Future-Proof Your Research: Designing for Replicability and Reproducibility

Isabel Brunkan
ibrunkan@gmail.com
Minerva Schools at Keck Graduate Institute

Zhuo Zhen
zhenz@uchicago.edu
University of Chicago

Kate Keahey
keahey@mcs.anl.gov
Argonne National Laboratory

Levent Toksoz
letoksoz@uchicago.edu
University of Chicago

ABSTRACT

Computer Science is Agile. Iteration after iteration, moving quickly to solve new problems, uncover new questions, find the next big thing. Hardware, software, libraries, datasets, experiments - technology becomes outdated almost as soon as it's released. So why save code? Why share code? For replication, to verify results. For education, to train the next generation. For variation, to discover new insights. We seek to understand how to package experiments to encourage experiment exploration and longevity by replicating an AlexNet reproduction.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; • **Information systems** → **Collaborative and social computing systems and tools**.

KEYWORDS

replicability, reproducibility, datasets, cloud computing, machine learning, neural networks

ACM Reference Format:
Isabel Brunkan, Zhuo Zhen, Kate Keahey, and Levent Toksoz. 2020. Future-Proof Your Research: Designing for Replicability and Reproducibility. In *Proceedings of SuperComputing (SC'20)*. 3 pages.

1 INTRODUCTION

Replicable research is a crucial step in the scientific process. Computer science research faces a unique challenge due to resource variety, availability, and continual upgrading, as well as original data and code accessibility. However, replicability, the "recreation of the same experimental apparatus and performing the same experiment", and reproducibility, defined as "implementing the same general idea" of an experiment, but with "newly created experimental apparatus", have practical and empirical value for confirming findings[2]. Additionally, experiments packaged for replicability provide an opportunity for education and experimentation. Clark et al. go as far as to label repeated research as "an important step in transferring the results of computer science research into production environments"[1]. They also note the difficulty of repeating research and propose an intermediary group repeat and polish the process before releasing experiments to readers. We propose a technology pairing that eliminates this intermediary, allowing researchers to simultaneously run the experiment and package it for future use.

This paper implements these ideas, attempting to replicate the original AlexNet experiment, credited with motivating research into applications of GPUs and CPUs for deep learning[5]. The AlexNet model trained on 1.2 million images with 2 GPUs over 5-6 days, obtaining top-1 (correct class) and top-5 (correct class in top 5 estimates) error rates of 37.5% and 17.0%[5]. Repeatability, "rerunning exactly what someone else has done using their original artifacts", is not possible due to resource availability and data accessibility[2]. However, a variation was found on Kaggle, a popular data science and machine learning community, applying the AlexNet model to a simplified dataset, the Stanford Dogs dataset [4]. In this work, we replicate this experiment and compare accuracy and performance results.

2 TECHNOLOGY

This experiment was conducted on Chameleon, a NSF-funded large-scale, reconfigurable testbed[3]. Its integrations make it an ideal reproducibility platform: Chameleon provides hardware resources, integrating with Jupyter Notebook to enable convenient experiment packaging, and Zenodo, to publish the final product with a DOI for citation.

3 REPETITION, REPLICATION, AND VARIATION

3.1 Repetition

As mentioned before, repetition is not possible due to the 2010 ImageNet subset data accessibility and hardware availability. The original AlexNet experiment uses 2 GTX 580 GPUs while Chameleon offers P100, K80, M40, RTX-6000 and V100 GPUs[5]. However, the Kaggle-hosted reproduction is repeatable as Kaggle provides 30 hours of P100 GPU usage per week for free. The experiment was repeated 10 times for a test set accuracy average, as it ranged from 23.4% to 38.5%. The average test accuracy, 34.1%, was similar to the original experiment's 31.6%[6].

SC'20, November 2020, Virtual

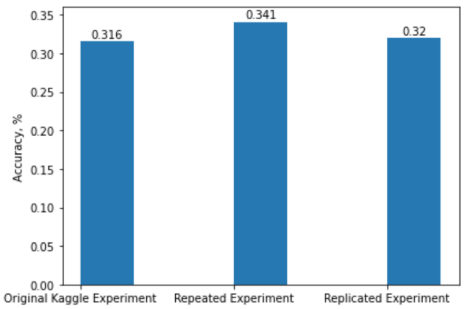

Figure 1: Test Accuracy for original, repeated and replicated experiment on Chameleon.

3.2 Replication

The Kaggle-hosted experiment was replicated in Chameleon, recreating the experimental setup with a P100 GPU. The data was uploaded to Chameleon using the Kaggle API and filtered to match the original experiment's selected 20 dog breeds. The replication achieved results within 0.4% accuracy of the original experiment.

3.3 Variation

Variation was introduced by running the experiment on different GPUs, with different dataset lengths, and additional pictures.

Table 1: Alternate methods to load the Stanford Dogs dataset

	Reproduction on Kaggle	For Variation on Chameleon
`Data Loading`	Uploaded from Operating System	TensorFlow-Datasets API
`Data Processing`	SciKit-Learn and NumPy	Enabled with TensorFlow-Datasets API
`Image Processing`	Python Imaging Library	TensorFlow's image.resize function

3.4 Container Variation: Repeating Experiments on Different Hardware

Variation is simplified by packaging container setup and the experiment in different scripts, allowing hardware and experiment variation to be introduced independently. The same experimental container can be deployed on different hardware without adjustments making replication faster.

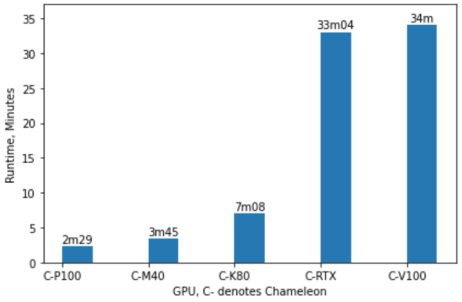

Figure 2: Training runtime per Chameleon GPU for 20 dog classes.

The AlexNet-Stanford Dogs replication was deployed on all Chameleon GPUs. While the test set accuracy variation was within 2%, the runtimes had pronounced variation, ranging from about 2 to 30 minutes.

3.5 Experiment Variation: Repeating Experiments with Manipulation

Similarly, minor adjustments to the experiment script introducing variation can be made without affecting container setup. Expanding from 20 dog classes to all 120 dog classes deploys the same container setup, but tests how different GPUs handle larger datasets.

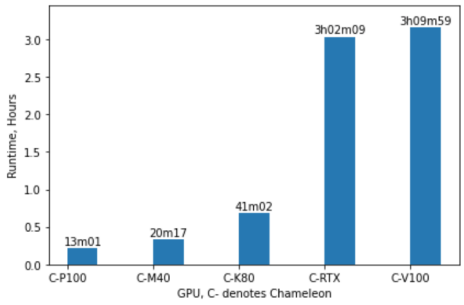

Figure 3: Training runtime per Chameleon GPU for 120 dog classes.

New data can also be easily introduced. Photos of the author's frenchton, a ¾ French bulldog and ¼ Boston terrier mix, were tested, posing a unique challenge as a dog mix and by not belonging to a training set class. The results were inaccurate, though consistent - all images classified as chihuahuas. This shows additional research and training the model on a larger, more diverse dataset is needed.

Figure 4: Sample images of the Frenchton, with his predicted labels.

4 CONCLUSIONS

True replication is often difficult due to hardware and data accessibility. However, packaging experiments with replication and reproducibility in mind can help ensure their longevity and increase experiment accessibility for education, extension, and technology transfer. While the original AlexNet paper was not directly replicable, a simplified reproduction was replicable.

The process identified best practices to package for replication and variation - separating container setup and experiment into different scripts - and indirectly created a template for future experiments. This helps experimental longevity, deploying the same container setup across newer hardware or adjusting the experiment script for a given trial. It also increases readability, minimizing visible code and allowing researchers to direct focus by expanding the container scripts, if focusing on hardware, or the experiment script.

Ultimately, structuring experiment packaging with clear divisions for container setup and experiment allows easier replication, readability and longevity.

REFERENCES

[1] Bryan Clark, Todd Deshane, Eli Dow, Stephen Evanchik, Matthew Finlayson, Jason Herne, and Jeanna Neefe Matthews. 2004. Xen and the Art of Repeated Research. In *Proceedings of the Annual Conference on USENIX Annual Technical Conference* (Boston, MA) *(ATEC '04)*. USENIX Association, USA, 47.

[2] Dror G. Feitelson. 2015. From Repeatability to Reproducibility and Corroboration. *ACM SIGOPS Oper. Syst. Rev.* 49 (2015), 3–11.

[3] Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons Learned from the Chameleon Testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

[4] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. Novel Dataset for Fine-Grained Image Categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (Lake Tahoe, Nevada) *(NIPS'12)*. Curran Associates Inc., Red Hook, NY, USA, 1097–1105.

[6] Sripooja Mallam. 2019. Dog images classification using Keras | AlexNet. Kaggle Notebook. Retrieved August 7, 2020 from https://www.kaggle.com/msripooja/dog-images-classification-using-keras-alexnet