发现一篇专门吐槽 NLP 内卷现状的 ACL 论文 ...

原创 Severus 夕小瑶的卖萌屋 2021-08-23 12:05

文 | Severus
编 | 小轶

随着大模型的发展，NLP领域的榜单可说是内卷到了无以复加，现在去浏览各大公开榜单，以至于各个比赛，随处可见BERT、RoBERTa的身影，甚至榜单中见到各大large模型的集成版也并非偶然。在发论文的时候，又要不断地去内卷SOTA，今天的SOTA在明天就有可能被打败，成为了过眼云烟。极端情况下，某一篇论文正在撰写，ArXiv上就突然刷新了SOTA，又足以让研究者们头疼应该怎样应对。

同时，参数规模的内卷，在去年GPT-3发布之后，上升到了百亿、千亿甚至万亿，参数规模的急剧上升自然也将榜单的分数提升了一大截，而这种大模型无论算力消耗、实验成本还是优化难度，都足以让广大研究者们望而却步。

一直以来，NLP的这种发展方式都存在一些指责的声音，而内卷程度达到了今天这种程度之后，自然也会有更多的工作停了下来，他们去自省：现在的热门工作的意义在哪里？过度以SOTA为标准来评审工作对这个专业是否是有利的？军备竞赛做出来这么多目前根本没法用的大模型意义何在？

今天要介绍的这篇 ACL'21 的文章，就是总结了当前NLP领域的一些问题，以及给出了相应的解决思路。文章作者总结列举了当前NLP领域研究的5个问题，分别为：

1. **过早地应用了未经充分分析理解的方法**
2. **偏好计算方法，却不考虑其局限性带来的风险**
3. **论文发表的偏好**
4. **因实验成本而导致不可能复现实验**
5. **模型的不可解释性**

以下是文章作者对这5个问题的详细阐述，以及分别提出了自己的解决方案，笔者也会逐条发散一下自己的看法。

论文标题：
On the Gap between Adoption and Understanding in NLP

论文链接：
https://aclanthology.org/2021.findings-acl.340.pdf

网址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 **【0823】** 下载论文PDF~

过早的应用

BERT发表之后，迅速席卷了NLP领域，将NLP研究的范式改变为Pretrain+fine-tune模式，但是作者认为，BERT（及在它之后的所有类似工作，如ERNIE、RoBERTa等）的应用未免太快，我们还没有充分理解它到底学到了什么，它就已经成为了几乎所有工作的基座模型，因为它在当时的理解榜单上迅速以压倒性的优势刷新了SOTA。而同时很多对BERT的分析工作表明，我们对预训练语言模型的能力有了过高的估计，例如BERT对否定的概念不敏感，例如在BERT上可以构造各种对抗样本去使其结果变差等等。也就是说，作者认为现在的方法中存在使用和理解的差距（gap between adoption and understanding，GAU）。

这种未对成果进行充分分析及研究，就过早地将成功应用起来，所造成的危害在其他的科学研究领域，已经造成了一些危害。例如文章中举出的例子，用于治疗孕妇失眠的药物康特甘，后来被证明有严重的副作用，甚至可能导致流产。在医药学领域，类似的例子还有许多，例如海洛因，最初研究目的是作为一种镇痛效力比吗啡更强，有不具备吗啡的成瘾性的药物，但是在实验阶段，忽略了动物的异常反应，而这个以“英雄”之意冠名的药物却成为了恶魔。类似的还有甲基苯丙胺。

NLP领域当然不会有这种风险（虽然我认为如果将现有的模型当成AGI滥用的话，其危害不会很小，好在现在大家都很清醒），过拟合也不会给人造成身体上的伤害。但是，考虑如果研究者A，发表了一个最好的方法X，那么方法X就会变成一个标靶，后来者的方法都会去参考它，试图打败它，对抗它，从而去刷新分数。但如果之后，方法X被证明是错误的，研究者A撤回了他的结果，那么对于X之后的相关研究可能就是毁灭性的。想象一下，如果BERT被证明是错误的，那么对于NLP领域造成的毁灭将会是什么样的？

又或许，研究者A发表了方法X，且方法X成为了一个经典方法之后，他也就满足了，也就不再继续推进研究，那么对于新的更好的工作，也造成了障碍，因为错误的方法没有被识别出来，变革性的方法反倒又导致了该领域的停滞不前。

作者认为，**需要创造一个可以去探索NLP方法缺点，及负向的发现的环境，而不是做事后诸葛亮**。其中关于负向结果的workshop[1]以及带有对抗性质任务的workshop[2]（build-it，break-it）是比较正确的方向。

笔者的看法

实际上，我想经典预训练语言模型（BERT、RoBERTa等）的作者们是应该是相当清醒的，实际上我们可以看到，无论是自回归还是自编码的语言模型，其预训练任务及方法都已经是经过多年实践的方法，且其理论依据也经过了充分的研究，而且大家都是可以充分理解这些语言模型，或者说是统计模型的局限。而在统计模型上，过度的苛求其在语料之外的人类知识类问题上的性能。当然，关于BERT通过统计共现，可能已经记忆到了什么语言学知识的相关研究，我认为是相当有意义的，它可以让我们更清楚地看到统计语言模型具备什么样的非凡能力，以及它局限在哪，或者可以有什么其他的用法。

除此之外，文章作者所提到的这个观点我是完全认同的。同时，我认为，**提出了错误的方法，却得到了高分的结果，则更加有可能是任务不可靠，或者数据不可靠，导致过拟合形成了高分**。在我与研究者们的交流中，研究者们往往也会指出数据上的问题，导致他们无法判定在固定任务场景之下，他们做出来的模型到底有什么意义。

所以，**构建更加可靠的数据，提出更加可靠的任务，以及[2]中所提到的对抗模式，可能一定程度上能够缓解这个问题**。当然也需要广大研究者们对统计模型的认识足够清楚。

计算类论文

NLP领域是方法驱动的，自然也会不断地去探索新的技术。然而这也使得论文数量失衡，对领域自省或其语言学上的动机的研究变少。这种发展起源于上世纪九十年代的统计革命，统计模型大幅占据了优势，方法导向的论文优于理论导向的论文。到现在，深度学习模型的统治地位仍未被撼动，这种思潮也就根深蒂固，那么自然提出新的模型，比单纯的语言学理论研究要受欢迎得多。不过，纯方法论的论文也更加容易客观评价，这也是事实（因为更加注重结果）。那么就引出了两个问题：

- 模型的结果比它语言学上的理解更加重要吗？
- 计算类的论文是否应该以不同的方法评估？

这也就是理性主义和经验主义的分歧，理性主义希望模型能够被理解，而经验主义则希望模型有用[3]。两种方向结合，才能够取得真正的进展。

每年都有无数的论文提出了新的模型，声称自己取得了新的结果，但现在却没有一种方式去认证这些结果，多数时间我们不了解这些论文的评估是否是合理、正确的。这其中最大的问题是很多发表出来的论文，却没有高质量的开源代码。很多论文中开源的代码可能是残缺的，可能仅仅是一个Jupyter，而没有环境参数、任务参数等等必要信息，甚至代码逻辑都是残缺的，也就无法将它复用在其他的任务上去验证效果。毕竟DNN模型是非常敏感的，batch大小、CUDA版本的变化、随机种子的变化等都可能大幅影响模型的效果。

同时，当一篇新的论文发表，代码开源了之后，评审员们可能也会要求比较，然而在GitHub上还是可以经常看到很多问题是没有回答的。

毋庸置疑，方法论的错误会导致延缓研究的进度，而文档健全的方法及代码可以让我们更加容易找到方法上的错误或者实验上的问题，所以作者认为应该类似[4]，发布开源代码的声明，明确约定发表论文的同时，应该发表什么样的代码，至少应该是易于使用且文档完备的代码。因为与使用实验来证明假设一样，代码也是科学研究中重要的组成部分。例如HuggingFace等机构，SentenceBERT等工作就做了相当好的示范。

发表偏好

由于多数教职职称或学生毕业会以论文发表数量作为硬性指标，绝大多数研究者会倾向于在A+类会议或Q1期刊上发表论文。所以发表论文数量和论文的引用量相比于其他方面，就更加重要。

所以，一些研究者们就会去抨击这种现象，认为不应当以论文数量为唯一的评价标准，他们主张“慢科学（slow science）”。但虽然这种想法理论上值得称赞，但致力于慢速，却并不符合多数研究者的需求。并且，实际上以论文数量来作为评价依据，或许是目前能找到的最为公平的一种方式了，毕竟这还是一个明确公共的指标，论文能否发表至少还是由第三方审稿人给出意见，并且是双盲评审，而如果不以上述为硬指标，则初级研究者就更难有出头之日。

然而现在每年A+区投稿的论文越来越多（网传斐波那契投稿法），审稿人也就有越来越少的时间去评审一篇论文，这也自然导致了很多优秀的论文却没有得以发表。

所以研究者们就只能要么让论文更加易于阅读，从而易于评判（适用于前文提到的描述方法的论文），或者找别的地方发表论文。所以很多研究者选择在ArXiv上先发表论文，以建立发表的记录。也多亏ArXiv的存在，研究者们可以在线分享自己的成果。但也由于ArXiv上未经审核，以现在NLP领域的研究步伐，早晚有一天，ArXiv上NLP领域也会被大量有偏置的模型淹没，GAU仍然会占据着整个领域。

所以，短期来看，一个比较可能的解决方式是加强审核的标准，发表真正有价值的工作。

算力不可获取

这一切还是要归于以GPT及BERT为起始的transformer系列的预训练语言模型。从GPT，到BERT，到GPT-2，到T5，到GPT-3，模型参数越来越大，打榜、比赛都进入了军备竞赛的时代，好像正如Sutton教授所说，大力真的出了奇迹。可是，BERT系列的模型（包括BERT、RoBERTa、ERNIE1/2等）都还可以进行科研实验，fine-tune做任务的代价并不是那么大，但是已经很难应用到实际的应用中，尤其是由高吞吐需求的线上应用。但到了T5-11B这种模型上，虽然分最高，但应用起来已经很吃力了。到了以GPT-3为首的千亿/万亿模型上，别说用起来了，找到一个硬盘去存储这个模型都很难。所以在EurNLP 2019的一次小组讨论中，Phil Blunsom提出：未来的NLP不在于更大的模型，而在于更大的想法。

> 在比赛战场上，CCKS2020的workshop中，我们可以看到，榜一和榜二几乎没有做任何算法上的优化，用着大数据、大模型加上集成，就大幅超越了其他的工作。榜一使用了25个RoBERTa-large集成，榜二使用了15个base和large模型集成，而榜三没有用任何的预训练模型和集成技术，生生用算法做出了榜单第三。那么相比来讲，是不是榜三才应该是更加有价值的工作呢？

这种Pretrain+fine-tune的范式，自然也决定了，只有豪横的公司、学校等研究机构，才有财力去支持庞大的算力，参与这个内卷的战场，而财力相对不足的机构，则因算力紧缺难以快速做出来实验。同时这也就导致了大模型实验在其他的地方难以复现。实验不可复现对于任何领域的科学研究都是具有风险的，例如社会心理学就因为实验不可复现而导致整个学科的声誉被玷污。

NLP领域的研究者们是希望看到的论文是可复现的，然而，[4]统计了506篇工作，发现其中只有15.61%的工作是可以复现的，与NLP领域相当高的数据共享比例形成了鲜明对比，而高共享的数据本该导向更高的可复现比例的。

笔者的看法

如前文所说，我们也不应该一味批评大模型所带来的资源浪费，以及给后来人所带来复现实验上的困难。大模型本身给我们展示了统计模型+海量数据能够展现出什么样的能力，研究者们对BERT进行的各种分析实验，包括延续着BERT诞生的RoBERTa，也表明了BERT类统计模型能够捕捉到的多元特征。的确BERT本体很难在各类线上应用使用，但并不代表它没有任何的应用价值。例如模型蒸馏，大模型就是非常卓越的teacher model，它带来的丰富特征可以大幅提升线上应用的小模型的使用。

GPT-3的出现，则是回应了另一个问题，当统计模型的参数继续上升，使用的数据量级持续扩大，统计模型又将是什么表现？它让我们看到了，凭借着记忆力，统计模型可以做到什么神奇的事情，同时也向我们暴露了统计模型的局限性在哪里（仅仅是记忆而不是理解，只能在语料内泛化，但泛化不可控，而事实不能泛化）。我认为，GPT-3所引起的后续一系列讨论才让大家冷静了下来，真正去反思大规模统计模型的局限性。还是如前文所说，我认为GPT-3的开发者们，LeCunn等大佬为首的讨论者们对此认知都相当清醒，但是如果没有GPT-3这样一个模型出现，又怎么样能够实实在在地去说服大家呢？

我在工作环境中发表看法的时候，对于使用集成模型来打比赛刷分的确是深恶痛绝的，一方面因为我的训练资源被挤占了，一方面我认为这对于个人参赛者，学界参赛者就是不公平的。可是，如果仅仅是从应用角度上，多个集成模型都作为teacher model，用来蒸馏一个应用模型，我认为也是相当可行的思路。

不可解释的方法

模型的可解释性在深度学习兴起之后，就一直是老生常谈的问题了。尤其GPT-3出现了之后，其在自然语言生成的表现相当抓眼球，一时间也让人们认为这种大模型已经能够当一个可以乱真的作者。而实际上，GPT-3生成的文章也是经过其大量输出编辑而成的最终结果，看上去一致性比较好，如果读者去试用它，则会很容易发现它的不可控泛化的case。研究者们对GPT-3模型生成的假新闻的担忧，以及模型生成结果对性别、种族上的偏见在去年也有广泛讨论。

我们说DNN模型结果的解释，也仅仅能说它反映了模型的训练样本中有什么现象，但完全没有办法去说明模型的结果到底是怎么来的。那也就是说，统计模型的可解释性本身就是个难以解决的问题，尽管有各种各样的研究去试图解释DNN模型，但那些工作给出的也更像是实验给出了一定的关联，没有得到明确的解释（实际上DNN模型的参数几乎是没有办法解释的，因为它始于随机，每一步训练进行的纠错究竟是在纠正哪些部分，或者模型学习到了样本的哪些关联或偏置，都是不清楚，且可能是多变的）。实际上我们也仅仅能够通过干预训练样本的分布去干预模型的表示，如我写的上一篇推文我删了这些训练数据…模型反而表现更好了！？，通过去除重复的训练样本，解决语言模型复读的问题。

现有的模型存在所谓的种族偏见、性别偏见之类的，反倒是最好解释的一种：训练样本中本来就存在这种偏差。

这种不可解释性，也注定无法将模型应用到需依赖于过程解释结果的领域，如法学和医疗领域。

然而，研究者们知道这一点，但到了PR工作上，又往往会对模型的能力有“报喜不报忧”的现象，最终使得公众对DNN模型的能力有了过高的估计。例如几年前某对话模型在机器多轮对话中出现了无意义的乱码，媒体的标题是机器发明了新的语言；例如AlphaGo Zero出现的时候，媒体的标题是无监督学习的胜利；例如ERNIE-3.0/GPT-3在PR的时候，使用了千挑万选的好结果，让大家认为大规模DNN模型就是无所不能。

所以最终笔者完全认同本文作者的观点，我们要更好地与媒体和公众接触，以确保来自这个领域的消息不仅仅是关于惊人的可能性的重大新闻，虽然让公众去理解我们的工作的局限性很困难，这些不够抓眼球，对于公众来讲很无聊，但这是确保公众去理解无法解释的模型的所有可能的结果的唯一方法。

> 有鉴于此，我在做我的开源项目宣传的时候，就喜欢极力避免读者有过高的估计，被认为有了我们的项目之后就可以直接端到端做到一些事情，以免起到反效果。

萌屋作者：Severus

Severus，在某厂工作的老程序员，主要从事自然语言理解方向，资深死宅，日常愤青，对个人觉得难以理解的同行工作都采取直接吐槽的态度。笔名取自哈利波特系列的斯内普教授，觉得自己也像他那么自闭、刻薄、阴阳怪气，也向往他为爱而伟大。

作品推荐

1. 深度学习，路在何方？
2. 数据还是模型？人类知识在深度学习里还有用武之地吗？
3. 在错误的数据上，刷到 SOTA 又有什么意义？

[1].Rogers A, Sedoc J, Rumshisky A. Proceedings of the First Workshop on Insights from Negative Results in NLP[C]//Proceedings of the First Workshop on Insights from Negative Results in NLP. 2020.

[2].Ettinger A, Rao S, Daumé III H, et al. Towards linguistically generalizable NLP systems: A workshop and shared task[J]. arXiv preprint arXiv:1711.01505, 2017.

[3].Norvig P. On Chomsky and the two cultures of statistical learning[M]//Berechenbarkeit der Welt?. Springer VS, Wiesbaden, 2017: 61-83.

[4].Belz A, Agarwal S, Shimorina A, et al. A systematic review of reproducibility research in natural language processing[J]. arXiv preprint arXiv:2103.07929, 2021.