Google Decimeter Challenge

Report of the final project of the course Machine learning for mechanics

Adrian Klink

Torben Wissel

Table of Contents

[1. Challenge Task 3](#_Toc173684700)

[2. Provided Data 4](#_Toc173684701)

[2.1. Ground Truth 4](#_Toc173684702)

[2.2. GNSS Data 4](#_Toc173684703)

[2.3. IMU data 4](#_Toc173684704)

[3. Data Preparation 5](#_Toc173684705)

[3.1. Data Filtering 5](#_Toc173684706)

[3.2. Data Normalization 5](#_Toc173684707)

[3.3. Data Sorting and Padding 5](#_Toc173684708)

[3.4. Time Series preparation and batching 5](#_Toc173684709)

[4. Models 7](#_Toc173684710)

[5. Results 8](#_Toc173684711)

# Challenge Task

The Google Decimeter Challenge [1] is a competition held out by google on the Kaggle platform. Its goal is to improve the Global Navigation Satellite System (GNSS) positioning accuracy of smartphones using machine learning. The challenge started in September 2023 and closed in May 2024. At the date of this writing, the challenge was already closed. This task is thus a challenge to ourselves with the goal of improving our understanding and skills in machine learning topics. Out goal is to implement a Long Short-Term Memory (LSTM) model to accurately predict the smartphone position.

Over 300 people in 279 Teams participated in the challenge [1] with the goal of achieving sub-meter accuracy by only using the provided GNSS and Inertial Measurement Unit (IMU) data. The expected generated data is the latitude and longitude of the smartphone only.

# Provided Data

In total more than 16Gb of data is provided for the challenge. It is split up in a test set only containing GNSS and IMU data and a train set that also contains a ground truth. The train set itself is made up of three files per recorded trajectory. One containing the ground truth, one the GNSS data and the other containing the IMU data.

## Ground Truth

The ground truth holds the latitudinal and longitudinal position of the smartphone which are the main parameters the model shall be able to calculate. Additionally, information like speed, altitude, and bearing is provided that could be used to train the model additional skills.

The ground truth can be assigned to a timestep of the input data using the timestamp in Unix time milliseconds which is provided in the ground truth data as well as in the GNSS data.

## GNSS Data

The GNSS data files hold 58 different features per satellite datapoint. They include the raw GNSS data received by the smartphone, including the time difference between sent and received messages, the satellite positions and speeds and received correction factors. Also, some calculated data of the smartphone like the resulting pseudorange to the satellite and an estimated position calculated with the weighted least squares method is provided in an Earth Centered Earth Fixed (ECEF) coordinate system is provided. Many of the raw datapoints are not recorded and the columns in the files are actually empty.

The data is provided in a long list with all satellite data points received at the same timestep appearing in groups under each other in the list. If time series of trajectory points are required, the data must first be prepared such that the satellite data received at the same time step is moved to the same line in the list

## IMU data

The provided IMU data consists of the integrated accelerometers and gyroscopes of the smartphones. The file holds the raw measurements of the x,y and z axis as well as the bias of the measurements. The sampling rate is much higher than the sampling rate of the provided GNSS data and the ground truth. Also the accelerometer and gyroscope recordings are stored in alternating order in the same list, as they are also not recorded at the same timestep.

# Data Preparation

In order to use the provided data for a LSTM model, the data must first be cleaned and sorted. Additionally, some pre-calculations are performed, and time series and batches are prepared for training the model.

## Data Filtering

Out of the 58 provided GNSS features only 13 are effectively used, one of them being the timestamp. Many of the features are dumped as they are empty in almost all the provided files, they don’t provide useful information or low variance in the information. The raw pseudorange already gets corrected with the ISRB factor that is also provided in the GNSS data. As the satellite positions and pre-calculated WLS positions reference to a ECEF system, they are transformed into latitude, longitude and height.

From the IMU measurement file the raw IMU measurements are used while the biases are dumped. As the sampling rates are much higher than the GNSS sampling rates the raw measurements are used to calculate the mean and variance of the measurements of each axis of each sensor between the GNSS data points. This results in a total of 12 additional features that are used to train the model.

## Data Normalization

## Data Sorting and Padding

As the provided GNSS and IMU data is not provided in a list where one axis can be interpreted as the time axis the data must be sorted, so that it can be fed to a LSTM model. To achieve this, all of the satellite data points with the same timestep are moved into the same line, while cutting of all timestamps as they don’t provide useful information to the model. Also, the precalculated WSL positions are only added to the end of the line once, as they are the same for all timesteps. The same is done for the calculated IMU data, which is also added once to the end of the line just before the pre-calculated WSL position.

As the number of satellite data points per timestep is not fixed and depends on the recorded trajectory and timestep, the data must be padded and later passed through a masking layer, as the model expects constant dimensions of the input data. This is done by padding the lines of the GNSS data to the maximum number of features in the entire dataset on hand after appending all satellite data of the same timestep to one line. This excludes the features of the IMU measurements and the pre-calculated WSL positions. After padding these features are appended to the end of the line, so that they are in a constant position within the line.

## Time Series preparation and batching

# Models

# Results

References

|  |  |
| --- | --- |
| [1] | Google, "Kaggle," 12 September 2023. [Online]. Available: https://www.kaggle.com/competitions/smartphone-decimeter-2023. |