



A Statistical Framework for Deepfake Detection

Shannon Gallagher

Dominic Ross

Jeffrey Mellon

Catherine Bernaciak

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Document Markings

Copyright 2022 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.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

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.

Carnegie Mellon® and CERT® are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM22-0674

Deepfakes are “believable media generated by neural networks” (Mirsky and Lee 2021)

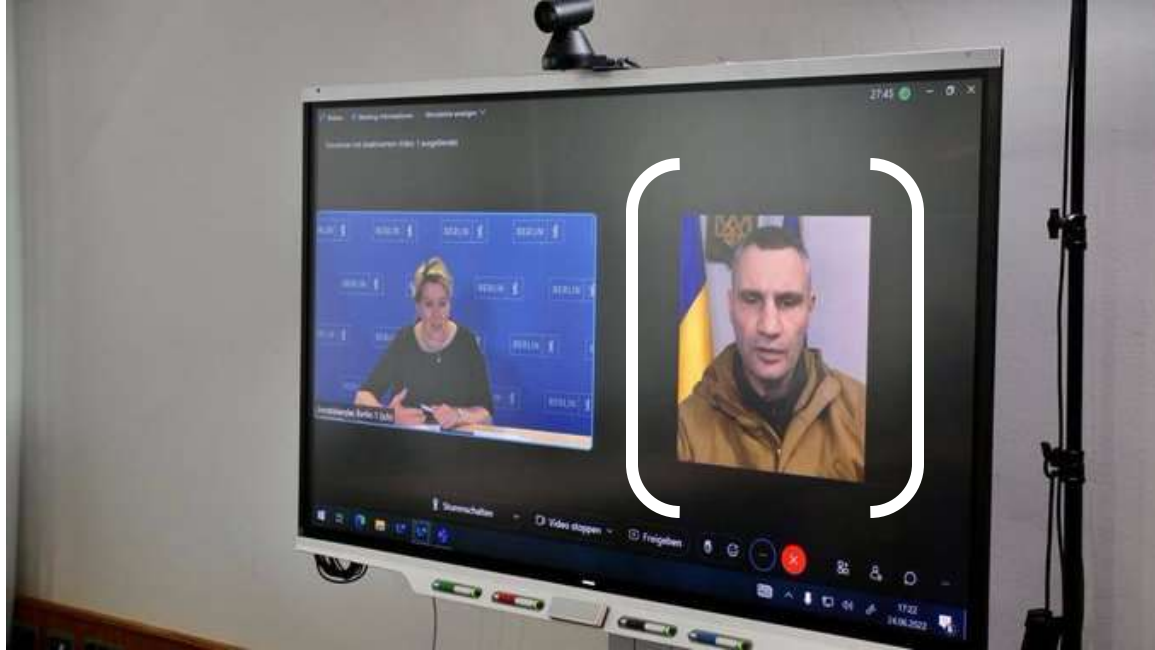


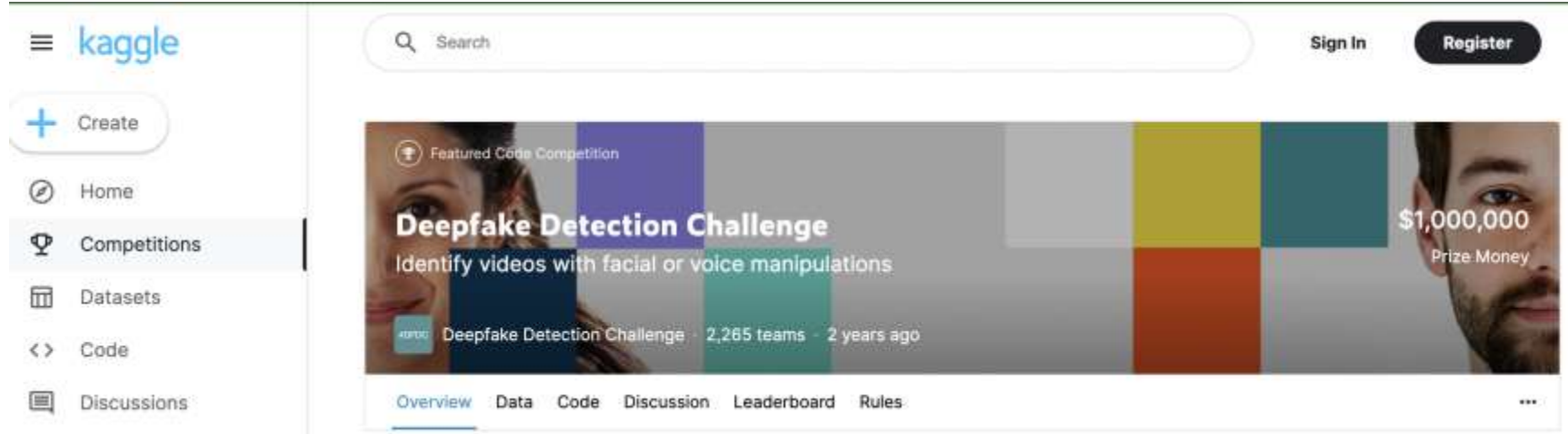
Image from DW.com. The righthand side image is an example of a deepfake used to impersonate the mayor of Kyiv. Brackets are ours.

Potential Dangers:

- Impersonation of political figures and celebrities
- Defamation of citizens
- Extension of ‘catfishing’
- >100k hours of video uploaded to web every day!

We need robust detectors!

No shortage of methods, but reproducibility is difficult

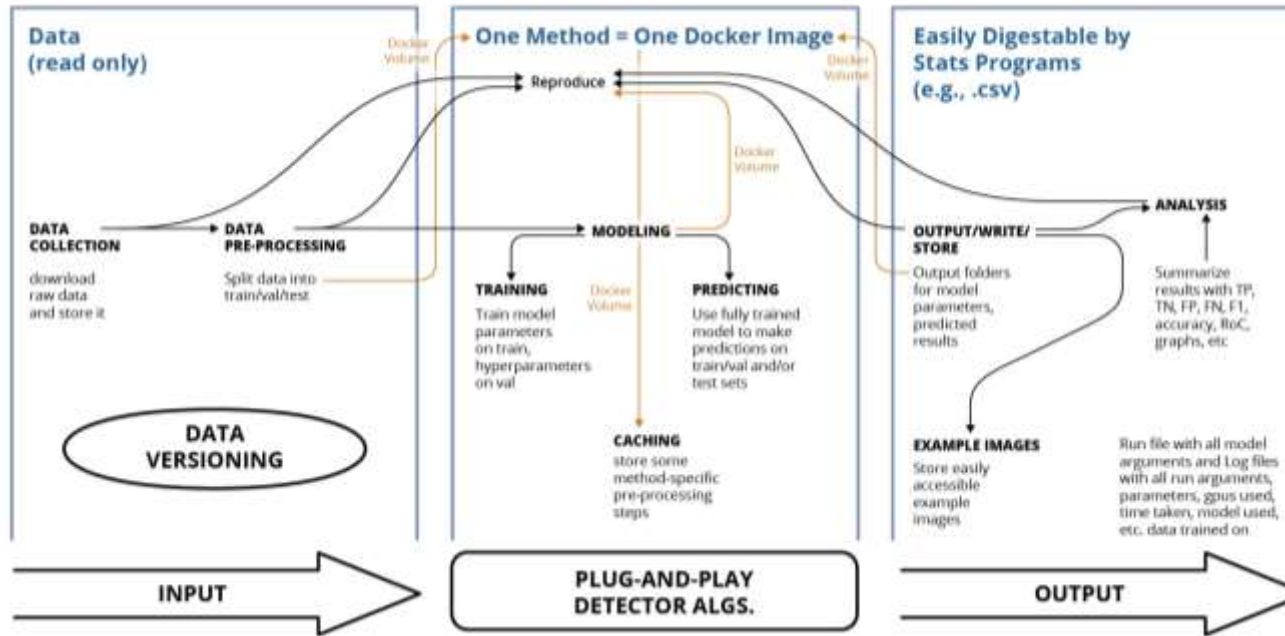


Challenges:

- Data formats (storage, image vs. video, resolution, source)
- Non-generalizable code (will work only in given scenario)
- Dynamic software (tensorflow, pytorch, opencv, pillow)
- Hardware (GPUs, CPU, data storage, cost)
- Uninterpretable algorithms (usually work well on one data set but not others)

Our solution: Deepfake Detection Pipeline (DDP)

End-to-End Process



DDP is reproducible, portable, and modular

DDP's backend is SEI's Juneberry – a publicly accessible tool!

Some preliminary results: generalization is difficult

Accuracy (%) of fine-tuned ResNet

Tested on

<i>Trained on</i>	Data Set	Celeb DF v1	Stylegan2	Stylegan3-r	Stylegan3-t	DFDC Pt. 0
	Celeb DF v1	99.1	44.2	44.0	44.2	
	Stylegan2	24.1	98.7	48.4	52.9	
	Stylegan3-t	16.7				
	Stylegan3-r	16.9	68.0	97.2	89.0	
	DFDC Pt. 0					

Looking forward: Ensembles, Data, and GANs

Ensembles

- Time (real time pred.?)
- Accuracy and other metrics
- Feature-based

Data

- Different orgs. have different data with different levels of privacy
- Improved generated data

GANs

- Generative adversarial networks
- Can use our detectors against us
- Should we release detectors?
- How can we defend detectors?