Bardia Mojra – 1000766739

02/28/2022 – Assignment 02 – Application 01

Supercomputer #4 – JUWELS Booster Module Supercomputer

An HPC-Driven Data Science Platform to Speed-up Time Series Data Analysis of Patients

with the Acute Respiratory Distress Syndrome

In recent years, the use of large and readily available datasets and cutting-edge machine learning (ML) methods have become popular among medical researchers as it has shown to result in increased accuracy and overall treatment of the patient. More importantly, recent advents in deep learning (DL), biomedical engineering, and High-Performance Computing (HPC) have allowed researchers to train DL models capable of learning the dynamics of large multivariate time series medical datasets. The deployment of Recurrent Neural Networks has been shown to predict patient conditions with relatively high accuracy. This is particularly helpful in the analysis, diagnosis, and treatment of patients with diseases related to the respiratory system e.g. Acute Respiratory Distress Syndrome (ARDS).

Multiple solutions have been proposed; but, as the authors point out, using such tools and datasets is difficult for medical professionals for multiple reasons. First, although there is an abundance of publicly available medical data, these datasets are scattered across the web and are often stored in different Electronic Health Records (EHR) formats. Second, deploying ML and DL analysis methods and prediction tools on large datasets requires expert knowledge of the field and programming skills that medical professionals are not widely trained on. Third, the growing amount of data available through EHRs requires an HPC-enabled application that is both scalable over time and seamlessly integrates into the ICU environment and provides patient-specific recommendations.

With a 40% mortality rate, ARDS is a life-threatening medical condition that on average involves 12% of mechanically ventilated (MV) ICU patients. Per the Berlin definition, the onset of ARDS is defined as a prolonged ratio of arterial oxygen potential to fraction of inspired oxygen (P/F ratio) of 300 mmHg or less. Multiple papers have pointed out the correlation between early detection of the onset of ARDS and the patient’s rate of survival. Moreover, several papers have demonstrated the effectiveness of MV protocols, e.g. “low tidal volume” and “high Peak End-Expiratory Pressure (PEEP)”, for ICU patients suffering from ARDS. However, practical use of such techniques requires a well-integrated system within the hospital ICU that can analyze patient data and make inferences on the onset of ARDS. To address this, the authors propose an HPC-based application that is tailored to support clinical researchers in understanding ARDS, and in the future, it could potentially be deployed to real ICUs and clinical research i.e. German Smart Medical Information Technology for Healthcare (SMITH) project.

This application is designed to trains its DL model on two HPC systems, the DEEP prototype, or the JUWELS system for increased computational loads. The JUWELS Booster Module (subject of this assignment) is the world’s 8th highest-performing supercomputer, per the TOP500 list. The system is built in Germany by Atos and utilizes an AMD EPYC processor and NVIDIA A100 for accelerated GPU operations. The system is the most powerful in Europe with 44.1 Pflop/s.

It is important to note that the trained model could be downloaded with a much smaller size and stored locally to make inferences on the onset of ARDS in ICU patients. This is an important factor as it keeps the patient training data in a secure location by the host HPC system which avoids unnecessary distribution. Moreover, the model could be used for inference and diagnosis at local hospitals and clinics without reliance on the cloud which could not be reached if communication deteriorates. Thus, they define two main activities or applications *ARDS Time Series Analysis and Model Training* and *ARDS Time Series Analysis and Model Inference*.

This cloud-based ARDS early diagnosis application uses DataLad, a git-based data management system that features transparent and traceable access to patients' datasets on the premise of the hospital. For development purposes, the authors used the Medical Information Mart for Intensive Care - III (MIMIC-III) database for training their prediction model and stored in the Scalable Storage Service Module (SSSM) and could be used by researchers to train new models without the need to download the database. Their platform supports an array of local and cloud-based deep learning training tools, e.g. DeepSpeed and Jupiter Notebook, that could be used by researchers to develop more innovative models or run GUI-based analysis on a local machine.

The authors trained their model on datasets collected from nearly 20,000 ICU patients who received MV during their stay. First, they make predictions for some missing values, and in some cases with many data points missing they completely drop the feature. They discovered the following six medical metrics are well preserved, Respiratory Rate (RR), Heart Rate (HR), Systolic Arterial Pressure (SAP), Diastolic Arterial Pressure (DAP), Mean Arterial Pressure (MAP), and Blood Oxygen Saturation (SpO2). As mentioned earlier, the diagnosis of a patient depends on the ratio of arterial oxygen potential (PaO2) to the Fraction of Inspired Oxygen (FiO2). The FiO2 is known since it is set on the ventilator and it is recorded automatically and the goal is to predict the value of SpO2 which is directly related to PaO2.

|  |
| --- |
| Figure 1 |

The author deployed three learning and prediction approaches to study the performance of the proposed HPC-based ARDS prediction system. First, they used the Gated Recurrent Unit (GRU) model which consists of two GPU layers with 32 units each, followed by a dense layer. The hyperparameters that define the structure of the model were chosen empirically and the authors point out the crucial role HPC resources played. The system is able to train a total of 10,209 parameters with Mean Absolute Error (MAE) converging to 0.7432 after about 15 epochs, all under 405 seconds. The second method deploys a One-Dimensional Convolutional model that consists of three convolution layers with 128 filters each, a stride of 9, and a 1D max-pooling layer after each layer except the last which has a 1D global max-pooling layer. This model has a total of 302,337 parameters and the MAE stabilizes at 0.725 after only 40 seconds of training. The third approach uses a mixed model with two 1-D convolution layers with 64 filters each, followed by a GRU layer with 32 units, followed by a dense output layer. This model has a total of 49,889 trainable parameters with stabilizing around similar MAE losses in 45 seconds. Figure 1 shows the performance for all three models. The authors highlight the benefits of their mixed model which combines quick training time and high accuracy shown by the other two models.

The authors point out the quick training time provided by the DEEP and JUWELS HPC systems and the role they played in finding the optimal model structure in all three cases. Furthermore, they have made their HPC-based platform available to the public and researchers can train new models for ARDS early detection applications in seconds using only a few lines of code. Moreover, they show the promising results 1D convolution models produce and outline how such a system could be one day integrated into the ICU environment seamlessly.

Corresponding Authors:

[c.barakat@fz-juelich.de](mailto:c.barakat@fz-juelich.de) (C. Barakat), [sfritsch@ukaachen.de](mailto:sfritsch@ukaachen.de) (S. Fritsch), [morris@hi.is](mailto:morris@hi.is) (M. Riedel), [sb@hi.is](mailto:sb@hi.is) (S. Brynjólfsson)

References:

1. Baraka, Fritsch, Riedel, and Brynjolfsson, “An HPC-Driven Data Science Platform to Speed-up Time Series Data Analysis of Patients with the Acute Respiratory Distress Syndrome,” 44th International Convention on Information, Communication and Electronic Technology (MIPRO), 2021, PP. 311–316, IEEE.

02/28/2022 – Assignment 02 – Application 02

Supercomputer #4 – JUWELS Booster Module Supercomputer

Super-Resolution of Large Volumes of Sentinel-2 Images

with High-Performance Distributed Deep Learning

In this paper, the authors use the JUWELS booster module high-performance computing (HPC) system to train a deep learning (DL) model for satellite super-resolution (SR) applications. More precisely, they propose a novel distributed DL model for remote sensing (RS) image SR application of data gathered by the Sentinel-2 satellite. They use GPU accelerators and synchronous data parallelism to achieve state-of-the-art performance while achieving substantial speed-up during model training. For their model, they mostly rely on the self-attention mechanism, residual learning while keeping the model size relatively small.

The spatial resolution of the satellite images is measured by Ground Sampling Distance (GSD) which represents the distance between two sequential pixel centers measured on the ground. High-resolution satellite imagery and RS applications have become an integral part of life in the 21st century and the demand for higher resolution steadily increases. The problem arises from hardware limitations in various ways and the inability to make frequent updates. Sensors and communication hardware onboard satellites are examples of real-life bottlenecks that limit high-resolution image capture and transmission. Thus, the need for super-resolution post-process applications is also increasing. Moreover, satellites acquire vast amounts of data that are hardly supported on HPC platforms. For example, the twin Sentinel-2 orbiters capture up to about 23TB/day of multispectral images. It could take weeks or months to train and optimize a DL model with a month’s worth of data if it were to run even on the most powerful desktop computers. Thus, an HPC-based solution becomes even more necessary as the significant training speedup allows researchers to explore more innovative models with different structures in pursuit of an optimal DL model. In this paper, the authors propose an HPC-based DL model for satellite imagery super-resolution applications which is smaller in relative size and state-of-the-art performance.

Sentinal-2 is deployed by the European Space Agency to capture and communicate multi-spectral optical observations over land and sea surfaces. It is considered a high revisit frequency probe which makes it capable of capturing temporal data. It has 13 spectral bands on board that can be divided into three sets based on their GSD, set A 10m bands, set B 20m bands, set C 60m bands. It is important to note set A has the highest resolution and is used as ground truth data. For the purpose of this study, they super-resolve 20m and 60m bands to 10m bands from A, the two models are designated S2X and S6X respectively.

Both models share similar network architecture with a few differences that are mentioned later in this document. The S2X network begins with upsampling the 20m bands by two times using bilinear interpolation. Then, the upsampled signal is fused with the 10m bands through a band fusion module. Then, the network proceeds with a residual self-attention module (RSA), followed by a final convolutional layer that learns the residual difference between the high-resolution 10m bands and the interpolated and upsampled 20m bands. Convolutional Neural networks cannot learn global patterns among input pixels due to the nature of their convolutional process. To mitigate this and learn global patterns, six residual blocks were deployed in the SRA module, with a self-attention module to capture long-range dependency over the input feature maps. The S6X model has similar architecture, except it upsamples by a factor of six and has an additional input channel from the 20m bands, as well as 60m bands.

Per published literature on the topic, there are two general paradigms for parallelization of training for DL models, *model parallelism* and *data parallelism*. In this paper, the authors apply *synchronized data parallelism* to speed up the training process, which specifically takes advantage of large GPU on clusters on an HPC system. They split batches of data and divide the workload among the available GPU workers to be processed in parallel. At each optimization epoch, they synchronize by averaging and aggregating the gradients over multiple workers. They use a library called, Horovod, which deploys a *Ring Reduction Mechanism* to scale up, which has been proven to be *bandwidth optimal.* To improve the model convergence of mini-batches when the training is scaled, they present a *modified linear learning rate scale rule* to avoid suffering from *explosive loss*. The HPC scale rule states the initial learning rate should be increased by the same factor as which the mini-batch size has scaled up and decayed to half every SGD iteration.

To train and evaluate the proposed models, the authors use the Dsen2 dataset and use the Wald’s protocol to generate the ground truth. The mentioned models are implemented with Tensowflow with an ADAM optimizer and a modified linear learning rate scale rule of 64,000. The models were trained on the JURON and JUWELS HPC systems. The JUWELS supercomputer (which is the subject of this assignment) is the world’s 8th best HPC system with each of its compute nodes are equipped with four Tesla V100 GPUs that are interconnected via NVLink with an all-to-all structure.

|  |
| --- |
| Table 1 |

The authors use the following 6 metrics in the evaluation of their proposed models, the root-mean-square error (RMSE), signal-to-reconstruction error (SRE), spectral angle mapper (SAM), Erreur Relative Globale Adimensionnelle (SSIM) and peak signal to noise ratio (PSNR). They compare the results with naive bicubic interpolation and the results provided the model published alongside the Dsen2 dataset. They show that they beat state-of-the-art performance when trained both models in 4 GPU configuration while training four times faster. The performance begins to suffer slightly as the GPU configuration is increased to 8 and 16 while speeding up as expected. Table 1 shows the performance S2X model with different GPU configurations as well naive bicubic interpolation and DSen2 for reference. The drop in performance is attributed to communication overhead among the GPUs. The S6X model also shows similar performance.

Thus, the authors were able to speed up the DL training process for a multi-spectral super-resolution application using the JURONS and JUWELS HPC platforms and reduce actual training time significantly. They were able to train their DL models faster and achieve state-of-the-art performance in a 4 GPU configuration. Moreover, they showed that HPC-based DL models can utilize higher GPU counts to achieve higher speedup rates for training without suffering from a significant decrease in performance.

Corresponding Authors:

[zhang@aices.rwth-aachen.de](mailto:zhang@aices.rwth-aachen.de) (R. Zhang), [g.cavallaro@fz-juelich.de](mailto:g.cavallaro@fz-juelich.de) (G. Cavallaro), and [j.jitsev@fz-juelich.de](mailto:j.jitsev@fz-juelich.de) (J. Jitsev)

References:

1. Zhang, Cavallaro, and Jitsev, “Super-Resolution of Large Volumes of Sentinel-2 Images with High Performance Distributed Deep Learning,” International Geoscience and Remote Sensing Symposium (IGARSS), 2020, PP. 617–620, IEEE.