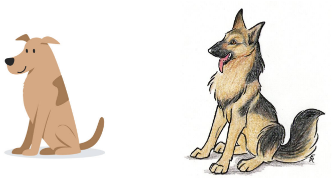
Hierarchical Inference with an Offload Queue

Srinivas Nomula, Parimal Parag,
Ganesh Ayalvadi, Jaya Prakash Champati

Electrical Communication Engineering,
Indian Institute of Science

Problem setup



Data Samples

Edge device

Workload type: DL Inference

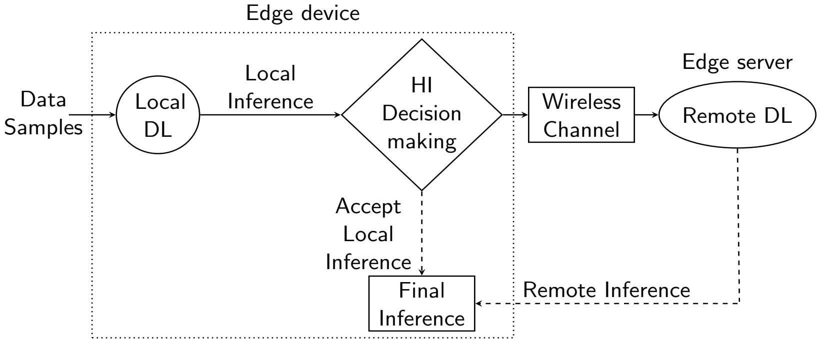
- **Objective**: Obtain accurate outputs

State-of-the-art techniques

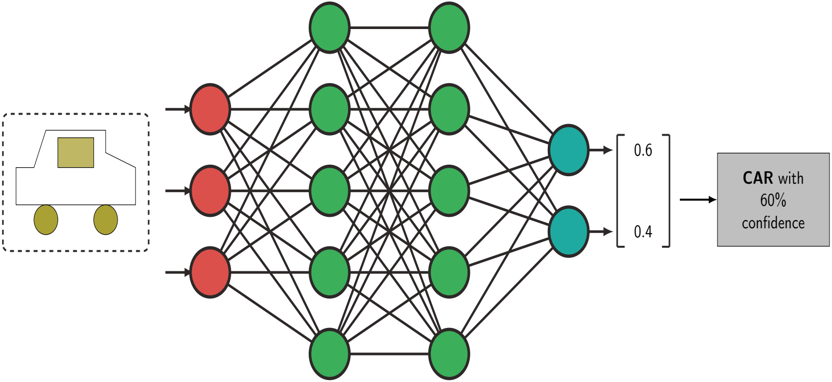
DL Inference

- On-device inference [Iborra et al., 2016]
  - Utilize only local DL model to infer the incoming data
- DNN partitioning: [Tang et al., 2017]
  - Partition the whole DL model into two partitions subject to their capacities
- Inference-load balancing: [Wang et al., 2018], [Fresa et al., 2023]
  - Offload a part of data to the remote server subject to resource availability

Hierarchical Inference



Confidence

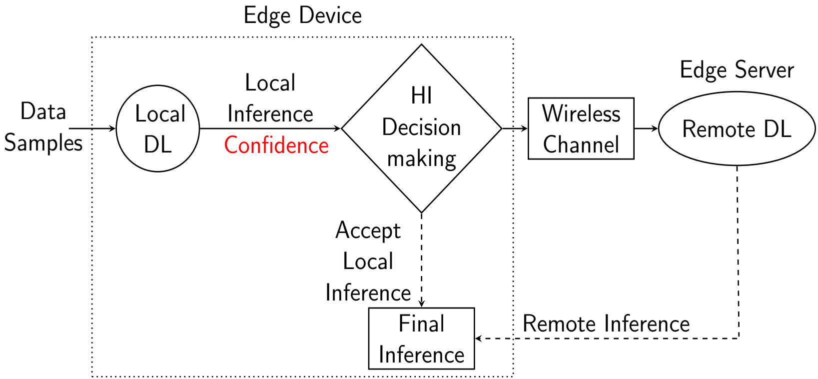


Maximum softmax value of Deep Learning (DL) model output

Remark

Confidence value sequence are independent of the input task sequence

Thresholding confidence



For a task n with confidence Z_n and threshold $\theta \in [0, 1]$ the offload decision is given by,

$$I_n \triangleq \mathbb{1}_{\{Z_n \leqslant \theta\}}$$

Thresholding policies

$$I_n \triangleq \mathbb{1}_{\{Z_n \leqslant \theta\}}$$

Fixed threshold policy

A stochastic threshold policy with a fixed threshold

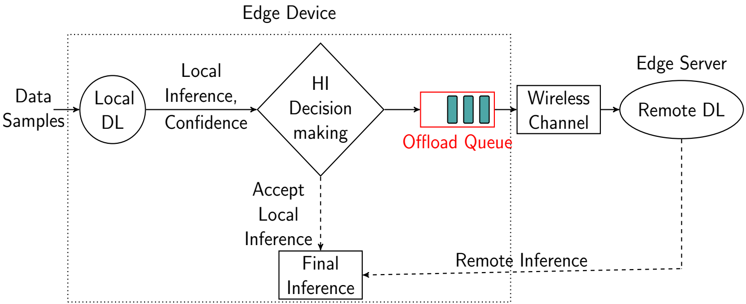
Random threshold policy

A policy where the threshold is a random variable with distribution π_t at any time t

Piecewise random threshold policy

A stochastic threshold policy where the threshold θ is independent for each task n and is identically distributed for all times

System with offload queue[5]



- Total arrivals A_t, total offloads O_t, max service tasks C_t
- Misclassification error Φ_n, Total errors:

$$E_t \triangleq \sum_{n \in A_t} \bar{I}_n \Phi_n$$

- Queue evolution: $Q_t \triangleq \max\{Q_{t-1} - C_t + O_t, 0\}$

Objective

Minimize the limiting-average errors in the system by ensuring queue stability

$$\min \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} E_t$$

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} Q_t = 0$$

Unknown

- Confidence distribution is unknown
- Confidence-to-mismatch relationship is unknown

Control parameter

- Offloading thresholds for every incoming task

Online learning

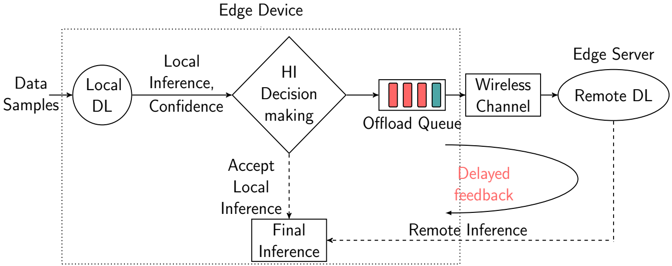
$$E_t \triangleq \sum_{n \in A_t} \bar{I}_n \Phi_t$$

Partial feedback

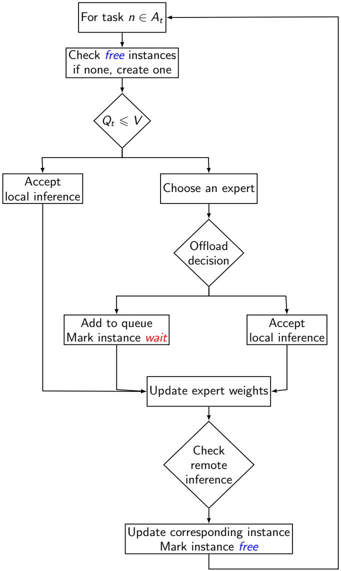
- Φ_n for local accepted tasks are unknown

Delayed feedback

- Φ_n for tasks in queue revealed after some delay



Blackbox Online learning with delayed feedback (BOLD)-Hedge-Q policy



Limiting bounds

Regret analysis

The regret of Black-box Online learning under Delayed Feedback (BOLD)-Hedge-Q policy is given by

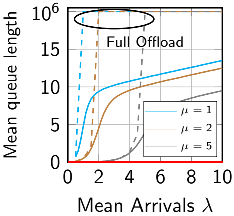
$$R_T^H = O\left(\lambda_m\left(\left(\frac{V}{\lambda_m}+1\right)\log|\mathcal{Z}|\right)^{\frac{1}{3}} T^{\frac{2}{3}}\right)$$

Total errors

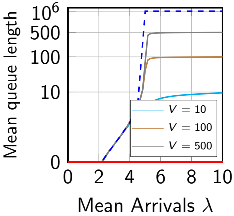
Let $\bar{E}_0 \in \mathbb{Z}_+$ denote the limiting error observed when used the optimal threshold $\theta_0 \in \mathcal{Z}$. The limiting error under BOLD Hedge-Q is upper-bounded as

$$\lim_{T\to\infty} \mathbb{E}\bar{E}_T^H \leqslant \left(\bar{E}_0 + \frac{B}{V}\right) \wedge 1$$

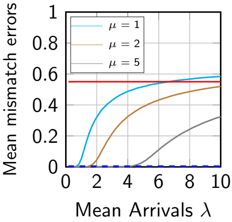
Simulation results



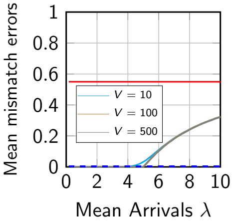
(a) Queue length vs λ; $V = 10$



(b) Queue length vs λ; $\mu = 5$



(a) Mismatch err. vs λ; $V = 10$



(b) Mismatch err. vs λ; $\mu = 5$

ImageNet: Average queue length and mismatch errors of BOLD Hedge-Q for varying arrival rate λ with $\epsilon = 0.4$ and $\eta = 0.1$. Dashed lines represent Full-offload; solid red lines represent On-Device

Conclusion

- Utilized Hierarchical Inference framework for better accuracy
- Edge device framework included an offload queue
- Modeled into a constrained-optimization problem of minimizing total mismatch errors subject to queue stability
- Noticed partial and delayed feedback problems
- Proposed *BOLD-Hedge-Q policy* to achieve our objective
- Derived robust performance bounds for any arrival and service

References

[1] *R. Sanchez-Iborra and A. F. Skarmeta, "TinyML-enabled frugal smart objects: Challenges and opportunities," IEEE Circuits Syst. Mag., vol. 20, no. 3, pp. 4-18, Aug. 2020*

[2] *Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," ACM SIGARCH Comput. Arch. News, 2017*

[3] *A. Fresa and J. P. Champati, "Offloading algorithms for maximizing inference accuracy on edge device in an edge intelligence system," IEEE Trans. Parallel Distrib. Syst., vol. 34, no. 7, pp. 2025–2039, Jul. 2023*

[4] *J. Wang, Z. Feng, Z. Chen, S. George, M. Bala, P. Pillai, S.-W. Yang, and M. Satyanarayanan, "Bandwidth-efficient live video analytics for drones via edge computing," in IEEE/ACM Symp. Edge Comput., 2018*

[5] *Yang et al., 2024. ADTO: A Trust Active Detecting-based Task Offloading Scheme in Edge Computing for Internet of Things. ACM Trans. Internet Technol. 24*

[6] *Michael J. Neely., Stochastic Network Optimization with Application to Communication and Queueing Systems, Morgan and Claypool Publishers, 2010*