The SOREL-20M dataset

The first production scale malware detection dataset with complete malware binaries

Rich Harang (Duo Security)
and
Ethan M. Rudd (Mandiant)

An important note

This work was funded by Sophos, and performed by the authors while employed there, with generous support and cooperation from ReversingLabs.

The least you need to know:

1. **Who?** Rich Harang and Ethan Rudd
2. **What?** The Sophos-ReversingLabs 20 million sample malware dataset (SOREL-20M), including metadata and full EMBERv2.0 features for ~20 million benign/malware samples, and 10 million disarmed malware samples with complete binaries
3. **When?** Released December 14, 2020
4. **Where?** https://github.com/sophos-ai/SOREL-20M and
   s3://sorel-20m/09-DEC-2020/
5. **Why? and How?** For SCIENCE™! See the next few slides.

Why?

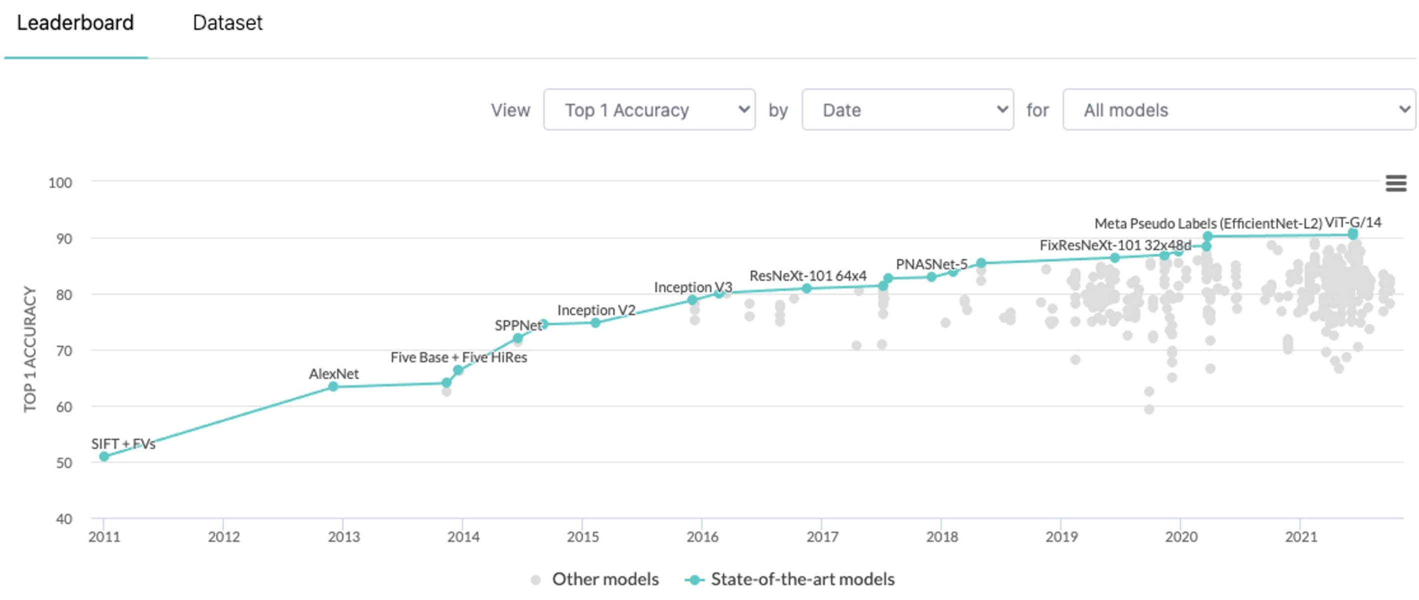
I’m going to say just one word to you. Just one word. Are you ready?

I’m going to say just one word to you. Just one word.
Are you ready?

How do you measure progress in ML?

Image tasks:

- ImageNet

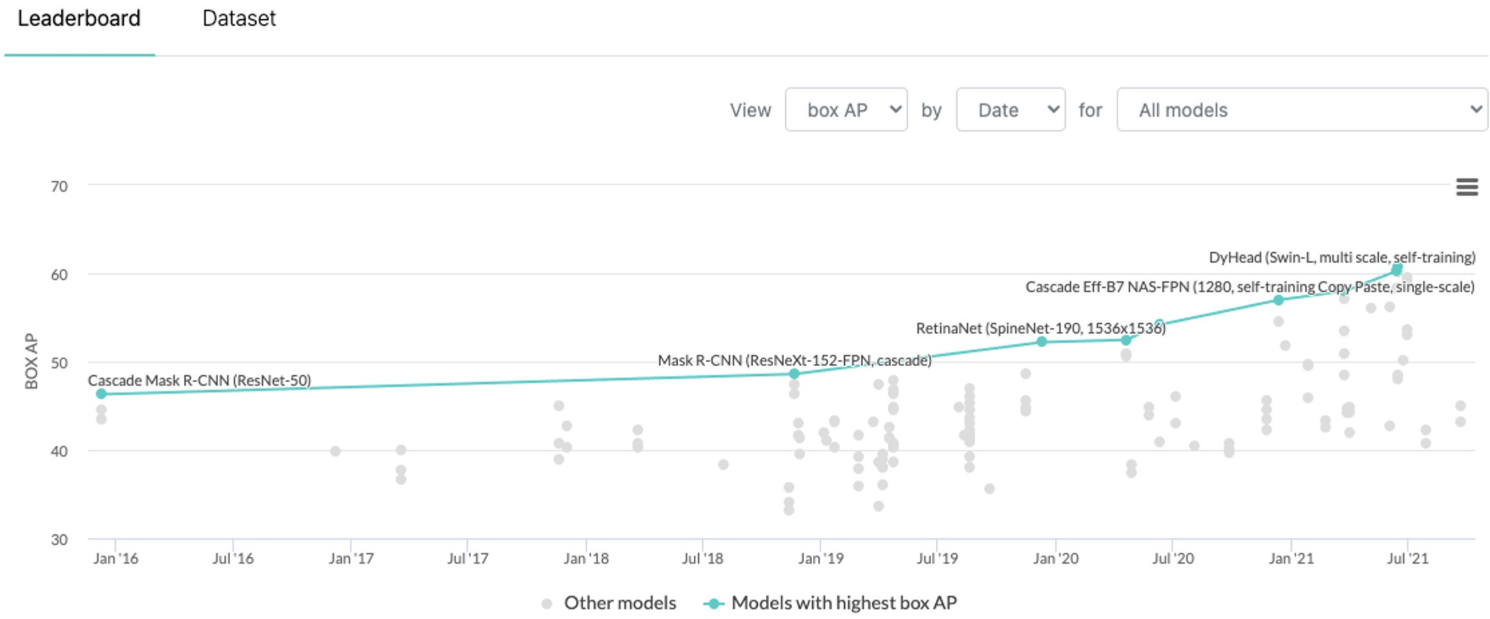


View Top 1 Accuracy by Date for All models

How do you measure progress in ML?

Image tasks:

- ImageNet
- COCO



Leaderboard   Dataset

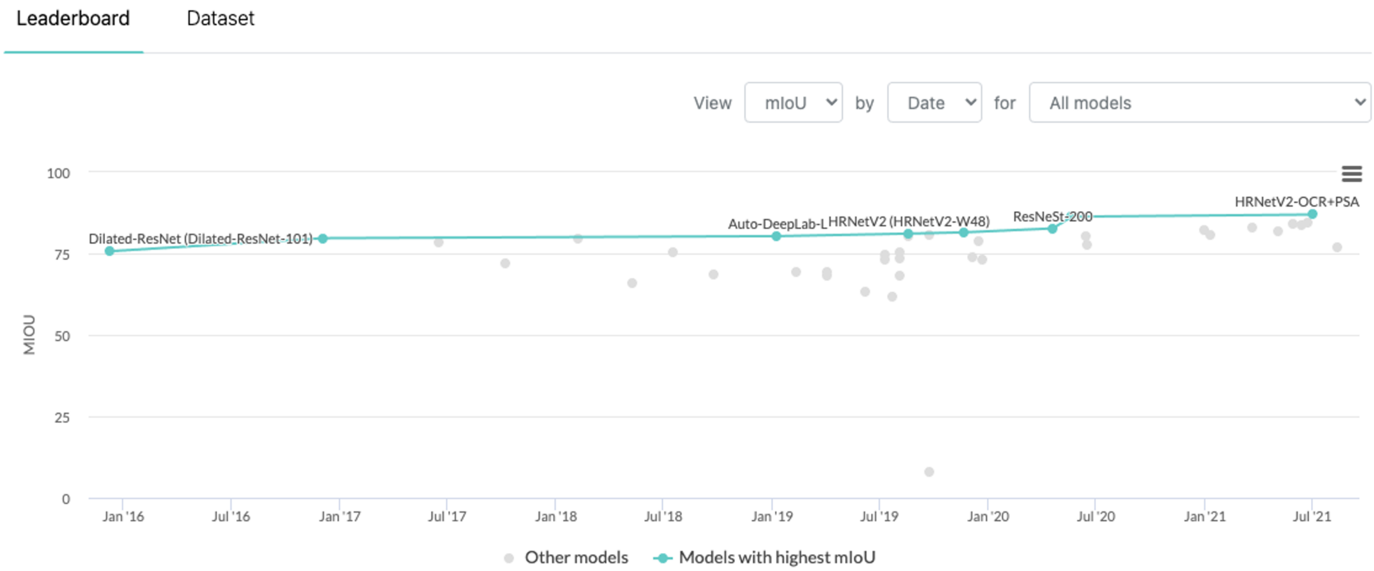
View box AP by Date for All models

70
60
50
40
30
BOX AP
Cascade Mask R-CNN (ResNet-50)
Mask R-CNN (ResNeXt-152-FPN, cascade)
RetinaNet (SpineNet-190, 1536x1536)
Cascade Eff-B7 NAS-FPN (1280, self-training Copy Paste, single-scale)
DyHead (Swin-L, multi scale, self-training)
Jan '16
Jul '16
Jan '17
Jul '17
Jan '18
Jul '18
Jan '19
Jul '19
Jan '20
Jul '20
Jan '21
Jul '21
Other models
Models with highest box AP

How do you measure progress in ML?

Image tasks:

- ImageNet
- COCO
- Cityscapes

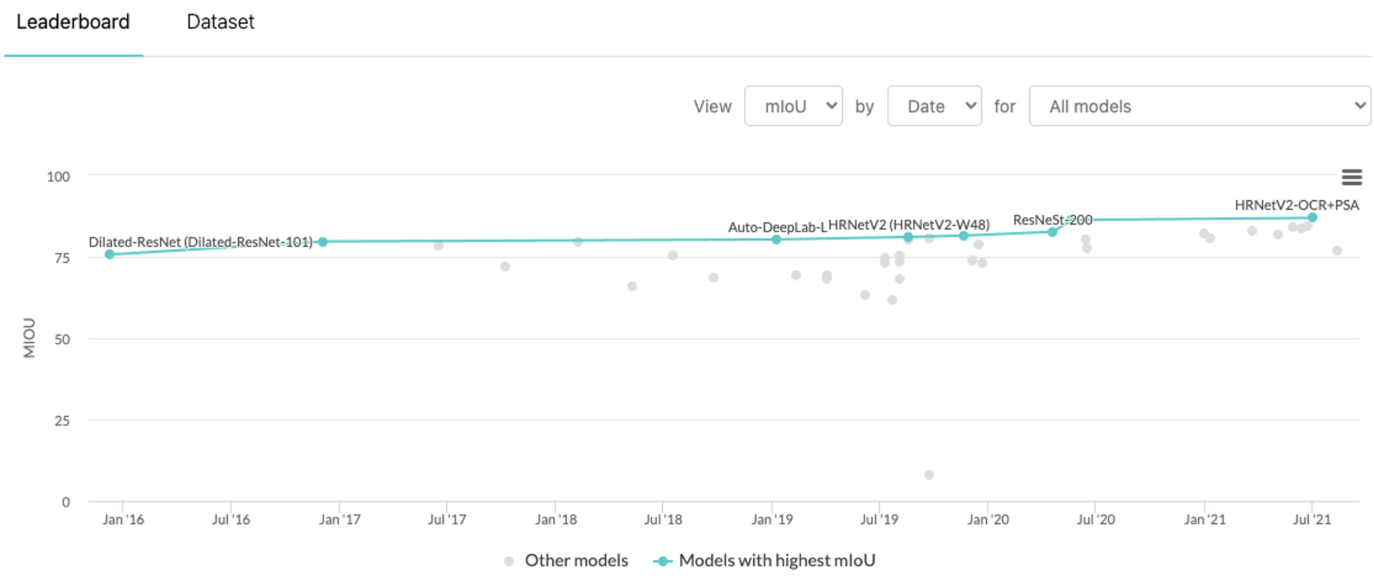


View mIoU by Date for All models

How do you measure progress in ML?

Image tasks:

- ImageNet
- COCO
- Cityscapes
- ...and more



View mIoU by Date for All models

How do you measure progress in ML?

Malware

- Ember
- ...

How do you measure progress in ML?

Malware

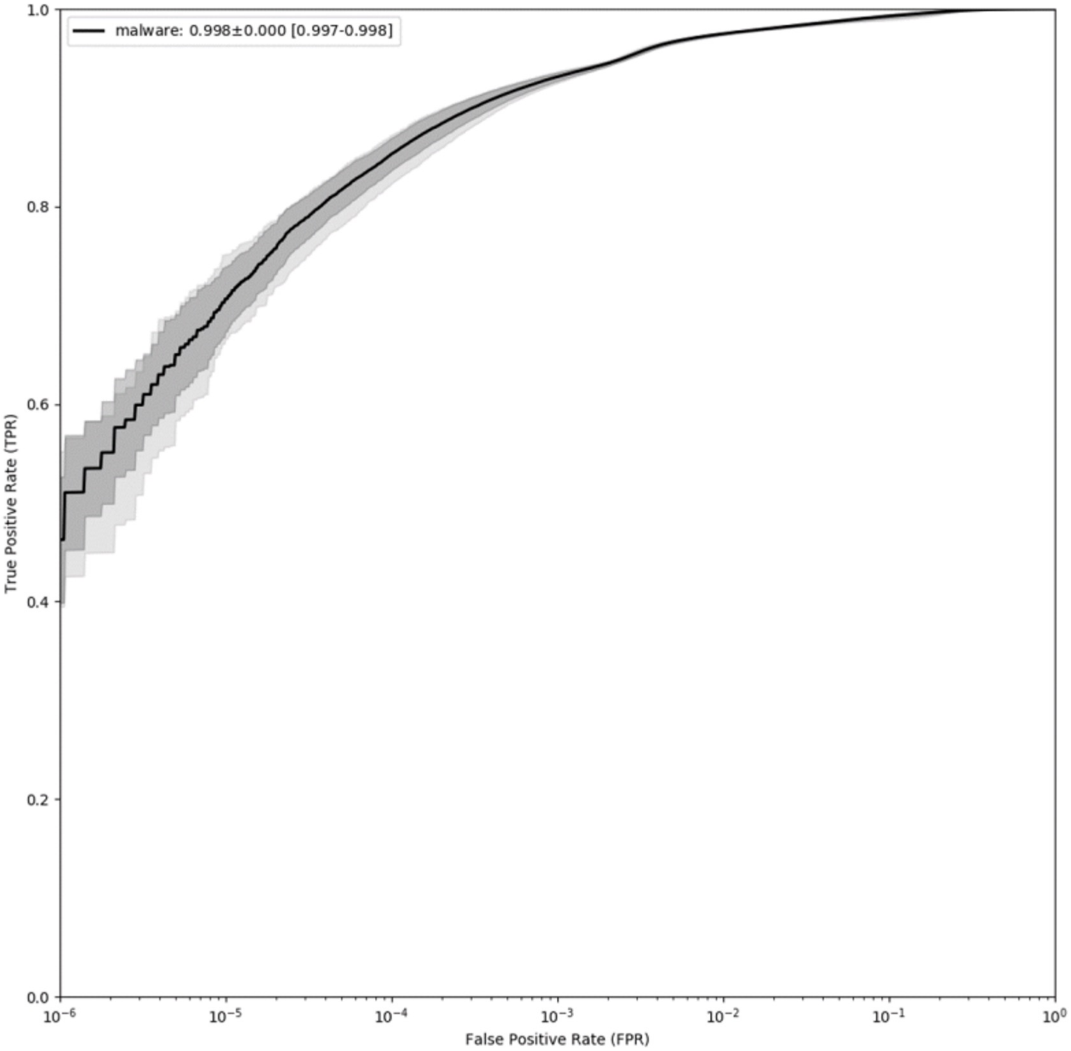
- Ember
- New: SOREL-20M

OK, but why a *new* data set?

Robust evaluation

Key performance metrics for malware classification are things like TPR at a FPR of 1 in 1000

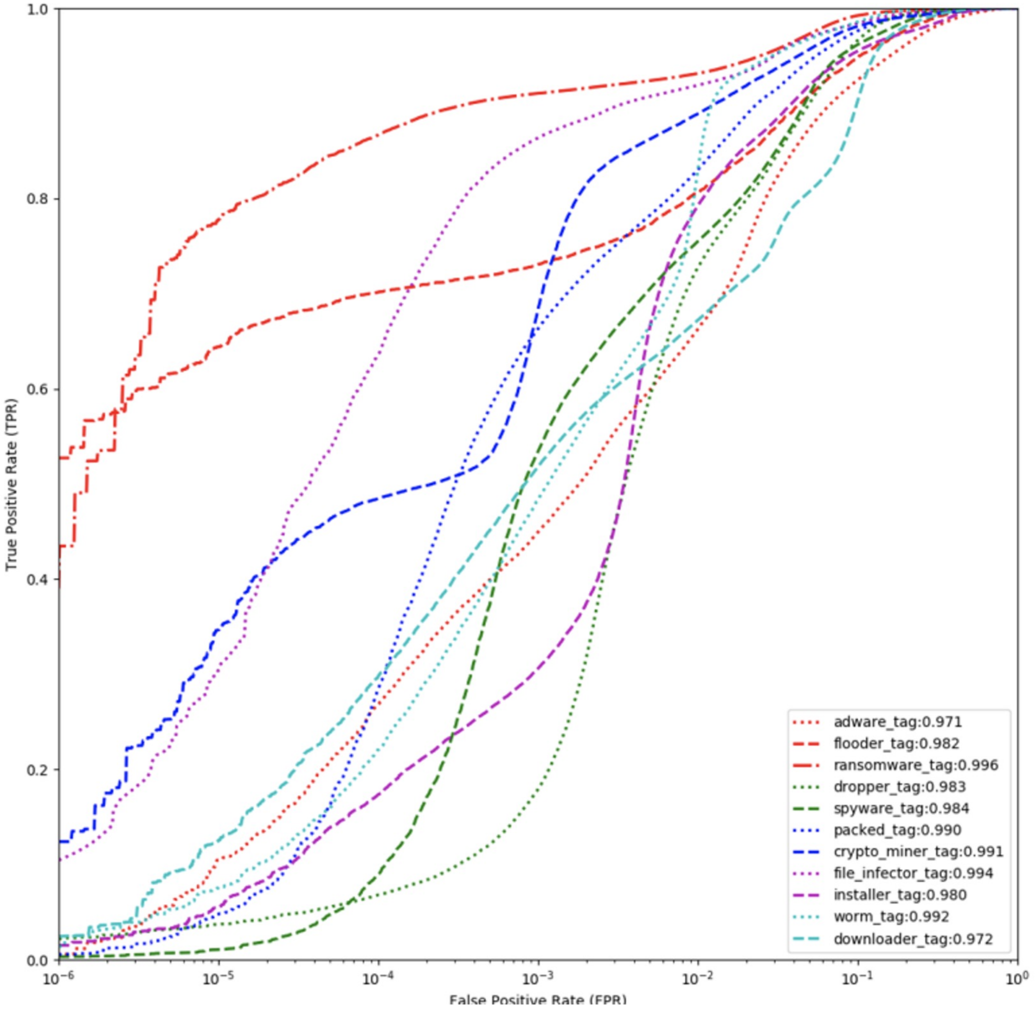
- Ember's limited validation sample size (200,000) makes this difficult to evaluate with low variance
- SOREL-20M contains 2.5M validation and 4.2M test samples; low sample variance at low FPRs



Multiple targets

Good benchmarks are multi-modal

- Ember contains binary labels
- SOREL-20M contains binary labels, behavioral tags, and detection counts



High-quality labels generated with proprietary techniques

Good malware labels are hard (and often proprietary)

- Ember uses a simple thresholding rule
- SOREL-20M uses in-house malware labeling combining multiple data feeds and proprietary techniques.

We believe these are some of the cleanest labels out there for malware today

Full binary samples (of malware), rich data

Feature exploration is critical

- Ember provides only pre-extracted features
- SOREL-20M provides ‘disarmed’ malware binaries and PEfile metadata for all samples

Any limitations?

1. No benignware samples
2. Disarming of files prevents dynamic analysis
3. Disarming of files causes some issues with less robust static analysis tools
4. Malware landscape moves fast, newest sample in the data is already 28 months old.

Recap: SOREL-20M...

1. Production-scale sample size, allows production-level evaluation
2. Multiple targets to allow exploration of multi-task learning, family classification, etc.
3. Extremely high-quality “production” level labels
4. Ember2.0 features with complete PE metadata
5. Full (disarmed) malware binaries for feature exploration.

Thanks once more to...

Questions?