Automation and Machine Learning (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2023.040107
ISSN 2516-5003 Vol. 4 Num. 1

Research on Big Data Analysis and Prediction System Based on Deep Learning

Ziwei Guo

Dickinson College, Carlisle, Pennsylvania, USA

Keywords: Big data; deep learning; prediction system

Abstract: Along with the rapid development of information and communication technology, the amount of global data has shown explosive growth. How to effectively analyze the huge amount of complex data, dig to realize the potential value in it and use it reasonably is one of the important topics at present. The booming development of deep learning has led to an increasing need for its use in various industries. However, the threshold of using deep learning is relatively high for general industry users, which requires a lot of time cost for learning to use and writing complex underlying models, as well as a lot of hardware cost for building deep learning frameworks such as computing servers. Based on the above-mentioned research background and current situation, this paper designs and implements a big data analysis and prediction system based on deep learning, aiming to support the use of deep learning and reduce the cost and operational complexity of users.

1. Introduction

Along with the rapid development of information and communication technology, the volume of data worldwide has exploded. The world has entered the era of big data as huge amounts of data drive more effective decision-making and become a major driving force for the efficient and sustainable development of countries, enterprises and society as a whole. One of the most important issues that needs to be considered and addressed is how to effectively analyse the huge amount of complex data, uncover its potential value and utilise it wisely. Effective analysis of data cannot be achieved without the support of data processing and computing, machine learning and other systems. Traditional analysis systems are based on structured data for online analysis and processing and online transaction processing systems [1]. These traditional analysis tools adopt the streaming data processing mechanism in the data analysis process is extremely effective, but due to the limitations of the single machine operation mode, for large data level data processing reveals the processing time is too long, performance is insufficient and other defects.

Deep learning, which has become increasingly popular in recent years, is the key to solving this problem [2]. Deep learning has not only changed previous machine learning methods, but also opened up new horizons and has made breakthroughs in applications such as speech recognition, image understanding, natural language processing and video recommendation. Building big data analytics prediction systems is an extremely convenient and popular solution to improve the efficiency of data analysts and reduce the cost of use, and all parties have designed and implemented big data analytics systems with different architectures for different objectives. Based

on the above research background and current situation, the research direction of this paper is a big data analysis and prediction system based on deep learning, aiming to design and implement a system that supports the use of deep learning, reduces the user learning cost and simplifies the complexity of user operation.

2. Design of a big data analysis and prediction system functional requirements

This system uses open source frameworks Spark, Hadoop, Tensorflow, etc [3]. As the underlying foundation framework to develop a big data analysis and prediction system based on deep learning. Among the main functions implemented in this system are as follows.

2.1 Data uptake

Data interaction: Support data interaction between the internal data warehouse of the system and external data sources, including the functions of importing external database into the internal data warehouse of the system, exporting internal data warehouse of the system to external database, and importing internal data warehouse of the system by FTP.

Data pre-processing: It provides various data pre-processing operations for data from external data sources and supports data pre-processing and importing into the system internal data warehouse. Data pre-processing operations support the use of various types of data processing frameworks, including pyton, Spark, MapReduce, etc.

2.2 Big Data Analysis and Forecasting

Visual analysis and prediction task model construction: provide users with visual operation interface and basic data analysis and prediction algorithm model components, support users to use visual drag and drop to combine basic algorithm model components to form a custom analysis task corresponding algorithm model workflow task in the form of DAG diagram, where each node in the DAG diagram is a basic model component, each directed edge is a data dependency between components and passed in the form of Json [4]. Each node in the DAG diagram is a base model component, each directed edge is a data dependency between components, and is passed in the form of Json to the backend for parsing and execution.

Component configuration: Each basic data analysis and prediction algorithm model component has a configuration panel, including input data, output results and component properties.

Workflow task parsing and scheduling: Receive the workflow task information in Json form from the frontend, parse and restore the original DAG graph structure information, and parse the data dependencies between nodes through topological sorting and other methods to obtain the scheduling node sequence, and then submit the algorithm model component corresponding to the next node to the unified resource management framework for execution when the upstream dependency node is completed.

Workflow task monitoring: Monitor the running status of each workflow task node during the workflow task scheduling process, including running, process and failed, and display the running information of each node to users in real time to understand the status of workflow task information in real time.

2.3 Real-time query

SQL query: Provides users with the ability to query and analyse data stored in the system's internal data warehouse through the use and execution of SQL statements. Data visualisation:

provides users with the ability to visualise data stored in the system's internal data warehouse through the use of visualisation components such as tables, histograms, pie charts and line graphs.

2.4 Performance Requirements

The system performance requirements are mainly as follows.

(1) Operational stability: the system can run stably without external force majeure influence.

(2) Big data analysis and prediction efficiency: the system established big data analysis and prediction function to achieve the use of algorithm models with more efficient, excellent performance.

(3) Convenient operation: the system can support new users to learn to operate and get started quickly.

3. Overall Architecture Design

In order to facilitate users to conduct rapid deep learning predictive analysis, improve the granularity of support for deep learning algorithm models, and simplify the complexity of deep learning model construction, this paper deconstructs big data predictive analysis tasks from data visualization, data management, task execution and other aspects, and builds and designs a big data analysis and prediction system based on deep learning, in which the architecture contains big data absorption module, data warehouse module, hybrid computing processing, algorithm model library, visualization interface module and other functional parts [5]. The overall design of the specific architecture is shown in Figure 1, which contains five layers: data absorption, system framework, algorithm model, integration components, and service application.

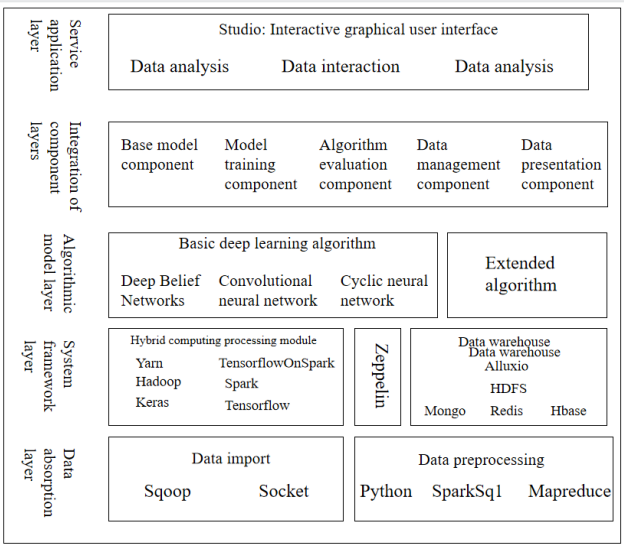

Figure 1: Big data analysis and prediction system architecture diagram based on deep learning

3.1 Data absorption layer

The Data Absorption Layer processes all types of data from different data sources, including numerical data, string data, unstructured text data, image data, etc. As an intermediate interaction between the system and external data sources, this module supports the interaction between the upper-level data warehouse module and external data sources, providing functions such as importing various types of data after data pre-processing and exporting data from the data warehouse to assist users to operate and use the data more easily. In addition, the data absorption

layer provides various types of data pre-processing functions, which are responsible for pre-processing the data information imported into the system from external data sources. The module will also introduce a distributed framework to optimise the data pre-processing and data interaction functions, incorporating distributed features to improve efficiency.

3.2 System framework layer

The system framework layer is the basic part of the deep learning-based big data analysis and prediction system, integrating all kinds of open source frameworks and technologies used in the system, divided into two major parts: the hybrid computing processing module and the data warehouse.

As the amount of data used and processed by deep learning algorithm models is often very large, it must be stored persistently before and after processing, and the image data, video data, audio data and other data involved in deep learning algorithms are often not suitable for storage in common relational databases. Therefore, this paper is designed to build a data warehouse based on a distributed file system, which divides the stored data into two types of data: structured and unstructured, and stores them in the appropriate storage format. At the same time, on top of the underlying distributed file system, an in-memory storage engine is added to speed up the interaction between the upper-level functional modules and the data warehouse module by means of fast in-memory reads, thus increasing the throughput of the data warehouse.

The current processing frameworks used for deep learning algorithm model implementation include Tensorflow, Cafe, Torch, Keras, etc. At the same time, various distributed computing frameworks are added to the deep learning based big data analysis and prediction system to introduce distributed features to model training, data processing and other functions, resulting in multiple different types of computing and processing frameworks in the system operation process The system is running simultaneously. Each framework has a different resource management system and a different resource allocation process, which may lead to resource allocation conflicts between tasks running on different frameworks, resulting in exceptions. To achieve the overall standardised management and scheduling of hybrid computing frameworks, the system will design a unified resource management and allocation module for the computing frameworks involved, to achieve unified resource management for each computing framework, to provide unified resource allocation and scheduling for upper-level applications, to avoid conflicts between resource allocations, and to support multiple computing frameworks to run simultaneously without conflicts.

3.3 Algorithmic model layer

The algorithmic model layer, which contains a wide variety of algorithmic models, is the core part of the whole deep learning-based big data analysis and prediction system, and is the basis for supporting the upper layer of big data analysis tasks. This paper will provide users with a rich set of algorithmic models through pre-integration, eliminating the need for users to learn and write the underlying code to implement the algorithms, making it easy for users to use. The algorithm models will include RNN, CNN, BP and other basic deep learning models to support users to define their own task models, and will also integrate LeNet-5, ResNet-50 and other classical convolutional neural networks to provide users with direct access to them.

3.4 Integration of component layers

The algorithmic models required by users often vary according to their needs, and the development of deep learning often comes with new algorithmic models, making the library of

algorithmic models not static, but progressively more complete. The algorithms used for the predictive analytics tasks performed by users are not entirely different, and parsing each task into the set of algorithmic models used often presents more or less intersection between tasks. That is, by combining different algorithmic models, different analytic tasks can be accomplished, then using a certain amount of basic algorithmic models can basically satisfy most of the tasks implemented, and can reduce the coupling between algorithmic models and improve the reuse of each model. To address these needs, the Big Data Analysis and Prediction System introduces hot-plugging features to implement a collection of base algorithm models based on a hybrid computing processing framework, which can achieve the required algorithms for analysis tasks by combining a limited number of base algorithm models, and each base algorithm model will be built in a component form to achieve its high reusability. In addition, the Big Data Analytics and Prediction System will build a workflow engine to support the functioning of the components and avoid conflicts when reusing components.

3.5 Service application layer

The service application layer is the main way for users to interact directly with the system to achieve data absorption, big data analysis and prediction, real-time queries and other functions to support user use. The user can perform various system operations using command line, scripting, visual graphics and other methods, of which visual operation is simple, fast, intuitive and other characteristics, is the most appropriate way to facilitate the user, providing convenient operation. With this in mind, the Big Data Analysis and Forecasting System will be designed to be operated visually and provided as a web-based service that can be accessed and operated by the user through a browser, with the specific computational processes delivered to the system's computing server. The interface provides user-operable algorithm components and the ability to drag and drop and combine them at will, as well as supporting real-time task status monitoring and graphical display of results to help users understand the status of tasks and the corresponding results.

4. Conclusion

In this paper, we first gave a design requirement analysis of a deep learning based big data analysis and prediction system, followed by the overall architecture design of the whole system according to the two parts of functional requirements and performance requirements in the design requirements, and finally, according to the architecture design of the system, a five-layer architecture system was obtained from the bottom up, which constitutes a set of deep learning based big data analysis and prediction system integrating data absorption, data query, big data analysis and testing, data interaction and other functions.

References

[1] Ning Y. Research on the Application of Big Data Technology in Network Security Analysis[C]. Journal of Physics: Conference Series. IOP Publishing, 2021, 1955(1): 012015.

[2] Zhang Junyang, Wang Huili, Guo Yang, et al. A review of research related to deep learning [J]. Computer Application Research, 2018, 7:1-12.

[3] He Qing, Zhuang Fuzeng, Zeng Li, et al. PDMiner: a parallel distributed data mining tool platform based on cloud computing [J]. Chinese Science: Information Science, 2014, 44:871-885.

[4] Wang Yueqing, Dou Yong, Lu Qi, et al. DLPF: A parallel deep learning programming framework based on heterogeneous architectures [J]. Computer Research and Development, 2016, 53(6):1202-1210.

[5] Zou Y., Jin X., Li Y., et al. Mariana: Tencent deep learning platform and its applications [J]. Proceedings of the VLDB Endowment, 2014, 7(13): 1772-1777.