Unsupervised Unknown Unknown Detection in Active Learning

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

<table>
<thead>
<tr>
<th>Known knowns</th>
<th>Known unknowns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low entropy, low diversity</td>
<td>High entropy, high diversity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unknown knowns</th>
<th>Unknown unknowns</th>
</tr>
</thead>
<tbody>
<tr>
<td>High entropy, low diversity</td>
<td>Low entropy, high diversity</td>
</tr>
</tbody>
</table>
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

U3DAL Results

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

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

- 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

- 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

<table>
<thead>
<tr>
<th>No. of data points used for training</th>
<th>Random</th>
<th>Uncertainty</th>
<th>Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IF</td>
<td>LOF</td>
<td>U3DAL</td>
</tr>
<tr>
<td>1000</td>
<td>4</td>
<td>17</td>
<td>35</td>
</tr>
<tr>
<td>2000</td>
<td>9</td>
<td>29</td>
<td>59</td>
</tr>
<tr>
<td>3000</td>
<td>16</td>
<td>30</td>
<td>82</td>
</tr>
<tr>
<td>4000</td>
<td>23</td>
<td>33</td>
<td>122</td>
</tr>
</tbody>
</table>

- 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