Graduate Research Plan Statement
Time-Series Interpolation of Scientific Data via Machine Learning

Background: Big Data is a major challenge for computational simulations in High Performance Computing (HPC). Traditionally, these simulations save their state at regular intervals to disk, which enables post hoc analysis and visualization after the simulation has run. However, while disk capacity and bandwidth are increasing on each new generation of supercomputer, they are increasing much more slowly than the ability to compute data. For example, over the last decade, the Texas Advanced Computing Center (TACC) supercomputers have increased disk bandwidth by approximately a factor of ten, while compute power (i.e., the ability to compute data) has increased by a factor of one hundred. As a result, the traditional method for post hoc analysis and visualization is rapidly falling out of favor. Instead, simulations are increasingly incorporating in situ processing, i.e., analyzing and performing visualization tasks while the data is being generated. In situ methods allow analysis to be performed on a subset of the simulation data and processed results are then stored in a reduced form. Further analysis can then be explored in a post hoc manner. However, this can have major drawbacks — scientists cannot re-analyze or come back to parts of simulations after they have thought more deeply about problems or gained insights from other work. In essence, scientists have to hope that they got the analysis correct the first time, otherwise have to go back and re-perform part or all of the simulation again. To solve this problem, we need to develop in situ techniques that maximize the information extracted while the simulation is running. This is an open research problem.

Fortunately, the explosion in machine learning research has generated many techniques that can potentially help the computational simulation community. While several relevant works have shown that machine learning can play a role in scientific visualization [1, 2, 3], additional research is needed to explore how machine learning can be used to address the problem of maximizing the in situ extraction of information.

Research Question: Machine learning's role for in situ visualization is still an open question. The focus of my proposed research is asking how machine learning might be used to interpolate time-series scientific data. This question leads to asking many sub questions. One such sub question is, how can simulation data be encoded in a manner useful to machine learning algorithms? This involves determining what features are relevant to simulation data, and how to compress or encode the relevant information. Additionally, can the relevant features for HPC simulation data be stored on modern GPUs? And if so, how? What methods might be used to perform interpolation on highly irregular data? Can machine learning algorithms learn enough physics to make accurate interpolations with meaningful temporal differences? Additionally, which machine learning techniques execute quickly enough to be used in situ? These questions will need to be answered to determine what role machine learning can play for in situ visualization.

Work Plan: My work plan for this fellowship addresses my research questions. To achieve this I will break down the project into five tasks that will be spread out over a three year plan. The tasks are to: T1) find feature reduction techniques, T2) research compression and encoding algorithms to fit data into GPU memory, T3) explore techniques for time-series interpolation on scientific data, T4) study the generalization and limitations of the developed algorithms, T5) utilize online algorithms so the learning can be done in situ. I describe my specific plans for these tasks in the remainder of this section.

T1, the motivation for feature reduction techniques for scientific data is that HPC simulations can

produce petabytes of data a day, which is much larger than data sets conventionally used in machine learning algorithms. Current GPU architectures, which are essential for machine learning, are limited in memory. I will explore established techniques, such as t-SNE and Principle Component Analysis, to reduce the feature space and determine which features are necessary.

The motivation for T2, researching compression and encoding techniques to fit the required feature space and architecture into GPU memory, is to make simulation data as small as possible to reduce training time and computational load. One of the breakthroughs in modern machine learning is using it for compression. It has been shown that machine learning can be used to perform complex encodings of data that can be used to regenerate the original set. This type of encoding will be important for several reasons. A good encoding can reduce memory burden and free up resources for competing computations. This can also enable more checkpointing for the simulations, where a checkpoint can allow for a simulation to be run from that point instead of starting again from the beginning. With the completion of this milestone, scientists will be able to more easily restart simulations and be able to save more data out for post hoc analysis.

The goal for T3, researching techniques for time series interpolation methods, is to develop architectures that can predict the underlying physics and perform interpolations over large time steps. Some of the newest machine learning techniques provide methods that imply this possibility. Modern architectures, such as Recurrent Neural Networks and Transformers, demonstrate the ability for the algorithms to become contextually aware. An important aspect of interpolating scientific data is to understand how parts of the data move within a given system. Conventional methods use momentum of data points to determine how different subsections of data evolve compared to others. Modern machine learning architectures present the possibility for similar contextual knowledge to be learned by the algorithm.

T4, researching the generalizability of these interpolation algorithms, generalization is one of the most difficult tasks in machine learning. With any algorithmic development it is important to understand its applications and limitations. A deep analysis will need to be performed to determine the extent of generalization and how much fine tuning will need to be done.

T5, convert our algorithms to online algorithms so that fine tuning can happen in situ, which adds flexibility for scientists and decrease the computational demands of the algorithms.

Intellectual Merit: This proposal aims to apply machine learning to scientific data sets for computer simulations. It endeavors to do fundamental research on how large data sets can be mapped into machine learning frameworks and how these frameworks can be used to solve a specific problem – time series interpolation.

Broader Impacts: Success of this research would result in a large reduction in memory consumption of HPC applications while also allowing for better understanding of data. HPC is being used to tackle some of the toughest scientific challenges facing humanity, namely climate research, clean energy, and medical research. Enabling this science to be done more efficiently and faster directly results in a higher quality for people living on the planet. Through NSF this research can be public and help scientists everywhere have better tools when facing the most challenging problems.

References

[1] BERGER, M., LI, J., AND LEVINE, J. A. A generative model for volume rendering. *CoRR abs/1710.09545* (2017).

[2] HE, W., WANG, J., GUO, H., WANG, K.-C., SHEN, H.-W., RAJ, M., NASHED, Y. S. G., AND PETERKA, T. Insitunet: Deep image synthesis for parameter space exploration of ensemble simulations. *IEEE Transactions on Visualization and Computer Graphics* (2019), 11.

[3] LEDIG, C., THEIS, L., HUSZAR, F., CABALLERO, J., AITKEN, A. P., TEJANI, A., TOTZ, J., WANG, Z., AND SHI, W. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR abs/1609.04802* (2016).