Large-Scale Bandwidth and Power Optimization for Multi-Modal Edge Intelligence Autonomous Driving

Xinrao Li, Tong Zhang, Member, IEEE, Shuai Wang, Member, IEEE, Guangxu Zhu, Member, IEEE, Rui Wang, Member, IEEE, and Tsung-Hui Chang, Fellow, IEEE

Abstract—Edge intelligence autonomous driving (EIAD) offers computing resources in autonomous vehicles for training deep neural networks. However, wireless channels between the edge server and the autonomous vehicles are time-varying due to the high-mobility of vehicles. Moreover, the required number of training samples for different data modalities, e.g., images, point-clouds, is diverse. Consequently, when collecting these datasets from vehicles to the edge server, the associated bandwidth and power allocation across all data frames is a large-scale multi-modal optimization problem. This letter proposes a highly computationally efficient algorithm that directly maximizes the quality of training (QoT). The key ingredients include a data-driven model for quantifying the priority of data modality and two first-order methods termed accelerated gradient projection and dual decomposition for low-complexity resource allocation. Finally, high-fidelity simulations in Car Learning to Act (CARLA) show that the proposed algorithm reduces the perception error by 3% and the computation time by 98%.

Index Terms—Autonomous driving, edge intelligence, large-scale optimization.

I. INTRODUCTION

EDGE intelligence autonomous driving (EIAD) is a promising paradigm to ease the conflict between the resource-hungry model training and the resource-limited vehicle platforms [1]. Compared to cloud-assisted approaches, EIAD achieves better privacy protection and lower latency by providing computing resources in close proximity to autonomous vehicles [1]. Among others, model training is the most fundamental task in EIAD systems, which consists of dataset generation, transmission, calibration, annotation, and processing. However, since EIAD systems need to train an ensemble of deep neural networks (DNNs) for learning semantic, geometry, and motion representations, the datasets are multi-modal, and the trained DNNs are heterogeneous [2]. Therefore, the required number of training samples is diverse, disabling conventional throughput-oriented approaches.

Another challenge of EIAD is the time-varying wireless channels between the edge server and the autonomous vehicles due to high mobility, which makes the coherent time (i.e., a unit of time block for resource allocation) very small [3], [4], [5], [6]. Hence given a common data volume of an AD dataset, e.g., 100 GB, the number of time blocks for transmission can be very large. Consequently, EIAD systems require a fast large-scale optimizer for wireless resource allocation, and conventional convex optimization methods, e.g., the interior point method, are no longer suitable.

In this letter, we would like to shed some light on the above important issues. Specifically, this letter presents a new design objective, i.e., quality of training (QoT), for resource allocation, e.g., bandwidth and power allocation, in multi-modal EIAD systems. Specifically, the QoT is defined as the overall perception accuracy (or planning efficiency) of all trained DNNs. The QoT metric is monotonically increasing with the communication throughput, but their relationship is nonlinear. Hence, the QoT-oriented approach directly maximizing the QoT would result in fundamentally different designs, compared with that from conventional throughput-oriented approaches. Furthermore, despite the QoT-oriented problem being nonlinear and non-smooth, we leverage the accelerated gradient projection (AGP) and dual decomposition methods to optimally solve it in a highly efficient way with low complexity. The designed AGP achieves the fastest convergence rate, and the designed dual decomposition yields semi-closed-form solutions. The superior performance of the proposed algorithm has been verified by high-fidelity Car Learning to Act (CARLA) simulator [7].

Finally, we would like to emphasize that QoT-oriented scheduling was studied in [8], [9], [10], [11], which quantifies the importance of data uploaded from different mobile users by fitting a parametric model to experimental data. Our work exploits a similar principle but focuses on domain-specific AD datasets and time-varying channels rather than general-purpose datasets and static channels. In addition, EIAD resource allocation was extensively investigated for emerging scenarios such as multicast, federated learning and space-air-ground networks [3], [4], [5], [6]. However, these methods ignore the multi-modality issue and thus fail in achieving high QoT under resource constraints. In contrast, our method can potentially improve their performance by integrating the QoT-oriented scheduling into the EIAD resource allocation.
II. SYSTEM MODEL AND PROBLEM FORMULATION

As shown in Fig. 1(a), we consider the scheduling of data uploading from \( K \) connected autonomous vehicles (CAVs) to \( L \) edge servers for DNN model training in an EIAD system. In particular, the procedure includes the following four stages: 1) Sensing: each CAV senses the surrounding environment and stores the sensing data locally; 2) Communication: the edge server collects datasets from CAVs via uplink transmission; 3) Training: the edge server annotates the data, trains DNNs with the labeled data, and releases the models to CAVs; and 4) Inference: the performance of the trained DNNs depends on the number of uploaded samples, thus on the bandwidth and power allocated in the communication stage.

The data modality at CAV \( k \) is denoted by \( M_k \) (with \( M_k = 1 \) representing Lidar point cloud and \( M_k = 2 \) representing images) and the size of its data sample is \( D_k \). Each point-cloud with 12800 Kbits is used to train a sparsely embedded convolutional detection (SECOND) network, for Task 1 (object detection). Each image with 5600 Kbits is used to train a convolutional neural network (CNN) for Task 2 (weather classification). The SECOND architecture is shown in Fig. 1b, and the CNN architecture has four layers with \( 32 \times 64 \times 28 \times 10 \) units. Communication time is split into \( N \) time slots, where the duration of each time slot is \( T \). The channels are assumed to be quasi-static during each time slot, and vary in different time slots. All channel power gains are assumed to be predictable, since the CAV routes and the traffic map, i.e., mobility pattern, are known in advance. Let \( h_{l,k,n} \) denote the uplink channel power gain from CAV \( k \) to BS \( l \) at time slot \( n \). The achievable rate between CAV \( k \) and BS \( l \) at time slot \( n \) is

\[
R_{l,k,n}(w_{l,k,n}, q_{l,k,n}) = w_{l,k,n} \log_2 \left( 1 + \frac{h_{l,k,n} q_{l,k,n}}{N_0 w_{l,k,n}} \right),
\]

where \( q_{l,k,n} \) and \( w_{l,k,n} \) denote the transmit power and bandwidth of CAV \( k \) in BS \( l \) at time slot \( n \), respectively, and \( N_0 \) denotes the additive white Gaussian noise (AWGN) power spectral density. Note that there is no inter-cell interference in the denominator of equation (1), as adjacent BSs adopt different frequency bands for multiplexing while remote BSs causes random interference that can be included in AWGN. On the other hand, each CAV can be associated to only one BS at a certain time slot. Due to limited coverage of BSs and high mobility of CAVs, it is necessary to perform handover during the entire dataset collection procedure. To be specific, let \( x_{l,k,n} \in \{0, 1\} \) with \( \sum_l x_{l,k,n} = 1 \) for all \( (k, n) \) denote the association state between CAV \( k \) to BS \( l \) at time slot \( n \), where \( x_{l,k,n} = 1 \) represents connection and \( x_{l,k,n} = 0 \) represents disconnection. Consequently, the number of samples uploaded for the \( k \)-th modality is given by

\[
v_k = \sum_{l=1}^{L} \sum_{n=1}^{N} \frac{T x_{l,k,n} R_{l,k,n}}{ND_k}.
\]

For EIAD perception, the QoT is defined as the perception accuracy or one minus the perception error [1]. Generally, it is difficult to characterize the relationship between the perception error and the number of data samples analytically. Fortunately, based on the research results of [8], [12], this relationship can be approximately characterized by \( \Psi_k \approx a_k v^{b_k}_k \), where \( \Psi_k \) denotes error rate of the DNN at CAV \( k \), and \( a_k, b_k > 0 \) are hyper-parameters representing task difficulty. It is further indicated by [8], [12] that \( a_k, b_k \) can be obtained from the curve fitting of experimental data. Our goal is to optimize the association, bandwidth, and power of all time slots, denoted by \( X = \{x_{l,k,n}\}, W = \{w_{l,k,n}\}, \) and \( Q = \{q_{l,k,n}\} \), respectively, such that the average perception error is minimized. It can be formulated as the following optimization problem.

\[
P_0 : \min_{X, W, Q} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{n=1}^{N} \frac{T x_{l,k,n} R_{l,k,n}}{ND_k} - b_k
\]

s.t. \( \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} q_{l,k,n} \leq P_k, \quad \forall k, \) \( \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{n=1}^{N} q_{l,k,n} \leq P_{total}, \) \( \sum_{k=1}^{K} w_{l,k,n} = B_{total}, \) \( \forall l, n, \) \( \sum_{l=1}^{L} x_{l,k,n} = 1, \quad \forall k, n, \) \( \sum_{l=1}^{L} x_{l,k,n} \in \{0, 1\}, \quad \forall l, k, n, \)

where (3a) is the time slot average individual power constraint for each CAV, (3b) is the time slot average total power constraint for all CAVs, and (3c) is the total bandwidth constraint. The challenges of solving \( P_0 \) are two-fold: (1) the discontinuity of the CAV-BS association variables; (2) the curse of dimensionality brought by numerous time slots \( N \).

III. PROPOSED FIRST-ORDER ALGORITHM

A. Optimal CAV-BS Association

In order to address challenge (1), we propose the following proposition.

**Proposition 1**: The optimal \( \{x_{l,k,n}^\star\} \) to \( P_0 \) satisfies \( x_{l,k,n}^\star = 1 \) if \( h_{l,k,n} \neq \max \{h_{l,k,n}\} \) and \( x_{l,k,n}^\star = 0 \) otherwise.

**Proposition 1** can be proved by contradiction. Specifically, assume that there exists some \( x_{l,k,n'}^\star = 1 \) for \( h_{l,k,n'} \neq \max \{h_{l,k,n}\} \). Then, we can always construct another solution by setting \( x_{l,k',n'} = 0 \) and \( x_{l,k',n'} = 1 \) with \( h_{l,k',n'} \neq \max \{h_{l,k,n}\} \) while keeping other variables unchanged. This would reduce the objective of \( P_0 \), which contradicts the optimality of \( \{x_{l,k,n}^\star\} \). Similarly, we can show that the bandwidth \( w_{l,k,n} \) and power \( q_{l,k,n} \) should be zero if \( x_{l,k,n} = 0 \); otherwise those resources can always be allocated to another link.
Based on Proposition 1, we substitute \( x_{i,k,n} = x_{i,k,n}^* \) into \( P_0 \), which would not change the solution of \( P_0 \). By setting \( x_{i,k,n} = \sum_i x_{i,k,n}^*u_{i,k,n} \), \( U = \{ u_{k,n} : u_{k,n} = \sum_i x_{i,k,n}^* u_{i,k,n} \} \), and \( P = \{ p_{k,n} : p_{k,n} = \sum_i x_{i,k,n}^* q_{i,k,n} \} \), PROBLEM \( P_0 \) is equivalently transformed into

\[
P_1 : \min_{U,P} \sum_{k=1}^{K} a_k \left( \sum_{n=1}^{N} T u_{k,n} \log_2 \left( 1 + \frac{g_{k,n}P_{k,n}}{N_0 u_{k,n}} \right) \right)/N D_k \quad -b_k
\]

s.t. \( 1/N \sum_{n=1}^{N} p_{k,n} \leq P_k, \forall k \),

\( 1/N \sum_{n=1}^{N} \sum_{k=1}^{K} p_{k,n} \leq P_{total}, \forall n \)

\( \sum_{k=1}^{K} u_{k,n} = B_{total}, \forall n \)

\( p_{k,n} \geq 0, u_{k,n} \geq 0, \forall k, n \).

PROBLEM \( P_1 \) is convex due to the following reasons:
- All the constraints are linear;
- Function \( a_k(\cdot)^{-b_k} \) is non-increasing and convex;
- Function \( T u_{k,n} \log_2 \left( 1 + \frac{g_{k,n}P_{k,n}}{N_0 u_{k,n}} \right) \) is convex due to the following reasons:

\[ \text{Why Acceleration?} \]

When the bandwidth allocation vectors is \( \{ u_{i,k,n} = u_{i,k,n}^\circ \} \), where \( u_{i,k,n}^\circ \) denotes the given value of bandwidth of the \( k \)-th CAV in BS \( i \) and the \( n \)-th time slot, PROBLEM \( P_1 \) is converted to PROBLEM \( P_3 \) by fixing bandwidth allocation

\[
P_3 : \min_{\{ p_{i,k} \}} \sum_{k=1}^{K} a_k \left( \sum_{n=1}^{N} T u_{k,n} \log_2 \left( 1 + \frac{g_{k,n}P_{k,n}}{N_0 u_{k,n}} \right) \right)/N D_k \quad -b_k
\]

s.t. (3a), (3b), and (3d).

Applying dual decomposition to PROBLEM \( P_3 \) yields

\[ D(\lambda) : \max_{\lambda \geq 0} \min_{\{ p_{i,k} \}} \sum_{k=1}^{K} a_k \left( \sum_{n=1}^{N} T u_{k,n} \log_2 \left( 1 + \frac{g_{k,n}P_{k,n}}{N_0 u_{k,n}} \right) \right) -b_k
\]

\[ + \lambda \left( \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} p_{k,n} - P_{total} \right), \]

where \( G_k = \{ p_k : \frac{1}{N} \sum_{n=1}^{N} p_{k,n} \leq P_k, p_{k,n} \geq 0 \} \). The dual of PROBLEM \( P_3 \) is a bilevel optimization problem, where the outer problem is an unconstrained nonsmooth maximization problem and the inner problem is a constrained but decomposable problem. In the outer problem, the dual variable \( \lambda \) can be updated via the sub-gradient descent method as

\[ \lambda[i+1] = \lambda[i] + \xi \left( \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} p_{k,n} - P_{total} \right), \]

where \( \lambda[i] \) and \( p_{k,n}^{[i]} \) are dual variable \( \lambda \) and power allocation \( p_{k,n} \) of the \( i \)-th iteration, respectively, and \( \xi \) is the step size. At the \( i \)-th iteration, the inner problem for fixed \( \lambda[i] \) can be
Algorithm 1 Proposed First-Order Algorithm for Solving $\mathcal{P}_1$

1: Initialize $\eta = 10^4$ and $\xi = 10^{-3}$.
2: Repeat
3: Set $i = 1$ and $c[0] = 1$.
4: Repeat
5: Calculate $c[i]$, $Q[i]$, $U[i+1]$ based on (5)-(7).
6: Update $i \leftarrow i + 1$
7: Until: The stop criterion is satisfied.
8: Set $i = 0$ and $\lambda[0] = 0$.
9: Repeat
10: Calculate $P_k[i]^*$ based on (13).
11: Update $\lambda[1+i]$ based on (10).
12: Update $i \leftarrow i + 1$
13: Until: The stop criterion is satisfied.
14: Until: The stop criterion is satisfied.

equivalently decomposed into $K$ sub-problems, given by

$$
P_{i}^{[\cdot]}(k) : \min_{p_k[i] \in G_k} a_k \left( \sum_{n=1}^{N} \frac{T u_{k,n}^{0}}{N D_k} \log_2 \left( 1 + \frac{g_{k,n} P_k^{[i]}}{N_0 u_{k,n}^{0}} \right) \right)^{-b_k} + \lambda[i] t_k
$$

\text{s.t.} (3a), (3d),
$$
\frac{1}{N} \sum_{n=1}^{N} p_{k,n}[i] = t_k.
$$

We have the following proposition on the optimal solution of Problem (12).

Proposition 2: Given $t_k$, the optimal $p_k^*$ to Problem (12) is

$$
p_k^{[i*]}(\mu) = \left[ \frac{\mu N D_k \ln 2 - N_0 u_{k,n}^{0}}{g_{k,n}} \right]^{+}, \forall k, n,
$$

where $\mu > 0$ satisfies $\sum_{n=1}^{N} p_k^{[i*]}(\mu) = N \min(P_k, t_k)$.

Proof: Please refer to [15, Appendix B].

Remark 2: Proposition 1 indicates that the optimal $p_k^*$ can be found via one-dimensional search over $t_k$. Since $t_k \in [0, P_k]$ and the objective function is uni-modal w.r.t. $t_k$, the optimal $t_k^*$ in (13) can be found by bisection search within $[0, P_k]$. The iteration complexity of bisection is $O(\log(1/\epsilon))$.

D. Complexity Analysis

The entire procedure of the proposed method is summarized in Algorithm 1. It can be seen that Algorithm 1 involves two levels of iterations. In the outer-level AO iteration, the AGP method is first adopted to solve $\mathcal{P}_2$, which executes (5)-(7) iteratively. The computation is dominated by equation (5), which requires a complexity of $O(KN^2)$ (computing each element in $\nabla \Xi(Q[i])$) needs a complexity of $O(N)$ and there are $KN$ elements. Consequently, with $O(1/\epsilon)$ iterations, the AGP method costs a complexity of $O(KN^2/\epsilon)$. Then, to solve $\mathcal{P}_1$, dual decomposition is adopted, which executes (13) for all $(k, n)$ and (10) iteratively. The computation cost of (13) for all $(k, n)$ is given by $O(KN)$. Therefore, with $O(1/\epsilon)$ iterations for sub-gradient update and $O(\log(1/\epsilon))$ iterations for bisection search, the dual decomposition method requires a computation complexity of $O(\log(1/\epsilon)KN/\epsilon)$. In summary, the total complexity of Algorithm 1 is given by $O(\text{ITER}(KN^2/\epsilon + \log(1/\epsilon)KN/\epsilon))$, where ITER is the number of iterations for AO to converge.

IV. SIMULATION RESULTS

The simulations were done on the CARLA simulator [7], [16] with $K = 2$, where the first CAV with a 64-line LiDAR on top of the car is a point-cloud data collector, while the second CAV with an RGB camera in front of the car is an image data collector. The Adam optimizer is adopted for training SECOND and CNN. The communication parameters are given by $N = 1000$, $B_{tota}=20$ MHz, $P_1 = P_2 = 1$ W, $P_{total} = 2$ W, $N_0 = -110$ dBm/Hz. We simulate $L = 10$ BSs and the CAV-BS distance is generated randomly from 5 m to 150 m. The channels are generated by using a distance-dependent path-loss model with 30 dB loss at a unit distance of 1 m.

Fig. 2(a) shows the convergence behaviour of the proposed algorithm. It can be seen that the algorithm converges very fast within 10 iterations. Fig. 2(b) compares the computation time of the proposed algorithm and CVX on a desktop with 17-7700 3.6GHz CPU and 64G RAM. The proposed algorithm significantly reduces computation time. Notably, when $N = 1000$, the proposed first-order algorithm reduces the computation time of CVX (i.e., interior-point method) by 98.2%, and the gain increases with the number of time slot $N$.

Next, we compare the proposed algorithm with benchmark schemes as follows: 1) Scheme 1: equally allocating bandwidth and power across all CAVs; 2) Scheme 2: maximizing the total communication throughput via water-filling [4]; 3) Scheme 3: QoT-oriented power optimization with equal bandwidth allocation [8]; 4) Scheme 4: QoT-oriented bandwidth and power allocation ignoring time-varying channels [9]. Figs. 2(c) and 2(d) compare the number of uploaded samples and the average perception error of the proposed algorithm with those of the benchmark schemes, respectively. It can be seen from Fig. 2(c) that the proposed algorithm leads to a significantly smaller perception error, i.e., a higher QoT, than those of other benchmarks as shown in Fig. 2(d).
Compared with Scheme 1, the proposed scheme reduces the perception error by 3%, which implies that resource allocation is crucial to the EIAD systems. Moreover, Scheme 2 leads to the second-worst performance among all the simulated schemes, meaning that the objective function have a more significant impact on EIAD than other factors such as the choice of design variables and the input channels. Finally, by comparing the proposed method with Schemes 4 and 3, we find that ignoring the time-varying feature of wireless channels would degrade the system performance inevitably.

The simulation results of Figs. 2(c) and 2(d) are further visualized in Fig. 3(a) and Fig. 3(b). In particular, no matter which algorithm is chosen, the trained CNNs always distinguish different weather. This is because task 2 has a fast learning process, and tens of images are enough for realizing accurate perception. On the other hand, the SECOND trained with the proposed algorithm successfully detects objects on the road. In contrast, other schemes yield missing or inaccurate detection results due to the insufficient number of point clouds. This is because the proposed algorithm automatically allocates more resources to task 1, which has a more significant learning curve, for QoT maximization.

V. CONCLUSION

This letter has studied the large-scale bandwidth and power allocation problem in EIAD. A first-order accelerated algorithm with linear complexity has been proposed. The proposed algorithm achieved a smaller perception error than the state-of-the-art, and a lower complexity than interior point method.

REFERENCES

[1] J. Zhang and K. B. Latief, “Mobile edge intelligence and computing for the Internet of Vehicles,” Proc. IEEE, vol. 108, no. 2, pp. 246–261, Feb. 2020.

[2] D. Feng et al., “Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges,” IEEE Trans. Intell. Transp. Syst., vol. 22, no. 3, pp. 1341–1360, Mar. 2021.

[3] H. Xiao, J. Zhao, Q. Pei, J. Feng, L. Liu, and W. Shi, “Vehicle selection and resource optimization for federated learning in vehicular edge computing,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 8, pp. 11073–11087, Aug. 2022.

[4] X. Pei, H. Yu, Y. Chen, M. Wen, and G. Chen, “Hybrid multicast/unicast design in NOMA-based vehicular caching system,” IEEE Veh. Technol., vol. 69, no. 12, pp. 16304–16308, Dec. 2020.

[5] S. Yu, X. Gong, Q. Shi, X. Wang, and X. Chen, “EC-SAGINs: Edge-computing-enhanced space-air-ground-integrated networks for Internet of Vehicles,” IEEE Internet Things J., vol. 9, no. 8, pp. 5742–5754, Apr. 2022.

[6] B. Liu, L. Wang, M. Liu, and C.-Z. Xu, “Federated imitation learning: A novel framework for cloud robotic systems with heterogeneous sensor data,” IEEE Robot. Autom. Lett., vol. 5, no. 2, pp. 2509–2516, Apr. 2020.

[7] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “CARLA: An open urban driving simulator,” in Proc. Ist Annu. Conf. Robot Learn., 2017, pp. 1–16.

[8] S. Wang, Y.-C. Wu, M. Xia, R. Wang, and H. V. Poor, “Machine intelligence at the edge with learning centric power allocation,” IEEE Trans. Wireless Commun., vol. 19, no. 11, pp. 7293–7308, Nov. 2020.

[9] L. Zhou et al., “Learning centric wireless resource allocation for edge computing: Algorithm and experiment,” IEEE Trans. Veh. Technol., vol. 70, no. 1, pp. 1035–1040, Jan. 2021.

[10] T. Zhang, S. Wang, G. Li, F. Liu, G. Zhu, and R. Wang, “Accelerating edge intelligence via integrated sensing and communication,” in Proc. IEEE ICC, 2022, pp. 1–7.

[11] X. Li, S. Wang, G. Zhu, Z. Zhou, K. Huang, and Y. Gong, “Data partition and rate control for learning and energy efficient edge intelligence,” 2021, arXiv:2107.08884.

[12] C. White, A. Zela, R. Ra, Y. Liu, and F. Hutter, “How powerful are performance predictors in neural architecture search?” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 28454–28469.

[13] L. Grippo and M. Sciandrone, “On the convergence of the block nonlinear Gauss–Seidel method under convex constraints,” Optim. Res. Lett., vol. 26, no. 3, pp. 127–136, 2000.

[14] W. Su, S. P. Boyd, and E. J. Candès, “A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 2510–2518.

[15] X. Li, T. Zhang, S. Wang, G. Zhu, R. Wang, and T.-H. Chang, “Large-scale bandwidth and power optimization for multi-modal edge autonomous driving,” 2022, arxiv.abs/2210.09659.

[16] S. Wang et al., “Federated deep learning meets autonomous vehicle perception: Design and verification,” IEEE Netw., early access, Dec. 12, 2022, doi: 10.1109/MNET.104.2100403.