<table>
<tr><th rowspan="2">cases</th><th>doc_1</th><th>doc_2</th><th>decision</th><th>id</th></tr>
<tr>
<td>
<table>
<tr><th>authors</th><td>Alexander Meinke
Julian Bitterwolf
Matthias Hein</td></tr>
<tr><th>title</th><td>Provably Robust Detection of Out-of-distribution Data (almost) for free.</td></tr>
<tr><th>publication_date</th><td>2021-06-08 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.OPENALEX</td></tr>
<tr><th>journal</th><td>arXiv (Cornell University)</td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td>None</td></tr>
<tr><th>urls</th><td>https://openalex.org/W3172596993</td></tr>
<tr><th>id</th><td>id-264729082788196000</td></tr>
<tr><th>abstract</th><td></td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td>
<td>
<table>
<tr><th>authors</th><td>Matthias Hein
Julian Bitterwolf
Alexander Meinke</td></tr>
<tr><th>title</th><td>Provably Robust Detection of Out-of-distribution Data (almost) for free</td></tr>
<tr><th>publication_date</th><td>2021-06-08 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.PAPERS_WITH_CODE</td></tr>
<tr><th>journal</th><td></td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td>https://arxiv.org/pdf/2106.04260v2.pdf
https://github.com/AlexMeinke/Provable-OOD-Detection
https://openreview.net/pdf?id=qDx6DXD3Fzt</td></tr>
<tr><th>id</th><td>id-9201529456528127826</td></tr>
<tr><th>abstract</th><td>The application of machine learning in safety-critical systems requires a reliable assessment of uncertainty. However, deep neural networks are known to produce highly overconfident predictions on out-of-distribution (OOD) data. Even if trained to be non-confident on OOD data, one can still adversarially manipulate OOD data so that the classifier again assigns high confidence to the manipulated samples. We show that two previously published defenses can be broken by better adapted attacks, highlighting the importance of robustness guarantees around OOD data. Since the existing method for this task is hard to train and significantly limits accuracy, we construct a classifier that can simultaneously achieve provably adversarially robust OOD detection and high clean accuracy. Moreover, by slightly modifying the classifier's architecture our method provably avoids the asymptotic overconfidence problem of standard neural networks. We provide code for all our experiments.</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td>
<td>DUPLICATES</td>
<td>117</td>
</tr>
</table>