**SEQUENCE PREDICTION USING DEEP-LEARNING TECHNIQUES**

Project submitted to the

SRM University AP, Andhra Pradesh

for the partial fulfilment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

**SAJITH SALU** **AP21110011498**

**VIJAY KUMAR GANDEM AP21110011507**

**SURYA PRAKASH YALLA AP21110011524**

**BHARGAVA RAYADU RAVVA AP21110011531**

A picture containing text

Description automatically generated

Under the Guidance of

**Ms Anusha Nalajala**

**SRM University-AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh - 522 502**

**Certificate**

Date: 02-DEC-2023

This is to certify that the work present in this Project entitled “**SEQUENCE PREDICTION USING DEEP LEARNING TECHNIQUES**” has been carried out by **G Vijay Kumar**  under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

**Supervisor**



(Signature)

Ms Anusha Nalajala

Assistant Professor,

Department of Computer Science and Engineering

**Acknowledgements**

We would like to express our heartfelt gratitude to the following individuals for their invaluable support and guidance throughout this project:

I. Head of the Organization (Vice Chancellor and Dean): Thank you to the esteemed Head of the Organization, the Vice Chancellor, and the Dean for their visionary leadership and unwavering support. Their commitment to fostering a culture of excellence and innovation within the institution has been instrumental in making this project possible.

II. Faculty Mentor of the University: We are deeply grateful to our Faculty Mentor Ms Anusha Nalajala for giving us this opportunity to do this UROP under his guidance. His continuous encouragement and expert supervision have led us to complete this project and learn a lot of new things. Specially, the valuable feedback in each step of the work has been tremendous in refining the project and enhancing its academic rigor.

III. We also extend our sincere gratitude to our supervisor, classmates, technical staff, and our friends for their invaluable support and contributions to the completion of this project report on Sequence Prediction using Deep Learning Techniques. Their guidance, assistance, feedback, and encouragement have been instrumental in our academic and personal growth throughout this project. We are truly grateful for their unwavering support and the contributions that have greatly enriched our understanding and completion of this project report..

# 

# Table of Contents

1. Abstract

2. Abbreviations

3. Introduction

3.1. Problem Statement

4. Related work

5. Methodology

5.1. Data Collection

5.2. Libraries used

5.3 Preprocessing steps

5.4 Model Training

6. Experimental Analysis and Results

7. Conclusion

8. Future works

9. References

**1. Abstract:**

Sequence prediction, the task of predicting the next element in a sequence based on the preceding elements. It is a very important task in various domain such as Natural Language Processing (NLP), Speech Recognition and Anomaly Detection. Long-Short-Term memory (LSTM) networks, a type of Recurrent Neural Network (RNN) has emerged as a powerful tool for sequence prediction tasks due to their ability to capture long range dependencies and handle complex patterns. LSTM addresses draw-backs of traditional methods by incorporating memory cells that can retain information over long periods. These memory cells allow LSTM networks to effectively model the relationships between distant elements in a sequence. In this project we have a dataset of raw text file containing information about the files and file blocks that were accessed by the client application programs in the distributed file system environment. This report explores the application of LSTM networks to predict the sequence of Block id in the File id which is accessed in near future based on the frequency of each Block id in that File id. It also outlines the research problem, methodology, experimental analysis, results, and future directions for this research.

**2. Abbreviations:**

**NLP-Natural Language Processing**

**LSTM-Long Short-Term memory**

**RNN-Recurrent Neural Network**

**Rid-Rack id**

**Did-Data node id**

**Fid-File id**

**Bids-Block ids**

**3. Introduction:**

**The ability to predict and analyse about the subsequent elements if the data of preceding elements is given is the main goal of intelligent decision-making systems. It is an easy task for a human if the data is not so large, but imagine a data of popular search engines such as Google, Mozilla and other Social media platforms where the data is in GB’s and TB’s. Here comes the use of Deep learning techniques. Deep learning based models fit perfectly into learning and help us by predicting the subsequent outcomes or elements.** Long-Short-Term memory(LSTM) networks, a type of Recurrent Neural Network(RNN) has emerged as a powerful tool for sequence prediction tasks due to their ability to capture long range dependencies and handle complex patterns. In this project we are going to generate sequence from the given sequence.

**In sequence prediction problem we will find the subsequent element with the given past data whereas Sequence to sequence prediction is a challenging extension to the sequence prediction problem. A new sequence is predicted which may or may not have the same length of the previous one. This type of problem has recently seen a lot of study in area of automatic text translation. In our project we are given with sequence of blocks for every file and various files are stored in different data nodes and all the data nodes are stored in different racks. If the input is given as file id then the model would give a sequence of blocks as output. The main parameter for the sequence prediction we used here is frequency. In this paper we will get to know how the LSTM model is used to predict the sequence and train the machine to generate the sequences of the blocks.**

**3.1 Problem Statement:**

**Predicting the Sequence of blocks that will be requested by the client applications in future in the respective caches of data nodes present in each rack. The output sequence should be obtained on the basis of frequency of the Block id of each file and the accuracy of the model.**

**4. Related Work**

**Several studies have explored the application of deep learning techniques for sequence prediction tasks. Recurrent neural networks (RNNs), such as LSTM networks, have demonstrated promising results in capturing long-range dependencies and handling complex temporal patterns within sequential data.**

**Handwriting Recognition: A notable study by Graves et al. (2005) successfully applied LSTM networks to the task of handwriting recognition, achieving significant improvements over traditional methods. LSTM networks are able to capture the complex patterns and temporal dynamics of handwritten characters, making them well-suited for this task.**

**Machine Translation: Similarly, Cho et al. (2014) demonstrated the effectiveness of LSTM networks for machine translation, showing that LSTM-based models could outperform traditional statistical machine translation approaches. LSTM networks are able to capture the long-range dependencies and contextual information in language, making them well-suited for machine translation tasks.**

**Stock Price Prediction: The application of LSTM networks for sequence prediction has extended beyond natural language processing to other domains as well. Zheng et al. (2017) employed LSTM networks to predict stock prices, demonstrating the ability of LSTM models to capture complex temporal patterns in financial data. LSTM networks are able to identify trends and patterns in historical stock prices, making them useful for predicting future price movements.**

**5. Methodology**

**5.1 Data Collection**

**We have generated a synthetic log with 100000 entries using Medisyn technique. We considered 10000 files and 1000 blocks in the distributed file system environment with 10 racks and 10 data nodes per each rack.**

**5.2 Libraries Used**

**The code utilizes several libraries to perform various tasks, including data manipulation, sequence processing, model building, and evaluation. Here's a breakdown of the libraries and their roles**

**5.2.1 Numpy**

**NumPy is a fundamental numerical computation library for Python. It provides efficient data structures, such as arrays, and functions for numerical operations, enabling efficient data handling and manipulation.**

**5.2.2 Pandas**

**Pandas is a versatile data analysis library built on NumPy. It provides powerful data structures, such as Data-Frames, and tools for data cleaning, transformation, and analysis. In this code, Pandas is used to read and manipulate the dataset containing file IDs and block ID sequences.**

**5.2.3 Scikit-Learn**

**scikit-learn is a comprehensive machine learning library that offers various tools for data preprocessing, model fitting, and evaluation. In this code, scikit-learn's Label-Encoder and train\_test\_split functions are used for data preprocessing.**

**5.2.4 Keras**

**Keras is a high-level neural network library that simplifies the process of building and training deep learning models. In this code, Keras is used to construct, train, and evaluate the LSTM model for sequence prediction.**

**5.3 Preprocessing Steps**

**5.3.1 Converting text file to csv file**

**In this project we use read\_csv function present in Pandas library to convert the text file to csv and dropped unwanted columns Aid, Uid and TS. Sorted the values by grouping Rid ,Did and Fid . Selected the Rows of R1 and D1.**

**5.3.2 Dividing the block ids into separate rows**

**The bids attribute is first divided using function str.split( ) and then explode( ) function divides the block ids into rows.**

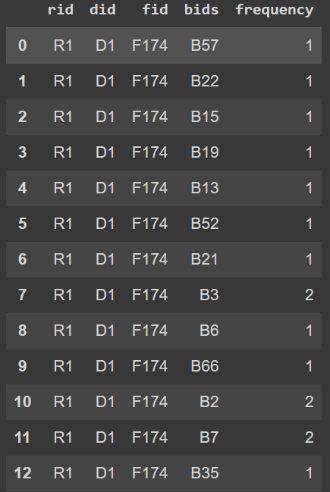
**5.3.3 Adding Frequency Column**

**The frequency column is added by grouping fid and bid. Then used transform( ) function to calculate the frequency.**

**5.3.4 Encoding fid and bids for LSTM**

**Label-Encoder plays a crucial role in preparing the data for the LSTM model. By converting categorical labels into numerical representations, Label-Encoder allows the model to understand and process the data effectively. Without proper encoding, the model would struggle to learn patterns and make meaningful predictions from categorical data.**

**Bids and fids were encoded using a combination of one-hot encoding and embedding layers. One-hot encoding was used to convert each unique bid or fid into a binary vector, preserving the categorical nature of the data. The one-hot encoded vectors were then passed through embedding layers, which transform the high-dimensional binary vectors into lower-dimensional dense representations. This process allows the model to capture the semantic relationships between different bids and fids, enhancing its ability to predict the next block ID in the sequence.**



**Fig 1 Fig 2**



**Fig 3**

**Fig 1 and Fig 2 depicting the data of Rid, Did, Fid, Bids and frequency. Fig 3 is the encoded data.**

**Label encoder uses numeric value between (o and total\_number\_of rows – 1) in the Fig 1 and Fig 2 the total\_number\_of rows are 17 so in Fig 3 the values obtained are o to 16**

**5.4 Model Training**

**5.4.1 Loss function**

**It is averaging the squared difference between the predictions and the target values, the Mean Square Error (MSE) or L2 loss is a loss function that measures the amount of error between a machine learning algorithm's estimated output and an actual output.**

**Loss = (1/T) \* Σ(Xi - X’)²**

**T= Total number of samples**

**Xi= Predicted value of the i-th sample**

**X’= Targeted value of the i-th sample**

**5.4.2 Optimizer**

**Adam, an adaptive learning rate optimizer, was used to update the model's parameters during training. Adam adjusts the learning rate dynamically based on the gradient information, ensuring efficient convergence and improved performance.**

**5.4.3 Training Process**

**The LSTM model was trained on the pre-processed dataset, with the optimizer updating the model's parameters based on the loss function. The model's performance was monitored on a validation set to prevent overfitting and ensure generalization to unseen data.**

**6. Experimental Analysis and Result:**

**In this section we discuss about the Accuracy , Precession , Recall and results obtained from the simulation.**

**6.1 Accuracy**

**It is one of the most often used metrics in activities involving classification. Its definition is the percentage of true positives and true negatives among all instances evaluated that were correctly predicted.**

**In this project we got an accuracy of 86.53%. Fig-7**

**6.2 Precision**

**The percentage of true positive predictions—that is, the positive class correctly anticipated by the model—among all of the positive predictions the model made is known as precision.**

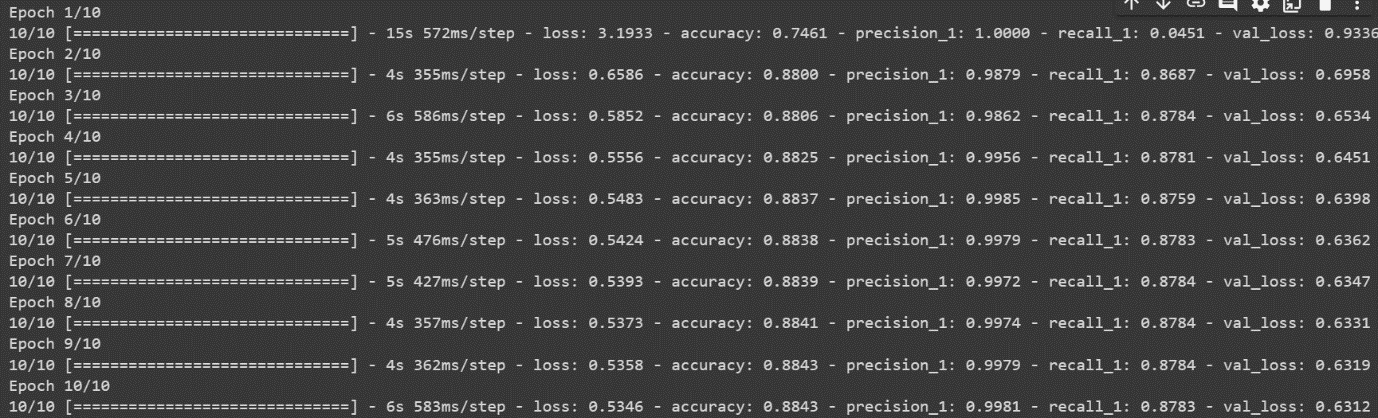
**In this project we got a precision of 0.99. Fig-7**

**6.3 Recall**

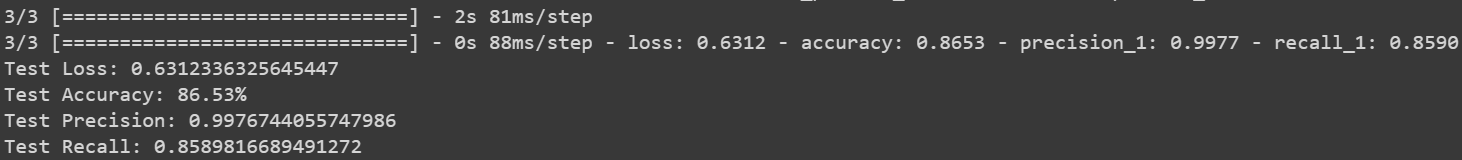
**The percentage of true positive predictions among all actual positive cases is called recall, which is often referred to as sensitivity or true positive rate.**

**In this project we got Recall of 0.85. Fig-7**

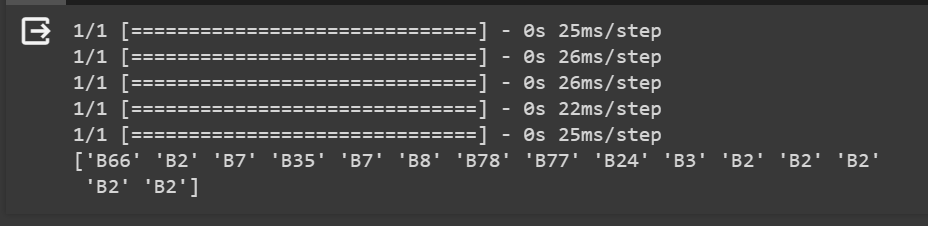
**The Fig-6, Fig-7, Fig-8 depicts result of a file “F174” of Rack-1 and Data-Node-1.**



**Fig-6**



**Fig-7**



**Fig-8**

**7. Conclusion:**

**The conclusions section summarizes the key findings of the research, including:**

**The Effectiveness of LSTM Networks: Summarize the effectiveness of LSTM networks for sequence prediction tasks involving block ID sequences.**

**Optimal LSTM Model Architecture: Identify the optimal LSTM model architecture and hyperparameters for achieving high prediction accuracy.**

**Relative Performance of Deep Learning and Traditional Methods: Discuss the relative performance of deep learning and traditional sequence prediction methods.**

**Applicability to Real-World Problems: Emphasize the applicability of the proposed LSTM-based approach to real-world sequence prediction problems.**

**8. Future Work:**

The future works section outlines potential directions for future research:

**Exploring Different LSTM Variants:** Investigate the performance of different LSTM variants, such as bidirectional LSTMs and attention-based LSTMs, for sequence prediction tasks.

**Addressing Data Imbalance:** Address the issue of data imbalance, which can occur in sequence prediction datasets, to improve the model's performance on minority classes.

**Incorporating Domain Knowledge:** Integrate domain knowledge into the LSTM model to enhance its ability to capture the specific patterns and relationships within the domain-specific data.

**Developing Explainable LSTM Models:** Develop explainable LSTM models that provide insights into the model's decision-making process, enabling better interpretability and trust in the model's predictions.

**9. References:**

# 1. [Chul Ho Kim , Marie Kim , Yu Jin Song], (2021). “Sequence-to-sequence deep learning model for building energy consumption prediction with dynamic simulation modeling.” Journal of Building Engineering. Volume 24. <https://doi.org/10.1016/j.jobe.2021.102577>

# 2. [Gyula Dorgo, Janos Abonyi], (2019).” Learning and predicting operation strategies by sequence mining and deep learning”. Computers and Chemical Engineering.

# Volume 128 and 174-187. <https://doi.org/10.1016/j.compchemeng.2019.06.006>

3. [Jason Brownlee] “LSTM networks with python.” Develop Sequence Prediction model with Deep learning