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 have saved 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 is increasing on each new generation of supercomputer, it is increasing much more slowly than the ability to compute data. For example, the Texas Advanced Computing Center (TACC) supercomputers have increased disk bandwidth by approximately a factor of ten over the last decade, while compute power (i.e., the ability to computer data) has gone up by a factor of one hundred. As a result, the traditional method for post hoc analysis and visualization is rapidly falling out of favor, and is being replaced by in situ processing, i.e., analyzing and performing visualization tasks while the data is still being generated. With this mode, simulation data can be significantly reduced, and then stored in a reduced form. 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 have 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 problem is an open research problem.

Fortunately, the explosion in machine learning research has generated many sophisticated and complex 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 [3, 1, 2], additional research is needed to explore how machine learning can used to address the problem of maximizing the in situ extraction of information.

Research Question:

Machine learning's role in computational simulation is still an open question. One such question is, how can simulation data be stored in a manner useful to machine learning algorithms? This involves determining what features are relevenat to simulation data, and how to compress or encode the relevant information. What methods might be used to perform interpolation on highly irregular data? Can machine learning algorithms learn enough physics to make accurate interpolations on meaningful timesteps? As well as if machine learning can be used in situ.

Work Plan:

With this fellowship I plan on researching machine learning algorithms to perform time-series interpolations on scientific data. To achieve this I will break down the project into five tasks that will be spread out over a three years plan. The tasks are: 1) to find feature reduction techniques, 2) research compression and encoding algorithms to fit data into GPU memory, 3) explore techniques for time-series interpolation on scientific data, 4) study the generalization and limitations of the developed algorithms, 5) convert to online algorithms so the learning can be done in situ.

The first task is to research feature reduction techniques for scientific data. In an HPC setting, simulations can produce petabytes of data a day, which is much larger than data sets conventionally used in machine learning algorithms. Current GPU architectures cannot store this much data but are essential for machine learning. 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 second task is to research compression and encoding techniques to fit the required features space and architecture into GPU memory. Making simulation data as small as possible reduces training time and reduces the 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. Once we have a good encoding we can reduce our 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 third task is to research techniques for time series interpolation methods. The goal 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 possible. Architectures like Recurrent Neural Networks (RNNs), Transformers, and Attention 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. These architectures have provided great leaps forward in Natural Language Processing with the ability to construct large scale text that is nearly indistinguishable from human writing [4]. This same contextual awareness can theoretically be applied to the language of scientific analysis. These architectures provide a necessary component that helps enable algorithms to understand the structure of time within scientific computing.

The fourth task to to research 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.

The fifth task is to convert these algorithms to online algorithms so that fine tuning can happen in situ. This will add flexibility for scientists and decrease the computational demands of the algorithms.

Intellectual Merit

Success of this research would lead to a reduction in the search space for simulation scientists. It would also lead to a reduction of data storage and potentially lead to greater capabilities in both in situ and post hoc analysis. Success would result in less computational resources required as well as decrease the time that scientists need to perform their research.

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).

[4] RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., AND SUTSKEVER, I. Language models are unsupervised multitask learners.