Batch Adaptative Streaming for Video Analytics

徐振轩

2022.12.28

[1] Lei Zhang , et al. "Batch Adaptative Streaming for Video Analytics. " In Proceedings of IEEE INFOCOM 2022，May 2-5 2022, pp. 2158-2167.

目录

- 背景介绍
- 系统模型
- 解决方法
- 实验结果
- 研究进展

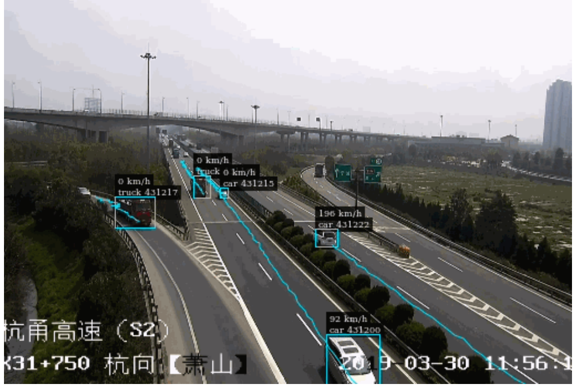
背景介绍

- 趋势：实时视频分析应用愈发广泛

工业安全

目标监控

失踪人口寻找

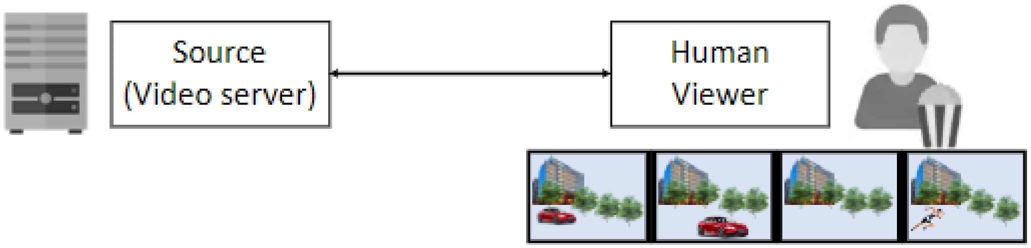


交通检测

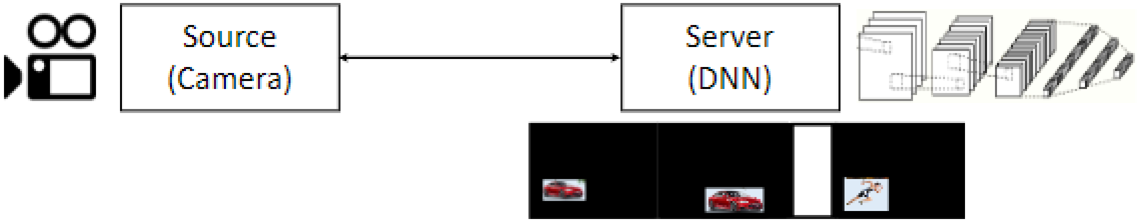
实时视频流分析，指通过计算机视觉算法对一个或多个摄像头产生的视频流内容自动进行分析和理解，从而在视频流录制和传输的同时完成目标识别、异常检测等复杂任务。

背景介绍

- 在适应有限资源的同时最大化服务质量



(a) Video streaming for human viewers



(b) Video streaming for computer-vision analytics

☑ **算法准确率**　　☒ **清晰程度**

☑ **延迟**　　☒ **流畅程度**

现有工作

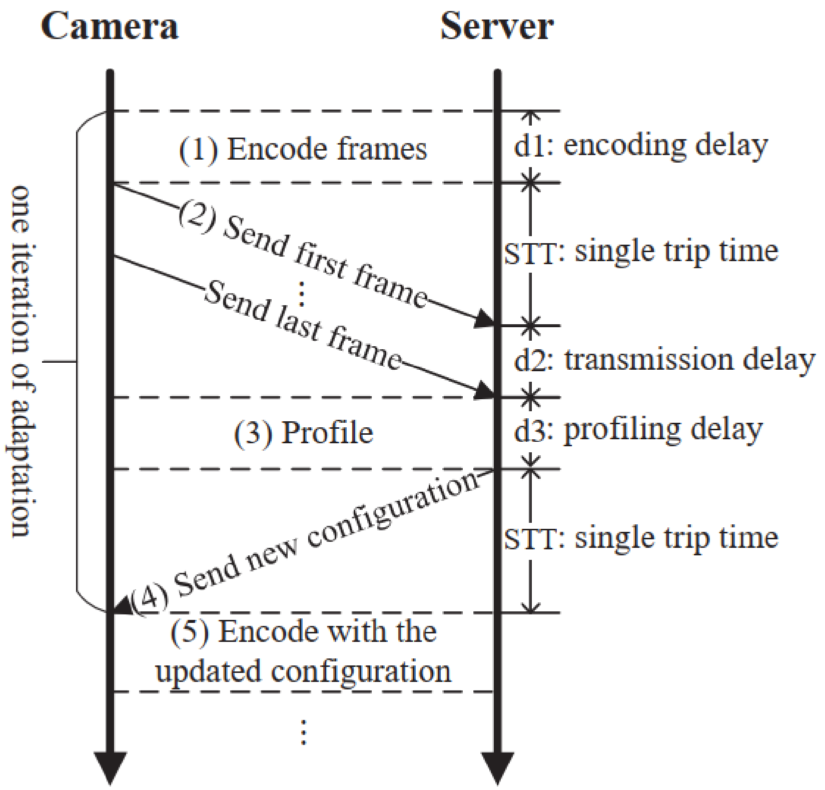
■ The basic workflow of server-driven adaptation

(1) encodes the raw frames

(2) sends the encoded video chunk to the server

(3) profiles the received chunk to select a configuration update with the best bandwidth-accuracy trade-off

(4) sends the adaptation decision back to the source camera



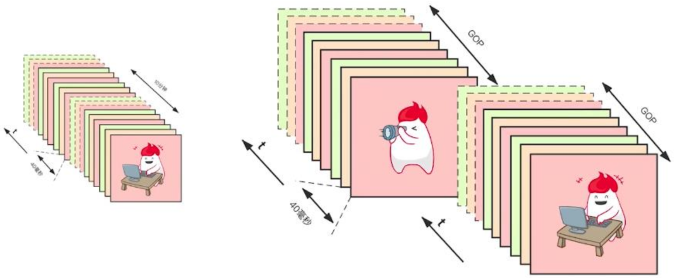
[1] K. Du, A. Pervaiz, X. Yuan, A. Chowdhery, Q. Zhang, H. Hoffmann, and J. Jiang, "Server-driven video streaming for deep learning inference," inProceedings of ACM SIGCOMM 2020, pp. 557–570.

Motivation

- **Batch Transmission and Batch Processing**

Devices: ——消耗更少带宽
frames are encoded and transferred in batches

GOP



Servers: ——提高GPU吞吐率
multiple frames as a batch input to ML algorithm

Motivation

■ Impact of Batch Size on Bandwidth and Response Delay

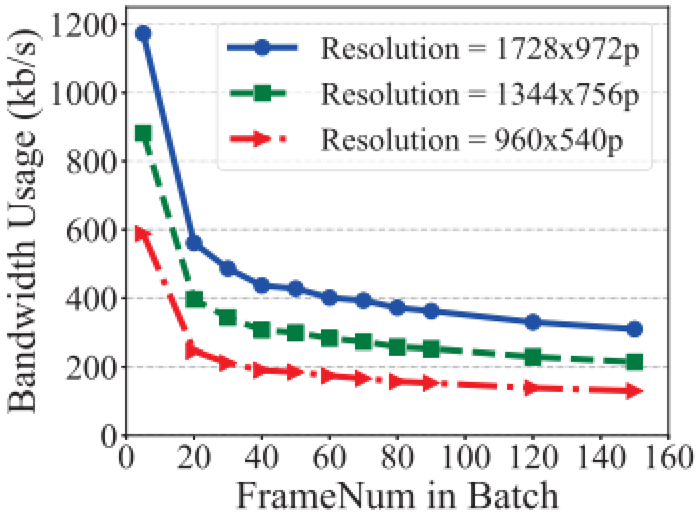


Fig. 2. Batching effect on transmission efficiency

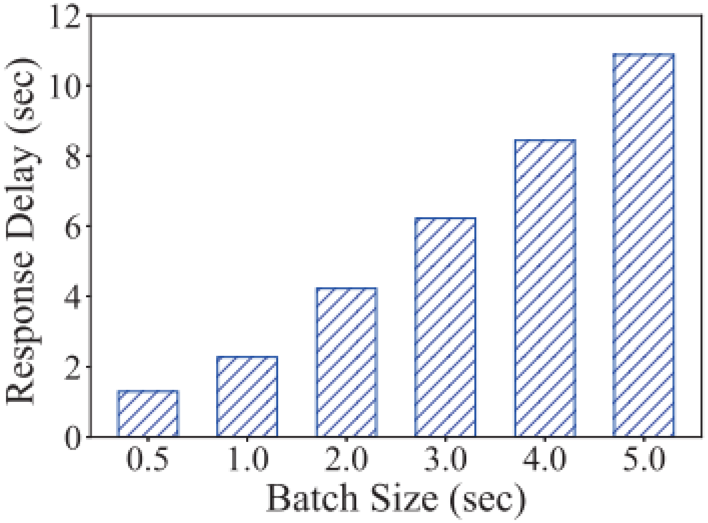


Fig. 3. Batching effect on response delay

Batchsize越大，带宽消耗越小

Batchsize越大，响应时延越大

Motivation

■ Impact of Batch Size on Inference Accuracy

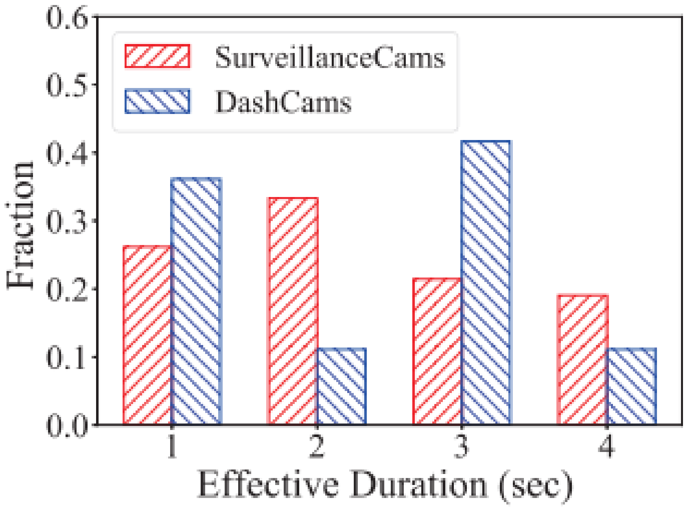


Fig. 4. Distribution of Effective Duration

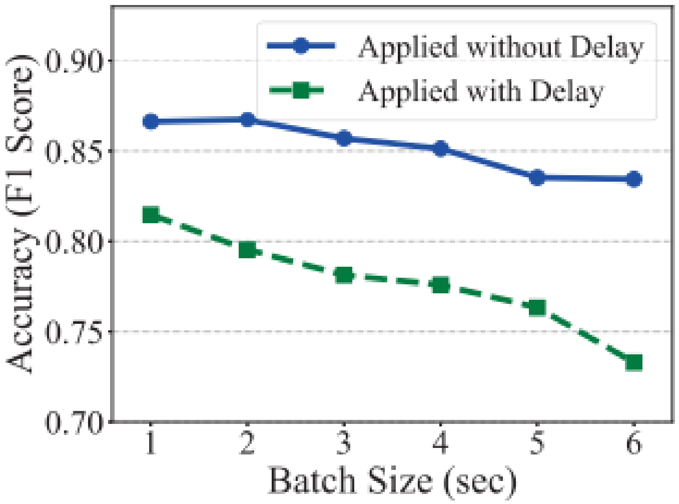


Fig. 5. Accuracy with different batch size

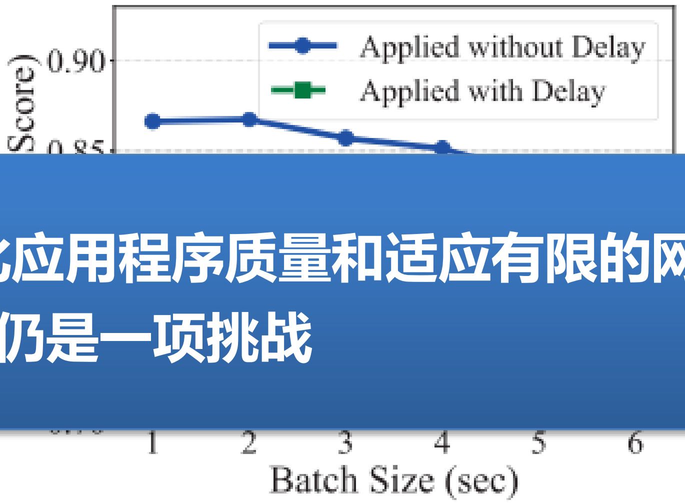
大多数自适应配合的有效期很短

存在响应延迟严重影响推理的准确性

Motivation

■ Impact of Batch Size on Inference Accuracy

0.6
0.5
0.4
▨ SurveillanceCams
▨ DashCams
1
2
3
4
Effective Duration (sec)



挑战：如何调整batchsize，在最大化应用程序质量和适应有限的网络资源之间取得平衡仍是一项挑战

Fig. 4. Distribution of Effective Duration

Fig. 5. Accuracy with different batch size

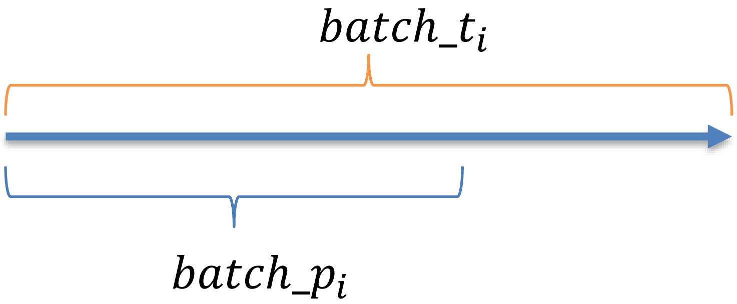
大多数自适应配合的有效期很短

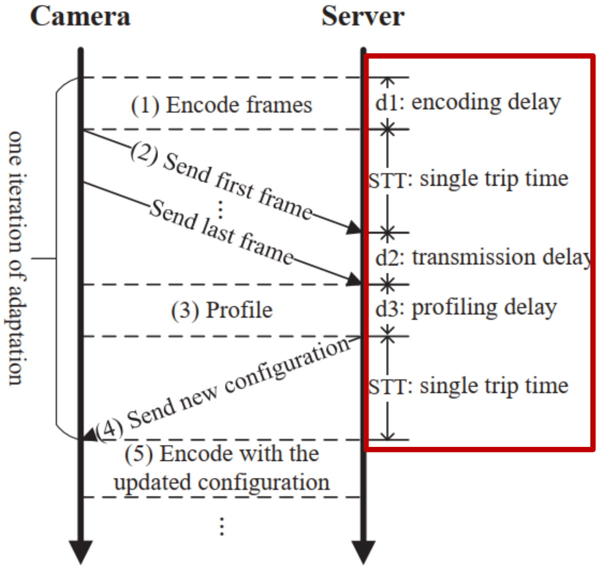
存在响应延迟严重影响推理的准确性

系统模型

- Delay Analysis

符号	定义
$batch\_t_i$	第i个batch占用的时间
$batch\_p_i$	做自适应决策的时间
$c_i = \{c_i^r, c_i^f, c_i^b\}$	传输视频设置
$g(c_i, batch\_t_i)$	编码时间
$f(c_i, batch\_t_i)$	一个batch传输的数据量
C_i	对于batchi服务器帧处理速度
B_i	batchi传输的可用带宽





$$d_i^1 = g(c_i, batch\_t_i).$$

$$d_i^2 = f(c_i, batch\_t_i)/B_i.$$

$$d_i^3 = c_i^f \cdot batch\_p_i/C_i.$$

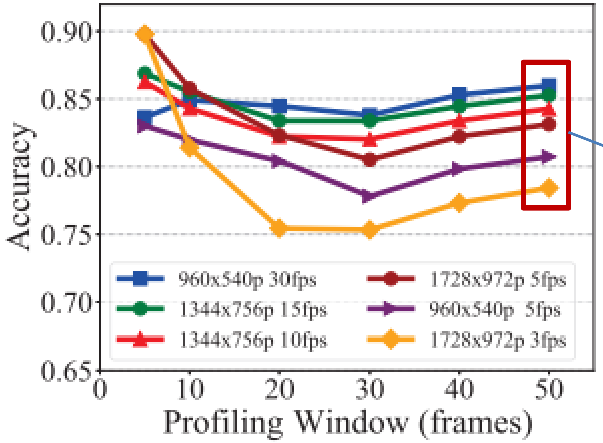
$$d_i = d_i^1 + d_i^2 + d_i^3 + RTT.$$

系统模型

- Optimization problem

$$\mathbb{C} = \text{Profile}(c_i^f, batch\_p_i)$$

帕累托最优集：不存在除了集合内的配置消耗更低的带宽获得更高的准确率



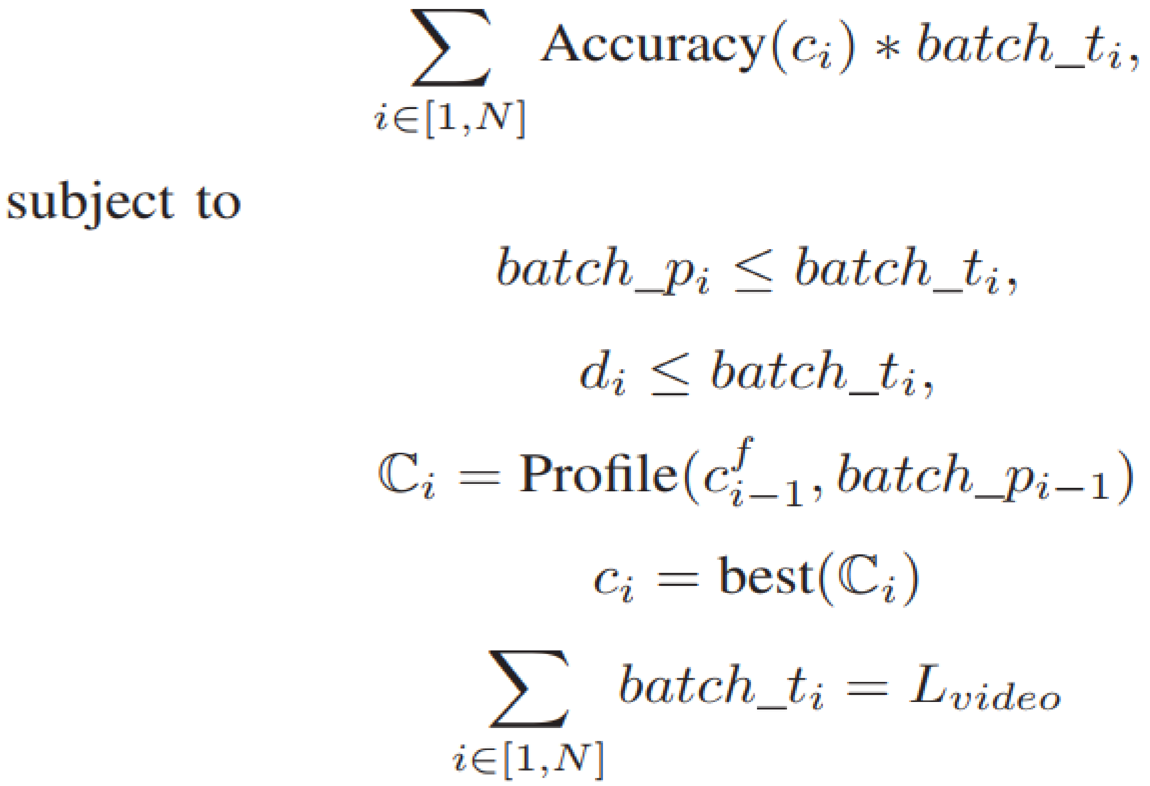
找到最优解c’

$$\text{Accuracy}(c_i)$$

反映设置为c_i时推断的准确率

系统模型

- Optimization problem



$$\sum_{i\in[1,N]} \text{Accuracy}(c_i) * batch\_t_i,$$

subject to

$$batch\_p_i \leq batch\_t_i,$$

$$d_i \leq batch\_t_i,$$

$$\mathbb{C}_i = \text{Profile}(c^f_{i-1}, batch\_p_{i-1})$$

$$c_i = \text{best}(\mathbb{C}_i)$$

$$\sum_{i\in[1,N]} batch\_t_i = L_{video}$$

选择最佳的配置**batch_p**以及总共**batch_t**使得总体准确率最大

系统模型

- Optimization problem

$$\sum_{i\in[1,N]} \text{Accuracy}(c_i) * batch\_t_i,$$

subject to

$$batch\_p_i \leq batch\_t_i,$$

$$d_i \leq batch\_t_i,$$

$$\mathbb{C}_i = \text{Profile}(c^f_{i-1}, batch\_p_{i-1})$$

$$c_i = \text{best}(\mathbb{C}_i)$$

$$\sum_{i\in[1,N]} batch\_t_i = L_{video}$$

1) 搜索空间过大

2) 两个相邻batch之间有联系

3) 实时准确率难以计算

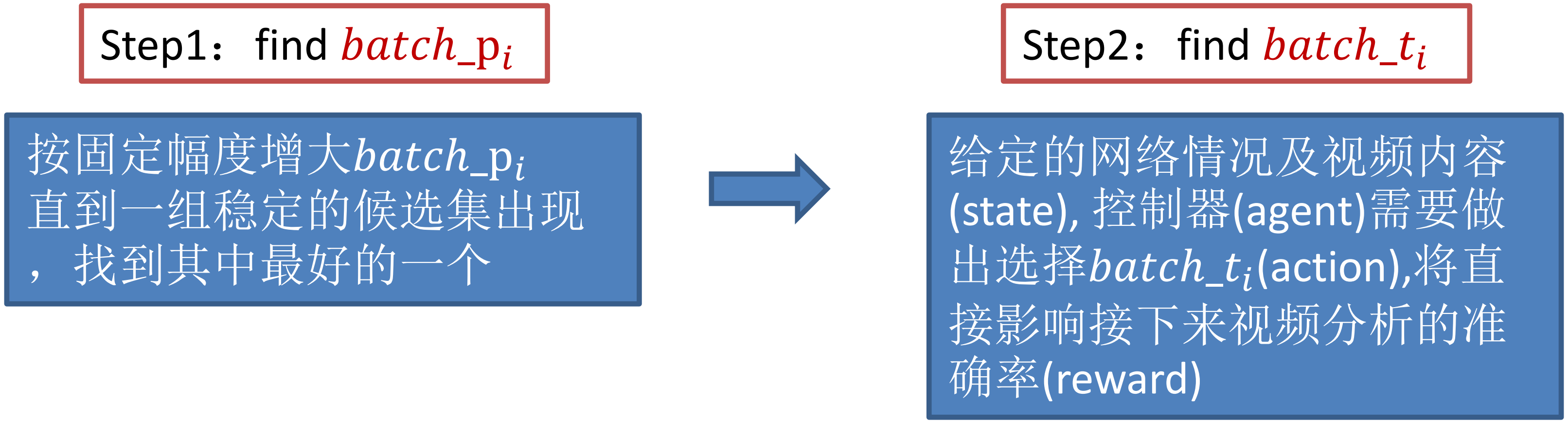
↓

问题难以解决

解决方法

- 总体思路

A two-step solution



Step1： find $batch\_p_i$

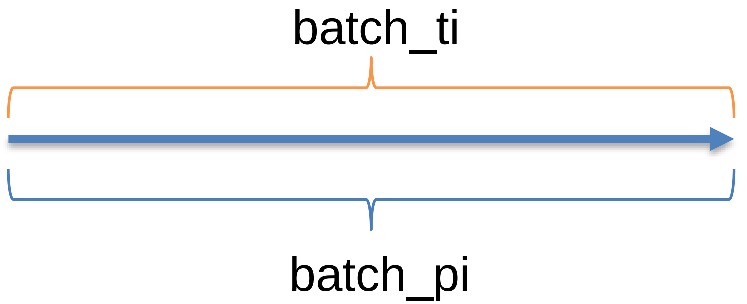
按固定幅度增大$batch_p_i$直到一组稳定的候选集出现，找到其中最好的一个

Step2： find $batch\_t_i$

给定的网络情况及视频内容(state), 控制器(agent)需要做出选择$batch_t_i$(action),将直接影响接下来视频分析的准确率(reward)

解决方法

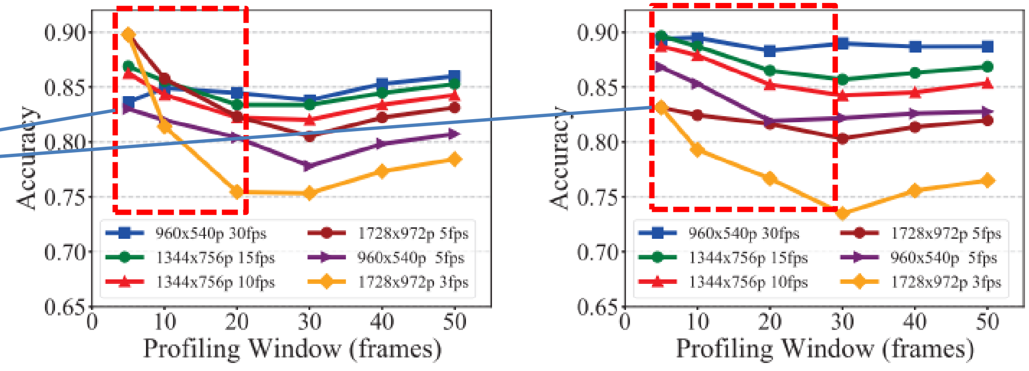
- Profiling with Early Quit



batch_pi = batch_ti ✗

配置窗口的前几帧变化较大

随着窗口增大变化逐渐平缓，并且前几个高准确率的配置方法的排名不变



(a) Video 1 (SurveillanceCams)

(b) Video 2 (DashCams)

解决方法

- Model for Batch Size Selection（DRL）

Agent Profiler

State Space $s_i = \{\overrightarrow{c_i}, \overrightarrow{B_i}, \overrightarrow{batch\_t_i}\}$

Action Space

batch size原本是一个连续值，这里将其离散化（间隔0.1s）

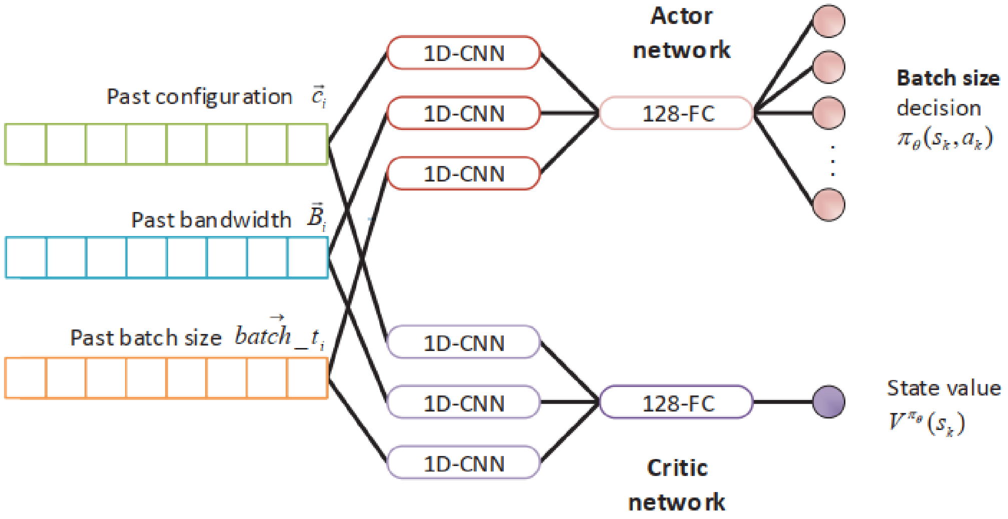
Reward Setting $r_i = batch\_t_i * Accuracy(c_i)$

Learning Architecture Actor-critic architecture

解决方法

■ Model for Batch Size Selection

Learning Architecture



Past configuration $\vec{c}_i$
Past bandwidth $\vec{B}_i$
Past batch size $\overrightarrow{batch\_t}_i$
1D-CNN
1D-CNN
1D-CNN
Actor network
128-FC
Batch size decision $\pi_\theta(s_k,a_k)$
1D-CNN
1D-CNN
1D-CNN
128-FC
Critic network
State value $V^{\pi_\theta}(s_k)$

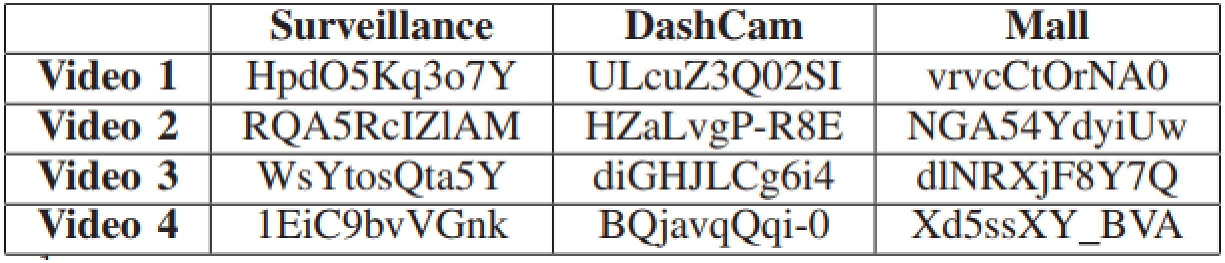
Critic network : $$\theta_v \leftarrow \theta_v - \beta \sum_k \nabla_{\theta_v} (R_k^{(n)} - V^{\pi_\theta}(s_k))^2$$

Actor network : $$\theta_a \leftarrow \theta_a + \alpha \nabla_\theta \log \pi_\theta(a_t|s_t)$$

实验结果

■ 实验设置

1) Data Sets:



	Surveillance	**DashCam**	**Mall**
Video 1	HpdO5Kq3o7Y	ULcuZ3Q02SI	vrvcCtOrNA0
Video 2	RQA5RcIZlAM	HZaLvgP-R8E	NGA54YdyiUw
Video 3	WsYtosQta5Y	diGHJLCg6i4	dlNRXjF8Y7Q
Video 4	1EiC9bvVGnk	BQjavqQqi-0	Xd5ssXY_BVA

2) Settings:

- object detection model： YOLOv3
- Server： Intel(R) Xeon(R) Gold 6152 CPU @ 2.10GHz and Tesla T4 GPU x 4
- Frame rate： 30fps
- batch searching size： 10frames
- Top-3 decide the profiling window

3) Methods for Comparison

- **Chameleon**
- **AWStream**

实验结果

- *Effectiveness of Profiling with Early Quit*

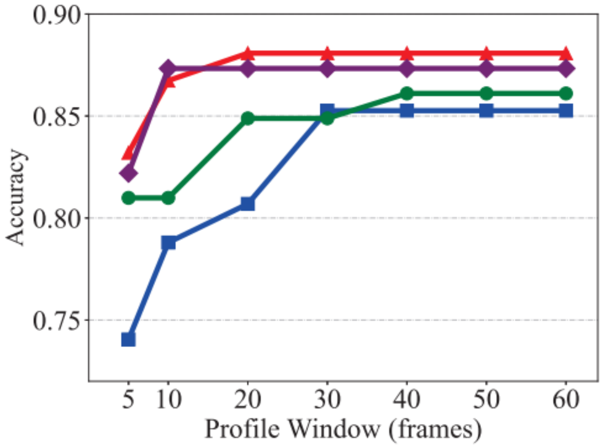


Fig. 8. Configuration's Goodness *vs.* Profile Window

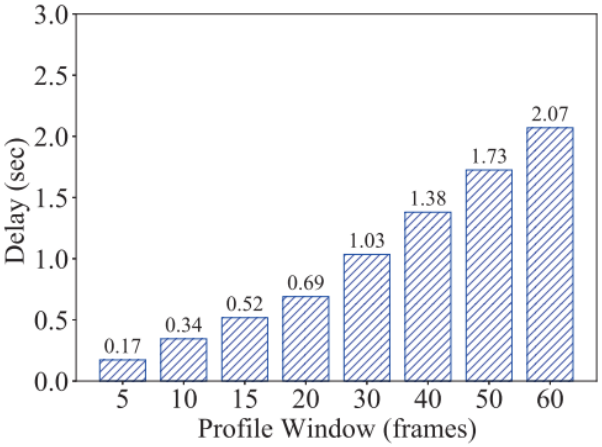


Fig. 9. Delay *vs.* Profile Window

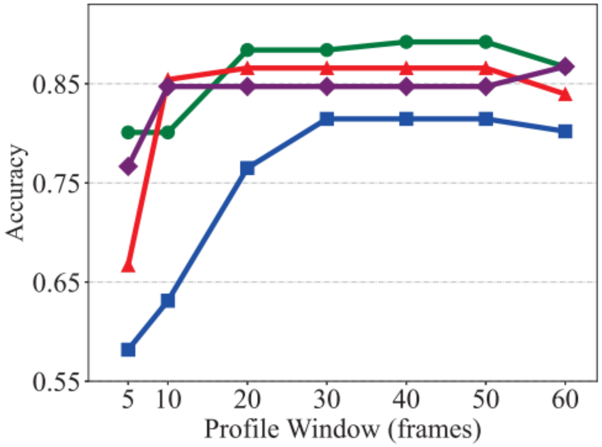


Fig. 10. Actual Accuracy *vs.* Profile Window

配置建议的质量首先提升很快，然后随着窗口增大不再变化

配置窗口越大，会导致更长的响应时间

实际的推断准确率随着窗口增大不断提高，然后保持不变，最后因为过长时延准确率下降

体现了早退出机制的优越性

实验结果

- *Accuracy–Bandwidth Trade-off*

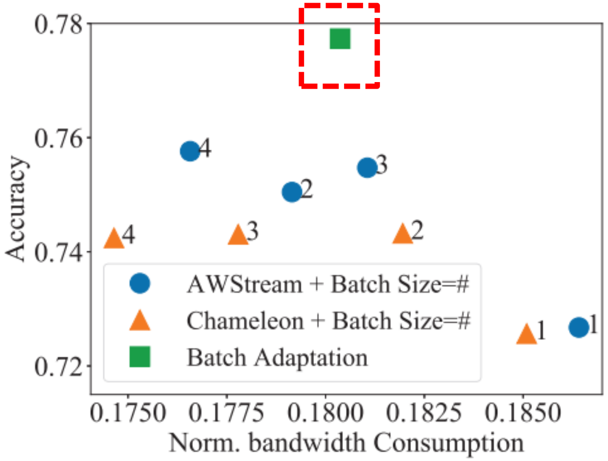


Fig. 11. Accuracy *vs.* Bandwidth for Surveillance Videos

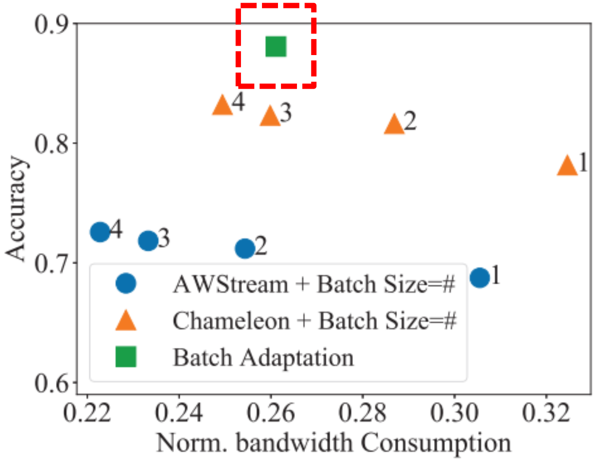


Fig. 12. Accuracy *vs.* Bandwidth for DashCam Videos

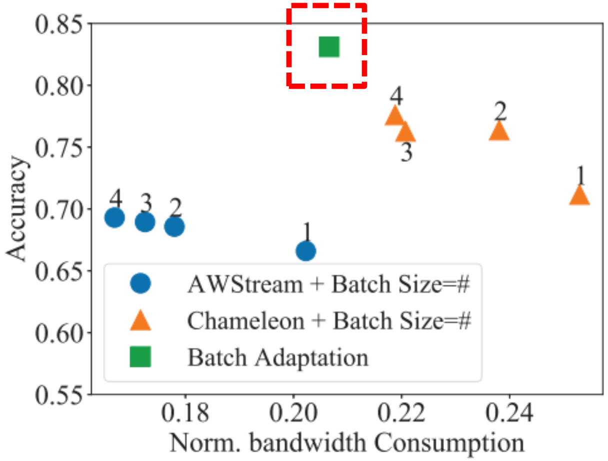


Fig. 13. Accuracy *vs.* Bandwidth for Mall Videos

☑ **准确率最高**

☑ **消耗的带宽适中**

实验结果

- *Response Delay of Server-Driven Streaming*

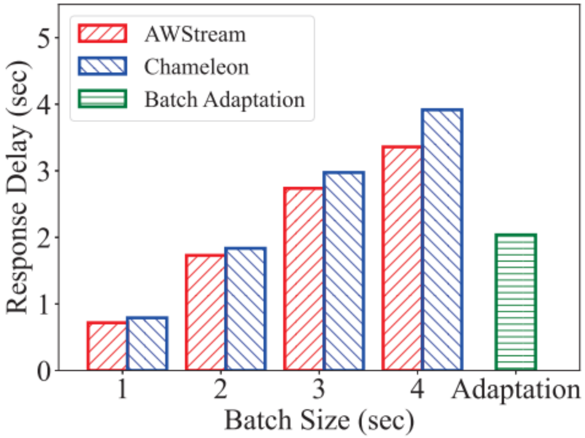


Fig. 14. Response Delay for Surveillance Videos

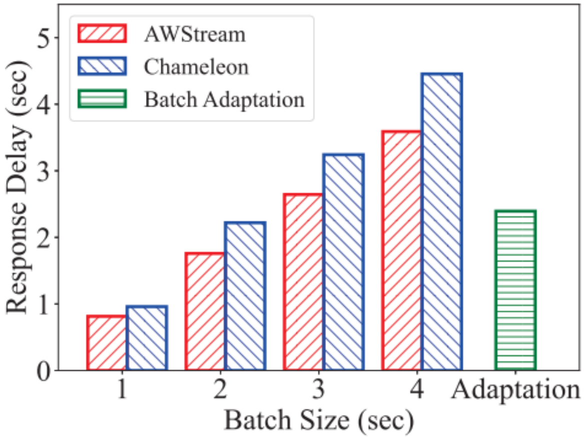


Fig. 15. Response Delay for DashCam Videos

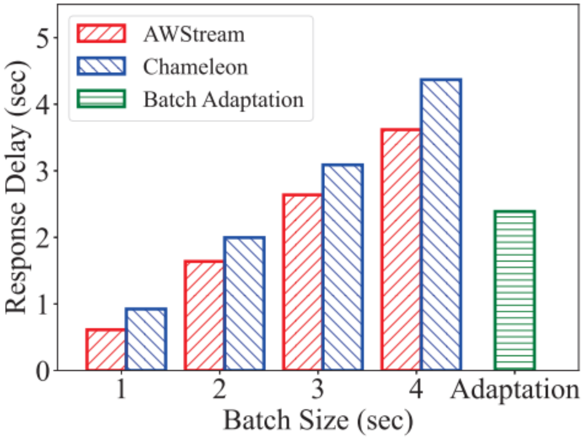
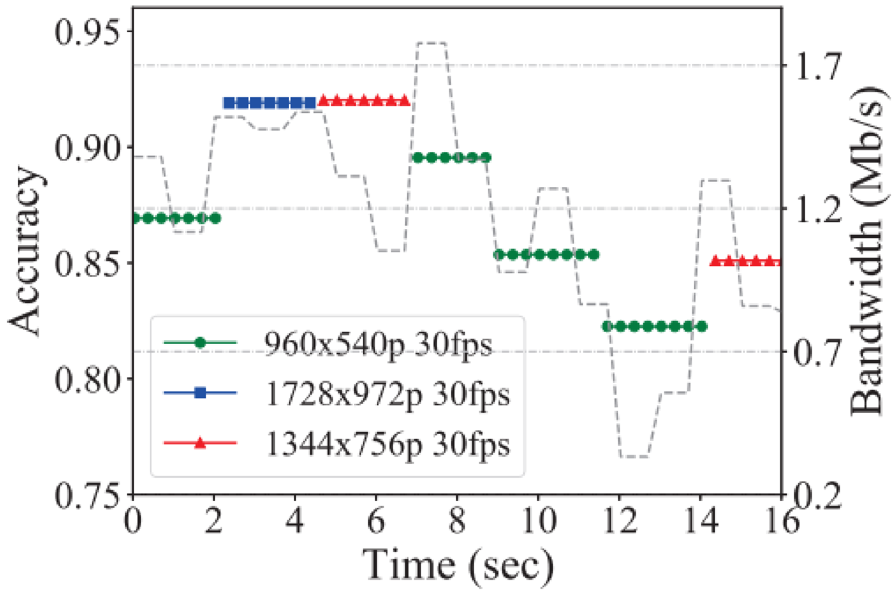


Fig. 16. Response Delay for Mall Videos

☑ **响应时延适中**

实验结果

- *Effectiveness of Batch Size Adaptation*



A snapshot trace of streaming adaptation in Video 1 of Type 2

- ✓ 视频配置随着网络波动以及视频内容发生变化
- ✓ 网络环境变差，通过调整batchsize准确率波动很小

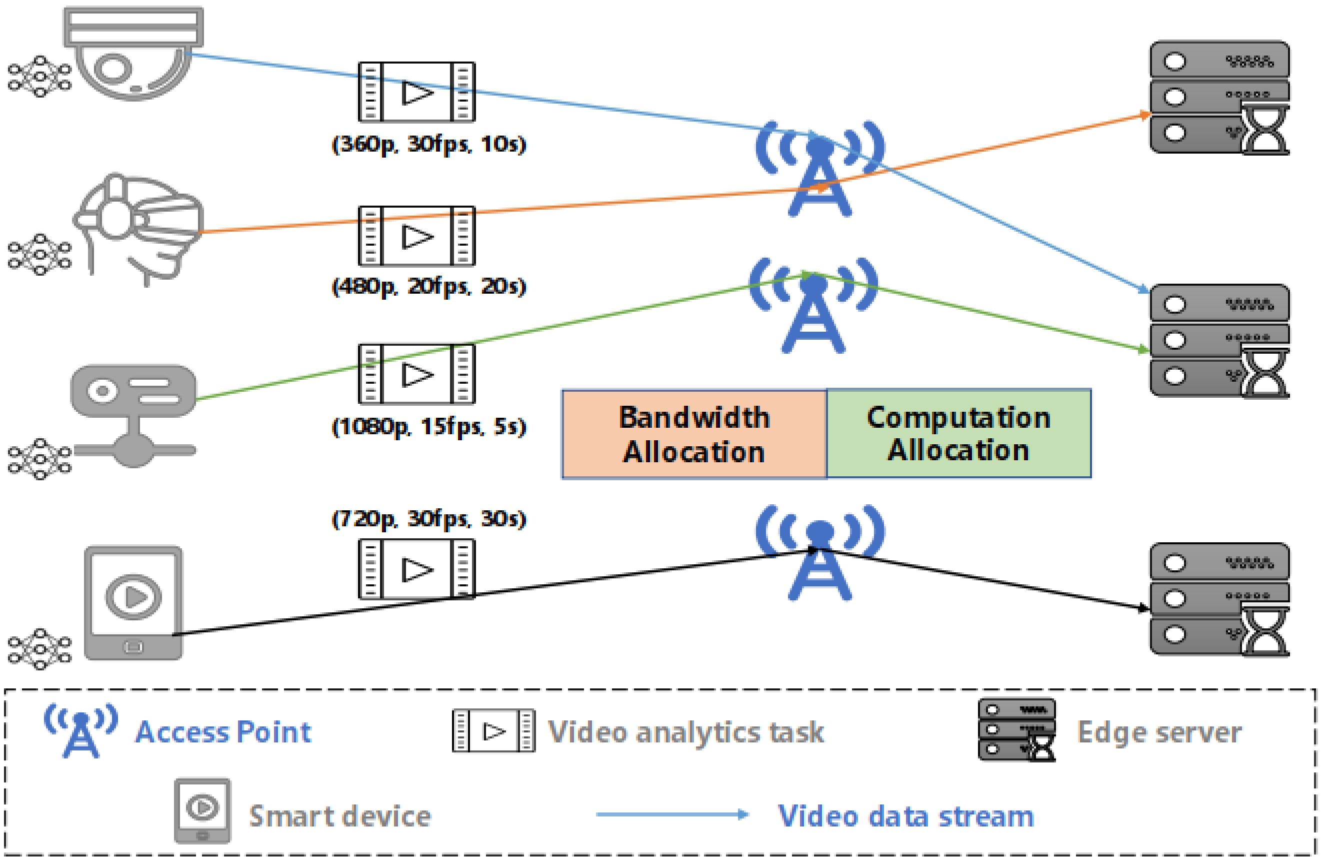
结论

- **优点**
  - 第一次抓住batchsize进行研究，探索其影响
  - 提出了一种提前退出机制，既减少响应延迟又提高了准确率

- **缺点**
  - 理论性不够，很多方法缺乏理论证明
  - 候选集方法没有详细说明
  - 深度强化学习的设置及说明过于草率

研究进展

- 时间敏感视频分析任务卸载 （cscwd2023）



研究进展

$$T_{i,j} = (1 - x_{i,j})\frac{\Gamma(r_i)f_i l_i}{C_{loc}^{i}} + x_{i,j}[\underbrace{\frac{\alpha r_i^2 f_i l_i}{B_j b_{i,j}}}_{\text{通信时延}} + \underbrace{\frac{\Gamma(r_i)f_i l_i}{C_j c_{i,j}}}_{\text{计算时延}}]$$

$$\boxed{\min \frac{1}{N}\sum_{i=1}^{|\mathcal{U}|}\sum_{j=1}^{|\mathcal{S}|} T_{i,j}}$$

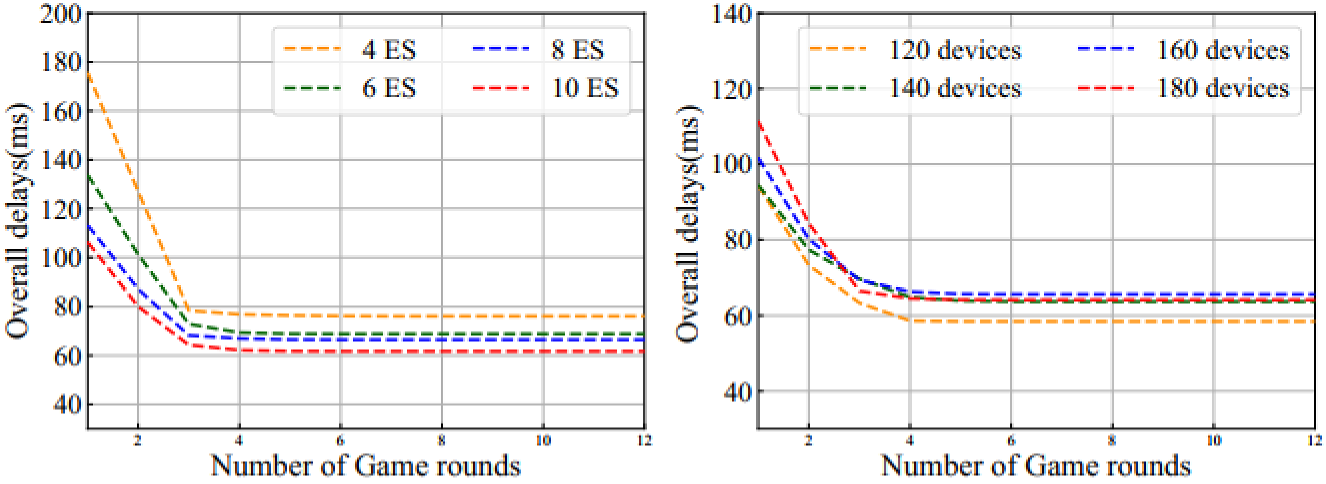
$$\text{s.t.} \quad \sum_{i=1}^{n} b_{i,j} = 1, j \in \{1, 2, \cdots, m\}$$

$$\sum_{i=1}^{n} c_{i,j} = 1, j \in \{1, 2, \cdots, m\}$$

$$x_{i,j} \in \{0, 1\}, \quad b_{i,j} \geqslant 0, \quad c_{i,j} \geqslant 0$$

研究进展

■ **实验结果**



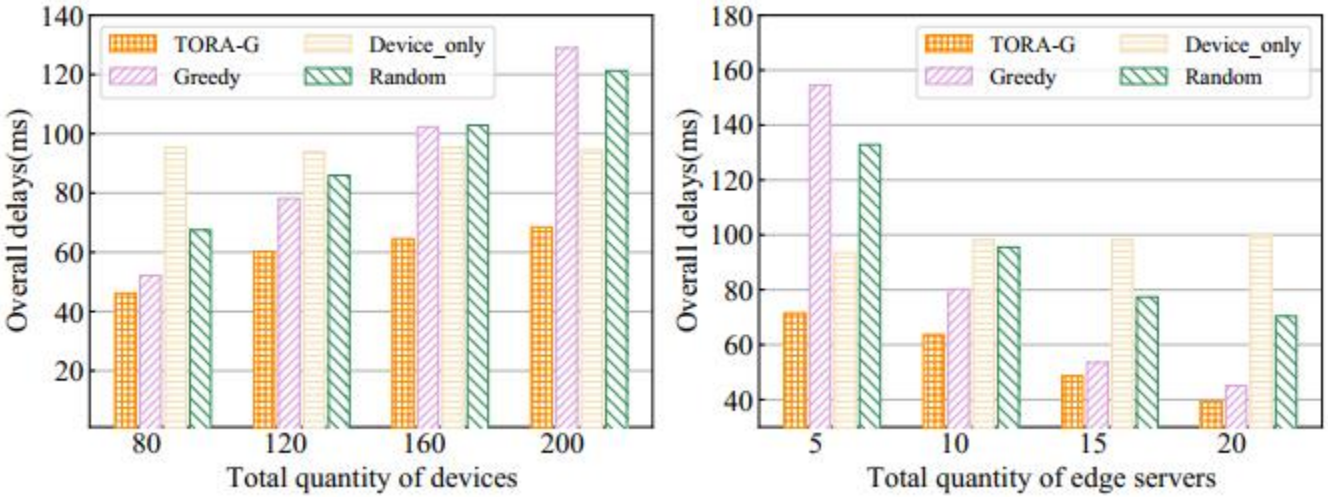
(a) Nash Equlibrium : $\mathcal{N} = 150$ (b) Nash Equlibrium : $\mathcal{M} = 10$

Fig. 4: Results for Nash Equilibrium.

经过及轮迭代，系统达到纳什均衡

研究进展

■ **实验结果**

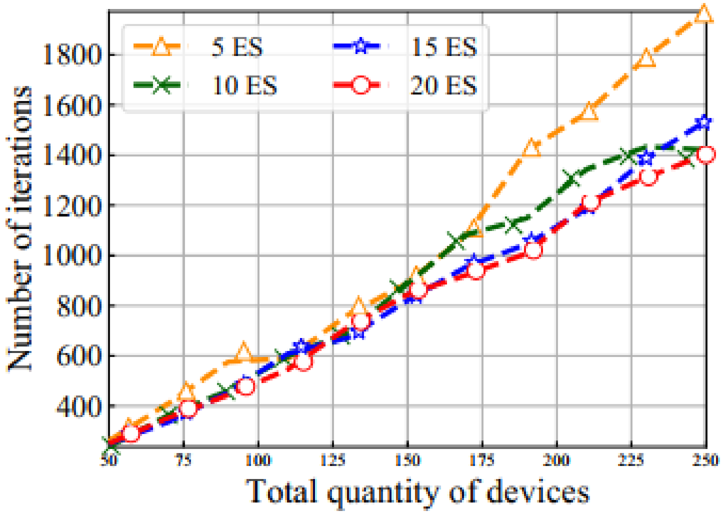


(a) Comparison when $\mathcal{M} = 10$ (b) Comparison when $\mathcal{N} = 120$

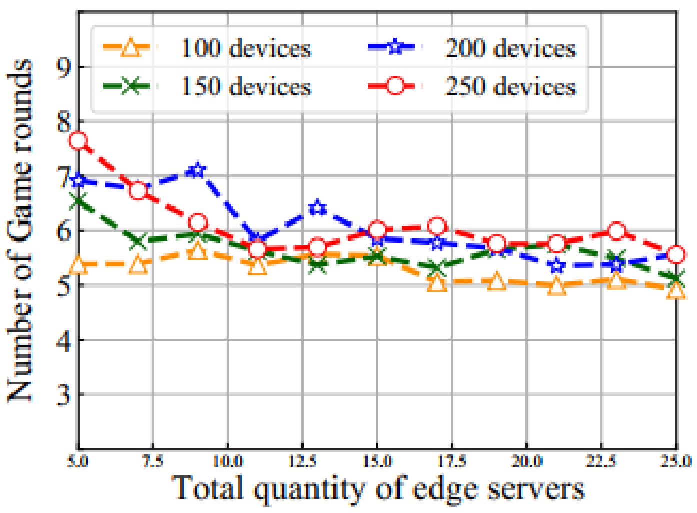
分别固定device数量和service数量，TORA-G效果最好

研究进展

- **实验结果**



(a) Iterations' linear variation



(b) Game rounds' nearly constant

迭代轮数随终端数量几乎线性增长

博弈回合数几乎为常数级别，收敛很快

算法效率很高

研究进展

后续研究计划

研究点2：视频分析任务的卸载决策优化问题

改进1：

实验由仿真改成真实系统的实验

改进2：

离线的优化改成基于DQN的在线方法

研究点1：模型层/ 应用层面的加速优化

调研学习中…

Thank You !

2022.12.28