AI Safety 2023
21 August 2023

Unsupervised Unknown Unknown Detection in Active Learning

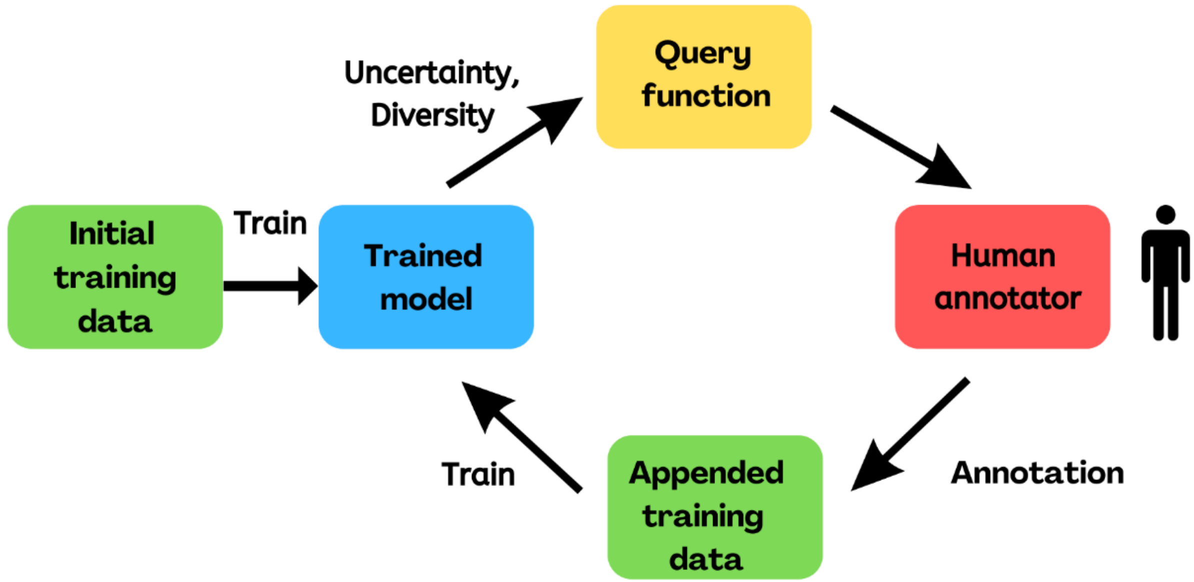
Prajit T Rajendran (CEA LIST), Huascar Espinoza (ECSEL JU),
Agnes Delaborde (LNE), Chokri Mraidha (CEA LIST)

AI SAFETY 2023

Active learning

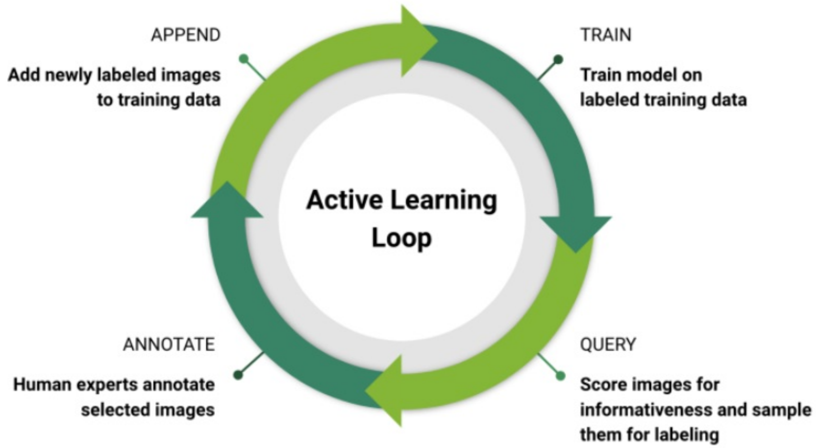
Active learning

- Semi-supervised ML where only a subset of the training data is labelled
- Human queried interactively to label data points of interest from the unlabelled set
- **PROS:** Reduces data labelling requirement
- **CONS:** Selecting the right points to query is important
- **QUERY TYPES:** Random, uncertainty, diversity, consistency

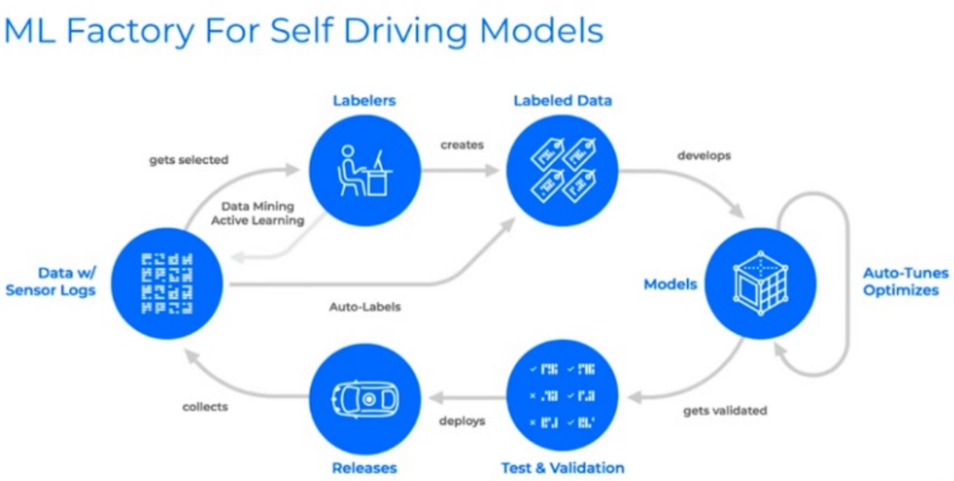


Active learning approaches in companies

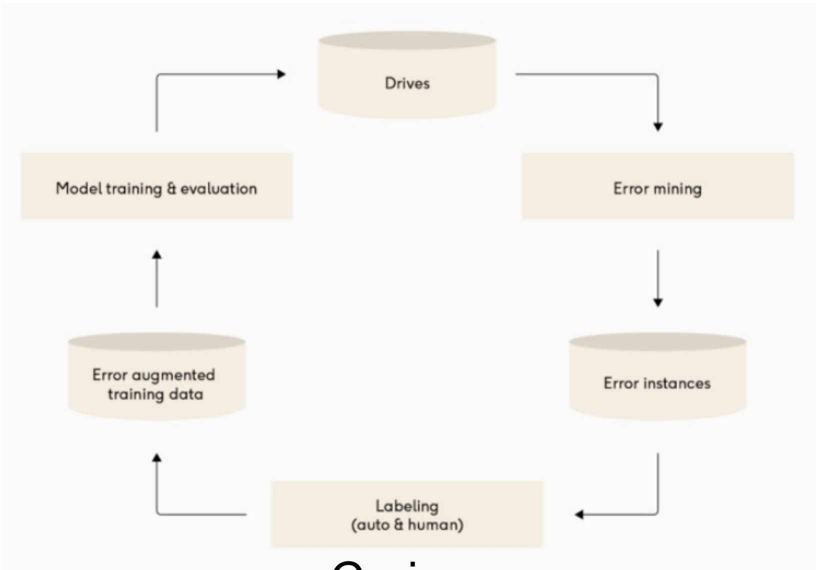
Credits: https://kargarisaac.github.io/blog/



NVIDIA



Waymo



Cruise

Waabi

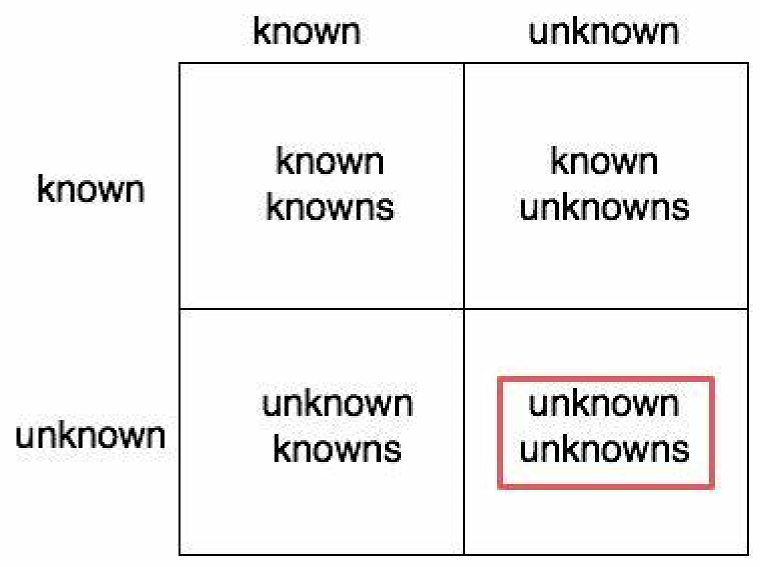
Unknown unknowns

- In machine learning, unknown unknown (UU) data points typically involve **rare and unexpected scenarios** where the models may make wrong predictions, potentially leading to catastrophic situations

- Closely tied to concepts of **anomalies, outliers** in datasets; Difference being UUs are high confidence mispredictions

- Detecting UUs is essential to ensure machine learning systems' reliability and robustness and avoid unexpected failures in real-world safety-critical applications

- **QUESTION:** How can we detect unsafe data points + unknown unknowns in a stream-based setting + can this be feasible in active learning approaches? *(Safety, data efficiency tradeoff)*

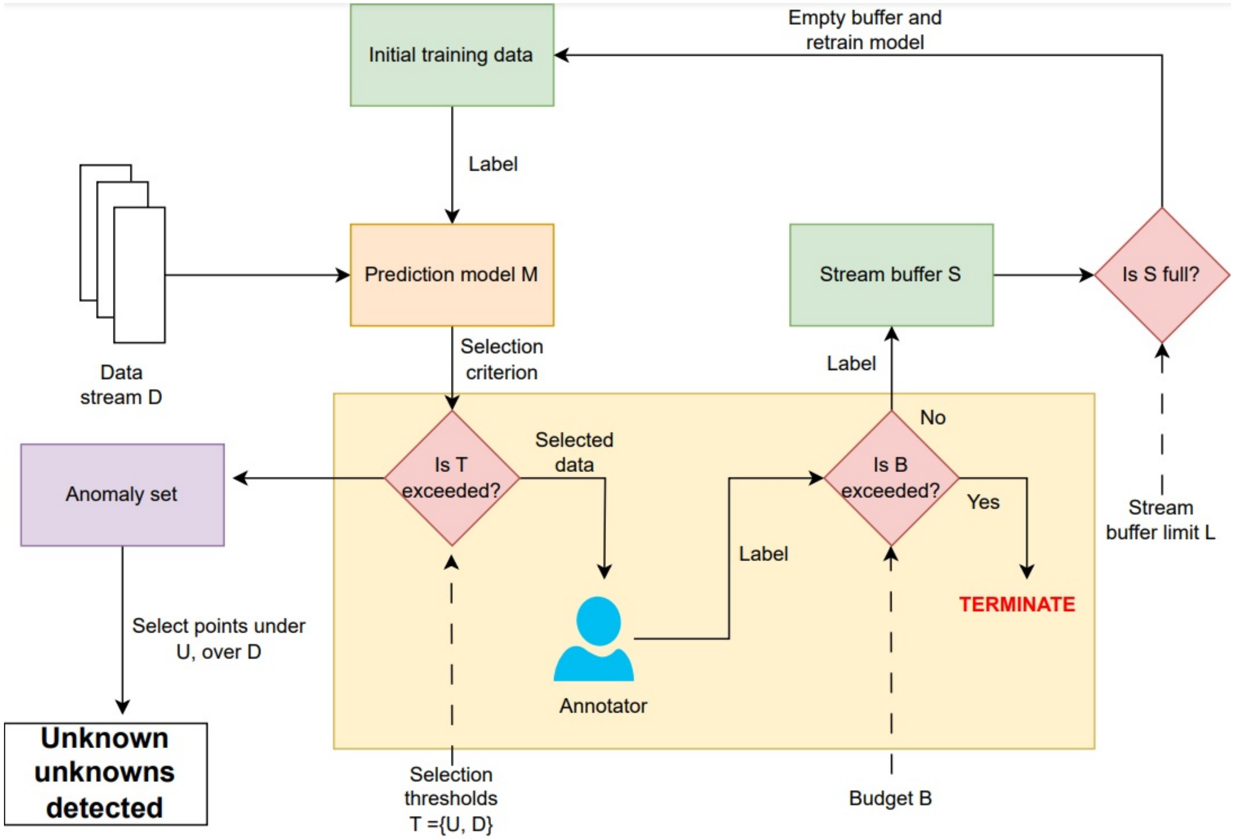


Unknown Unknown Detection in Active Learning (U3DAL)

- Active learning requires uncertainty and diversity thresholds
- Low entropy, high diversity points can be captured by thresholds
- These points may constitute unknown unknowns
- HYPOTHESIS: Active learning thresholds may be used to determine unknown unknowns

Known knowns Low entropy, low diversity	**Known unknowns** High entropy, high diversity
Unknown knowns High entropy, low diversity	**Unknown unknowns** Low entropy, high diversity

U3DAL Block Diagram



- Model M is trained with some initial available labelled data
- Data stream arrives and at each point, a decision is made to accept or reject for labelling based on a threshold
- If both **uncertainty and diversity metrics are high**, the data point is sent to be **annotated**
- If the thresholds for uncertainty and diversity have been set, **low diversity and high uncertainty points** are detected as unknown **unknowns**

U3DAL Experiments

- Mini ImageNet dataset, filtered out 15 classes, 9000 images corresponding to confusing points from ImageNet-A [1]
- ImageNet-A is a set of images labelled with ImageNet labels that were obtained by collecting new data and keeping only those images that ResNet-50 models fail to correctly classify
- 1000 initial training points, 759 "confusing" points from ImageNet-A
- Rest of data shuffled, fed as stream
- Baselines: Local outlier factor [2], isolation forest [3] which are used to detect outliers

Ladybug

Ladybug – confusing point

[1] Hendrycks, Dan, et al. "Natural adversarial examples." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
[2] Breunig, Markus M., et al. "LOF: identifying density-based local outliers." *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000.
[3] F. T. Liu, K. M. Ting and Z. -H. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17.

U3DAL Results

Table 1
Classification accuracy over the train set and anomaly set for different acquisition functions (15-class problem)

No. of data points	Random		Uncertainty		Diversity	
used for training	Validation set	Anomaly set	Validation set	Anomaly set	Validation set	Anomaly set
1000	0.261	0.052	0.261	0.052	0.261	0.052
2000	0.387	0.085	0.417	0.076	0.385	0.088
3000	0.449	0.105	0.432	0.081	0.404	0.096
4000	0.516	0.118	0.506	0.098	0.428	0.113

- Table 1 illustrates that the performance on the validation set and the "anomaly set" were very different
- There was a significant increase in accuracy of the classification task in the validation set as the number of training points increased, as is the expected behaviour
- For the anomaly set, the performance remained poor, demonstrating that they consist of mainly confusing anomalous points which could be potentially unknown unknowns

U3DAL Results

Threshold	D=0.5	D=0.6	D=0.7
U=0.5	91	69	56
U=0.6	116	85	68
U=0.7	122	88	70

Table 2
Variation of number of unknown unknown data points detected as a function of the uncertainty threshold (U) and diversity threshold (D), acquisition function = Random

Threshold	D=0.5	D=0.6	D=0.7
U=0.5	84	62	46
U=0.6	96	69	52
U=0.7	108	77	58

Table 3
Variation of number of unknown unknown data points detected as a function of the uncertainty threshold (U) and diversity threshold (D), acquisition function = Uncertainty

Threshold	D=0.5	D=0.6	D=0.7
U=0.5	90	69	55
U=0.6	100	76	57
U=0.7	104	78	59

Table 4
Variation of number of unknown unknown data points detected as a function of the uncertainty threshold (U) and diversity threshold (D), acquisition function = Diversity

- Tables 2, 3 and 4 illustrate the effect of the uncertainty and diversity thresholds on the number of UUs detected
- In the case of all acquisition functions, U=0.7 and D=0.5 were observed to be the best. This illustrates that having different dimensions for each makes sense rather than a combined equal threshold
- The thresholds are specific to each dataset, model, type of uncertainty score, diversity score used
- Adaptive thresholds, which change based on the arriving distribution could hypothetically increase the detection rate

U3DAL Results

Table 5
Comparison of the number of unknown unknown data points detected by LOF, Isolation forest, U3DAL

No. of data points	Random			Uncertainty			Diversity		
used for training	IF	LOF	U3DAL	IF	LOF	U3DAL	IF	LOF	U3DAL
1000	4	17	35	5	18	55	15	17	48
2000	9	29	59	22	24	58	30	26	66
3000	16	30	82	27	29	93	37	31	82
4000	23	33	122	38	35	108	44	44	104

- Table 5 compares the performance between Isolation forest, local outlier factor and U3DAL in UU detection

- IF and LOF perform better when diversity based measure is used to select new data points for labelling because they are diversity based detection methods themselves

- U3DAL outperforms IF and LOF in all acquisition functions because the confusing data points in the "anomaly set" aren't just different in terms of diversity scores/distance but also in terms of the model's knowledge or lack thereof

Summary

- Proposed a simple and novel method- U3DAL to detect unknown unknowns in an unsupervised manner in a stream-based active learning setting
- Conducted experiments on the Mini ImageNet and ImageNet-A datasets to determine efficacy of UU detection
- Results demonstrate that U3DAL outperforms existing methods like isolation forest and LOF in identifying confusing anomalous data points
- **Future work:** Impact of adaptive thresholds for uncertainty and diversity in UU detection

Thank you

Commissariat à l'énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr