Benchmarking Deep Learning Models on Time Series/Spatiotemporal Data

Adelle Driker, Bo Yan, Yuan Hu

Advisor: Professor Rose Yu

Challenges

A. Diverse application domains and models

Time series/spatiotemporal data has a broad range of applications with various types of learning tasks and models. We need to consider the trade-offs between different models.

B. Diverse systems and performance requirements

Time series/spatiotemporal applications can involve multiple kinds of hardware systems and software systems. The combination of hardware and software can help us gain insights on performance optimization not only for machine learning and deep learning, but also for places that these models may not reveal.

C. Performance measurement and workload characterization

The performance analysis for a deep learning benchmark can be considered the measurement of end-to-end system performance, which can be measured using satisfying and optimizing metrics. Workload characterization is another aspect which helps understanding performance bottlenecks and drives optimizations. The existing performance analysis approaches can be adopted by new benchmarks for time series/spatiotemporal deep learning workloads. However, other specific characteristics need to be considered in detailed design.

D. Open and standard datasets

The selection of datasets is also an important part in deep learning benchmarks. Scalability and dimensions of well designed real production datasets can strongly influence the progress of a deep learning field.

E. Learning curve for deep learning framework (PyTorch) and algorithms

PyTorch is a new framework for us to learn. Even though it is easy to learn when compared to TensorFlow because the syntax is very much similar to the traditional Python programming language and it is easier to experiment with if you already know Python. With a more object-oriented style and straightforward data handling, the learning curve for PyTorch is easier compared to the one for TensorFlow. Though it might not be feature-rich, the ease of learning makes it one of the most loved deep learning frameworks among beginners, unlike TensorFlow where you need to understand the functions of sessions and placeholders. For deep learning models and algorithms, these are also new to us. We may need more time to get familiar with various models and algorithms.

Opportunities

A. Spatiotemporal/Time series data

Throughout spatiotemporal/time-series data exploration, model building, performance tuning and evaluating, the Capstone will provide a benefit for us to gain a better understanding of the characteristics of spatiotemporal/time-series data and models. This process will help us work as data scientists in companies with various needs because it is being utilized in a variety of industries such as econometrics, finance, marketing, engineering, etc.

B. Scalability of data

Another interesting opportunity for us in this project is the chance that we may experience working with a No-Sql database. Additionally, we hope to interact with high performance computing resources such as AWS and utilize distributed storage systems.

C. Deep learning

This will provide us opportunities to acquire deep learning skills, develop robust technical skills, and problem solving skills. Our previous courses in the DSE Program did not focus on this aspect of Machine Learning and we are looking forward to learning more about it.

D. Open-source contributions

This capstone potentially provides us with collaboration with the open-source community from two aspects. First, we are heavily relying on open source frameworks such as pytorch and torchTS, we may contribute to its ecosystem as our skills grow. Secondly, one of our objectives eventually is to publish a spatiotemporal/time series benchmarks dataset to the open-source community to benefit others.

E. Leadership and collaboration skills

As a team, this Capstone project also helps us build our team's values, core, and spirit. It's all about people. We work together. We are one team, we respect and trust each other. This project helps us to learn the true reality of an end-to-end project. We will need to collaborate with team members who have various levels of experience on different parts of the project.

Data Sources

For the purposes of this project, we will be utilizing several data sources to cover a variety of domains and industries where both spatiotemporal and time series data may be used.Once our product is available to the community, users will be able to select the functionality they need based on the dataset they are working with. Our preliminary list of data sources consists of the following:

A. Air Quality PM2.5, which tracks air quality and weather attributes including humidity, pressure, and temperature in the Beijing area. [https://drive.google.com/drive/folders/1cA0fwMbpX84j3WgQqzazpPWSus96ZWKO]

B. PEMS Traffic Data, which is obtained from CalTrans in real-time spanning freeways around metropolitan areas in California.Historical data from the past ten years is available as well.

[https://dot.ca.gov/programs/traffic-operations/mpr/pems-source]

C. US Weather Events (2016-2020), which covers 6.3 million weather events such as rain, snow, and storms across 49 states.

[https://www.kaggle.com/sobhanmoosavi/us-weather-events]

While the list is short at the moment, we plan to incorporate additional time series and spatiotemporal datasets so that our product is able to cater to a wider variety of uses.

Approach

A. We will refer to Open Graph Benchmark as a guide. For each dataset, we provide a unified evaluation protocol using meaningful application-specific data pre-processing, data splits, and evaluation metrics. In addition to building the datasets, we also perform extensive benchmark experiments for each dataset.
B. We will train each dataset in different deep learning models and compare performance.
C. We will create an evaluation tool to evaluate the performance of each datasets and each model.
D. We will also create a leaderboard to keep track of state-of-the-art methods and encourage reproducible research. Once users have developed their own model and got results, users can submit their test results to our leaderboards.
E. The pipeline will include datasets selection, data loader, deep learning model, performance evaluator, and leaderboard.
F. This is a general approach to solve this problem, it will probably change in future.

Team Roles and Responsibilities

Based on each team member’s strengths and experience, we have assigned our roles in the following way: Bo, as the Record Keeper as well as the Software/ML/DL Engineer, will manage the project GitHub repo and help provide additional understanding in the Deep Learning domain with her previous experience in the field. Yuan will be the Budget Manager and Data Engineer, and will be in charge of tracking resources used by the team, mapping out the ETL and code automation process, and providing insight for the visualization portion of the project. Finally, as the Project Coordinator/Manager and Data/Business Analyst, Adelle will be responsible for maintaining contact between the Group and the Advisor, tracking project progress, and aiding with coding/analytical tasks.

Please note that throughout the duration of the Capstone, our roles and responsibilities may slightly evolve based on the state of the project. For instance, if additional assistance is needed on a certain section, team members will be available to assist.

Project Coordination and Communication Plan

One of our main channels for communication is Discord, where we plan to keep in touch via instant message to share quick updates and plan future meetings. Professor Yu also invited us to her Slack channel, where we may send an IM whenever we may have a brief question for her. Initially, we intend to begin with weekly recurring meetings (about half an hour long) via Zoom with our advisor, Professor Yu, and will schedule additional meetings if/when necessary. As the Project Coordinator, Adelle will be the main point of contact between the Group and the Advisor. Professor Yu has additionally advised our team to create a comprehensive timeline of additional tasks/deliverables that will be critical to our project.

Initially, we had hoped to use project management software such as Asana or Jira, but the free versions are quite limited in functionality and amount of time allowed to use the application. Instead, we aim to create a simple version of a project tracking system using a Google doc to make sure our team is progressing as planned and minimize any blockers.

Team Member Contributions

Bo:

- Created Github and google drive for our team to share any resources and information.
- Summarized the minutes of our meetings (group meeting and meeting with prof Yu).
- Worked on Challenges and Approach sections.
- Formatted report document.

Yuan:

- Proposed and tested project management tools such as Asana.
- Initiated a plan on isolating the working environment for team members to avoid a version or dependency compatibility issues.
- Worked on the Opportunities section.
- Formatted report document.

Adelle:

- Reached out to the Advisor on behalf of the team.
    - Coordinated and scheduled team meetings and kick-off meeting with Advisor.

- Set up the skeleton template for the first report.
- Worked on Data Sources, Team Roles and Responsibilities, Project Coordination and Communication Plan sections.
- Formatted report document.

References

1. Hu, Weihua, et al. "Open graph benchmark: Datasets for machine learning on graphs." arXiv preprint arXiv:2005.00687 (2020).

2. X. Huang, G. C. Fox, S. Serebryakov, A. Mohan, P. Morkisz and D. Dutta, "Benchmarking Deep Learning for Time Series: Challenges and Directions," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 5679-5682, doi: 10.1109/BigData47090.2019.9005496.

3. Zhang, Aston, et al. “Dive into Deep Learning.” Dive into Deep Learning - Dive into Deep Learning 0.17.2 Documentation, https://d2l.ai/index.html.