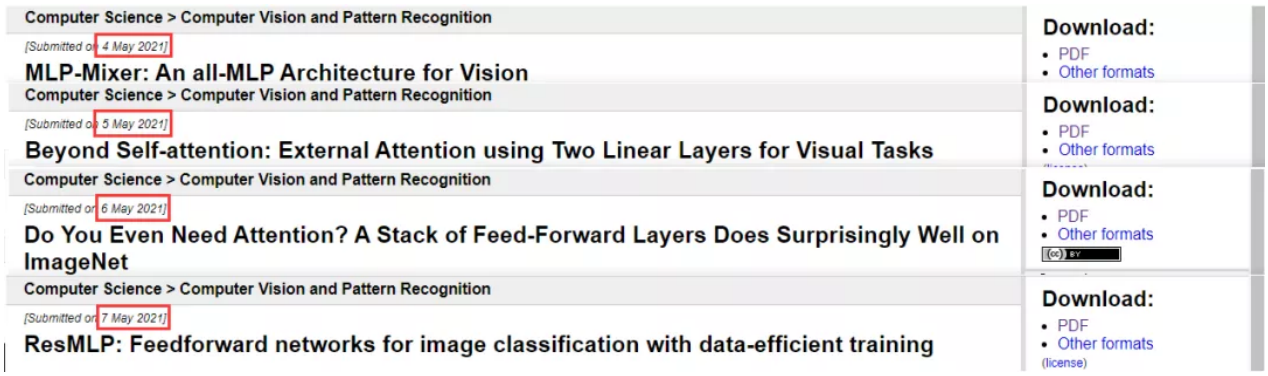
论文投稿新规则，不用跑出SOTA，还能“内定”发论文？！

原创 Sheryc_王苏 夕小瑶的卖萌屋 2021-05-25 18:00

文 | Sheryc_王苏

从5月初开始，CV圈似乎开始了一阵MLP“文艺复兴”的热潮：在短短4天时间里，来自**谷歌**、**清华**、**牛津**、**Facebook**四个顶级研究机构的研究者分别独立发布了4篇关于MLP结构在图像任务上取得不错效果的论文。虽然研究本身令人兴奋，但发表的过程却让人一言难尽：来自牛津的小哥Luke就在reddit上抱怨到他正在进行的实验被谷歌的MLP-Mixer**抢先发表**（scoop）了，所以他正在撰写的实验总结也只能以实验报告的方式尽快挂在arXiv上。



做研究的过程中，同样的想法被其他研究者抢先发表是家常便饭。或许我们已经对此习以为常，但这真的没有办法解决吗？在NAACL 2021上，就有研究者从心理学、药学等其他学科研究中被普遍使用的“**预注册**”机制（Pre-registration）出发，为NLP领域量身定做了一套预注册机制，希望能用一套新的研究和投稿流程让作者**不再担心被抢先发表**，**不再担心好的研究因为没有SOTA被拒**，**不再担心慢研究赶不上快节奏**...根据作者在文中绘制的蓝图，这种机制虽然简单却好处多多，它究竟能够为未来的NLP研究带来什么样的改变呢？

论文题目：

Preregistering NLP Research

论文链接：

https://arxiv.org/abs/2103.06944

Arxiv访问慢的小伙伴也可以在 **【夕小瑶的卖萌屋】**订阅号后台回复关键词 **【0525】** 下载论文PDF~

什么是预注册？

预注册

所谓**预注册**，指的是在**进行一项研究之前**，将自己的详细研究计划在**预注册网站**[1]上进行注册。预注册的内容反映了在**研究开始之前需要考虑的所有事项**，例如自己的研究假设、模型设计和实验方案。

在内容上，预注册很像是在申请项目资金时撰写的申请材料，或是研究开始前导师要求撰写的研究计划，不过虽然内容类似，预注册的不同之处在于“**注册**”二字：这些内容会被提交至网站上，并打上提交之时的*时间戳*，证明作者在某个时刻已经完成了实验设计。在网站上预注册的内容可以作为*实验的初始设计记录*、作为实验取得进展或发生变化后*记录进度*的仓库，甚至可以直接作为“**注册报告**”（Registered reports）向期刊或会议**直接投稿**。这种机制早在2018年就已在Science上发文[2]进行过相关讨论，但目前在AI领域还并不常见。

注册报告投稿制度

上文中提到的“**注册报告**”**投稿制度**正是线上预注册制度的主要副产物。在这种审稿制度下，**同行评议过程被分为两个阶段**：

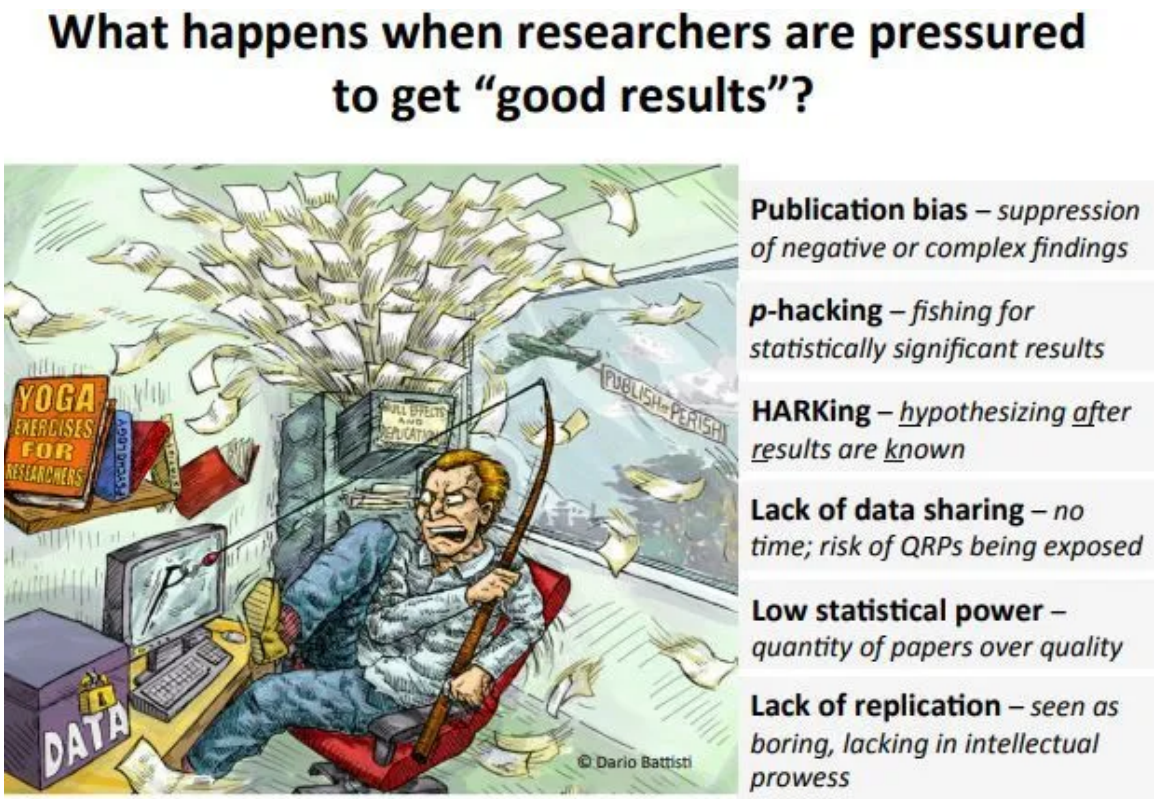
1. 在*研究开始前*，作者给审稿人提交一份**预注册的研究计划**。随后，审稿人根据研究计划进行评审，经过多轮修改，决定**是否接收该研究计划**。
2. 在*研究计划被接受后*，开始根据已提交的研究计划进行相关实验，撰写论文，随后向审稿人进行**第二轮提交**。随后，审稿人根据论文进行评审，经过多轮修改，决定**是否接收该论文**。

对于注册报告投稿制度，一般在第一阶段的研究计划被接受后，作者便得到了来自主办方的保证，即**只要按照研究计划完成的论文即可被接收，无论实验效果好坏**。因此，如果实验计划被认为有意义，即使最终被证明方法无效也可以被接收。

在进行实验时，作者可以随时对研究计划进行修改，但对研究计划的任何修改都需要在最终报告中体现。目前，注册报告投稿制度已经被包括Nature子刊在内的数百种期刊采用。

预注册有哪些好处?

看起来，预注册似乎只是将自己的详细研究方案在开始着手试验之前挂在网站上。但是，这样简单的操作却能带来新的研究范式，改变现有研究中的诸多问题：



1. 让自己**提前完整设计研究方案**。在进行研究之前，通过预注册网站上提供的一系列问题列表让自己在动手之前就可以从多个维度仔细思考研究的细节和意义，**避免进行无意义的探索**。同时，正如NLP大牛Jason Eisner所建议的[3]，预注册过程中撰写的报告或许可以**作为最终论文的一部分**，让自己在开始研究之前就着手撰写论文。
2. 区分**探索性分析**（Exploratory）和**验证性分析**（Confirmatory）。所谓探索性分析指通过实验结果产生新的假设，而验证性分析指通过更多实验和分析验证先前已提出的假设。在一些实际研究中，研究者往往**将探索性分析伪装成验证性分析**，即首先通过实验得出结果，再通过结果*反推*出一个假设，说明自己的实验验证了反推出的假设，这种行为被简称为**HARKing**（Hypothesizing after results are known），会导致产生不严谨的假设。如果是根据注册报告进行评价，则在得到实验结果前就需要对探索性分析和验证性分析进行区分，避免错误假设的出现。
3. **避免发表偏见**[4]。所谓发表偏见，指的是会议或期刊**偏好于发表现象显著、结果好的研究**，而不偏好现象不明显、结果较差的研究。实际上，有些研究即使结果不好也有其发表价值，但是为了能够让论文发表，研究者倾向于**压缩负面的发现**，着重强调好的结果。这也难怪为何近年来NeurIPS上还会有“I Can't Believe It's Not Better! Workshop”[5]这样专发没效果的模型的workshop了。但如果根据注册报告投稿制度，根据论文本身的**设计**来确定接受与否，就可以让研究者在看到哪条路可行的同时，更多的接触到那些前人试过但**不可行**的方案了。
4. **避免被他人抢先发表**。注册报告制度使得**先提出实验方案的人拥有优先权**。即使不实行注册报告制度，预注册报告上的**时间戳**也可以证明提出类似想法的时间先后。预注册报告可以随时选择是否公开，让他人难以直接通过预注册报告剽窃方案。
5. **鼓励慢科学**[6]。当注册报告被同意接受之后，作者不需要担心被提前发表，因此可以不用在尽可能短的时间赶完文章，而是可以选择用更长时间**仔细打磨**自己的想法，让论文更具深度。

NLP的预注册有什么特点？

不同领域有着不同的研究特点，甚至同一领域内的不同种类论文也有不同的研究方法。这篇论文的最大贡献就在于其对NLP领域内的各类论文分别提出了一个初版的**预注册表格**，作者在预注册时需要回答表格内的若干问题。

对于NLP领域，作者根据COLING 2018的论文分类将研究分为了3种：**计算辅助的语言学分析**（Computationally-aided linguistic analysis），**NLP工程实验**（NLP engineering experiment paper），以及**复现/资源/立场/综述**（Reproduction/Resource/Position/Survey paper）。其中，作者对于除Position Paper以外的各类研究都给出了推荐的预注册表格。

下面以最为常见的NLP工程实验论文为例，在预注册过程中需要回答以下问题：

1. 你的*研究目的*是什么？
2. 你的*研究假设*是什么？
3. *独立变量*有哪些（例如：模型结构）？*非独立变量*有哪些（例如：模型输出好坏）？
4. 以上变量将被*如何衡量*？

5. 实验包含*几种语料或任务*?
6. 你将使用哪些*软件库*?
7. 你将使用何种*硬件*?
8. 你将使用何种*参数设置*?
9. 你将使用什么样的*数据*?
10. 如果实验数据不存在，请回答关于资源类论文（Resource paper）关于*收集数据*的预注册问题（原文附录A.6）。如果实验数据存在，你对实验数据的*熟悉程度*是？你的*实验假设*在多大程度上与该数据*相关*？这在多大程度上影响了你方法在其他数据上的*泛化性能*？你是否准备*收集更多数据*来验证自己的方法?
11. 为何*选择该数据*？这些数据有哪些*关键性质*?
12. 这些数据是如何被划分为*训练集/验证集/测试集*的?
13. 你将如何*分析结果*并测试自己的假设？如果是自动评测，你将使用什么样的*指标*和实现？它们被如何设置？如果是人工评测，请回答关于*人工评测*设置的预注册问题（原文附录A.8.1）
14. 你是否会进行*错误分析*？如果是，请回答关于错误分析设置的预注册问题（原文附录A.8.2）
15. 你是否有其他需要进行预注册的信息?

其他种类论文的预注册表格可以在论文的附录中找到。

可以看到，预注册的问题着重瞄准自己的方法如何能**论证/反驳自己的假设**，而非如何在数据集上取得更好结果。虽然需要回答的问题很多，但这些问题基本涵盖了NLP实验论文的全部设计细节，在着手试验之前想清楚以上所有问题对于实验的整体把握会有相当大的帮助。

总结

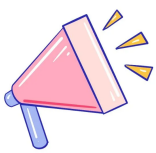
预注册和注册报告制度虽然已经被心理学等学科的顶刊作为标准流程，它的可行性和影响依然在探索之中。不过，对于我们一直以来所诟病的“刷SOTA”、抢创意、堆算力的行为，使用注册报告制度或许能够带来缓解；同时，预注册制度也能让研究者从在单一数据集上追求模型性能的过程中提前跳出来，从更高的角度和更多元的视角上探索自己的方法**对整个领域的推动作用**。即使预注册不是标准流程，在研究开始前仔细思考和记录预注册问题的结果也能提前避免一些弯路，让后续的实验更有效率。在AI领域日益火爆的当下，对于研究和投稿流程的改进同样是一项重要课题，几年后NLP的研究范式会变成什么样，着实让人期待呢(=•ω•=)

萌屋作者：Sheryc_王苏

北航高等理工学院CS专业的市优秀毕业生，蒙特利尔大学/MILA博士生，资深ACG宅，目前作为实习生在腾讯天衍实验室进行NLP研究。虽主攻NLP，却对一切向更完善的智能迈进的系统和方向充满好奇。如果有一天N宝能真正理解我的文字，这个世界应该会被卖萌占领吧。（还没发过东西的）知乎ID：Sheryc

作品推荐：

1. NLP未来，路在何方？12位巨佬联名指路！
2. 这几个模型不讲“模德”，我劝它们耗子尾汁



参考文献

[1].常用的预注册网站，不妨去看看：Open Science Framework: https://osf.io/prereg/ AsPredicted: https://aspredicted.org/

[2].Science对于预注册的讨论：More and more scientists are preregistering their studies. Should you?. Science. https://www.sciencemag.org/news/2018/09/more-and-more-scientists-are-preregistering-their-studies-should-you

[3].研究之前先开始写：Write the Paper First by Jason Eisner. https://www.cs.jhu.edu/~jason/advice/write-the-paper-first.html

[4].发表偏见：Publication Bias - Wikipedia. https://en.wikipedia.org/wiki/Publication_bias

[5].慢科学与快科学：Research Fast and Slow by Min-Yen Kan. http://bit.ly/kan-coling18

[6].有关预注册机制的更多细节：https://www.cos.io/initiatives/prereg?_ga=2.218660505.1451147193.1621172626-420219689.1621172626