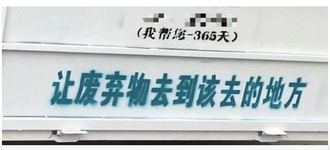
两个月，刷了八千篇Arxiv，我发现……

原创 白鹤鸣 夕小瑶的卖萌屋 2021-07-22 12:05



文｜白鹤鸣
编｜小轶

从五月初到现在，大约刷了八千篇Arxiv之后，我发现我有毛病。

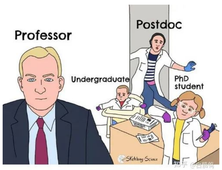
你好像有那种大病

当然，这是读论文上头时的牢骚，不是真心话，只是说，**我在Arxiv上投入的精力的努力，与我预计的收获不成正比。**

故事的起因是这样的：

作为一个搏一的萌新，学校和导师不会直接让你上手科研，而是先上上课。确保来自不同学校的同学们能够拥有相近的知识背景，互相认识认识。但是，当你的日常是上课的时候，看着学长学姐们学术讨论，实验跑得风生水起，人总是会慌的。

"天呐我已经是一个成熟的研究生了，为什么每天还像本科生一样课课课，我也要搞科研！"



然后想想除了课程和作业报告，似乎确实没什么整块的时间可以静下心来研究，最可行的只有每天看看论文了。下定决心的时候是5月，由于各种课程的大作业并始陆续下发，最终，实际能干的事情，就是通过RSS订阅[1]，开始遍历Arxiv和一些领域相关Journal的论文。

在开始做这件事情的时候，我充满了干劲和对论文的美好期许：

"每天能够接触到所有研究者最新的idea和发现，我就是时代的弄潮儿！"

"那么多新发的方法，研究与研究之间都是相通的，可以把数理领域的前沿成果拿过来实现我们领域的研究问题，这效果绝对杠杠的！"

"顺带还可以练一练英语速读能力，文科理科两手都抓，太机智了！"

现在回头望去，我就像个戏台上的老将军——浑身插满了Flag。

Arxiv是北京时间每天上午九点更新，美国时间的周末不更新。我订阅了CS领域下 *人工智能AI，机器学习ML，计算机视觉CV，信息理论IT* 四个方向的论文。平均每天加起来这些领域会更新150篇上下，周一会更多一点，因此我每周大约会接收到800～900篇论文推送，5月到7月中旬，加起来推送的总量必然有8k以上。

作为一个理智尚存的成年人，我采取的策略是首先速览题目，对于研究相关的、或者看起来很有意思的文章，瞅一眼Abstract。如果Abstract搞不出毛病，再打开原文更详细地阅读。

- **综述类文章**：这类文章的价值是介绍一个方向的研究进展和前沿技术，并总结研究难点痛点，几乎不具有创新性。写得好的往往会直接投稿给期刊，因此在Arxiv上出现频率不高。一篇好的综述除了方法，更重要的是指出有待研究的空白。因此，对于只是罗列方法，总结不足的综述我都不会进一步阅读。
- **理论/观点型文章**：显然，这类文章最重要的就是它的观点和论证过程。一定要搞清楚文章的假设是哪些，限制在哪，如果不合理的话就不用看下去了。论证一般要么靠逻辑，要么靠公式推导，想很快把公式搞懂显然是不现实的，但可以看看是基于哪些数学方法来决定是否值得细看。
- **方法型文章**：这类文章的常见结果中包括"我们做到了xxx方面的SOTA"，但是，模型的评估指标有哪些，和什么样的参考如何比较得出了这个SOTA，往往暗藏玄机。所以，看一眼模型构造，如果不是瞅熟的缝合怪，再看一下实验，实验没有太大问题，再瞄一眼结果，到底进步了多少，有没有机理分析，这些全齐活了，文章的具体方法才可能具有可信度。

浏览方法是合理的，实施过程是痛苦的，我看到了五花八门标题类丽，开头让人心神荡漾，实验结果或者方法一言难尽的文章。还有些投稿，只描述了作者想达到的效果，方法刚写了一小段，实验还没跑，导致我最终养成了开文章先看眼页数，免得被画饼欺骗感情的好习惯。

这两个月里，各式各样的SOTA我见了上百篇，近期的few-shot，explanable AI，看起来都是研究热点，然而最终，这大约8k篇的论文中，我精挑挑拣，目前下载导入Mendeley打算好好研究的只有不到100篇。这样做的时间成本是多少呢？

- 假设每天我稳定读了150个标题，这大约需要半个小时。
- 这150个标题中，有10篇能引起我的兴趣，我花十五分钟，过了一下它们的摘要。
- 作为一个新手，我对于摘要的判断能力还不是很强，因此，这10篇文章中我需要仔细地阅读5～8篇文章的intro，result，conclusion。这至少需要半个小时。
- 最后，由于我连续读了这么久文章，我笑励自己就地躺平一刻钟。

所以，**在Arxiv上刷文章，我每天需要花一个半小时左右，能够获取1～2篇可能有价值的文章。** 而作为一个新手，我的研究嗅觉未必足够灵敏，也就是说，在这些决定精读的这些文章中，有50%以上的概率，在继续阅读1～2小时之后，我仍将一无所获。而 **如果利用这些时间有目的地定向搜索特定领域的文章，** 参考文章的引用量，**我将更可能在同样的时间内了解更有价值的研究成果。**

在Arxiv上，作为一个研究领域的新手面临的问题是选择太多了，难以甄别有效信息。最初我试图从数理领域获得新的方法的设想并不成功，**数理领域的breakthrough出现概率并不高，而且，想要将其他领域的方法迁移到自己的领域，一方面，获取方法的时间成本会成倍地增长；另一方面，踩雷的风险绝不低。**

作为一个能够流畅读写论文的研究生，绝对不要指望用Arxiv能对英语水平有多少提升。 很简单，因为Arxiv上的论文，在没有经过会议和期刊对语言的筛选打磨时，英语质量着实参差不齐。目前英语词汇量在1w左右的我感受到的瓶颈，主要来自词汇的使用不够多样化导致的语言生硬，以及做不到快速逐行阅读。而论文能让人锻炼快速阅读的部分并不多，很多内容都是要边思考边看的。论文作者也未必是Native speaker，很可能写文章的时候也词穷。对于这个个人问题，最近摸索的结论是，**看CNN和BBC的新闻，对语言的提升效果远好于读论文。**

总而言之，**Arxiv上良莠不齐，对于研究领域的新手（博一博二及以下）来说，并不应该以刷Arxiv作为信息获取的主要渠道。**我的导师在听说我的计划的时候，曾经劝阻过我：

> *"你现在不应该大量漫无目的地阅读文献。而是应该努力寻找可能给你提供新的研究灵感，或者教会你研究方法的论文。"*

也就是说，**搜索特定词条下的论文和Tutorial对我这个阶段的研究生帮助会更大。Arxiv在现阶段更适合作为检索是否存在idea撞车的数据库，**而非图书馆。至于领域中的老手，刷Arxiv的时间成本应该是著降低（很多方法只要大致浏览就能理解），但若要紧跟研究潮流，每天1～2小时的阅读应该还是少不了的。具体细节，就等我能看到他们眼中的风景时再来和各位分享吧。

不过呢，Arxiv上乐子还是不少的。可以这么说：如果回到两个月前，我不会开始刷Arxiv：但在经历这么多痛苦，逐渐摸索到一些门道之后的现在，我还是打算继续刷下去的。希望接下去，Arxiv能提升我甄别论文的能力，此外，我会对有价值的论文做一些笔记，从而提升自己的理解概括能力。

本文描述的读文献方式"导师见打"，非搞英人士请勿模仿！

萌屋作者：白鹤鸣

白鹤鸣（jī líng）是一种候鸟，天性决定了会横跨很多领域，已在上海交大栖息四年，进入了名为博士的换毛期。目前以图像语义为食，但私下也对自然语言很感兴趣，喜欢在卖萌屋轻松不失严谨的氛围里凌~~形~~飞~~翔~~
知乎ID也是白鹤鸣，欢迎造访。

作品推荐：

1. NLP太卷，我去研究蛋白质了~
2. 谷歌40人发表59页长文：为何真实场景中ML模型表现不好？
3. 学术&工业界大佬联合打造：ML产品落地流程指南

[1] ^RSS (Really Simple Syndication) 是一种消息来源的格式规范，网站可以按照这种格式规范提供文章的标题、摘要、全文等信息给订阅用户，用户可以通过订阅不同网站 RSS 链接的方式将不同的信息源进行聚合。在一个工具里阅读这些内容。