BlockFITS: A Federated Data Augmentation Modelling for Blockchain Based IoVT Systems

Bhrigu Kansra[1][0000−0002−0288−7536], Harshita Diddee[2][0000−0002−0852−7371], Ashish Khanna[1], Deepak Gupta[1][0000−0002−3019−7161], and Joel J. P. C. Rodrigues[3][0000−0001−8657−3800]

[1] Maharaja Agrasen Institute of Technology bhrigukansra98@gmail.com, deepakgupta@mait.ac.in, ashishkhanna@mait.ac.in
[2] Bharati Vidyapeeth's College of Engineering harshita.bvcoend@bvp.edu.in
[3] Federal University of Piauí joeljr@ieee.org

Abstract. In Intelligent Transport Systems (ITS), the collection of diverse data is a major practical roadblock; Not only can their data be personally identifiable i.e private, but the lack of incentive for entities to participate in any kind of collaborative training is also severely limited due to the added computational expense of training collaborative models locally. In this paper, we propose BlockFITS: A Vehicle-to-BlockChain-to-Vehicle (V2B2V) federated learning enabled model training paradigm for ITS entities. In addition to which we propose a data augmentation scheme that operates with cooperative training to generate an incentive for entity participation. The immutability and decentralized features of the Blockchain system leverages the federated-like averaging of synthetically generated data samples that generate incentives for the participation of entities in such a training setup. BlockFITS can be practically deployed in future ITS systems to improve the autonomous driving system, pedestrian safety, and vehicular object detection or more due to its model-constraint-free characteristics which provide access to a synthetic and global data whilst maintaining data privacy.

Keywords: blockchain · intelligent transport systems · internet of vehicular things · federated learning · data augmentation

1 Introduction

Given the paucity of covariate data in IoVT systems, developing a trustworthy and decentralised technique to augment local data can be massively useful in enhancing the robustness of these systems. The deep learning models used in such systems are data hungry and benefit from the provision of data obtained from different sources. It is also generally observed that the performance of deep learning models increases exponentially, if more representative data is provided at the training stage. The data used in IoVT systems is generally acquired using mounted cameras, sensors, embedded global positioning systems(GPS), Light

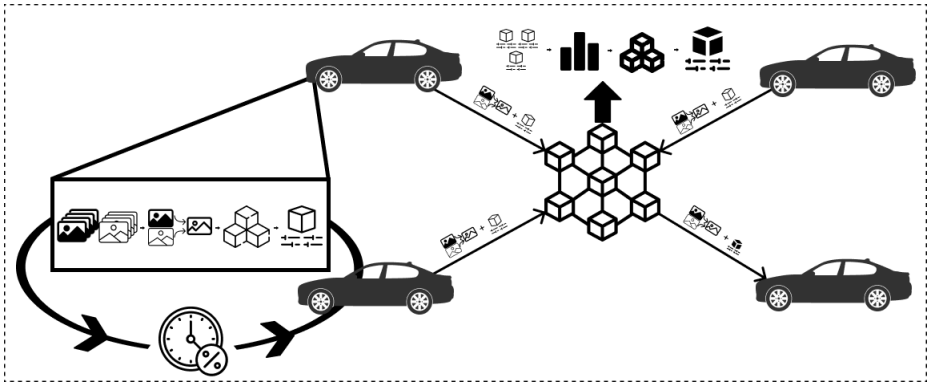

Fig. 1. Distributed ITS entities connected to BlockFITS IoVT system

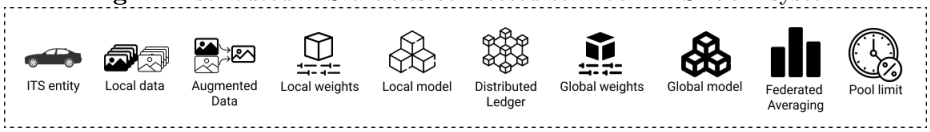

Fig. 2. Notation of the illustrated BlockFITS Model

Detection & Ranging (LIDAR) sensors, accelerometers, optical sensors, etc. As with other domains of deep learning, augmentation of data derived through these sources can improve the performance of such systems significantly, especially since, a large of amount of data being used in these systems is real-time and local context dependent. These sources can collect important useful data in real time which can be used to train locally deployed models with applications in autonomous driving, vehicles detection, pedestrian detection, traffic signs recognition, lane detection etc [20]. Given the dependence of IoVT systems on such dynamically changing data, It seems natural to seek a method which can effectively use locally collected data points via onboard sensors, to train robust deep learning models for IoVT use cases. Generally in the given , The data collected via local sensors is pre-processed, augmented, and then an edge deployed model is trained on the ITS entity itself. This limits the utilization of data collated by the entity since the neighbouring entities obviously do not have access to the data collected by entities individually. If this data were to be shared with a central authority, the possibility of data breaches and malicious manipulation endanger the safe and efficient storage of such data [14]. Hence, The need to keep data private as well as shareable while not depending on a single central organisation, calls for the use of a distributed-ledger enabled system which can alleviate the problem of having a single point of failure/control, as well as provide incentives to the participatory entities using an ledger-enabled application such as Blockchain. Additionally, the need to train a set of locally deployed models in a privacy preserving manner encourages the use of a decentralized and anonymized method of collaborative model training - Federated Learning. Lastly, Under collaborative

and distributed learning systems - the lack of a proper incentive provision may discourage entities from participating in cooperative learning. To counter this, several incentive strategies have been explored, even in the context of ITS Systems; The additional advantage of using distributed ledgers ensure that workers are voluntarily joining the network and no single company overhead costs are involved in it hence making it a decentralised system. In accordance with the following principle motivations, we define BlockFITS with the aim to provide the following contributions:

- To tackle the paucity of quality and quality data in IoVT systems, we propose a privacy-preserving data sharing technique that uses synthetically augmented to enhance the quality of the locally acquired data with ITS entities.
- A federated learning enabled distributed ledger mechanism which allows a decentralized network of all ITS entities to share weights of their locally trained model to allow entities on the ledger to leverage the learning of the local models hosted by other entities.
- To encourage participation in such a collaborative method of learning, we specify an incentive mechanism that defines the utility function of each entity with its local accuracy; Based on its performance, the entity is given a proportionate "quota" of augmentation to compute - thus allowing relatively better performing entities to reduce their computational strain due to the added data augmentation step.

2 Literature Review

2.1 Internet of Vehicular Things (IoVT)

Since the last few decades Intelligent Transport Systems(ITS) have been an important nexus for the next generation automated world. The availability of cost effective as well as reliable sensors networks gave a boost to Internet of Things (IoT) [16].
Internet of Vehicular Things (IoVT) is an application of the Internet of Things (IoT which focuses on making human intervention to a minimal and seamless extent whilst making vehicles more intelligent and connected. It does so by exchanging information, increasing reliability and efficiency and maintaining safety of the end user [1].
While exploring the current research about security in IoT systems it is observed that the major focus is in evolving techniques to maintain a physical level safety of the end user but we want to focus on less investigated area of safety in domain of maintaining end users privacy and keep the user data secure and still using it to create a more intelligent system [5] [6] [13].

2.2 Blockchain

Blockchain is collection of congruent blocks with each block containing a cryptographic hash value of the last block. By design, a Blockchain is impervious to

Ref.	Objective	Privacy	Incentive	Data
[9]	Vehicular communication management.	Conditional	Mission Based Reward	Broadcasted between vehicles
[10]	Privacy preserving carpooling	Conditional	None	Encrypted and sent to a central server
[26]	Vehicular data sharing	Complete	Data Coins Rewards	Shared and stored on a private BlockChain
This	Vehicular Model and Augmented data sharing	Complete	Accuracy	Store locally and Share only Augmented Data

Table 1. A comparison between BlockFITS and other blockchain-based IoVT systems

change of data once written on the block. Blockchain is widely used as a distributed ledger system managed via a P2P (Peer2Peer) network of nodes while adhering to a predefined protocol for connectivity between blocks.[17] In recent work in the field of Blockchain based Vehicualar systems there have been several consensus algorithms to achieve different objectives. In [9] [25] the authors propose a vehicle communication management system which uses practical Byzantine Fault Tolerance(pBFT) consensus. The authors of [10] proposes a privacy preserving car-pooling system which a permissioned and decentralized network based on PoS consensus. The authors in [26] propose a Vehicular ad-hoc network (VANET) which leverages Consortium and decentralized network with main objective to share data and storage between vehicle using a pBFT consensus.

2.3 Federated Learning

Modern ITS entities have potential to access abundance of data for deep learning models which can drastically improve the user interaction and experience. However, this widely accessible data is often user privacy sensitive, and available in billions of independent entities. Federated Learning advocates on keeping the private user data on distributed entities itself, otherwise logged into a centralised data center, and learns a shared model by aggregating the locally determined updates[12]. Most of the recent publications [23], [3], [11], [7] about application of federated learning in vehicular systems are based on aggregation of collected local weights on a centralised server; This is a partially centralised approach and solely focuses on securing the user data but lacks in providing an incentive to the participating entities which makes it to hard to predict if the contributions of the participating nodes would be sustained .

2.4 Data Augmentation:

The synthetic generation of data for several deep learning tasks has resulted in enhanced performance; [19] provides an overview which aims to validate the same in the context of time series use cases. The data collected from mounted camera on an ITS entity provides only a small set of local information ergo leads to lack of accurate judgment at global scenario. This calls for a way of

introducing more data for training the system. Data Augmentation is one of the technique which can used to do the same without the need of burdening the entity to generate more data. Data augmentation of spatiotemporal data collected in ITS entities can be augmented using by various methods like image stitching techniques proposed by [18] or using comotion algorithm to land-points from adjacent cameras, and construct a homography matrix constructed as done in [21].

3 System model

In this section, we first delineate the definitions and notations that will be used to describe the working of the proposed, V2B2V IoVT model. Following which we explain the detailed working of the system along with the necessary implementation details. The system working is broadly divided into 2 broad sections - The first one is specific to the on-device computation and model training on the ITS entity and the subsequent transfer of the augmented data and the local weights to a ledger. The second section follows to define the pooling and aggregation of the locally derived data on a randomly selected ledger node and then its subsequent distribution on the participating entities using the specified incentive mechanism.

3.1 Requisite definitions

In this section we attempt to explicate the frequently used terms and notation used in our system model description:

- **Local Data or *LD***: This defines the set of real-time data samples collected by the ITS entity. This data may be acquired through any of the sources specified in Section I.
- **Augmented data or *LD****: This defines the set of synthetically augmented set of data samples that will be generated by the ITS Entity.
- **Local Model**: Leveraging the collaborative model training traits of federated learning, the ITS entity may host an edge deployed local model; This model may be a secondary utility model for the ITS entity, such as traffic flow prediction, travel time estimation, multiple trajectory prediction and congestion control prediction model for a wireless daisy chain network.
- **Local weights** : Local weights are the learnable parameters of the on- device deep learning model; This model is trained on LD and LD*.
- **Pool limit or** α : Pool limit is the minimum number of augmented data samples that the entity must generate before it can relay this set to the ledger. The motivation of introducing this term derives from the fact that due to computational constraints on the ITS entity and the practical constraints on the constant relaying of information across to a ledger due to network constraints - the synthetic data at each ITS entity will not be relayed to a common ledger until it acquires a certain minimum number of data samples i.e the pool limit number of samples.

- **Pooling Frequency or λ** : This refers to the frequency at which augmented data, LD*, is pooled at a commonly decided ledger node.
- **Image Stitching**: Image stitching refers to the concatenation of images, similar to the generation of panoramic image generation, using local data samples.This type of augmentation is carried out to specifically eliminate the identification and consequent reporting of duplicate objects in the same image. [18]
- **Idle Time:** Since the assignment of computational resources during the operation of the ITS Entity for the task of data augmentation may not be the optimal choice bearing in mind the low priority of the data augmentation in comparison to the primary tasks of the ITS entity (including resource assignment to any deep learning model hosted locally that may be crucial to navigation, destination or traffic prediction model) - The augmentation for the proposed model will be carried out when the entity is not involved in any other resource-intensive activity - mostly when the entity is non-operational for a significantly long period of time such as when it is parked or being charged. [2]

3.2 Workflow of the System:

Generation of LD* from LD: After the acquisition of local data samples - LD* is populated as follows: Augmentation manipulations such as brightness enhancements, horizontal and vertical shifts, shadow casting, flipping and subsampling are already used to enhance the quantity of locally available dataset. In addition to these techniques, we use Image Stitching to generate more synthetic images. Note that this augmentation will take place during the idle time only [15].

Training of the local model: Data samples from both, LD and LD* are used to train the local model on device. Note that the decision of what proportion of augmented samples to use in the training of the local model is a hyper parameter i.e the training data split [between locally acquired data samples and synthetically generated/received] data samples may be modified experimentally to achieve the split that achieves the maximum local accuracy.

Transfer of weights (and LD*, If applicable) to the ledger Due to their inherent characteristic of hosting immutable data which whilst being immune to malicious manipulation, still supports the transparent viewing of the artificially augmented data that is generated by each node - ledgers provide an ideal platform for the established setup. The weights of the local model are relayed to the ledger node after running a predefined set of epochs on the local models (The number of epochs being run on the local model is also a hyper parameter). This step will be accompanied by the relaying of LD* as soon as the number of samples in LD $> \lambda$. Note that the relayed LD* may be directly transferred to the pooling node [4].

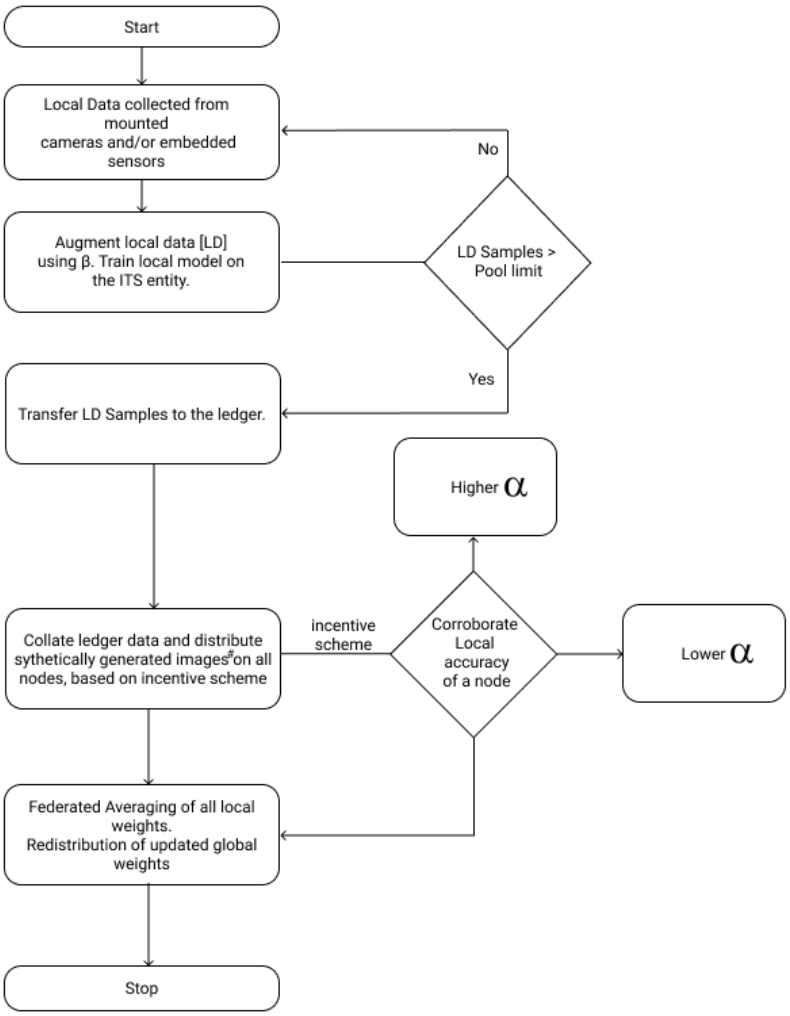

Fig. 3. Flowchart depicting the cycle of BlockFITS

Aggregation and Redistribution of Local weights Since there is no central supervising entity in such a system, the aggregation of weights hosted at all ledger nodes is carried out on a randomly selected node in the ledger that is referred to as the pooling node. This aggregation, done in accordance with the FedAVG algorithm [12], is done without any additional pre-processing, this aggregation may be carried out using a smart contract as well. To test the performance of these aggregated and averaged weights, the aggregating node, relays these aggregated nodes to the ITS entity it is connected to, which computes the accuracy of the model on its local test set. If the loss assumes a converging trend, the ledger distributes these aggregated weights to all the other entities on the network.

Establishing Incentive based Distribution of Pooled data to the entities Similar to the fashion of redistribution of local weights, the LD* hosted at each node are pooled at a randomly selected node. This pool of synthetically generated samples are shuffled and relayed to all the nodes on the ledger, which are finally relayed to the ITS entities. This redistribution of samples is governed by an incentive scheme which operates as follows: The scheme is inspired majorly by the Individual Profit Sharing Scheme proposed by [22], which maps the utility function of the scheme as the local accuracy of each ITS Entity (equation 1):

$$U_i(t) = Vi \times \alpha 0 \leq Vi \leq 1\alpha \, \epsilon \, R \tag{1}$$

In essence, Higher the local accuracy of the system, higher is the utility of the contributing entity. Using a smart contract, the local accuracy of all the ITS entities can be related to their utilities. The utility of the entity decides how much of its resources would it be asked to sacrifice during the generation of LD* in the next iteration i.e A high utility entity will receive an updated α or pooling limit - which will allow it to relay its LD* without having augmented a LD* commensurate to its poorer performing peers on the network. This ensures that high performing entities enjoy the access to the collaborative training as well as the diverse pool of synthetic data without compromising on their computational resources consistently, which will be necessary if they are too augment LD* from LD [24] [8].

4 Conclusion

In this paper, we present BlockFITS, A federated learning based, Blockchain enabled data augmentation system that allows participating ITS entities to leverage a rich set of locally gathered yet artificially pruned data to collaboratively train their deep learning models. Unlike most existing systems - BlockFITS establishes an incentive mechanism focused around the primary factor that affects local model accuracy i.e the data that an entity uses. Moreover, it attempts to give consideration to the practical idea that all entities participating in the collaborative training do not have massive resource capabilities and hence, must be advantaged, if their contributions benefit the network at large.

5 Future Work

This work must be further analyzed to identify and mitigate the caveats that arise from hosting the synthetic data on a Blockchain node. Additionally, the incentive mechanism may be enhanced to include other data quality driven metrics such as one that accounts for the ITS entity that provides the set of data with the highest sample diversity or the entity that provides data that is representative of some temporal trend.

References

1. Chavhan, S., Gupta, D., Garg, S., Khanna, A., Choi, B.J., Hossain, M.S.: Privacy and security management in intelligent transportation system. IEEE Access **8**, 148677–148688 (2020)
2. Chmiel, W., Dańda, J., Dziech, A., Ernst, S., Kadłuczka, P., Mikrut, Z., Pawlik, P., Szwed, P., Wojnicki, I.: Insigma: an intelligent transportation system for urban mobility enhancement. Multimedia Tools and Applications **75**(17), 10529–10560 (Sep 2016). https://doi.org/10.1007/s11042-016-3367-5, https://doi.org/10.1007/s11042-016-3367-5
3. Du, Z., Wu, C., Yoshinaga, T., Yau, K.A., Ji, Y., Li, J.: Federated learning for vehicular internet of things: Recent advances and open issues. IEEE Open Journal of the Computer Society **1**, 45–61 (2020)
4. Goel, A., Agarwal, A., Vatsa, M., Singh, R., Ratha, N.: Deepring: Protecting deep neural network with blockchain. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2821–2828 (2019)
5. Janušová, L., Čičmancová, S.: Improving safety of transportation by using intelligent transport systems. Procedia Engineering **134**, 14–22 (12 2016). https://doi.org/10.1016/j.proeng.2016.01.031
6. Janušová, L., Čičmancová, S.: Improving safety of transportation by using intelligent transport systems. Procedia Engineering **134**, 14 – 22 (2016). https://doi.org/https://doi.org/10.1016/j.proeng.2016.01.031, http://www.sciencedirect.com/science/article/pii/S1877705816000345, tRANSBALTICA 2015: PROCEEDINGS OF THE 9th INTERNATIONAL SCIENTIFIC CONFERENCE. May 7–8, 2015. Vilnius Gediminas Technical University, Vilnius, Lithuania.
7. Kang, J., Xiong, Z., Niyato, D., Xie, S., Zhang, J.: Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. IEEE Internet of Things Journal **6**(6), 10700–10714 (2019)
8. Kang, J., Xiong, Z., Niyato, D., Ye, D., Kim, D.I., Zhao, J.: Toward secure blockchain-enabled internet of vehicles: Optimizing consensus management using reputation and contract theory. IEEE Transactions on Vehicular Technology **68**(3), 2906–2920 (2019)
9. Li, L., Liu, J., Cheng, L., Qiu, S., Wang, W., Zhang, X., Zhang, Z.: Creditcoin: A privacy-preserving blockchain-based incentive announcement network for communications of smart vehicles. IEEE Transactions on Intelligent Transportation Systems **19**(7), 2204–2220 (2018)
10. Li, M., Zhu, L., Lin, X.: Efficient and privacy-preserving carpooling using blockchain-assisted vehicular fog computing. IEEE Internet of Things Journal **6**(3), 4573–4584 (2019)

11. Lu, Y., Huang, X., Dai, Y., Maharjan, S., Zhang, Y.: Federated learning for data privacy preservation in vehicular cyber-physical systems. IEEE Network **34**(3), 50–56 (2020)
12. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data (2016)
13. Regan, M.A., Oxley, J.A., Godley, S.T., Tingvall, C.: Intelligent transport systems: safety and human factors issues. No. 01/01 (2001)
14. Sakiz, F., Sen, S.: A survey of attacks and detection mechanisms on intelligent transportation systems: Vanets and iov. Ad Hoc Networks **61** (03 2017). https://doi.org/10.1016/j.adhoc.2017.03.006
15. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**(1), 60 (Jul 2019). https://doi.org/10.1186/s40537-019-0197-0, https://doi.org/10.1186/s40537-019-0197-0
16. Sładkowski, A., Pamuła, W.: Intelligent transportation systems-problems and perspectives, vol. 303. Springer (2016)
17. Swan, M.: Blockchain: Blueprint for a new economy. " O'Reilly Media, Inc." (2015)
18. Tsao, P., Ik, T.U., Chen, G.W., Peng, W.C.: Stitching aerial images for vehicle positioning and tracking. pp. 616–623 (11 2018). https://doi.org/10.1109/ICDMW.2018.00096
19. Wen, Q., Sun, L., Song, X., Gao, J., Wang, X., Xu, H.: Time series data augmentation for deep learning: A survey (2020)
20. Wu, W., Yang, Z., Li, K.: Internet of Vehicles and applications, pp. 299–317 (12 2016). https://doi.org/10.1016/B978-0-12-805395-9.00016-2
21. Wu, Y., Liu, C., Lan, S., Yang, M.: 3d road scene monitoring based on real-time panorama. Journal of Applied Mathematics **2014**, 403126 (Aug 2014). https://doi.org/10.1155/2014/403126, https://doi.org/10.1155/2014/403126
22. Yang, G., He, S., Shi, Z., Chen, J.: Promoting cooperation by the social incentive mechanism in mobile crowdsensing. IEEE Communications Magazine **55**(3), 86–92 (2017)
23. Ye, D., Yu, R., Pan, M., Han, Z.: Federated learning in vehicular edge computing: A selective model aggregation approach. IEEE Access **8**, 23920–23935 (2020)
24. Yeow, K., Gani, A., Ahmad, R.W., Rodrigues, J.J.P.C., Ko, K.: Decentralized consensus for edge-centric internet of things: A review, taxonomy, and research issues. IEEE Access **6**, 1513–1524 (2018)
25. Yuan, Y., Wang, F.: Towards blockchain-based intelligent transportation systems. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). pp. 2663–2668 (2016)
26. Zhang, X., Chen, X.: Data security sharing and storage based on a consortium blockchain in a vehicular adhoc network. IEEE Access **PP**, 1–1 (01 2019). https://doi.org/10.1109/ACCESS.2018.2890736