Inferring modes of transportation from raw unlabelled GPS data using Convolutional Neural Networks

Szymon Ignacy Padlewski,
University College London, Movement Strategies

Background and Motivation

The comprehension of the peoples' transportation patterns, travel behaviours and the modal split of individuals' journeys are essential insights used for demand analysis, traffic management, or optimisation of travel time. Gathering data on commuters' behaviour in cities has been traditionally, conducted through household interviews or phone questionnaires. Nonetheless, collecting modal statistics utilising such methods is laborious, financially expensive and troublesome, due to low response rate, data accuracy and reliability issues. However, the ubiquitous presence and widespread application of Global Positioning System (GPS) technology in everyday life devices have open new possibilities for gathering accurate information regarding peoples transportation choices. With the enormous market penetration of this location-sensing technology, vast datasets are being registered daily , allowing for new applications and individuals' travel pattern analysis.

The aim of this work is to investigate whether modes of transport can be reliably identified from unlabelled GPS data using state-of-the-art classification methods such as a convolutional neural network trained on the dissimilar labelled GPS dataset. The objective of such an approach is to establish a unique framework enabling for wider application of existing transportation modes identification techniques on enormous raw GPS datasets gathered daily through various smartphone apps. Additionally, this thesis will also attempt to create a methodology that would contribute to knowledge by setting an architecture for future testing of new classifiers, refining results obtained by this research.

Data and Methods

In this project, the user's modes of transport are inferred using data from everyday devices with GPS capabilities through the utilisation of a convolutional neural network (CNN). Moreover, this work enhances the already established techniques relying solely on motion features extracted from GPS data for training the model by proposing the inclusion of proximity features, calculated using transportation networks characteristics points' locations. The CNN models are trained on a separate set of features which predictions are inserted in the logistic regression model to derive the ultimate classification. Additionally, in this dissertation, the unlabelled GPS trajectories go through various phases, such as cleaning pre-processing, extraction of single-mode stages, and data segmentation to match the CNN classifier's input layer requirements. Furthermore, the proposed unique methodology enables to tests whether users' modes of transport can be reliably identified from raw unlabelled GPS data gathered using a heterogeneous set of GPS devices and from a dissimilar region than the dataset used for training the models.

Key Findings

The proposed CNN approach combining two kinds of classifiers is based on an ensemble technique called stacked generalisation and proved to be beneficial for the inferring process. The model's accuracy level achieved while solely utilising motion features for training and classification has been strengthened from 82.3 per cent to 83.6 per cent by employing the proposed technique. The 1.3 per cent improvement compared to the existing transportation modes classifiers using CNN depicts a notable potential of the suggested technique. Since this study uses a CNN's architecture proposed by Dabiri and Heaslip (2018), the obtained results might be enhanced by rigorously testing a wide range of CNN layer configurations to establish a more suitable CNN setup for a model using proximity features. However, the effectiveness metrics calculated after manually validating samples of unlabelled GPS trajectories classified by the CNN model mentioned above haven't been sufficient to prove the truthfulness of our core hypothesis. The best overall accuracy of the validation sample measured in this study reached 49 per cent, with an average f-score of 48 per cent.

Value of the research

The proposed method in this dissertation has demonstrated a notable potential, before unexplored to

the author's knowledge, in enhancing transportation modes identification through the stacking process of CNN models relying on different sets of users' trajectories' characteristics. Such an approach mitigates potential problems with feature selection and dimensionality reduction. Moreover, this study has depicted and performed a detailed pre-processing phase enabling the usage of CNN classifiers on unlabelled raw GPS trajectories gathered by a heterogeneous set of GPS devices. The unlabelled GPS data has been transformed into a format complying with CNN's input layers requirements through noise reduction and data cleaning, change stop detection, single-mode stage identification, and feature extraction.

Even though the established results haven't been sufficient enough to consider the process reliable, with the overall accuracy of the validation sample of 49 per cent and the average f-score of 48 per cent, nevertheless, this dissertation with its unique methodology framework can be seen as a proof of concept, and a starting point for future work. In the end, this dissertation has successfully met all the objectives set out at the beginning of the study and contributed to knowledge by helping in paving a path towards the real-world application of transportation modes detection using raw unlabelled GPS data.

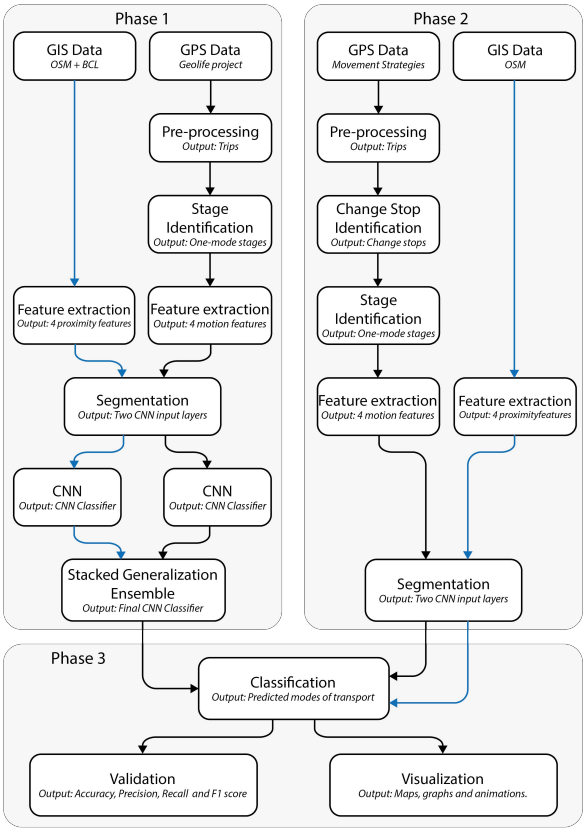

Fig 1– Shows the diagrammatic representation of the three-phased methodology established for this dissertation.

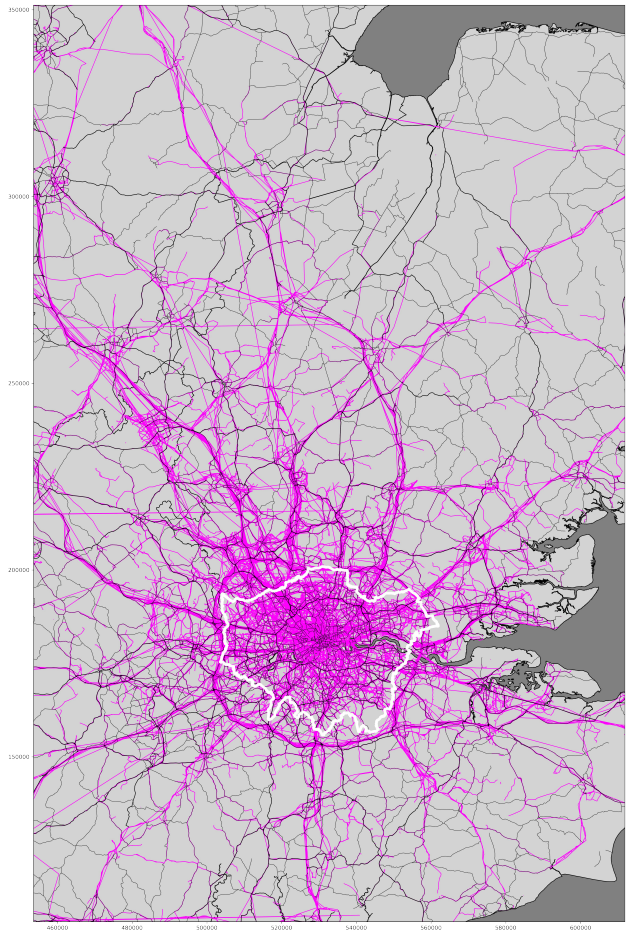

Fig 2– Depicts a map of London and Surrounding Areas with Group three's segments classified as a driving mode by the CNN model in pink and major roads in black.

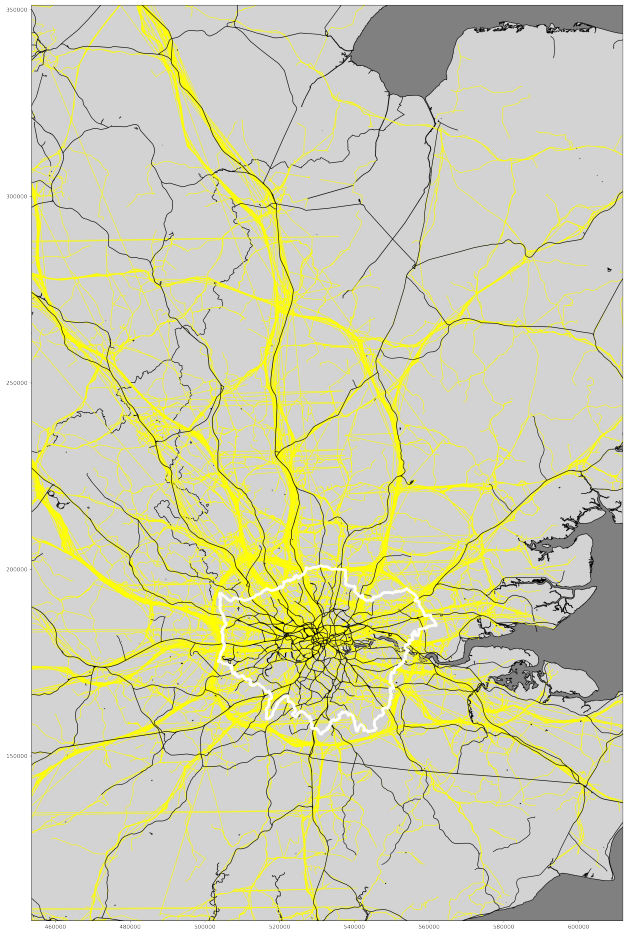

Fig 3– Depicts a map of London and Surrounding Areas with Group three's segments classified as a train mode by the CNN model in yellow and railway network in black.