An Ensemble of ConvTransformer Networks for the Sussex-Huawei Locomotion-Transportation (SHL) Recognition Challenge

Aosheng Tian*

Ye Zhang*

tianaosheng20@nudt.edu.cn

zhangye18@nudt.edu.cn

National University of Defense Technology

Changsha, Hunan Province, China

Chao Ma

National University of Defense Technology

Changsha, Hunan Province, China

machao0408@nudt.edu.cn

Huiling Chen

National University of Defense Technology

Changsha, Hunan Province, China

chl@nudt.edu.cn

Shilin Zhou

National University of Defense Technology

Changsha, Hunan Province, China

slzhou@nudt.edu.cn

ABSTRACT

The task of the SHL recognition challenge 2021 is to recognize eight modes of locomotion-transportation based on radio sequential data collected by smartphones. These data includes GPS reception, GPS location, WiFi reception and GSM cell tower scans. In this challenge, our team (Transformers) presents a recognition scheme. First, a deep model (ConvTransformer) composed of convolutional and Transformer subnets is proposed. The convolutional subnet captures local features, and the Transformer subnet constructs long-term dependencies. Then, the ConvTransformer network is used to recognize different locomotion and transportation modes through location data. Finally, since the *still* and subway categories are easily confused, our team uses another ConvTransformer network to further classify them through cells and location data. Through cross validation techniques on the training dataset, our proposed scheme achieved a F1 score of 0.6838 on the validation dataset.

CCS CONCEPTS

• **Human-centered computing** -> **Ubiquitous and mobile computing**; • **Computing methodologies** -> **Neural networks**;

KEYWORDS

Locomotion and Transportation mode Recognition, Smartphones, ConvTransformer

ACM Reference Format:

Aosheng Tian, Ye Zhang, Huiling Chen, Chao Ma, and Shilin Zhou. 2021. An Ensemble of ConvTransformer Networks for the Sussex-Huawei Locomotion-Transportation (SHL) Recognition Challenge. In *Adjunct Proceedings of the*

*Both authors contributed equally to this research.

This work is licensed under a Creative Commons Attribution International 4.0 License.

UbiComp-ISWC '21 Adjunct, September 21–26, 2021, Virtual, USA

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8461-2/21/09.

<https://doi.org/10.1145/3460418.3479383>

2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (UbiComp-ISWC '21 Adjunct), September 21–26, 2021, Virtual, USA. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3460418.3479383>

1 INTRODUCTION

In recent years, the locomotion-transportation recognition mode recognition based on mobile devices has attracted increasing attention. This is due to the great significance of the analysis of traffic law and the development of transportation demand plan. The 2021 SHL Challenge is to perform locomotion and transportation recognition based on the SHL challenge dataset [7, 8]. This dataset is publicly available, and contains various data sources including GPS (Global Positioning System) reception, GPS location, WiFi (Wireless Fidelity) reception and GSM (Global System for Mobile Communications) cell tower scans.

To tackle the locomotion-transportation recognition problem, various works have been proposed. These works fall into two categories: traditional methods and deep learning methods. We start the introduction from traditional methods. Timmermans extracted the features of time, distance and acceleration from GPS and accelerometer data. Then the Bayesian algorithm was used to recognize 5 modes (bike, car, bus, subway, motorcycle) [1]. Rasmussen [2] utilized the fuzzy theory to recognize transportation modes. Through the fuzzy theory, 4 different transportation modes (car, bus, bike, walk) were recognized. Then, the car and bus were further recognized through the bus route information and station distance information. In the last few years, deep-learning has eliminated the traditional methods based on complicated hand-crafted features, due to its effective feature representation, automatic classification schemes and inspiring performance. Wu [3] combined CNN and LSTM models to perform the transportation model prediction process. Das[4] divided the trajectories into smaller trajectory segments. Then, a deep network was used to perform classification on each segment. The final result was achieved through a voting mechanism. In [5], Byon fed the artificial features of GPS data to a deep network for travel track recognition.

The transportation mode recognition task of the SHL challenge 2021 can be regarded as a multivariate time series classification problem. To tackle this problem, this paper presents a recognition scheme. First, a deep model (ConvTransformer) composed of convolutional and Transformer subnets is proposed. To the best of our knowledge, deep CNNs are good at capturing local features, while lacking the ability in long-term correlation. On the contrary, the Transformers better construct long-term dependencies, while performing badly in local feature extraction. Consequently, these two models are complementary. Our ConvTransformer network was used to perform classification based on locomotion data. Finally, due to the network did not recognize subway and *still* well, we used another ConvTransformer network to classify these two categories. The second network was trained through locomotion and cell tower scans data, due to these data better distinguishes the two different scenes of inside-subway and outside-subway.

The rest of this paper is organized as follows. Section II introduces 2021 SHL dataset. Pre-Processing is described in Section III. We describe the technical details of the ensemble of ConvTransformer networks in Section IV. Experimental results are presented in Section V and the Conclusion in Section VI.

2 DATASET

The purpose of the SHL2021 challenge is to use radio-data for transportation mode recognition. There are eight different categories of radio-data, including *Still*, *Walk*, *Run*, *Bicycle*, *Car*, *Bus*, *Subway* and *Train*, respectively. The original radio-data was recorded by the Huawei Mate 9 mobile phone. These data includes location data, GPS data, Cells data and WiFi data. The whole dataset has been divided into training, validation and testing datasets. The training dataset contains the 59-day recorded data of User 1. The validation dataset contains the 6-day recorded data of User 2 and User 3. Moreover, the testing dataset contains 14-day recorded data of User 2 and User 3. The radio-data is asynchronously sampled by the sensors with a time-varying sampling rate of approximately about 1Hz. The time of the labels is integer seconds. Moreover, due to the limitation of the environment and equipment, the data is incomplete. Compared to the previous data of the motion sensors [6], the radio-data for transportation mode recognition brings novel challenges.

3 PREPROCESSING

We selected location data (latitude, longitude, altitude) and cells signals (3GPP RSRP) for locomotion-transportation recognition. This is because, the location data provides the motion information of transportations. As supplementary to the location data, the signal intensities of cells data help to distinguish the scenes of above- and below-ground. The preprocessing procedures of these data are introduced as follows.

Step1 We processed the selected data to be correlated with the timestamps of the label, and sorted all data in chronological order. The time periods which lose the GPS or location data for more than 10 seconds were regarded as 'GPS off' period. These periods were deleted. Then, the missing values of location data were replaced with zeros. The same operation was also performed on the cells signals. Finally, the cells signals were interpolated through the nearest

Figure 1: Structure of model

neighbor interpolation method. The location data was interpolated through the cubic interpolation method.

Step2 The processed latitude and longitude data of the Step 1 was first transformed to the plane coordinates. Then, the transformed data of adjacent moments were subtracted, and converted to the navigation coordinate system (ENU). Finally, the missing values were replaced with negative ones. Since the missing data problem was usually caused by underground scenes, these negative ones were used to mark these scenes in the sensory data.

Step3 These discontinuous data was first divided into fragments. Each fragment was further segmented with a sliding window of 180 seconds. Then, the segments less than 180 seconds in length were deleted. Finally, the window stride of the training dataset was set to 10 seconds considering training costs and actual conditions. Moreover, the window stride of the validation dataset was set to 1 seconds, which was consistent with the test dataset.

Step4 The label of each segment was the locomotion and transportation mode that appeared the most frequently in the window. Moreover, the Z-score normalization was used to process all data.

4 METHOD

As above mentioned, the locomotion-transportation recognition task can be converted to the multivariate time series classification problem. To tackle this problem, various deep learning networks have been proposed, such as CNNs [13], RNN [11], LSTMs[12] and GRUs. However, these networks have obvious defects in capturing both short-term and long-term dependencies. In this challenge, a deep model (ConvTransformer) composed of convolutional and Transformer [15] subnets is proposed. The convolutional subnet extracts local features, and the Transformer subnet captures long-term dependencies. Consequently, the advantages of these two networks are combined. Figure 1 presents the overall architecture of our proposed model. It consists of two networks, one with a classifier of eight categories and the other with a classifier of two categories. Network1 is used to classify 8 locomotion and transportation modes through location data. Since the *still* and subway categories are easily confused, Network2 further classify these two categories through cells data. Consequently, all the samples identified as subway are reclassified into subway and *still* categories.

Figure 2: Structure of Network1

Figure 3: Structure of Network2

4.1 NETWORK

We propose two ConvTransformer networks that have similar network structures. Both of them are composed of six convolution layers, a transition layer and two transformer layers.

The architecture of Network1 is shown in Figure 1. First, the location data (180×3) are fed to a six-layer convolutional subnet. It is used to aggregate the isolated timestamps into the local features. Each convolutional layer consists of a convolutional operation, a Batch Normalization and a RELU activation function. The convolutional layer consists of 96 filters of 1×3 in size. These 1D-CNNs are superior to the traditional 1D-CNNs, due to its reducing training parameters and enhancing the connection among different univariate time series. Then, the transition layer is followed after the convolutional subnet. This layer comprises $96 \times 1 \times 1$ conv-filters, a Batch Normalization (BN) and a RELU activation function. It is used to achieve appropriate feature dimensions for the Transformer layers. Finally, the processed features are fed to two Transformer layers for long-term correlation. Each Transformer block consists of a self-attention and a feed forward layer with a dropout rate of 0.2.

The architecture of Network2 is shown in Figure2. Compared with the first network, the second model is different by supplementing the modality attention mechanism. First, the location (3×180) and cells data (3×180) are fed to two branches of convolutional subnets separately. Then, the extracted feature vectors of different

modalities (locomotion data and cell signals) are weighted by the modality attention [10, 14]. Finally, the weighted features are concatenated, and input to the following subnets which are the same as the first network. Note that the second network has the same parameters as the first network.

4.2 PARAMETER

In our model training phase, the proposed model was optimized through Focal loss [9] and Adam optimized for 80 epochs. The learning rate was initialized $5e-4$, and decreased according to the training loss. Besides, the batch size was set to 256.

4.3 COMPUTATIONAL RESOURCES

- Physical Memory (RAM) : 32GB
- Language: Python 3.8
- GPU: RTX 2080Ti GPU
- Framework: PyTorch 1.7.1

The size of the trained model is about 3 MB. It took to train the model for about 1.5 hours. It took to evaluate the test dataset for about 4 minutes.

Table 1: Confusion matrix of Network1

	1	2	3	4	5	6	7	8
1	9387	217	0	0	131	213	390	5275
2	310	13836	600	883	47	256	100	1350
3	0	632	1342	257	0	0	0	0
4	0	470	709	8513	0	192	2	0
5	151	91	96	458	13504	1956	268	1016
6	46	8	0	531	769	4297	703	1639
7	422	0	0	46	1213	872	3775	6443
8	47	8	0	145	78	828	2	16043

Table 2: Confusion matrix of Network2

	1	8
1	15402	4510
8	284	19085

Table 3: Confusion matrix of model

	1	2	3	4	5	6	7	8
1	13114	217	0	0	131	213	390	1548
2	1072	13836	600	883	47	256	100	588
3	0	632	1342	257	0	0	0	0
4	0	470	709	8513	0	192	2	0
5	335	91	96	458	13504	1956	268	832
6	1214	8	0	531	769	4297	703	471
7	3326	0	0	46	1213	872	3775	3539
8	747	8	0	145	78	828	2	15343

5 RESULTS

After the training, an F1-score of 0.6664 could be achieved by the 1st network on the validation set. However, it is difficult for this network to distinguish from the *still* and subway categories. Therefore, the location data and the cells data are fed into the 2nd network for further classification of these two categories and an F1-score of 0.8779 is achieved. Then, the Network1 and Network2 are ensemble to obtain a final F1-score of 0.6838 on the validation dataset. In our experiment, the original training dataset and validation dataset are shuffled to build a larger training dataset. The model corresponding to the minimum training loss during 80 epochs was selected to report the final results. The confusion matrices of Network1, Network2 and the whole model are shown in Table 1 to Table 3 respectively. The 8 class activities are: 1 - *Still*; 2 - *Walk*; 3 - *Run*; 4 - *Bike*; 5 - *Car*; 6 - *Bus*; 7 - *Train*; 8 - *Subway*.

6 CONCLUSION

This paper presents a recognition scheme for SHL challenge 2021. First, a deep network named ConvTransformer is proposed. It is composed of convolutional and Transformer subnets. Then, the location data is input to this network for recognizing different locomotion and transportation modes. Finally, since the *still* and subway categories are easily confused, another ConvTransformer

network is trained to further classify these two categories through cell and location data. Our proposed scheme achieves an F1 score of 0.6835 on validation dataset. Unfortunately, we failed to introduce WiFi and GPS data into our scheme for further improvements for improving the classification readiness rate during the competition. We would like to involve this issue as our future works. The recognition result for the testing dataset will be presented in the summary paper of the challenge [16].

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant (No.61903373) and (No. 62002372).

REFERENCES

- [1] Vij, Akshay, André Carrel, and Joan L. Walker. "Incorporating the influence of latent modal preferences on travel mode choice behavior." *Transportation Research Part A: Policy and Practice* 54 (2013): 164-178.
- [2] Rasmussen, Thomas Kjær, et al. "Improved methods to deduct trip legs and mode from travel surveys using wearable GPS devices: A case study from the Greater Copenhagen area." *Computers, Environment and Urban Systems* 54 (2015): 301-313.
- [3] Wu, Yunkai, and Huachun Tan. "Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework." *arXiv preprint arXiv:1612.01022* (2016).
- [4] Das, Rahul Deb, Nicole Ronald, and Stephan Winter. "A Simulation Study on Automated Transport Mode Detection in Near-Real Time Using a Neural Network." *Research at Locate* (2015).
- [5] Byon, Young-Ji, et al. "Real-time transportation mode identification using artificial neural networks enhanced with mode availability layers: A case study in Dubai." *Applied Sciences* 7.9 (2017): 923.
- [6] Wang, Lin, et al. "Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2020." *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 2020.
- [7] Gjoreski, Hristijan, et al. "The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices." *IEEE Access* 6 (2018): 42592-42604.
- [8] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, and D. Roggen, "Enabling reproducible research in sensor-based transportation mode recognition with the Sussex-Huawei dataset." *IEEE Access* 7 (2019): 10870-10891.
- [9] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." *Proceedings of the IEEE international conference on computer vision*. 2017.
- [10] Vaswani, Ashish, et al. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).
- [11] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).
- [12] Guan, Yu, and Thomas Plötz. "Ensembles of deep lstm learners for activity recognition using wearables." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.2 (2017): 1-28.
- [13] Ha, Sojeong, and Seungjin Choi. "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors." *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016.
- [14] Ma, Haojie, et al. "AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition." *IJCAI*. 2019.
- [15] Liu, Zhouyong, et al. "ConvTransformer: A Convolutional Transformer Network for Video Frame Synthesis." *arXiv preprint arXiv:2011.10185* (2020).
- [16] L.Wang, M.Ciliberto, H.Gjoreski, P.Lago, K.Murao, T.Okita, and D.Roggen. "Locomotion and transportation Mode Recognition from GPS and radio signals: Summary of SHL Challenge 2021", *Proceedings of the 2021 ACM International Joint Conference and 2021 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2021.