Learning from Smartphone Location Data as Anomaly Detection for Behavioral Authentication through Deep Neuroevolution

Mhd Irvan, Tran Phuong Thao, Ryosuke Kobayashi, Toshiyuki Nakata and Rie Shigetomi Yamaguchi
Graduate School of Information Science and Technology, The University of Tokyo, Japan

Keywords: Behavioral Authentication, Machine Learning, Deep Neuroevolution.

Abstract: Passwords and face recognition are some examples of many approaches to authenticate smartphone users. These approaches typically authenticate users at an initial log-in or unlock session, and there are risks of an unauthorized person using the authenticated account if the smartphone owner lose their device while still in unlocked status. Because of this reason, there is a necessity to continuously authenticate from time to time. Passwords and biological biometrics-based authentication procedures are impractical for this kind of situation because they require constant interruption. In this early research we are applying a behavioral authentication approach implementing location history data to implicitly authenticate users. Traits derived from users' movements are easy to monitor and hard to fake. Previously visited locations represent patterns within people's daily behaviors and in this paper we are proposing deep learning method evolved by genetic algorithms to recognize such patterns and to correctly authenticate people that match the patterns.

1 INTRODUCTION

Smartphones play important roles in many people's daily life. People commonly use smartphone applications to take photos, send messages, book rides, or shop online. It is not unusual for those applications to ask private information (such as names, gender, or credit card information) from their users to improve the quality of their service. The sensitive nature of those private information requires application developers to properly secure access to their service.

A very popular way to secure such access is by asking passwords from users during login process. However, passwords and other knowledge-based authentication methods such as PIN (personal identification number) codes carry great risk as users tend to use the same passwords across multiple services. Thus, many services currently require additional possession-based authentication method before granting access (Zviran, 2006). A typical way of this implementation is by sending a unique code through SMS (short message service) to users' phone numbers. This extra step is known as 2-factor authentication (2FA) or multi-factor authentication (MFA) (Banyal, 2013).

Unfortunately, possession-based authentication methods bring potential inconveniences to users because they may have to carry additional devices which can be easily lost. Many users also use the same smartphone to input passwords and receive 2FA codes. Thus, if their smartphone is stolen, attackers can bypass 2FA checks (Velásquez, 2018).

To further secure users against such situations, many smartphone makers offer the possibility to use inherence factors to authenticate users. These inherence factors often take the form of biological biometrics information such as fingerprints or facial expressions. Biometrics authentication methods offer more seamless authentication because users only need to provide information that they already have and carry all the time (Ogbanufe, 2018).

All previously mentioned authentication methods typically grant a one-time unlock session during login process, and if the smartphones are stolen while being in unlocked status, whoever steals the device could have access to private data contained inside. Because of this, many services ask to re-authenticate when a certain amount of time has passed. Nevertheless, the period between those re-authentications is still prone to attacks.

Such threats introduce the necessity to regularly authenticate users from time to time. This approach is known as continuous authentication (Feng, 2017). It requires users to continuously provide credentials to

Irvan, M., Thao, T., Kobayashi, R., Nakata, T. and Yamaguchi, R.
Learning from Smartphone Location Data as Anomaly Detection for Behavioral Authentication through Deep Neuroevolution.
DOI: 10.5220/0010395407230728
In *Proceedings of the 7th International Conference on Information Systems Security and Privacy (ICISSP 2021)*, pages 723-728
ISBN: 978-989-758-491-6

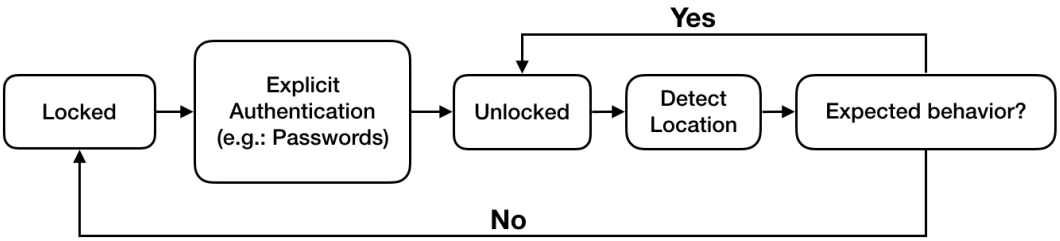

Figure 1: Flow of authentication.

prove their authenticity. Passwords, 2FA codes, and biometrics-based authentication methods are considered to be not suitable to continuously authenticate users due to the inconvenience of manually inputting those information multiple times. The laborious process of inputting such information led researchers to look for implicit factors to be applied into continuous authentication.

Implicit factors, such as user behaviors, are suitable for continuous authentication because they do not require constant interaction from users and can be done unobtrusively in the background without interrupting user activities. Furthermore, behaviors are unique to each person and hard to mimic by another (Sitová, 2015). Methods to authenticate users by learning from their past behaviors are often referred as behavioral authentications methods. Behavioral authentications also have their own issues. The unique nature of people's behaviors means that it is challenging to recognize the patterns that define them.

In this paper, we propose a method that recognizes users behaviors through their location history data gathered though their smartphone's built-in GPS (Global Positioning System) sensor. Our method continuously authenticates users based on their past behavioral patterns. When our proposed method recognizes that the owner is no longer accompanying the phone, smartphone developers may use those information to lock access to the phone to prevent further access by asking for explicit re-authentication, such as passwords or facial expression (figure 1).

We implement deep neuroevolution models (Such, 2017), which combine Deep Neural Network (DNN) architectures (Szegedy, 2013) with Genetic Algorithm (GA) operations (Goldberg, 2006), to learn from users location history and find patterns inside their moving behaviors to regularly authenticate users based on their current location. Through a collaborative research project between our affiliated university and various commercial companies, behavioral data from over 7,000 smartphone users were collected.

To evaluate the feasibility of our proposed method, we conducted early experiments on a small number of users inside the dataset. Our early findings from the experiments demonstrate that our model can be used to detect anomaly in expected users' locations with relatively high accuracy.

2 RELATED WORK

Hsieh and Leu (Hsieh, 2011) proposed an authentication scheme which exploits One-Time Passwords (OTPs) based on the time and location information of the mobile device to authenticate users while accessing Internet services, such as online banking services and e-commerce transactions. Their research demonstrated that location information can be used to correctly authenticate genuine users. However, their research is applicable only for a one-time authentication session, instead of continuously repeated. This limitation comes from the requirement to manually enter SMS-based OTPs based the time and location.

Ghogare et al. (Ghogare 2012) also showed that location can be used as one of the credentials to give access to data only to legitimate user. However, their system was not designed for smartphone users in mind. They implemented dedicated GPS devices to get the location information of users. The location information is transferred during an explicit authentication session so their location-based authentication approach is not suitable to implicitly authenticate users in the background without interrupting user activities.

Zhang et al. (Zhang, 2012) applied a location-based authentication for mobile transaction using smartphones. Similar to Ghogare et al. research, their authentication is also applied during an explicit authentication session, instead of implicitly. However, the location information in their research comes from users smartphones, instead of separate GPS devices. They showed that since users typically carry their smartphones everyday and everywhere, the amount of location information is richer and contribute towards stronger location-based authentication.

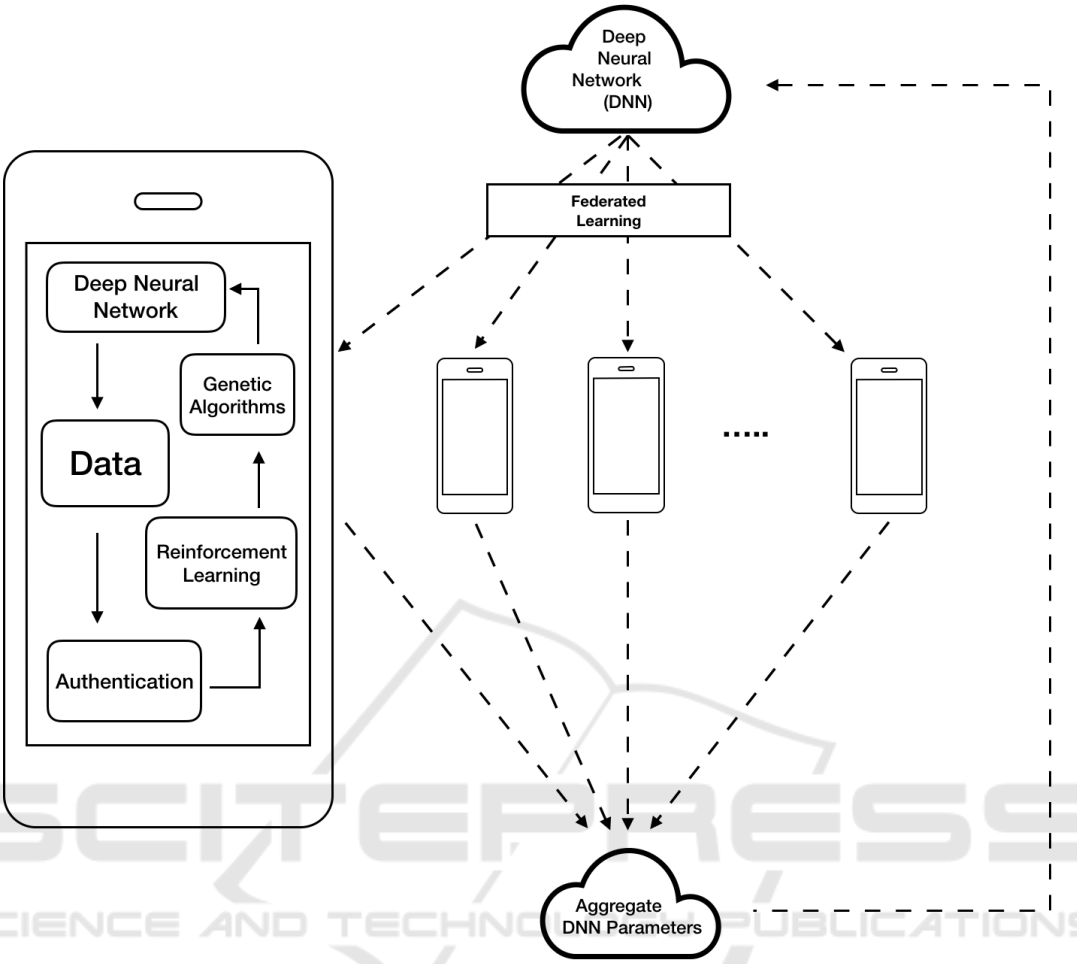

Figure 2: Proposed approach.

3 PROPOSED APPROACH

Deep neuroevolution concept is proposed as our authentication algorithm. Deep neuroevolution is a concept of evolving Deep Neural Network (DNN) models with Genetic Algorithm (GA) operators instead of with standard gradient descent method (Such, 2017). It founds success in problem areas where patterns in data continuously change and finding an optimal DNN model in such situation is a laborious process. Similarly, learning from location data is challenging because people's behaviors occasionally change and a good DNN model for past behaviors may not necessarily produce good results for the changed behaviors (Miikkulainen, 2019). For this reason, the DNN parameters of our authentication models are regularly evolved by GA operators to adapt to the dynamics of users' behaviors.

Location history carries rich information. Places typically visited on certain time represents unique behaviors of a person. Furthermore, in addition to regularly visiting favorite places, people also visit new places from time to time. Correctly authenticating at new places is challenging. Meanwhile, discovering a good neural network for a particular kind of behavior is a laborious process. There is no general ideal number of layers and nodes when designing a neural network for location data. Researchers have recently found that GA can be used to automate this design process. GA operators, such as mutation and crossover, can be used to autonomously evolve a neural network into a shape suitably work with the data it is fed with. Consequently, our DNN will continue to evolve to match new behaviors.

To maintain privacy, our model is designed to work exclusively with locally kept data inside each smartphone. At first, an initial global model of DNN

is distributed across users' device. This DNN feed exclusively on the device's user location history as inputs and learn the movement patterns behind their travel. GA operators will then regularly change the DNN parameters, namely number of hidden layers, number of nodes at each hidden layers, and weights between them. Consequently, each device maintains their own evolved DNN to match its own user behaviors. Furthermore, Reinforcement Learning (RL) is used to give feedbacks the system based on the authentication accuracy produced by DNN. Finally, the values of DNN parameters from each smartphone are aggregated and averaged to build a new global DNN model in a similar fashion to Federated Learning (FL) to be re-distributed across users' device again (figure 2).

Inputs for the DNN are defined as pairs of "time" and "place". Since people's behavior may vary greatly between each day, as well as between weekdays and weekends, "time" is defined as a collection of labels of "day", "weekday/weekend", "hour", and "minutes". Meanwhile, "place" is defined as a pair of "latitude" and "longitude" information.

DNN is initially trained with the first week of location history, where regularly visited locations at particular time are assumed as usual paths, and the less frequently visited locations are assumed as unusual paths. When the authentication system encounters unusual locations at particular time during the learning process, it observe the deviation distance between the expected location and time and assumes authenticity for a while. This is essential for DNN to grasp the nature of its user's change in behavior. Reinforcement learning is used to give feedbacks to the neural network about the consequence of its observation and GA operators are tasked to update the parameters based on those feedbacks. Each device will evolve the neural network independently to better fit their user's data. Overtime, the DNN will understand better whether the deviation is a change of behaviors or an anomaly due to theft (figure 3). After a predetermined number of evolution rounds, every device reports to the remote server informing their evolved design and the remote server will then average the evolution parameters (number of layers, nodes, and the respective weights between them) found in all evolved neural networks sent by the clients to generate a new global neural network. This new global neural network is again dispatched to all clients to repeat the evolution process and evolve further. The information reported to the remote server contains strictly only machine learning parameters. No actual location data is being transmitted into the server to maintain privacy.

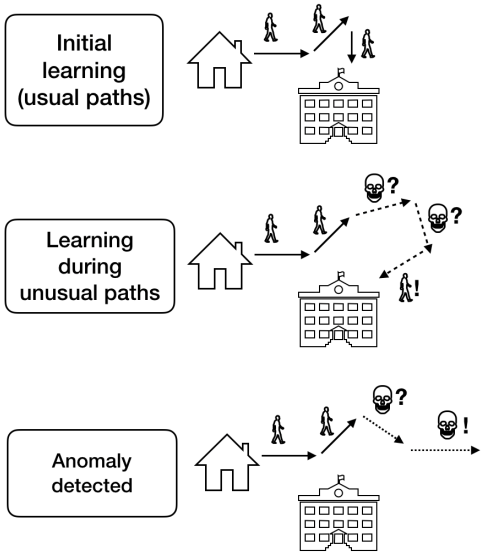

Figure 3: Observing change of behaviors.

4 EARLY EXPERIMENT

We conducted experiments by giving a real-world dataset collected from regular smartphone users as inputs to our proposed method. This dataset reflects their actual lifestyle.

4.1 Data Collection Method

Our dataset came from a collaborative project between our affiliated university and several commercial companies which develop and distribute smartphone applications in Japan. These applications are relatively popular amongst smartphone users in Japan. Users of the applications were presented with an option to participate in a university project that study and analyze smartphone users' lifestyle habit. A website address to the detailed project description hosted on the university website was also linked. By agreeing to participate in the project, users agree to share data about their smartphone, application usage, and their location data. Users were clearly informed that their data would be limited to analysis by the university laboratory and their privacy would be strictly protected. The project passed an extensive review by the

Table 1: An example of collected data from a user.

ID	Date	Time	Latitude	Longitude
1	2017XXXX	21 : XX	31.XXXXXX	139.XXXXXX
1	2017XXXX	21 : XX	31.XXXXXX	139.XXXXXX
1	2017XXXX	21 : XX	31.XXXXXX	139.XXXXXX
1	2017XXXX	21 : XX	31.XXXXXX	139.XXXXXX
1	2017XXXX	21 : XX	31.XXXXXX	139.XXXXXX
...	...	...	...	...

Table 2: Summary of result for each experiment.

Number of random users	Training time	Authentication time	FAR	FRR
100 users	1 hour	0.8 seconds	0.94%	1.74%
100 users	2 hours	2.0 seconds	0.71%	1.21%
200 users	1 hour	0.8 seconds	0.98%	1.93%
200 users	2 hours	2.1 seconds	0.84%	1.62%

university's ethical committee and deemed to be appropriately implemented.

4.2 Dataset

The dataset used in this paper contains information about GPS coordinates every 5 to 10 minutes gathered from 7,236 smartphone users over a period from February 2017 to April 2017. Table 1 illustrates the contents of data of a user. Because this research is still in its early phase, and we we would like to initially validate our approach, we only used small subsets of the whole data.

For our early experiments, we created two groups of users. A group of 100 randomly selected users and a group of 200 randomly selected users. We trained our model by initially feeding the first week of location history as training data and the system is tasked to continuously authenticate users every 5 minutes during the following week. We gradually increased the size of training data by a week each time up until the second to last week of the collected data.

4.3 Early Results and Discussions

We run our experiments two times for each group. In the first experiment, DNN was trained and evolved by GA for one hour, while in the second experiment DNN was trained and evolved by GA for two hours. Our experiments showed that after two hours training the model, GA evolved DNN into a size so large that it requires 2 seconds or more to authenticate users on average. Meanwhile, training our model for 1 hour, produced a DNN that is still capable to authenticate users within 1 second. Table 2 summarizes the average False Accept Rate (FAR) and False Reject Rate (FRR) for each of our experiments.

As we can see from table 2, DNN models trained and evolved for 2 hours produced better FAR and FRR in both groups of users. They did, however, require a longer time to authenticate than the models trained in shorter time. While in a traditional explicit authentication methods, such as face recognition, authentication time needs to be as short as possible, we believe our behavioral authentication method does not need to address this necessity as much. The reason is because our authentication method is done implicitly in the background and does not need to interrupt users' activities. Users may not necessarily notice the extra seconds taken to implicitly authenticate them.

These results demonstrate that DNNs evolved through GA can successfully learn the behavioral patterns contained within traces of location history of smartphone users. Mobile devices can be confident that their owners are no longer the ones who accompany them when the DNN outputs a reject behavioral authentication signal. In cases where false rejects do happen, the owners can simply authenticate through explicit authentication methods (such as passwords or face recognition).

5 FINAL REMARK

Our research is still in its early phase and initial results demonstrated that deep neuroevolution models where DNNs are evolved by GA can be a good approach to implicitly authenticate users through their behaviors. Although our results showed that this approach is still not accurate enough to fully replace traditional explicit authentication methods, we believe it can be a suitable alternative for authenticating additional behavioral factors after the initial unlock through explicit authentications.

We also see a potential scenario for our approach to be implemented for smoother mobile payments through smartphones without requiring additional interaction, (such as providing PIN codes) from users. While ideally the learning process for this kind of application should be done continuously on the mobile device itself, it would rapidly drain the battery power. As such, the learning process itself could be limited during the time when the phone is connected to a power outlet. We are currently investigating this feasibility.

To also further validate our approach, we are planning to conduct more experiments using whole users' information from the dataset. Parameter optimization to achieve better FAR and FRR is also being planned.

REFERENCES

Banyal, R. K., Jain, P., & Jain, V. K. (2013, September). Multi-factor authentication framework for cloud computing. In 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation (pp. 105-110). IEEE.

Feng, H., Fawaz, K., & Shin, K. G. (2017, October). Continuous authentication for voice assistants. In Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (pp. 343-355).

Goldberg, D. E. (2006). Genetic algorithms. Pearson Education India.

Ghogare, S. D., Jadhav, S. P., Chadha, A. R., & Patil, H. C. (2012). Location based authentication: A new approach towards providing security. International Journal of Scientific and Research Publications, 2(4), 1-5.

Hsieh, W. B., & Leu, J. S. (2011, July). Design of a time and location based One-Time Password authentication scheme. In 2011 7th International Wireless Communications and Mobile Computing Conference (pp. 201-206). IEEE.

Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., ... & Hodjat, B. (2019). Evolving deep neural networks. In Artificial Intelligence in the Age of Neural Networks and Brain Computing (pp. 293-312). Academic Press.

Ogbanufe, O., & Kim, D. J. (2018). Comparing fingerprint-based biometrics authentication versus traditional authentication methods for e-payment. Decision Support Systems, 106, 1-14.

Sitová, Z., Šeděnka, J., Yang, Q., Peng, G., Zhou, G., Gasti, P., & Balagani, K. S. (2015). HMOG: New behavioral biometric features for continuous authentication of smartphone users. IEEE Transactions on Information Forensics and Security, 11(5), 877-892.

Such, F. P., Madhavan, V., Conti, E., Lehman, J., Stanley, K. O., & Clune, J. (2017). Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. arXiv preprint arXiv:1712.06567.

Szegedy, C., Toshev, A., & Erhan, D. (2013). Deep neural networks for object detection. In Advances in neural information processing systems (pp. 2553-2561).

Velásquez, I., Caro, A., & Rodríguez, A. (2018). Authentication schemes and methods: A systematic literature review. Information and Software Technology, 94, 30-37.

Zhang, F., Kondoro, A., & Muftic, S. (2012, June). Location-based authentication and authorization using smart phones. In 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (pp. 1285-1292). IEEE.

Zviran, M., & Erlich, Z. (2006). Identification and authentication: technology and implementation issues. Communications of the Association for Information Systems, 17(1), 4.