CS 181AI
Lecture 15

Final Project Review

Arthi Padmanabhan

Mar 8 2023

Logistics

- Project survey (~10 min) out Wednesday after class, due **Friday**
- Week after spring break: one day will be working session for project proposal (due 3/27)
  - Will post a template on course webpage by Friday
- No paper reading the week after spring break (we'll cover a paper on Monday but I'll do it)

Project: Rough Timeline

- Total: 6 weeks (including writing final report)
- 2 weeks: Set up and motivation
  - Description of the problem and why it's important
  - What would be the simplest solution? **Implement the simplest solution** and show that it's not ideal. Collect results to show this and add figures to final report
- 3 weeks: Your idea
  - **Implement your idea** and compare it against the simple solution. Collect results to compare and add figures to final report
- 1 week: Write it up!
  - Similar to proposal, template will be on course webpage when this gets closer

Expectations

- You convince the reader that there is a problem and the simple solution is insufficient
- You are resourceful in finding ways to implement what you needed (as necessary, learned about different libraries, read blog posts, etc.)
- You run the necessary experiments to thoroughly compare your idea against the simple solution and presented your data clearly
- NOT an expectation: your solution works as expected and is better
  - However, you are expected to explain why you think your solution did or did not work

Project Days in Class

- The following dates will be working sessions. I will check in with each group
  - 3/22
  - 4/12
  - 4/19
  - 4/17 pending (I might be away in which case it will be a working day)

Allow User to Specify Resource Limits

Computational cost: $2000

Accuracy: >= 90%

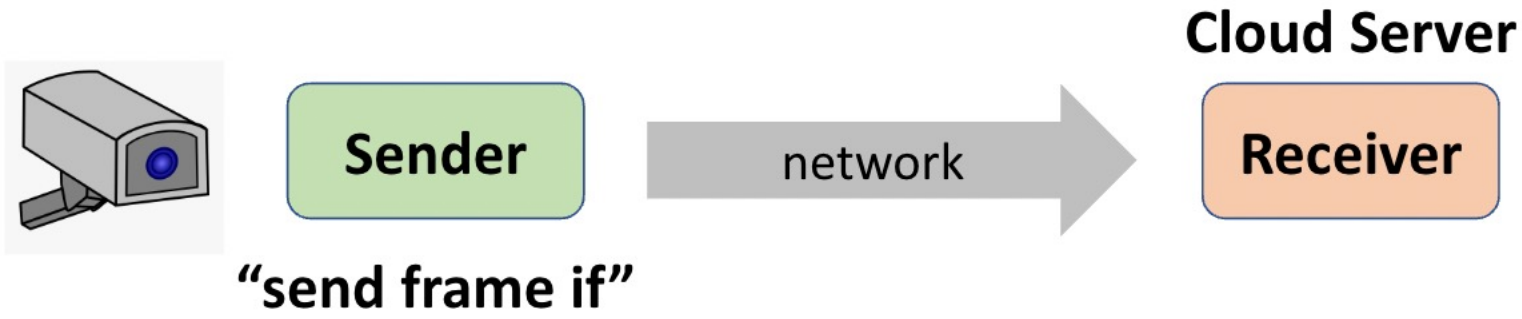
Memory: <4GB

Energy?

Here is the model that best suits your needs

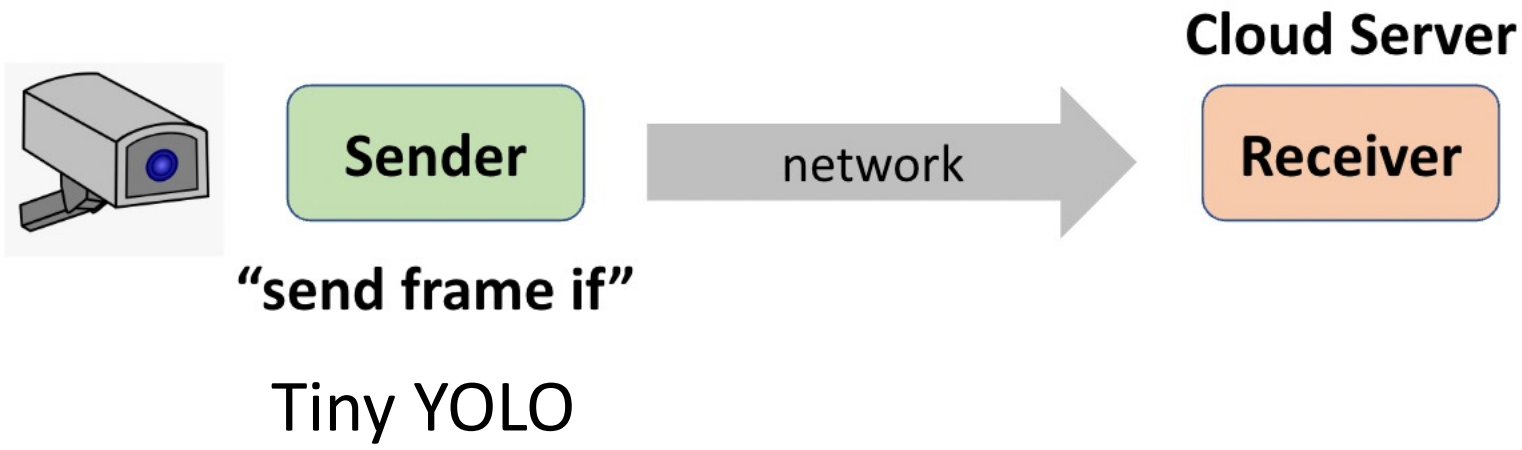
Filtering

- Develop your own frame filtering system to be run at the edge to lower the bandwidth of sending all frames to the cloud



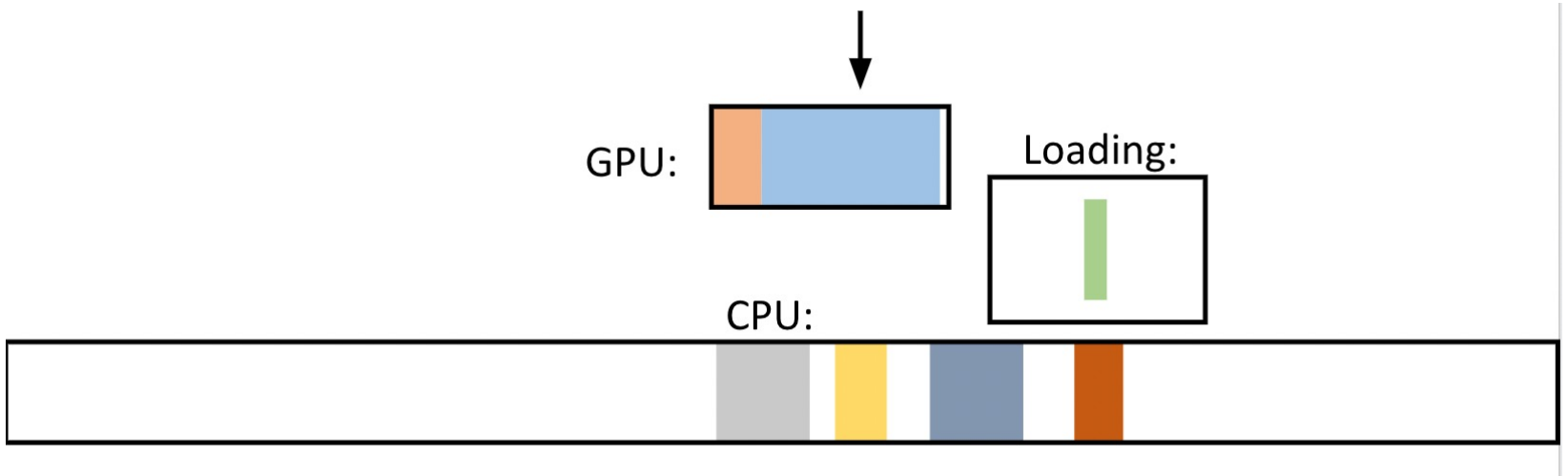
Using a Small Model at the Edge

- If your filtering system was a small version of a model (Tiny YOLO), how would you use the results of that to reduce *total* computation?



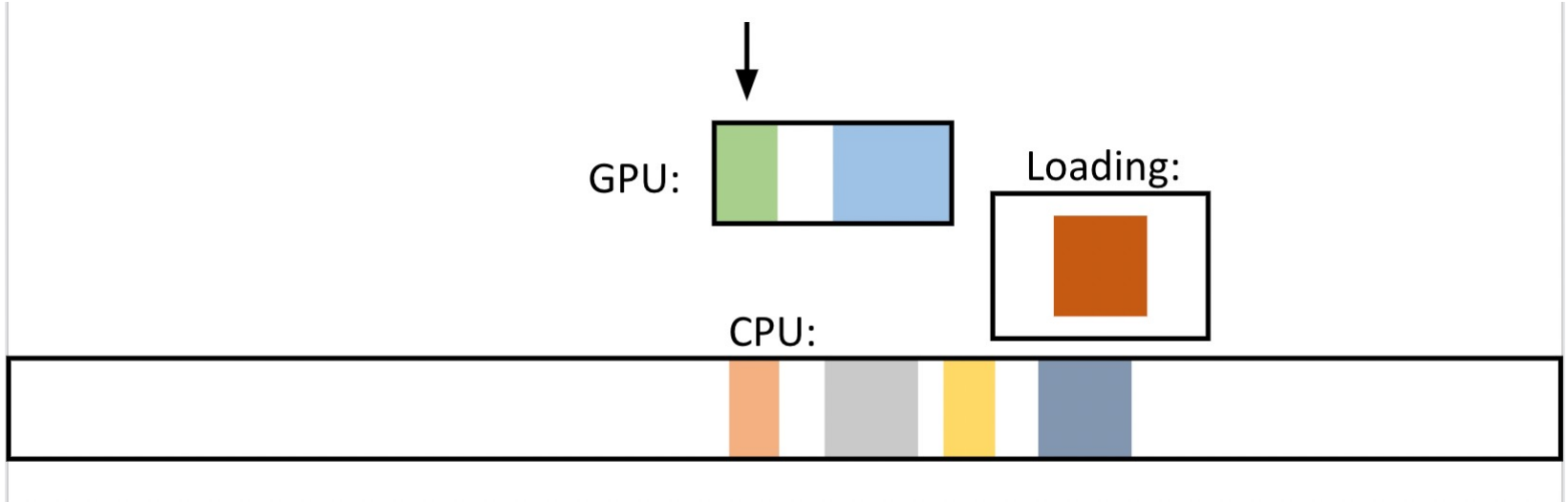
Scheduling Algorithm for Running ML Workload

- Goal: find a way to schedule ML jobs that achieves high throughput and fairness (or low latency)
- Consider: memory to load and run each type of model at several possible batch sizes, time to load and run each type of model at several possible batch size, memory capacity, potentially SLAs (service level agreements about latency)



Scheduling Algorithm for Running ML Workload

- Goal: find a way to schedule ML jobs that achieves high throughput and fairness (or low latency)
- Consider: memory to load and run each type of model at several possible batch sizes, time to load and run each type of model at several possible batch size, memory capacity, potentially SLAs (service level agreements about latency)



Assess Energy Usage

- Try to implement Monday's paper's method for assessing energy usage and do a study on which types of layers use the most energy across a variety of models

Chameleon w/ triggered profiling

- Chameleon (last Monday's paper) profiled to the get the best configuration (frame rate, resolution, etc) but their profiling was done periodically
- Can you find cues in the video that would imply that profiling should happen again (i.e., what aspects of the video make it so that the best configuration changes?)

Dashcam footage

- Take some existing filtering mechanism (e.g., from your response to Assignment 4) and compare its performance between still and mobile cameras (dashcams)
- Give an assessment of how to improve performance and reduce noise so more frames can be filtered on a dash cam?

Exploration of Object Trackers

- So far we have looked largely at classifiers and object detectors
- Explore object trackers and ways of improving the performance and resource usage tradeoff there

Proposing Your Own

- You may propose your own project (ideally with other students)
- The goal should be focused on a performance/resource usage tradeoff (e.g., “we made this process run faster in this scenario”), not pure ML (e.g., “we trained a model to do this task”, or “we made a model more accurate”)
- You are welcome to move outside the CV space with regards to the models you choose, but I will be far less able to help you ☺

Have a Good Break!