Low-Bandwidth Self-Improving Transmission of Rare Training Data

Shilpa George, Haithem Turki, Ziqiang Feng, Deva Ramanan,

Padmanabhan Pillai[†], Mahadev Satyanarayanan

Carnegie Mellon University and [†]Intel Labs

© 2023 M. Satyanarayanan

Many things in ML simplified if you already have a good training set

But what if you are trying to assemble that training set?

from data only found in remote and inaccessible places?

of a new, rare event?

© 2023 M. Satyanarayanan

Extreme Sensor to Backhaul Mismatch

4K Video Camera → 30 Mbps demand

future higher resolution, multispectral cameras will demand even higher bandwidths

Unmanned probes often have very poor wireless backhaul connectivity

- **deep space and inter-planetary networks** (10 – 100 kbps, 10^2 – 10^6 s one-way latency)
- **underwater acoustic networks** (10 – 100 kbps)
- **LoRa networks** (1 – 100 kbps)

Many exciting discoveries await us in these remote locations

© 2023 M. Satyanarayanan

A Perfect Storm

Convergence of three factors

- **Extreme mismatch of sensing vs transmission data rates**
  can’t blindly ship all data
- **Rare unlabeled events**
  < 0.1% of frames, possibly much rarer, can’t do random sampling
- **New phenomenon**
  no pre-built accurate detector/classifier, data is unlabeled

How to Retrieve Almost All Events Seen?
“Event” = True Positive (TP)

© 2023 M. Satyanarayanan

Our Solution: Live Learning

Interative human-in-the-loop workflow that combines

- **Semi-Supervised Learning (SSL)**
- **Active Learning**
- **Transfer Learning**

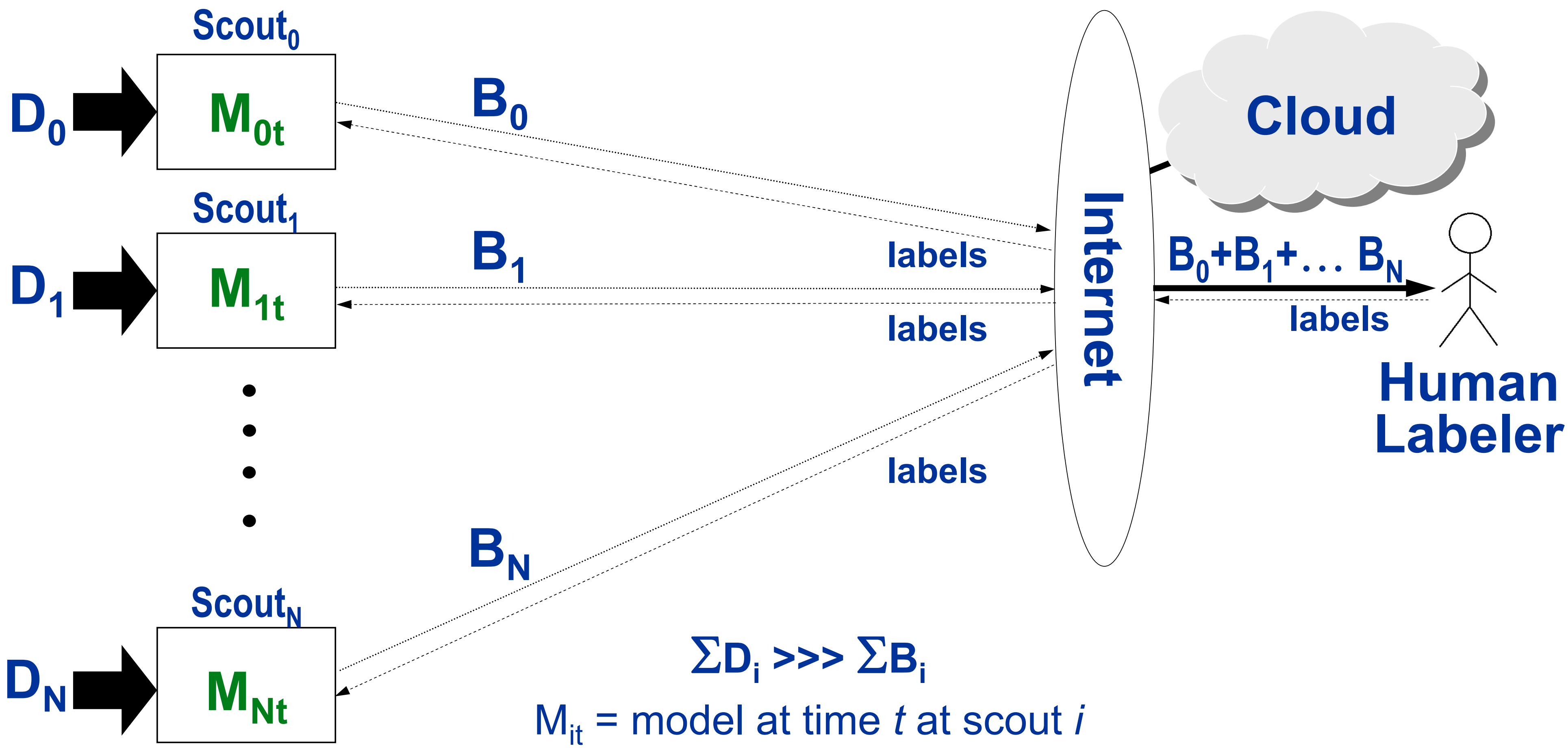
Pipeline sensing, inferencing, transmission, labeling, and training

Key steps in pipeline

1. **Bootstrap with weak initial model** (few-shot learning)
2. **Grow training set with newly-discovered TPs** (human confirmation for every TP)
3. **Train new model and replace current model asap** (cloud or edge training site, bandwidth-adaptive)
4. **Iterate steps 1–3 during mission**

© 2023 M. Satyanarayanan

Live Learning Overview



$\Sigma D_i >>> \Sigma B_i$

M_{it} = model at time *t* at scout *i*

© 2023 M. Satyanarayanan

Hawk: Open Source Implementation

https://github.com/cmusatyalab/hawk

Based on *ZeroMQ* delay-tolerant messaging

Completely model-agnostic (easy plugin of new DNN models)

Paper reports extensive investigations re detailed design choices

- Top-K vs MaxEnt selective transmission
- Hybrid SVM-DNN model evolution vs pure DNN evolution
- Revisit policy to collect missed positives
- Importance of tiling high-res data
- …

© 2023 M. Satyanarayanan

Experimental Evaluation

1. *In spite of extreme low bandwidth, can scouts discover most TPs encountered?*
2. *How close is Hawk to an ideal system?*
   - **Oracle   (perfect precision and recall)**
   - **BruteForce (imperfect precision and recall)**
     - model with the same architecture but trained in advance on fully labeled incoming data.
     - grossly overfitted to the data that will be seen during the mission.
     - requires all incoming data to be seen in advance, and transmitted to the cloud for labeling and training.
     - may not have perfect precision and recall.
3. *Can Hawk use additional bandwidth effectively?*
4. *Is Hawk DNN-agnostic?*

… many more questions …

© 2023 M. Satyanarayanan

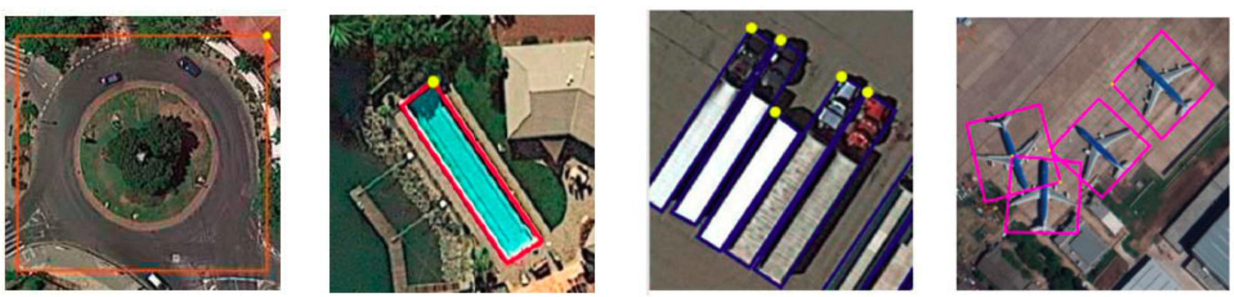
Dataset: Aerial Drone Surveillance

DOTA: Dataset for Object deTection in Aerial Images (published in 2018)

Consists of 2806 fully labeled images across 15 classes

Image Resolution: Ranges from 800x800 to 4000x4000

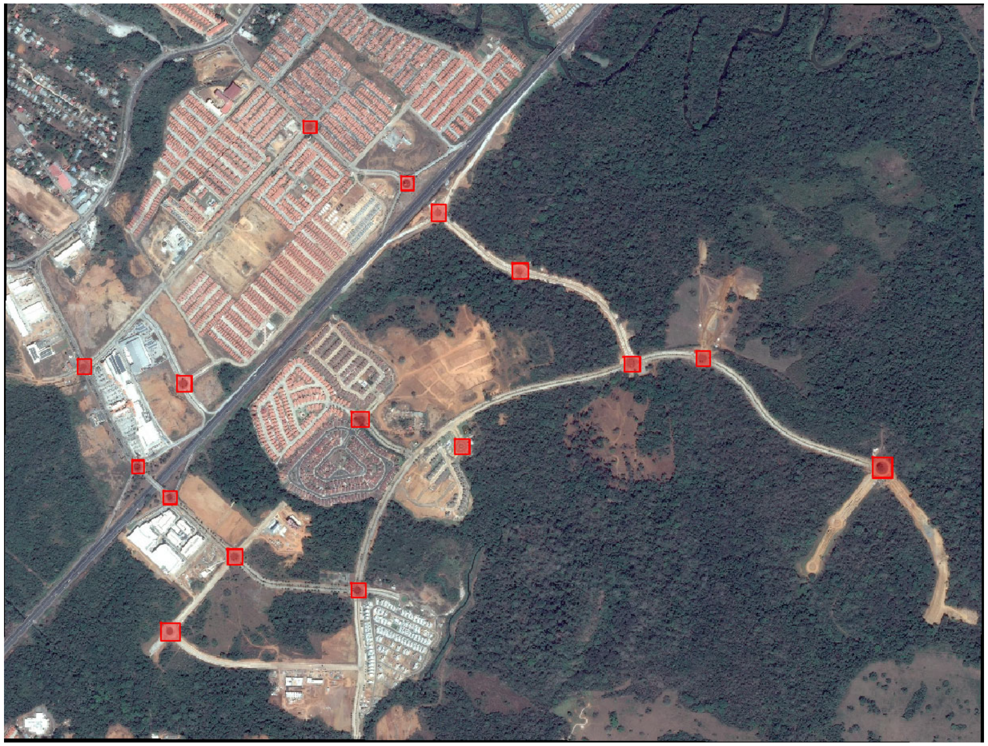
Derived dataset has 252231 labeled tiles having base rate of 0.1%



(a) Roundabout (TPs=336) (b) Swimming Pool (TPs=335) (c) Large Vehicle (TPs=357) (d) Airplane (TPs=350)

256x256 tiles from large 4K images

© 2023 M. Satyanarayanan



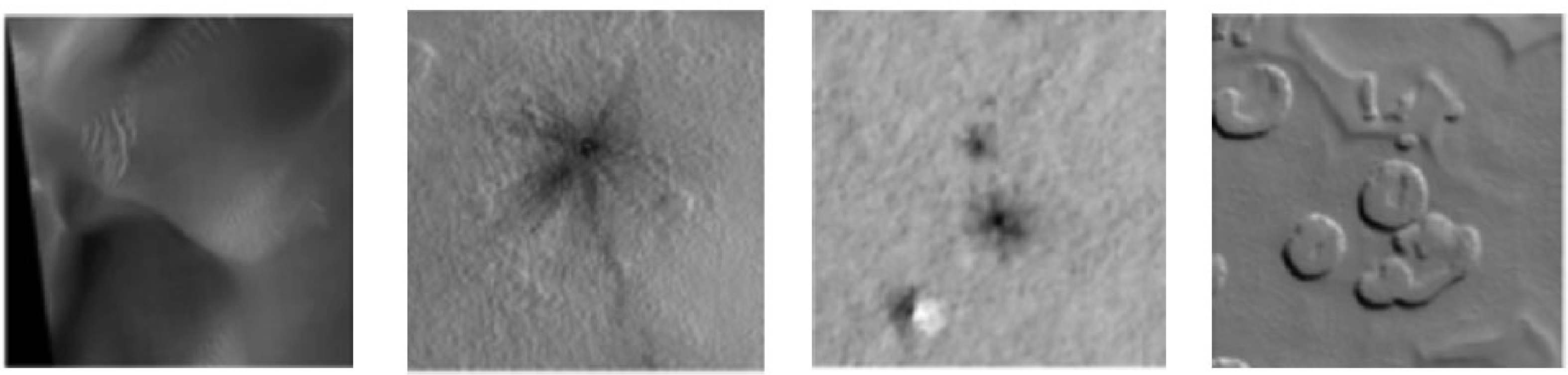
Dataset: Planetary Exploration

HiRISE: High Resolution Imaging Experiment from Mars (published in 2019)

Images collected by Mars Reconnaissance Orbiter

Dataset has 7 classes of landmarks on Martian terrain

Consists of 73,031 labeled images of size 227x227



(a) Dark Dune (TPs=64)
(b) Impact Ejecta (TPs=64)
(c) Spider (TPs=64)
(d) Swiss Cheese (TPs=64)

© 2023 M. Satyanarayanan

Dataset: Underwater Sensing

Brackish: Marine dataset (published in 2019)

Images of marine animals in a brackish strait with varying visibility

Consists of 14,518 labeled images of 1080p resolution

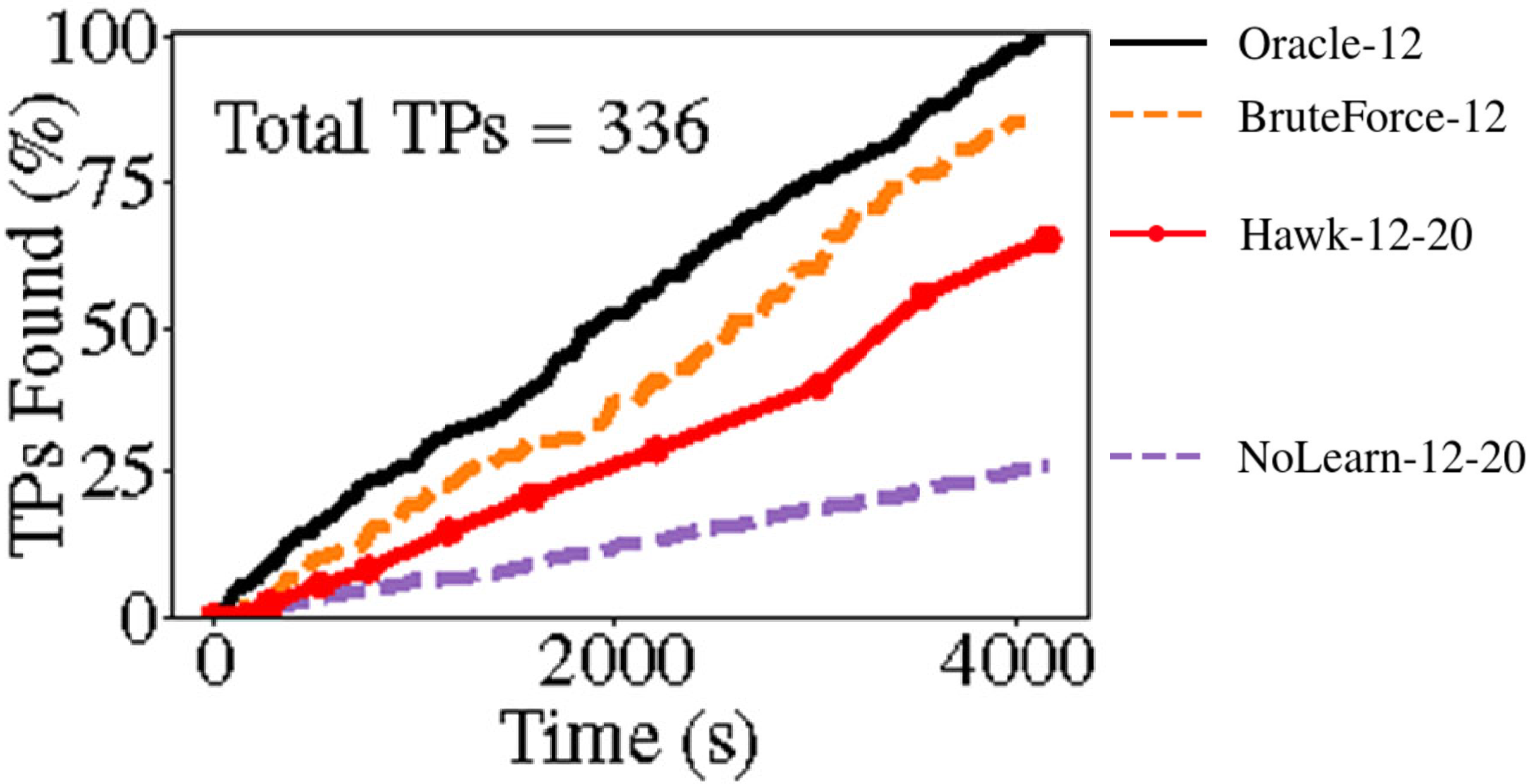
Derived dataset has 563,829 tiles across 6 classes with target baserate of 0.1%

(a) Starfish (TPs=370) (b) Shrimp (TPs=564) (c) Small Fish (TPs=564) (d) JellyFish (TPs=584)

© 2023 M. Satyanarayanan

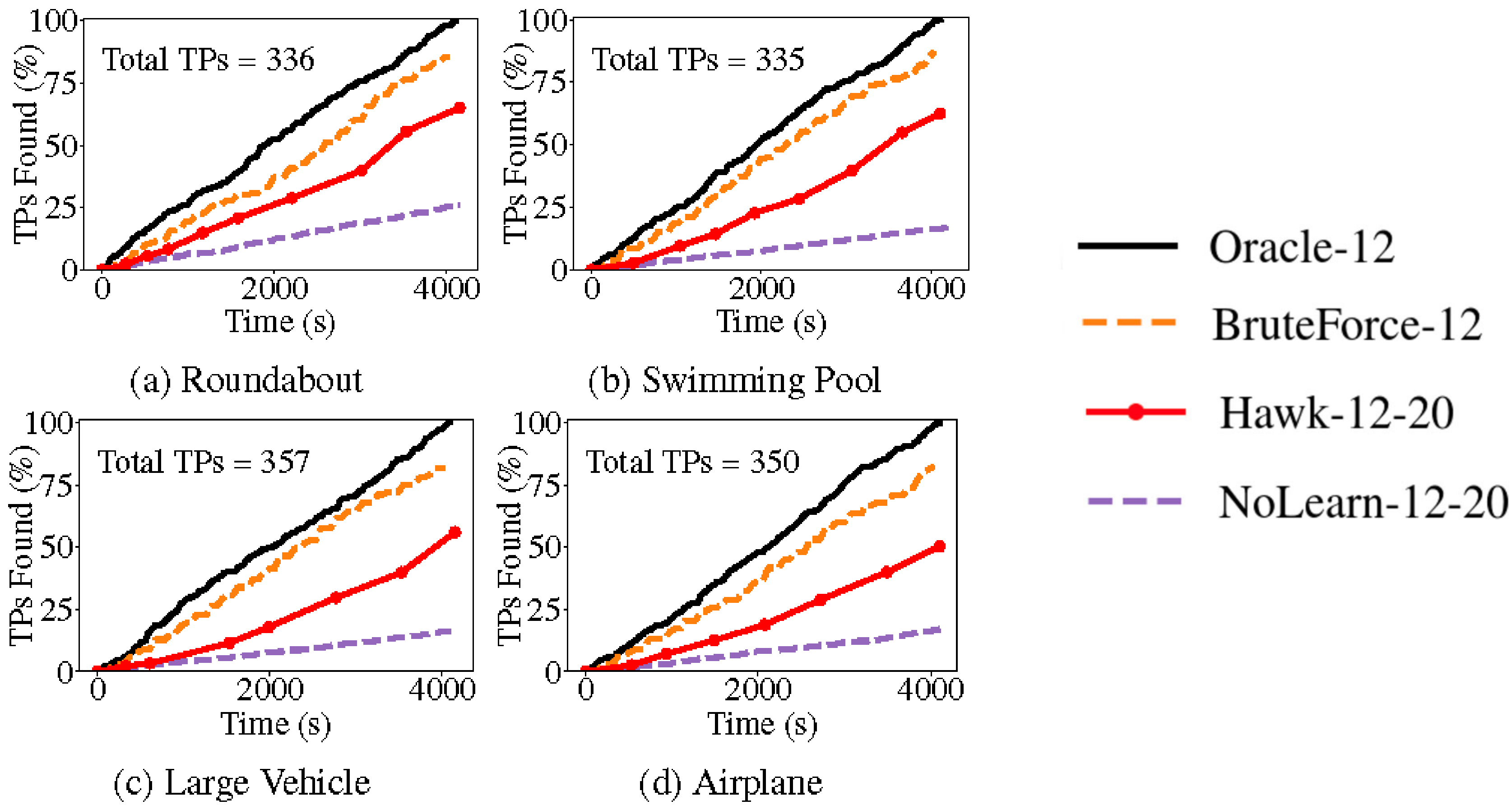
Scout-based Training - 12kbps (DOTA)

Class: Roundabout



© 2023 M. Satyanarayanan

Scout-based Training - 12kbps (DOTA)



See Paper for Many More Results

1. **Robust results across many datasets and classes** (aerial drone, Mars, underwater)
2. **Value of revisiting discard pile** (result caching of old scores)
3. **Live Learning is DNN agnostic** (ResNet-50, YOLOv4, ExtremeNet results)
4. **Ability to use higher bandwidth effectively** (12 kbps, 30 kbps, 100 kbps)
5. **Dynamic choice of cloud training versus scout training**
6. **Diversity Sampling to improve recall**
7. **Integration with Few-Shot Learning**

© 2023 M. Satyanarayanan

Take-Away Message

Gross bandwidth mismatch in remote sensing will grow worse

Live Learning is a viable solution to this problem

Key idea: *Integrate Learning with Selective Transmission & Human Labeling*

Hawk discovers up to 87% of the TPs discovered by BruteForce

Bonus: *Hawk also helps with limited human bandwidth*

© 2023 M. Satyanarayanan