AutoFR

Automated Filter Rule Generation for Adblocking

SCAN ME

ProperData UCI UCDAVIS


Hieu Le
levanhieu.com


Salma Elmalaki


Athina Markopoulou


Zubair Shafiq

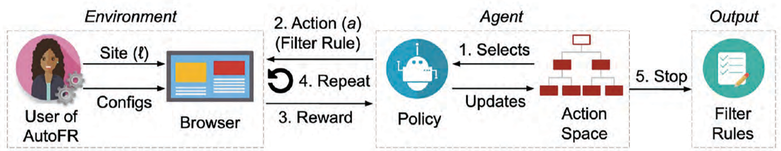
Motivation

- Filter rules power privacy-enhancing technologies (PET) to block ads and tracking, but they are still maintained by human experts.
- Researchers utilize machine-learning (ML) approaches to replace or assist filter rule creation. Yet, they rely on existing rules to label their ground truth for ML models, creating a circular dependency.

AutoFR Contributions

- **Framework:** Using reinforcement learning (RL) to automate URL-based rule generation that block ads without using existing rules.
- **User preference (w)**: A knob to express user preference of avoiding visual breakage (missing images/text), automatically.
- **Practical and Scalable**: AutoFR is a tool that takes 1.6 minutes per-site and can scale to millions of sites and over time.

Reinforcement Learning Framework



(b) **AutoFR (Automated) Workflow.** AutoFR automates these steps as follows: (1) the agent selects an action (*i.e.,* filter rule) following a policy; (2) it applies the action on the environment; (3) the environment returns a reward, used to update the action space; (4) the agent repeats the process if necessary; (5) the agent stops when a time limit is reached or no more actions are available to be explored. The human FL author only provides site ℓ and configurations (*e.g.,* threshold w and hyper-parameters).

Breakage

$$\mathcal{B} = \frac{\widehat{C}_I + \widehat{C}_T}{2}$$

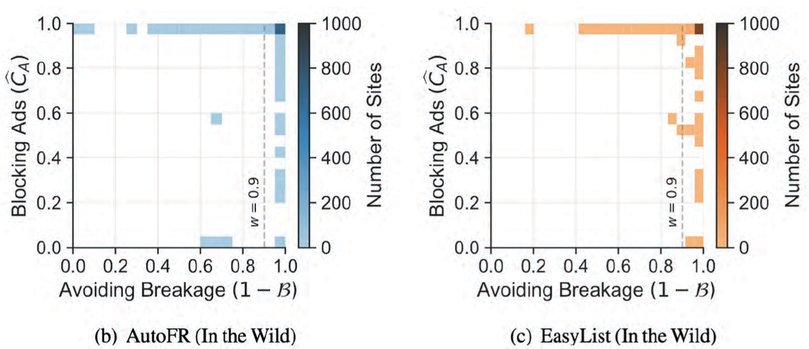
Reward

$$\mathcal{R}_F(w, \widehat{C}_A, \mathcal{B}) = \begin{cases} -1 & \text{if } \widehat{C}_A = 0 & \text{(3a)} \\ 0 & \text{if } \widehat{C}_A > 0\,,\ 1-\mathcal{B} < w & \text{(3b)} \\ \widehat{C}_A & \text{if } \widehat{C}_A > 0\,,\ 1-\mathcal{B} \geq w & \text{(3c)} \end{cases}$$

Effectiveness of a Filter Rule: Is measured by the proportion of ads ($\mathbf{C_A}$) that were removed (good) vs. how much legitimate content, such as images ($\mathbf{C_I}$) and text ($\mathbf{C_T}$), went missing (bad), when applying the rule on a site. Threshold **w** is a user given preference to denote how much the user cares about avoiding breakage (**B**).

Evaluation: AutoFR vs. EasyList

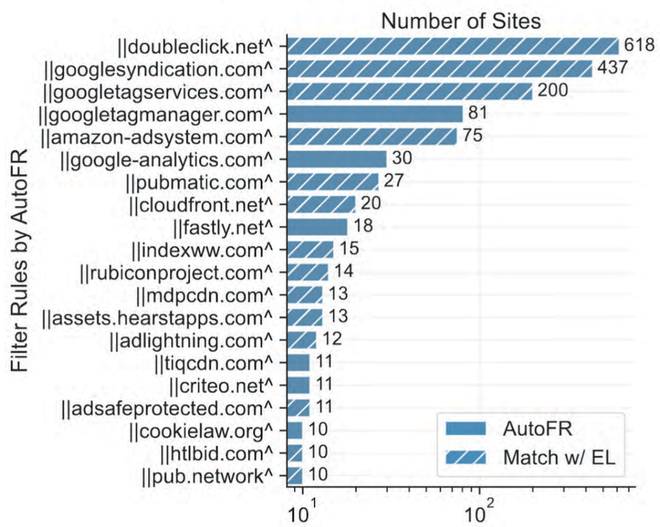
Performance



(b) AutoFR (In the Wild) (c) EasyList (In the Wild)

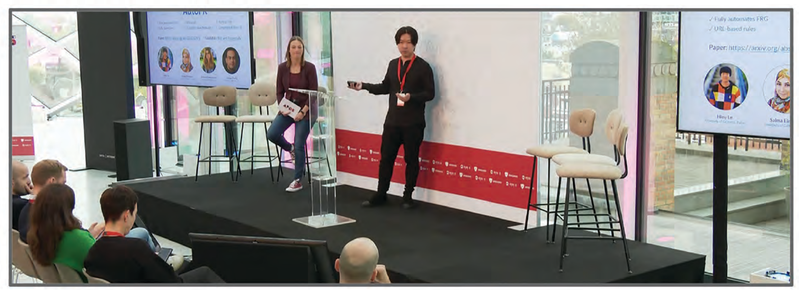
- **86% of sites (AutoFR) vs. 87% (EasyList):** We apply AutoFR to the Top-5K websites in the wild and find that it achieves comparable results for blocking ads while avoiding visual breakage within the given threshold (w=0.9) to EasyList, the state-of-the-art filter list used by PET.

Filter Rules Generated



- **361 Rules Generated:** 70% of the Top-20 rules generated by AutoFR appear in EasyList (EL). 40% of these rules match with the Top-20 rules from EasyList.

Broader Impact



- **Ad-Filtering Dev Summit:** Hieu Le presented AutoFR to the PET community on Oct. 2022 and is working closely with the community to get AutoFR adopted by industry players.

Future Directions

- **Other Platforms**: Extend AutoFR to platforms such as mobile, smart TVs and Oculus VR.
- **Tracking and Functionality**: Create rules to block tracking and avoid functionality breakage (forms, scrolling).