CHAPTER 1
INTRODUCTION

1.1. Background

The transportationas become a human need to support mobility, there are 3 types of Transportation, namely land transportation, sea transportation, and air transportation. Transportation that is often found and used is land transportation because the price is relatively cheap and easy to reach by the public. The means of transportation are divided into 2 parts, namely public transportation and private transportation. Public transportation until now has become one of the most efficient means of transportation because it is able to carry many passengers at once and is quite practical. Public transportation that is often used by people to go somewhere is a bus, especially if you want to go out of town, the bus is a good choice. The bus is one of the modes of land transportation that can accommodate many people at once. However, bus companies currently have several obstacles in providing their fleets from one city to another because of the highly dynamic demand from passengers, bus companies must be able to analyze which routes will have a lot of demand so that bus companies can provide more fleets on the routes that will have high demand. Unfortunately the bus company is currently still unable to predict which routes will be in high demand, at this time the bus company can only guess. Currently, to overcome this, the bus company has collected data which will later be analyzed.

Time series a collection of data where the data are indexed by date or time sequentially. Multi variate time series is a type of timeseries data but has many data variables. Bus route ticket sales when collected over time and indexed by date and time it becomes time series data. There are many methods and algorithms that can be used to analyze the time series data such as artificial intelligence, machine learning, data mining to deep learning. Machine learning is a study that try to make the computer learns from huge data set to solve specific problems that provided. Until now machine learning model has solved many problems and contribute to the development of the world. Now day's machine learning is used to translate a language to another language, predicting time series data, classifying problems and clustering data from big data. Big data is a stream of data that have extremely high velocity, large variety, and large volume.

Deep learning is a part of machine learning that learns about how the human brain works and implements it into artificial intelligence, in short, deep learning is an algorithm that imitates the human brain. Deep learning is becoming popular because of its excellent performance in studying data and using the learning outcomes to make predictions and analyses based on the data it has learned. Recurrent Neural Network mimics the Frontal cortex of the brain, it is the part of the brain that controls memory. The things that distinguish RNN from DNN (Deep Neural Networks) is that RNN has an input node and an output node, but the RNN output node may return to the previous sequential so that RNN is an algorithm that can solve problems related to sequential data such as speech recognition, text data, time-series and others. However, RNN has drawbacks, especially regarding vanishing gradients, to overcome this problem a LSTM or Long Short Term Memory is formed.

LSTM is an improvisation of RNN which unlike RNN, LSTM can overcome the problem faced by RNN, namely vanishing gradient. The vanishing gradient problem can be overcome by the LSTM because LSTM has an input gate, forget gate and output gate. The forget gate on the LSTM serves to determine whether the data from the previous time step is still relevant or not then the gate decides either the data must be continued or not. With the forget gate, the LSTM can handle dynamic sequential data such as time-series data, time series data is highly dynamics but has its own seasonal and trends. Bi-LSTM is a model architecture that makes the models learn backwards and forwards at the same time. Support Vector Machines, Convolutional Neural Networks, Long Short Term Memory, and even statistical methods such as Auto Regressive Integrated Moving Average can all be used to perform time series forecasting. But in this study only LSTM will be used to forecast bus route demand.

Time-series forecasting is highly depends on historical data. Besides that, time-series data have many variables that can affect the results of the forcasting, therefore a method before forecasting is probably needed to extract features from the time series data. There will be some variable that will impact much bigger than other variable for forecasting bus route demand, and from one variable we can get more information that should impact the forcasting. The hypothesis is the data of event each date will impact the most and from date many new information can be extracted out from it, like semester break (students are more likely to go home to their hometown), or the Christmas and new year, for example. Without a doubt, LSTM does have good performance in handling continuous data, but LSTM itself

probably cannot handle data that have very large dimensions like this multivariate time-series, therefore there is a model architecture that can help LSTM to make predictions called autoencoder. Autoencoder is an architecture that has 2 stages, namely the compression stage and the decompress stage, this is done so that the model tries to make a copy of the original data to extract features on very large dimensions. In this research bi-LSTM-Autoencoder and bi-LSTM models will be comapred to see the effect of autoencoders in bus route demand forcasting problems. Since LSTM can have a memory of past events, LSTM models could use to solve the forecasting problem.

Since the deep learing method is relatively new for bus company to predict the bus route demand, the feasibility study explores new method to make the bus company more profitable by optimizing bus fleet deployment for each route. By having ticket sales history data from a minimum of 2 years and by having calendar events data, it is possible to predict ticket sales on a date in the future to forecast bus route surge. This model can later be implemented in many bus companies.

1.2. Problem Formulation

1. Is the surge in demand for bus routes predictable?
2. What data will help to forecast the bus route demand?
3. How autoencoders architecure efect on time series forcasting problem

1.3. Scope

This research will use python as the programming language, Tensorflow and sickit learn will also be used to help to create algorithms and perform analysis. The algorithm that will be created are LSTM-autoencoders-bi-LSTM hybrid and bi-LSTM. The data used are data from natural resource bus companies taken from the company database, but only 5 route will use to perform forcasting, the models will evaluate with MSE as loss function and RMSE as metrics. This research output will be used to analyze deep learning models on bus route demand problem.

1.4. Objective

The main goal to be achieved from this research is to create a system to analyze and predict bus routes that will have a high demand for which prediction results are obtained from pre-existing data. This research outcome would make a deep learning models to help company make a decision on how to optimize spread of bus fleets.