外，延迟与规模化等问 题也进一步阻碍开发团队将其引入实际生产系统。

根据从 MVP 当中汲取到的经验，领英建立起一套可扩展机器学习服 务系统，其负责利用一套富特征集对评论内容进行个性化排名。评论特 征——包括评论的实际内容、获得的参与度水平以及评论发布者的相关信 息等——都将通过 Samza 流处理工具进行提前运算。在此之后，其特征 将被预先判定，并快速与反射索引系统（FollowFeed）中的评论进行关 联。当某一请求要求为特定查看者获取相关评论时，预先计算完成的特征 与查看器将实现对接，而后运行一套机器学习模型以提供符合个性化要 求的评论排名。在过去几个月中，已经通过实践证明，这套系统每秒能够支

持数千次查询（简称 QPS），平均时长为 60 毫秒，且第 99 百分位延迟为 190 毫秒。

在接下来的内容当中，作者将就这套系统的具体构建细节进行探讨。

## 了解领英的对话机制

在馈送方面，每条评论通常都会产生一条 Comment Viral Update（即评论更新提醒），并通过评论者连接进行发布。技术团队利用这种评论更新提醒的方式建立领英新闻馈送中最具吸引力的更新术语，具体效果如下图所示。

The screenshot shows a LinkedIn comment feed. At the top, it says "Jeff Weiner commented on this". Below that is a comment by "Jonathan (Jasper) Sherman-Presser" (Media + Technology, Product + Strategy, 6d ago). The comment text is: "Hypothesis: the likelihood that a product succeeds is inversely proportional to how clever it seemed when you came up with it. Product folks, thoughts?". It has 1,663 Likes and 150 Comments. Below the comment are "Like", "Comment", and "Share" buttons. A link to "Show previous comments" is present. A reply from "Jeff Weiner" (5d ago) follows: "Agreed. Put another way, always felt like the products that seemed most obvious after the fact, i.e. "Why didn't I think of that?" were the ones most likely to succeed." This reply has 244 Likes and 29 Replies.

当您发布在某篇文章或者某条状态下的评论生成大量参与度——包括点赞或者后续评论——时，更新评论更新提醒会显示在您的馈送内容当中。在全部馈送更新中，评论更新提醒会得到最高的参与度。每条评论更新所生成的交互次数相当于关联更新的 2.5 倍（这里的更新是指您个人专业网络中出现的新连接），亦相当于连接内点“赞”参与度的 1.8 倍。2017 年，领英迎来了创纪录的参与水平。与馈送内容相关的社交操作（包括点赞、分享以及评论等）实现 60% 以上的同比增长。如今领英会员们正以前所未有的热情进行交互，而这些交流线索也将带来极为可观的价值。然而，规模庞大的评论内容也给技术团队带来了新的挑战。

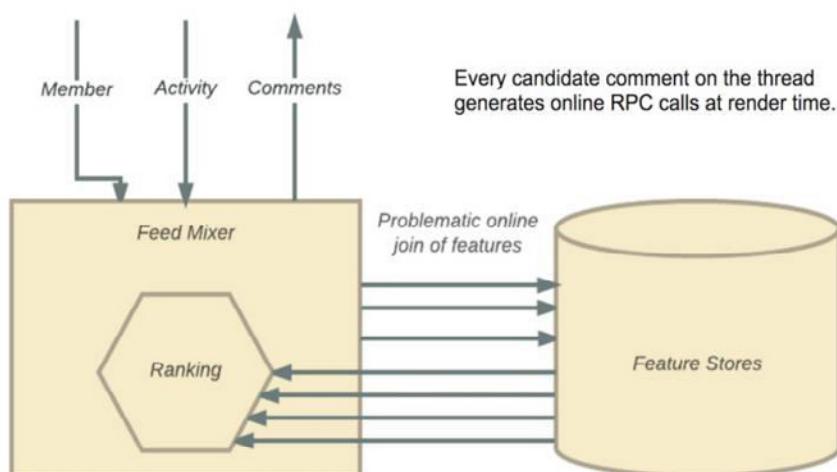
与互联网上的其它事物类似，评论线索也存在“长尾效应”。从列表的角度来看，少量长期拥有讨论热度的职位持续出现在会员们的眼前，并极大占用了其在网站上浏览的时间。具体来讲，1% 此类长线索吸引到超过 40% 的会员进行访问。很明显，以反向时间排序方式从头查看成百上千条评论内容绝对不会是令人愉快的使用体验。领英需要一种更为个性化的评论排名方法，以确保每位会员都能够从其关注的长线索评论当中获得最大价值。

## MVP 架构

那么，领英是如何最终实现为会员提供相关信息的目标的？在回答这个问题之前，需要首先聊聊最低可行产品（简称 MVP）的实际架构。

### 设计

MVP 的架构其实非常简单。



iOS 与 Android 移动应用会与 Voyager-API（一个基于 REST 的应用层）进行通信，旨在为特定馈送内容上的某项活动请求相关评论。Voyager-API 会将此请求转发至 Feed Mixer（排名与混合层），用以生成一份相关评论列表。

Feed Mixer 通过向用户信息库扇出请求的方式获取对方的一级连接，从而实现 MP 评论关联性算法。而评论线索库则将获得一份关于特定线索的完整评论列表。该评论线索库为领英 NoSQL 存储库 Espresso 的真相库来源。对于该线索之上的每一条评论，Espresso 都会存储一项示例特征（即合并后的点赞次数）。各评论依靠这样的特征进行排名，并被发送回 Voyager-API 以实现面向观看者的结果显示。

## 可扩展性挑战

通过架构图可以看到，当用户等待相关评论列表时，会出现在线特征加入的问题。在这种情况下，系统会因延迟问题而导致可排名评论数量受限。极高的计算量与由评论特征检索及处理带来的延迟会令 Espresso 库的后台延迟超过 800 毫秒，这无疑会引发严重的站点性能下滑。可以肯定的是，领英无法将此作为长期性解决方案。

尽管如此，此 MVP 仍然算得上成功。其能够显示排名评论的价值，而不仅仅是根据时间顺序进行简单呈现。另外，其也暴露出这种简单处理方式的弱点——高质量评论可能单纯因为得到的赞较少而被低质量评论所淹没，而且新近发布的评论由于没有足够的时间积累赞与回复而在线索判断层面处于劣势。凭借着积累到的这些宝贵经验，技术团队开始着手对该系统进行生产化调整。

## 可扩展生产架构

2016 年 8 月，领英决定推出一套不存在上述在线特征生成问题的架构方案。他们还借此机会尝试进一步提升评论排名的实际效果。

### 特征管道

在起步阶段，技术团队选择了一套能够有效描述评论的特征列表以及作为特征获取来源的数据源。其规模要远远超过只具备单一功能的 MVP。为了实现这项目标，他们从以下三个角度着手：

- 关于评论者的特征；

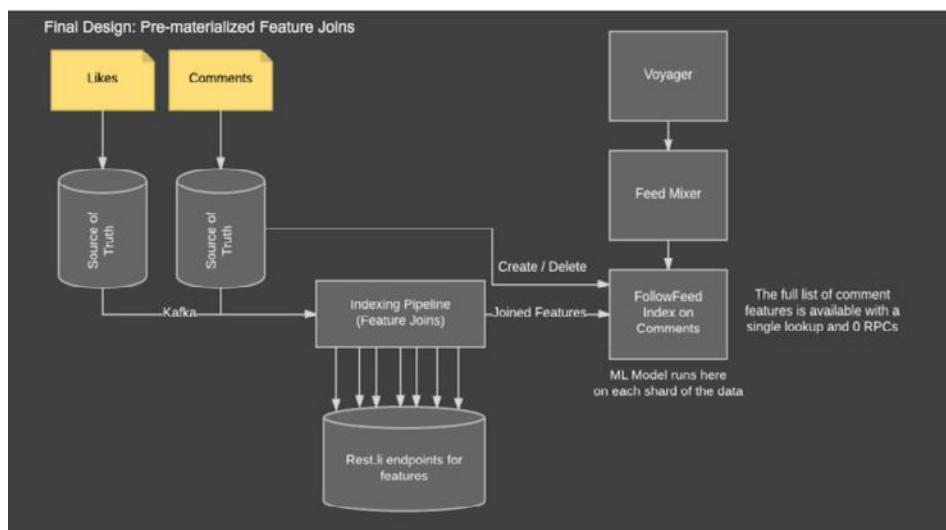
- 关于评论内容本身的特征；
- 关于评论参与度的特征。



上述各项特征皆通过对各数据存储库进行近实时查询的方式获取。这些特征通过一条特征添加管道进行推送，而此添加管道又由评论创建与交互进行驱动。另外，这一特征添加管道利用 Apache Samza 构建而成（领先的近线流处理系统）。

最终，技术团队得到了一份已添加特征的可用列表，其中包含每条评论的对应交付系统的单一 SSD/ 内存查找结果。另外，以上全部特征皆驻留在 HDFS 当中以备离线分析与模型训练。

## 交付基础设施



为了保证评论内容的相关性，需要一套能够满足以下要求的交付子系统：

- 这套系统应具备一份索引，用以立足某一评论线索（快速）检索全部评论内容。
- 立足该线索快速访问添加特征列表当中的每条评论。
- 有能力以规模化方式支持成千上万 QPS 并立足馈送内容为每条请求生成相关评论。

虽然（1）与（2）两项任务比较简单，但（3）才是这项设计的核心驱动因素。以规模化方式交付成千上万 QPS 大大缩小了可选范围，因此最终选项被固定在两套子系统当中：Galene（文档划分式搜索堆栈）以及 FollowFeed（术语划分式馈送堆栈）。

这两套系统在各自的领域中皆拥有良好的实际表现：Galene 为领英的搜索流量与多站点功能（例如工作推荐、人员搜索等）提供支持，而 FollowFeed 则支撑着馈送体系中的全部用户生成内容。在经过审议与一系列基准测试之后，研发人员决定采用 FollowFeed，因为其已经与馈送生态系统完美结合起来。不过这又带来了其它一些值得讨论的设计取舍。

这里解释一下，FollowFeed 是一套术语划分系统（其中每个叶节点负责存储与某一主要术语相关的文档）。在 FollowFeed 当中，各术语属于围绕 actor（例如会员、企业等等）建立而成的概念以及一份由该 actor 所执行社交操作的列表（评论、赞、分享等）。为了保证 FollowFeed 始终返回相关评论，需要对基础进行重新调整。

### 重整系统以接入可进行评论的各类条目

与能够在领英生态系统内部生成评论内容的少数 actor 不同（例如会员、高校、企业），这里研发团队面对的是更为广泛的会员可评论对象（例如文章、长篇帖子、分享、纪念活动以及视频等等）。FollowFeed 的核心数据结构能够将每个 MemberID 同一系列活动关联起来。技术专家们以这套数据结构为基础，而后构建起另一套能够将帖子 ID 同一份评论活动列表相关联的数据结构。从概念层面讲，这是一项小小的调整，但在大型生产系统当中实现这项调整却需要耗费远超想象的时间。

## 修正前 N 条与扇出概念

在馈送领域当中，技术团队面临的挑战在于如何为特定用户挑选最适合的前 N 条帖子。他们需要根据特定会员及其连接列表生成这些与其需求最契合的前 N 条推荐内容。具体来讲，需要提取该会员的连接集，并根据该会员的每一条连接扇出请求，最终给出前 N 条帖子推荐结果。

不过在评论方面，需要将以上提到的 1: N 扩大到 M: NM。其中的 M 为帖子 ID，而每篇帖子都需要生成前 N 条最佳评论。

尽管可以在 FeedMixer 层上进行 N 次请求扇出并对 FollowFeed 进行 N 次查询，但这显然不是最佳解决方案。

方差和法则告诉技术团队，N 个独立延迟量彼此为加和关系。这意味着：

$$\text{Variance}(X_1 + X_2 + \dots + X_n) = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \dots + \sigma_{X_n}^2$$

因此在 FeedMixer 层上进行请求扇出将极大提升延迟水平。最终的尾延迟不仅包含由 FollowFeed-Storage（最底层交付基础设施层）产生的方差，亦包括来自 FollowFeed-Query 层的方差。然而，如果首先对 FollowFeed-Query 执行单批次请求，而后再对 FollowFeed-Storage 进行一次扇出，则可有效控制尾延迟水平。

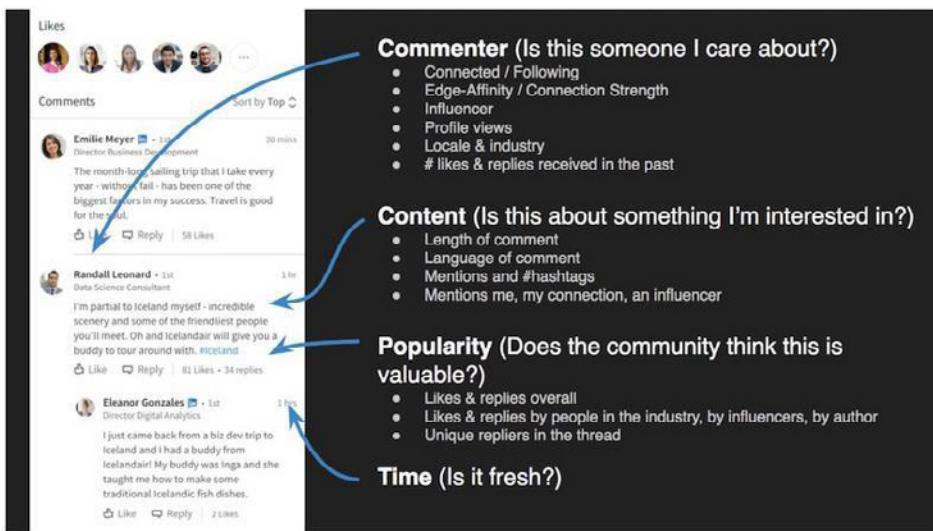
### 提供、发布及访问新特征，并利用其进行评论排名

这套系统尚处于早期开发阶段，因此还无法顺畅处理各类特征。在各个交付节点之上设置并使用特征存储库，在各节点间合理发布数据，并根据特定键进行划分。这一流程在基础设施工程层面非常简单：在出现数据时，技术专家们会利用 FollowFeed 出色的模型执行能力对来自机器学习模型内的相关评论进行排名。

## 领英使用的机器学习模型

为了实现个性化评论排名，领英训练出一套逻辑回归模型，用以预测查看者对于各条评论的参与度水平。这套模型利用领英 Photon ML 库当

中的广泛特征储备训练而成。



这套机器学习模型从查看者、评论者以及评论内容处获取特征。任何查看者特定特征（例如评论者与查看者间的亲密度）都会以每日方式被添加至离线 Hadoop 工作流内。这些特征被从 HDFS 中推送至一套在线 Voldemort 存储库以供查询。评论相关特征（例如评论的语言表达）则在评论创建时由一款近线 Samza 处理工具负责生成。如前文所述，这些特征将在 FollowFeed 中进行索引以实现在线交付，并通过 ETL 流程引入以供离线模型训练使用。

评论者特征包含每一位评论者的声誉以及受欢迎程度（基于其个人资料视图中的查看次数及影响力状态等）。领英的这套模型还会根据行业、职位以及其它共享属性对评论者与查看者进行匹配。在领英，能够凭借各类成熟的机器学习信号资源发现两位会员之间的交互关系。技术专家们会考量双方的联系 / 关注关系，其各自个人资料的相似性以及以往馈送内容中的互动记录。这些信号属于关键性输入信息，能够帮助技术团队为每位查看者选定最具个性化考量的高质量评论内容。

而在实际评论内容层面，领英的技术专家利用自己的内部自然语言处理（简称 NLP）库来表达语言、评论长度、语法结构、主题标签的存在 / 不存在以及其它内容特征。他们还尝试推断评论当中是否有提及其它领英

会员或者其它职能实体。

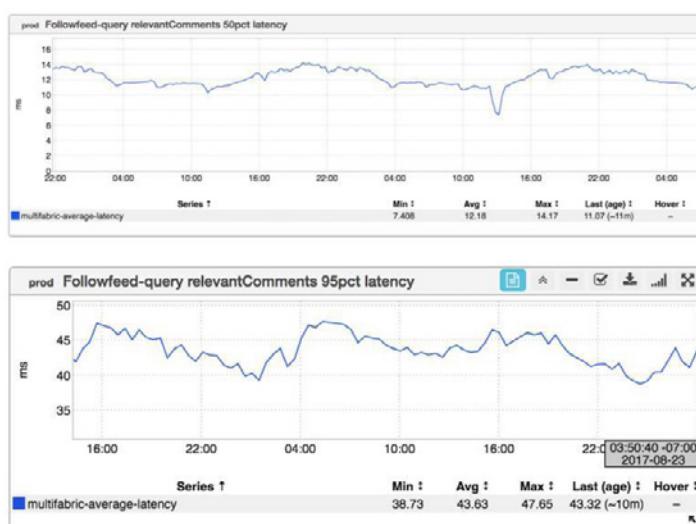
馈送信息当中的社交参与特征会根据不同行业进行细分，旨在保证机器学习模型能够准确找到只对特定一部分会员具有吸引力的评论内容。

评论新鲜度特征则源自针对当前评论的近期操作。技术专家们会捕捉评论的创建时间戳、最后回复以及最后点赞。查看者一般更倾向于阅读新鲜评论或者最近进行讨论的话题。

说到这里，还仅仅涉及这项任务的表面。在捕捉并实现在线排名功能的过程当中，实际使用到近 100 项特征，并利用机器学习模型进行特征训练以准确预测会员们对特定评论内容的参与度。对于每一位会员，都会利用其它机器学习模型对评论中的垃圾信息与低质量内容进行分类与检测，并最终选出最适合查看者的内容。

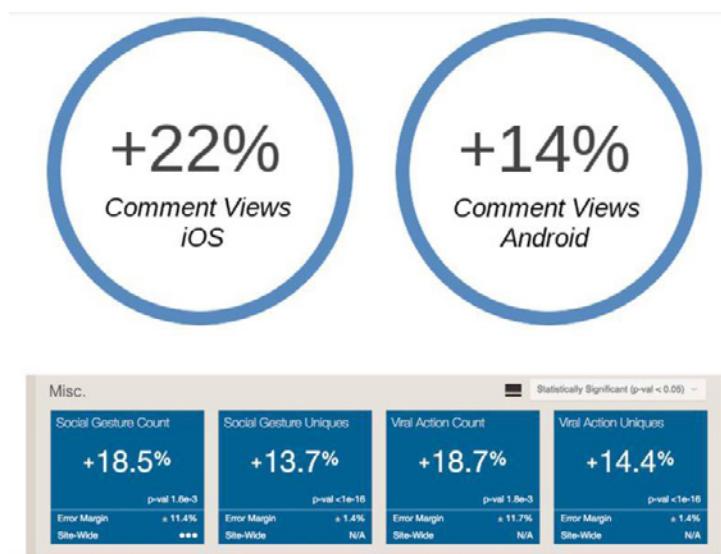
## 性能指标

这套系统拥有稳定的运行效果：可以看到，其第 50 百分位尾延迟水平为 15 毫秒，第 99 百分位延迟则为 65 毫秒。在获取会员特征时，整体系统的中位数延迟为 60 毫秒，而第 99 百分位端到端延迟则为 190 毫秒。具体来讲，这套系统相较于原本的 MVP，能够仅利用四分之一时间生成高达两倍的评论排名结果——这无疑解决了后者原本的最大短板。



## 结论

评论关联性这类长期基础设施项目拥有一段有趣的生命周期：首先是孕育期，而后是一段时间的持续评估，接下来投入具体执行，并最终进行稳定生产。回顾此前完成的这一系列工作，同时考虑到领英超过 5 亿会员得以借此获取价值，通过观察，会发现会员们阅读评论、参与内容馈送以及同领英生态系统当中其他会员进行交互与对接的效果皆有所改善。系统的个性化与资讯启发式发现能力得到了会员们的肯定，领英技术团队的工作也取得了良好的反响。



在调整之后，回复当中“赞”的数量开始快速提升。目前 iOS 平台上馈送内容的评论数量增长了 22%，而 Android 平台则提升 14%。

# 爱奇艺视频场景下的自然语言处理应用

作者 Moment



## 一. 引言

NLP 涉及的面非常广，包括语音识别 / 合成、信息检索，信息抽取，问答系统，机器翻译、对话系统等。

在爱奇艺，自然语言处理团队专注于以下 7 个方向：

1. 词法分析和知识图谱
2. 打标签 (Tag Recommendation)
3. 查询理解
4. 热门事件发现和聚合

5. 语音助手
6. 舆情分析
7. 电影票房和电视剧 VV (video view) 预测

从而实现更好地理解视频 / 图文内容，用户的搜索意图和用户的评论，为搜索、推荐、广告、社交、舆情监控的智能化提供基础服务和技术支持，并探索 NLP 的直接应用业务。

## 二. 词法分析和知识图谱

我们的词法分析作为文本分析的基础服务，已广泛引用于多个亿级流量的业务线。



**图 1 词法分析平台**

图 1 显示了现阶段的词法分析功能：

分词、词性标注、词权重、新词发现、实体识别 / 链接功能等，采用的技术主要包括 CRF、L2R、CNN、CNN+CRF、LSTM+CRF。

其中，实体识别是词法分析中的重点也是难点。除了通用的人名、地名、组织机构名的识别，我们还特别关注娱乐领域的影视剧名、游戏名、文学作品名、游戏解说名等的识别。

上述的娱乐领域的实体识别挑战较大，主要包括：

1. 目前工业界和学术界还鲜有相关工作的介绍。

2. 实体本身的规律性弱。任何一个词都有可能是实体的一部分，例如“杀破狼”、“西游记之孙悟空三打白骨精”等；
3. 实体词与实体词之间、以及实体词与普通词之间的歧义性大，如电影“十二生肖”、“功夫”、“长城”、电视剧“解密”，既是普通名词也是实体词，“非诚勿扰”即可能是电影，也可能是综艺或普通词。
4. 缺乏训练语料

我们首先在训练语料的准备上做了大量的工作，包括：

1. 使用启发式规则自动构建了 100 万句弱标注的视频语料。
2. 人工方式标注了几万句的精准标注的视频语料。

在实体词典的构建上，使用数据挖掘技术实时地从全网挖掘影视剧名 / 角色名 / 艺人名 / 游戏名等领域词典。

算法上不仅在传统的 CRF 模型上做了很多的尝试，也在深度学习方法进行了一些探索。CRF vs. CNN vs. LSTM 等对比实验表明，如图 2 所示的双层 CNN+CRF 模型获得较优的性能。在 2 个不同的测试集上，我们的模型对剧名识别的 f-score 分别是 82.1% 和 72.6%。

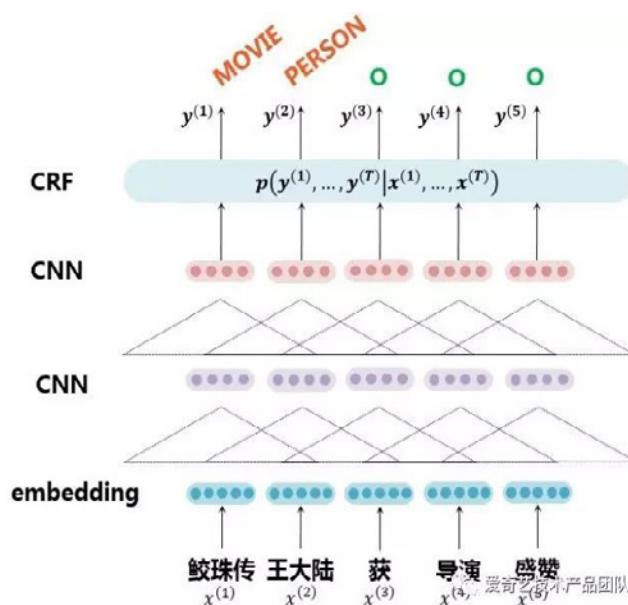


图 2

实体识别 / 链接服务除了作为底层的 NLP 基础服务应用于各业务线，我们也开发了多个直接应用业务。

- 通过实体识别进行泡泡圈子的内容分发：即从图文 / 视频 / 图集中的文本中准确地识别明星、剧名；对识别出的明星、剧名按照和内容的匹配度进行重要性打分、并根据重要性将内容自动分发到对应的明星圈子和影视剧圈子。
- 通过实体识别将 feed 流中的视频和电影票（图 3 左）、游戏（图 3 中）、电商（图 3 右）、漫画和文学等垂线业务进行关联。可在不损伤用户体验的情况下，提高对垂直业务线的导流，进而实现一键购买电影票、下载游戏和下单电商等。



**图 3 基于实体识别 / 链接的 feed 流视频与垂线业务的关联**

在视频领域知识图谱，我们可以分析出视频与视频、视频与人、人与人，人与视频的关系。再结合精确的语义分析、实现了用户查询的精确回答（图 4）。未来我们还要继续挖掘游戏、文学等更多垂直领域的实体属性和实体关系。

### 三. 打标签

标签 是从对内容（视频、图文、或图集）的描述（标题、摘要、或正文）中提取可表示内容的一种元数据（关键词或术语）、有助于更好的个性化内容推荐、更高效的内容编辑。

标签可以是一个封闭的预定义分类体系（我们称之为类型标签），也

可以是从内容中提取的开放的关键词集合（内容标签）。



**图 4 基于知识图谱的问答系统**

标题	范爷辣眼睛新街拍, 难道减肥真的可以无止尽吗?
内容标签	范爷 ( <a href="http://www.iqiyi.com/lib/s_200044305.html">http://www.iqiyi.com/lib/s_200044305.html</a> ) 街拍 减肥
类型标签	娱乐 明星 内地

**表 1 视频描述 (标题) 及其类型和内容标签**

类型标签 采用的是基于 SVM 的分类算法，特征包括字的 n-gram、词的 n-gram、主题语言模型特征、词典特征等。

传统的内容标签抽取方法分二步走：

1. 基于启发式规则的候选标签生成。
2. 基于无监督 (TextRank, ExpandRank) 或有监督 (Maui, CeKE) 算法的候选打分，并输出概率最大的作为系统标签。

按我们经验和对业务的了解，我们将基于打分或者分类的内容标签任务转化为一个序列标注任务，并采用 CRF 模型。该算法具有：

1. 可以抽取任意长度的词组作为标签
2. 不再需要单独的候选抽取模块
3. 可以获得最佳的性能

目前、标签服务已经应用于视频推荐、爱奇艺头条、泡泡、视频编辑等业务等。

## 四. 查询理解

查询理解包括个性化的默认搜索词、查询补全、查询纠错和查询分类等。

其中个性化的默认搜索词是在用户发生搜索行为前，通过用户在爱奇艺的历史行为猜测用户可能感兴趣的 query。其本质是一个推荐系统，方法是计算用户画像和 query 的相似度。优秀的个性化默认搜索词可以增加用户黏性，提高用户体验，进一步地引导用户行为。

查询补全是在用户发生搜索行为的过程中，通过用户不完整的输入（我们称之为 token）与 query 的匹配度，query 的点击量、专辑与否、freshness 等提示用户一些可能感兴趣的 query，提高搜索效率。

## 五. 语音助手

我们的语音助手已落地在爱奇艺 VR 一体机和爱奇艺 APP 上。通过 VR 语音助手，可以实现和 VR 一体机的虚拟女友 Vivi 进行 40 多种交互，包括视频播放 / 搜索、天气查询、和 Vivi 的互动、VR 设备设定（亮度调高、音量调低）等。

在 APP 里，语音助手可实现便捷地购买 VIP 会员（我要买爱奇艺 VIP 会员），下载游戏（我想下载爱奇艺斗地主游戏）、直接观看电视剧的某一集或电影等。

语音助手简单来说，即是 把用户说的话 (utterance)，转换为结构

化的语义表示，从而执行相应的动作（action），分为如图 6 所示的 3 个大模块：语音识别、语音识别纠错、语义解释。



试试这样说

我要买会员

播放楚乔传第十集

下载王者荣耀

按住 说出你想搜的

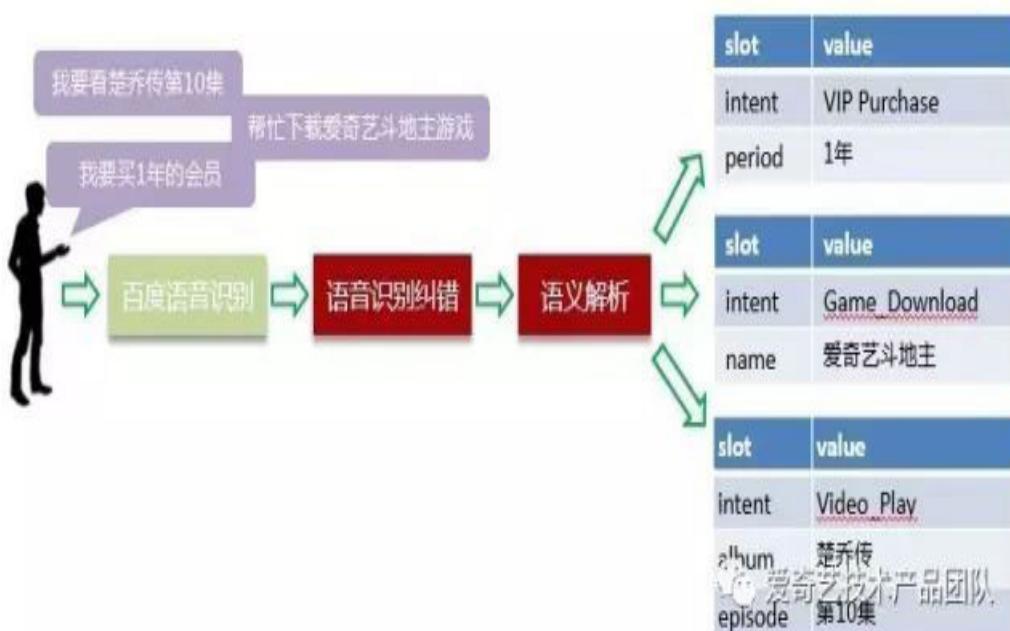


图 5/6 APP 上的语音助手 / 语音助手框架图

语义解析模块又进一步分为 意图分类（intent classification）和要素抽取（slot filling）。

需要说明的是，第二节介绍的词法分析，特别是影视剧名识别，和第四节介绍的基于爱奇艺全网搜索的查询纠错让我们的语音助手鲁棒性，特别是在影视娱乐、游戏领域得到较大的优化。

## 六. 舆情分析

**舆情分析** 可直观反映观众对剧和艺人的关注焦点和态度，为版权方和自制剧的内容运营、内容营销策略制定、营销趋势把握提供参考。

我们使用自然语言处理中的句法分析技术，从 UGC 内容（用户评论、弹幕、泡泡圈子）中抽取评价对象、评价词以及情感色彩，从而形成对用户观影评论、社交互动的多维度结构化舆情分析。

图 7 是对电影“战狼 2”从 视觉效果、场面、演员 三个维度的舆情分析结果。



图 7 电影“战狼 2”的部分舆情分析结果

带情感的热词分析效果可以访问爱奇艺指数网站 <http://index.iqiyi.com/>，其中 词的大小反应提及频度， 词的颜色反映情感色彩。

## 七. 电影票房和电视剧 VV 预测

无论是票房还是 VV 的预测，都面临很大的挑战，包括：

- 提前时间长（提前 1 年 / 半年等），可获取的信息有限；
- 上线前影响因素较多（如同期影片、突发事件）；
- 训练样本少（少于 1000 部）；
- 站内外多个数据源的数据融合、清洗等。

基于大数据和机器学习算法，我们对电影票房、电视剧 vv (video view)、综艺 vv 等提前 60 天、180 天、360 天等多个时间窗口预测，为版权剧采购立项、自制剧立项、广告售卖等提供科学的数据支撑。

为获得较好的性能，我们在数据清洗和特征工程上做了很多尝试。最后采用了包括时间类、题材类、播放平台和方式类、指数类、ip 类、前作类、趋势类等 100 多维特征、并对丢失特征的补全和部分特征的变换。

模型上对比了线性模型、SVM、随机森林、GBDT、DNN、stacking 集成方法等。

在最近的 90 部版权电视剧上最优的 R2 准确率为 85%。vv 超 10 亿的头部剧预测误差在 30% 以内的占 67%，误差在 50% 以内是 100%。

图 8 是部分剧的预测 vv 与真实 vv 的比较。



图 8 部分头部版权剧提前 180 天、60 天预测值和真实值的对比

## 八. 总结

基于用户弱标注和人工精准标注数据、使用机器学习和深度学习的自然语言处理技术更好地理解视频、理解用户，从而让搜索、推荐、数据挖掘更智能，为用户提供智能化的专业视频体验。

接下来，我们要进一步优化上述功能模块，并拓展在视频场景下的更多应用。

在算法上，将进一步探索更有效的深度学习模型、文本和图像的融合、迁移学习等提高系统的性能。

本文系 爱奇艺技术产品团队 原创文章，已经授权 InfoQ 公众号转发传播。

## 作者介绍

**Moment**，2016 年至今任职于爱奇艺技术产品中心 - 搜索广告部，主要负责自然语言处理（NLP）和商业系统的研发和管理工作。博士毕业于中科院自动化所自然语言处理方向，先后在日本 ATR 研究所、日本情报通信研究机构（NICT）、英国爱丁堡大学（短期访问）和索尼中国研究院担任自然语言处理、语音识别与机器翻译等研发工作。曾在 EMNLP、COLING、CIKM、INTERSPEECH、ICASSP、Computer Speech & Language 等国际会议和期刊发表文章 20 余篇。

# 小红书里的秘密：机器学习如何帮助十人算法团队快速达成目标

作者 赵晓萌



本文要为大家分享四点内容：

首先介绍下小红书、小红书的人工智能团队、以及小红书在机器学习上的应用。

第二点，举一个深入的例子介绍我们怎么理解用户在小红书上产生的内容。

第三点介绍下人工智能在推荐搜索中的应用以及在小红书的应用。

第四点是结合我的经验介绍下如何在像小红书这样一个比较初期的人工智能团队，这样比较小的公司里更好地应用人工智能。

我早上听了几位讲师的分享，特别留意到他们公司有自己的机器学习

训练框架和训练平台，他们在算法上做了非常多的优化，我们作为一个小公司并没有这些，我很羡慕，希望有一天我们也会有。去年我们的算法团队大约从 6 个人发展到了 10 个人，以下介绍的是去年一年的时间内，尤其在后半年完成的一些工作，希望我们的经验能够给人员和资源都比较早期的公司一些借鉴。



## 小红书与人工智能

我先从介绍小红书开始，介绍下我们算法团队需要解决的问题。小红书是一个分享社区加电商的 APP。

首先看下分享社区，分享社区以女性为主，是一个有少量话题引导的，但基本上是自然形成的，关于分享精致好生活的社区。这里主要分享的内容包括美装、穿搭，喜欢去的餐馆，新发现的旅行地点、酒店，最新的母婴和家居生活的内容也在增加。小红书今天有五千万的注册用户，月活跃用户超过千万，这些用户帮助我们在平台上产生了九百多万篇非常高质量的分享。这么多的内容，我要如何转发分发给用户，让他们看到想看的，这是一个算法组需要解决的问题。

小红书的另外一面是福利社，就是电商，这个问题很简单，就是你在



社区里看到的，在福利社希望能买到。如何提高福利社的商品购买转换率，这个也是算法组需要解决的问题。



小红书独一无二的地方，就在于它把社区和电商融合到了一个 APP 里，非常重要的一点是它造就了女生人数占比最高的互联网公司。在我们公司里面女生男生占比是 7: 3，其他公司里比例可能最高 3: 7，或者 1: 10 都是有可能的，这是小红书独特的地方。更重要的一点是，它造就了 1 加 1 大于二的效应。



这是什么意思呢？我们社区是提供用户黏性的，它为我们电商引流，电商这部分把流量变现，在我们 APP 里形成了一个闭环，这两个是互相推动的。对于算法团队来说，因为我们有这样独特的形式，我们有社区的用户数据，同时也有用户在福利社的行为数据，我们如何把两边的行为连接起来，更好地理解用户，这是一个非常独特的挑战。



我们算法优化的目标是什么？优化的核心目标也是两个，分别对应社区和电商。社区的目标是用户增长，我们衡量的是在社区的深度交互，这是一个间接的，但是离我们更近的一个 metrics(指标)。对于电商我们要做到的是驱动盈利，福利社的加车购买，是我们关注的指标。机器学习大概是从去年年初开始在小红书慢慢地发展起来，截止到去年年底，整体

效果还不错，我们需要达到的深度交互，以及电商的购买转换，都有非常不错的提高。

## 深度理解小红书们产生的内容

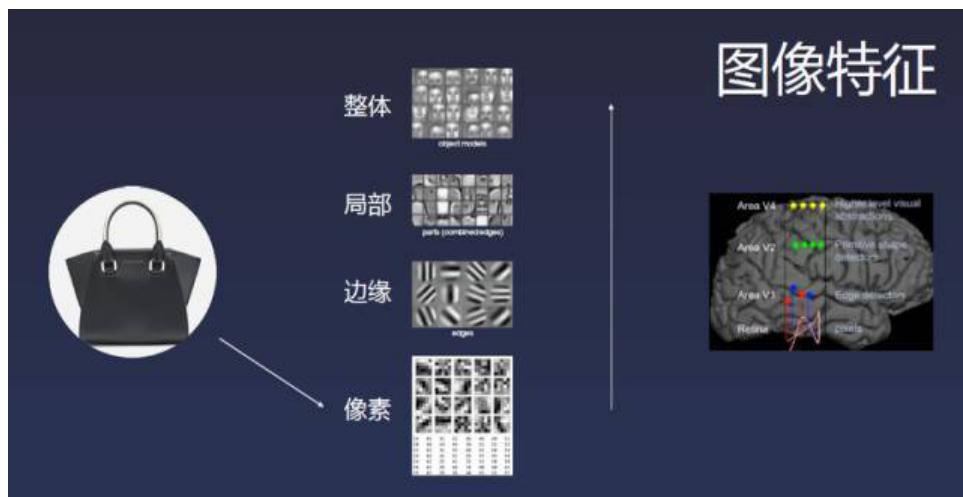
刚才介绍了小红书和我们算法要解决的问题，现在举个具体的例子介绍怎么理解小红书的内容。刚才提到五千万用户的九百万篇推荐笔记，是我们最重要的内容，我们花了非常大的精力来理解内容。



首先我们看一下这些内容大概是什么样子。很简单，是图文并茂的。用户产生的内容图片多，而且质量非常高，同时是非常详细的种草文（推荐物品的文章），这个文章正常情况下不是横过来的，横过来是为了让大家看到这个文章很有长度，写得非常仔细，吸引眼球，而且有感情有干货。



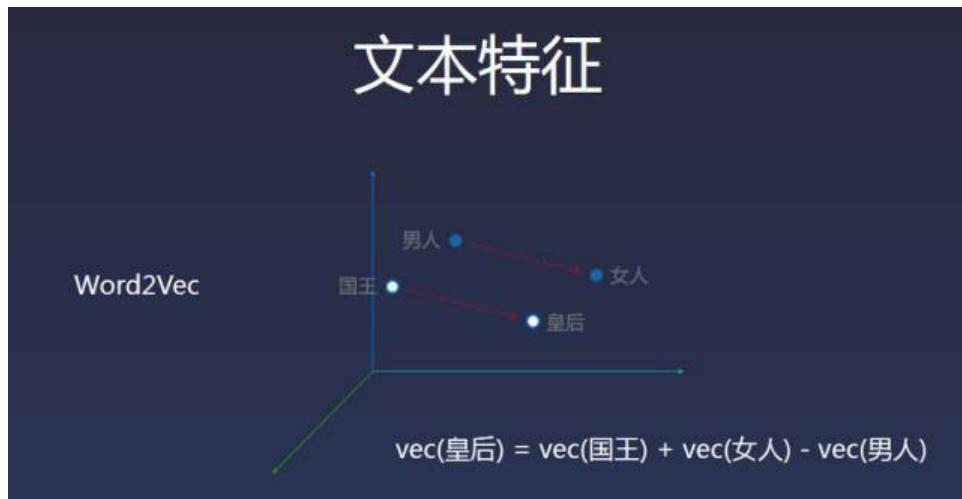
因此需要把文本和图片结合起来去理解文章内容。我们通过机器学习把笔记的主题分到人工标定的上百个主题里。我们用 CNN（卷积神经网络）提取图像特征，用 Doc2Vec（文本到向量模型）提取文本特征，通过一个简单的分类器就能把用户笔记分到主题中。接下来具体介绍下图像特征的提取。



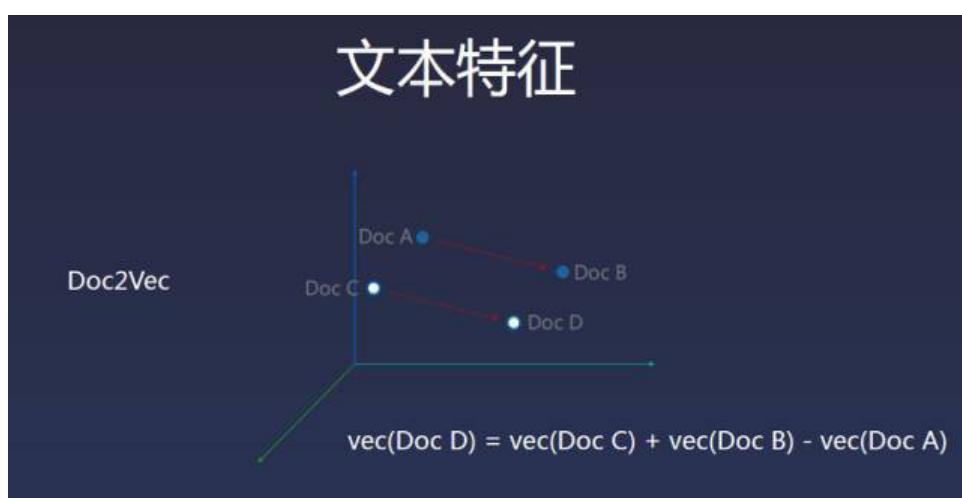
我们用的是卷积神经网络。

卷积神经是深度神经网络，层次比较多，是 feedforward 神经网络。简单解释下它的原理，它模拟了我们大脑处理图像的过程。什么意思呢？如果我们选一个稍浅的神经网络，把这个多层神经网络每一层的输出打出来，那我们大概就能够了解这个神经网络在做什么。最下面的输入层是我们图像的像素，第一层、第二层的输出我们可以看到这个神经网络能够提取出一些边缘的信息，再往上四五层的时候，它把边缘的信息组合起来了，我们会看到转角、圆圈还有网格这样一些形状上局部的信息，再往上到第六第七第八层的时候，就能看到一些整体的概念被抽象出来了。经过这个卷积神经网络层层的抽取和抽象，在像素之上会形成概念最有用的一些特征，这个就是我们拿到的图像的特征。通过这个卷积神经网络我们把一张图变成一个 4096 维的向量，这个向量是这个图在高维空间里的一种表示，它是有空间意义的，这个意义是指相似的图片，或者说图片上的相似特征在这个空间里是距离接近的。

我们这一套神经网络是在 Caffe Model Zoo 的很多已经预先训练好的模型里选出来的一个 VGG 的 16 层神经网络，它已经在 ImageNet 上训练好了，我们不需要花太多时间去训练它，我们只是标注了少量小红书上的图片，把它的主题标上去，然后我们再 fine-tune 这个神经网络，最后就达到我们期望的效果。



讲完了图像，我再讲一下文本的向量表示，文本的向量表示有非常多种，其中一个比较有名的向量表示叫做 Word2Vec，是 Google 提出来的，它的原理非常简单，它其实是一个非常浅的浅层神经网络，根据前后的词来预测中间这个词的概率，优化预测的时候模型就得到了词的向量表示。同样的这个词的向量表示在空间里也是有意义的，相似的词也处在相近的



空间里。这个模型比较有意思的是，把向量拿出来随时可以做向量运算，比如图中，女人到男人之间的那个指向的向量，和皇后到国王之间是一样的，所以你知道其中三个，就能算出另外一个。

把文字的 Word2Vec Model 往上提一层时就会得到文本到向量的 Doc2Vec Model。那么怎么用它呢？其实就是把小红薯，就是我们的用户的笔记上的标签，它提到的商品的品牌品类加上笔记本身的内容一起放到这个模型中，我们就得到了一个描述笔记的向量，然后再提取。这个模型也是开源的，我们也直接拿来用就可以了。

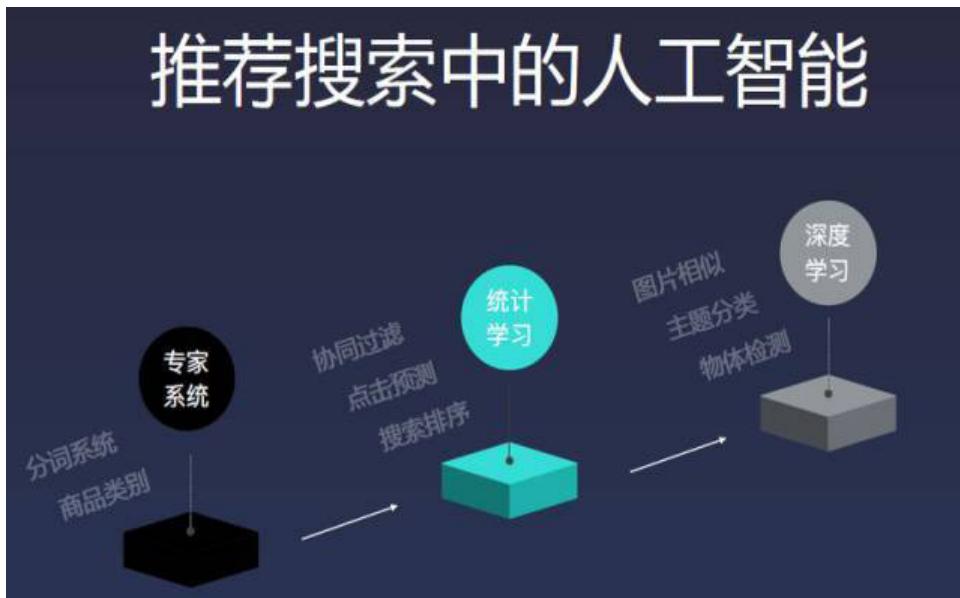


当我们现在有了图像和文本的向量表示以后，我们就有监督地去训练一个分类器，把它分布到我们标注出来的主题上，我们用的分类器是一个，有一个隐层的全连接的神经网络，能达到我们要的效果。之前我说过小红薯是一个非常视觉的社区，图像很多，我们只用图像提取特征就已经达到良好的效果，准确率大概是 85% 时覆盖率能达到 73% 左右，加上文本以后效果更好，准确率达到 90%，覆盖率达到 84%。

## 人工智能在推荐搜索中的应用

上面是一个我们用文本和图像特征来理解我们用户产生内容的具体实例。下面给大家概括介绍下小红薯机器学习使用的情况。

做这方面的同行应该知道人工智能有三个阶段，最早的专家系统，到



统计学习，到深度学习现在慢慢流行起来。对于我们这样一个小的公司来说，我们非常注重算法的实际效果，远超过我们看这个算法先进不先进，比如刚才我们讲的提取图像特征的模型，并不是 CNN 里效果最好的，而是一个相对简单的模型，16 层神经网络对我们来说相对简单，容易理解，比较能 Hold 住。

人工智能的三个阶段产物，我们都各有各的应用。

专家系统是指我们需要依靠人对问题的理解来设计规则，比如中文搜索中特殊的分词，在小红书早期时，搜索中大约 80% 的问题都是因为中文分词分得不太准，导致用户搜不到想要的东西。我们通过人工的一些工作，从人工加词、加词典维护，到我们做了新词发现的一套半自动的系统，这些问题就被解决掉了，实际上分词有更好的解决方法，现在已经有了基于深度学习的方法，有可能以后考虑替换这套系统。

有些专家系统确实不太好替换，比如对于电商来说，它的品类（商品的分类）系统是非常需要行业知识的，需要对商品品类有深刻的理解，而且需要结合公司电商发展的阶段来设计这套系统，系统要随着公司发展阶段的变化去不断地迭代。这套系统特别重要，因为当品类不对时，在品类下的推荐就不准，搜索的筛选可能会做得不太好，或者品类的粒度分得不

够细不够准时，推荐和搜索的算法都会受到非常严重的影响，所以这套专家系统可能会一直存在下去。这是最左边的早期的专家系统，相当于早期第一个版本的人工智能，最新版本的人工智能就是现在比较火的深度学习，这点刚才也讲了，主要讲的是主题分类，同样我们可以做到图片相似，以及物体检测等。

剩下中间这块统计学习，统计学习是我们使用量最大也是实践中最常用的，这一套系统当中，最重要的就是统计的信息，往往需要用到大量的统计特征。



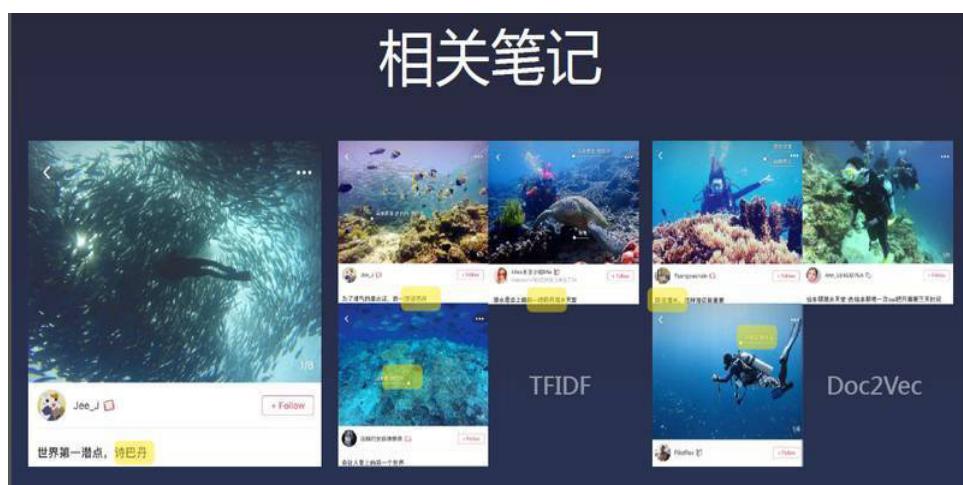
比如在我们推荐的 GBDT 模型中，模型的产品目的是个性化，我们希望用户可以看到想看的笔记，这属于我之前讲的笔记分发的问题。在这张表里，大家可以看到，我们有非常多的用户行为统计，产生了一些静态的信息，用来描述用户或者笔记。

我们通过用户画像和人口统计信息来描述用户，比如性别年龄等常用的静态信息。笔记分作者和内容两个维度，比如作者打分和笔记的质量、标签以及刚才介绍的主题等。还有一些我们实验过的更复杂的统计信息，被我们放弃了，因为虽然复杂但是效果并不显著。比如我们会算用户的行为趋向，是趋向在社区里花的时间比较多，还是趋向于在福利社电商花的时间比较多，我们尝试过统计用户的生命周期，这个用户是新用户、老用

户、经常回来的用户、还是可能快流失的用户，还有用户的活跃时段，用户是在早上比较活跃还是晚上比较活跃，用户购买力，用户对折扣的敏感度，在社区里的活跃度，这些统计信息我们都尝试过，这些都比较静态地去描述一个用户。

还有一个非常关键的信息是动态特征，虽然动态特征并不多，但是很重要。动态特征包括用户在浏览和搜索过程中有没有点击、有没有深度行为等类似的用户反馈，这些交互的数据有一个实时的 pipeline 从线下直接放到线上的模型里，在线上会利用这些数据对交互的质量，比如点击率进行预测，以及通过协同过滤得到用户和笔记的隐性分类，这在推荐当中也是有用的。我想讲的是用户反馈的数据，即使是简单的统计都是非常有用的。我们在使用复杂模型之前先用简单的统计方法把用户的反馈数据放到模型预测中，可能就能达到想要的 80% 的效果，这是非常重要的。

还有一点是我们有两部分的数据，社区和电商的行为数据，用户在社区的行为和电商的行为是不太一样的，而且是有点互补的。比如用户在社区的行为是比较高频的，用户会在这里搜、看、点击、点赞，可能因为好奇进行点赞和点击，而在电商的数据是低频的，比如产生最后的购买。高频数据我觉得统计信息是非常有用、非常准确的，对于电商，我们认为比较昂贵的行为，比如购买和加心愿单的行为，是非常可信的，这两种数据是互补的，我们试着把这两边的数据融合起来用到特征里。



下面再举一个比较具体的例子，另外一个从文本中提取特征的例子。

之前讲的提取特征，是为了判别文本的主题，我们用的是 Doc2Vec 文本到向量的方法，向量越接近，文本越接近。现在这个场景叫相关笔记，相关笔记的要求是什么呢？推荐的笔记和用户在看的笔记最好讲的是同一个东西，比如说，同一款口红，同一个旅行目的地，同一家酒店，同一家餐馆，有可能不是同一家餐馆，是类似的餐馆，或者说同一件衣服，但是也有可能是不同款但是相似的衣服。

相关笔记的要求的是首先相关性非常强，第二在这个基础上稍微有些扩展。实际实验时发现，如果用 Doc2Vec（文本向量表示）选出来的笔记不太能满足相关性的要求，比如上图的例子，讲的是世界第一潜点诗巴丹，Word2Vec 的结果不太在乎具体地点是哪里，在最右边的例子里，我们可以看到，它找的是附近的地方，比如越南芽庄，它会把相似的地方找出来把它替换掉。在这样一个场景当中，我们选择了另外一個词向量的方法 TDIDF，一个简单的统计学方法。这个就能比较好地解决相似性的问题，因为它本来就是用于信息抽取和信息检索。

有一点让我觉得比较惊喜的是，TFIDF model 虽然基本要求词是一样的，但它可以把一类笔记找出来，就是讲用户心理、描述用户心情的笔记，因为用户描述心情用的词汇很接近，所以这个方法也会把扩展的内容找出来。举这个从另外一个应用场景来选文本向量表示的例子的意思是我们的

## 更智能的未来

- 深度学习的加速普及
- 更易用，更有效的算法
- 更有创意的应用

算法选择在小公司里需要非常接地气，需要考虑具体要求，而且是实验性质的，如果不做实验尝试就没法知道哪个方法更能满足具体场景的需求。所以对于小公司来说，团队能不能快速试错、实验和迭代，这个能力可能比某个模型的质量或者模型本身能力的局限更加重要。

之前讲的几个具体的例子是我们已经实现过的，接下来展望下小红书未来机器学习团队需要做的事情。

之前举的几个都是统计的例子，主要想说明我们如何选择算法，如何注重开发的成本、速度和最后的效果。其实深度学习的效果在推荐预测上已经渐渐超过了之前讲的一些统计模型，随着机器学习平台的成熟，以及相关模型的开源，我们也会考虑把之前在推荐里用的 GBDT 模型替换掉。

深度学习有什么好处呢？首先讲深度学习的一个缺点，就是抽取的特征比较没有解释性，人工特征比较好解释是因为预先设计了特征，然后再去构造。机器学习的特征在抽象完之后，仍然保留了很多信息，虽然不好解释，但是有一个好处，留给应用想象的空间很大，可以实现一些比较有创意的应用。比如去年有一段时间小红书上突然流行分享治痘，就是脸上有很多痘痘，怎么把它治好的这种文章，用户会秀很多自拍的、脸上长很多痘痘的照片，但不是每个用户都喜闻乐见脸上充满痘痘的照片，所以我们需要识别出这些照片，把它推荐给合适的目标人群，这也可以通过我刚才讲的 CNN model 来实现。我们尝试做这件事，发现它对全脸露出的、半脸、1/4 脸甚至脸上只有少量的脸部器官，都能识别为脸部图像，而且能够识别脸上有没有痘痘。CNN 还可以很好地识别这张图里是不是文本占了绝大多数，比如是不是一个截图，对 AntiSpam（反作弊）会有帮助。未来 CNN 还可以帮我们做更多，比如我们想做一些风格上的尝试，希望通过用户买的东西和经常看的东西能够知道用户穿搭的风格。

## 如何在初创公司合理使用人工智能

刚刚讲完了小红书的故事，现在结合我的经验介绍下如何在类似的初创公司合理地使用人工智能。

## 初创公司的人工智能应用

- 将人工智能融入公司业务，注重实用性
- 有别于Google，Facebook，阿里，百度
- 有别于主业基于人工智能的创业公司

我指的小公司是怎样的小公司？首先是希望人工智能能够融入公司业务，是非常实用主义的小公司。它不是 Google，Facebook，阿里百度这样的大公司，大公司研究的更多是人工智能的平台和框架，提供什么样的服务，专注在训练和算法效率上的提升。同时也有区别于人工智能创业公司，这些公司人工智能是它的主业，比如视觉识别、自动驾驶，它们专注于算法的创新突破，算法准确率需要有比较大的提升。我讲的这些小公司，为了将人工智能融入公司业务，它需要更多的是被验证过的算法，它关注的是算法的实用性和开发维护的成本，对于这样的小公司，我们有经验，我想从两个方面谈一下，它应该怎么看待在自己的公司应用人工智能这件事情。

## 加速降低的成本

- 运算力，机器学习平台的开放
- 模型的成熟，开源
- 理论知识的相对重要性在降低，工程（学习）能力的重要性在上升

首先我觉得第一点非常重要，就是越来越多的人意识到机器学习、人工智能的应用成本加速降低，小公司要抓住这个机会，抓住人工智能发展提供给自己的红利。运算力、机器学习平台的开放大家都知道，Tensorflow、Caffe 以及 MXNet 都想扩大自己的 Community，希望能够有更多的公司和更多的开发者用他们的平台。

另外一点是模型的成熟和开源，这对于深度学习阶段尤其重要，因为在统计学习阶段模型也是开放的，但是这个阶段算法的核心不在模型，而在模型里使用的特征，特征工程是当时的核心。到了深度学习，情况发生了改变，因为深度学习的核心就是模型，模型能够抽取特征，能够很快地在分类、推荐、预测得到应用。

对一个小公司来说，我们是非常需要开源的，因为从头搭建自己的这套模型非常耗时、耗力，比如我们刚才用的神经网络，它有多少层，它的层和层之间应该怎样卷积，需不需要使用 dropout，这些都是需要花大量的时间做实验，对于小公司来说这样的投入并不值得，投入太大而产出效果可能不太好。而且深度学习好的模型通用性非常强，比如我刚才举的例子，可以用它处理很多图像识别方面的问题。

最后一点是理论知识，我认为小公司如果有效地利用这点就能够比较快地享受到人工智能带来的利益。理论知识的相对重要性在降低，工程学习能力的重要性要求在上升，这说明个人力成本在下降。几年前在推荐预测上要做到比较好的结果还是很难的，大家可以参考 09 年的时候 Netflix 做了一个挑战奖金是一百万美金，想提高推荐系统的效率，全世界当时有相关知识而且能够把这些理论知识用到推荐系统里的人非常少，人力成本非常昂贵。那么到了现在呢，我记得 Google I/O 上 TensorFlow Team 有个分享，说 2005 年大约一个研究室 6 个月想要做到的基于神经网络分类的效果，在今天一个优秀的 Python 开发者有一些 TensorFlow 的背景知识，大概需要几天就能远远超过当时的效果。

小公司对理论知识非常强的深度学习大牛的依赖程度大幅度降低了，而且一个小公司请个大牛并不划算，第一大牛特别贵，第二因为大牛一般

## 整个团队的投入

- 核心业务数据（社区，电商）的设计，整理，迭代（产品后端）
- 用户行为打点，自动埋点（产品前端）
- ETL，数据分析，挖掘（数据、算法）
- 训练数据标注
- AB 测试，Dashboard

有自己的研究方向，尤其到了今天这个时候，他可能有自己的计划，并不一定愿意花时间在小公司这种已经成熟的应用上面，这就不是每个小公司都能承担的起的，第三招个大牛并不代表一段时间内业务会有很大的提升，因为人工智能需要大量工程师协同完成。现在学习能力和工程能力强的团队，能够把已经验证过的算法在快速迭代中优化，能够实现自己的场景，这样的公司反而有比较大的优势。从平台到模型开发、到开发人员的成本，这些成本都在降低，由于这些成本降低，更多的人会进来，促使这些成本加速降低，这点是现在的小公司和初创的公司可以很好利用的。

算法应用和算法团队是整个人工智能应用的冰山一角，我们需要动用整个公司研发团队的力量来支持人工智能的落地。

首先是核心业务数据的设计，对我们的电商来说刚才说的归类商品的品类系统，这是非常重要的，以及这些系统的整理迭代，保证数据的清洁。我们内部想要做这样一件事情，因为我们的搜索经常会被一些不太好的数据干扰，我们需要设计一套系统，让我们内部的人针对我们的搜索引擎做一些 SEO（搜索引擎的优化），帮助我们清理业务的核心数据，并且让他们负责地去输入新的业务数据。

第二在前端对用户行为打点，甚至为了提高效率，要做到自动埋点，比如用户反馈行为，这个信息在整个机器学习应用中非常重要，信息的收集需要产品前端和我们配合，把打点做好。数据收集之后的分析和挖掘是

数据、算法团队需要帮我们做的。下一点是训练数据的标注，这是算法团队需要做的事情。当我们需要大量做实验时，AB 测试的框架，实验上线后所有的指标是不是能够反映到 Dashboard 里，这都是需要非常多的人去投入的。

## 结语

- 人工智能正加速从学术走向日常
- 更多的公司，更多的工程师实践机器学习
- 期待看到更多激动人心的智能应用

因为成本的降低人工智能正在加速从学术走向日常，希望能看到更多的公司和工程师进入实践机器学习的领域，期待看到更多激动人心的智能应用。

## 作者介绍

**赵晓萌**，小红书算法架构师。曾任微软 Bing 必应搜索 Core Ranking Team 的 program manager，分管用户点击模型在 dynamic ranking 中的应用。曾任 Twitter Performance Ads, Tech Lead，负责移动相关的精准广告投放。现任小红书算法架构师，负责机器学习应用。

# 分析海量视频中的违规内容，七牛如何构建弹性深度学习计算平台

作者 彭垚



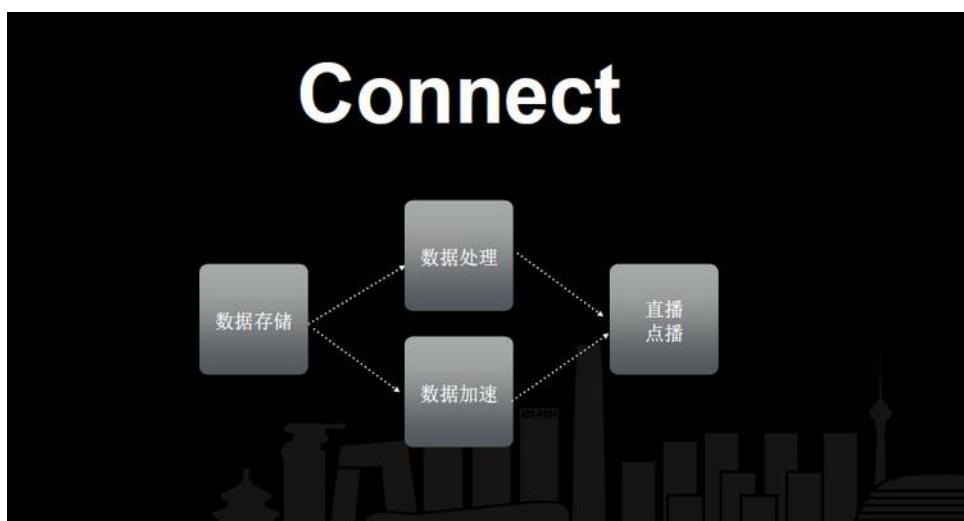
我是七牛云人工智能实验室的负责人彭垚，我们七牛云人工实验室是去年 6 月份创立的。我先介绍一下七牛为什么在这个时间点开始做人工智能。

今天演讲的主要内容包括人工智能实验室的前因后果，现在在做的深度学习主要是机器视觉方面研发的成果和近况，以及深度学习计算平台的框架架构。大家可以在会后跟我深度沟通，因为这一块我们才做了大概不到一年的时间。

七牛是云存储起家，服务移动互联网已经五年的时间了。这几年移动互联网变成了一个富媒体的时代。从社交网站上的图片开始到短视频，

今年短视频又开始复苏，包括去年非常火的直播。我们七牛一直跟着这股风潮在服务我们平台上广大的用户。

前面这五六年我们一直在做一件事情，我们把这件事情统一地叫做一个词“Connect”，就是连接。我们连接主要做的事情，最早做的是数据存储，就是让大家把各自 APP 上用户上传的图像、视频、音频内容存放在七牛云存储上。我们基于云存储又做了一些富媒体的编解码、图像处理和其他数据处理等，之后我们又给大家做了 CDN，使大家得到更好的用户体验，能够更好地访问这些数据、浏览这些数据。



去年我们又给大家提供了直播和点播云。我们一直在做的关键的事情，就是让用户和用户连接起来。那么怎样把用户体验做好，我们这么多年一直在做的事情就是用户体验，这个用户体验体验在什么地方？就是我们把人跟人之间的连接，把基础服务提供给 APP，提供给我们的客户，这是我们这么多年一直在做的事。

后来我们发现每天用户上传的数据非常多，每天用户上传的图像超过 10 亿张，有超过万亿小时的视频在云存储上。我们存储了海量的内容，大部分的存储数据都是图像、视频和音频。

我们问自己，这么多客户在我们的云存储上存了这么多内容，我们接下来该如何给用户提供更好的用户体验。于是我们去问客户需不需要知道这三种内容具体是什么，就是图像、视频、音频的具体内容。你的客户通

过 APP 上传，每天在浏览，在分享的内容到底是什么，所以我们就开始思考这个问题，然后发现有这么几件事情，其实他们自己已经在做了。



第一件事是很多 APP 有自己专门的内容审核团队，审核客户上传的东西内容是不是合法，有没有涉黄、涉及反政府的信息在传播。

其次，对这些图像、视频、音频的内容，已经有客户有自己的数据运营团队去分析 APP 客户上传的具体内容，可能用抽样的方法，或者机器学习的方法去分析。

内容分析说起来很简单，就是你上传一个图像具体是什么，但是实际上又很复杂，很难说清楚，内容是什么？

比如我拿出一张图片，每个人描述一张图片里面有什么东西，这个叫图片描述。每个人的描述可能都不一样。主要问题是我们在看到东西，听到东西的时候，我们做出的反应，做出的事情跟我们大脑处理的任务相关。所以内容总结起来其实是跟内容最后的目的相关的。

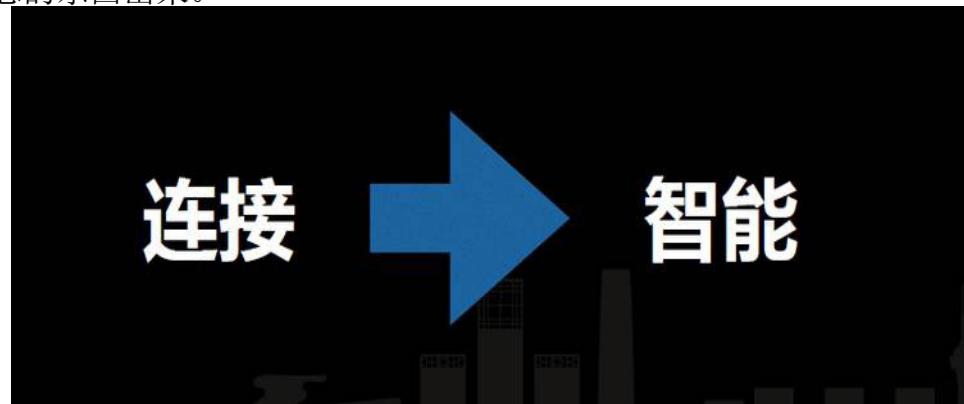
我们怎么理解这个内容。首先我们可以去把内容解析成很多目的。第一个是分类，分类是基本内容的解析，比如判别这个图片是不是黄色图片。第二个就是检测，比如检测这个视频里面有没有人脸，这些人脸是谁，里面出现了哪些物体，有没有车，车的型号是什么。还有分割，比如说一个画面里面，这个人的形状是怎样的，他跟背景的界限在哪里，这就是一个很简单的分割问题。然后就是跟踪，比如说一个视频中，我们有人脸在走

动，这就是一个跟踪问题。以及一个视频的描述，一个视频每一段里出现了什么事件，每一段里面有多少人物，这些是一个描述。还有搜索，我看了很多图片之后搜关键的信息出来，再之上可能就是分析，还可能做很多的处理。



其实我们去解读 content，最关键的是内容的目的。我们首先会去看对这些内容需要做哪一些事情，我罗列的就是我们经常做的一些项目的相关内容。

我们从去年开始做了一个很大的转变，我们从连接基础服务的提供商，变成去给客户做智能的提供商，也就是说我们希望帮助客户去做智能，去提供一些智能的解决方案，让客户去做一些更智能、更互动性的，更了解自己内容的一些行为。这就是我们提出要把我们的连接生意做成智能的生意。我们现在有海量的数据，而图像和视频的泛化能力是很强的，我们通过平台上的数据跟用户一起共建，一起训练，就可以得到很多有价值、有意思的东西出来。



现在这个时代经常提人工智能，智能这个词语到底是什么意思？其实很久以前图灵机的时候就已经有智能这件事情了，而到现在大家对智能还没有一个准确真实的答案，怎么样算是一个智能，我个人理解的智能是类似于人一样直觉型地思考反馈很多的东西，这可能就是最基本初级的智能。

其实我们现在做人工智能，要具备泛化的能力。比如要用深度学习解



决像机器视觉这样的问题，首先要解决的最重要的两个问题，一个是大数据的问题，还有一个就是深度学习，也就是机器学习算法的问题。每天我们平台上传处理的图像非常多，可能超过 10 亿，我们不可能把所有的上传图像都拿来学习一次，所以大数据的处理能力非常重要。其次就是我们不可能把所有图像都拿去人工做标注，这个工程量非常大。所以我们会结合很多算法做一些半监督的机器学习，再加上标注，再加上深度的神经网络取得最终的结果。也就是说人工智能实验室在解决两个问题：一个是大数据，另外一个是机器学习的问题。

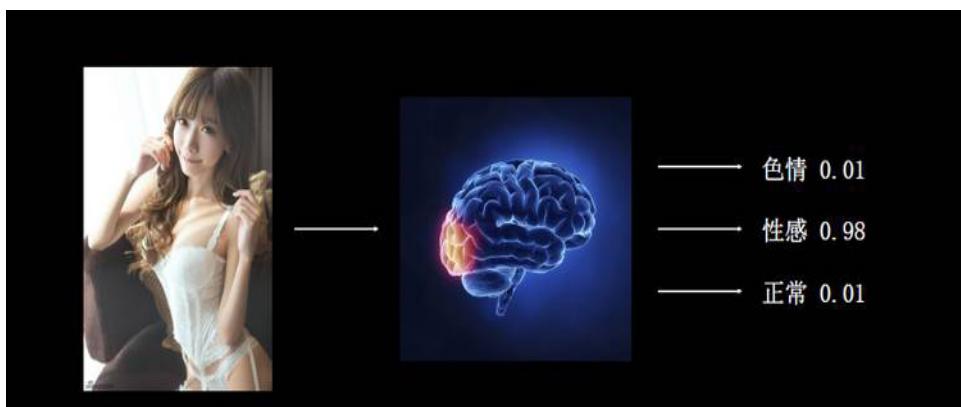
图中是我们去年成立的实验室 Ataraxia AI Lab。这个名称来源于一



个古希腊的哲学学派，这个学派是个怀疑论的，Ataraxia 是指人对世界的认知是有缺陷的，你永远不可能了解事物的本质，就像我刚才提出来智能这个问题，其实每一个阶段都有人提出智能的含义，图灵认为智能能用机器制作出来，后面有希尔乐等等人反驳了他，其实智能这些东西跟用机器模仿出来的东西完全不一样。

我们做人工智能、做认知这件事情，我们一直在质疑自己，最终想达到的境界就是 Ataraxia 的境界，一直在不停地追求永远达不到的一个境界，这个就是古希腊文翻译出来的一个哲学的单词。

接下来介绍一些我们之前做的事情。我们做的第一件事情就是把一张图片扔进 CNN 的网络，识别这张照片是色情、性感还是正常的。如果有搞机器视觉的朋友就会觉得这是一个非常常见、非常基础的一个分类问题。但是这个分类问题，它其实不那么好解决。因为会有各种各样的图像表述它是色情的，是性感的，所以模型需要去学习、去标注的内容非常多。我们在去年刚建实验室的时候，有很多实习生在实验室每天标注这些色情内容。当然现在已经少了，因为我们每天会有半监督打标的迭代过程，我们一直在优化鉴别色情暴恐的系统。如果大家有兴趣可以去我们实验室看一下，我们一直固定有人在做图像标注。包括有一些兼职的，在学校里面在帮我们做的，我们自己做了一套网络上的标注系统。



我们线上已经有超过 700 万的样本一直在滚动，每天新增的数据就有一两万，一直往样本中添加，还需要做大量的评估，以及过滤掉大量不

需要打标学习的数据。我们对算法的要求已经固化了，算法基本停止了迭代，但是数据还在不停地迭代，鉴黄项目是一个数据量很大，要滚动起来自动迭代的一个项目。

第二个是识别图片具体内容的项目，就是人脸识别。需要对人脸提取特征，然后对大量的图片进行人脸聚类。比如说标注它是 id1 类的人，可以做一些特征的分类，像戴不戴眼镜、年龄、性别、颜值。后面就是场景识别，场景识别现在支持 300 多类场景的识别。户外的场景识别准确率非常高，室内会有很多误判，比如说教室和办公室等等。因为如果学习一个单一任务，可能会有疏漏，比如如果一张图片里有学生，场景是教室的概率就会非常高，成为 Office 的概率就会非常低。现在基本的分类算法，如果要提升背景的准确率，图像里面的人物内容都要结合学习。

还有就是审查，我们能够审查判定图像内容是非色情、非暴力、很健康的。



还有一些跟图像描述相关，就是通过 CNN 提取特征，通过 RNN 做图像和视频描述相关的内容，比如我们在跟广电相关的一部分工程上做尝试，对一些球赛做分析，会学习很多名人的人脸，大概有 5000 多类名人的人脸。我们一直在搜集、迭代这些数据库以及对球赛的动作去做学习和描述，这就是我前面提到的描述。

第三个就是视频，视频的识别涉及到场景的概念。什么叫场景？你可以想象我们在拍电影，大家就会非常容易理解镜头，就是 Shot 这种概念。比如我们在拍摄这几个人在说话做事情，突然切了一个场景大家在户外开



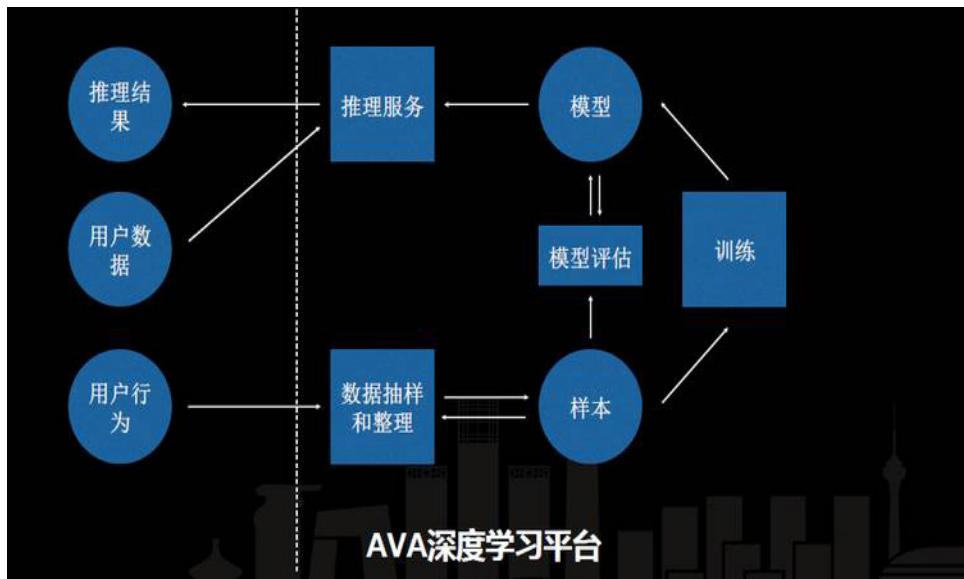
摩托，这就是场景的变化。它最根本的是对人脸和物体的跟踪，如果突然发现这些东西没了，那就说明场景切换了，这就是基本的场景识别。我们会把视频根据场景先切开，切开以后会把场景中的事件 1、事件 2 列出来，比如说有人在打棒球，有人在开摩托车这样的事件罗列出来。



之后会检测视频里的人脸，做一些人脸的识别加跟踪。视频是每帧图像之上持续的表述，一般会用 CNN 识别图像特征，图像特征之上会用 RNN 网络做时序学习。

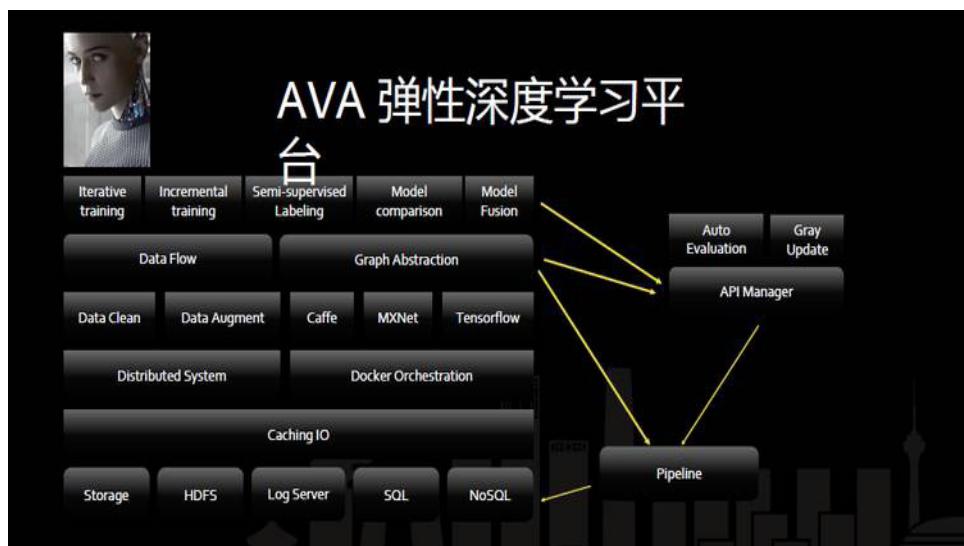
前面我提到，今天我的主题是讲深度学习和计算平台，接下来跟大家介绍一下计算平台。计算平台同时在解决两个问题，一个是大数据，一个是深度学习算法，抽象来讲计算平台在做一些什么事情呢。首先有一些用户的行为，这些用户的行为会产生很多上传的图像、视频，包括调整相册这些动作，会告诉抽样整理模块，这些图像标注的信息是什么，或者说系统需要搜集这些信息，而抽样整理模块是分布式的富媒体处理模块，会不停地处理抽样和调整的工作，抽样调整完了之后就可以生成目标样本集。

通过抽样整理不停地迭代整个样本，得到这个样本集之后我们就会继续上传到训练集群里。



训练集群完成后会生成线上的模型，我们的样本集也会有一部分持续投到模型评估的模块里，模型会根据一套 API 生成器自动上线到推理服务上。最后利用用户数据去访问推理服务，会得到相应的推理结果，这是比较简单的 AVA 的一个基本逻辑。

这是我们现在的 AVA 整体的架构图。最底层通过七牛云存储了大量线上的图像、视频、音频的数据，这些数据会通过统一的 IO 接口做统一读



写管理，这之上我们有两套系统。一套系统专门用于数据抽样和数据整理。Data Flow 里会做数据的清洗，以及数据的放大，数据放大是指对图像的二次加工，通过把同一张图像做裁剪、旋转等操作增加数据样本。

另一套是基于 Docker 的编排系统，这套编排系统与 Kubernetes 有点像，也是七牛很早之前在做的事情，和 Kubernetes 出来的时间差不多，七牛很多线上的图像处理一直在用。Docker 编排系统支撑的是 DataFlow 大数据分布式系统以及支撑了 Caffe、MXnet 、TensorFlow 三个主要的机器视觉框架。模型训练结束以后会自动通过 API Manager 的自动代码生成器生成线上的 Inference API，Inference API 生成自动评估模块以及做自动化的灰度发布。

最上一层我们基于下面的基础系统做了几个 APP 应用系统，第一个就是自动迭代的训练系统，这套自动迭代系统主要用于持续学习的项目。我们每天会有很多新增数据投到训练数据池中。我们会定期地，比如到上一个模型迭代周期结束之后，把这些数据自动化投进训练池中重新清洗，清洗之后重新训练，这就是迭代系统。

还有一个自增长数据集系统。比如鉴黄系统，针对每天都会增长的数据，我们会采取流式的深度学习训练模式，系统在某一个 snapshot 的时候引进一个新的数据集，然后会用这批新的数据再去学习。这个系统可以解决一些对训练出模型频率要求比较高的问题，比如最近比较热的黄色信息。

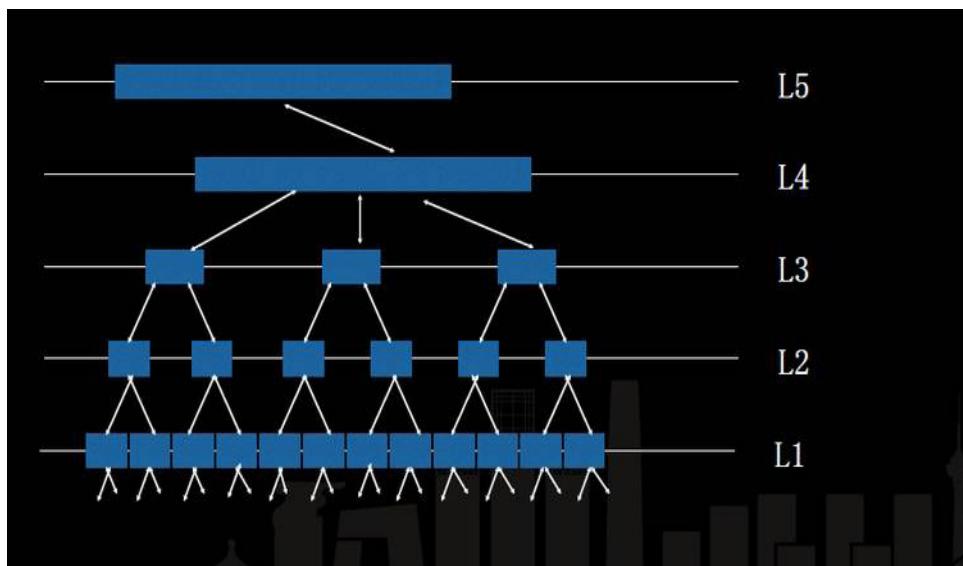
另外是做了一个半监督打标的系统，这套系统跟我们的打标软件连接。我们用一些轻量的模型，甚至 svm 这种小的分类器先做自动的图像预标注，跟我们的分类器的中心做比较，比较出来之后，拿出一部分的数据再去学习，投入到我们应该要学习的样本中。这其实也是模型融合的一点。

我们做了大量的模型融合。我们会选不同的 CNN 网络，在一些大一点的和小一点的不同的情况下做模型的融合。模型融合确实比较有效，但是它比较费资源，费人力，所以我们就把这个单独做成一个 APP 自动化地运行，有时候在一些特定的场景还是需要模型融合的方法才能把准确率

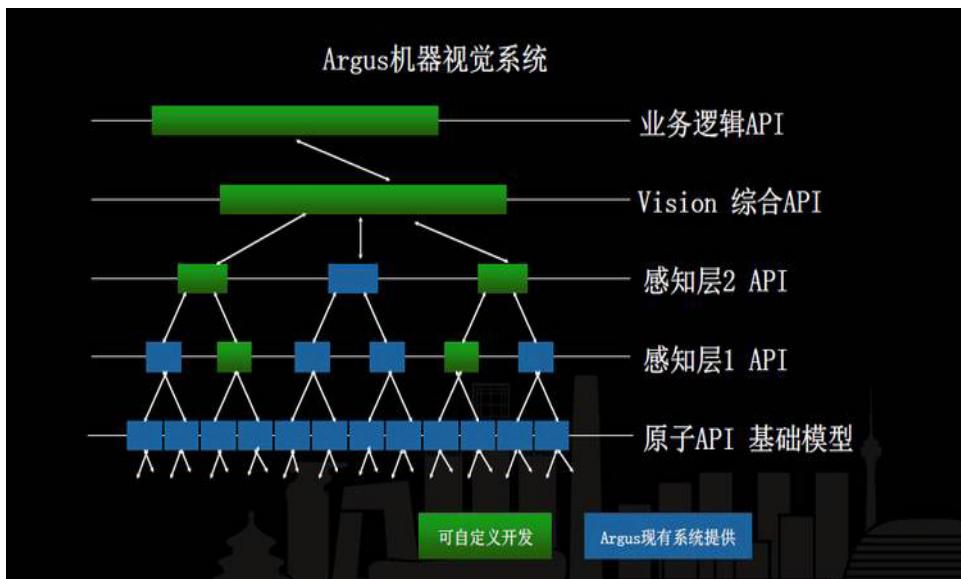
优化到能达到商用。

训练的过程还有一块是 Pipeline，这个 Pipeline 其实是对日志做迭代收集，做 transform，到不同的存储结构上，这些可能是一些图像的标签，视频的标签这些内容，这就是我们整体的 AVA 平台的架构。

这里我没有提到 multi task。实际上它的处理比较复杂，不像鉴黄那么简单，大部分问题都不会这么简单。举个特别简单的例子，比如说人脸聚类，也有三个小模型，首先要检测到图像里人脸的位置，其次要用机器学习抽取图片的人脸特征，之后利用这些特征做聚类。至少需要三个模型。



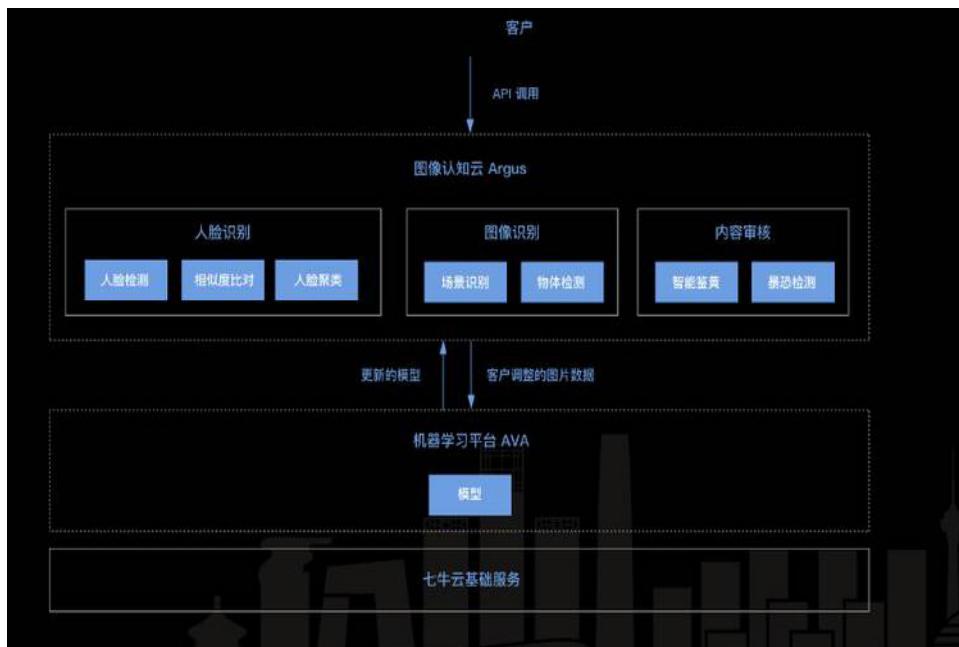
这个其实和人脑也很像，人脑解决问题是像这样的图。图中有 L1 到 L5，大脑皮层每一层都是这么处理问题的。信息从最底层扔给几个基础的模型，去做一些抽象、完成一些任务，到第二层的时候再去解决更高维的一些任务，比如像聚类这样感知型的任务，再上面做一些更具体的任务，比如搜索、判别这类事情。最高维就是在做一些预警，一些业务层的事情。已有的 AVA 只能解决单一的问题，不能满足整个人工智能的设计框架。所以我们做了一套 Argus 系统，实际上就是 API 的整体网状管理系统，它支持 Pipeline，也支持并行处理。可以直接用 Pipeline 的语义解决这种事情。



Argus 系统最底层是通过 AVA 训练出来的原子 API，有了原子的 API 之后上层是感知层，感知层会做基于原子 API 的抽象做一些复杂任务，比如聚类。再之上是一些高级的任务，最后是一些与视觉相关的综合 API，再往上是业务逻辑大数据分析，在 Vision 层我已经不管了，我把这个东西扔到抽象层结构化数据，或者说 vision 跟语言相关的加了一些 RNN 把语意描述出来之后就扔给业务逻辑处理了。所以现在 API 的 framework 整体设计成这个样子。设计成这套系统后，有很多是我们新研发的，Argus 系统现有的是蓝色的，原子 API 是通过 AVA 训练出来的，AVA 还没有公开，原子的基础视觉 API 都是我们自己研发的。我们希望之后跟大家公开用 AVA 训练出来的特定的一些识别模型。我们也在尝试性地找一些想做这个事情的长期合作伙伴。

上面业务层的 API 客户可以独立开发应用，包括像感知层、综合的整体业务逻辑的 API，直接可以通过我们 user-defined 图像处理模块，直接写一些简单的 docker 处理镜像 load 进来参与到 Argus 的机器视觉系统里。也就是说高层的业务层或者说智能的大数据分析能力是开放给客户的。

这是我们现在整体上 Argus 的图像认知，有很多基础服务，包括一些业务层的比如人脸检测、相似度比对、人脸聚类、鉴黄、暴恐，这些基



础的模型之下，有一个一直在迭代运算的 AVA 深度学习平台，它一直不停地产出一些基础的原子 API 给 Argus 系统，Argus 系统跟客户走得更近，让客户可以自己在 Argus 上编 Docker 镜像，load 上来，一起完成智能的任务。

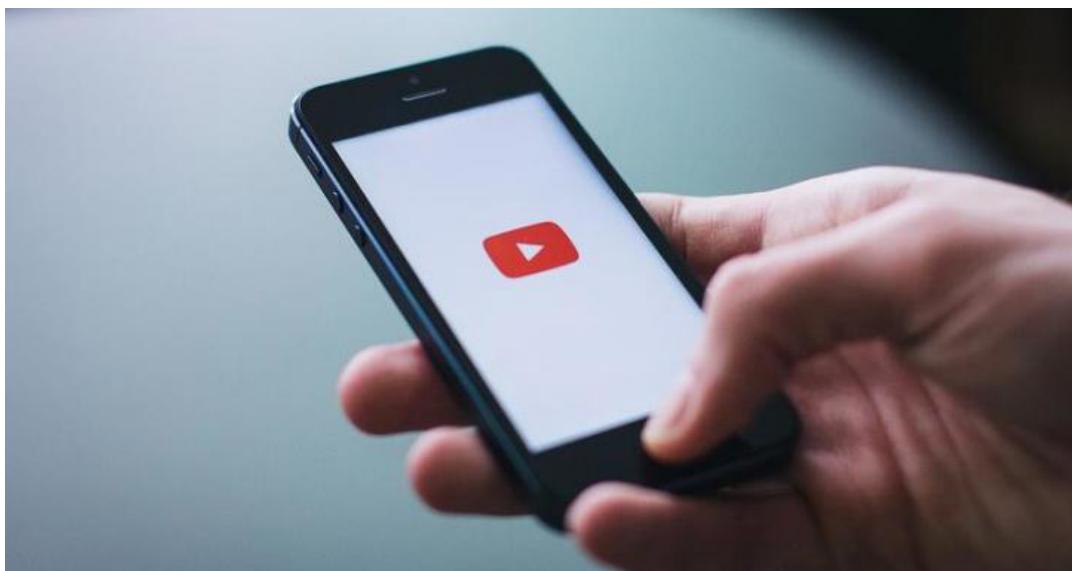
最后我想说 time to be an AI company，希望七牛之前的合作伙伴以及将来大家愿意跟我一起合作，可以一起在 Argus 系统上添砖加瓦，做很多有意思的机器视觉系统。

## 讲师介绍

**彭垚**，七牛云技术总监，人工智能实验室发起人和负责人，主导了七牛云人工智能和机器学习云的架构和发展，在分布式计算存储，富媒体海量数据分析与深度学习领域有超过 10 年的产品研发经验，曾担任 IBM 系统与科技实验室研发架构和管理工作多年，在美国、法国发表数篇专业领域发明专利。

# YouTube 整合 Google Brain 推荐算法， 视频播放量提升 20 倍

作者 Casey Newton 译者 NER



去年某一天，当我在玩一款名为《冤罪杀机 2》的电子游戏时，我在 YouTube 上常规搜索了一下，看看这个游戏中一个棘手的部分怎么通关。像往常一样，我找到了回答我这个问题的一个视频。但当我下次打开 YouTube 的时候，网站却给我推荐了更多更复杂的关于如何玩这个游戏的视频：玩家如何玩这个游戏而不被敌人发现的视频剪辑；玩家用高超的方法杀死每一个敌人的视频剪辑；采访游戏制作者的视频；精彩绝伦的讽刺评论。我只是去 YouTube 搜索一个问题的答案，它却像我展示了一个全新的宇宙。

后来，我发现自己每天都会访问 YouTube 好几次。大多数时候，我

打开这个网站都没有什么特别的目的，我已经习惯了被动地让网站自动给我推荐点什么我可能喜欢的东西。一月份的时候，我开始痴迷于一个叫 Pinegrove 的民谣乐队，几个礼拜的时间里，YouTube 给我推荐了几乎所有上传到它服务器的这个乐队的现场演出视频。当我春天住进一个新公寓的时候，开始越来越多地做饭，在我搜索了一次如何做意大利面包沙拉之后，YouTube 很快就让我认识了它的家庭主厨阵营：Byron Talbott、Serious Eats 频道的 J. Kenji López-Alt，以及 Tasty 等等。

YouTube 总是很有用，它从 2005 年创办以来，就成了互联网的支柱之一。但在过去的几年里，对我来说，YouTube 变得出奇地棒。它开始极端准确地预测出我可能感兴趣的视频剪辑是什么，比它过去所做的要强得多。到底是什么发生了改变？

在过去的 12 年时间里，YouTube 已经把自己从一个搜索驱动的网站转变成了一个为自己目的服务的网站。要到达自己的目的地，它需要成百上千种尝试、大量的重新设计，以及在人工智能方面的巨大飞跃。但真正让 YouTube 提升的还是它朝 Feed 的方向进化。

现在想起来很难记得了，但最开始，YouTube 只是一个基础设施。它提供了一种方便地把视频植入其他网站的方法，通过这种方法，你最有可能看到这些视频。随着网站壮大，YouTube 成了一个寻找过期电视节目剪辑的地方，它会随时跟进最新的午夜喜剧，用来观看最近的病毒式传播视频。和维基百科一样，YouTube 或许也是互联网上最臭名昭著的地方。你的同事在茶水间提了一句 Harlem Shake 搞笑视频，然后你就登录 YouTube 看了一晚上。

同时，Facebook 发明了我们这个时代标准格式：News Feed，一种根据你的兴趣定制的永不间断的信息流。Feed 占领了整个消费者互联网，从 Tumblr 到 Twitter、Instagram 和 LinkedIn。YouTube 早期像个人定制化的发展很有限，它主要是让用户去订阅频道。这个主意是从电视那里借用的，它拥有混杂的复杂结果。根据 ComScore 的数据，在 2011 年，YouTube 一次大规模的推送取得了一些成功，但人均观看 YouTube 视频的

时间却和之前持平。

“频道”已经不能像以往一样主宰 YouTube 了。现在，打开你手机上的 YouTube，你会发现“频道”被隐藏在了一个单独的菜单里。取而代之，这个应用会根据你的兴趣量身定制一些混合的视频，并以 Feed 的形式呈现给你。这些视频中当然也有你订阅过的频道的视频，但其中也包括和你以前看过的视频相关的视频。

这就是为什么，当直接搜索了关于《冤罪杀机》视频之后，我就开始看那些推荐的通关视频和刻薄评论。YouTube 开发了各种工具，让它的推荐不仅是个人定制化的，还是高度准确的，这些就最终提升了整个网站的观看时间。

YouTube 推荐的技术带头人 Jim McFadden 说：“我们知道，人们来 YouTube 是来找他们想要的东西，但我们还希望，在人们不知道他们想要找什么的时候，同样满足他们的需求。” McFadden 从 2011 年就加入了这家公司。

我第一次拜访 YouTube 也是在 2011 年，就在 McFadden 加入这家公司几个月之后。就是那时候，YouTube 开始让用户花更多时间来看它们的视频，现在，这也依然是 YouTube 的核心目标。在那个时候，事情还进展得并不是非常顺利。McFadden 说：“YouTube.com 作为一个主页，它并没有带来大量的娱乐性。我们就想，好吧，那我们就把让它具备大量娱乐性作为转型目的。”

这家公司什么事情都尝试了一下：它为顶级的创作者购买了专业的摄像设备，发起了“leanback”功能，它可以在你观看视频的时候，自动排列新的视频给你。YouTube 重新设计了它的主页，以此强调订阅频道，而不是看单独的视频。

每个用户观看的时长依然持平，但有一个变化出现了，那就是它们的推荐算法并非基于有多少人点击了视频，而是基于人们花了多长时间来观看这些视频，正是这个变化驱动了接下来那个春天发生的剧变。

几乎是一夜之间，那些受益于误导性标题和视频略缩图的视频创作者

就看到他们的观看数字急转直下。质量较高的视频往往和更长的观看时间相关，它们开始急剧上涨。在接下来的三年里，YouTube 的观看时长每年都增长了 50%。

我订阅了一些频道，并且自认为是个 YouTube 的普通用户。但是它要成为一个一天内多次访问的目的地，还需要一系列的新工具，那些在过去 18 个月内成为可能的工具。

当我这个月拜访 YouTube 办公室的实话，McFadden 向我介绍了 YouTube 精确推荐视频的根源：Google Brain，它是 YouTube 的母公司 Google 的人工智能部门，YouTube 从 2015 年开始使用它。Google Brain 并不是 YouTube 第一次尝试使用 AI，YouTube 此前曾把 Google 建立的 Sibyl 系统中的机器学习技术应用到推荐算法中。然而，Google Brain 引入了一种见无监督学习的技术，它的算法能在不同的输入中寻找到联系，这是软件工程师们从未曾想过的。

McFadden 说：“最关键的一点是它能够普及推广，在此之前，如果我看了一个喜剧的视频，推荐算法会说，又有一个人喜欢了这个视频。但是 Google Brain 的模型会识别出类似于此的其他喜剧，但又不是完全相同，它们拥有更毗邻的关系。它能够识别出不那么明显但相似的模式。”

举个例子，一个 Google Brain 算法会给一个移动应用用户推荐短小的视频，但给 YouTube TV 的用户推荐长一些的视频。它猜测，根据平台的不同推荐不同长度的视频会最终提升观看时长，它是正确的。YouTube 在 2016 年实施了 190 多个类似这样的改变，而今年计划要做出 300 个改变。YouTube 发现小组的产品经理 Todd Beaupre 说：“现实就是，它是随时间推移累积起来的一大批微小的改进。对每一个改进来说，你都要尝试 10 件事最终实施一件事。”

Google Brain 的算法比 YouTube 之前的算法要更快。公司表示，在过去的几年里，一个用户行为要经过好几天才会被整合进未来的视频推荐中，这样就很难识别出趋势。Beaupre 说：“如果我们希望把用户吸引过来了解当下在发生什么，我们就必须修补这个问题，现在，延迟被设定在

几分钟或几小时的时间里，而不是几天。”

把 Google Brain 整合到 YouTube 中有一个重要的影响：人们在 YouTube 上看视频的时间，现在有超过 70% 都来自 YouTube 的推荐算法。每一天，YouTube 会推荐两亿个不同的视频给用户，涉及语言有 76 种。和三年前相比，人们在 YouTube 主页上看视频的总时长增长了 20 倍。

这也基本上和我个人的用户行为相符合。几年前，我基本上通常只在午饭休息的时候访问 YouTube 主页，一边吃饭一边看点什么。但他们的推荐实在太好了，我开始用更多的空闲时间看视频。这礼拜，我在 PS 4 上登录了 YouTube，这样我就能用我最大的屏幕来看它推荐的视频了。

这就是一个真正强大的个人定制化 Feed。对我来说，令人惊讶的是，YouTube 对我数字生活的改变比其他任何东西都要强大。Facebook 的 Feed 是基于你的朋友们发了什么东西，还有你喜欢的主页发了什么内容。知道谁订婚了或者生小孩了很有用处，但除了这些里程碑意义的事件外，我从朋友们发表的内容中没找到什么乐趣。Twitter 会给你看你关注的人们的推文，还有这些人选择转发的东西。作为一个记者，我必须依赖于 Twitter，即使有些时候我的时间线真是看似没有尽头，充满了焦虑的呐喊。

每一个 Feed 都有长度限制，虽然 2017 年取消了这个限制。在 Twitter 上，不管你关注谁，关于政治的争论永远主宰者讨论。Facebook 对于“事件”和“团体”这些功能的短暂热情让 Feed 每周都以令人震惊的方式发生变化，这让我感觉和每一个朋友的连接都更少了。（以图片为重的 Instagram 看起来就像一片绿洲，也难怪这个应用还在如此迅速地增长。）

Facebook、Twitter 和 Instagram，看起来那些 Feed 都要求人们不断地为它们表演点什么。而 YouTube 很显然是表演驱动的，但很少一部分用户会给它上传视频，而且 YouTube 也从来没有强迫用户去上传。YouTube 可以被人们被动地享受着，就像它那么努力地去尝试取代的电视频道所做的一样。在我们这样一个疯狂的年代，能不被询问我们对某个新闻的看法，这真是让人感到平静啊。

YouTube 对你可能喜欢的相关视频的强调，意味着它的 Feed 和其他 Feed 相比更宽广，更具有好奇心。它越是寻求不同的内容，就越让人觉得它在逃离其他 Feed 的模式。在一个黑暗的时代，我更倾向于选择 YouTube 的逃避主义。

在 2013 年，《大西洋月刊》上有一篇文章，在那篇文章里，Alexis Madrigal 假设我们所知的 Feed 有其顶峰。他认为，未来会属于有限的经历：电子邮件的 newsletter、Medium 的合集、10 集长度的 Netflix 剧集。毕竟，无穷无尽的信息流内容让人感到疲惫。Madrigal 说：“当媒体宇宙的秩序被彻底击败，自由并不会来填补空白，拥有其自身逻辑的新秩序将会取代旧的秩序。我们发现信息流已经展现出它的强迫性和控制性。更快！更多！更快！更多！更快！更多！”

从那四年之后，YouTube 的方向只说明了 Feed 模式在变得更重要。一种前所未有的视频存储增长，辅以前所未有的个人化定制技术，将会创造出让人难以拒绝的东西。YouTube 现在会调查用户有多喜爱它们推荐的视频，长此以往，调查的结果会让 YouTube 更加智能，从而让更多视频内容被消费。

Beaupre 向我描述了这个过程，说它就像跨越一条鸿沟那样。“有些内容和你已经喜欢的内容有高度的契合，而有些内容会代表着趋势和流行的内容，而在这两者之间，就是充满魔力的地帶。”如果 YouTube 的竞争对手找不到跨越这条鸿沟的方式，它们就会发现竞争举步维艰。

# Uber 的机器学习平台：从打车到外卖，一个平台如何服务数十个团队？

作者 JEREMY & MIKE



提到米开朗基罗你会想到什么？著名的雕塑作品《大卫》？还是梵蒂冈西斯廷大教堂震撼的穹顶画《创世纪》？优步（Uber）的技术团队为机器学习领域带来了“米开朗基罗”——一款机器学习即服务平台，旨在实现机器学习技术民主化，并调整 AI 规模以保证商务人员能够像使用优步叫车服务一样轻松地借机器学习之力满足自身需求。

优步工程技术团队致力于为客户创造无缝化且极具影响力的产品性技术成果。我们正逐步加大对人工智能（简称 AI）与机器学习（简称 ML）的投入，希望借此实现这一发展愿景。在优步，我们在机器学习领域的一大贡献正是 Michelangelo（米开朗基罗）项目——这是一套内部机器学

习即服务平台，旨在实现机器学习技术民主化，并调整 AI 规模以保证商务人员能够像使用优步叫车服务一样轻松地助机器学习之力满足自身需求。

米开朗基罗项目使得优步公司的各内部团队能够以规模化方式无缝构建、部署及运行各类机器学习解决方案。该项目涵盖各项端到端机器学习工作流程，包括数据管理，模型的训练、评估与部署，作出预测并监控预测结果等等。这套系统亦能够支持多种传统机器学习模型、时间序列预测以及深度学习等方案类别。

米开朗基罗项目已经在优步公司的生产环境中运行了约一年时间，并已经成为我们的工程师与数据科学家们的首选机器学习实现系统——目前有数十个团队利用其构建并部署各类模型。事实上，米开朗基罗被部署在多座优步数据中心当中，配合专用硬件并成为我们目前负载强度最高的各项在线服务提供支持。在今天的文章中，我们将介绍米开朗基罗项目、探讨其相关产品用例，同时深入了解这套强大的机器学习即服务系统的整个工作流程。

## 米开朗基罗项目背后的开发动机

在米开朗基罗项目诞生之前，我们面临着一系列实际挑战——包括立足优步公司庞大的运营规模与强度构建并部署各类机器学习模型。尽管数据科学家们一直利用多种工具以建立预测模型（包括 R、scikit-learn 以及各类定制化算法等等），但各工程技术团队仍然需要频繁构建一次性定制化系统以适应这些模型的具体需要。因此，优步公司的机器学习在影响力方面始终受到限制，导致仅有部分数据科学家及工程师能够在短时间内利用多种开源工具完成任务。

具体来讲，当时的优步公司还不具备能够构建可靠、统一且可重复的管道系统来打造并管理规模化数据训练与预测任务的系统方案。在米开朗基罗项目出现之前，数据科学家们根本无法对超出台式计算机承载能力的模型进行训练，不具备标准化结果存储平台，也无法利用简单方式将不同

实验的具体分析结果进行比较。最重要的是，生产部署模型当中不存在确定的路径——在大多数情况下，相关工程技术团队需要制定仅适用于特定项目的定制化服务容器。另外，Scully 等研究人员还发现并记录下一系列机器学习的反模式场景（即不适用机器学习的模式）。

米开朗基罗项目的设计初衷在于对各团队所使用的工作流程及工具进行标准化调整，从而解决上述差距。通过这套端到端系统，全球用户将能够轻松构建并扩展机器学习系统。我们的目标是在解决这些直接问题的同时，打造出一套能够与业务规模同步发展的系统。

在 2016 年年初着手建立米开朗基罗项目时，我们首先需要解决将可扩展模型训练及部署任务引入生产性服务容器的挑战。在此之后，我们专注于构建更出色的系统与特征共享管道。最近，我们又将关注重点进一步转向提升开发人员生产力水平——即如何加快设计思路的第一生产模式以及随后的快速迭代工作。

在接下来的章节内，我们将着眼于一基示例应用，了解米开朗基罗项目如何构建并部署机器学习模型。虽然我们需要强调这里的 UberEATS 仅属于一项具体用例，但米开朗基罗平台同样可以管理其它企业在实际预测场景当中所使用的类似模型方案。

## UberEATS 建立的交付时间预测模型

UberEATS 当中包含多套运行在米开朗基罗平台之上的模型，其功能涵盖外卖交付时间预测、搜索排名、搜索结果自动填充以及餐厅排名等等。交付时间模型预测需要在订单发布之前完成，从而告知用户食物需要多长时间准备、又需要多长时间能够送到您的手中。

估算交付时间（简称 ETD）绝非易事。当 UberEATS 客户下达订单时，餐厅首先需要接受订单并着手准备餐点，而这项任务在订单情况复杂且忙碌饭点期间往往需要更长时间才能办理就绪。接下来，优步公司的派送人员会前往餐厅收取饭菜——具体包括抵达餐厅、找到停车位、走进餐厅拿取食物、回到车里、开车前往客户位置（具体时间取决于路线及交通情况

等因素）、找到停车位，而后将食物交至用户手中。对整个流程的时长作出预测需要将多个复杂的执行阶段纳入考量，而期间每一个阶段的变化都会给最终交付时长造成影响——意味着系统必须重新进行计算。

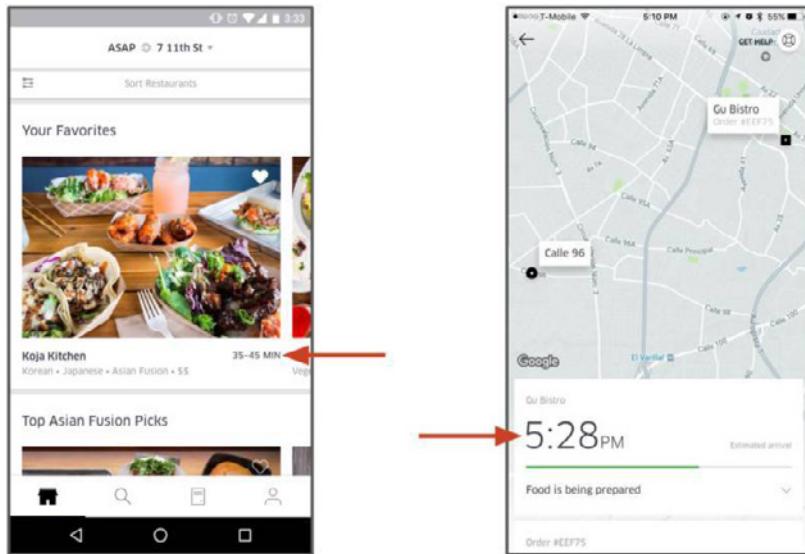


图 1:UberEATS 应用凭借着由米开朗基罗平台承载的多套机器学习模型提供交付时间预测功能。

在米开朗基罗平台上，UberEATS 的数据科学家们利用渐变增强型决策树递归模型来预测这种端到端交付时间。该模型的特征包括请求信息（例如当前时间段、交付位置等）、历史特征（例如过去七天中的平均食物筹备时长等）以及近实时特征计算（例如过去一小时内的平均食物筹备时长）。各模型被部署在不同的优步数据中心之内，而米开朗基罗则负责为模型提供支持容器并经由网络请求对 UberEATS 内的各项微服务加以调用。这些预测结果会在餐厅接到订单之前被首先显示给用户，并在得到其认可后方允许餐厅进行食物的筹备与交付。

## 系统架构

米开朗基罗当中包含有多种开源系统及内部开发组件。其中的主要开源组件包括 HDFS、Spark、Samza、Cassandra、MLLib、XGBoost 以及 TensorFlow。总体而言，我们更倾向于尽可能使用成熟的开源选项，同时

根据需求进行 fork、定制以及贡献——具体包括在开源解决方案不符合用例实际需求时自行构建系统方案。

米开朗基罗平台建立于优步的数据与计算基础设施之上，后者提供一套数据湖以存储优步公司的全部交易与记录数据；Kafka 中间人负责汇总来自全部优步服务的记录信息；此外其中还囊括有 Samza 流计算引擎、受控 Cassandra 集群以及优步自主开发的各类服务配置与部署工具。

在下一章节中，我们将以 UberEATS ETD 模型为例，通过深入探讨米开朗基罗系统内的各层以勾勒出其详尽技术面貌。

## 机器学习工作流

在优步公司，无论具体负责解决哪些挑战——包括分类、回归以及时间序列预测——所有机器学习用例皆采用同样的一般性工作流程。由于这一工作流不受具体实现方式的影响，因此能够轻松进行扩展以支持新的算法类型与框架——例如新型深度学习框架。另外，其同时适用于不同的部署模式，例如在线与离线（特别是车载与手机内用例）预测用例。

在米开朗基罗平台的设计当中，我们希望能够借此提供可扩展、可靠、可重现、易使用且自动化工具选项，最终解决以下这套六步式工作流程：

- 数据管理
- 模型训练
- 模型评估
- 模型部署
- 预测制定
- 预测监控

接下来，我们将详细介绍米开朗基罗架构如何一步步实现上述工作流。

## 数据管理

发现正确特征往往是机器学习工作当中最具难度的部分。我们同时发现，数据管道的构建与管理则通常是机器学习解决方案之内成本最高的部

分。

一套平台应该能够提供标准化工具以构建数据管道，并借此生成特征以及标签数据集，进而利用这些成果实现预测所必需的训练（与重新训练）及纯特征数据集。这些工具应当深入集成至企业的数据湖或者数据仓库当中，同时全面与企业的在线数据服务系统相对接。这类管道必须具备可扩展性与良好的成效表现，拥有数据流与数据质量综合监控能力，同时支持在线与离线两种训练与预测模式。在理想情况下，其应以跨团队共享的方式生成特征，同时减少重复工作量并提高数据质量。另外，这些工具还应提供强有力的保护性约束与控制措施，旨在鼓励并批准用户采用各项最佳实践（例如保证在训练周期与预测周期内使用相同的数据生成 / 筹备流程）。

米开朗基罗平台中的数据管理组件按照在线与离线两类管道进行划分。目前，离线管道主要用于批量模型训练与批量预测任务的馈送；而在在线管道则负责在线低延迟预测任务的馈送（不久之后，我们还将引入在线学习系统）。

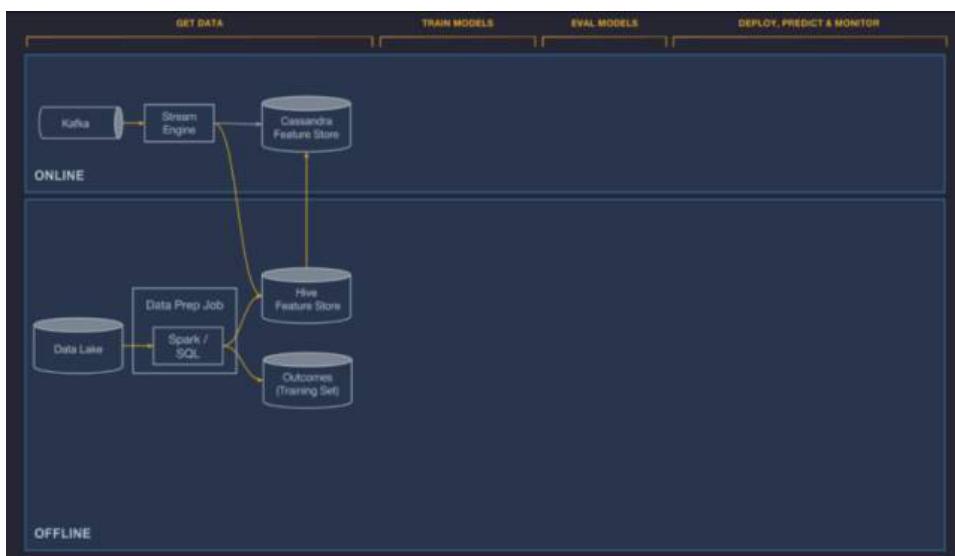


图 2：数据筹备管道将数据推送至特征存储表与训练数据库当中。

另外，我们还向其中添加了一个数据管理层，其作为特征存储机制以帮助各团队对解决机器学习问题所必需的高准确度特征集进行共享、发现

与使用。我们发现，优步目前面临的大多数建模难题都存在着共通性或者说相似性，因此确保各团队能够在不同部门间分享各自项目心得以及特征成果无疑具有重大的实际价值。

## 离线

优步公司的交易与记录数据流被引入一套数据湖内，并可通过 Spark 以及 Hive SQL 计算任务轻松进行访问。我们提供各类容器与调度方案以定期运行特征计算任务，而这些特征将可根据项目实际进行内部保留或者被发布至特征库（Feature Store，详见下文）中进行跨团队共享。而批量处理任务则按计划或者配合触发机制运行，结果将被整合至数据质量监控工具处以快速根据本地或上游代码或数据问题对该管道中的问题进行回溯。

## 在线

以在线方式部署的模型无法访问存储在 HDFS 当中的数据，而且其通常也很难直接面向用于支持优步生产服务的在线数据库进行高成效特征计算（举例来说，我们无法直接查询 UberEATS 订单服务以计算某家餐厅在特定时段之内的平均食物筹备时长）。相反，我们允许对各在线模型所需要的特征进行预计算，并将结果存储在 Cassandra 当中——如此一来，这些结果即可在预测阶段以低延迟方式接受读取。

我们支持两种面向在线特征的计算选项，分别为批量预算算与近实时计算，具体说明如下：

- 批量预算算。第一种计算选项为定期进行批量预算算并从 HDFS 中将历史特征加载至 Cassandra 内。这是一种简单且有效的方法，通常适用于处理历史特征。该系统保证始终利用同样的数据与批量管道进行训练与服务。UberEATS 在计算“餐厅在过去七天内的平均食物筹备时长”等指标时使用的就是这类系统。
- 近实时计算。第二种选项是将相关指标发布至 Kafka 处，而后

运行基于 Samza 的流计算任务。接下来，这些特征将被直接写入 Cassandra 并随后交付及记录至 HDFS 当中，以供后续训练任务使用。与批量系统类似，近实时计算能够确保在训练与服务过程中使用同样的数据集。为了避免冷启动，我们专门利用一款工具面向历史记录运行批量任务，从而实现数据“回填”并生成训练数据。UberEATS 在计算“餐厅过去一小时内的平均食物筹备时长”指标时，使用的就是近实时特征管道。

## 共享式特征库

我们发现建立一套集中式特征库非常重要，这意味着优步公司的各内部团队将能够创建并管理其实际使用的各类特征，并将其与其他团队进行共享。从宏观角度来看，这种作法能够带来以下两种助益：

1. 其允许用户轻松向共享特征库内添加特征，且仅需要为各特征添加少量额外元数据（例如拥有者、描述以及 SLA 等）即可实现内部及特定项目使用需求。
2. 一旦特征被添加至特征库当中，各团队将能够轻松以在线及离线方式对其进行消费——包括引用特征在模型配置当中的简单规范名称。配合这部分信息，该系统将能够协同正确的 HDFS 数据集以实现模型训练或批量预测，并在在线预测场景下从 Cassandra 处获取正确数值。

就目前来讲，我们的特征库中包含约 1000 项适用于机器学习项目加速的特征，而公司内的各个团队仍在不断向其中添加更多新特征。特征库中的各项特征会得到自动计算，且每天进行一次更新。

着眼于未来，我们希望能够进一步探讨构建一套自动化系统的可能性——我们希望利用它对特征库进行完整搜索，从而发现能够解决某一特定预测问题的重要实用特征。

## 特征的选择与转换

一般来讲，由数据管道所生成或者来自客户端服务的特征并不具备正确的格式，且其中部分数值可能未得到有效填充。另外，模型本身可能只需要所提供特征中的一个子集。在某些情况下，模型可能更适合将具体时间戳转换为当天内时段或者当周内某天的表达方式。再有，我们也可能需要对各特征值进行归一化（例如减去平均值并除以标准差）。

为了解决这些问题，我们创造了一种 DSL（即域特定语言）以供建模人员在模型训练及预测阶段对发出的特征进行选择、转换以及组合。该 DSL 作为 Scala 的一个子集实现，且拥有完整的常用功能选项。利用该 DSL，各团队还能够向其中添加自己的用户定义功能。具体包括从当前情景（例如离线模式下的数据管道或者在线模式下的客户端当前请求）或者特征库当中提取特征值的能力。

值得强调的是，DSL 表达式亦属于模型配置中的组成部分，且会在训练阶段及预测阶段被用于帮助确保生成同样的最终发出特征集。

## 模型训练

我们目前能够支持超大规模离线分布式决策树训练、线性与逻辑模型、无监督模型（k-means）、时间序列模型以及深层神经网络。我们会定期添加新的算法以应对客户的实际需求，而这些算法皆由优步公司 AI 实验室以及其他内部研究人员负责开发。另外，我们还允许客户团队通过提交定制化训练、评估及交付代码的方式添加自己的模型类型。这套分布式模型训练系统可向上扩展至支持数十亿项示例，亦可向下收缩至处理小型数据集的快速迭代。

所谓模型配置，是指模型类型、超参数、数据源引用、特征 DSL 表达式以及计算资源要求（包括机器数量、内存容量以及是否使用 GPU 等）等指标。我们利用模型配置对训练任务进行配置，并将其运行在 YARN 或者 Mesos 集群之上。

在模型训练完成之后，我们会计算其成效指标（例如 ROC 曲线以及 PR 曲线）并将结果合并至模型评估报告当中。在训练末尾，我们会将初

始配置、学习到的参数以及评估报告保存至模型库当中以供分析与部署。

除了训练个别模型之外，米开朗基罗还支持面向全部模型类型以及分区模型的超参数搜索。利用分区模型，我们能够自动根据来自用户的配置进行训练数据划分，而后立足各个分区训练对应的单一模型，并在必要时将其回滚至母模型状态（例如为每座城市训练一套模型，并在城市级模型的准确度无法达标时将其回滚至国家级模型版本）。

训练任务可通过一套 Web UI 或者 API 进行配置与管理，我们一般选择使用 Jupyter notebook。也有不少团队利用 API 与工作流工具按计划对其模型进行定期重新训练。

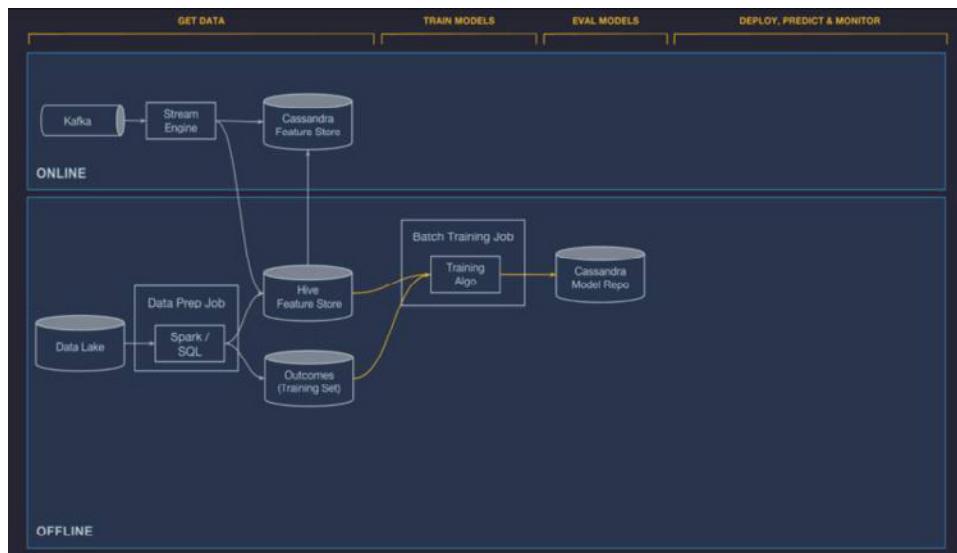


图 3：利用特征库与训练数据库内的数据集执行模型训练任务，而后将其推送至模型库当中。

## 模型评估

作为方法探索流程中的组成部分，我们通过模型训练以发现特征集、算法以及超参数等能够为特定问题建立最佳模型的因素。在为特定用例找到最佳模型之前，我们往往需要对数百套模型进行训练。尽管其中绝大多数模型不会被应用于最终生产环境，但这些模型的实际表现会给工程师们带来启发，帮助他们借此找到能够实现最佳模型成效的配置方案。追踪这

些已训练模型（例如谁在何时对其进行训练、配合怎样的数据集、使用怎样的超参数，等等）、对其进行评估并将结果作出相互比较往往难度极大，特别是考虑到模型自身的庞大数量及其给米开朗基罗平台带来的可观数值规模。

对于每一套在米开朗基罗平台上进行训练的模型，我们都会在 Cassandra 中的模型库内为其保存一个版本控制对象，其主要负责记录：

- 谁训练了该模型
- 训练任务的开始与结束时间
- 完整模型配置（所使用特征与超参数值等）
- 引用的训练与测试数据集
- 每项特征的分布与相对重要性
- 模型准确度指标
- 每种模型类型的标准图表与图形（例如 ROC 曲线、PR 曲线以及二进制分类器的混淆矩阵）
- 该模型的完整学习参数
- 模型可视化摘要统计

用户可以通过 Web UI 以及编程化 API 轻松访问这些信息，并利用其检查单一模型中的细节并将其与一套或者多套模型加以比较。

## 模型准确度报告

回归模型的模型准确度报告将显示出标准准确度指标与相关图表。分



类模型则将显示出不同的信息集合，具体如图 4 与图 5 所示：

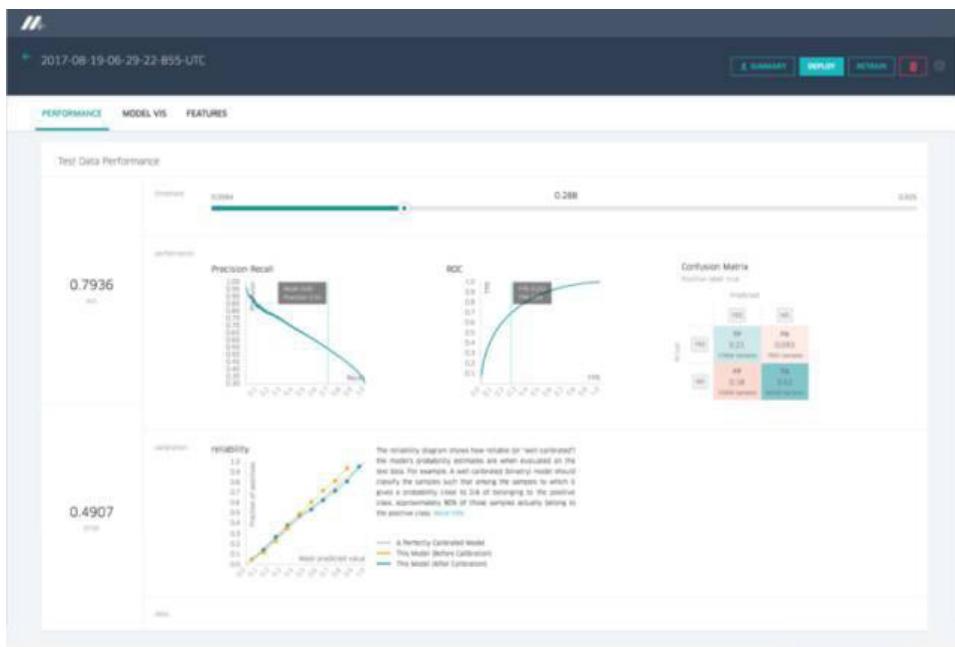


图 4：回归模型报告中显示出与回归相关的成效指标。

图 5：二进制分类成效报告显示与分类相关的成效指标。

## 决策树可视化

对于重要的模型类型，我们亦提供更为复杂的可视化工具以帮助建模者了解模型的行为原理，并在必要时对其进行调试。在决策树模型用例当中，我们允许用户浏览其中的每一单独树状结构，查看其对于整体模型的相对重要性、各分割点、每项特征对于特定树状结构的相对重要性，以及各分割点的数据分布情况等。用户还可以指定特征值，并以可视化方式查看决策树因此形成的触发结果、各树状结构的预测结论以及模型的整体预测结果，具体如图 6 所示。

## 特征报告

米开朗基罗平台提供的特征报告功能可根据对整体模型的重要性、部分依赖关系散点以及分布直方图按序显示每一项特征。如果选定两项特征，

用户则可了解二者间的双向交互依赖关系，具体如下图所示：

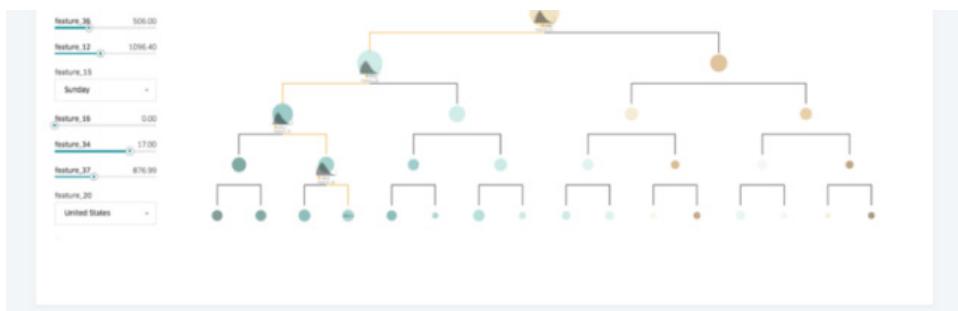


图 6：树状模型可配合强大的树图可视化机制进行探索。



图 7：通过特征报告查看两种特征、其对模型的影响以及二者间的交互关系。

## 模型部署

米开朗基罗平台以端到端方式通过 UI 或 API 支持受控模型部署流程，且模型可通过三种模式进行部署。

### 离线部署

模型被部署至一套离线容器当中，同时立足 Spark 任务运行以按需或者根据重复计划生成批量预测结果。

### 在线部署

模型被部署至一套在线预测服务集群当中（一般处于负载均衡器之后且包含数百台设备），其中各客户端能够以网络 RPC 调用的方式发送单一或者批量预测请求。

## 库部署

我们计划提供新的选项，允许模型以服务容器的方式进行部署，并将其作为库嵌入至其它服务内且通过 Java API 进行调用。（图八并未包含这种情况，但其工作原理类似于在线部署。）

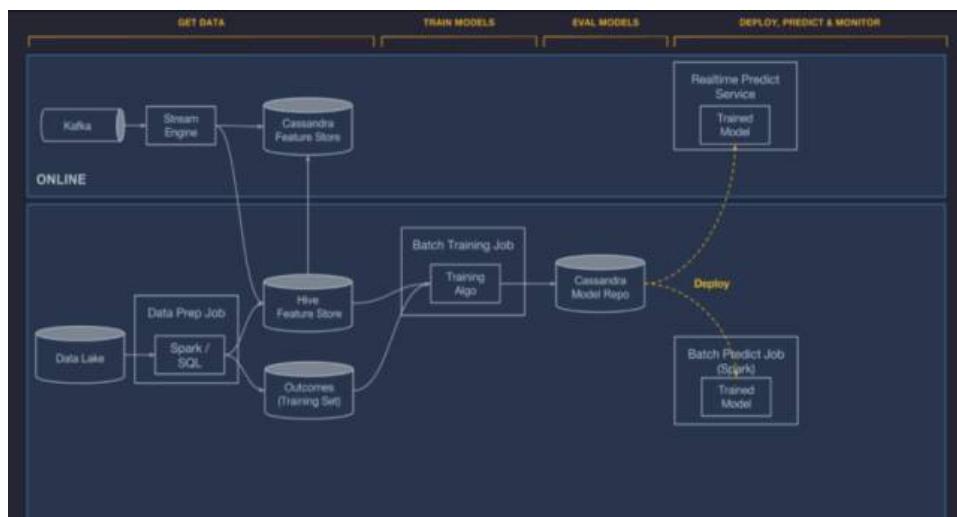


图 8：来自模型库的模型被部署至在线及离线容器当中。

无论如何，一切必要模型组件（包括元数据文件、模型参数文件以及已编译 DSL 表达式）皆被封闭在 ZIP 归档文件内，并被复制至我们的标准代码部署基础设施当中。预测容器会自动从磁盘处加载这些新模型，而后利用其处理对应预测请求。

目前大多数团队都拥有自动化脚本，可用于通过米开朗基罗的 API 定期执行计划内模型重新训练以及部署。在 UberEATS 的交付时间西东发中，数据科学家及工程师们可以通过其 Web UI 手动触发训练与部署操作。

## 制定决策

一旦模型由服务容器部署并加载完成，其即可用于根据加载自数据管

道或者直接提取自客户端服务的特征数据作出预测。各原始特征通过已编译 DSL 表达式进行传递，其负责修改这些原始特征并 / 或从特征库处提取其它特征。最终特征向量会被建立并传递给模型以进行评分。在使用在线模型的情况下，预测结果会被返回至客户端处。而在离线模型场景下，预测结果将被写回至 Hive 并供下游批量任务进行消费，或直接通过基于 SQL 的查询工具供用户访问，具体如图 9 所示。

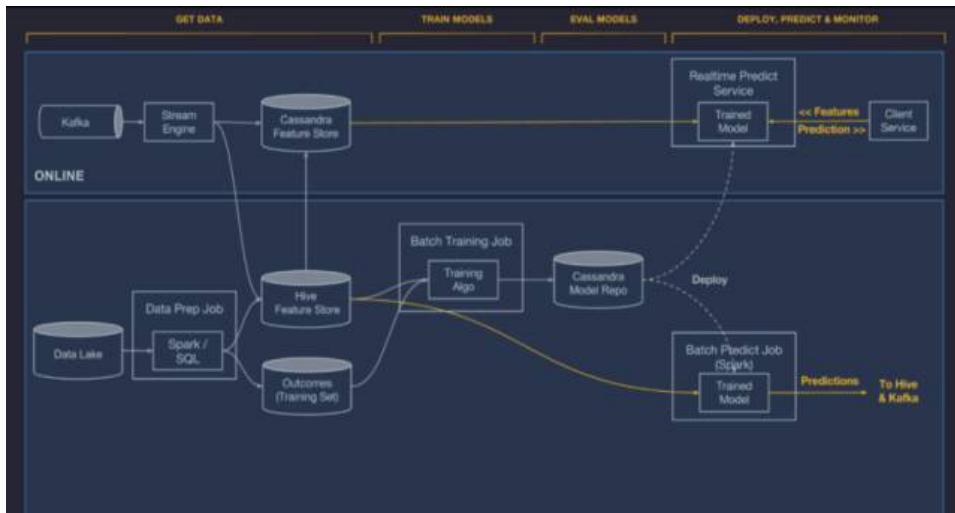


图 9：在线与离线预测服务使用特征向量集以生成预测结果。

## 模型引用

我们可以在同一给定容器当中同时部署多套模型。通过这种方式，我们将能够以安全方式实现旧版本到新版本的过度，并实现模型的并行 A/B 测试。在交付期间，模型根据其 UUID 进行身份识别，同时配合在部署阶段所指定的可选标签（或者昵称）。在使用在线模型的情况下，客户端服务会发出特征微量以及对应的模型 UUID 或者希望使用的模型标签；如果使用标签方式，则该容器会利用最近以该标签进行部署的模型生成预测结果。而在批量模型的情况下，所有已部署模型都将被用于对各批量数据集进行评分，而预测记录当中则包含对应的模型 UUID 及可选标签，消费方可据此对结果进行筛选。

如果在部署新模型以替换旧模型时，发现两套模型拥有同样的签名（即

指向同样的特征集），用户则可继续在新模型中沿用旧模型的标签，而容器则会立即开始使用新模型。如此一来，客户即可在无需变更其客户端代码的前提下对模型进行轻松更新。另外，用户还可以单纯利用 UUID 实现新模型部署，而后在客户端内或者通过中间服务修改配置以将指向旧模型 UUID 的流量切换至新模型处。

至于模型 A/B 测试，用户可以直接通过 UUID 或者标签部署竞争模型，而后立足客户端服务之内利用优步的实验模型向各模型发送流量区段，同时追踪其实际成效表现。

## 规模与延迟

由于机器学习模型具有无状态且不共享属性，因此其能够在在线与离线两种交付模式下轻松实现向外扩展。在线模型情况下，我们能够直接向预测服务集群中添加更多主机，并允许负载均衡器对负载进行分发。而在离线预测情况下，我们则可添加更多 Spark 执行器并允许 Spark 管理相关并发性任务。

在线交付模式的延迟取决于模型类型、整体复杂度以及该模型是否需要从 Cassandra 特征库内提取特征。如果不需要从 Cassandra 特征库处提取特征，那么 P95 常规延迟通常低于 5 毫秒（5 ms）。但如果需要从 Cassandra 特征库处提取特征，那么 P95 延迟通常低于 10 毫秒。目前我们流量最高的模型能够每秒处理超过 25 万条预测请求。

## 预测监控

在对模型进行训练与评估时，我们全程皆使用历史数据。因此，为了确保模型能够在未来继续发挥预期作用，我们必须随时考量生产环境的变化，并根据一切可能影响模型准确度的指标作出调整。

为了解决这个问题，米开朗基罗将自动记录并有选择地保留特定比例的预测结果，而后将这些预测结果与数据管道生成的观察结果（或标签）相整合。利用这些信息，我们将能够对模型的持续性实时准确度进行衡量。

在回归模型情况下，我们会将 R 平方 / 确定系数、均方根对数误差（简称 RMSLE）、均方根误差（简称 RMSE）以及平均绝对误差发送至优步的时间序列监测系统当中，以便用户可以随时间推移分析图表并设置阈值警报，具体如图 10 所示。

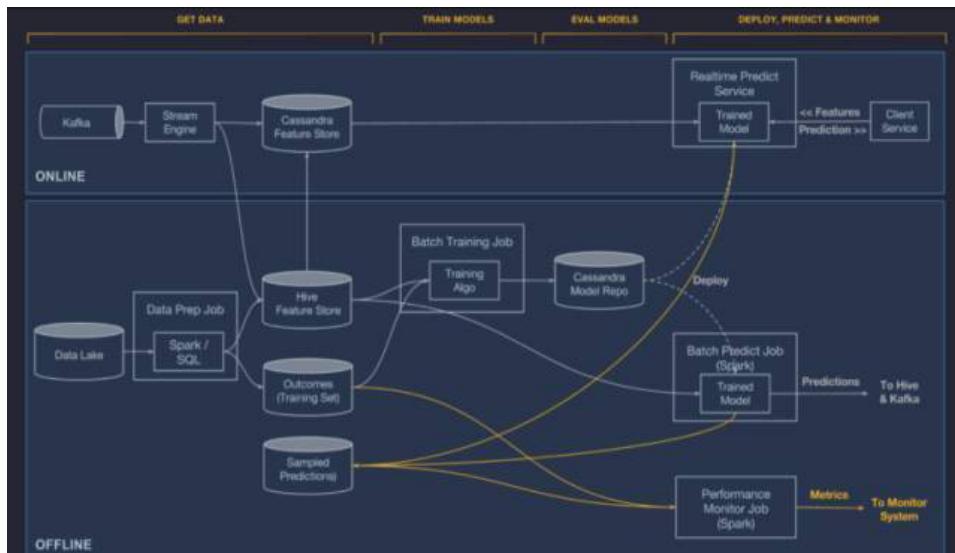


图 10：对预测结果进行采样，并将其与观察结果进行比较。

## 管理面板、API 与 Web UI

这套系统中最后一个重要的部分为 API 层，其亦堪称这套系统的大脑所在。API 层由一款管理应用构成，其负责提供 Web UI 与网络 API，同时与优步的系统监控与警报基础设施进行集成。该层还用于协调负责批量数据管道、训练任务、批量预测任务以及批量与在线容器模型部署的工作流系统。

米开朗基罗平台的用户可以直接通过该 Web UI、REST API 以及其它监控与警报工具同这些组件进行交互。

## 米开朗基罗平台的下一步发展计划

在未来几个月中，我们计划进一步对现有系统进行规模化扩展与增强，旨在支持我们的团队以及优步整体业务的规模增长。随着该平台的不断成

熟，我们还计划投入更多高水平工具及服务以推动机器学习技术的民主化进程，从而更好地支持自身业务。

## AutoML

这是一套用于自动搜索及发现模型配置（包括算法、特征集以及超参数数值等）的系统，旨在针对特定建模问题提供成效最佳的模型选项。这套系统还将自动建立生产数据管道，用以生成模型运作所必需的特征与标签。我们已经利用特征库、归一化离线与在线数据管道以及超参数搜索功能解决了其中大部分问题。这套系统将帮助数据科学家们指定一组标签与目标函数，而后立足隐私与安全感知要求选择最佳模型。其设计目标在于提供更易于使用的智能化工具以提升数据科学家们的生产力水平。

## 模型可视化

对于模型的理解与调试能力正变得愈发重要，特别是在深度学习领域。尽管我们已经通过可视化工具为树状模型提供了良好的理解起点，但未来还需要推出更多相关解决方案以帮助数据科学家理解、调试及调整模型。

## 在线学习

优步的大部分机器学习模型都会以实时方式直接影响到优步产品。这意味着其必须运行在复杂的动态物理世界当中并面临不断变化的周遭环境。为了保证环境变化不会对模型的准确度造成影响，我们必须对模型进行及时更新。目前，各团队会立足米开朗基罗平台定期对模型作出调整。未来，我们需要建立一套完善的平台解决方案，旨在轻松更新模型类型、加快训练速度、评估其架构与管道、自动进行模型验证与部署，同时支持更为复杂的监控与警报系统。这无疑是一项庞大的计划，但目前我们已经获得了可喜的初步成果，这也证明我们的思路值得继续坚持下去。

## 分布式深度学习

越来越多的优步机器学习系统开始采用深度学习技术。由于需要特殊的平台支持能力，因此深度学习模型上的定义与迭代工作流与标准工作流

存在显著差别。深度学习用例通常会处理规模更为庞大的数据，且对硬件的要求也相对更高（例如 GPU），这意味着我们必须对分布式学习方案作出进一步投资，并将其同一整套灵活的资源管理方案作出更为紧密的集成。

# Geekbang

极客邦科技

整合全球优质学习资源，帮助技术人和企业成长

InfoQ

技术媒体

EGO NETWORKS

职业社交

StuQ

职业教育