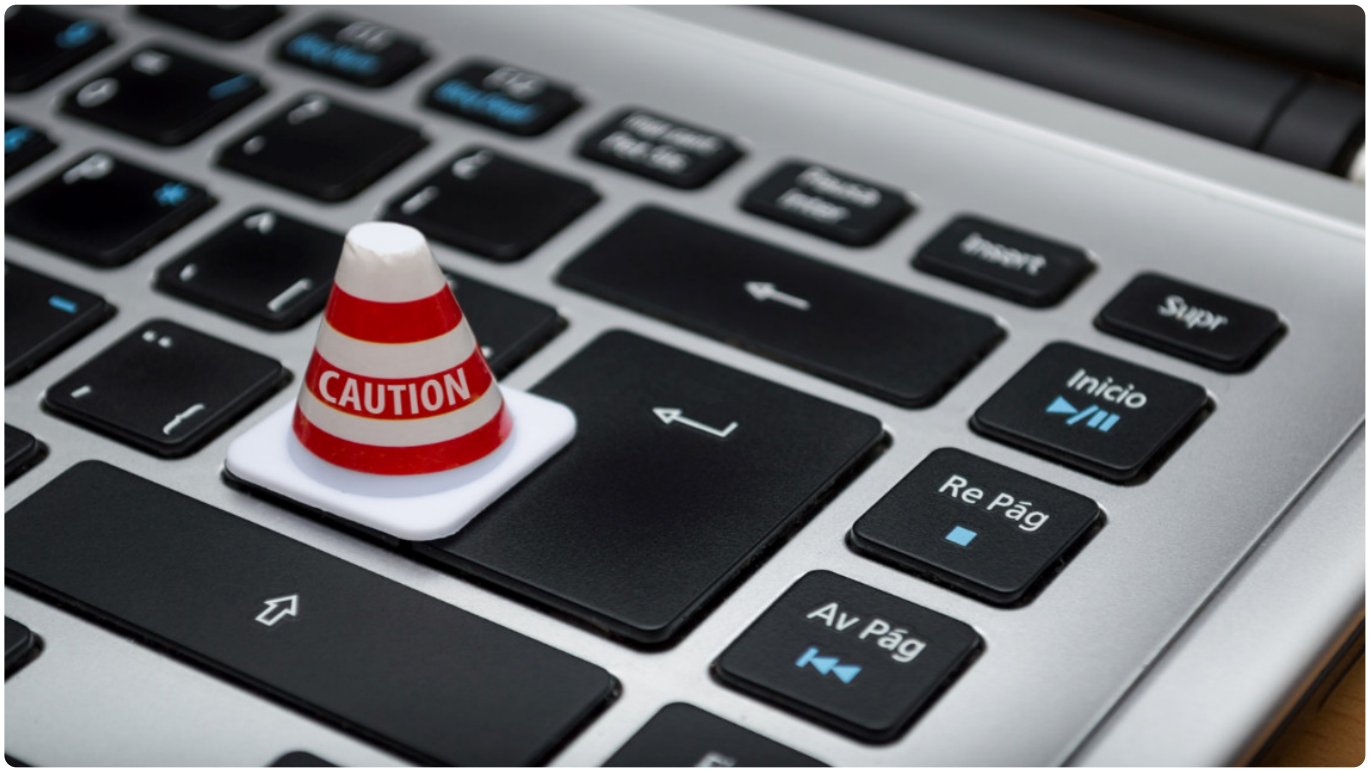
137 | 如何做好人工智能项目的管理?

2018-08-17 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 07:11 大小 4.12M

关于数据科学团队养成这个主题，在之前的分享中，我们已经聊了数据科学团队招聘以及一些高级话题，主要是围绕如何组建一个高效的团队，包括数据科学家的绩效评定、培养以及如何构建水平和垂直的组织架构这些内容。

在接下来的几篇分享里，我们重新回到数据科学团队的本源，来看一看数据科学团队在整个公司的位置，以及在数据科学团队的发展和运行中，又有哪些至关重要的环节。

今天我们就来聊一聊运行数据科学团队里面的一个核心问题，就是如何针对人工智能项目进行管理，从而保证团队运行的项目能够顺利完成，同时能够真正帮助企业利用人工智能技术来推动产品的发展。

一说到项目管理，很多成熟的工程师或者项目经理可能会不以为然，觉得不需要对人工智能项目进行额外的关注。但是在实际工作中，如何运作数据科学项目关系到整个产品的推进，甚至可能对公司的发展都会有不小的影响。

那么，通常情况下，针对人工智能项目会有哪些项目管理的模式呢？我们先来看看两种极端模式。

把人工智能项目当作“研究项目”

第一种模式是把人工智能项目完全当作是研究项目。很多从学术界转到工业界的研究人员和工程师，在处理人工智能项目时就很容易陷入这种状态。那么研究项目有哪些特点，或者说，如果我们把人工智能项目完全当作是研究项目，会带来什么问题呢？

首先，在很多状态下，**研究项目并没有特别明确的目标**。有些项目看似是利用一个现成的方法来解决一个实际问题，但是做了一阵子才发现问题的定义并不清晰，而那个现成的方法可能需要重新修改，才能在新的问题上使用。而且，修改这个方法也许还需要进行一些理论推导，即便做了所有这些步骤以后，依然没有人可以保证这个方法对新的问题一定有效。

也就是说，研究项目的每一个步骤都充满了不确定性。**这种过程和结果的不确定性，可以说是研究项目最大的特点**。不确定的特点带来的结果，往往就是不太好控制整个研究项目的范围。

举一个例子，如果我们要针对一组图像构建一个分类器。这个项目其实可大可小，可快可慢。如果我们直接用现成的模型，然后利用迁移学习的办法，不去重新训练模型，仅仅是把模型直接应用到新数据上，那快则一天，慢则一个星期，可能就完成了这个任务。

然而，这么做对分类精度是没有任何保证的。我们可能会发现分类器的精度比我们想象的要低得多。那么，这个时候就会达到一个比较危险的时刻。原始项目范围内需要做的任务都已经完成了，但是没有达到效果，后面可以做的事情，范围可能就会非常大，也没人能说得清楚，做了这些额外的任务之后，是不是一定能提高分类器的精度。

因此，把人工智能项目完全看作研究项目的弊端就很明显了，我们无法很好地把握整个项目的周期，特别是完成时间。同时，种种的不确定还可能造成项目范围的不可控。显然，这种局面是工程项目中最不愿意看到的。

把人工智能项目当作“软件工程项目”

另外一个人工智能项目管理的极端模式，就是完全按照软件工程的模式来进行管理。我们这里不讨论具体的软件工程管理方法，我们仅从宏观上来讨论软件工程管理模式对于人工智能项目管理的弊端。

软件工程管理的一个核心思想，就是能够**把一个大的任务拆分成一些细节的任务，然后假设如果能够完成小的细节任务，那么大的项目也就能顺利完成。**同时，软件工程管理还有一个重要的假设，那就是**工程的绝大多数步骤都是确定的，没有过多的变数。**

遗憾的是，正如我们刚才提到的，人工智能项目的一个特征，就是不确定性。因此，**按照软件工程进行管理，就容易做一些看似有意义，但其实对工程进展并没有真正帮助的任务。**这些任务往往是数据科学家或者工程师凭空制造出的，目的就是为了符合软件工程的管理流程。过于细节的任务划分，往往就会把整个项目真正的目标给迷失掉，从而无法针对是否达到目标很好地进行控制。

回到上面那个图像分类器项目的例子，如果我们采用纯粹的软件工程管理方式，那步骤很可能是这样的：需要先写一个计划书，再对数据进行描述，然后找责任相关方来探讨是否需要重新训练模型等等。这些步骤耗费了大量时间，但是对于能否构造出高精度的分类器并没有帮助。

人工智能项目的管理

那到底该如何来管理人工智能项目呢？人工智能还处于发展的初期，目前其实并没有一个完全成型的项目管理方法论和一个放之四海而皆准的框架。不过，通过刚才对两种极端情况的讨论，相信我们可以在真实的项目管理过程中想出一些办法。

首先，我们需要有一个迭代的思路。迭代思路是为了能够有效地**管理项目的范围。**还是回到我们所说的图像分类器项目，如果我们利用迭代的思路来进行项目管理，就不会把一个绝对的模型精度当作是项目的唯一目标，而把提高精度当作目标，但事先不会针对精度有过分苛责的追求，那么每一天每一周需要做的工作就相对比较容易可控。

其次，我们需要分清楚项目中哪些部分是相对可控的，而哪些部分是比较不容易控制的。当一个模型被训练出来后，要把这个训练流程形成一个**每天可以更新的工作流**（Workflow），这个任务是相对比较可控的。可控的部分我们就可以利用软件工程的项目管理方法了，来对这些任务进行细分。那不可控的任务呢？比如希望提高当前模型的精度，或者是数据量大了十倍以上，依然希望能够进行操作等等。针对这些任务都没有直接的答案，寻找解决方法的过程充满了不确定性，那就无法真正利用软件工程的项目手段了。

最后，是人工智能工程项目管理的一个“小窍门”，那就是设置完整的“交工日期”（Deadline）。不同的交工日期往往意味着完全不同的解决方案，甚至这些解决方案之间会有非常大的精度区别。我们的交工日期要建立在迭代思想上，这样就能保证我们的项目在每一个交工日期都有一个成型的结果。如果模型的精度还有提升的空间，我们就可以依赖下一次迭代去完成精度的提高。

小结

今天我为你讲了数据科学团队的一个核心问题，那就是如何针对人工智能项目进行管理。

一起来回顾下要点：第一，我简单介绍了什么是人工智能项目管理；第二，我们分析了两种极端的项目管理模式以及各自的弊端；第三，我们讨论了如何利用两种极端模式来寻求中间路线的办法。

最后，给你留一个思考题，你自己的经验里，人工智能项目在运行过程中，哪些步骤或者说是流程是最消耗时间的？

欢迎你给我留言，和我一起讨论。

上一篇　136 | ACL 2018论文精读：什么是“端到端”的语义哈希？

下一篇　138 | 数据科学团队必备的工程流程三部曲

精选留言 (1)

写留言

廉明
2018-08-31

1

我们是这样做的
整体使用模块化。确定性模块和不确定模块，都先定义好接口。先用baseline代码实现功能，确保数据和整体流程在限定日期可以跑起来。然后各模块分头开发。整合时再进行准确率调优和性能调优。这样项目才能可控。