AI Safety 2022, IJCAI-ECAI
25 July 2022

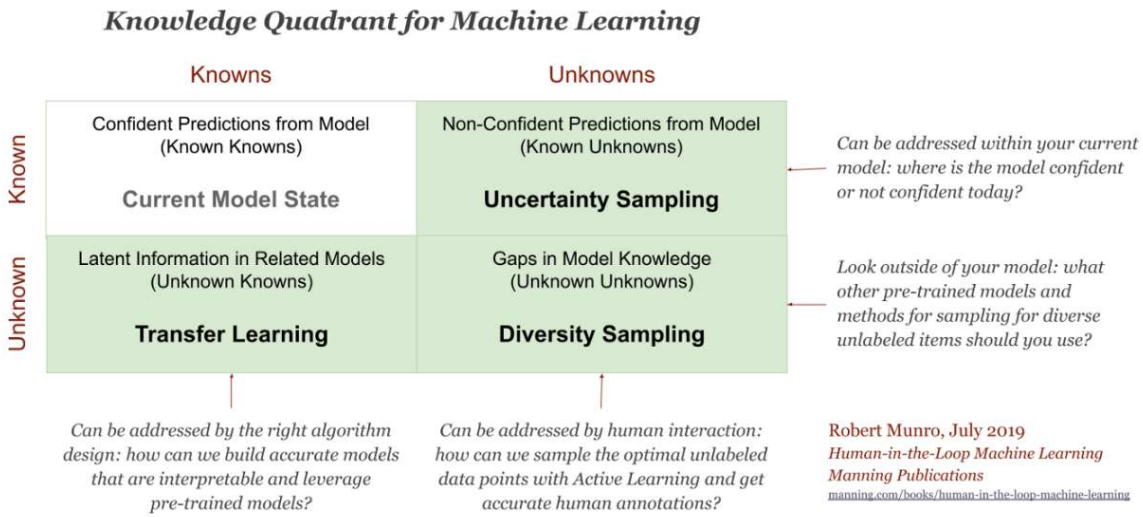
Safety-aware Active Learning with Perceptual Ambiguity and Criticality Assessment

Prajit T Rajendran (CEA), Guillaume Ollier (CEA), Huascar Espinoza (KDT JU), Agnes Delaborde (LNE), Morayo Adedjouma (CEA), Chokri Mraidha (CEA)

AI SAFETY 2022

Objectives

- AI-based autonomous systems should accurately and dependably operate even when confronted with anomalies and unexpected perturbations
- AI-based autonomous should be able to handle unknown unknowns – high confidence wrong predictions
- Discover unpredictable failure situations combining human + machine knowledge



Source: « Human-in-the-loop Machine Learning » by Robert Munro

Blindspots

Blindspots = Deficiencies in a model which may be detrimental to its performance and adaptability to unknown and uncertain situations

- **Model blindspots:** Data points where the model is highly uncertain about or unsure of its predicted label constitute the model blindspots (uncertainty)
- **Data blindspots:** The areas of the feature space that are not covered in the training set constitute the data blindspots (diversity)
- **Human-identified blindspots:** Conceptual gaps leading to unexpected results, which can be identified by human oracles (unknown unknowns)
- **Safety blindspots:** Data points whose misclassification by the specific trained model component could compromise the safety of the system which the component is a part of

Blindspot examples

Distance

Obstruction

Distracting features

Ambiguity

Source: Small Traffic Lights dataset

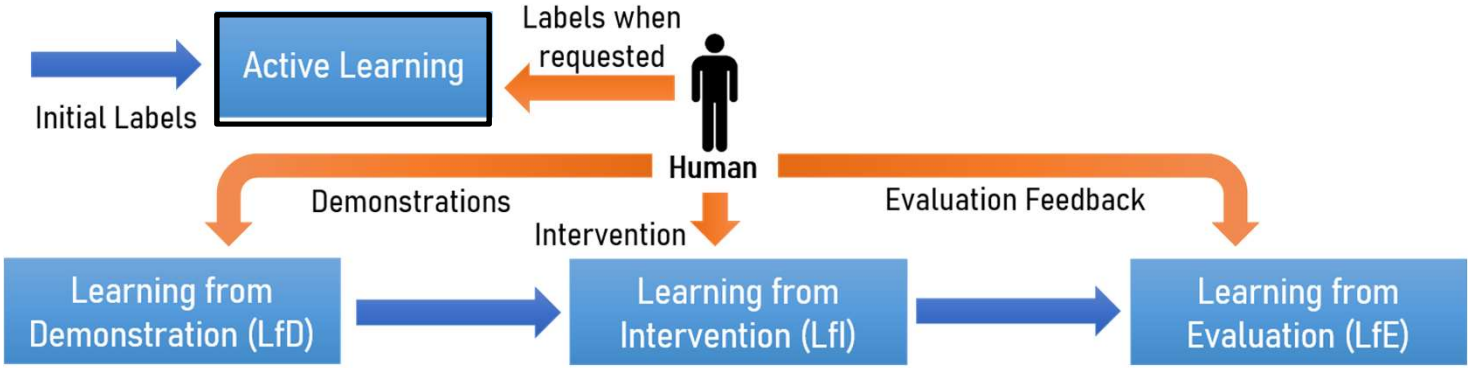
Perceptual ambiguity and criticality

- **Perceptual ambiguity** = What is the degree of ambiguity in this data point?

Level	Explanation
Very low	Unambiguous image, label easy to identify
Low	Distracting features but easy to classify
Medium	Some ambiguities in identifying the label
High	Occlusions and ambiguities, hard to classify
Very high	Corner case with safety implications

- **Criticality & Severity** = What are the consequences of mispredicting this data point?
  - Annotator asked to provide severity estimate ‘s’; If ‘s’ is high ask additional questions to obtain ‘f’ with each question about scenario with exposure ‘e’
  - Compute the criticality score with the formula $c = (\sum f * e) * s$

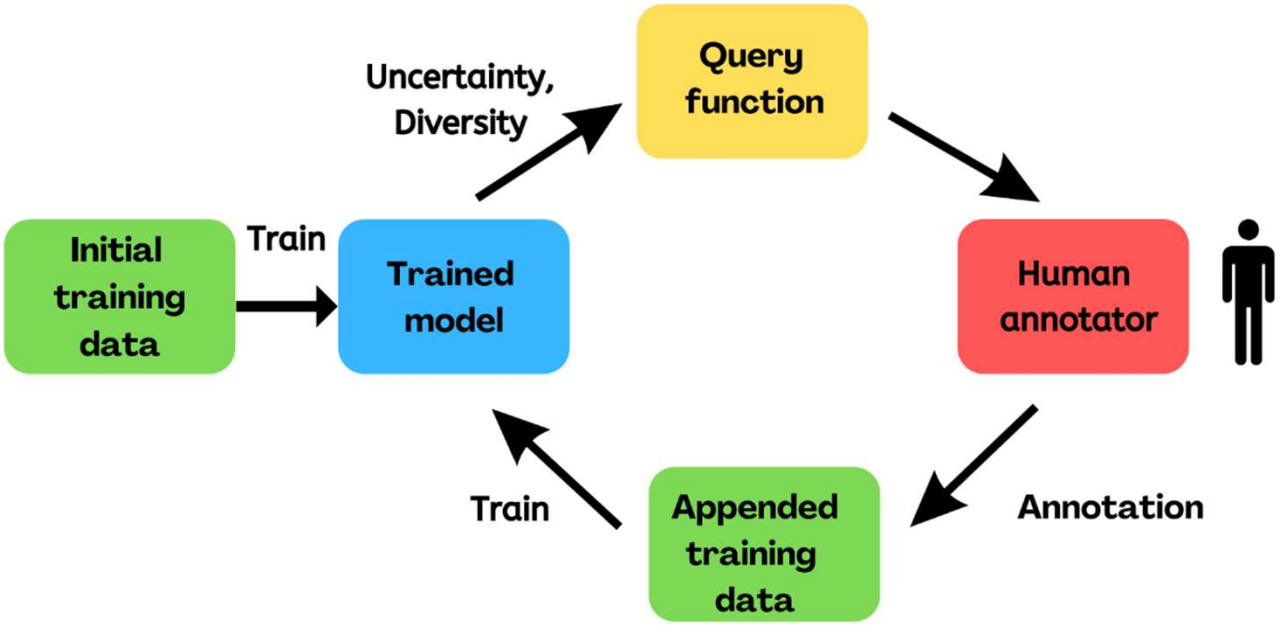
Human-in-the-loop learning



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
- **Random sampling:** Strategy where we pick random samples from the unlabeled pool of data as query points
- **Uncertainty sampling:** Strategy identifying unlabeled items that are near a decision boundary in the trained model
- **Diversity sampling:** Strategy identifying unlabeled items that are underrepresented or unknown to the ML model
- **PROS:** Reduces data labelling requirement
- **CONS:** Selecting the right points to query is important

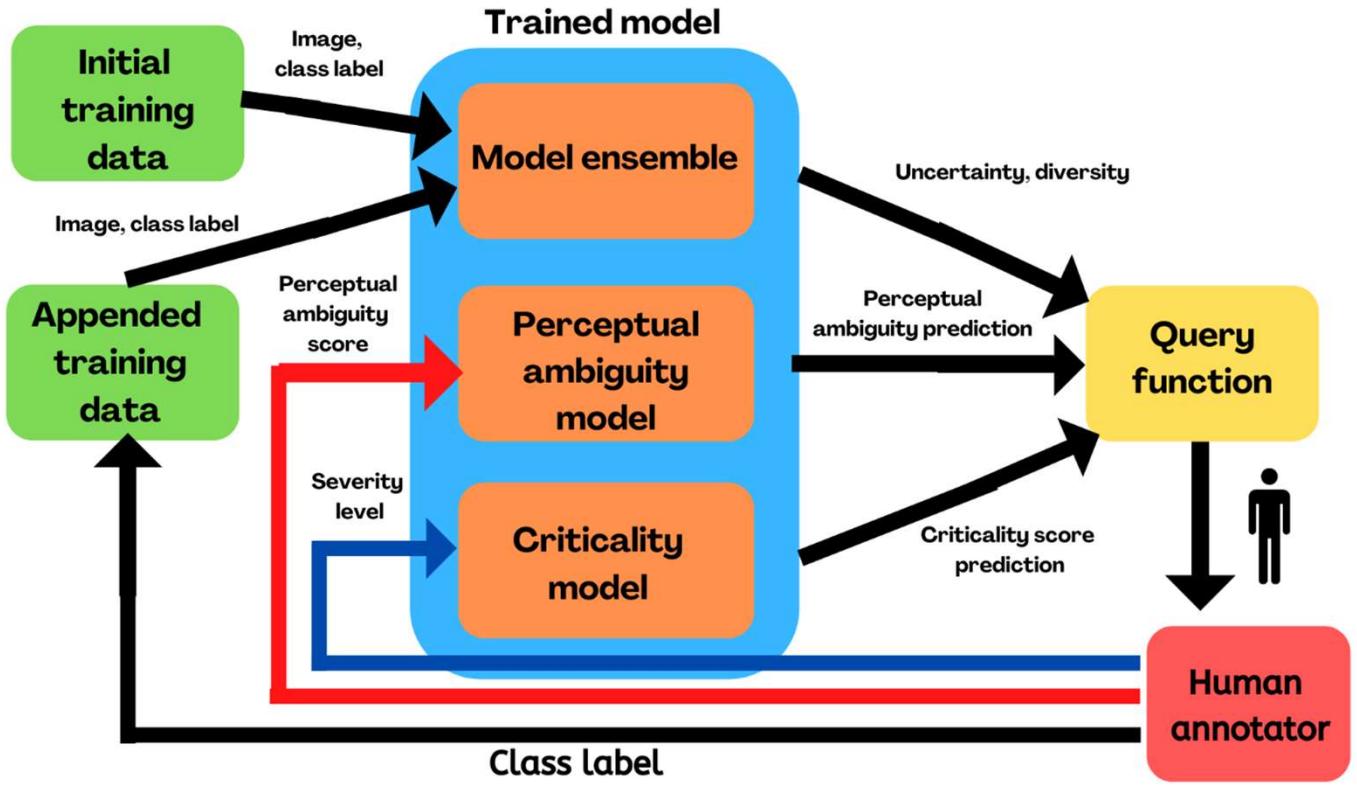
Active learning



Research questions:

1) How to incorporate/embed safety into a learning component to deal with risks such as unknown unknowns?
2) How to reduce data labelling requirements without compromising on safety?
3) How to select data points for querying based on their perceived impact on safety (query function) ?
4) How to measure the impact of the new query function in the active learning task?

Proposed approach



- Uses an ensemble model for class prediction (eg. traffic light – red, yellow, green)
- Uses a continual learning approach for criticality & perceptual ambiguity (very low-very high)
- Check for differences in model prediction to human feedback
- Update models based on differences, hold best possible representation so far to avoid catastrophical forgetting

Experiment plan

- Run tests on a larger dataset with about 13000 traffic light images- Bosch Small Traffic Lights dataset
- Need class labels, perceptual ambiguity and criticality scores, so we need some human labelling
- Create a website to host images (locally) and allow humans to label them at random
- Use inter-annotator agreement to obtain a single class label and single safety relevance score for each data point; store in database
- In the experiment, when the acquisition function queries a particular data point, query from the database
- Test against different acquisition functions- random, uncertainty (U), diversity (D), criticality (C), perceptual ambiguity (P) and their combinations

Possible evaluation metrics for experiment

- **Query relevance** (for acquisition function)
- **Uncertainty over test set** (for component)
- **Alignment of perceptual ambiguity & criticality predictions** (for component)
- **Accuracy** (for component)
- **Corruption robustness to random perturbations** (for component)
- **Adversarial robustness** (for component)
- **Collision / traffic rule violation incidents** (for system)
- **ODD exits** (for system)
- **Number of human interventions** (for system)

Some evaluation metrics in detail

After each round of updating model(s):

- For **uncertainty**- Calculate average entropy in validation set
- For **diversity**- Calculate feature space coverage in training data used thus far
- For **query relevance-** Calculate difference in average entropy of automated labelled v/s human labelled subset in each round
- For **safety**- Hold wrong predictions in buffer, retrain when buffer overflows, conditioned on not forgetting the learnt correct predictions (continuous evaluation by human)
- For **performance**- Accuracy, F1-score, FPR

Predicted trends on unseen test set:

- For **uncertainty**- After each round, average entropy should decrease
- For **diversity**- After each round, feature space coverage should increase
- For **query relevance-** After each round, the difference in average entropy should increase
- For **safety**- After each round, the number of wrong (mismatched) predictions should decrease
- For **performance**- Accuracy and F1-score should increase, reverse for FPR

Next steps

- Experiment with the Bosch Small Traffic Lights dataset
- Extension plan: Experiment with another application like pedestrian detection
- Link idea more concretely with existing safety concepts in the design stage
- Experiment with active learning for sequence of images (scenario based)

Thank you

Prajit Thazhurazhikath Rajendran
prajit.thazhurazhikath@cea.fr

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

Établissement public à caractère industriel et commercial | RCS Paris B 775 685 019