TRUSTWORTHY MODEL EVALUATION ON A BUDGET

Iordanis Fostiropoulos, Bowman Brown, Laurent Itti
University of Southern California
Los Angeles, CA, USA
{fostirop,bnbrown,itti}@usc.edu

ABSTRACT

Standard practice in Machine Learning (ML) research uses ablation studies to evaluate a novel method (Meyes et al., 2019). We find that errors in the ablation setup can lead to incorrect explanations for which method components contribute to the performance. Previous work has shown that the majority of experiments published in top conferences are performed with few experimental trials (less than 50) and manual sampling of hyperparameters Bouthillier & Varoquaux (2020). Using the insights from our meta-analysis, we demonstrate how current practices can lead to unreliable conclusions. We simulate an ablation study experiment on an existing Neural Architecture Search (NAS) benchmark and perform an ablation study with 120 trials using ResNet50. We quantify the selection bias of Hyperparameter Optimization (HPO) strategies to show that only random sampling can produce reliable results when determining the top and mean performance of a method under a limited computational budget.

1 INTRODUCTION

Machine Learning (ML) experimental results are sensitive to variations in the training process that can affect the conclusions of the analysis (D'Amour et al., 2020). Stochasticity in the training process can lead to results that are difficult to reproduce (Ahmed & Lofstead, 2022) and statistical outliers that can be misleading (Picard, 2021). The effect can be further exaggerated by the computational budget of the experiment (Shwartz-Ziv & Armon, 2021).

Ablation studies are a type of ML experiment where a method is evaluated with different components removed to identify the component's contribution to the performance of the method (Meyes et al., 2019). In contrast to hyperparameter optimization (HPO), which seeks the hyperparameter values that lead to the best performance, ablation studies are used to explain a method's performance (Lipton & Steinhardt, 2018).

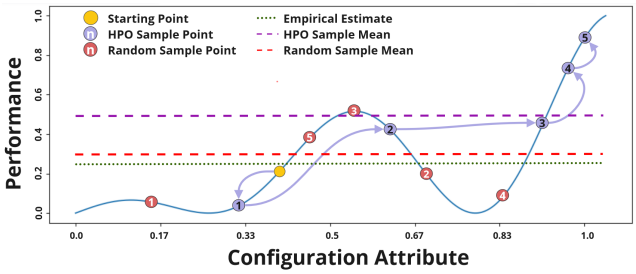

Figure 1: Comparison of two hyperparameter selection strategies, Random and HPO. An HPO strategy selects points based on previous samples and is *biased* towards a small fraction of the search space that shows good performance. HPO can overestimate the mean performance, making statistical analysis such as ANOVA inapplicable. The bias is exacerbated by the computational budget, since HPO methods may sample from a small range of attributes.

ML experiments are composed of trials where each trial defines the training process and evaluation of a single model, and a budget that specifies the number of trials allocated to the experiment. Vari-

ation between the results of different trials means that a limited number of trials may not accurately represent the mean performance of an experiment Nagarajan et al. (2018).

Several works (Balaji et al., 2018; Bouthillier et al., 2019; 2021) analyze current experimental practices in studies published in top conferences. They find that most experiments are run with a small number of trials (a small budget) and are evaluated with hyperparameters that are sampled manually (Bouthillier & Varoquaux, 2020) or with HPO. As we show in this work, these practices can lead to unreliable conclusions if evaluated incorrectly; fig. 1.

We evaluate current experimental practices for ablation studies and compare the effect of computational budget and hyperparameter selection strategy on the reliability of the analysis. Based on our observations, we propose specific experimentation strategies for ablation studies under a limited computational budget. We will open-source the code used for our experiments once the review process concludes.

2 Related Work

Bouthillier & Varoquaux (2020) conduct a survey of publications at top ML conferences (NeurIPS and ICLR) and report that $\sim$45.65% of experiments perform manual hyperparameter selection while only $\sim$7.44% of experiments perform random search. Additionally, $\sim$65.37% of publications performed fewer than 50 experimental trials. Bouthillier & Varoquaux (2020) recommend the use of HPO as an alternative to manual sampling.

Turner et al. (2021); Moosbauer et al. (2021) show that the choice of hyperparameter selection strategy can have a statistically significant impact on the results where under-explored regions of the hyperparameter space can lead to unreliable conclusions.

Kurach et al. (2019); Lucic et al. (2017); Eggensperger et al. (2021); Liao et al. (2021) demonstrate that the computational budget can have a significant impact on the results of the analysis. Similarly, Picard (2021) find that with a sufficiently large budget, statistical outliers can out-perform state-of-the-art.

Similar to previous work, we find that current experimental practices can be responsible for errors in the analysis. Our results suggest that the interaction effects between computational budget and hyperparameter selection strategy can further exacerbate the unreliability of the results. In contrast to previous work, we find that evaluations using the mean performance and random sampling are reliable estimates of a method's performance even with a small budget.

3 Experiment

3.1 Number of Trials

We perform an experiment on the NATS-Bench benchmark (Dong et al., 2021) dataset that includes 7,776 neural network architectures composed of an exhaustive search over 6 topological components, where each component can be one of 5 different layers. The benchmark evaluated the performance of network architectures on the CIFAR-10 dataset (Krizhevsky, 2009) with multiple random repetitions. We define two model variants in the NATS-Bench for models with skip connections (ResNet) (He et al., 2015) and networks without (Inception) (Szegedy et al., 2014); $\mathcal{M} = [\text{'Res'}, \text{'Inc'}]$. The exhaustive search of configurations and multiple repetitions of each trial provide a good approximation of the *empirical mean* performance of each model variant, which can be used as the ground-truth for our experiments.

We evaluate how the sampling bias can lead to incorrect analysis as the number of trials increases for different HPO strategies. We use surrogate modeling of the accuracy score using a Gaussian-Process (GP) (Paleyes et al., 2019) with different acquisition functions such as sampling from regions with the highest expected performance, 'greedy' ('GP_G'), highest expected variance ('GP_V'), and both high expected performance and high variance, ('GP_UCB') Srinivas et al. (2009). We evaluate Tree-structured Parzen Estimator ('TPE') Bergstra & Bengio (2012) where the trials are sampled based on the Expected Improvement. Lastly, we use Latin Hypercube ('Quasi'-random) and pseudo-random ('Random') sampling.

We simulate experiments of varying budgets for each HPO strategy for 10 repetitions and with different random seeds. We quantify the relative error as the difference between the estimated mean performance of a method and the **empirical mean** performance of the method from the dataset: $\Delta\mathcal{A}_{\mathcal{M}} = \mathcal{A}_{\mathcal{M}} - \mathcal{A}'_{\mathcal{M}}$. We define the **best** performance error $\Delta\mathbf{A}_{\mathcal{M}}$ similarly. We calculate the **bias** of the sampling method as $\Delta\mathbf{P}_{\mathcal{M}} = P(\mathcal{M}) - P'(\mathcal{M})$ where $P(\mathcal{M}) = \frac{N_{\mathrm{M}}}{N - N_{\mathrm{M}}}$ is the empirical sampling probability of $\mathcal{M}$ in the NATS-Bench dataset and P' is the sampling method estimate.

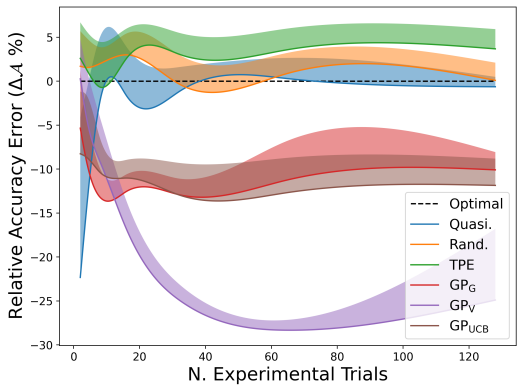

Figure 2: Sampling bias in different selection strategies on the NATS-Bench dataset. Random provides the least biased estimate and is closest to the optimal while remaining sample efficient. The majority of strategies lead to sampling sub-optimal architectures (negative relative error) from regions with unstable training dynamics and poor performance. TPE leads to sampling more often from optimal regions of model performance (positive relative error).

The *empirical* performance of ResNet architectures is $\mathcal{A}_{\text{Res}} = 90.20 \pm 4.33\%$ and Inception architectures is $\mathcal{A}_{\text{Inc}} = 81.50 \pm 20.07\%$. The dataset is biased towards ResNet variants by $P(\textbf{Res}) = 1.48$. Table 1 shows the relative error as the average for all methods $\mathcal{M} = [\text{'Res'}, \text{'Inc'}]$.

Our results from fig. 2 indicate that the relative bias and variance decrease as we increase the computational budget for Random. In contrast, the relative difference increases for the 'TPE' and GP methods. TPE samples architectures from only the optimal regions leading to above-optimal mean performance, while GP methods sample from regions of high variance that can lead to sampling from regions of sub-optimal mean performance.

When evaluating with the best-performing trial in table 1, our results show an equivalent top performance between all sampling methods. Additionally, the incremental benefit of additional trials beyond 64 begins to decline. Our results suggest that with a moderate budget of 64 trials, an unbiased estimate of model performance can be obtained with Random when evaluating the mean performance of a method. The relative error when evaluating the best trial was similar for TPE and Random at 1.21 and 1.15. Our results suggest that a Random sampling strategy might be a more reliable option under more general conditions to ablation experiments.

Table 1: Increasing the computational budget ('Trials') reduces the relative mean error $\Delta\mathcal{A}$ and Resnet bias $\Delta\mathbf{P}_{\text{Res}}$ for Random ('Rand.'). The opposite is true for HPO methods ('TPE' and GP_{UCB}). The relative error of the top-performing trial ($\Delta\mathbf{A}$) falls as the budget increases for all methods. As such, we conclude that Random is the only choice that can produce reliable, unbiased results.

	$\overline{\Delta\mathcal{A}} \downarrow$			$\overline{\Delta\mathbf{A}} \downarrow$			$\Delta\mathbf{P}_{\text{Res}} \downarrow$		
Trials	Rand.	TPE	GP_{UCB}	Rand.	TPE	GP_{UCB}	Rand.	TPE	GP_{UCB}
< 16	2.08	2.00	4.49	3.64	4.10	5.78	-31.54	-1.19	-25.70
< 64	1.33	2.82	20.91	1.21	1.15	3.88	-2.45	46.71	-31.36
128	0.56	3.58	26.61	0.78	0.76	4.29	0.68	263.01	-26.82
Avg.	1.32	2.81	17.34	1.88	1.99	4.65	-11.10	102.84	-27.96

3.2 Selection Bias

We perform two ablation experiments and further compare Random and TPE strategies. We evaluate the sensitivity of the learning rate $[1.0e^{-4}, 1.0e^{-1}]$ and the batch-size $[32, 1024]$ to the performance (batch-size generalization gap (Hoffer et al., 2018)). A total of 60 trials for each strategy are run using a ResNet50 (He et al., 2015) network to classify images in a modified version of the CIFAR-10 (Krizhevsky, 2009) dataset with only images of cats and dogs as a binary classification problem.

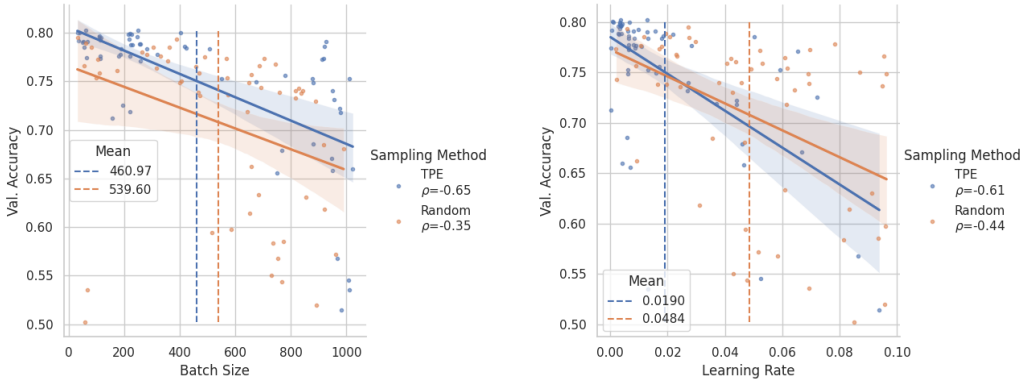

Figure 3: Comparison of the Random and TPE hyperparameter selection strategies. Random evenly samples from all learning rates, including the ones that result in poor performance. TPE is biased towards a small fraction of the search space and skews sensitivity analysis towards sub-regions of the hyperparameter space.

Figure 3 shows TPE sampling from a small region of hyperparameters, where on average 68% of sampled hyperparameters are below the mean configuration value. Additionally, when computing Pearson's correlation ρ between the two strategies, TPE suggests a large negative correlation between the hyperparameters and the accuracy that is about x1.5 higher than the empirical estimate. For an equivalent budget, both methods reach similar best-trial performances, where the best-trial performance was 0.80 for TPE and 0.79 for Random. The relative error of the best trial compared to the empirical value ($\Delta\mathbf{A}$) was -0.00225 for TPE and -0.00710 for Random with a budget of 16 trials. Our results suggest that the benefits of TPE for evaluations that use the best performance might be small, and results can be unreliable when evaluating and exploring a model's performance, as TPE is biased towards sub-regions of the hyperparameter space.

4 Discussion

A reliable estimate of the performance of a method requires multiple experimental trials. In this work, we find that even under a limited budget, a Random sampling strategy used to sample hyperparameters may be the most reliable. Additionally, we observe that using an HPO strategy to select hyperparameters may lead to errors in comparisons that use the mean performance with statistical tests like ANOVA.

Our experiments highlight the need for new community standards specific to model evaluation. HPO exploits a limited range of hyperparameters to find the best performance and can be an unreliable estimate of the model's performance under general conditions. Poor experimental practices can lead to mistrust towards scientific achievements as the results are often contradictory. For ML research, evaluating the generalization ability of a model remains an open problem.

5 Conclusion

Through a meta-analysis, we identify incorrect experimental practices for evaluating the performance of a method. We simulate ablation experiments on a NAS benchmark dataset and use the empirical performance of the dataset as a ground-truth to quantify the reliability of the sampling strategy. Additionally, we perform an ablation study on the sensitivity between learning rate and batch size. Our results demonstrate that the reliability of the study declines sharply for different hyperparameter selection strategies other than Random. Finally, we find that even when used as intended, HPO provides little benefit when finding the best configuration compared to random sampling.

ACKNOWLEDGEMENT

This work was supported by C-BRIC (one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA), DARPA (HR00112190134) and the Army Research Office (W911NF2020053). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

REFERENCES

Hana Ahmed and Jay Lofstead. Managing randomness to enable reproducible machine learning. In *Proceedings of the 5th International Workshop on Practical Reproducible Evaluation of Computer Systems*, P-RECS '22, pp. 15–20, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393133. doi: 10.1145/3526062.3536353. URL https://doi-org.libproxy2.usc.edu/10.1145/3526062.3536353.

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/647bba344396e7c8170902bcf2e15551-Paper.pdf.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Xavier Bouthillier and Gaël Varoquaux. Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020. Research report, Inria Saclay Ile de France, January 2020. URL https://hal.science/hal-02447823.

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. In *International Conference on Machine Learning*, pp. 725–734. PMLR, 2019.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In A. Smola, A. Dimakis, and I. Stoica (eds.), *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021. URL https://proceedings.mlsys.org/paper/2021/file/cfecdb276f634854f3ef915e2e980c31-Paper.pdf.

Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395, 2020. URL https://arxiv.org/abs/2011.03395.

Xuanyi Dong, Lu Liu, Katarzyna Musial, and Bogdan Gabrys. NATS-Bench: Benchmarking nas algorithms for architecture topology and size. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. doi: 10.1109/TPAMI.2021.3054824. doi:10.1109/TPAMI.2021.3054824.

Katharina Eggensperger, Philipp Müller, Neeratyoy Mallik, Matthias Feurer, René Sass, Aaron Klein, Noor H. Awad, Marius Lindauer, and Frank Hutter. Hpobench: A collection of reproducible multi-fidelity benchmark problems for HPO. *CoRR*, abs/2109.06716, 2021. URL https://arxiv.org/abs/2109.06716.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Karol Kurach, Mario Lučić, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A large-scale study on regularization and normalization in GANs. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of

Proceedings of Machine Learning Research, pp. 3581–3590. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kurach19a.html.

Thomas Liao, Rohan Taori, Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship, 2018.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study, 2017. URL https://arxiv.org/abs/1711.10337.

Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks, 2019.

Julia Moosbauer, Julia Herbinger, Giuseppe Casalicchio, Marius Lindauer, and Bernd Bischl. Explaining hyperparameter optimization via partial dependence plots. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2280–2291. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/12ced2db6f0193dda91ba86224ea1cd8-Paper.pdf.

Prabhat Nagarajan, Garrett Warnell, and Peter Stone. The impact of nondeterminism on reproducibility in deep reinforcement learning. 2018.

Andrei Paleyes, Mark Pullin, Maren Mahsereci, Neil Lawrence, and Javier González. Emulation of physical processes with emukit. In *Second Workshop on Machine Learning and the Physical Sciences, NeurIPS*, 2019.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision, 2021. URL https://arxiv.org/abs/2109.08203.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *CoRR*, abs/2106.03253, 2021. URL https://arxiv.org/abs/2106.03253.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. URL http://arxiv.org/abs/1409.4842.

Ryan Turner, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In Hugo Jair Escalante and Katja Hofmann (eds.), *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pp. 3–26. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/turner21a.html.