Investigation into the Efficiency and Effectiveness of Diffusion Modeling in Predicting Calorimeter Particle Showers
(Technical Paper)

The Social Cause and Effect of Bad Science
(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia | Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Luke Ostyn
Fall 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Advisors
Geoffrey Fox, Department of Computer Science
Travis Elliot, Department of Engineering and Society

Technical Paper - Investigation into the Efficiency and Effectiveness of Diffusion Modeling in Predicting Calorimeter Particle Showers

Introduction

Calorimeters are a staple of particle physics, used ubiquitously to determine the energy associated with various particles. When these particles enter the calorimeter they break apart rapidly into multitudinous lower energy particles, a so-called particle shower. Traditionally, attempts to model these showers have been overly laborious and computationally expensive, meaning that the development of new, more lightweight methods is of great import (Mikuni & Nachman, 2022). For my technical project, I explored the use of deep learning methods to produce high accuracy predictions of these calorimeter showers. Specifically, I investigated the efficacy of normalizing flows models by expanding on the CaloScore diffusion model of Mikuni and Nachman (2023). Beyond just looking to increase the accuracy of the model, I also attempted changes designed to increase the speed with which the model could be trained. Although the model once trained is significantly faster than traditional methods, the training can certainly be a roadblock especially when lacking access to a distributed system. As such, I thought it pertinent to look into methods to abbreviate the training without substantially compromising model accuracy. The original model and my subsequent alterations were built to perform against the datasets available from the Fast Calorimeter Simulation Challenge 2022 hosted by Kaggle.

The Model

Broadly speaking, diffusion models are trained by taking our data and then through an iterative process adding and removing noise; we take our data, add noise to it, and then try and train a model that is able to recover that original data. When we then pass something resembling

noise back through the model it should produce an output similar to the data we trained it on (Chang et al.). For the base CaloScore model, this involves training just under two and a half million different parameters on data passed throughout a series of different networks. This data is composed of the input energies associated with the particle entering the calorimeter and then the energies associated with individual physical voxels and layers during shower generation within the calorimeter. As explained by Mikuni and Nachman, in a given training cycle, noise is added to the layer and voxel data before they are input to Resnet and U-Net architectures respectively. The Resnet architecture consists largely of a series of convolutional layers while upsampling, residual, and downsampling layers comprise most of the U-Net architecture (2023).

Methods

The methods developed for testing variations on this model were in part constrained by limitations of the system upon which they were trained. Because the original model was designed for a distributed system, utilizing parallelization techniques which were removed for the purpose of training on my personal laptop, the runtime for a single training epoch ballooned quite considerably, up to approximately two hours per epoch. With the training process taking up to three hundred epochs to complete, obviously fully training the network for multiple different variations was simply unfeasible. Instead, I opted to look at loss statistics after an epoch or two. Moreover, for similar motivations related to computational intensity, I opted to test these changes using solely the first dataset available from the Kaggle challenge.

Generally, my attempted tweaking of the model took one of two different forms: adjustments to either the architectures themselves or to the hyperparameters in charge of controlling the process by which the weights associated with these architectures are updated. In the first category are changes to both the U-Net and Resnet as well as the greater CaloScore

network which frames them. For the U-Net, these changes include an increase to the numbers of residual layers and a separate decrease in the number of sampling layers (both upsampling and downsampling). As for the Resnet, I experimented with adding a significant number of convolutional and dropout layers. Finally, I modified the process by which noise is added, generally decreasing the degree of noise appended. As for hyperparameter modifications, I tested various learning rates and data sampling time steps, observing in large part whether convergence could be faster induced.

Results and Discussion

After the completion of a training epoch, several different loss performance metrics are reported. These include loss, voxel loss, and layer loss, where loss is just the sum of the voxel loss and layer loss. As the names suggest, voxel and layer loss each report a square loss associated with the independent voxel and layer models. These statistics are reported in Table 1 for a myriad of different modifications to the default CaloScore model. Critically this data will

Table 1

Loss Statistics for Assorted Modifications to the CaloScore Normalizing Flows Model

Modification	Loss	Voxel Loss	Layer Loss
Base	1.1396	0.4949	0.6477
Removed U-Net Sampling Layers	1.2400	0.5935	0.6464
Increased U-Net Residual Layers	1.1330	0.4913	0.6417
Increased Number of Resnet Convolutional Layers	1.1354	.4971	.6383
Increased Learning Rate	1.0310	0.4286	0.6024
Fewer Time Steps	1.1396	0.4949	0.6447

Decrease Noise	2.4349	1.5328	0.9021

help us to answer the dual questions of whether these modifications were able to improve performance by either decreasing loss or keeping loss stable and increasing speed.

Let us first examine those changes made to the CaloScore architecture. More specifically, we will start with those made to the U-Net voxel model. The first alteration made was to remove the sampling layers, which reduced the number of trainable parameters in the overall network by a factor of about four. Despite this significant drop in network complexity, we failed to see any serious reduction in epoch training time. This fact, in addition to the loss and voxel loss numbers being worse than the baseline, indicates that these sampling layers are necessary and should not be excluded. To the contrary, when we increased the number of residual layers in the same U-Net, we observed a slight decrease in voxel loss and the attendant decrease in total loss. While the change obviously did not help to reduce training time, the data suggests that boosting the number of residual layers can produce small payoffs. Similarly, increasing the number of convolutional layers within the Resnet prompted the total and layer loss values to exhibit small improvements. In the same vein as the U-Net modification, this shows that an increase in trainable layers can lead to small dividends in decreasing loss. Finally, attempting to decrease the amount of noise added to the model produced loss statistics significantly worse than that of the normal model. Moreover, the training lost witnessed a significant plateau over the course of the run, suggesting that little improvement was occurring.

As for the final two changes, these changes were largely implemented with the goal of reducing training time. With the increased learning rate, we can discern that after a single epoch all of the model's loss statistics were superior to that of the base model, which is indicative of the fact that the higher learning rate induces faster convergence to a reasonable model. While the

final result after hundreds of epochs might be inferior to that produced by the base model with the stunted learning rate, for the purposes of faster training, increasing the learning rate proved effective. Secondly, I looked into drastically reducing the number of time steps involved in creating samples during the training and testing stages of the CaloScore model. Curiously, doing so resulted in no change at all to the loss statistics but decreased the runtime significantly, demonstrating that at least for the first kaggle dataset, cutting down the number of time steps involved in sample generation is a method worth exploring.

Conclusion

Calorimetry, an evergreen field within particle physics, is notoriously difficult to build models for because of the exceptional computational power needed. This is where sophisticated deep learning techniques enter the fray, with the hope that we can use them to train vast networks and unlock previously untapped modeling capabilities. One such technique, normalizing flows, was explored in this paper through the lens of the CaloScore module. The primary focuses were to both improve its effectiveness and the speed at which it could be trained by altering various parameters. The results obtained submit that greater effectiveness may be possible when both the U-Net and Resnet components of the model are expanded to include more residual and convolutional layers respectively while increasing training speed follows from a heightened learning rate or a cutback on the number of time steps during sample generation. Future research should be geared towards testing these modifications for a greater number of epochs on a more capable distributed system.

STS Paper - The Social Cause and Effect of Bad Science

Introduction

Consider the eugenics movement of the early 1900's: reprehensible and abhorrent, in the United States it led to widespread sterilization of marginalized communities and in Nazi Germany it provided the foundation for the Holocaust. Of course, underpinning these atrocities was the scientific consensus of the time. So, do we think this science was built on good data, data produced by experiments adhering to proper implementation of the scientific method. Maybe more importantly, was the science guided by well-intentioned individuals? The answer to both of these questions is undoubtedly no. It follows that if we desire to prevent such moral catastrophes we must also discontinue such bastardizations of science. More broadly, we will look to comprehend what compels scientists to falsify data, to cut corners, or to manufacture results: in general, to conduct bad science. To this end, it is imperative that we understand both the prevailing social forces and, when negative, whether it is feasible to mitigate or alter them.

However, bad actors and falsified data is not where this story ends. Good people with the best of intentions can still produce results which are flawed, in some cases incredibly so. Across science, especially in psychology and related domains, this situation is referred to as the replication crisis; in myriad instances, studies generate results which future studies at minimum fail to recreate and in the extreme completely contradict. This paper will seek to understand both of these phenomena, science that fails deliberately and science that fails accidentally.

Frameworks

While these issues are undoubtedly connected, we will largely tackle them in isolation. That said, it is fortunate that both should be tractable when investigated under the auspices the Co-production of Science and Social Order, a framework which, as explained by Jasanoff, emphasizes "this self-conscious desire to avoid both social and technoscientific determinism in

S&TS accounts of the world" (2004). Within this framework, we reject any notion that the flow of control between technology and society is unidirectional. Rather, society and scientific knowledge mutually establish and create the other (Swedlow, 2011). In the case of scientific dishonesty, this bidirectionality is especially important because it allows us to shed light on the societal factors which impel this illegitimacy (social factors influencing science) and also the damaging downstream effects of this fabricated knowledge (science influencing society). The latter and largely more subtle phenomenon, science unintentionally erring, promises to be similarly amenable to this Coproduction of Science and Social Order framework. We will look at social influences in concert with flaws in the scientific method itself to determine what goes wrong while again observing societal impacts of these mistaken ideas. In addition, an emphasis will be placed on looking at both successful and failed theories, so that comparisons can be drawn between their conceptions.

Scientific Dishonesty

When it comes to scientific dishonesty, there are two main forms of research misconduct which we want to highlight: falsification and fabrication. (Plagiarism is the third major type, but it falls outside the purview of this paper.) While similar, there are some important distinctions between the two. Falsification refers to adjusting values measured during experiment while fabrication is when results are completely manufactured (i.e. the experiment was never actually performed) (Bik, 2019). In modern history, a good case study is that of Bell Labs researcher Jan Hendrik Schön, who between 1998 and 2001 fabricated data for sixteen different studies within the fields of organic electronics, superconductivity, and nanotechnology. Upon discovery of these infringements, he was promptly fired (Service, 2002). Even more recent are the developments coming out of Palo Alto, where the former president of the university has stepped down amid

distressing reports surrounding his past research. Despite the original claim that he had falsified data being walked back, they still allege a negligence in correcting mistakes within his papers and a complicity in fostering a culture within his lab agnostic towards the topic of research integrity (Kaiser, 2023).While not as extreme as the first example, the story indicates that there is still much to be desired when it comes to establishing research norms conducive to the production of good science. So, what might drive Jan Hendrik Schön to lie about his results and why might Marc Tessier-Lavigne not have felt that unwavering commitment to truth that we would expect? The conventional wisdom is to say that they felt a "pressure to perform"; they believed that if they did not continuously pump out research, then it might have been their head next on the chopping block. However, according to Fanelli et al., this hypothesis is not completely borne out by the data. Instead they point to the most crucial factor in predicting falsification and fabrication to be the degree of social control present in the research setting (2017). Scientific dishonesty manifests not when researchers feel pressure, but when they feel like they can get away with it.

Reproducibility and Replicability

To begin, it is necessary that we understand the difference between reproduction and replication. Whereas replication requires the entire study to be repeated, reproduction refers to when only the analysis is repeated. Essentially, the difference is that replication requires new data while reproduction does not. To understand why they are important, we can hearken back to the words of philosopher Karl Popper: "we have seen that non-reproducible single occurrences are of no significance to science" (Popper 66). Without reproducibility, we have nothing. Good results are those that hold up under scrutiny and can be repeated ad infinitum. And in many scientific fields we generally have this; there will be the occasional bad apple, but this is really

just part of the process of doing science. As we are able to make more fine-grained measurements, we discover that our theories aren't as airtight as we once hoped, so we evolve them accordingly and perform new experiments to see if our updated theories do indeed pass muster. Unfortunately, this is not the case within every field. In a remarkable study performed by Nosek, it was shown that across a sample of one hundred psychology studies only thirty nine were able to be reliably replicated (2015). The idea that over fifty percent of studies within psychology are not to be trusted is shocking and clearly indicative of structural issues within the greater scientific apparatus.

Looking Forward

Going forward, the focus will be first on continuing to evaluate the various factors which drive the creation of defective scientific knowledge. For that knowledge created through scientific dishonesty these factors will largely be social; for that created through methodological failures, the factors will likely be more fundamentally intertwined with how we do science itself. Secondly, there will be an emphasis on observing how knowledge when improperly arrived at can detriment society. In general, this will be done by looking at multiple case studies and attempting to both draw out the factors that contributed to their success or failure and the impact that success or failure had on society. Finally, if possible, a prescription will be offered to try and reduce the likelihood that similar failures continue to occur.

Conclusion

When science fails it can have disastrous effects for society, and as such, it is imperative that we attempt and figure out why they happen. In the cases of both unintentional failures and falsification and fabrication, we will call on Co-production of Science and Social Order. Beyond helping us to outline the existing social pressures underlying such scientific frustrations, this

framework will also assist in elucidating their short and long term effects. Thus, we can understand both what happens when we make a methodological blunder and how that blunder was allowed to be made. We understand why we desire to arrive at truth and what we can do to make that a reality.

Works Cited

Bik, E. (2019, May 29). *What is research misconduct? part 3: Fabrication*. Science Integrity Digest. https://scienceintegritydigest.com/2019/05/29/what-is-research-misconduct-part-3-fabrication/

Chang, Z., Koulieris, G. A., & Shum, H. P. H. (2023, October 19). *On the design fundamentals of diffusion models: A survey*. arXiv. https://arxiv.org/abs/2306.04542

Fanelli, D., Costas, R., Fang, F. C., Casadevall, A., & Bik, E. M. (2017, April 12). *Why do scientists fabricate and falsify data? A matched-control analysis of papers containing problematic image duplications*. bioRxiv. https://doi.org/10.1101/126805

Jasanoff, S. (2004b). *States of knowledge: The co-production of Science and Social Order*. Routledge.

Kaiser, J. (2023, July 19). *Stanford president to step down despite probe exonerating him of research misconduct*. Science https://www.science.org/content/article/stanford-president-to-step-down-despite-probe-exonerating-him-of-research-misconduct

Mikuni, V., & Nachman, B. (2022, October 19). *Score-based generative models for calorimeter shower simulation*. arXiv. https://arxiv.org/abs/2206.11898

Mikuni, V., & Nachman, B. (2023, August 7). *CaloScore V2: Single-shot calorimeter shower simulation with diffusion models*. arXiv. https://arxiv.org/abs/2308.03847

Nosek, B. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Popper, Karl (2005). *The Logic of Scientific Discovery.* Taylor & Francis. http://philotextes.info/spip/IMG/pdf/popper-logic-scientific-discovery.pdf

Service, R. F. (2002, September 25). *Physicist Fired for Falsified Data*. Science. https://www.science.org/content/article/physicist-fired-falsified-data

Swedlow, B. (2011). Cultural coproduction of four states of knowledge. *Science, Technology, & Human Values*, *37*(3), 151–179. https://doi.org/10.1177/0162243911405345