Expanding Access to ML Research through Student-led Collaboratives

Deep Gandhi
University of Alberta
drgandhi@ualberta.ca

Raghav Jain
Indian Institute of Technology, Patna
raghavjain106@gmail.com

Jay Gala
Unicode Research
jaygala24@gmail.com

Jhagrut Lalwani
Veermata Jijabai Technological Institute
jhagrutpradeep@gmail.com

Swapneel Mehta
New York University; One Fact Foundation
swapneel@onefact.org

Abstract

We present a model of a student-led community of researchers to highlight the impact of pursuing collaborative machine learning research on the group's members individually as well as towards achieving shared goals. We provide concrete examples of the guiding principles that led to the evolution of the collaborative from a reading group into a research group and eventually launching a non-profit software product to help non-technical stakeholders leverage artificial intelligence (AI), improving access to advanced technologies, and promoting open science. Our goal is to lay out a template to launch similar small-scale collaborative organisations at different institutes around the world.

1 Introduction

Machine learning (ML) has been a field that has historically borrowed heavily from information theory and often independently rediscovered known theories including the famous backpropagation algorithm first published, to our knowledge, by a collaboration between Bryson and Ho, an aeronautical engineer and a mathematician respectively. Collaboration often accelerates the (re)discovery of theories that were well-known in other fields and drives meaningful applications beyond the areas of impact envisioned initially. Cross-domain research collaborations in machine learning have been a cornerstone of various theoretical and practical advances in the field over the years, in particular, math and physics[1]Bronstein et al. [2021], Croitoru et al. [2022]. In a fast-growing field like machine learning, multiple individuals often study the same research questions. This has often led to the 'leaderboard-style' approach towards gathering a collective to solve a well-structured problem; often the development of a model to improve the test performance on a dataset of interest given a publicly available training dataset. The Netflix Prize [2] was a famously successful example of such an approach. In many of these situations, collaboration provides an effective middle-ground for the 'cooperation versus competition' choice that individuals may have to make.

The machine learning community has grown by leaps and bounds in the past decade with 9122 full paper submissions in NeurIPS 2021[3], since the first demonstration of AlexNet's superior performance in the ImageNet challenge, occasionally referenced as the 'watershed moment' for the deep learning community. This precipitated the creation of research labs and facilities dedicated to the pursuit of fundamental machine learning research both in industry and academia. As the field matured and

[1] https://jaan.io/how-does-physics-connect-machine-learning/

[2] https://en.wikipedia.org/wiki/Netflix_Prize

[3] https://www.vinai.io/an-overview-of-neurips-2021s-publications/

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

data became a first citizen for information technology, many of these research labs shifted their focus towards collaborations that advanced scientific progress such as the AlphaZero [Silver et al., 2017], AlphaGo[Silver et al., 2016], and AlphaStar[Vinyals et al., 2019] models for games and the AlphaFold[Jumper et al., 2021] for protein folding. These models arguably revolutionized their respective domains of applications, serving as examples of successful collaborations at the highest echelons of science. Importantly, though, these models were not initially released as completely open-source nor as an API service. So for the average researcher who does not have access to the compute to train and replicate these large models, there was limited utility from these advancements. To address this issue, there emerged novel collaboratives like the Leela Zero community[Pasqualini et al., 2022] to replicate the success of these models in the public domain, using crowdsourced compute, a novel system of using volunteers' devices to distribute training. This approach became quite popular with the advent of large language models like OpenAI's GPT-3 which was initially deemed "too dangerous to open-source"; with the EleutherAI community[4] training an open-source equivalent of the model. Clearly, collaboratives have contributed nontrivially to the democratisation of machine learning even ignoring the intangible impact via the upskilling of individual contributors and AI literacy efforts that the community often undertakes. A lot of the participants in these collaboratives are students at various levels of education. There has been little research conducted on the principles that guide the launch and growth of such focused collaboratives to help replicate these models in other domains of applied machine learning. In this perspectives paper, we present an evidence-based study of the guiding principles for a small student-led collaborative called Unicode Research through the lens of how it was able to support the requirements of individual contributors to the group using 'non-expert' mentors as the cornerstone of the organisation.

Many students that were part of this organisation came in without sufficient exposure to novel machine learning technologies and a large part of the experience related to how we provided a shared mentorship experience through non-expert mentors. The 'non-expert' refers in this case to peers who have some additional experience in the field, but do not hold a formal academic position at the postdoctoral position or higher; and are not highly-cited researchers in the field. This is arguably an imperfect metric for 'expertise', but provides a better way to quantify expertise than leaving it open-ended for the readers. A source of motivation to participate in this collaboration then is the provision of non-expert mentorship to every individual within the community. A student with a year of experience working in natural language processing research, for instance, could serve as a non-expert mentor to another whose experience is limited to computer vision, to help them pursue a topic modeling project.

2 Related Literature

Mentorship has been often studied under various contexts, in education, psychology, research pedagogy, and other domains. The authors of [Rose, 2013] studied how effective teachers are in the learning process of the student. Various studies have been covered by Rose [2013] which dealt with identifying how factors such as the home environment and family [Todd and Wolpin, 2007] affect student progress. Furthermore, there have been multiple factors including school resources [Hedges et al., 1994], peer diversity [Kimball et al., 2004], and teacher preparation [Goldhaber and Brewer, 2000] that affect the educational achievements of the student. The authors of [Rose, 2013] also put forth the point that the average student performance is not a good indicator of the effectiveness of teachers. These results broadly influence how the effectiveness of non-expert mentorship in the case of students can and should be measured.

In the rest of this perspectives paper we focus on the Unicode Research organisation and the influence the group has had on the current members through an anonymous survey collecting the group's opinions. These are not causal claims, and likely suffer from a sampling and selection bias. They are intended to provide an early and largely informal view preceding principled research that we plan to conduct to understand the impact of providing non-expert machine learning mentorship to students.

[4] https://www.eleuther.ai/

3 Launching a Research Collaborative

Unicode Research started as a distributed research collaborative group to guide students, and early-career artificial intelligence enthusiasts, and researchers to conduct impactful ML research and help them in starting their careers in the field of AI/ML. Unlike many well-known universities across the world where students have access to quality resources and guidance in addition to well-structured programs like undergraduate research opportunities (UROPs), such research is incredibly challenging to conduct at most universities that our students come from where you have not only a lack of experts to guide research but also a lack of adequate resources to conduct it. To bridge the knowledge gap and help put together resources to guide the students, we started Unicode Research as a medium to connect undergraduate students and early career AI researchers with more experienced mentors and researchers.

In the first year of the Unicode Research organisation, we supported students in developing a research mindset, and understanding existing research papers. As the next step, we created a pathway to learn more about advanced statistics and machine learning through a class based on the *Advanced Data Analysis from an Elementary Point of View* textbook by Prof. Cosma Rohilla Shalizi[5]. Our goal at Unicode Research is to help students conduct quality research and finding relevant and meaningful research areas. However, it is challenging to identify a research area of interest for someone without prior experience in ML research, let alone in the area itself. To guide students to find a suitable research direction, the first order of business we laid out was to have them jointly present a few papers in the area of interest, and answer questions from the group forcing them to think hard about other problems in the area. This led to a series of reading group presentations where we:

- Study research papers in smaller groups
- Discuss the strengths and limitations of existing literature
- Conduct brainstorming sessions for generating new research ideas

We received a grant to teach machine learning to over 100 other undergraduate students at Tier II and III universities through an online ML summer course lasting 10 weeks[6]. This helped us provide a well-compensated opportunity for students to serve as a section leader or a teaching assistant in the summer course. They were also project advisors for a set of undergraduate teams with 7 teams of 4-6 students ultimately presenting on Demo day. We believe that these opportunities starting with reading groups leading up to independently teach ML to their peers helped the students grow into independent researchers with a strong set of software and communication skills. A large number of participants joined Unicode Research at the end of the free ML course delivered by the group. These members were then trained to improve their research aptitude by the existing members.

4 Feedback from Participants in Unicode Research

We conducted an open-ended survey of members who have worked as a part of the Unicode organization to get a cursory understanding of the impact and shortcomings of such an organisation. While a larger undertaking is necessary to obtain representative opinions, we selected six team members at different skill levels to provide a sufficiently descriptive study of a small research collective.

4.1 Participants' Background

While each participant was a skilled software developer, most had very limited access to research opportunities due to the lack of access to resources at the institutional level. All of the surveyed members started out in the field of ML through traditional channels–taking courses and developing relevant projects–during which each of them faced some obstacles. Lack of access to quality research labs at their institute, lack of resources to work on certain ideas involving larger-scale data analysis than is possible on a single machine slowed their progression in the field.

[5] https://www.stat.cmu.edu/ cshalizi/ADAfaEPoV/

[6] https://djunicode.github.io/umlsc-2021/

4.2 Experience with Collaborative ML Research

Every participant preferred to work in a smaller team rather than in a larger group. This underscores how a smaller, motivated group could prove a better fit to pursue hard problems than larger ventures that tend to lose steam quickly. Most of the participants shared positive experiences while working with collaborators at Unicode Research. However, some negative experiences that the participants faced were delays on the collaborators' part in the completion of shared tasks and difference of opinions among collaborators. A majority of these 'blockers' were considered to have been handled with discretion and the overall experience of participants seemed to be positive. With respect to the evolution of research agenda, the individuals suggested that the current direction of work being done by the group highly aligns with their research interests.

Learning: Members of Unicode Research are spread across 4 continents as it is a remote research collaborative group. Most members are either full-time students or pursuing full-time employment with part-time research. We primarily execute projects with a high dependence on online tools [7]. As a remote community, we found asynchronous work is accelerated by dedicated documentation of active projects that helps onboard new members quickly. We have found that having a dedicated 'handbook' with a detailed walkthrough that participants can follow is extremely useful.

4.3 Key Takeaways

Some people valued the connections developed while working on different projects, while the others highlighted how Unicode Research taught them about working on their soft skills such as client interactions. An important skill that some participants picked up was structuring and completing large-scale projects, which included UMLSC, a summer Machine Learning course for undergraduate students hosted by the group and supported by Google Research India.

4.4 Future Directions

A large number of the participants involved with Unicode Research are currently collaborating towards building a social enterprise that models how information flows online. The project studies online audiences using social networks and Google Analytics data. They are helping investigative journalists to better understand the impact of online news. This includes the detection of artificially promoted state-sponsored media and monitoring of risks from toxic speech arising from different social network communities.

Some of the senior participants are putting into place a governance structure to expand the group into a larger cooperative with focused collectives inside the organisation working on research-focused software, conversing with external stakeholders and policymakers, and project management. As organizers, we are actively scaling the Unicode Research initiative so that more participants can have access to computing resources, research mentorship, and career growth opportunities in the sub-field of their choice.

5 Conclusion

In this paper, we discussed how small and focused research collaboratives can help students from a variety of semi-technical backgrounds develop into ML researchers whilst driving real-world impact. While dealing with a diverse group of participants with different research agendas, the model that Unicode Research provides illustrates the impact of such collaboratives on defining and achieving shared goals. This paper provides a comprehensive breakdown of a collaborative research organisation that can help to create a research culture at other institutes that may or may not be dealing with similar challenges. We are using this template and advising on the launch of a student organisation at New York University that has similar goals[8].

[7] https://www.when2meet.com/

[8] We would like to thank Priyambi Hiran for helping us with the data collection and the One Fact Foundation for their support throughout our initiative.

References

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*, 2022.

D. D. Goldhaber and D. J. Brewer. Does teacher certification matter? high school teacher certification status and student achievement. *Educational evaluation and policy analysis*, 22(2):129–145, 2000.

L. Hedges, R. Laine, and R. Greenwald. Docs money matter? a meta analysis of studies of the effects of differential school inputs on student achievement. *Educational Researcher*, 23:9–10, 1994.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

S. M. Kimball, B. White, A. T. Milanowski, and G. Borman. Examining the relationship between teacher evaluation and student assessment results in washoe county. *Peabody Journal of Education*, 79(4):54–78, 2004.

L. Pasqualini, M. Parton, F. Morandin, G. Amato, R. Gini, and C. Metta. Leela zero score: a study of a score-based alphago zero. *arXiv preprint arXiv:2201.13176*, 2022.

R. A. Rose. *Teacher effectiveness and causal inference in observational studies*. PhD thesis, The University of North Carolina at Chapel Hill, 2013.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

P. E. Todd and K. I. Wolpin. The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human capital*, 1(1):91–136, 2007.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.