A Vision-based Deep On-Device Intelligent Bus Stop Recognition System

Gautham Krishna Gudur
Ericsson R&D, India

Ateendra Ramesh
University at Buffalo

Srinivasan R
SSN College of Engineering, India

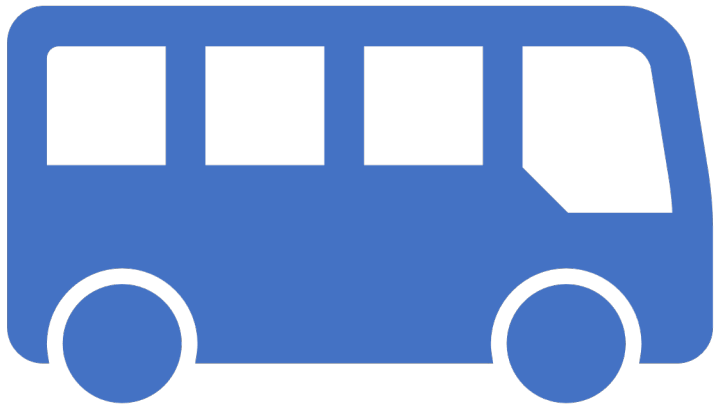
BUS STOP RECOGNITION

Intelligent Public Transportation Systems - cornerstone to any smart city, particularly ***Autonomous Vehicles***.

Over 47% of people use buses as preferred public transport mode in the United States.

~76% of buses have automatic bus stop announcements.

In India, buses used to take over 90% of public transport in Indian cities.



BUS STOP RECOGNITION IN INDIA

- Second largest road network in the world.
- Conventionally, bus conductor intimates (*whistles*) when a bus stop arrives and announces its location aloud.
- Driver halts the bus.
- New sophisticated buses consist of pre-defined queues – the sequence of bus stops are pre-loaded.
- Easier alternative to the conductor's manual announcement of bus stops.

CHALLENGES IN AUTOMATIC BUS STOP RECOGNITION

Really difficult to identify the arrival of a bus stop on-the-fly for buses to appropriately halt and notify its passengers.

Global Positioning System (GPS) look-up can be used for bus stops identification, however latency issues in the network.

Hard to localize, identify and halt the vehicle right in front of the bus stop using GPS.

CHALLENGES IN AUTOMATIC BUS STOP RECOGNITION IN INDIA

The landscapes and surroundings of bus stops dynamically change and evolve over time.

Rural and sub-urban Indian bus stops predominantly do not have bounded or localized spaces/lanes.

Bus routes are periodically revised and repurposed based on demand and traffic patterns.

Necessary to not only know the sequence of bus stops, but also intelligently perceive the location of bus stop in order to halt at right location.

GOALS OF THE PROPOSED SYSTEM

To employ novel vision-based techniques to recognize bus stops on-device and eliminate the latency of a GIS/GPS look-up

The model should be able to handle new and existing bus stops

Incremental Learning/Data Augmentation

- Existing Bus Stops
  - To incrementally handle the ever-changing surroundings of existing bus stops
- New Bus Stops
  - To incorporate new bus stops with minimal/no human intervention, without performance degradation due to class imbalance

To automatically handle real-time ground truthing/labeling of bus stops

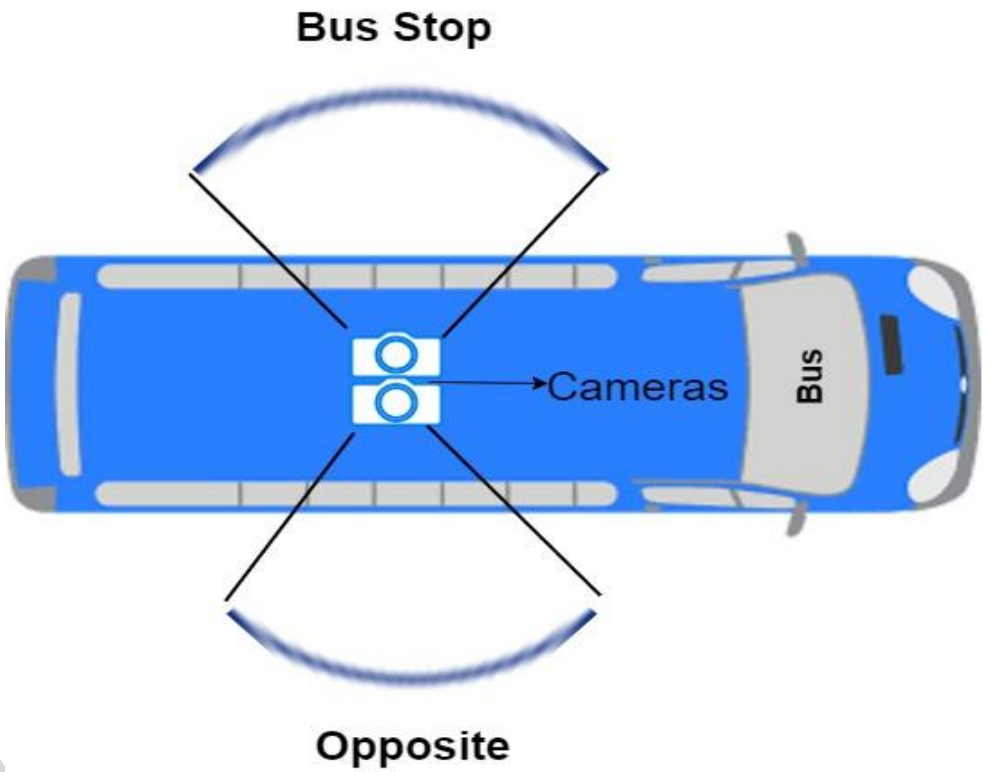
Crowdsourcing/Active Learning

REAL-TIME DESIGN AND INFERENCE

- Unnecessary overheads in capturing images during the whole route during classification of bus stop (inference).
- Hence, the images are captured and classified only when speed of the bus is below a certain ideal threshold.
- The real-time speed can be acquired from the speedometer of the bus.
- The ideal minimum threshold (10 km/hr for instance) is subject to locality and traffic conditions.
- We propose two different classifiers – day and night classifier (differentiated using a a light sensor).

DATASET

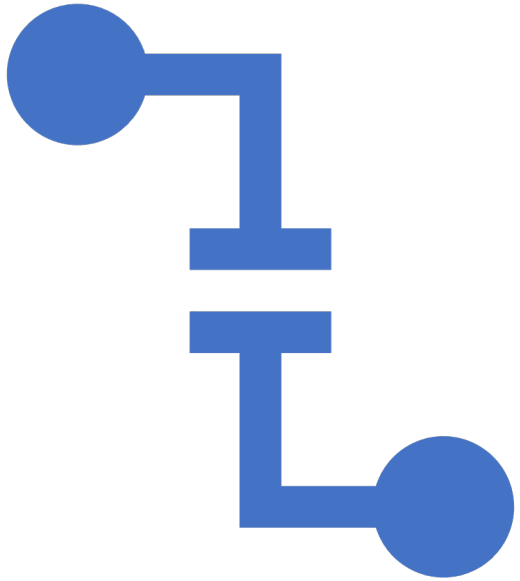
- The images of the bus stops were collected in and around the city of Chennai, India.
- Images were acquired using two 5 MP cameras placed in opposite directions.
- 8 bus stops during the day – 5 public urban and 3 rural bus stops.
- Images from 3 urban bus stops were also collected during the night.
- 90 images in each bus stop, 45 in each direction were collected.



EXAMPLES FROM THE DATASET

BAYESIAN CONVOLUTIONAL NEURAL NETWORKS

- CNNs are powerful mechanisms for distinctive spatial representations and offer automatic, effective feature learning capabilities.
- The acquired bus stops in real-time requires identification of discriminative features to handle evolving landscape changes.
- This essentially is a scene classification problem, which CNNs are known to effectively handle.
- Bayesian (Convolutional) Neural Networks are probability distributions (Gaussian priors) instead of point estimates, and this helps in modelling uncertainties.



MODELING UNCERTAINTIES USING DROPOUT

- **Dropout** - a stochastic regularization technique can perform approximate inference over a deep Gaussian process
- Learns the model posterior uncertainties **without high computational complexities** over few stochastic iterations at both train and test times
- Termed Monte-Carlo Dropout **(MC-Dropout)**
- Equivalent to performing Variational Inference
- $p(y*|x*,D_{train}) = \int p(y*|x*,\omega)\ p(\omega|D_{train})\ d\omega$

Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, ICML ‘16

MODELING UNCERTAINTIES USING DROPOUT

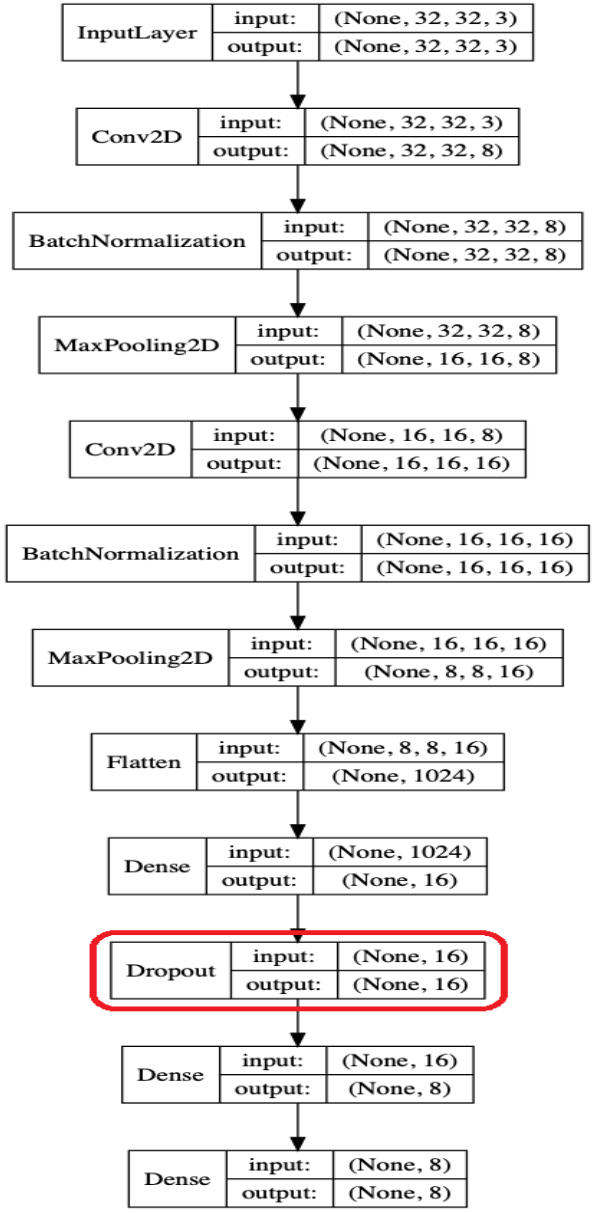
- **Dropout** - a stochastic regularization technique can perform approximate inference over a deep Gaussian process
- Learns the model posterior uncertainties **without high computational complexities** over few stochastic iterations at both train and test times
- Termed Monte-Carlo Dropout **(MC-Dropout)**
- Equivalent to performing Variational Inference
- $p(y*|x*,D_{train}) = \int p(y*|x*,\omega)\, p(\omega|D_{train})\, d\omega$

  Posterior

Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, ICML ‘16

Bayesian CNN Architecture

- The images were resized to 32 × 32 × 3 and normalized (divided RGB pixel values by 255 for easier model convergence).
- Utilize the CNN architecture, and treat it as a Bayesian Neural Net (with Dropout).
- We utilize 2 stacked CNN layers with BatchNorm and MaxPool layers between each layers, followed by two Fully-Connected Dense layers, with dropout of probability 0.3 between each FC layer.
- This is followed by a Linear Softmax layer, governed by the categorical cross-entropy loss.



INCREMENTAL LEARNING

- Used to update existing model with recent bus stop images which might have evolved with landscape changes.
- Will emphasize on learning the most recent and salient features of that bus stop.
- The inherent bias towards the model updated with the recently acquired images is favorable.

But, Why?

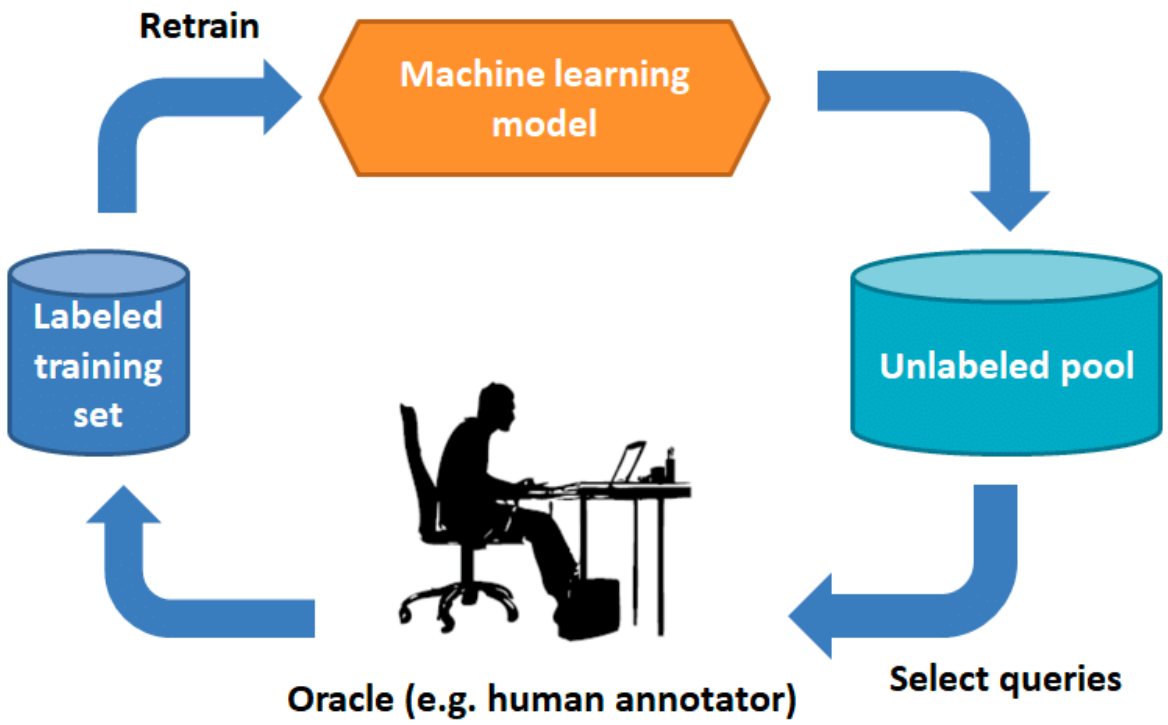
INCREMENTAL LEARNING

- Used to update existing model with recent bus stop images which might have evolved with landscape changes.
- Will emphasize on learning the most recent and salient features of that bus stop.
- The inherent bias towards the model updated with the recently acquired images is favorable.

But, Why?

- In a single bus stop, only the recently acquired data is actually sufficient to make accurate predictions, since the information from newer images are added periodically to the model.
- Hence, the memory and neural footprint of older images is not necessary.

ACTIVE LEARNING



- A big challenge in many real-time applications is obtaining labelled data.
- Active Learning/Crowdsourcing, over unsupervised techniques, is used predominantly to substantiate the confidence on the queried data points.
- Instead of labelling hundreds of bus stop images, an ideal system should query few labels in each bus stop.

ACQUISITION FUNCITONS

- Uncertainty measures from Bayesian CNN need to be quantified
- Arriving at most efficient set of data points (select k from n) to query from D_{pool}

ACQUISITION FUNCITONS

- *Max Entropy:* Maximize predictive entropy

$$H[y|x,D_{train}] := -\sum_c p(y = c|x,D_{train}) \log p(y = c|x,D_{train})\ c$$

- *BALD:* Maximise mutual information between predictions and model posterior

$$I[y, \omega|x, D_{train}] = H[y|x, D_{train}] - E_{p(\omega|Dtrain)} H[y|x, \omega]$$

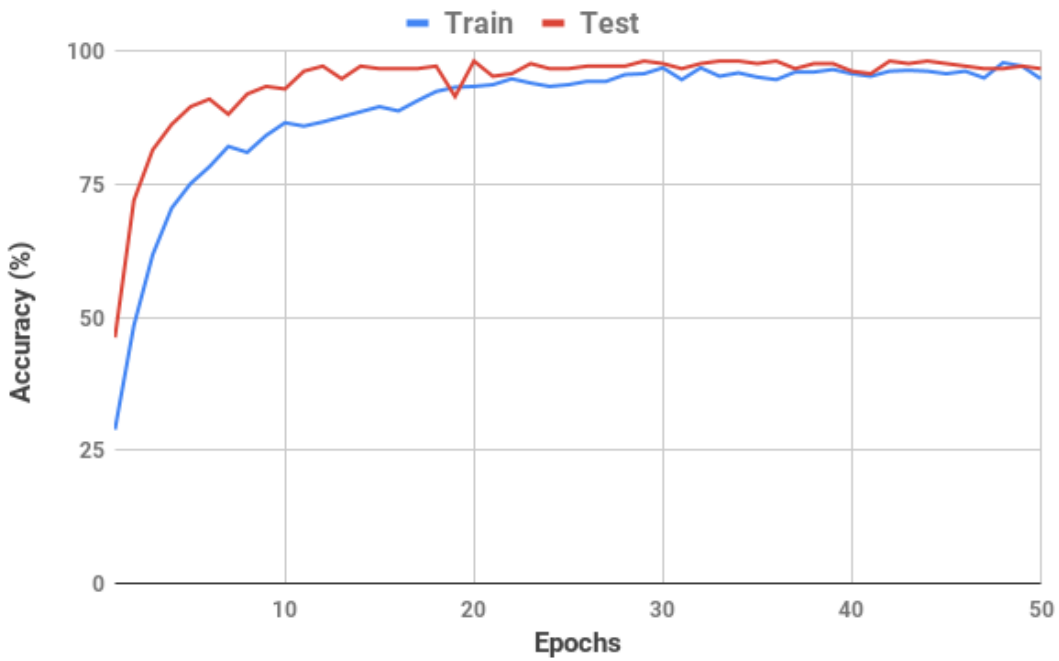
- Maximise *Variation Ratios*:

$$\text{variation-ratio}[x] := 1 - \max p(y|x, D_{train})\ y$$

- *Random Acquisitions:* Select data points from pool uniformly at random.

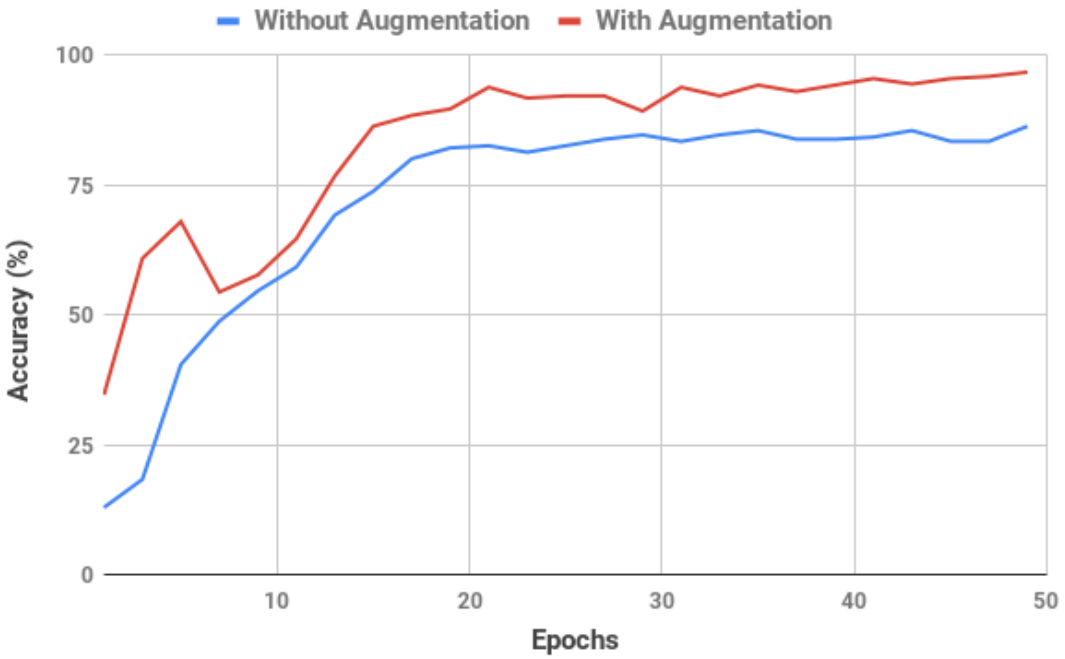
DATA AUGMENTATION

- When a new bus stop is added, it becomes a necessity for the model to scale and handle the incoming data of the new class.
- The acquired images from the new bus stop is predominantly lesser than that of existing classes which results in *class imbalance*.
- Techniques like zoom, shear and rotation (small angle) used for generating new images, which almost resemble the images acquired from a real-time camera to ensure stratified training across all classes.



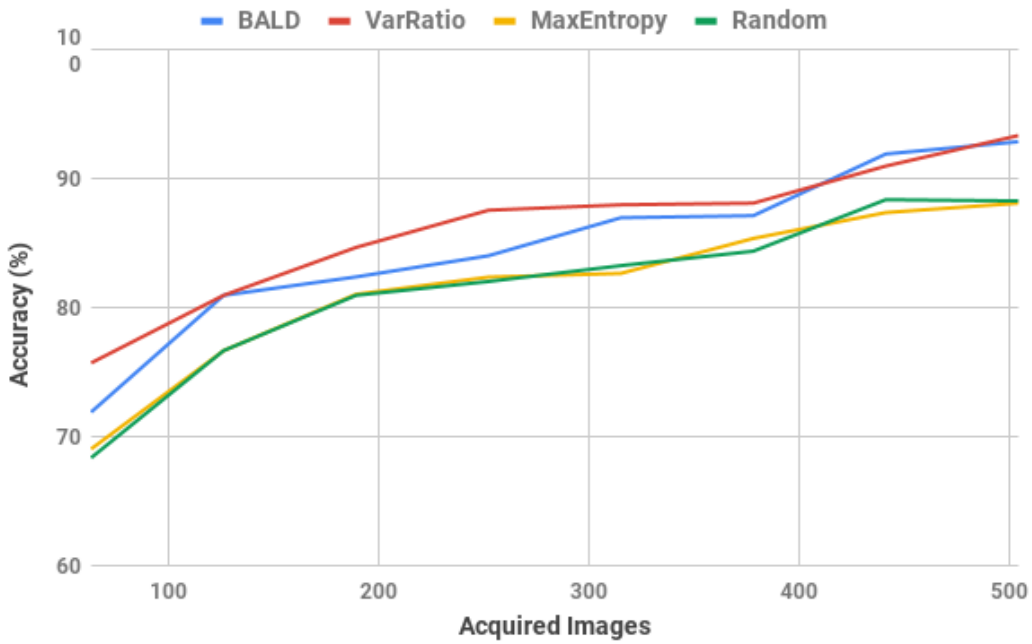
- The system is deployed on a Raspberry Pi 2 in real-time, and the stocked model weights are updated on-device in an incremental manner.
- Initially, we train the model with only 7 bus stops (B1, B2, … B7) and call them existing classes, while the 8th class (B8) is treated as the new unseen bus stop – illustrates scalability.
- Training and testing – high accuracies of ~97% and ~96% respectively.

EXPERIMENT AND BASELINE RESULTS



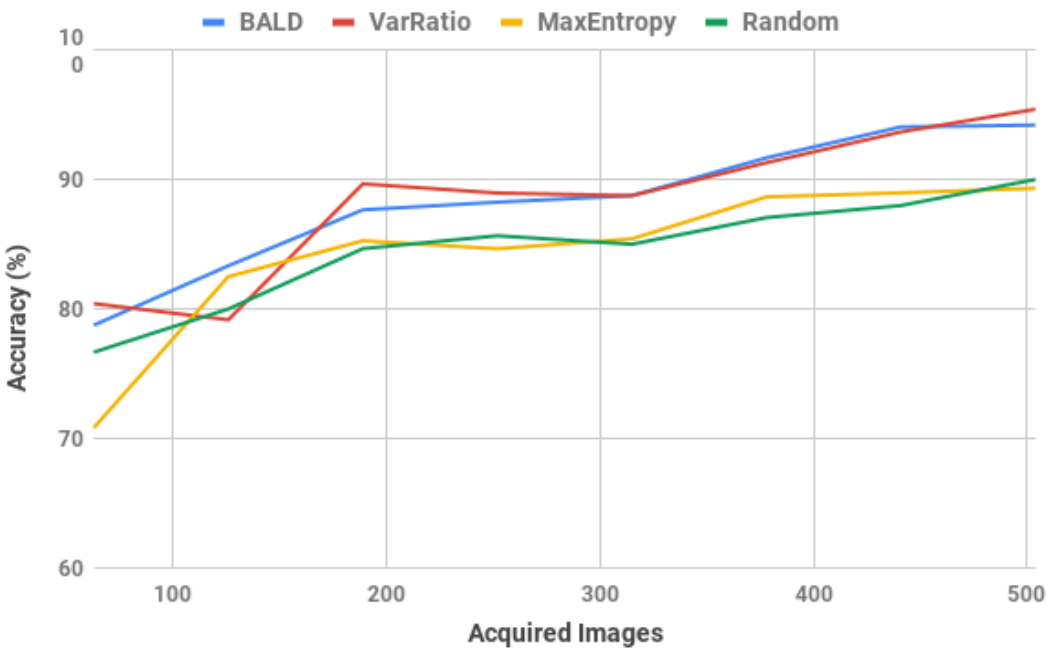
- Initially, we assume only 4 data points were collected from either side of the new bus stop.
- With just the 8 data points, accuracy – 86.25%.
- However, the model when augmented with new B8 images sufficiently overcome class imbalance.
- Achieves an accuracy of 96.7% which is a ~10.5% increase in accuracy.
- Effectiveness of data augmentation strategies for new classes – ensures scalability of bus stops.

RESULTS WITH & WITHOUT AUGMENTATION



- The training data with existing 7 classes (B1, B2….B7) are split into pool (D_{pool}) and train (D_{train}).
- The initial accuracy with just 20% of train data is observed to be 64.28%.
- Variation Ratios (VR) performs the best, achieving ~88% with just less than 250 data points (less than 50% of total D_{pool}).

INCREMENTAL ACTIVE LEARNING EXISTING CLASSES



- Similar training mechanism after data augmentation with (B1, B2….B8).
- Variation Ratios again performs the best again, with a classification accuracy of ~90% with just ~180 images (~37% of total D_{pool}).
- A good trade-off between accuracy and images actively acquired.
- After the first acquisition iteration, would typically require very few actively queried data points to achieve on-par classification accuracies of ~96%.

INCREMENTAL ACTIVE LEARNING AUGMENTED CLASSES

INTELLIGENT INFERENCE

- Ideology is to instill ***biomimicry*** -- just like a human brain towards human-like behavior – not like conventional classification.
- Model acquires and classifies multiple iterative bus stop images on demand, until it can assure a confidence of at least α – ratio of mode of predicted classes to n.
- Typically, the threshold for α is set to a majority among the classified ($\alpha > 0.5$ for 2 images, $\alpha \geq 0.67$ for 3 images, and so on).
- The number of images captured and classified during inference (n), is initially set to 2 and capped at 10 – ($2 < n < 10$).
- The proposed inference mechanism would steer the model towards near-100% accuracy.
- Termed misclassification only when $n > 10$, however even after numerous trials, did not falter with maximum value of n reaching 5.

INCREMENTAL ACTIVE LEARNING

Process	Computational Time
Inference time	11 ms
Incremental Learning per epoch	~1.7ms
Dropout iteration	~1.2ms

- The ConvNet takes a model size of 266 kB.
- T=10 stochastic dropout iterations (1.2 sec per iteration) were used.
- Can be customized depending on the locality and bus usage characteristics, like periodic per trip update, per day/night update, etc.
- Can be seamlessly integrated with vision-based autonomous vehicles.

Contact

Gautham Krishna Gudur

Let’s chat!

THANK YOU!

QUESTIONS?