Ryan Gross
G47667332
CSCI 6917
Guided Research Grad I
Project Report 2

As a refresher, my research topic is the application of Large Language Models (LLM) in prediction of maritime vessel trajectories. This topic is quantitative in nature as prediction accuracy or loss can be measured by the average distance between predicted points from actual points in the test dataset. After measuring average loss, I will then be comparing the loss of the LLM approach to other approaches as measured in the TrAISformer paper [1]. The TrAISformer paper modified an open source Transformer based on the GPT-2 code, and compared the results of this to several other sequence to sequence (seq2seq) methods. This provides the basis for my numerical comparison of loss to what I believe is the state-of-the-art in vessel trajectory prediction.

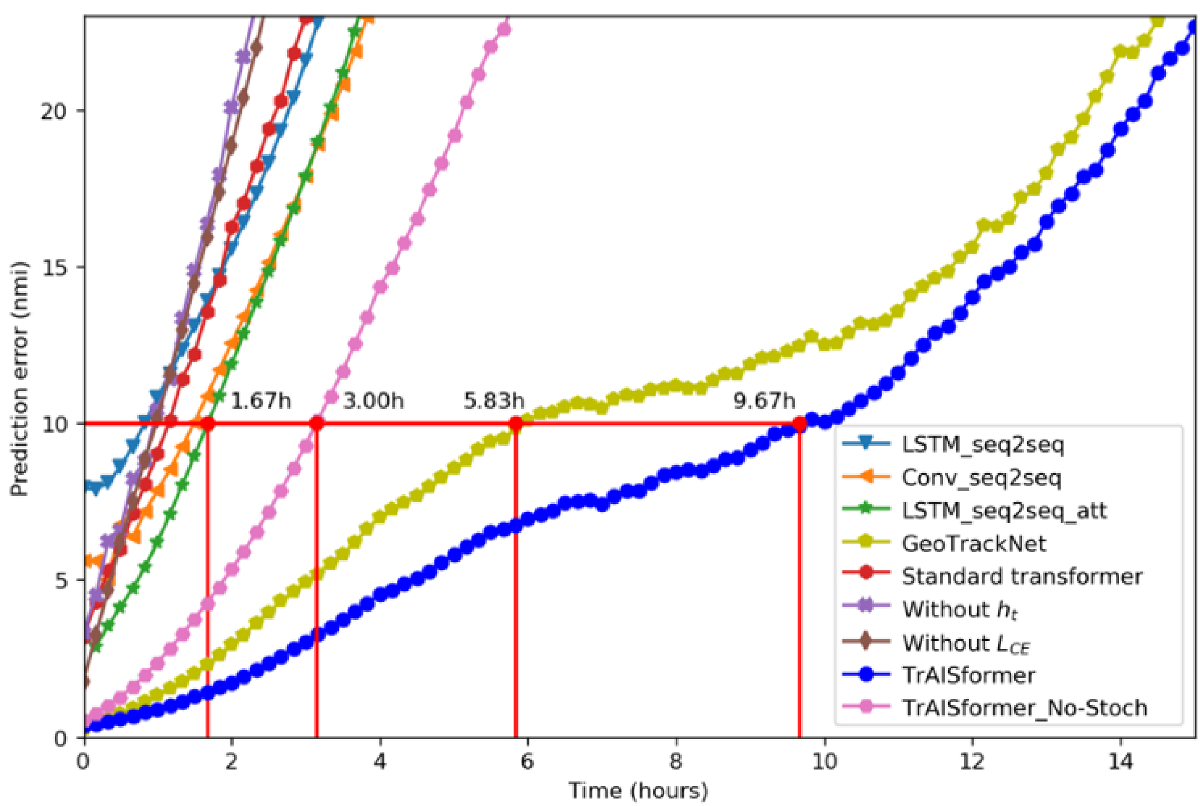

Figure 1. Many machine learning techniques attempting to predict vessel trajectories based on the same AIS training data. Results are measured in prediction error over time.

The dataset for this is AIS data from the Dutch government. The TrAISformer paper utilized this data set to do an apples-to-apples comparison of models, and therefore it is best for me to utilize this same dataset if I want to do a fair comparison of my results to theirs.The test and training dataset are python pickel files in my data directory of my github project. This dataset has the following data cleaning steps applied to it.

- Remove AIS messages with unrealistic speed values (SOG ≥ 30 knots);
- Remove moored or at-anchor vessels;
- Remove AIS observations within 1 nautical mile distance to the coastline;
- Split non-contiguous voyages into contiguous ones. A contiguous voyage [2], [3] is a voyage whose the maximum interval between two consecutive AIS messages is smaller than a predefined value, here 2 hours;
- Remove AIS voyages whose length is smaller than 20 or those that last less than 4h;
- Remove abnormal messages. An AIS message is considered as abnormal if the empirical speed (calculated by dividing the distance travelled by the corresponding interval between the two consecutive messages) is unrealistic, here above 40 knots;
- Down-sample AIS trajectory data with a sampling rate of 10-minute;
- Split long voyages into shorter ones with a maximum sequence length of 20 hours.

References

[1] Nguyen, Duong and Ronan Fablet. “TrAISformer-A generative transformer for AIS trajectory prediction.” ArXiv abs/2109.03958 (2021): n. pag..

[2] D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet, “A Multi-task Deep Learning Architecture for Maritime Surveillance using AIS Data Streams,” in 2018 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Oct. 2018.

[3] D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet, “GeoTrackNet-A Maritime Anomaly Detector using Probabilistic Neural Network Representation of AIS Tracks and A Contrario Detection,” IEEE Transactions on Intelligent Transportation Systems, Feb. 2021.