
AN ON-DEVICE FEDERATED LEARNING APPROACH FOR COOPERATIVE ANOMALY DETECTION

A PREPRINT

Rei Ito

Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Japan
rei@arc.ics.keio.ac.jp

Mineto Tsukada

Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Japan
tsukada@arc.ics.keio.ac.jp

Hiroki Matsutani

Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Japan
matutani@arc.ics.keio.ac.jp

February 28, 2020

ABSTRACT

Most edge AI focuses on prediction tasks on resource-limited edge devices, while the training is done at server machines, so retraining a model on the edge devices to reflect environmental changes is a complicated task. To follow such a concept drift, a neural-network based on-device learning approach is recently proposed, so that edge devices train incoming data at runtime to update their model. In this case, since a training is done at distributed edge devices, the issue is that only a limited amount of training data can be used for each edge device. To address this issue, one approach is a cooperative learning or federated learning, where edge devices exchange their trained results and update their model by using those collected from the other devices. In this paper, as an on-device learning algorithm, we focus on OS-ELM (Online Sequential Extreme Learning Machine) and combine it with Autoencoder for anomaly detection. We extend it for an on-device federated learning so that edge devices exchange their trained results and update their model by using those collected from the other edge devices. Experimental results using a driving dataset of cars demonstrate that the proposed on-device federated learning can produce more accurate model by combining trained results from multiple edge devices compared to a single model.

Keywords On-device learning · Federated learning · OS-ELM

1 Introduction

Most edge AI focuses on prediction tasks on resource-limited edge devices assuming that their prediction model has been trained at server machines beforehand. In this case, retraining or customizing a model for a given edge device later to reflect environmental changes is a complicated task, because the server machine needs to collect training data from the edge device, train a new model based on the collected data, and then deliver the new model to the edge device.

To realize training tasks at resource-limited edge devices, a neural-network based on-device learning approach is proposed in [1, 2], so that edge devices can train or correct their model based on incoming data at runtime. Its low-cost hardware implementation is also introduced in [2]. In this case, since a training is done independently at distributed edge devices, the issue is that only a limited amount of training data can be used for each edge device. To address this issue, one approach is a cooperative model update, where edge devices exchange their trained results and update their model using those collected from the other devices. Please note that edge devices share an intermediate form of their weight parameters instead of raw data, which is sometimes privacy sensitive.

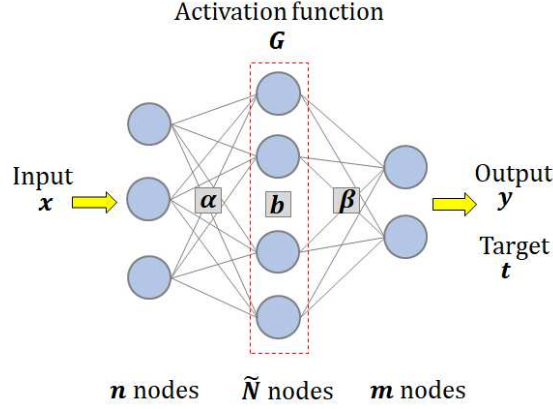


Figure 1: Single hidden-layer feedforward network (SLFN).

In this paper, we use the on-device learning algorithm [1, 2], which is based on OS-ELM (Online Sequential Extreme Learning Machine) [3] for sequential training for single hidden-layer neural networks and combined with Autoencoder [4] for unsupervised anomaly detection. It is then extended for the on-device federated learning so that edge devices exchange their trained results and update their model using those collected from the other edge devices. To do this, we employ Elastic ELM (E²LM) [5] that enables distributed training of neural networks, where intermediate training results are computed by multiple machines separately and then a final model is produced by combining these intermediate results. It is applied to the OS-ELM based on-device learning approach [2] for the on-device federated learning. Experimental results using a driving dataset of cars demonstrate that the proposed on-device federated learning can produce more accurate model by combining trained results from multiple edge devices compared to a single model.

The rest of this paper is organized as follows. Section 2 briefly overviews baseline technologies including ELM, E²LM, and OS-ELM. Section 3 proposes a model exchange and update algorithm for the on-device federated learning. Section 4 evaluates the proposed approach using a car driving dataset. Section 5 concludes this paper.

2 Preliminaries

This section briefly introduces (1) ELM (Extreme Learning Machine), (2) E²LM (Elastic Extreme Learning Machine), (3) OS-ELM (Online Sequential Extreme Learning Machine), and (4) Autoencoder.

2.1 ELM

ELM [6] is a batch training algorithm for single hidden-layer feedforward networks (SLFNs). As shown in Figure 1, the network consists of input layer, hidden layer, and output layer. The numbers of their nodes are denoted as n , \tilde{N} , and m , respectively. Assuming an n -dimensional input chunk $x \in \mathbf{R}^{k \times n}$ of batch size k is given, an m -dimensional output chunk $y \in \mathbf{R}^{k \times m}$ is computed as follows.

$$y = G(x \cdot \alpha + b)\beta, \quad (1)$$

where G is an activation function, $\alpha \in \mathbf{R}^{n \times \tilde{N}}$ is an input weight matrix between the input and hidden layers, $\beta \in \mathbf{R}^{\tilde{N} \times m}$ is an output weight matrix between the hidden and output layers, and $b \in \mathbf{R}^{\tilde{N}}$ is a bias vector of the hidden layer.

If an SLFN model can approximate m -dimensional target chunk (i.e., teacher data) $t \in \mathbf{R}^{k \times m}$ with zero error, the following equation is satisfied.

$$G(x \cdot \alpha + b)\beta = t \quad (2)$$

Here, the hidden layer matrix is defined as $H \equiv G(x \cdot \alpha + b)$. The optimal output weight matrix $\hat{\beta}$ is computed as follows.

$$\hat{\beta} = H^\dagger t, \quad (3)$$

where H^\dagger is a pseudo inverse matrix of H , which can be computed with matrix decomposition algorithms, such as SVD (Singular Value Decomposition) and QRD (QR Decomposition).

In ELM algorithm, the input weight matrix α is initialized with random values and not changed thereafter. The optimization is thus performed only for the output weight matrix β , and so it can reduce the computation cost compared with backpropagation based neural networks that optimize both α and β . In addition, the training algorithm of ELM is not iterative; it analytically computes the optimal weight matrix β for a given input chunk in a one-shot manner, as shown in Equation 3. It can always obtain a global optimal solution for β , unlike a typical gradient descent method, which sometimes converges to a local optimal solution.

Please note that ELM is one of batch training algorithms for SLFNs, which means that the model is trained by using all the training data available at that time. In this case, we need to retrain the whole dataset in order to update the model for newly-arrived training data. This issue is addressed by E²LM and OS-ELM.

2.2 E²LM

E²LM [5] is an extended algorithm of ELM for enabling the distributed training of SLFNs. That is, intermediate training results are computed by multiple machines separately, and then a final model is produced by combining these intermediate results.

In Equation 3, assuming that $\text{rank } H = \tilde{N}$ and $H^T H$ is nonsingular, the pseudo inverse matrix H^\dagger is decomposed as follows.

$$H^\dagger = (H^T H)^{-1} H^T \quad (4)$$

The optimal output weight matrix β in Equation 3 can be computed as follows.

$$\hat{\beta} = (H^T H)^{-1} H^T t \quad (5)$$

Assuming the intermediate results are defined as $U = H^T H$ and $V = H^T t$, the above equation is simplified as follows.

$$\hat{\beta} = U^{-1} V \quad (6)$$

Here, the hidden layer matrix and target chunk (i.e., teacher data) for newly-arrived training dataset Δx are denoted as ΔH and Δt , respectively. The intermediate results for Δx are denoted as $\Delta U = \Delta H^T \Delta H$ and $\Delta V = \Delta H^T \Delta t$.

Similarly, the hidden layer matrix and target chunk for updated training dataset $x' = x + \Delta x$ are denoted as H' and t' , respectively. The intermediate results for x' are denoted as $U' = H'^T H'$ and $V' = H'^T t'$. Then, U' and V' can be computed as follows.

$$\begin{aligned} U' &= H'^T H' = \begin{bmatrix} H \\ \Delta H \end{bmatrix}^T \begin{bmatrix} H \\ \Delta H \end{bmatrix} = H^T H + \Delta H^T \Delta H \\ V' &= H'^T t' = \begin{bmatrix} H \\ \Delta H \end{bmatrix}^T \begin{bmatrix} t \\ \Delta t \end{bmatrix} = H^T t + \Delta H^T \Delta t \end{aligned} \quad (7)$$

As a result, Equation 7 can be simplified as follows.

$$\begin{aligned} U' &= U + \Delta U \\ V' &= V + \Delta V \end{aligned} \quad (8)$$

In summary, E²LM algorithm updates a model in the following steps:

1. Compute U and V for the whole training dataset x ,
2. Compute ΔU and ΔV for newly-arrived training dataset Δx ,
3. Compute U' and V' for updated training dataset x' using Equation 8, and
4. Compute the new output weight matrix β using Equation 6.

Please note that we can compute a pair of U and V and a pair of ΔU and ΔV separately. Then, we can produce U' and V' by simply adding them using Equation 8. Similar to the addition of x and Δx , subtraction and replacement operations for x are also supported.

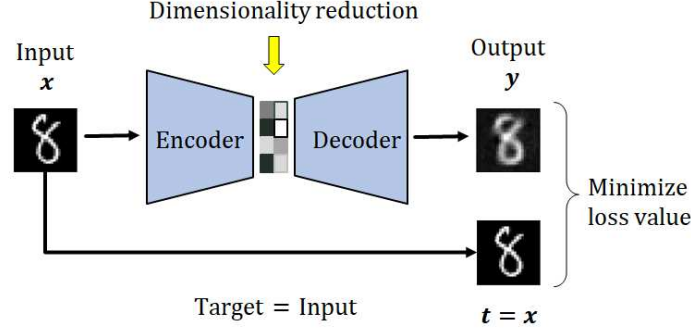


Figure 2: Autoencoder.

2.3 OS-ELM

OS-ELM [3] is an online sequential version of ELM, which can update the model sequentially using an arbitrary batch size.

Assuming that the i -th training chunk $\{x_i \in \mathbf{R}^{k_i \times n}, t_i \in \mathbf{R}^{k_i \times m}\}$ of batch size k_i is given, we need to compute the output weight matrix β that can minimize the following error.

$$\left\| \begin{bmatrix} H_0 \\ \vdots \\ H_i \end{bmatrix} \beta_i - \begin{bmatrix} t_0 \\ \vdots \\ t_i \end{bmatrix} \right\|, \quad (9)$$

where H_i is defined as $H_i \equiv G(x_i \cdot \alpha + b)$. Assuming $K_i \equiv \begin{bmatrix} H_0 \\ \vdots \\ H_i \end{bmatrix}^T \begin{bmatrix} H_0 \\ \vdots \\ H_i \end{bmatrix}$ ($i \geq 0$), the optimal output weight matrix is computed as follows.

$$\begin{aligned} \beta_i &= \beta_{i-1} + K_i^{-1} H_i^T (t_i - H_i \beta_{i-1}) \\ K_i &= K_{i-1} + H_i^T H_i \end{aligned} \quad (10)$$

Assuming $P_i \equiv K_i^{-1}$, we can derive the following equation from Equation 10.

$$\begin{aligned} P_i &= P_{i-1} - P_{i-1} H_i^T (I + H_i P_{i-1} H_i^T)^{-1} H_i P_{i-1} \\ \beta_i &= \beta_{i-1} + P_i H_i^T (t_i - H_i \beta_{i-1}) \end{aligned} \quad (11)$$

In particular, the initial values P_0 and β_0 are precomputed as follows.

$$\begin{aligned} P_0 &= (H_0^T H_0)^{-1} \\ \beta_0 &= P_0 H_0^T t_0 \end{aligned} \quad (12)$$

As shown in Equation 11, the output weight matrix β_i and its intermediate result P_i are computed from the previous training results β_{i-1} and P_{i-1} . Thus, OS-ELM can sequentially update the model with a newly-arrived target chunk in a one-shot manner; thus there is no need to retrain with all the past data unlike ELM.

In this approach, the major bottleneck is the pseudo inverse operation $(I + H_i P_{i-1} H_i^T)^{-1}$. As in [1, 2], the batch size k is fixed at one in this paper so that the pseudo inverse operation of $k \times k$ matrix for the sequential training is replaced with a simple reciprocal operation; thus we can eliminate the SVD or QRD computation.

2.4 Autoencoder

Autoencoder [4] is a type of neural networks developed for dimensionality reduction, as shown in Figure 2. In this paper, OS-ELM is combined with Autoencoder for unsupervised anomaly detection. In this case, the numbers of input and output layer nodes are the same (i.e., $n = m$), while the number of hidden layer nodes is set to less than that of

input layer nodes (i.e., $\tilde{N} < n$). In Autoencoder, an input chunk is converted into a well-characterized dimensionally reduced form at the hidden layer. The process for the dimensionality reduction is denoted as “encoder”, and that for decompressing the reduced form is denoted as “decoder”. In OS-ELM, the encoding result for an input chunk x is obtained as $H = G(x \cdot \alpha + b)$, and the decoding result for the hidden layer matrix H is obtained as $y = H \cdot \beta$.

In the training phase, an input chunk x is used as a target chunk t . That is, the output weight matrix β is trained so that an input data is reconstructed as correctly as possible by Autoencoder. Assuming that the model is trained with a specific input pattern, the difference between the input data and reconstructed data (denoted as loss value) becomes large when the input data is far from the trained pattern. Please note that Autoencoder does not require any labeled training data for the training phase; so it is used for unsupervised anomaly detection. In this case, incoming data with high loss value should be automatically rejected to be trained for stable anomaly detection.

3 On-Device Federated Learning

As an on-device learning algorithm, in this paper, we employ a combination of OS-ELM and Autoencoder for online sequential learning and unsupervised anomaly detection [1, 2]. It is further optimized by setting the batch size k to one, in order to eliminate the pseudo inverse operation of $k \times k$ matrix for the sequential training. A low-cost forgetting mechanism that does not require the pseudo inverse operation is also proposed in [2].

In practice, anomaly patterns should be accurately detected from multiple normal patterns. To improve the accuracy of anomaly detection in such cases, we employ multiple on-device learning instances, each of which is specialized for each normal pattern as proposed in [7]. Also, the number of the on-device learning instances can be dynamically tuned at runtime as proposed in [7].

In this paper, the on-device learning algorithm is extended for the on-device federated learning by applying the E²LM approach to the OS-ELM based sequential training. In this case, edge devices share their intermediate trained results and update their model using those collected from the other edge devices. In this section, OS-ELM algorithm is analyzed so that the E²LM approach is applied to OS-ELM for enabling the cooperative model update. The proposed on-device federated learning approach is then illustrated in detail.

3.1 Modifications for OS-ELM

Here, we assume that edge devices exchange the intermediate results of their output weight matrix β (see Equation 6). These intermediate results are obtained by $U = H^T H$ and $V = H^T t$, based on E²LM algorithm. Please note that the original E²LM approach is designed for ELM, which assumes a batch training, not a sequential training. That is, U and V are computed by using the whole training dataset. On the other hand, our on-device learning algorithm relies on the OS-ELM based sequential training, in which the weight matrix is sequentially updated every time a new data comes. If the original E²LM approach is directly applied to our on-device learning algorithm, all the past dataset must be preserved in edge devices, which would be infeasible for resource-limited edge devices.

To address this issue, OS-ELM is analyzed as follows. In Equation 10, K_i is defined as $K_i \equiv \begin{bmatrix} H_0 \\ \vdots \\ H_i \end{bmatrix}^T \begin{bmatrix} H_0 \\ \vdots \\ H_i \end{bmatrix}$ ($i \geq 0$),

which indicates that it accumulates all the hidden layer matrixes that have been computed with up to the i -th training chunk. In this case, U and V of E²LM can be computed based on K_i and its inverse matrix P_i of OS-ELM as follows.

$$\begin{aligned} U_i &= K_i = P_i^{-1} \\ V_i &= U_i \beta_i \end{aligned} \quad (13)$$

where U_i and V_i are intermediate results for the i -th training chunk. U_i and V_i can be sequentially computed with the previous training results. To support the on-device federated learning, Equation 13 is newly inserted to the sequential training algorithm of OS-ELM, which will be introduced in Section 3.2.

3.2 Cooperative Mode Update Algorithm

Figure 3 illustrates a cooperative model update of the proposed on-device federated learning. It consists of the following three phases:

1. Sequential training on edge devices,
2. Exchanging their intermediate results via a cloud server, and

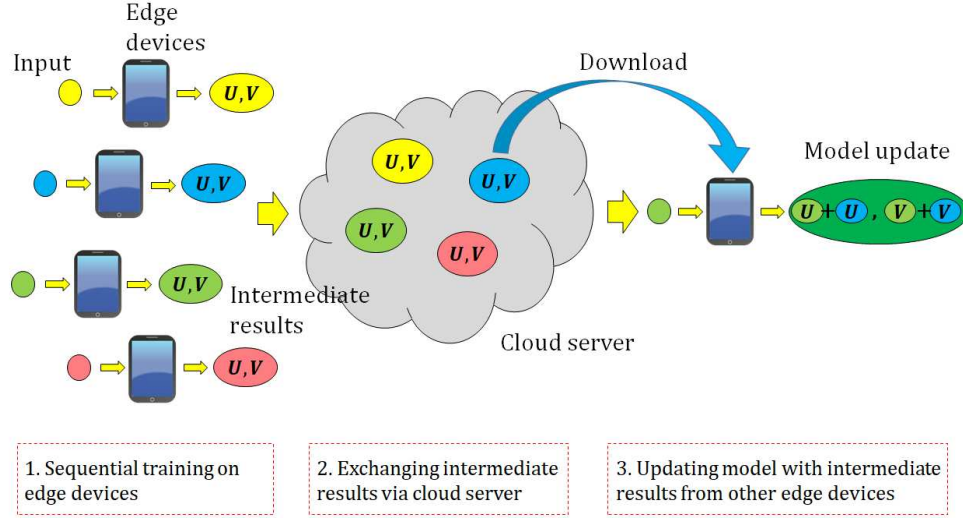


Figure 3: Overview of cooperative model update.

3. Updating their model with necessary intermediate results from the other edge devices.

First, edge devices independently execute a sequential training by using OS-ELM algorithm. They also compute the intermediate results U and V by Equation 13. Second, they upload their intermediate results to a cloud server. We assume that the input weight matrix α and the bias vector b are the same in the edge devices. They download necessary intermediate results from the cloud server if needed. They update their model based on their own intermediate results and those downloaded from the cloud server by Equation 8.

Figure 4 shows a flowchart of the proposed cooperative model update between two devices: Device-A and Device-B. Assuming Device-A sends its intermediate results and Device-B receives them for updating its model, their cooperative model update is performed by the following steps.

1. Device-A and Device-B sequentially train their own model by using OS-ELM algorithm. In other words, they compute the output weight matrix β and its intermediate result P by Equation 11.
2. Device-A computes the intermediate results U_A and V_A by Equation 13 to share them with other edge devices. Device-B also computes U_B and V_B . They upload these results to a cloud server.
3. Assuming Device-B demands the Device-A's trained results, it downloads U_A and V_A from the cloud server.
4. Device-B integrates their intermediate results by computing $U_B \leftarrow U_A + U_B$ and $V_B \leftarrow V_A + V_B$ by using E^2LM algorithm.
5. Device-B updates P_B and β_B by computing $P_B \leftarrow U_B^{-1}$ and $\beta_B \leftarrow U_B^{-1}V_B$.
6. Device-B can execute a sequential training of OS-ELM algorithm by using the integrated P_B and β_B .

Edge devices can share their trained results by exchanging their intermediate results U and V in the proposed on-device federated learning approach, which can mitigate the privacy issues since they do not share raw data for the cooperative model update. Please note that the intermediate results U and V in Equation 13 should be updated before sending them to a cloud server or the other edge devices; so there is no need to update them for every input chunk. Also, it is possible to compute U and V at the server, not edge devices.

4 Experiment

The behavior of the proposed on-device federated learning approach is demonstrated by combining trained results from multiple edge devices.

4.1 Experiment Environment

As a dataset, the experiment is conducted with UAH-DriveSet [8] that contains car driving histories of six drivers simulating three different driving patterns: aggressive, drowsy, and normal. It can be used for the aggressive driving

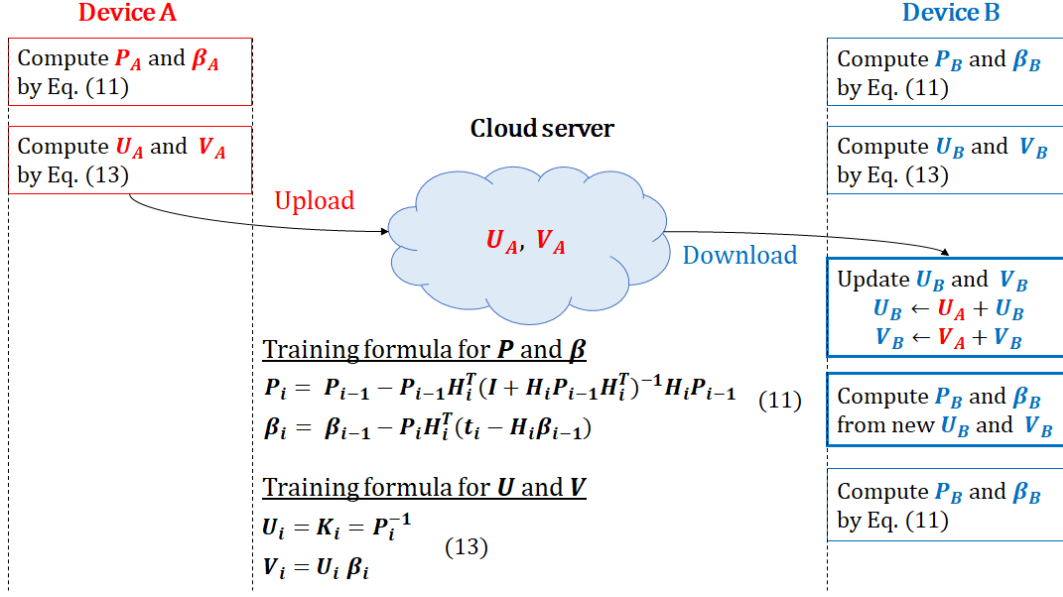


Figure 4: Flowchart of cooperative model update.

detection. Their car speed was extracted from the GPS data obtained from a smartphone fixed in their cars. The sampling frequency of the car speed was 1Hz. In this experiment, the car speed is quantized with 15 levels (1 level = 10km/h). Assuming a state is assigned to each speed level, we can build a state-transition probability table with 15×15 entries, each of which represents a probability of a state transition from one state to another.

Such a state-transition probability table is fed to the neural-network based on-device learning algorithm [1, 2] for anomaly detection. The numbers of input layer nodes n , hidden layer nodes \tilde{N} , and output layer nodes m are 225, 16, and 225, respectively. The activation function G is the sigmoid and the loss function L is the mean square error. The forget factor α is 1.00 (i.e., no forgetting) [2]. The number of instances k is 2 [7]. These instances are denoted as Device-A and Device-B in this experiment.

4.2 Experiment Result

Here, we compare loss values before and after the cooperative model update. Below is the experiment scenario using the two instances with the car driving dataset.

1. Device-A trains its model so that the aggressive driving pattern becomes normal. Device-B trains its model so that the normal driving pattern becomes normal.
2. Aggressive and normal driving patterns are fed to Device-B to evaluate the loss values. These results are denoted as “before the cooperative model update”.
3. Device-A uploads its intermediate results to a cloud server, and Device-B downloads them from the cloud server.
4. Device-B updates its model based on its own intermediate results and those from Device-A.
5. Aggressive and normal driving patterns are fed to Device-B to evaluate the loss values. These results are denoted as “after the cooperative model update”.

A low loss value means that a given input pattern is well reconstructed by Autoencoder, which means that the input pattern is normal in the edge device. Here, Device-A and Device-B are adapted to aggressive and normal driving patterns, respectively.

Figure5 shows the loss values of Device-B before and after the cooperative model update. X-axis represents the input patterns. Y-axis represents the loss values in a logarithmic scale. Blue bars represent loss values of Device-B before the cooperative model update, while orange bars represent those after the cooperative model update. Gray bars represent loss values of Device-A. In the case of aggressive pattern, the loss value of Device-B before the cooperative model update (blue bar) is high, because Device-B is trained with the normal pattern. The loss value then becomes quite low

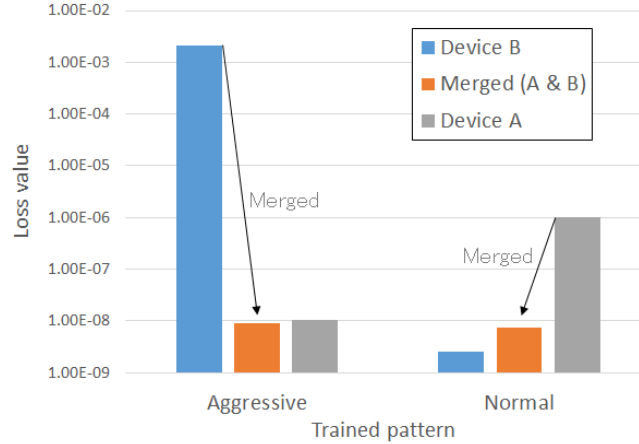


Figure 5: Loss values of Device-B before and after cooperative model update (blue and orange bars).

after integrating the intermediate results of Device-A to Device-B (orange bar). This means that the trained result of Device-A is correctly adapted to Device-B. In the case of normal pattern, the loss value before merging (blue bar) is quite low, but it slightly increases after the trained result of Device-A is merged (orange bar). Nevertheless, the loss value is still quite low. We can observe the same tendency for Device-A by comparing the gray and orange bars.

5 Summary

In this paper, we focused on a neural-network based on-device learning approach so that edge devices can train or correct their model based on incoming data at runtime in order to adapt to a given environment. Since a training is done independently at distributed edge devices, the issue is that only a limited amount of training data can be used for each edge device. To address this issue, in this paper, the on-device learning algorithm was extended for the on-device federated learning by applying the E²LM approach to the OS-ELM based sequential training. In this case, edge devices can share their intermediate trained results and update their model using those collected from the other edge devices. We illustrated an algorithm for the proposed cooperative model update. Experimental results using a driving dataset of cars showed that the proposed on-device federated learning can accurately merge trained results of the other edge devices. As a future work, we are planning to conduct more comprehensive evaluations of the proposed on-device federated learning in terms of accuracy, performance, and hardware amount.

References

- [1] Mineto Tsukada, Masaaki Kondo, and Hiroki Matsutani. OS-ELM-FPGA: An FPGA-Based Online Sequential Unsupervised Anomaly Detector. In *Proceedings of the International European Conference on Parallel and Distributed Computing (Euro-Par'18) Workshops*, pages 518–529, August 2018.
- [2] Mineto Tsukada, Masaaki Kondo, and Hiroki Matsutani. A Neural Network-Based On-device Learning Anomaly Detector for Edge Devices. *IEEE Transactions on Computers (TC)*, February 2020. Early Access.
- [3] Nan-Ying Liang, Guang-Bin Huang, P. Saratchandran, and N. Sundararajan. Fast and Accurate Online Sequential Learning Algorithm for Feedforward Networks. *IEEE Transactions on Neural Networks*, 17(6):1411–1423, November 2006.
- [4] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, July 2006.
- [5] Junchang Xin, Zhiqiong Wang, Luxuan Qu, and Guoren Wang. Elastic Extreme Learning Machine for Big Data Classification. *Neurocomputing*, 149:464–471, February 2015.
- [6] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'04)*, pages 985–990, July 2004.

-
- [7] Rei Ito, Mineto Tsukada, Masaaki Kondo, and Hiroki Matsutani. An Adaptive Abnormal Behavior Detection using Online Sequential Learning. In *Proceedings of the International Conference on Embedded and Ubiquitous Computing (EUC'19)*, pages 436–440, August 2019.
 - [8] The UAH-DriveSet. <http://www.robosafe.com/personal/eduardo.romera/uah-driveset/>.