Coffee Talk #9

March 15, 2023

Ozgur Taylan Turan

Modeling Machine Learning Multiverse[1]

[1] S. J. Bell, O. P. Kampman, J. Dodge, and N. D. Lawrence (Oct. 12, 2022). *Modeling the Machine Learning Multiverse*. arXiv: 2206.05985 [cs, stat]

Why this paper?

- Interesting effort...

Aim

- Present a principled framework for backed up claims...
- A step closer to reproducibility...

Introduction

Multiverse Analysis [1]

- Psychology background...
- Make all the possible choices at the same time!
- Mostly related to dataset construction.
- Different choices affect the outcome/conclusion!

[1] S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel (Sept. 1, 2016). "Increasing Transparency Through a Multiverse Analysis". In: *Perspectives on Psychological Science* 11.5, pp. 702–712. ISSN: 1745-6916. DOI: 10.1177/1745691616658637

Introduction

Machine Learning

- Again possible choices (batch size, learning rate architecture etc.)
- CLAIM: With this method we can investigate the effect of each choice...

Multiverse Exploration

search space $\mathcal{X}$ (claim: often continuous) and **evaluation function** $l \rightarrow$ multiverse
Due to search space being too large $\rightarrow$ GP surrogate and explore the space using Bayesian experimental design!

Multiverse Exploration

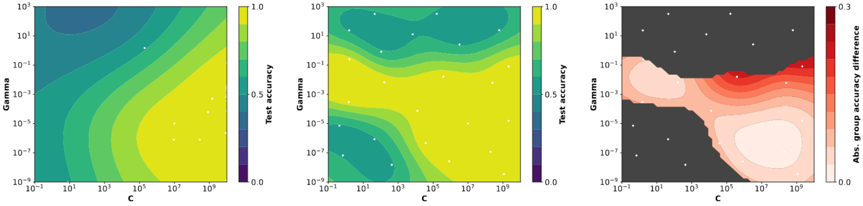
- Sample initial design $X_0 \sim \mathcal{X}$ and evaluate $Y_0 = l(X0)$ [They select Sobol sequence as initial design]
- Fit a GP model to X_0 and Y_0
- Use acquisition function a on f to sample and evaluate a new batch (X_i, Y_i)
- Repeat until a stopping criterion... [Bayes factor:=$\frac{P(X,Y|K_shared)}{P(X,Y|K_additive)}$]
- Sensitivity analysis

Multiverse Exploration

Caveat:

- a Integrated Variance Reduction $\rightarrow$ nasty integral $\rightarrow$ Monte-Carlo approx.
- Difference with standard optimization!

Optimization vs Exploration

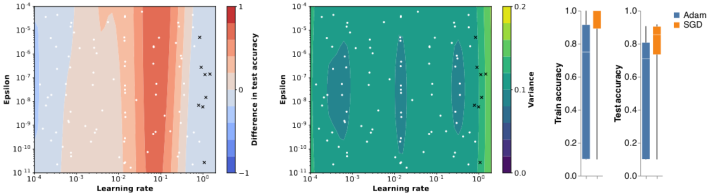


(a) UCB (b) IVR (c) IVR

Figure 1: Contour plot of GP-predicted mean test accuracy over search space of C and γ (Gamma) as explored by **(a)** UCB and **(b)** IVR acquisition functions. **(c)** Secondary objectives, e.g. minimizing group-level outcome differences, may vary along the IVR-revealed plateau.

- Premature optimization hinders our understanding...
- Not to throw shade at optimization!

When is Adaptive optimization helpful?



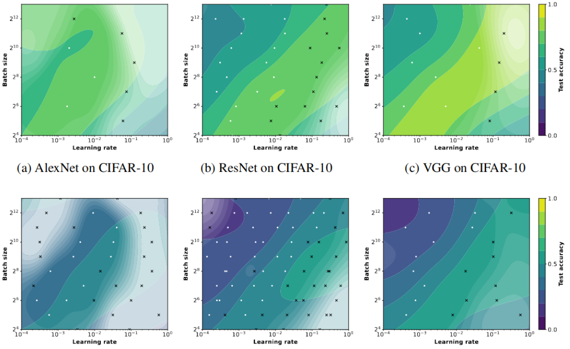
(a) Mean surface (b) Variance surface (c) Final accuracies

Figure 2: Contour plot of GP-predicted **(a)** mean difference in test accuracy (SGD - Adam) and **(b)** variance over the search space of learning rate and ϵ. Red regions indicate SGD with momentum outperforms Adam. White points are successful trials; black crosses failed. **(c)** Final train and test accuracies. Whiskers extend to min and max. Note SGD train accuracy has median, UQ and max 1.0.

- SGD with momentum vs Adam
- Opposing the claims of [1] SGD>Adam

[1] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht (May 21, 2018). *The Marginal Value of Adaptive Gradient Methods in Machine Learning*. arXiv: 1705.08292 [cs, stat]

Is there a large-batch generalization gap?



(a) AlexNet on CIFAR-10 (b) ResNet on CIFAR-10 (c) VGG on CIFAR-10

(d) AlexNet on CIFAR-100 (e) ResNet on CIFAR-100 (f) VGG on CIFAR-100

Figure 4: Contour plot of GP-predicted mean test accuracy over the search space of learning rate, batch size, dataset and model. White points are trials with training accuracy ≥ 0.99; black crosses were excluded. Overlayed translucent regions indicate high training error. For Tiny ImageNet see fig. S4; for variance see fig. S5. The discrepancy between contours and data points in **(a)** is due to the coregionalized model sharing information across functions.

- Many Researchers: batch size $\uparrow$ $\rightarrow$ generalization $\downarrow$
- Batch-size by itself does not explain the generalization performance!

Conclusions

- By using a multiverse analysis, researchers and practitioners gain more robust claims and better understanding of the consequences of their decisions.

THANKS!