Context-aware Data Operation Strategies in Edge Systems for High Application Performance

Tanmoy Sen and Haiying Shen
Department of Computer Science
University of Virginia
Intelligent Cognitive Assistants (ICAs): The Future

- ICAs assist working, learning, transportation, healthcare, and etc. in a smart city
  - Traffic accident prediction
  - Parking suggestion
  - Detect heart attack
  - Detect Covid-19

ICAs applications seamlessly collect data, process data and take actions.
Intelligent Cognitive Assistants (ICAs): The Future

**Challenge** for the marriage for ICAs:
- Achieve low job latency with low power and bandwidth consumption
- Focus on data operations

- constrained power and the bandwidth
- power and bandwidth consuming
Related Work and Novelty

• **Data placement**: where to store sensed data (ICFEC’17, ASAC’18, TC’19)
  - Only on source data
  - Still consumes high bandwidth and power

• **Data collection**: decrease the transmitted data samples (ICCPS’15, IACC’15, ICPADS’18, TMC’19)
  - Do not consider influence on AI accuracy

• **Novelty** of our system: **Context-aware Data Operation System (CDOS)**
  - Overcome the limitations
    - Data sharing and placement (CDOS-DP)
    - Context aware data collection (CDOS-DC)
    - Data redundancy elimination (CDOS-RE)
Data Sharing and Placement

Challenge
• Collecting source data and sharing source data still consume high power and bandwidth

Rationale
• Intermediate and final data results may be shared by many jobs
Data Sharing and Placement (cont.)

- Traffic accident prediction
  - Route suggestion
  - Current location
  - Traffic volume
  - Weather

- Traffic condition prediction
  - Traffic accident prediction
  - Traffic condition prediction

- Parking suggestion
  - Traffic volume
  - Current location

- Traffic condition prediction
  - Traffic condition prediction

- Traffic volume
  - Traffic condition prediction
  - Traffic condition prediction

- Weather
  - Traffic volume
  - Traffic condition prediction

- Traffic volume
  - Traffic condition prediction
  - Traffic condition prediction

- Current location
  - Traffic accident prediction
  - Traffic condition prediction
  - Traffic volume
  - Weather

- Weather
  - Traffic condition prediction
  - Traffic volume
  - Current location

- Traffic condition prediction
  - Traffic condition prediction
  - Traffic volume
  - Weather
Data Sharing and Placement (cont.)

Route suggestion
- Current location
- Weather
- Traffic volume

Traffic condition prediction
- Traffic accident prediction

Parking suggestion
- Current location

Traffic accident prediction

Route suggestion

Current location
Data Sharing and Placement (cont.)

Strategy

- **Storing** intermediate and final computation results for sharing
- **Use** dependency graph
- **Placement**: linear programming problem with aim to minimize communication overhead and latency

Communication overhead for storing and fetching data

\[
\min_{n_s \in N} \forall_{d_j \in D_g, n_g \in N_d} C(n_g, n_s, d_j, N_{d_j}^{d_j}),
\]

Time latency for storing and fetching data

\[
L(n_g, n_s, d_j, N_{d_j}^{d_j}) \times x(d_j, n_s)
\]

Selected node
Context aware Data Collection

Challenge:
- Reduce data sampling frequency without compromising AI accuracy

Context-related Factors
- Abnormality of data
- Priority of Events
- Data Weight on Computation Result
- Context of an Event
Context aware Data Collection (cont.)

Challenge

- Reduce data sampling frequency without compromising decision making accuracy

Context-related Factors

- Abnormality of data
- Priority of Events
- Data Weight on Computation Result
- Context of an Event

Traffic prediction

Frequency: Low

Car accident prediction

Frequency: High
Context aware Data Collection (cont.)

Challenge

• Reduce data sampling frequency without compromising decision making accuracy

Context-related Factors

- Abnormality of data
- Priority of Events
- Data Weight on Computation Result
- Context of an Event

Weight of each input

Weight: Time>temperature for traffic prediction
Context aware Data Collection (cont.)

Challenge

- Reduce data sampling frequency without compromising decision making accuracy

Context-related Factors

- Abnormality of data
- Priority of Events
- Data Weight on Computation Result
- Context of an Event

Sunny weather, light traffic
Frequency: Low

Rainy weather, moderate traffic
Frequency: High
Context aware Data Collection (cont.)

Strategy

• Change data collection frequency based on cumulative weight of the four factors

\[ W_{d_j} = \sum_{e_i \in E_j} w^1_{d_j} \cdot w^2_{e_i} \cdot w^3_{d_j,e_i} \cdot w^4_{e_i}, \quad (0 < W_{d_j} \leq 1) \]
Context aware Data Collection (cont.)

Strategy

• Change data collection frequency based on cumulative weight of the four factors

\[
W_{d_j} = \sum_{e_i \in E_j} w_{d_j, e_i} \cdot w_{e_i} \cdot w_{d_j, e_i} \cdot w_{e_i}^4, \quad (0 < W_{d_j} \leq 1)
\]

• Additive linear increase multiplicative decrease (AIMD) algorithm to tune the collection time interval

\[
T_{t+1} = \begin{cases} 
T_t + \frac{\alpha}{\eta} W_{d_j} \quad (\alpha \geq 1), & \text{all errors are within their limits} \\
\frac{T_t}{\beta + \eta W_{d_j}} \quad (\beta \geq 1), & \text{otherwise.}
\end{cases}
\]
Data Redundancy Elimination

Challenges
• Data transmission between nodes (edge, fog and cloud nodes) generate high bandwidth overhead and delay

Rationale
• Data redundancy in the data stream

Strategy
• Redundancy elimination
Experimental Setup

• Simulation on iFogSim: 5000 edge nodes

| Table 1: Simulation parameters. |
|----------------|----------------|----------------|----------------|
| Edge node (EN) | Fog node (FN1 & FN2) |
| Storage capacity | Storage capacity |
| 10MB- 200MB | 150MB- 1GB |
| Edge-FN1 network bandwidth | FN1-FN2 network bandwidth |
| 1 Mbps- 2 Mbps | 3 Mbps- 10 Mbps |
| Idle/Busy power | Idle/Busy power |
| 1/10 MW | 80/120 MW |

• Real device testbed
  • 5 Raspberry-Pi devices

• Compared methods
  • iFogStor (ICFEC’17)- finds data hosts that minimizes data transmission latency
  • iFogStorG (ASAC’18)- partitions the system to sub-graphs and finds the optimal data placement in each partition
  • LocalSense - each edge node senses all of its needed source data
Experimental Results (cont.)

- **23%-55% improvement**
- **21%-46% improvement**
- **18%-29% improvement**
Experimental Results

- 26% improvement
- 29% improvement
- 21% improvement
Conclusion

• **Motivation:** Reduce communication latency, job latency, power consumption and bandwidth consumption for AI jobs on the edge

• **Approach:** Context-aware **Data Operation System (CDOS)**
  - Data sharing and placement
  - Data collection
  - Redundancy elimination

• **Future work:** jointly consider job scheduling and data operations

Let ICAs assistant you!
Thank you!

Questions & Comments?

Tanmoy Sen
ts5xm@virginia.edu
University of Virginia