

顶会观察

NeurIPS 2021

南京大学人工智能学院副研究员 叶翰嘉

神经信息系统大会 (Conference on Neural Information Processing Systems, NeurIPS) 是机器学习领域的顶级会议之一, 在国内外有广泛影响力, 被评为 CCF-A 类会议。由于疫情影响, 第 35 届 NeurIPS 大会于 2021 年 12 月 6 日至 14 日通过线上会议形式举办, 包括 1 天的 Tutorial、4 天的正式会议以及 2 天的 Workshop。

一、会议亮点

线上会议形式: NeurIPS 已连续两年通过纯线上会议形式举办。NeurIPS 2021 为注册者提供与其他参会者互动的权限, 例如向演讲者提问、参加现场海报环节、与其他与会者聊天和交流等。与 2020 年类似, NeurIPS 2021 线上海报环节使用 GatherTown 系统, 通过 Rocket.Chat 实现参会人员和作者的交流。2021 年度会议免费直播 Tutorial 和主题报告 Keynote, 并在会议结束后提供所有视频。会议相关动态也随着会议的逐步进行在其官方博客 (<https://blog.neurips.cc>) 上同步。

审稿流程变化: 首先, 不同于往年使用 CMT 系统进行投稿、审稿, 今年 NeurIPS 使用了 OpenReview 系统。为促进更加透明的审稿过程, 所有录用论文的审稿意见公开 (包括匿名审稿人的意见、最终 Meta-Reviewer 的意见以及作者的回复), 并接收非匿名的评论, 未录用论文的审稿意见由作者决定是否公开; 其次, NeurIPS 2021 审稿去除了提前拒稿 (Desk Reject) 阶段。NeurIPS 2020 会议在开始审稿前, 领域主席会对部分论文阅读后进行直接拒稿, 而 2020 年审稿实验表明约有 6% 被提前拒稿的论文在进行完整的审稿流程后

能够被录用。因此 2021 年只针对违反投稿规则的论文进行提前拒稿; 最后, 2021 年审稿过程中允许审稿人和作者之间进行多轮回复。往年使用 CMT 系统时, 作者在 Rebuttal 阶段后, 无法继续跟进审稿人的后续问题, 而在 OpenReview 系统中, 作者和审稿人之间可以使用类似论坛的形式针对某些疑问进行反复讨论。

检查列表 (Checklist) 要求: 不同于往年在投稿系统中进行一些检查项的选择, 今年投稿中需要作者在文末附上检查列表。其设计主要考虑到论文的可复现、透明性, 论文的研究伦理和社会影响等方面。目的是鼓励作者记录、思考论文的潜在问题和局限, 并在写作时尝试有针对性地解决这些问题。希望作者填写对论文结果复现步骤的说明, 是否附上代码和数据。Checklist 仅在投稿中使用, 论文录用后不强制要求作为附录出现。

伦理审查改进: 伦理审查造成独立于技术问题的挑战, NeurIPS 也不断尝试对伦理审查的过程进行改进。NeurIPS 于 2020 年已试行了伦理审查流程, 而 2021 年伦理审查规模进一步扩大, 有 100 多名伦理审稿人对论文审稿人筛选出的 265 篇论文 (2.9%) 进行了伦理审查。伦理审查目的是促进论文在伦理方面的思考, 主要包括偏见和公平性问题、研究诚信问题、法律合规性、不适当的潜在应用和影响 (如人权问题) 等。

数据集投稿 Track: 考虑到以往大量 NeurIPS 研究工作以算法为主, 在寻找有效数据进行算法评估上有一定困难, 且大量算法评估时只考虑到小型数据集, 会产生有偏评估结果, NeurIPS 2021 新增了数据集和基准 (Datasets and Benchmarks) 投稿 Track, 鼓励论文

对数据和任务基准进行高质量评估。作者投稿时需要在附录中说明数据集的收集、获取、维护、伦理等问题。对于数据集的评审和对于算法的评审有所差别，例如对于数据集的评审是单盲 (single-blind) 审稿，并且共分两轮投稿，审稿会专门考虑数据集是否能完整地对照算法和问题进行评估和描述。数据集 Track 录用的论文也会在会议中进行报告，并被收录到 NeurIPS 单独文集。

二、录用情况

NeurIPS 2021 共收到 9122 篇有效投稿，最终接收 2344 篇论文，接收率为 26%，达到近几年来最高水平。相较于 2020 年，2021 年投稿量降低 3.5%，录用率提升 23.4%。录用论文中有 56 篇 Oral 论文 (2.39%)，282 篇 Spotlight 论文 (12.03%)，其占比相对于 2020 年度有所降低。所有评审意见的平均分为 6.36 (审稿在 1-10 区间打分，10 分为满分)，Oral、Spotlight 和 Poster 论文的平均得分分别为 7.56、7.01 和 6.36。

录用论文的研究热点方向包括强化学习、深度学习、表示学习、优化、图神经网络、Transformer、对抗/鲁棒性等。谷歌 (Google)、斯坦福大学、麻省理工学院、卡耐基梅隆大学、加州大学伯克利分校、微软等科研机构、高校院所录用论文较多。NeurIPS 2021 继续进行审稿实验，共随机抽取 298 篇最终录用的论文 (包括录用为 Oral、Spotlight 和 Poster 的论文)，先后进行了两组评审。两组评审后，只有 77 篇 (26%) 论文得到了一致的评审意见，而剩余的 221 篇论文第二组评分均有下降，其中有 199 篇被第二组评审评为拒稿。

数据集 Track 共有两轮征稿，收到 484 篇论文，录用 174 篇 (录用率 35.9%)，其中关于计算机视觉的论文最多 (占 20%)，其他热门主题包括自然语言处理 (15%)、强化学习和模拟环境 (15%)、语音识别 (7%) 和多模态数据 (6%)，也有约 15% 的论文关于元分析以及 AI 公平性。所有论文中，有 55% 的论文提出新的数据集，20% 的论文提出评测基准，25% 的论文两者皆有。

三、邀请报告

NeurIPS 2021 邀请了普林斯顿大学名誉教授、美国科学院院士、诺贝尔经济学奖得主 Daniel

Kahneman 进行关于人类和机器智能的访谈，也邀请了 7 位专家进行报告 (Invited Talk)，具体内容如下：

How Duolingo Uses AI to Assess, Engage and Teach Better. 多邻国 (Duolingo) 联合创始人和首席执行官 Luis von Ahn 介绍了 Duolingo 的 AI 使用情况。Duolingo 是世界上最受欢迎的语言学习工具之一，报告介绍了多邻国是如何利用人工智能技术增加用户粘性，提高学习效果。例如，应用 AI 技术向用户推荐练习测试以提升语言的学习效率，以及应用 AI 对语言翻译进行评分、针对不同方面进行语言能力测试等。

The Banality of Scale: A Theory on the Limits of Modeling Bias and Fairness Frameworks for Social Justice (and other lessons from the Pandemic). 微软研究院的高级首席研究员 Mary L. Gray 介绍了为社会正义构建偏见和公平性框架的局限性理论。报告通过为北卡罗来纳州边缘化黑人和拉丁裔社区提供服务的社区组织 (CBO) 的实践案例，分析机器学习模型在公正和包容性抽样方面的局限性。目前在衡量数据和决策系统中的偏见和公平性方面理论研究的不足，使我们更加关注于收集数据的价值，而非对社区的价值。报告主要论证关注收集社区成员数据的需求，并观察计算上难以衡量但在质量上无价的社会作用对于推进社会公正的机器学习的必要性，以及对如何将计算机科学和机器学习重新定位为更明确的数据权力共享理论和实践提出建议。

Do We Know How to Estimate the Mean? 巴塞罗那庞培法布拉大学经济与商业系的 ICREA 研究教授 Gábor Lugosi 在报告中讨论统计学中一个基本的问题——基于独立观测的均值估计。随着机器学习和数据科学应用的发展，可以从统计和计算的新角度看待这个问题。通过回顾最近关于平均估计器的统计性能的一些结果，发现这些估计器允许高维数据中存在严重的长尾和对抗性污染。

Benign Overfitting. 加州大学伯克利分校电子工程系和计算机科学系教授、澳大利亚科学院院士 Peter Bartlett 在报告中探讨了模型过拟合的问题。深度学习

即使没有任何显式的努力来控制模型的复杂性，模型仍然完美地拟合有噪声的训练数据，并能展现出优秀的性能。报告介绍了在概率环境下如何进行准确预测的方法。对于线性回归问题，最小范数插值预测具有近似最优的预测精度。在这种情况下，过度参数化对于良性过拟合是必不可少的，即参数空间中对预测不重要的方向的数量必须显著超过样本大小。报告还讨论了对抗样本对鲁棒性的影响，并介绍了岭回归的扩展以及通过依赖于模型的泛化界限来分析良性过拟合的障碍。

Optimal Transport: Past, Present, and Future. 苏黎世联邦理工学院 FIM 数学研究所的讲座教授和主任、2018 年菲尔兹奖得主 Alessio Figalli 报告了关于最优运输的相关内容。在 18 世纪末，Gaspard Monge 为了得到如何将材料从一个地方运输到另一个地方来建造防御工事的最优解，提出了最优运输问题。在过去 30 年里，这个理论衍生出不同形式，在数学的许多领域都有应用。目前最优运输问题在机器学习诸多领域中大放异彩。报告内容概述了最优运输问题及应用，详细介绍了最优运输在金融、随机矩阵、生成式对抗网络、单细胞基因组等问题的应用，以及最优运输的复杂度分析。

Gender, Allyship & Public Interest Technology. 纽约大学 Arthur L. Carter 新闻学院的教授 Meredith Broussard 探讨了为什么大型计算机系统被困在 1950 年代关于性别的观念中，以及更新社会技术系统需要什么。考虑到从 2021 年 10 月开始，X 正式成为美国护照上的性别选项，而为了适应这种更具包容性的性别选择，需要对如何进行计算的更改进行探讨。报告也同时探讨了如何利用公益科技来超越性别二元思维，审查软件系统，并创造对社会有益的代码。

The Collective Intelligence of Army Ants, and the Robots They Inspire. 哈佛大学计算机科学系教授、ROOT Robotics 联合创始人 Radhika Nagpal 在报告中介绍了如何基于军蚁的集体智慧启发机器人。在自然界中，数以千计的个体组成的群体纯粹通过局部相互作用来创造复杂的结构，在这些系统中，尽管个体能力有限，但足以实现巨大的复杂性集体。报告讨论了怎样才能创造出达到自然所能达到的规模和复杂性的人工集

体。还介绍了 Eciton Robotica 的一个项目，该项目为了研究这个问题，利用来源于集体生物灵感创造的机器人系统。受军蚁筑巢的启发，该项目想创造一个可以自组装的软攀登机器人团体，这项工作横跨软机器人、新的自组织自组装理论模型和新的生物学领域实验。

四、 热点论文

2021 年度共有以下 6 篇论文获得最佳论文奖 (Outstanding Paper Awards)。

A Universal Law of Robustness via Isoperimetry. 本文提出了一个理论模型来解释为什么许多最先进的深度网络需要非常多的参数以平滑地拟合训练数据。特别地，在关于训练分布的某些正则性条件下， $O(1)$ -Lipschitz 函数对标签噪声下的训练数据进行插值 (interpolate) 所需的参数数目为 nd ，其中 n 是训练样本的数目， d 是数据的维数。这一结果与传统结果形成了鲜明对比。传统结果指出，函数需要 n 个参数来插值训练数据。而为了平滑插值，额外的因子 d 似乎是必需的。该理论与一些对 MNIST 分类具有很强泛化能力的模型的实验观察一致。这项工作还提供了关于开发用于 ImageNet 分类的鲁棒模型所需模型大小的可信预测。

On the Expressivity of Markov Reward. 马尔可夫奖赏函数是应对不确定序列化决策和强化学习的主要框架。本文详细阐述了马尔可夫奖赏何时足以或不足以使系统设计人员根据他们对特定行为的偏好、或对状态和动作序列的偏好来指定任务。作者用简单例证说明，存在一些特定任务，它们不能指定马尔可夫奖赏函数来推导出期望的任务和结果。作者也证明了在多项式时间内可以确定对于期望的设置是否存在相容的马尔可夫奖赏。如果存在的话，那么在有限决策过程中构造这种马尔可夫奖赏的多项式时间算法也就存在。这项工作揭示了奖赏设计的挑战，并可能为研究马尔可夫框架何时及如何足以实现人类所希望的性能开辟新的研究途径。

Deep Reinforcement Learning at the Edge of the Statistical Precipice. 本文提出了针对深度强化学习算法精确性评估比较的更加严格的方法。具体而言，对新算法的评估应该提供分层的自助法 (Bootstrap)

置信区间、不同任务和不同批次下的性能曲线、以及四分位数之间的均值。本文强调过去的一些文献报告中深度强化学习在多个任务和多次运行中的结果会导致很难评估新算法是否比过去的方法具有一致和可观的进步，并用实例说明了这一点。本文提出的性能总结方式能在每个任务少量运行的情况下进行计算，这可能是许多计算资源受限下的科研所需要的。

MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. 本文提出了 MAUVE 方法，用于衡量模型生成的文本分布和人类生成的文本分布的偏差。论文想法很简单，对于需要比较的两个文本的量化表示，使用连续的(软) KL 散度进行度量。MAUVE 实质上是对一系列测量的综合，其目的是同时捕捉第一类错误(生成不真实的文本)和第二类错误(未捕捉所有可能的人类文本)。实验表明，与以往的度量相比，MAUVE 识别了模型生成文本的已知模式，并且与人的判断更加相关。在开放场景文本生成快速发展的背景下，论文的结论有一定的借鉴意义。

Continuized Accelerations of Deterministic and Stochastic Gradient Descents, and of Gossip Algorithms. 本文描述了 Nesterov 加速梯度法 (Nesterov' s accelerated gradient method)的“连续”版本，其中两个独立的向量变量在连续时间内联合演化。这一场景类似使用微分方程来理解加速度的方法，但更新时使用的是由泊松点过程 (Poisson point process) 确定的随机时间下产生的梯度。这种新的方法是一种随机的离散时间方法，具有以下特点：(1) 拥有与 Nesterov 方法相同的加速收敛；(2) 提供了利用连续时间论点的清晰且透明的分析，比以前的加速梯度方法更容易理解；(3) 避免了离散化连续时间过程的额外错误，这与之前那些试图理解并使用连续时间过程的加速方法的尝试形成了对比。

Moser Flow: Divergence-based Generative Modeling on Manifolds. 本文提出了一种训练黎曼流形上连续规格化流 (Continuous Normalizing Flow, CNF) 生成模型的方法。其核心思想是利用 Moser (1965) 的一个结果，这个结果刻画了一个 CNF 的解，

它使用具有几何正则性条件的一类受限常微分方程，并使用目标密度函数的散度来显式定义。本文提出的 Moser Flow 方法利用这一解的概念，提出一种基于参数化目标密度估计器 (可以是神经网络) 的 CNF 方法。训练相当于简单地优化密度估计器的发散性，绕过了运行 ODE 解算器 (而这正是标准反向传播训练所必需的)。实验表明，与以前 CNF 工作相比，该方法训练时间更短，测试性能更好，并且能够在具有非常数曲率的隐式曲面 (如 Stanford Bunny 模型) 上对密度进行建模。这种利用几何正则性条件绕过昂贵反向传播训练的概念很有价值。

NeurIPS 的时间检验奖 (Test of Time Award) 一般授予 10 年前 NeurIPS 会议的论文。由于在 2020 年委员会考虑了更广泛的论文且选择了 2011 年而不是 2010 年的获奖者，因此，本年度 NeurIPS 时间检验奖委员会同时考虑 2010 年和 2011 年的论文，且由于 2010 年度未有论文获奖，本次时间检验奖重点考虑 2010 年度的论文。最终，2021 年度的时间检验奖授予 NIPS2010 年论文《Online Learning for Latent Dirichlet Allocation》，作者为 Matthew Hoffman、David Blei 和 Francis Bach。这篇论文提出了一种基于随机变分梯度的推断方法用于在大规模数据集上训练 Latent Dirichlet Allocation (LDA) 模型。在理论方面，它表明训练过程收敛到局部最优值，随机梯度更新对应于证据下限 (Evidence Lower Bound, ELBO) 目标的随机自然梯度；实验方面，作者首次表明 LDA 可以在含数十万个文档的文本语料库上训练，使其成为解决“大数据”的实用技术。该想法在机器学习社区产生了巨大的影响，推进了一般化的基于随机变分梯度的推断方法在更多模型上的使用。

数据集和基准 Track 共有 2 篇论文获得最佳论文奖 (Best Paper Awards)。

Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. 该工作分析了数千篇论文，研究了不同机器学习社区中数据集使用的演变，以及数据集采用和构建之间的相互作用。论文发现：在大多数社区中，数据集随着时间的推移越用越

少，且这些数据集来自少数机构。这种演变是有害的，使得基准变得不普遍，数据集来源中存在的偏差可能会被放大，且新数据集更难被接受。这对整个机器学习社区来说是一个重要的“警钟”，提醒研究者要更加批判性地思考哪些数据集用于基准测试，并更加重视创建新的和更多样化的数据集。

ATOM3D: Tasks on Molecules in Three Dimensions. 本文收集了一组具有小分子和生物聚合物 3D 表示的基准数据集，用于解决单分子结构预测、分子功能的设计和工程任务等问题。通过与最先进的 1D 或 2D 表示模型进行比较，论文发现简单鲁棒的 3D 模型具有更好的性能。这项工作为如何为给定任务选择和设计模型提供了重要见解。论文不仅提供了基准数据集，还提供了基线模型和开源工具，大大降低了机器学习人员进入计算机生物学和分子设计的门槛。

五、 总结展望

NeurIPS 2021 中强化学习、图神经网络、表示学习以及注意力机制等领域仍保持热度。而对公平性、可解释性的关注有所提升。本年度 NeurIPS 的重要变化包括审稿方式的变化以及数据集 Track 的加入。一方面通过这些尝试使得审稿过程更加透明化，作者和审稿人之间更容易交流，同时也使整个机器学习研究社群关注数据集、基准的构建。

在会议的形式上，NeurIPS 2021 继续通过纯线上会议形式进行，通过 2 年的尝试，线上会议模式已有多项改进。根据目前的信息，NeurIPS 2022 将在新奥尔良会议中心(美国)以线下会议的形式举办，但同时也会保留线上会议的形式。希望这种混合模式能够让 NeurIPS 的参与者有新的体验。

责任编辑 魏秀参



叶翰嘉

南京大学人工智能学院副研究员。主要研究方向为表示学习、模型复用等。
Email: yehj@nju.edu.cn