Black Box Membership Inference Against Object Detection Models

Matthew Churilla

FEBRUARY 10, 2021

© 2021 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Legal

Copyright 2021 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

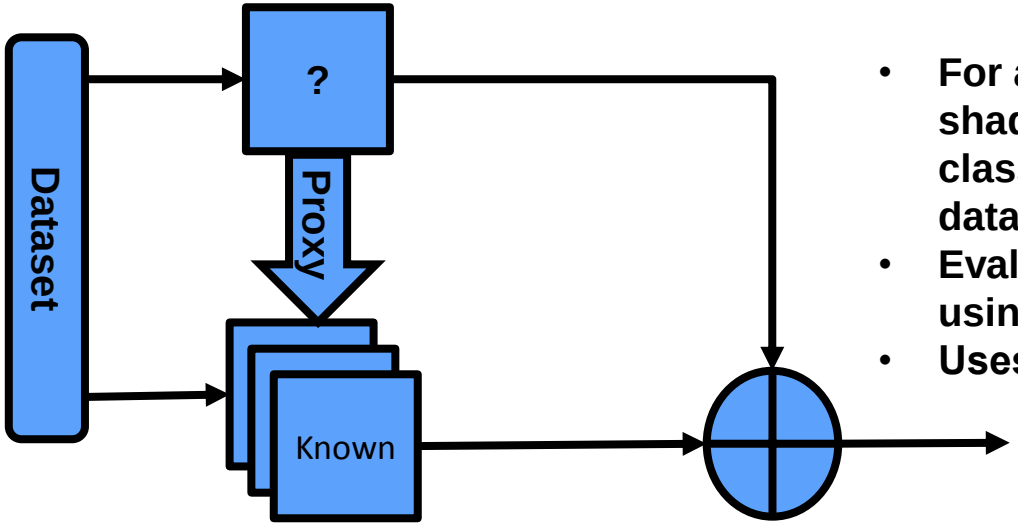
[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

DM21-0116

Shokri Membership Inference Attack

Determine if a model was trained on a dataset

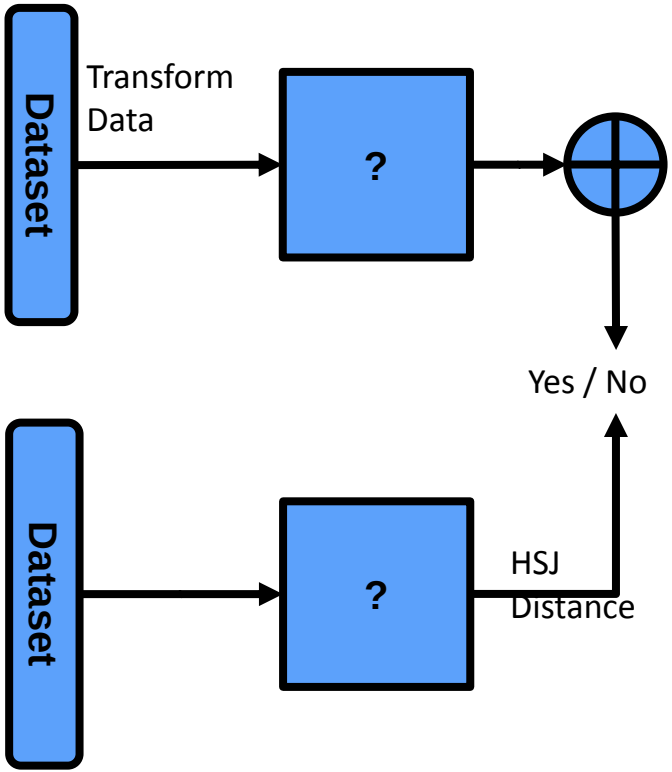


- For an unknown model use shadow models to create classifiers that will determine if data is training data or not
- Evaluate unknown model results using classifiers
- Uses Model's full prediction array

Membership Inference Attacks against Machine Learning Models, Shokri et al., 2016
[1610.05820] Membership Inference Attacks against Machine Learning Models (arxiv.org)

© 2021 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

Label-Only Membership Inference Attacks



Two methods:

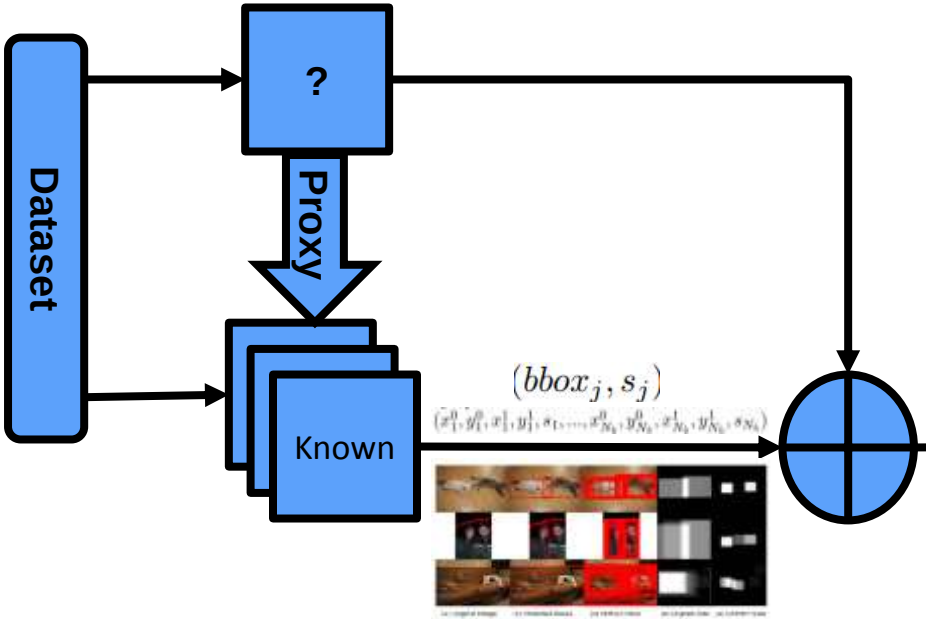
1. **Data Augmentation**
   - **Rotate and shift data**
   - **Detect Misclassifications**
   - **Create classifier on misclassifications**
2. **Distance**
   - **Use HopSkipJump to find decision boundary**
   - **Use distance as in / out proxy**

Only needs prediction values

Label-Only Membership Inference Attacks, Choo et al. 2020, [2007.14321] Label-Only Membership Inference Attacks (arxiv.org)

© 2021 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

Membership Inference Attacks Against Object Detection Models



Same shadow model process as Shokri but on bounding boxes not prediction vectors

Requires confidence and NMS off

Two methods:
XG-BOOST on tabular data

Canvas

Membership Inference Attacks Against Object Detection Models, Park and Kang 2020, [2001.04011] Membership Inference Attacks Against Object Detection Models (arxiv.org)

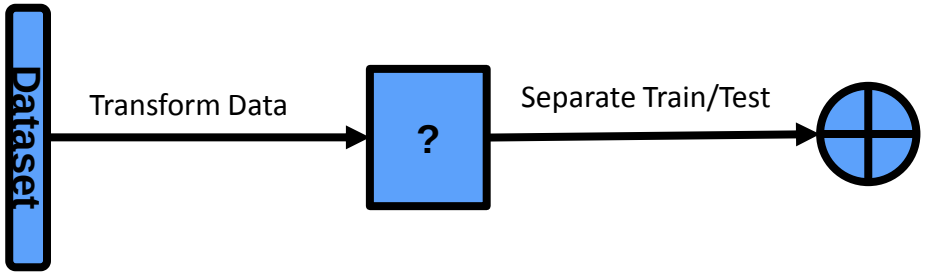
© 2021 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

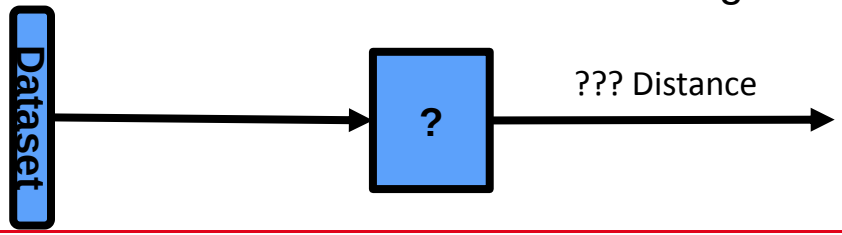
Our Hypothesis

Like classification, Membership Inference should be possible against object detection models in a label only manner.

Data Augmentation should surface differences in bounding boxes that a classifier can detect.



A distance based attack on bounding boxes should also be possible.



© 2021 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

Data Augmentation Needs

A set of transformers for images and bounding boxes

A way to measure differences in bounding boxes

Tensors of IOUs

```
tensor([[[0.1586, 0.0000, 0.0000,  ..., 0.0000, 0.0000, 0.0000],
         [0.0000, 0.0000, 0.0000,  ..., 0.0000, 0.6684, 0.0000],
         [0.0000, 0.0000, 0.0000,  ..., 0.0000, 0.0000, 0.4001],
         ...,
         [0.0390, 0.0000, 0.0000,  ..., 0.2497, 0.0000, 0.0000],
         [0.0000, 0.0000, 0.0000,  ..., 0.0000, 0.0000, 0.0000],
         [0.0222, 0.0000, 0.0000,  ..., 0.1835, 0.0000, 0.0000]])
```

Canvases of Bounding Boxes

© 2021 Carnegie Mellon University

[Distribution Statement A] Approved for public release and unlimited distribution.

Next Steps

Test separability of train / test based on IOU values

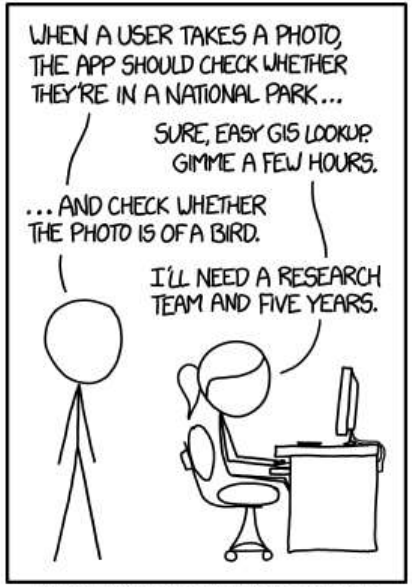
Develop XG-BOOST method for bounding box data

Refinement of Canvas method to account for lack of confidence vales

Create distance measurement for object detectors

© 2021 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.

Questions?



Tasks, xkcd, https://xkcd.com/1425/

© 2021 Carnegie Mellon University
[Distribution Statement A] Approved for public release and unlimited distribution.