CSN-341/361
Computer Networks
Critical Review

Group-6

Name	Enrollment No	Contribution
Meet Sindhav	22114053	Key findings
Mohammed Haaziq	22114055	Final documentation
Aditya Mundada	22114058	Summary, author details
Nayan Kakade	22114060	Positive and negative aspects
Sarvesh Baheti	22114087	Proposed solutions to various gaps
Roopam Taneja	22125030	Objective of paper

OpenData: a framework to train and deploy ML solutions in WANs

Authors:

1. **Sina Keshvadi**: Assistant Professor in Department of Software Engineering at Thompson Rivers University, Canada. Research focuses on Computer Networks and Internet Measurement.
2. **Shuihai Hu**: principal Research Engineer at Huawei. Research focuses on Networks and Machine Learning Systems.
3. **Geng Li**:Researcher in Huawei,Canada
4. **Yi Lian**:Interested in Machine Learning Systems.

Objective of the paper

Machine Learning solutions are being explored to solve various problems related to networking. For any data-driven application, quality of data is instrumental in determining the achieved results.

The paper describes how offline or synthetic data from simulators when used as training data does not produce highly accurate models which can be achieved through online data collected at real-time. It tries to understand the impact of data drift and data scarcity on a model's performance by a case study on a QoS forecasting model.

To address the issue of on-the-fly network data collection, the paper presents a generic framework named OpenData. OpenData was envisioned as a generic solution which would fulfill all data demands by its layered architecture divided into Application, Data and Infrastructure layers. The expected benefits from such a framework are explored.

The paper then looks at challenges that would occur to achieve such a solution which have not allowed such a solution to become a reality.

Through a detailed literature review, the paper explores different problem domains and current data-driven solutions along with their data requirements. This leads to a realization that achieving a one-fits-all solution may not be possible.

The paper concludes by providing possible directions of a domain-specific, model-specific or data-assisted approach to address the designing challenges while still providing the key benefits of data-driven applications.

Summary

Data driven solutions are known to improve the efficiency of various networking services and protocols. The biggest challenge in developing such solutions is to be able to collect good quality data as the accuracy and efficacy of ML models is the output of the quality and relevance of data that is used to train the models.

Current practices of acquiring data can be classified into two parts: offline and online data acquisition. Offline data acquisition refers to usage of accumulated historic data and online data means using synthetic data from simulators and network traffic generators.However, the heterogeneous nature of different networking services creates a need to be able to provide network specific data as synthetic data cannot model the environment that the ML model would operate in.

An alternative includes collecting real time data from the network itself and using this data to train and update the model. However, current methods to do this are expensive, create excess overheads and often hamper the privacy of the network traffic as each application must be able to monitor its own data.

The paper proposes OpenData: a generic framework for collecting real time data on the fly. OpenData framework comprises three layers: Infrastructure layer where the actual data is moving, the Data layer in which the most important part OpenData Controller resides which collects all the data from the infrastructure layer and processes as well as refines it. The apps and models running in the application layer then use API calls to use the collected data from the OpenData Controller. Using this framework addresses various challenges:

1. **Less Computation overhead**: Since we can now collect Data centrally and use it to immediately update the models we don't need excess storage and processing overhead for each individual application.
2. **Data Scarcity and Data Drift**: As we are using real time data so we are simply ensuring that the used data is not becoming redundant and irrelevant with time which addresses data drift. Also we now have enough data to address the scarcity issue.
3. **Scalability**: Using the 3 layer infrastructure abstracts apps from physical details of data collection. So many more applications can be deployed and the solution becomes scalable.

However, monitoring data on the fly makes a few difficult tradeoffs: It becomes difficult to label data in real time and hence we are mostly forced to shift to a Reinforcement learning approach rather than a Supervised Learning approach which generally creates better quality results.

We further observe that a generic framework like this cannot fit all kinds of networking domains where required parameters for model training are different. For e.g congestion control and video streaming require monitoring completely different network aspects. Hence, we further improve our solution by using the following approaches:

1. **Domain specific approach:** We try to create instances of this framework for each problem domain, only monitoring the aspects relevant to that domain. This greatly increases the efficiency of the deployed solution. Also as our problem domain is narrowed down hence labeling data becomes easier.
2. **Model specific approach**: It tells us to collect and preprocess data according to the approach that the application model would be interested in Supervised, Unsupervised or Reinforcement Learning.
3. **Data Assist solution**: In this solution we can use locally collected app data along with our real time data to analyze overall data trends and create effective solutions.

The paper also performs a Case Study on the Quality of Service (QoS) Forecasting model where it is concluded that a model trained on newer and smaller dataset performs better than on trained on older but larger dataset further strengthening the importance of collecting real time data using the proposed framework.

Critical Analysis

Positive Aspects of the Research Paper

- The framework OpenData is a quite innovative framework aiming to generate data-driven solutions for networking challenges.
- This model is three layered by which the intricate details of data collection in the infrastructure layer are abstracted from the application layer. Due to this the framework can be scaled in the future.
- The computation overhead of the model decreases due to the fact that the data collection is centrally controlled.
- It also solves the problem of data drift and scarcity by using the real time data.
- The improvised version of the framework proposes three different approaches.

These approaches cater to different networking domains which increases the framework's applicability in different contexts.

Negative Aspects of the Research Paper

- The improvised version which proposes different approaches requires large effort to draft solutions for each use case.
- Collecting real time data can raise security concerns.
- Shifting to reinforcement learning based solutions reduces the accuracy of the model.

Key Findings

Data-Driven Network Solutions: The paper proposes the use of network data itself to enhance the effectiveness of data-driven solutions, particularly in wide-area networks (WANs). The paper propose a framework called "OpenData," which is designed to train and deploy machine learning models directly in networks.

Challenges with Data Collection: The research identifies significant challenges in collecting high-quality data from production networks. Current practices rely heavily on synthetic data, which may not accurately represent the target environment. The paper suggests that models should be trained on data collected directly from the network they will operate in to improve performance. Data-driven systems must consider a large amounts of dynamic information to make decisions. However, testing on individual users requires each user to generate a tremendous number of connections but a user may generate little data.

Benefits of Online Data: The paper highlights the advantages of using online networking data for training ML models, including improved accuracy, better handling of data drift, and a more accurate representation of network conditions. This approach also helps in mitigating issues related to data scarcity and ensures the scalability of data-driven solutions.

OpenData Framework: The proposed OpenData framework addresses these challenges by enabling real-time data collection and processing within the production network. This allows for continuous training and updating of ML models, helping them stay relevant and accurate over time. The framework consists of three layers: Application, Data, and Infrastructure, with a centralized controller to manage data collection and processing.

Data Collection: The article discusses three possible directions for data collection within the OpenData framework, focusing on domain-specific, model-specific, and data-assisted solutions:

- **Domain-Specific Data Collection**: This approach carries forward data collection to the specific requirements of a particular domain or application. By focusing on the unique characteristics of a network, it ensures that the data gathered is relevant and can improve the performance of machine learning models.

- **Model-Specific Data Collection**: In this type, data collection is carried out based on the needs of a particular machine learning model. For instance, a framework can offer graph data structures and libraries specifically designed for deploying graph neural network models.
- **Data-Assisted Solution**: In a nutshell, A data-assist solution represents an approach where a dedicated framework enables a data-driven application to aggregate its local data with data obtained from other network elements.

Finding of Case Study: Model 2, which was trained only on Dataset B (most recent dataset), outperforms the other two models in predicting latency and packet-loss rate in most test cases. Model 1 (trained on the least recent dataset) performed the worst among these models, which indicates the vulnerability of data-driven systems to the data drift problem. This shows that data drift has huge impact on the efficiency of model and also shows the significance of using online data in place of historic data. Model 3 (trained on both dataset) only performed slightly better than Model 1, which indicates that re-training a model with new data does not necessarily adapt the model to the environment changes.

Limitations and Future Directions: Despite the benefits, the paper acknowledges that designing a general framework capable of supporting all data-driven applications is challenging.

Proposed solutions to address the gaps

- **Incorporating enhanced Security Measures**

  Implement robust encryption protocols and secure data transmission methods (e.g., TLS/SSL). Additionally, employ data anonymization techniques to protect user privacy and reduce the risk of data breaches.

- **Incorporate a hybrid learning model**

  Incorporate hybrid learning models that combine reinforcement learning with supervised learning where possible. Also, focus on optimizing the reward function in reinforcement learning to improve accuracy.

- **Domain-Specific and Model-Specific Data Collection**

  Enhance the framework with domain-specific plugins or extensions that can be activated based on the use case. For model-specific data collection, provide customizable pipelines that can be easily adjusted to the needs of different machine learning models.

- **A Two-Model Approach**

  In order to ensure the reliability of input data we can use another ML model that filters the most relevant data according to required context, temporal usage, depending on source and relevance within the OpenData Controller. This helps us reduce Data Drift Problems as our new Data filtering model helps utilize the real time data collected more efficiently.