NNU · 南京师范大学
NANJING NORMAL UNIVERSITY

2025 icassp HYDERABAD, INDIA
Celebrating Signal Processing

Volatile MAB-based Configuration Selection for Offloading Video Analytics Tasks to Edges

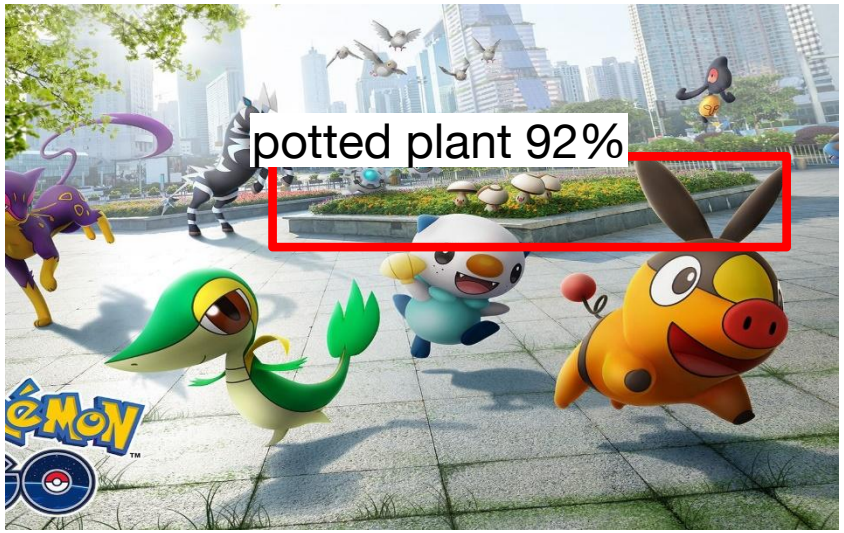
Yu Liang (Nanjing Normal University)
Sheng Zhang (Nanjing University)
Jie Wu (Temple University)

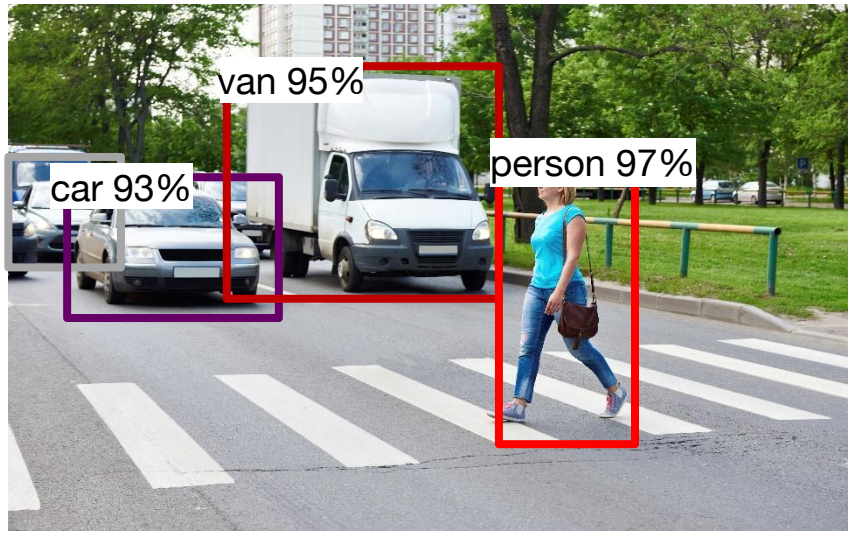
Outline

- Background
- System Model
  - Key Idea
  - Problem Formulation
- Solution
  - Algorithm: Design and Analysis
- Experiments

Real-time Video Analytics : Taking Object Detection as Example

- Target Scenarios :





Camera Analysis

Real-time Video Analytics

Offloading Real-time Video Analytics Tasks

- How to reduce computation and transmission costs at edges?

adjusting video transmission configuration has become an effective approach

configurations adjusting faces several challenges

Challenge #1 Specific video analytics models supported

each edge server with heterogeneous hardware only supports specific video analytics models .

- match the video configuration for transmission and the configuration supported by edge servers
- offloading decisions and configurations should be also dynamically adjusted over time

Challenge #2: trade-off between accuracy and energy consumption

more expensive transmission configurations lead to high accuracy analytics results as well as high computing and transmission energy costs

- this tradeoff decision must be made before offloading.
- dynamically adjusting transmission configurations over time while minimizing overall energy consumption and maximizing analytics accuracy becomes a critical issue.

Challenge #3: Difficult to estimate energy consumption in advance

The energy consumption for transmission and computation varies dynamically over time , making it difficult to estimate energy consumption in advance

- edge servers may experience energy depletion or movement, leading to uncertainty in the candidate server set

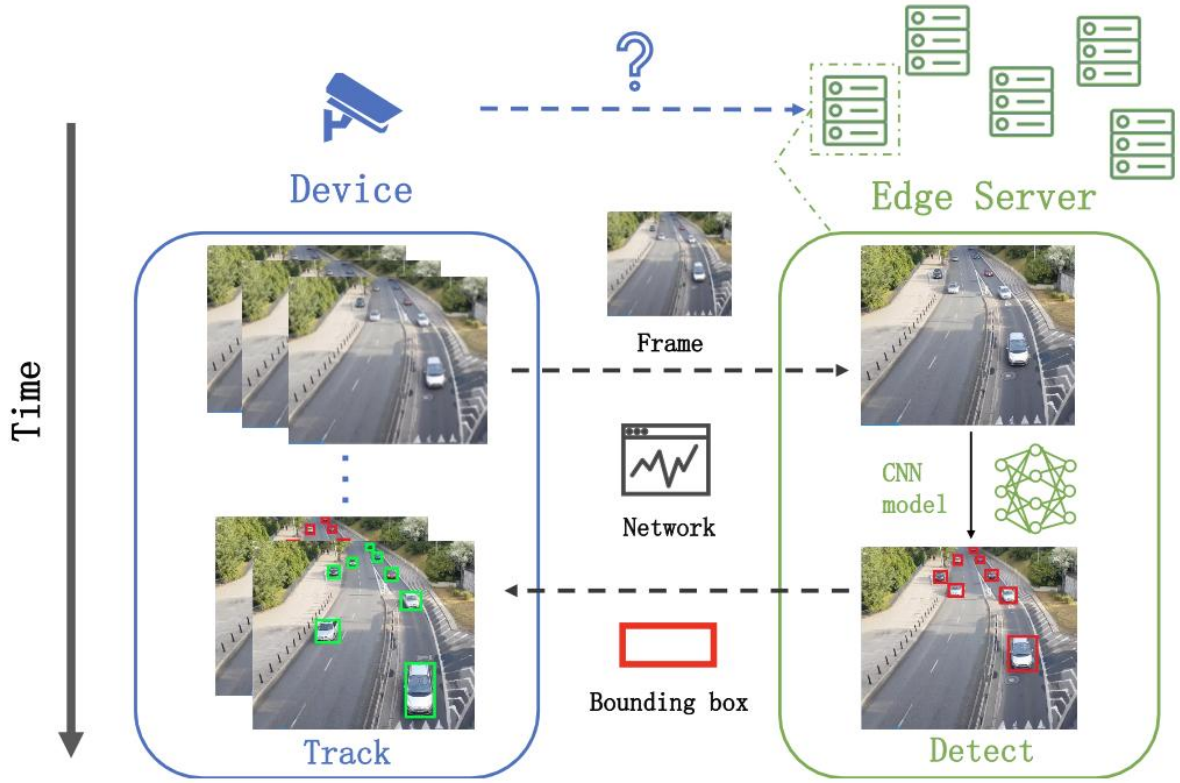
Prior Studies

- server-driven,
- offline + online,
- parallel decoding,
- super resolution

most of them are based on deterministic and known information of edge servers. do not take into account the dynamic candidate server set.

Our Solution: Key idea

- We utilizes the volatile multiarm bandit framework to capture the variations in the availability and performance of edge servers, predicting the utility rewards achievable by offloading to these servers and adjusting configurations to make task offloading decisions adaptively.



System Model

- a set of time epochs T = {1, .., t, .., T}, which are further divided into multiple time slots J_t.
- M = {mtIt = 1, 2, ..., T} denote the set of all video analytics tasks, where m_t represents the task at slot t.

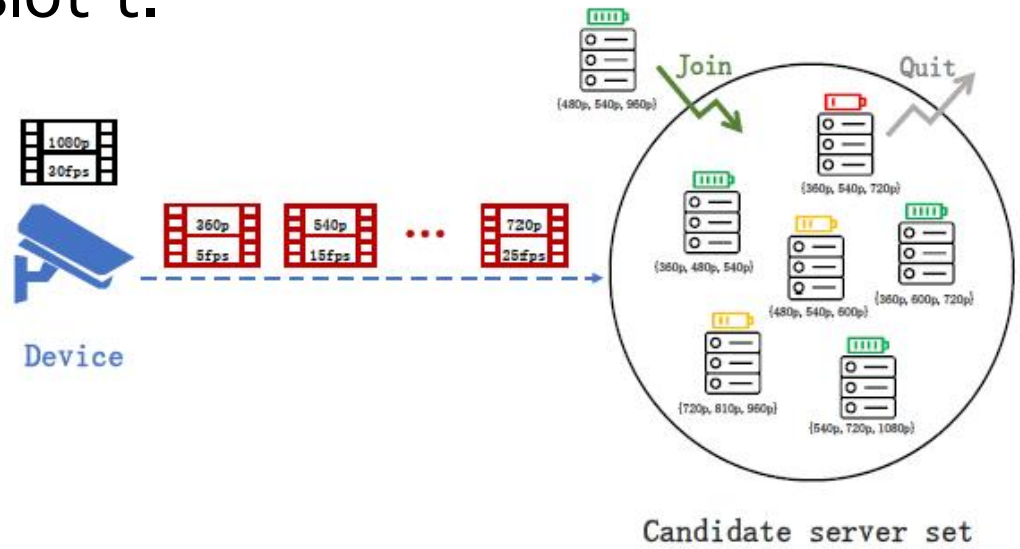


Fig. 1: Video offloading and configuration selection in multi-edge environment

System Model

- Detection Accuracy Model

The analytics accuracy $a_{k,t}$ of sub-task $m_{k,t}$ can be expressed as:

$$a_{k,t} = \epsilon_t\left(\sum_{s=1}^{|\mathcal{S}_{j,t}|} x_{k,s,t} r_{k,s,t}\right)\phi_t(f_{k,t}),$$

- Energy consumption Model

$$e_{k,t} = \sum_{s=1}^{|\mathcal{S}_{j,t}|}\left(e_{k,s,t}^{trans} + e_{k,s,t}^{pro}\right)$$

Processing energy

Transmission energy

System Model

- objective

The analytics accuracy $a_{k,t}$ of sub-task $m_{k,t}$ can be expressed as:

$$\mathcal{P}: \max \sum_{t=1}^{T} \sum_{k=1}^{K_t} \sum_{j=1}^{|\mathcal{J}|} (a_{k,t} - \omega e_{k,t})$$

$$s.t. \quad C_1: x_{k,s,t} \in \{0,1\}, \forall t, k \in \mathcal{K}_t, s \in \cup_j \mathcal{S}_{j,t},$$

$$C_2: \sum_{k=1}^{K_t} \sum_{s=1}^{|\mathcal{S}_{j,t}|} x_{k,s,t} = 1, \forall t, k \in \mathcal{K}_t, s \in \cup_j \mathcal{S}_{j,t},$$

$$C_3: f_{k,t} \in [\underline{f}, \bar{f}], \forall t, k \in \mathcal{K}_t.$$

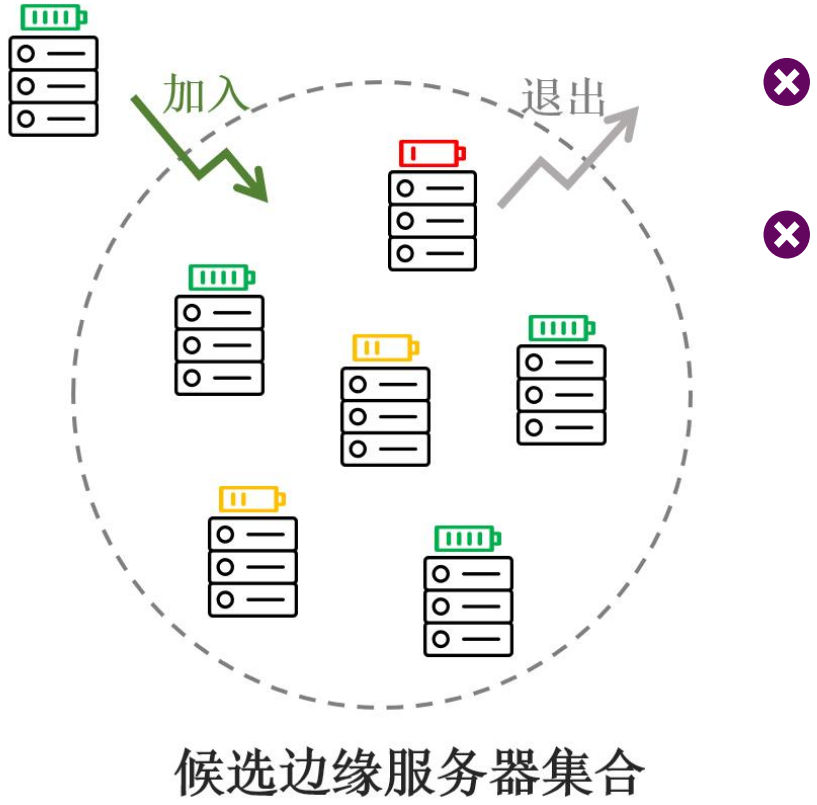
jointly limit that the device can only select one server for task offloading of video analytics tasks simultaneously

restricts the selection range of video frame rates

decision variable, indicate whether subtask $m_{k,t}$ is offloaded to edge server s

Analysis

- Before offloading tasks to server s, $\gamma_{s,t}$ and $\mu_{s,t}$ are stochastic.
  - “exploration”：
    - to offload video tasks to servers that have not been previously selected to gather more information
    - ⊗ the results obtained at the moment may not be optimal.
  - “exploitation”：
    - to make the best decision based on existing information
    - ⊗ may lead to the solution being trapped in a local optimum



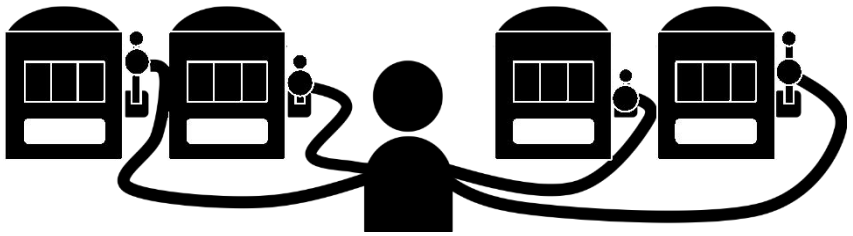
- The uncertainty of the candidate server set
  - ⊗ Traditional online learning algorithms will result in a large amount of redundant learning, reducing overall efficiency

Key ideas

Volatile Multi-Armed Bandit

- continuously “explores” the candidate edge server set by selecting edge servers that have not been chosen for offloading video tasks

- When the candidate edge server set becomes relatively stable, VACS “exploits” the acquired information to estimate the expected utility rewards of each server, making the optimal offloading decision



Algorithm

Algorithm 1: VACS

1 **for** $t = 1$ *to* T **do**
2 estimate accuracy parameters ϵ_t and ϕ_t;
3 **for** $j = 1$ *to* $|\mathcal{J}_t|$ *and each subtask* $m_{k,t}$ **do**
4 **if** $\exists$ *new edge server* s **then**
5 $u_{k,t} \leftarrow m_{k,t}$;
6 select server s, observe $\tilde{\gamma}_{s,t}$ and $\tilde{\mu}_{s,t}$;
7 **for** $r_{k,s,t}$ *in* $\mathcal{R}_s$ **do**
8 get frame rate $f_{k,t}$ according to $r_{k,s,t}$;
9 get best configuration $r^*_{k,s,t}$ and $f^*_{k,t}$;
 $\bar{V}^{s,r^*,f^*}_{k,t}(\bar{\gamma}_{s,t}, \bar{\mu}_{s,t}) \leftarrow \tilde{V}^{s,r^*,f^*}_{k,t}(\tilde{\gamma}_{s,t}, \tilde{\mu}_{s,t})$;
10 $\pi_{s,t} \leftarrow 1$;
11 **else**
12 select server s using Eq. (7), observe $\tilde{\gamma}_{s,t}$ and $\tilde{\mu}_{s,t}$;
13 **for** $r_{k,s,t}$ *in* $\mathcal{R}_s$ **do**
14 get frame rate $f_{k,t}$ according to $r_{k,s,t}$;
15 get best configuration $r^*_{k,s,t}$ and $f^*_{k,t}$;
16 $\bar{V}^{s,r^*,f^*}_{k,t}(\bar{\gamma}_{s,t}, \bar{\mu}_{s,t}) \leftarrow \frac{\bar{V}^{s,r^*,f^*}_{k,t}(\bar{\gamma}_{s,t},\bar{\mu}_{s,t})\pi_{s,t} + \tilde{V}^{s,r^*,f^*}_{k,t}(\tilde{\gamma}_{s,t},\tilde{\mu}_{s,t})}{\pi_{s,t}+1}$;
17 $\pi_{s,t} \leftarrow \pi_{s,t} + 1$;

Exploration

Choose the servers that join for the first time

Exploitation

selects an edge server for offloading based on existing information

Theoretical Analysis

- Define the regret of task m_t as:

the theoretical optimal utility function value obtained by selecting the optimal server s_* with future information

$$Reg(t) = \sum_{k=1}^{K_t} \mathbb{E}[V_{k,t}^{s^*,r^*,f^*}] - \mathbb{E}[V_{k,t}^{s,r^*,f^*}]$$

the utility obtained by VACS

- the upper bound of the regret for each analytics task m_t is:

$$\mathbb{E}[Reg(t)] \leq |\mathcal{J}_t| \sum_{s \neq s^*} \lambda(8\Delta_s^{-1} \ln K_t + \frac{8}{3}\Delta_s)$$

Experiments

Settings

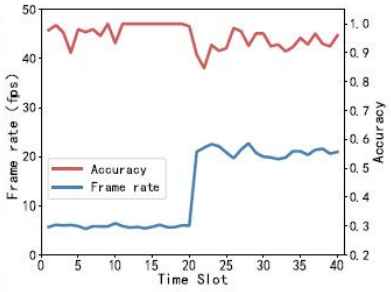
- Dataset： AI City2019
- Raw resolution and frame rate： 1080p， 30fps
- Object detection module： YOLOv5
- Supported input resolutions in edge servers ： 360p， 480p， 540p， 600p， 720p， 810p， 960p and 1080p（each edge server supports 3 of them）
- Baselines algorithms :

  (1)Accuracy-Optimal (AO), which maximizes accuracy and ignores energy consumption,
  (2)Energy-Consumption-Optimal (ECO),which minimizes energy consumption and ignores accuracy.

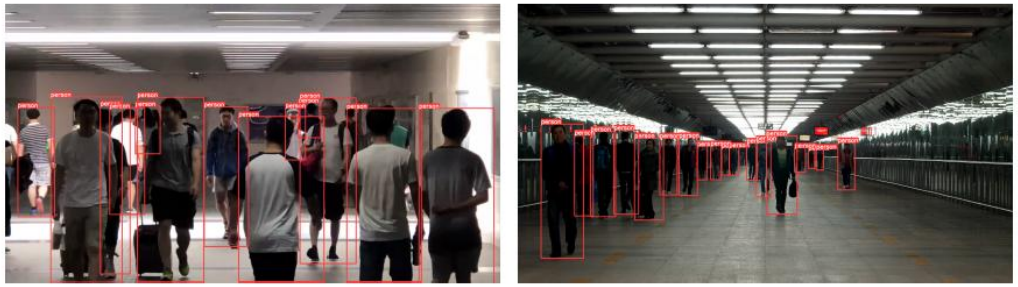
Experiment Result

- VACS dynamically adjust configuration

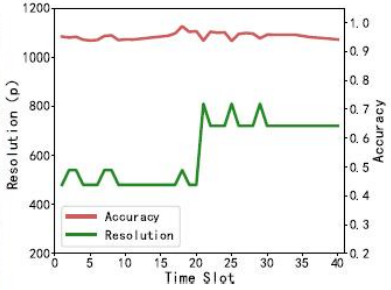
(a) Scenario 1: vehicle speed changes over time



(b) fps. vs. accu.



(c) Scenario 2: pedestrian size changes over time

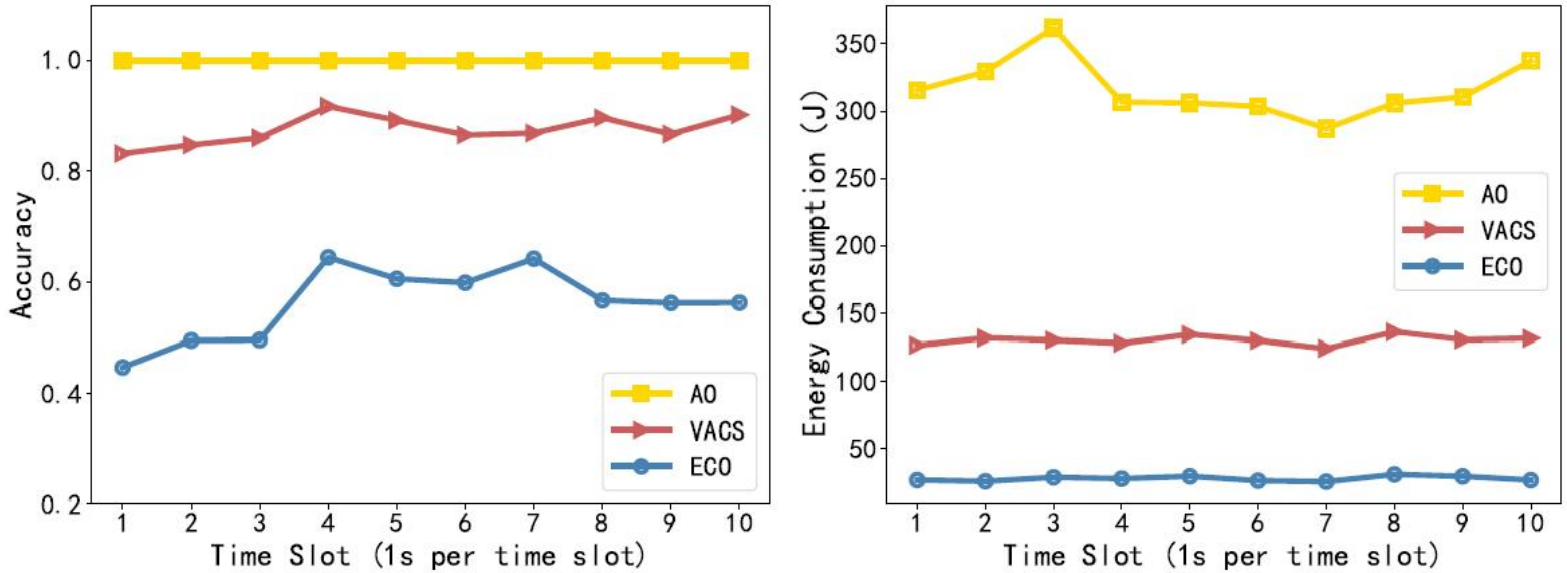


(d) res. vs. accu.

Fig. 2: VACS dynamically adjusts configuration

Experiment Result

- Performance of VACS:
  - achieves average accuracy of 87%, only a 13% reduction compared to AO, but it saves 59% in energy consumption
  - strike a good balance between accuracy and energy consumption.



(a) Accuracy comparison (b) Energy consumption comparison

Fig. 3: Comparison results

Conclusion

- **To maximize the analytics accuracy while minimizing the energy consumption.**
- VACS :**a volatile MAB-based configuration selection algorithm**
- **Theoretic Analysis**
- **Trace-driven Experiments**