Multi-source DNN task offloading strategy based on in-network computing

Lizi Hu*, Yuhao Chai*, Qin Li**, Weiyuan Li**, Yong Zhang*

*Beijing Key Laboratory of Work Safety Intelligent Monitoring, Beijing University of Posts and Telecommunications, Beijing, 100876, China

**China Mobile Research Institute, Beijing, China

lizihu@bupt.edu.cn, chaiyh@bupt.edu.cn, liqinyjy@chinamobile.com, liweiyuan@chinamobile.com, yongzhang@bupt.edu.cn

***Abstract*—As applications grow in scale, the centralized computing approach leads to excessive bandwidth requirements and high computational latencies. The traditional computing model regards the network as a transmission pipeline and has not fully explored the potential of network devices. At present, in-network computing is a new type of computing model that delegates application-layer processing functions to the network data plane. It can process traffic during transmission, reduce the cost of network bandwidth transmission, and alleviate the computing pressure and energy consumption of the cloud-side system. However, it needs to consider the redeployment cost. This paper proposes a multi-source DNN task offloading strategy based on in-network computing, which combines edge calculation, in-network computing and cloud computing. At the same time, it makes full use of the traditional routing nodes with no computing ability in the network. Particle swarm optimization (PSO) is used to solve the problem in the offloading scheme optimization. Service level agreement violation (SLAV) is introduced, and the network resources are offloaded in balance while the quality of service of users is satisfied. Simulated experiment results show that the proposed algorithm can reduce the cost and achieve convergence compared with the traditional offloading algorithm. In particular, we can find that there is an optimal deployment scheme in the network, which can make full use of the computing resources and bandwidth resources of the network, significantly reduce the computing pressure and transmission overhead of the whole network, and realize the balanced offloading of node resources.**

***Keywords*—in-network computing, edge-cloud computing, task offloading, particle swarm optimization**

Lizi Hu is working toward the master's degree from the Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include communication and resource allocation.

Yuhao Chai is working toward the master's degree from the Beijing University of Posts and Telecommunications, Beijing, China. His research interests include game theory, 5G networks and resource allocation.

Qin Li is a senior researcher of China Mobile Research Institute. She received her master's degree from Chongqing University and has engaged in technical and application research on core network and content network for more than 10 years. Her research interest covers network intelligence, future network architecture and digital twin network.

Weiyuan Li, is a researcher at China Mobile Research Institute. She received a BE degree from Beijing Jiaotong University and a PhD in Information and signaling processing from Chinese Academy of Sciences. Her current research is in intelligence network.

Yong Zhang received the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China in 2007. He is a Professor with the School of Electronic Engineering, BUPT. He is currently the Director of Fab.X Artificial Intelligence Research Center, BUPT. He is the Deputy Head of the mobile internet service and platform working group, China communications standards association. He has authored or coauthored more than 80 papers and holds 30 granted China patents. His research interests include Artificial intelligence, wireless communication, and Internet of Thing.