IEEE Hong Kong 6G Wireless Summit

11-12 September 2024
Hong Kong, China

IEEE ComSoc IEEE Communications Society | IEEE

Department of Electrical and Electronic Engineering, The University of Hong Kong

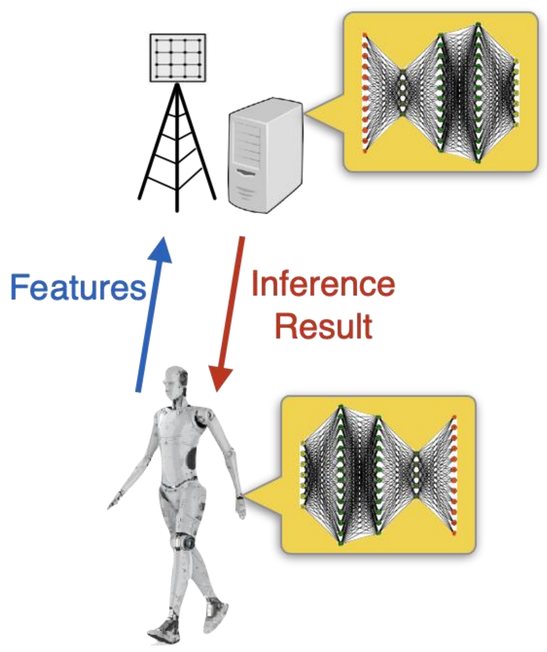
Ultra-Low-Latency Feature Transmission for Edge Inference

Presenter: Qunsong Zeng and Zhanwei Wang (Prof. Kaibin Huang's Research Group)

The next generation mobile networks will feature the widespread deployment of artificial intelligence algorithms at the network edge, which provides a platform for edge intelligence. In this work, we propose new air-interfaces targeting the edge inference systems, called ultra-low-latency (observation) feature transmission.

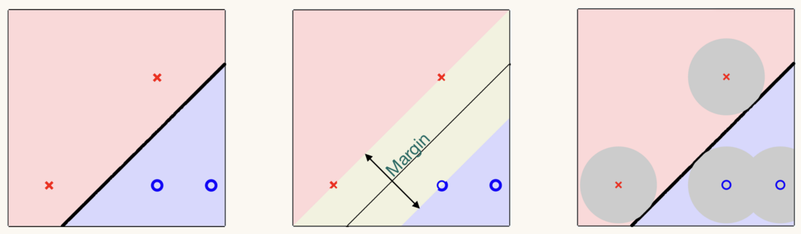
One framework consists of a novel transmission approach that exploits classifier's robustness, which is measured by classification margin, to compensate for a high bit error probability resulting from ultra-low-latency transmission (e.g., short packet and/or no coding). By utilizing the tractable Gaussian mixture model, we mathematically derive the relation between bit error probability and classification margin under constraints on classification accuracy and transmission latency. The result sheds light on system requirements to support ultra-low-latency feature transmission. Finally, experiments using deep neural networks as classifier models and real datasets are conducted to demonstrate the effectiveness of ultra-low-latency feature transmission in communication latency reduction while providing a guarantee of classification accuracy.

The other approach to realize ultra-low-latency inference is to leverage the hard deadline on the chronological occurrence of view collection and feature transmission. For this vision, we propose a novel framework of ultra-low-latency edge inference for latency-constrained distributed inference. The framework harnesses the interplay between short packet transmission and accuracy improvements from multi-view sensing to meet a stringent deadline while boosting the end-to-end inference performance. Under the latency constraint, a fundamental tradeoff between communication reliability and the number of views, controlled by the packet length, is revealed and optimized. The optimization is tackled by deriving accurate surrogate functions of the expressions for the end-to-end inference accuracy.

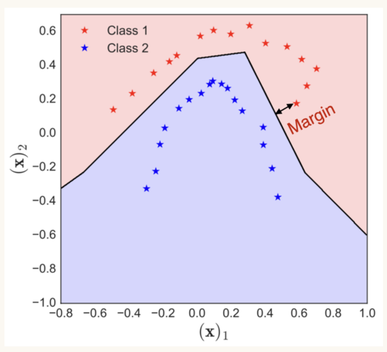


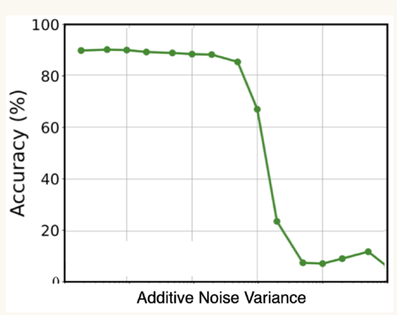
Ultra-Low-Latency Feature Transmission for Point-to-Point Inference

Principle: Robustness of Edge Inference



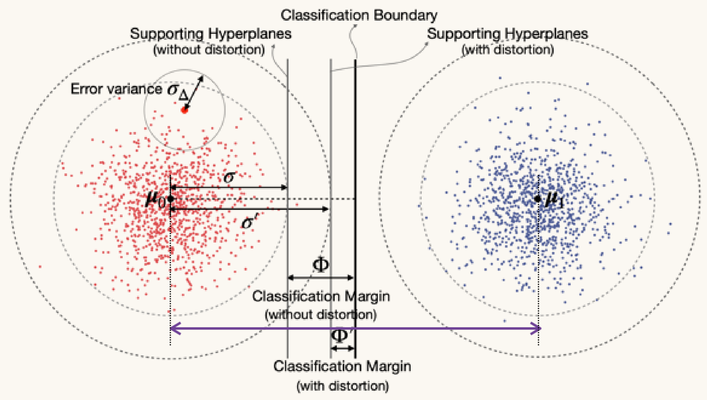
Margin of SVM classifier



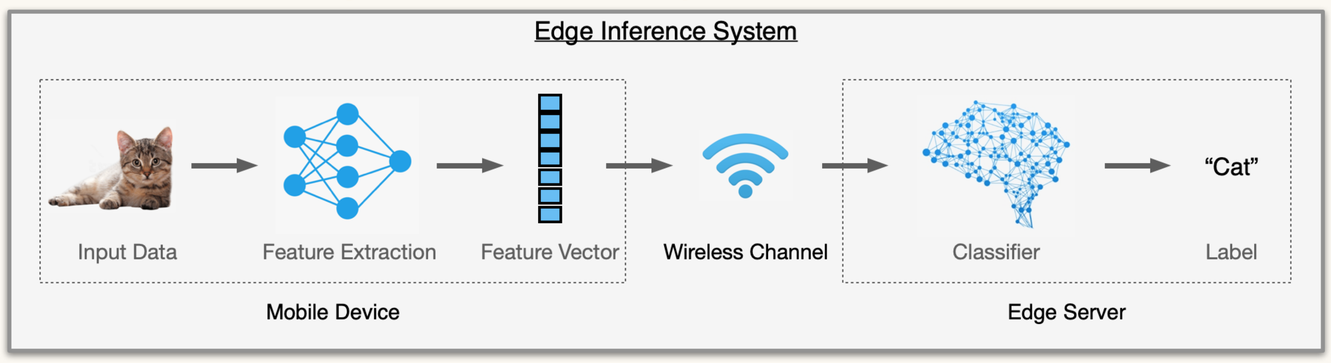


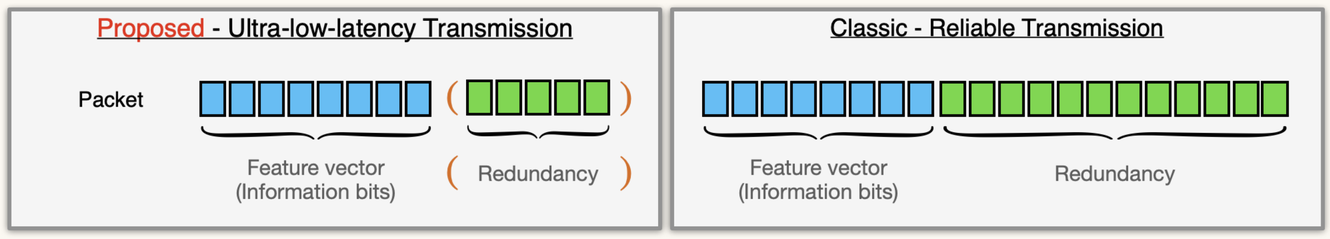
(Dataset "German Traffic Sign Recognition Benchmark")

Margin of DNN classifier



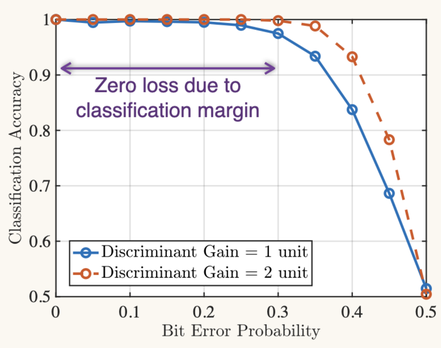
Design: Ultra-Low-Latency Feature Transmission



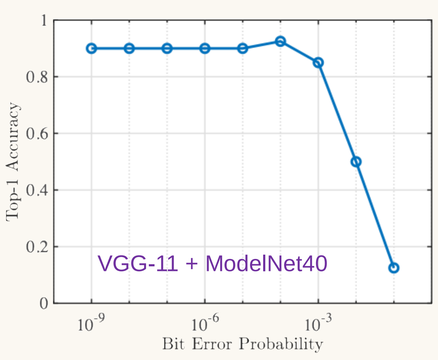


Performance Evaluation

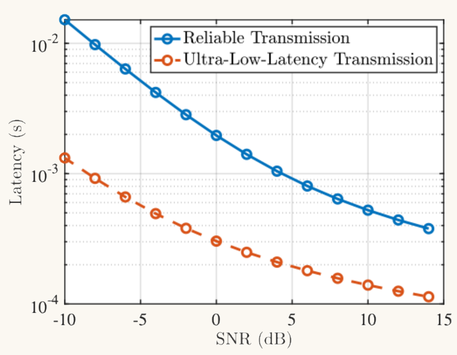
Gaussian Mixture Model



Deep Neural Network

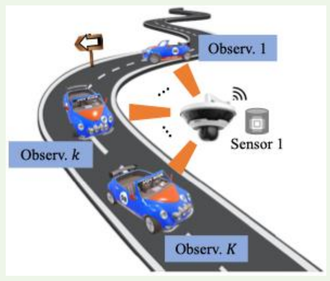


Feature Transmission Latency

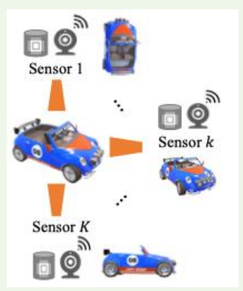


Ultra-Low-Latency Feature Transmission for Distributed Inference

Sensing Scenarios

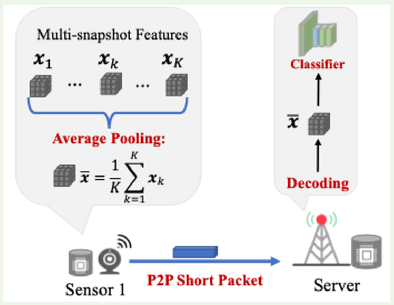


Multi-snapshot Sensing

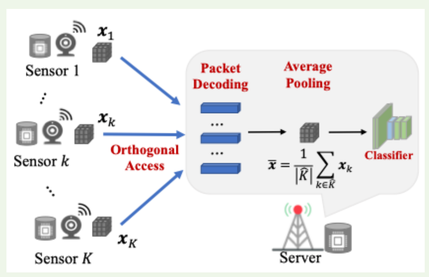


Multi-view Sensing

Multi-access Model

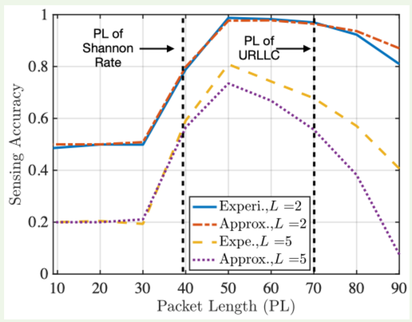


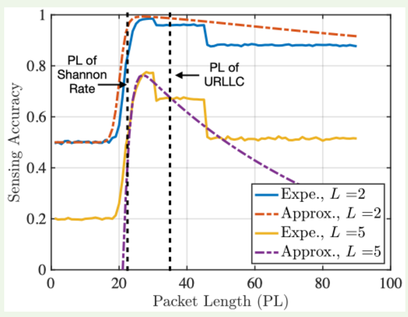
Point-to-Point



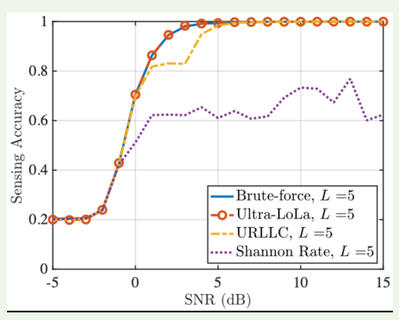
Orthogonal Access

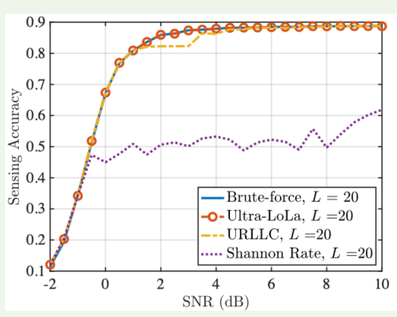
Tradeoff Demonstrations

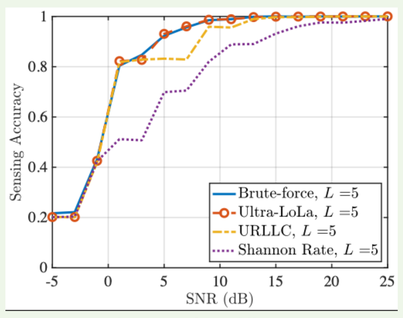




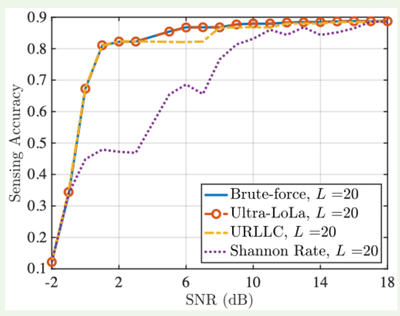
Performance Evaluation







Linear Classifier



CNN Classifier

[Journal] Q. Zeng, Z. Wang, Y. Zhou, H. Wu, L. Yang, and K. Huang, "Knowledge Based Ultra-Low-Latency Semantic Communications for Robotic Edge Intelligence", minor revision in *IEEE Trans. Commun.*, 2024.

[Conference] Q. Zeng, Z. Wang, Y. Zhou, H. Wu, L. Yang, and K. Huang, "Ultra-Low-Latency Feature Transmission for Edge Inference", accepted by *IEEE Global Commun. Conf. (GLOBECOM)*, Cape Town, South Africa, Dec. 8-12, 2024.

[Journal] Z. Wang, A. E. Kalør, Y. Zhou, P. Popovski, and K. Huang, "Ultra-Low-Latency Edge Inference for Distributed Sensing", submitted to *IEEE Trans. Wireless Commun.*, 2024.

Acknowledgment

This work was supported in part by the Hong Kong Research Grants Council under the Areas of Excellence Scheme Grant AoE/E-601/22-R