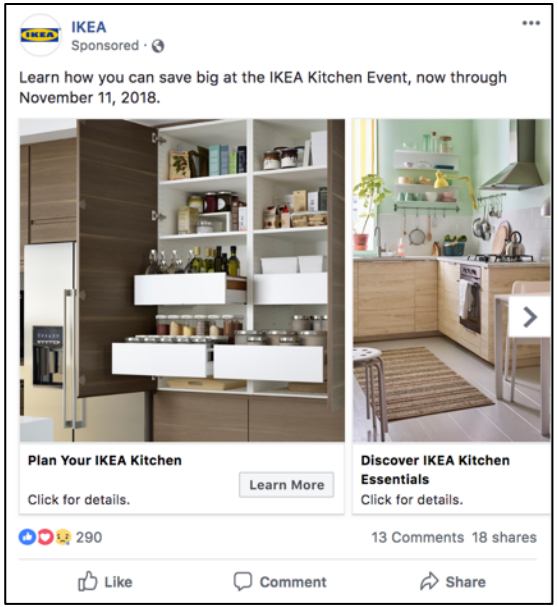
AdVersarial: Defeating Perceptual Ad Blocking with Adversarial Examples

Florian Tramèr

September 10th 2019

Joint work with Pascal Dupré, Gili Rusak, Giancarlo Pellegrino and Dan Boneh

The Future of Ad-Blocking?



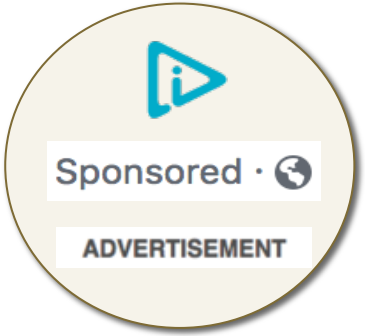
easylist.txt

…markup…
…URLs…

???

Human distinguishability of ads

- *Legal requirement (U.S. FTC, EU E-Commerce)*
- *Industry self-regulation on ad-disclosure*



Towards Computer Vision for Ad-Blocking

- Why not detect ad-disclosures programmatically?

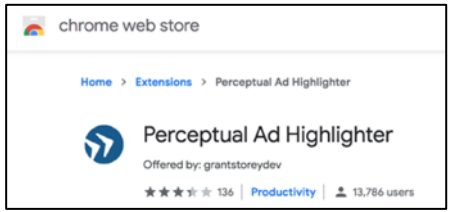
```
<a><span>
<span class="c1">Sp</span>
<span class="c2">S</span>
<span class="c1">on</span>
<span class="c2">S</span>
<span class="c1">so</span>
<span class="c2">S</span>
<span class="c1">red</span>
<span class="c2">S</span>
</span></a>
```

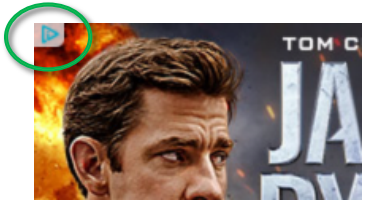
```
.c2 { font-size: 0; }
```

> *New arms race on HTML obfuscation*
> *E.g., Facebook vs uBlockOrigin: https://github.com/uBlockOrigin/uAssets/issues/3367*
  - ***1 year, 253 comments**, and counting...*

Perceptual Ad-Blocking

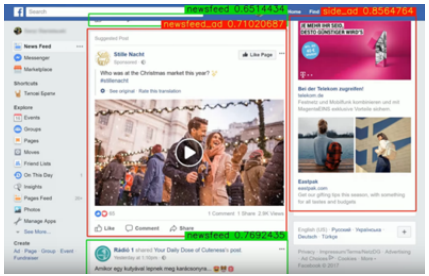
- **Ad Highlighter** [Storey et al., 2017]
  - > *Visually detects ad-disclosures*
  - > *Traditional computer vision techniques*
  - > *Simplified version in Adblock Plus*





- **Sentinel** by Adblock Plus
  - > *Locates ads in Facebook screenshots using neural networks*

- **Percival** by Brave [Din et al., 2019]
  - > *Neural network embedded in Chromium's rendering pipeline*



Perceptual Ad-Blocking

Business ▸ Policy

Will the MOAB (Mother Of all AdBlockers) finally kill advertising?

'Perceptual ad blocker' cannot be defeated, researchers claim

By Andrew Orlowski 19 Apr 2017 at 08:35 178 SHARE ▼

MOTHERBOARD

PERCEPTUAL AD BLOCKING | By Jason Koebler | Apr 14 2017, 10:47am

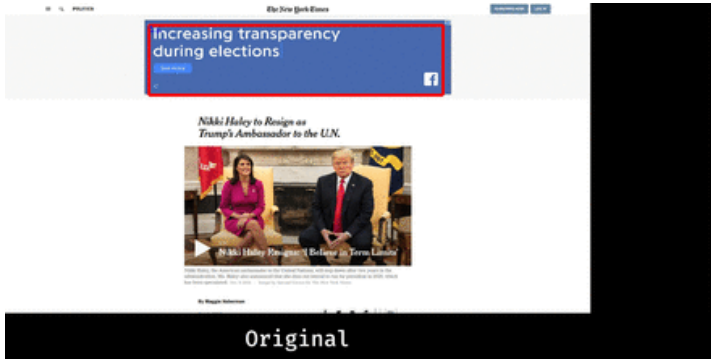
Princeton's Ad-Blocking Superweapon May Put an End to

BusinessWire
A Berkshire Hathaway Company

Adblock Plus Re-Invents Ad-Blocking Future Through People-Powered Artificial Intelligence

Adblock Plus launches AI-powered ad detector "Sentinel," and invites people worldwide to train neural network algorithms to understand what bad ads look like

How Secure is Perceptual Ad-Blocking?



Jerry uploads malicious content …

… so that Tom's post gets blocked

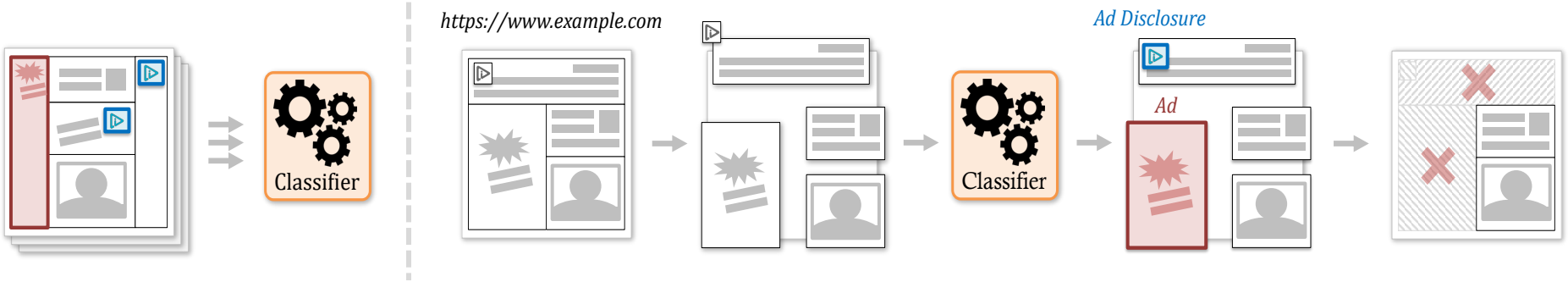
Outline

- Perceptual ad-blockers: how they work
- Attacking perceptual ad-blockers
- Why defending is hard

Outline

- **Perceptual ad-blockers: how they work**
- Attacking perceptual ad-blockers
- Why defending is hard

How does a Perceptual Ad-Blocker Work?

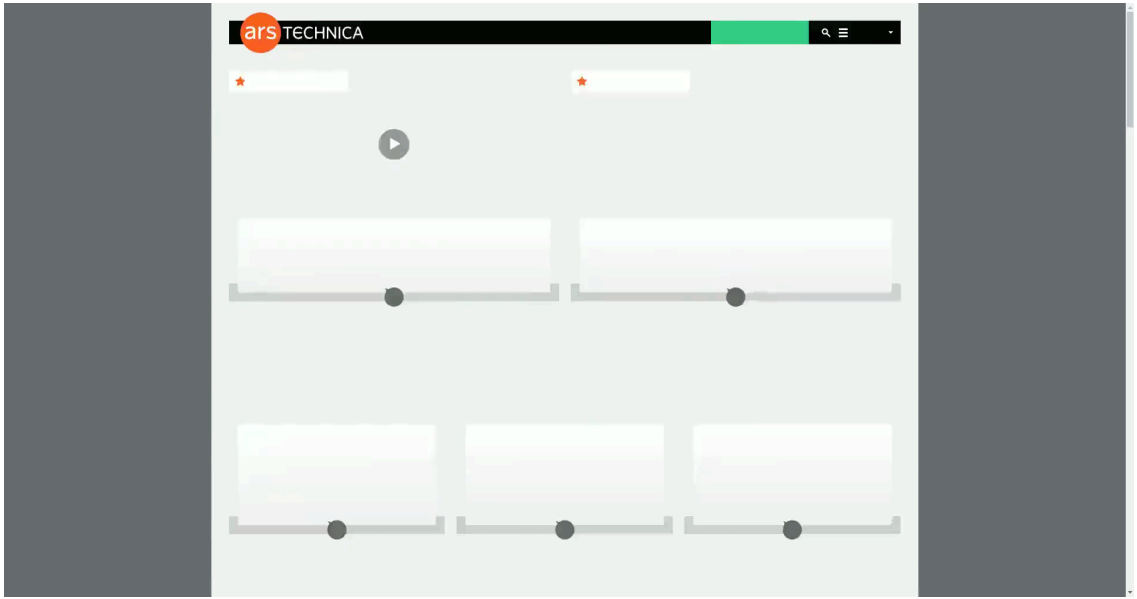


Template matching, OCR,
DNNs, Object detector networks

- **Element-based** (e.g., find all <img> tags) [Storey et al. 2017]
- Frame-based (segment rendered webpage into "frames" as in Percival)
- **Page-based** (unsegmented screenshots à-la-Sentinel)

Building a Page-Based Ad-Blocker

We trained a neural network to detect ads on news websites from all G20 nations



Video taken from 5 websites *not used during training*

Outline

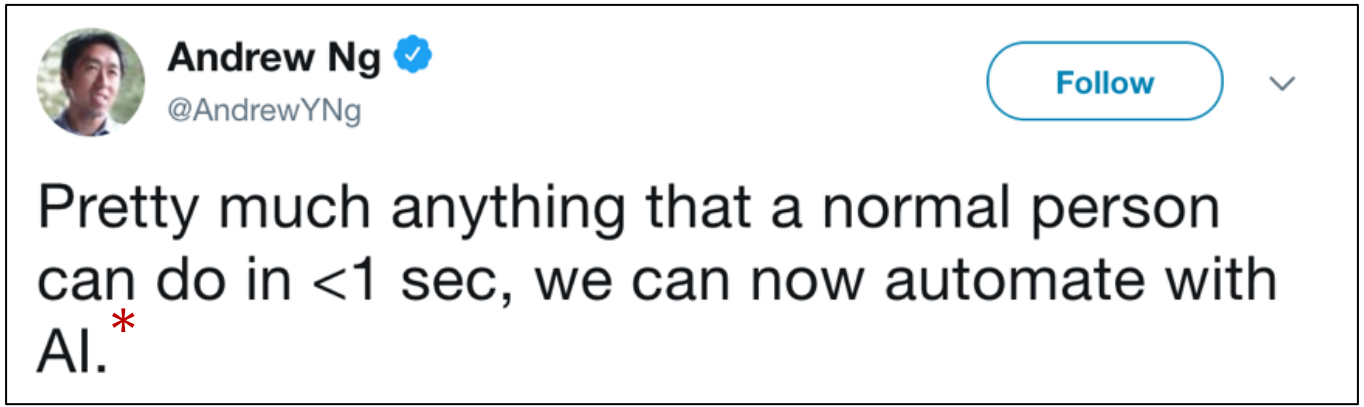
- Perceptual ad-blockers: how they work
- **Attacking perceptual ad-blockers**
- Why defending is hard

The Current State of ML

ML works well on average

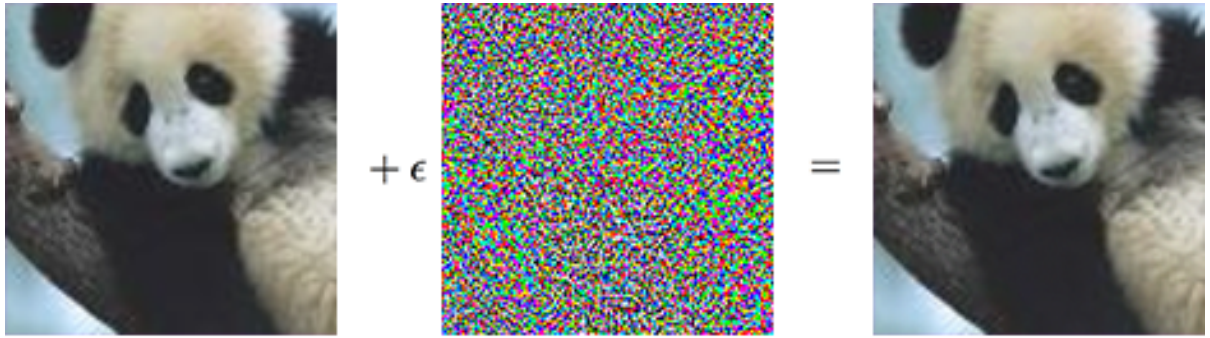
≠

ML works well on adversarial data



*as long as there is no adversary

Adversarial Examples



Szegedy et al., 2014
Goodfellow et al., 2015

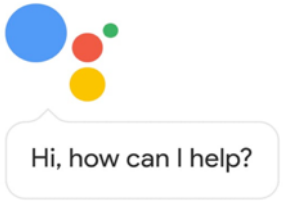
- **How?**
  - > *Training* ⟹ *"tweak model parameters such that* $f($  $) = panda$"
  - > *Attacking* ⟹ *"tweak input pixels such that* $f($  $) = gibbon$"

Adversarial Examples: A Pervasive Phenomenon

(Sharif et al. 2016)

(Kurakin et al. 2016)

(Athalye et al. 2018)



(Carlini et al. 2016,
Cisse et al. 2017,
Carlini & Wagner 2018)

(Eykholt et al. 2017)

(Eykholt **et al.** 2018)

(Meaningful) Defenses

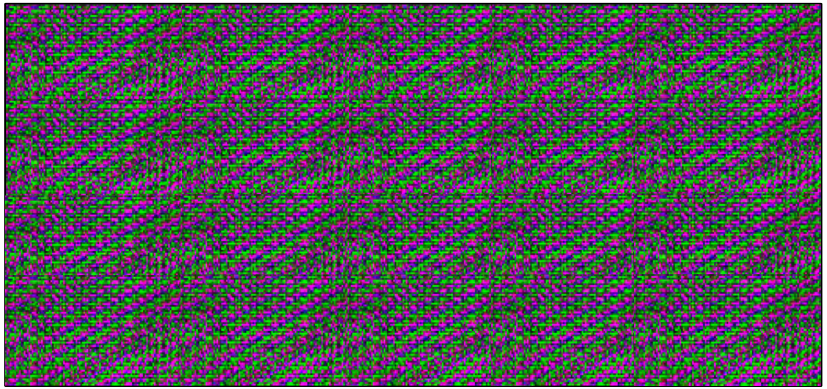
Adversarial Examples for Page-Based Perceptual Ad-Blockers

Ad-Block Evasion

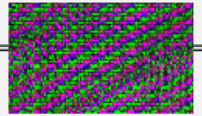
- **<u>Goal</u>: Make ads unrecognizable by ad-blocker**
- Adversary = Website publisher 
- Other adversaries exist (e.g., Ad-Network)

Evasion: Universal Transparent Overlay

- Web publisher perturbs every rendered pixel



```
<div id="overlay"></div>
```



```
#overlay {
  background-image:
    url("...");
  width: 100%; height: 100%; top: 0; left: 0;
  position: fixed; z-index: 10000;
  opacity: 0.01;
  pointer-events: none;
}
```

Use HTML *tiling* to minimize perturbation size (20 KB)

- 100% success rate on 20 webpages not used to create the overlay
- The attack is **universal:** the overlay is computed once and works for all (or most) websites
- Attack can be made more stealthy without relying on CSS

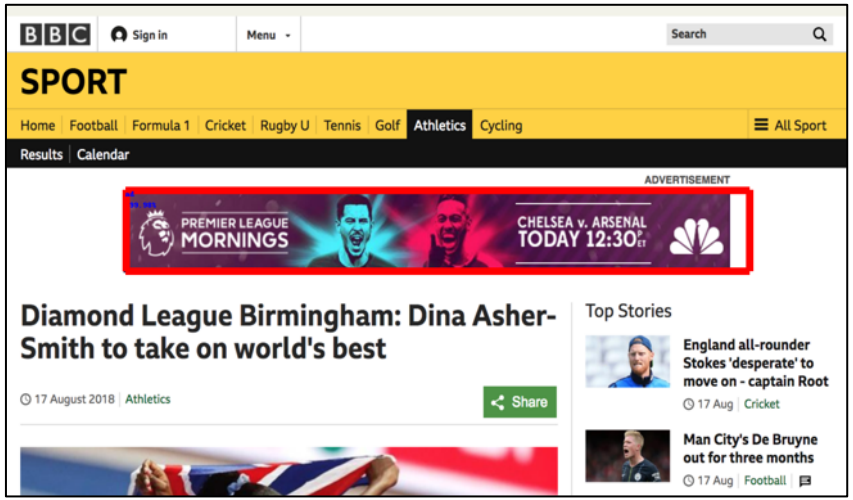
Ad-Block Detection

- **<u>Goal</u>: Trigger ad-blocker on "honeypot" content**
  - > *Detect ad-blocking in client-side JavaScript or on server*
  - > *Applicability of these attacks depends on ad-blocker type*

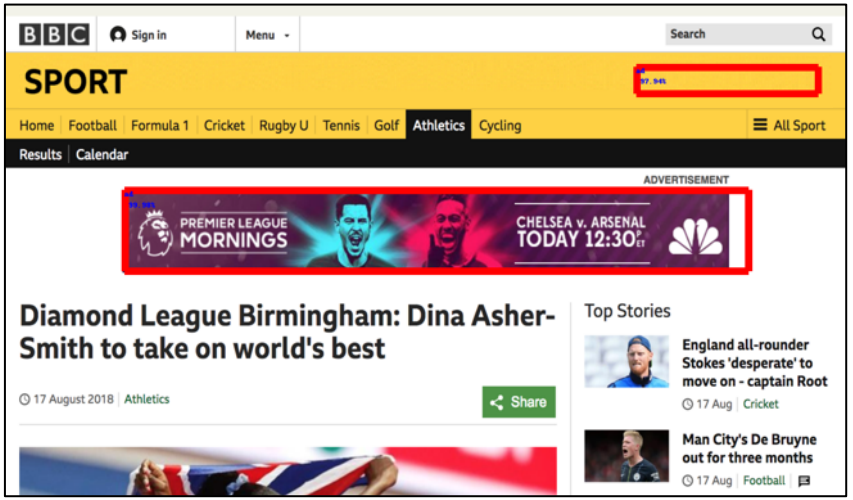
- Adversary = Website publisher
  - > *Use client-side JavaScript to detect DOM changes*

Detection: Perturb fixed page layout

- Publisher adds honeypot in page-region with fixed layout
  - > *E.g., page header*



original

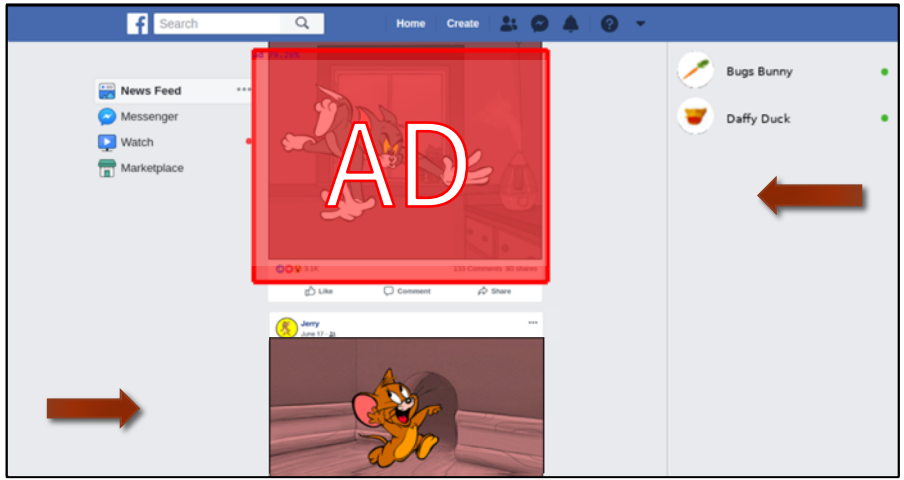


With honeypot header

New Threats: Privilege Abuse

- Ad-block evasion & detection is a well-known arms race. But there's more!

Jerry uploads malicious content …



… so that Tom's post gets blocked

What happened?

- *Object detector model generates box predictions from full page inputs*
- *Content from one user can affect predictions anywhere on page*
- *Model's segmentation is not aligned with web-security boundaries*

Outline

- Perceptual ad-blockers: how they work
- Attacking perceptual ad-blockers
- **Why defending is hard**

A Challenging Threat Model

- *Adversary has white-box access to ad-blocker*
- *Adversary can exploit False Negatives and False Positives in classification pipeline*
- *Adversary prepares attacks offline* ⇔ *The ad-blocker must defend against attacks in real-time in the user's browser*
- *Adversary can take part in crowd-sourced data collection for training the ad-blocker*

Defense Strategy 1: Obfuscate the Model

- Attacks are easy if the adversary has access to the ML model
  - *Solution: hide model from adversary?*

- Idea 1: Obfuscate the ad-blocker?
  - *It isn't hard to create adversarial examples for **black-box** classifiers*

- Idea 2: Randomize the ad-blocker?
  - *Deploy different models*
    - Adversarial examples that work against multiple models
  - *Randomly change page before classifying*
    - Adversarial examples robust to random transformations

Defense Strategy 2: Anticipate and Adapt

- If ad-blocker is attacked (evasion or detection), collect adversarial samples and re-train the model
  - *Or train on adversarial examples proactively*

- This is called **Adversarial Training** **(Szegedy'14)**
  - *New arms-race: The adversary finds new attacks and ad-blocker re-trains*
  - *Mounting a new attack is much easier than updating the model*
  - *On-going research: so far the adversary always wins!*

Adversarial Training: Current state of affairs

- Confer some robustness to a specific type of perturbation
  - > *CIFAR10:* *99% clean accuracy*
    50% accuracy at l_∞= 8/255
  - > *ImageNet:* *85% clean accuracy*
    45% at l_2 = 255 (1 px change)

- What about multiple perturbations? (with Dan Boneh, NeurIPS 2019)
  - > *Lose 5-20% accuracy points when training against two perturbation types*
  - > *We show provable tradeoffs in robustness for natural statistical models*

Defense Strategy 3: Simplify the Problem

- Storey et al: recognize ad-disclosures
  - > *Simpler computer vision problem than full-page ad-detection*
  - > *Light-weight and mature techniques (OCR, perceptual hashing, SIFT)*

- Adversarial Examples still exist

Take Away

- Emulating human detection of ads *could be* the end-game for ad-blockers
- But very hard with current computer vision techniques
  - *Resisting adversarial examples is a challenging open problem*
- Perceptual ad-blockers have to survive a strong threat model
  - *Similar attack for non-Web ad-blockers (e.g., Adblock Radio)*

ftramer / **ad-versarial**

- Train a page-based ad-blocker
- Download pre-trained models
- Attack demos

https://github.com/ftramer/ad-versarial

http://arxiv.org/abs/1811.03194