DenseNetX and GRU for the Sussex-Huawei Locomotion-Transportation Recognition Challenge

Yida Zhu
Beijing University of Posts and Telecommunications
Beijing, China
dozenpiggy@bupt.edu.cn

Haiyong Luo[†]
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
yhluo@ict.ac.cn

Runze Chen
Beijing University of Posts and Telecommunications
Beijing, China
chenrz925@bupt.edu.cn

Fang Zhao[†]
Beijing University of Posts and Telecommunications
Beijing, China
zfsse@bupt.edu.cn

Lin Su
Capital Normal University
Beijing, China
13207697853@163.com

ABSTRACT

The Sussex-Huawei Locomotion-Transportation (SHL) recognition challenge organized at the HASCA Workshop of UbiComp 2020 presents a large and realistic dataset with different activities and transportation. The goal of this human activity recognition challenge is to recognize eight modes of locomotion and transportation from 5-second frames of sensor data of a smartphone carried in the unknown position. In this paper, our team (We can fly) summarize our submission to the competition. We proposed a one-dimensional (1D) DenseNetX model, a deep learning method for transportation mode classification. We first convert sensor readings from the phone coordinate system to the navigation coordinate system. Then, we normalized each sensor using different maximums and minimums and construct multi-channel sensor input. Finally, 1D DenseNetX with the Gated Recurrent Unit (GRU) model output the predictions. In the experiment, we utilized four internal datasets for training our model and achieved averaged F1 score of 0.7848 on four valid datasets.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**; • **Hardware** → *Sensor applications and deployments*; • **Theory of computation** → *Design and analysis of algorithms.*

KEYWORDS

Activity recognition, Transportation mode recognition, Deep learning, DenseNet, GRU, Smartphone

UbiComp/ISWC '20 Adjunct, September 12—16, 2020, Virtual Event, Mexico
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8076-8/20/09…$15.00
https://doi.org/10.1145/3410530.3414349

ACM Reference Format:
Yida Zhu, Haiyong Luo, Runze Chen, Fang Zhao, and Lin Su. 2020. DenseNetX and GRU for the Sussex-Huawei Locomotion-Transportation Recognition Challenge. In *Proceedings of UbiComp/ISWC '20 Adjunct.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3410530.3414349

1 INTRODUCTION

In recent years, the ownership of mobile devices, especially smartphones, has grown rapidly. Hence, the smartphone has almost become an integral part of human life. Human activity recognition is one of the widely studied topics in recent decades. Knowledge about the specific human behavior in urban areas is beneficial to healthcare monitoring, safe driving, and journey planning. With the powerful perception capabilities and ever-increasing computing of mobile devices, human-activity-recognition-based smartphones attract much attention. Transportation mode recognition as a branch of human behavior recognition has a great influence on human life such as traffic management, route or parking recommendation, and footprint analysis.

The SHL recognition challenge Dataset [3, 12] contains multimodal locomotion and transportation data, which can be used in an activity recognition challenge. The goal of the 2020 SHL recognition challenge focuses on recognizing transportation modes in a user-independent manner with an unknown phone position. According to the experience in SHL recognition challenge 2018 [13] and SHL recognition challenge 2019 [11], we report that it is challenging to train a model using the smartphone data collected at a specific body position and test the model using the data collected from a new position. JSI_First [7] derived additional sensor streams from the existing ones and calculated a large body of features. They then used cross-location transfer learning via specialized feature selection and performed a two-step classification. Yonsei-MCML [2] proposed a deep multimodal fusion model. The sensor data are independently pre-processed via a convolutional neural network (CNN), and the results are combined with the EmbraceNet fusion algorithm. We-can-fly [14] employed a 1D DenseNet model working on the multi-channel sensor data simultaneously. Another challenge in SHL recognition challenge 2020 is the diversity of devices and users.

The mismatch between the different devices data in training and testing phase degrades the performance significantly.

In this paper, following with the 1D DenseNet model, we proposed a one-dimensional (1D) DenseNetX model for transportation classification. We independently extract higher-level features from magnetometer, acceleration, linear acceleration, gravity, gyroscope sensor by 1D DenseNetX. Then we slice the feature-map extracted by multi-sensor 1D DenseNetX and concatenate the features in chronological order. The concatenated features are fed into the GRU [1] model, and the output features of GRU are used together with the DenseNetX extracted features for the final classification. Furthermore, we employ the multi-model fusion to improve the performance. We participate in the SHL recognition challenge 2020 challenge under the team name of "we can fly".

2 SHL RECOGNITION CHALLENGE 2020 DATASET

The SHL recognition challenge dataset contains eight modes of locomotion and transportation, including still, walking, run, bike, car, bus, train, and subway. The 2020 challenge used a subset of the SHL recognition challenge Dataset, and all of the participant teams aim to accomplish the task below. The sponsor divided the data into three parts: train, validate, and test. The data comprises of 59 days of training data, 6 days of validation data, and 40 days of test data. The train, validation, and test data were generated by segmenting the whole data with a non-overlap sliding window of 5 seconds. Each sample contains 500 sensor values, which are acquired with a sampling rate of 100 Hz.

The train data contains the raw sensors data from one user (user 1) and four phone locations (bag, hips, torso, hand). The validation data contains the raw sensors data from the other two users (mixing user 2 and user 3) and four phone locations (bag, hips, torso, hand). Both the training and validation splits contain the ground-truth labels along with the sensor data, whereas the testing split does not. The goal of the SHL recognition challenge 2020 is to recognize the user activity from data coming from the phone of the "test" user (a combination of user 2 and user 3). The location of that phone on the "test" user is not specified. Recognizing modes of transportation in a user-independent manner with an unknown phone position is more challenging than before.

3 MODEL DESCRIPTION

1D Multi-Sensor DenseNet [14] has achieved success in the 2019 SHL recognition challenge. Hence, we modified the 1D Multi-Sensor DenseNet and named it 1D Multi-Sensor DenseNetX. Figure 1 shows the structure of the 1D Multi-Sensor DenseNetX model, including multiple sub-DenseNetX. Each sub-DenseNetX receives multi-channel sensor data as input. Six sensor data, including geomagnetic, acceleration, linear acceleration, gravity acceleration, gyroscope, and barometric, are used to extract transportation mode features. First, the sensor input passes through the input convolution block. Then the outputs of the input convolutional layer will feed into the first depth-wise separable dense block. Each depth-wise separable dense block is followed by a transition layer to transition and compress the channels. The output of the transition layer will feed into the next depth-wise separable dense block. In our model, we perform Batch Normalization (BN) [6] and SELU [9] before each convolution layer.

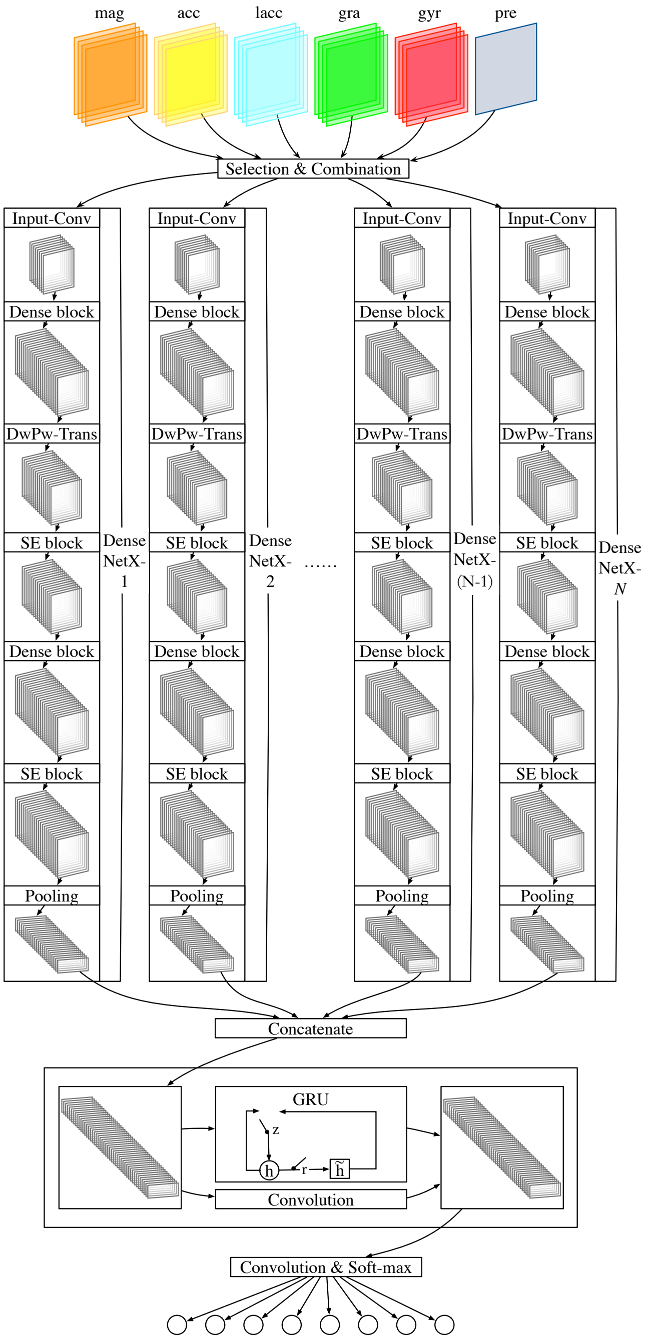

Figure 1: The system architecture of the proposed 1D Sensor DenseNetX and GRU model.

3.1 Input Convolutional Block

The input convolutional block consists of two different convolutional types which extract features from different receptive field.

We define the number of feature-map generated after each convolution as growth-rate k. The first type of convolution is a full channel convolution operation, which extracts 2k feature-maps. The full channel convolution operation is the standard convolution we used to exploit the features of all channels in the original input. The second type of convolution is to extract features by each channel independently, which we name this kind of convolution as SPC Conv. The input data comprise of multi-axis sensor data and each channel corresponds to one axis. SPC Conv focuses on exploit features on each axis and generates a total of 2k feature-maps. We then concatenate the output of full channel convolution and SPC Conv together to 4k feature-maps.

3.2 Depth-wise Separable Dense Block

The depth-wise separable dense block is the basic block in the DenseNetX model. In our configuration, each DenseNetX model contains 2 depth-wise separable dense blocks. Each depth-wise separable dense block consists of 12 “depth-wise-point-wise operation”. Each “depth-wise-point-wise operation” as a consecutive operation that continuously executes one-dimensional depth-wise convolution with a kernel size of 3 and a one-dimensional pointwise convolution with a kernel size of 1 that generates k feature-maps. The characteristics of DenseNet [5] is to connect each layer to every other layer in a feed-forward fashion in the basic block. We strengthen feature propagation and encourage feature reuse in our model. For each “depth-wise-point-wise operation”, the feature-maps of all preceding layers are used as inputs in each basic block.

3.3 Transition Layer

The transition layer connects two depth-wise separable dense blocks. We follow with [14] that design the transition layer to improve model compactness by reducing the number of feature-maps and the dimensional of each feature-map in the transition layer. The transition layer includes two parts, a “depth-wise-point-wise operation” and a max-pooling layer. The “depth-wise-point-wise operation” needs a compression rate to limit the output channel number, which is configured to 0.8 * input channel number.

3.4 SE Block

Inspired by [4], we introduce the attention mechanism to calibrate the features extracted by the convolutional operation. We perform the feature calibration on the features extracted after the first dense block is compressed by the transition layer and the features extracted from the last block.

3.5 Feature Combination

After obtaining the feature-maps of 1D multi-Sensor DenseNetX, we leverage adaptive max pooling to align the dimensions of all feature-maps. Since the features are extracted according to the positive sequence convolution in the time series according to a frame of data, the output feature-map can also be argued as obtained under different receptive field windows according to the time series. We splice the values of the same position among different feature-maps. The new features after splicing can be used as input features in each step of the GRU network. Thereby we introduce a GRU network that extracts the temporal features. Finally, each feature-map of 1D multi-Sensor DenseNetX will be transformed into a 1-dimensional feature by a learnable matrix. We then concatenated these features with the temporal features of GRU together to the classification layer with an output size of 8. The output of the classification layer will be processed using soft-max, and the most appropriate label will be selected.

4 EXPERIMENTS

In our pre-processing data phase, since the sensor reading is obtained in the coordinate system of the mobile phone, the sensor reading varies significantly in different phone locations. To eliminate the difference of sensor data caused by different phone locations, we convert geomagnetic, acceleration, linear acceleration, gravity, and gyroscope data from the phone coordinate system to the navigation coordinate system. In our model training phase, we set the growth-rate k to 12 and use cross-entropy as a function of loss and use Adam [8] to optimize during training. During the training process, we set a starting learning rate of 5e-4, every 5 epochs to reduce the learning rate to the previous 20%. In our parameter tuning phase, we leverage four datasets (train_bag, train_hand, train_hips, train_torso) to train our model and four valid datasets (valid_bag, valid_hand, valid_hips, valid_torso) to validate our model. Each training includes a total of 30 epochs, and we select the model parameters corresponding to the lowest loss in the valid datasets from 30 epochs as the final model. During our commit phase, we trained with all the data to improve the performance of the model.

4.1 Experimental Results on Valid Datasets

In the previous SHL recognition challenge, F1-score is adopted for evaluation metric. Therefore, we evaluate our model using the F1-score method by four validation datasets. We obtained F1 scores of 0.8419, 0.7337, 0.7553, and 0.8082 on valid datasets of valid_bag, valid_hand, valid_hips, and valid_torso respectively. From the comparison of the accuracy of the valid datasets of four different phone positions, we found that the accuracy of the hand dataset was the lowest, indicating that the influence of the device and user heterogeneity on the hand data was the greatest. On the contrary, the mobile phone position is less disturbed by other behaviors of users when it is in the bag, so the model can also obtain higher accuracy in the migration of different users and devices. Besides, Tables 1 to 4 show the confusion matrix on the four valid datasets. The 8 class activities are: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 – Subway. It can be seen from the confusion matrix that trains and subways are two kinds of traffic patterns that are difficult to distinguish in the hand position.

4.2 Computational Resources

We used 1 GPU on 1 server.

- 1 GPU (NVIDIA TESLA V100), Intel(R) Xeon(R) Gold 6132 CPU (2.6GHz 14cores/28threads), 128GB RAM
- Language: Python 3.8.0
- Framework: PyTorch 1.5

The size of the trained model is about 35 MB. It took to train the model for about six hours. It took to evaluate the test dataset for about 2 minutes.

Table 1: Confusion matrix of valid_bag

	1	2	3	4	5	6	7	8
1	5323	86	0	29	44	28	264	190
2	364	4683	10	55	0	7	56	55
3	1	22	355	176	0	0	0	1
4	68	65	0	2071	38	31	93	40
5	26	4	0	0	3128	806	70	60
6	4	13	0	4	95	1680	31	9
7	134	26	0	11	60	51	3757	323
8	26	11	0	3	10	3	509	3780

Table 2: Confusion matrix of valid_hand

	1	2	3	4	5	6	7	8
1	5155	77	1	51	67	103	382	128
2	348	4373	7	153	37	154	119	39
3	3	49	494	6	0	2	1	0
4	63	131	14	1287	31	691	144	45
5	257	28	0	69	1914	1230	542	54
6	44	12	0	3	183	1536	53	5
7	274	23	0	14	102	156	3325	468
8	128	14	0	6	118	18	829	3229

Table 3: Confusion matrix of valid_hips

	1	2	3	4	5	6	7	8
1	5491	50	0	13	27	38	234	111
2	441	4599	5	64	2	13	68	38
3	2	23	329	199	0	2	0	0
4	109	610	10	1444	11	80	109	33
5	244	4	0	0	1996	1277	511	62
6	48	17	0	0	56	1592	111	12
7	185	26	0	8	53	115	3710	265
8	88	11	0	0	5	4	756	3478

Table 4: Confusion matrix of valid_torso

	1	2	3	4	5	6	7	8
1	5459	36	0	53	38	17	226	135
2	442	4556	8	110	8	6	66	34
3	1	65	367	111	0	6	0	5
4	59	39	0	1554	10	31	610	103
5	143	5	0	13	3104	645	111	73
6	109	9	0	18	25	1585	81	9
7	323	22	0	6	6	15	3667	323
8	90	10	0	5	4	2	794	3437

5 CONCLUSION

We proposed 1D Sensor DenseNetX for the SHL recognition challenge. We construct a sub-DenseNetX for geomagnetic, acceleration, linear acceleration, gravity, gyroscope, and pressure sensor. Then we connect the output of all subnets and feed the features into GRU. After soft-max, we got the final classification result. As a result of classifying test data of the SHL recognition challenge, our model obtained F1 scores of 0.8419, 0.7337, 0.7553, and 0.8082 on valid datasets of valid_Bag, valid_Hips, valid_Torso, and valid_Hand respectively. Overall, the average F1-score across the four valid datasets was 0.7848. The above results indicate that the location of smartphones, especially hand position, still has a great influence on traffic pattern recognition. The recognition result for the testing dataset will be presented in the summary paper of the challenge [10].

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program under Grant 2018YFB0505200, the Action Plan Project of the Beijing University of Posts and Telecommunications supported by the Fundamental Research Funds for the Central Universities under Grant 2019XD-A06, the Special Project for Youth Research and Innovation, Beijing University of Posts and Telecommunications, the Fundamental Research Funds for the Central Universities under Grant 2019PTB-011, the National Natural Science Foundation of China under Grant 61872046, the Joint Research Fund for Beijing Natural Science Foundation and Haidian Original Innovation under Grant L192004, the Key Research and Development Project from Hebei Province under Grant 19210404D and 20313701D, the Science and Technology Plan Project of Inner Mongolia Autonomous Regio under Grant 2019GG328 and the Open Project of the Beijing Key Laboratory of Mobile Computing and Pervasive Device.

REFERENCES

[1] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. (2014), 1724–1734.

[2] Jun-Ho Choi and Jong-Seok Lee. 2019. EmbraceNet for Activity: A Deep Multimodal Fusion Architecture for Activity Recognition. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(UbiComp/ISWC '19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 693–698. https://doi.org/10.1145/3341162.3344871

[3] H. Gjoreski, M. Ciliberto, L. Wang, F. J. Ordonez Morales, S. Mekki, S. Valentin, and D. Roggen. 2018. The University of Sussex-Huawei Locomotion and Transportation Dataset for Multimodal Analytics With Mobile Devices. *IEEE Access* 6 (2018), 42592–42604. https://doi.org/10.1109/ACCESS.2018.2858933

[4] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. 2020. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 8 (2020), 2011–2023.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[6] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (2015), 448–456.

[7] Vito Janko, Martin Gjoreski, Carlo Maria De Masi, Nina Reščič, Mitja Luštrek, and Matjaž Gams. 2019. Cross-Location Transfer Learning for the Sussex-Huawei Locomotion Recognition Challenge. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(UbiComp/ISWC '19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 730–735. https://doi.org/10.1145/3341162.3344856

[8] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[9] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-Normalizing Neural Networks. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 971–980. http://papers.nips.cc/paper/6698-self-normalizing-neural-networks.pdf

[10] M. Ciliberto P. Lago K. Murao T. Okita L. Wang, H. Gjoreski and D. Roggen. 2020. Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2020. In *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 (UbiComp/ISWC '20 Adjunct)*. Association for Computing Machinery. https://doi.org/10.1145/3341162.3344872

[11] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2019. Summary of the Sussex-Huawei Locomotion-Transportation Recognition Challenge 2019. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(UbiComp/ISWC '19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 849–856. https://doi.org/10.1145/3341162.3344872

[12] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, and D. Roggen. 2019. Enabling Reproducible Research in Sensor-Based Transportation Mode Recognition With the Sussex-Huawei Dataset. *IEEE Access* 7 (2019), 10870–10891. https://doi.org/10.1109/ACCESS.2019.2890793

[13] Lin Wang, Hristijan Gjoreskia, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2018. Summary of the Sussex-Huawei Locomotion-Transportation Recognition Challenge. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 1521–1530. https://doi.org/10.1145/3267305.3267519

[14] Yida Zhu, Fang Zhao, and Runze Chen. 2019. Applying 1D Sensor DenseNet to Sussex-Huawei Locomotion-Transportation Recognition Challenge. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(UbiComp/ISWC '19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 873–877. https://doi.org/10.1145/3341162.3345571