

RESEARCH REVIEW 2022

**Carnegie
Mellon
University**
Software
Engineering
Institute

A Machine Learning Pipeline for Deepfake Detection

NOVEMBER 14–16, 2022

Shannon Gallagher
Dominic Ross

Jeffrey Mellon
Catherine Bernaciak

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

©2022

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.

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.

DM22-0838

Deepfakes Detection Team



Shannon Gallagher

Data Scientist



Dominic Ross

Multi-Media Design and
Communications Lead



Jeffrey Mellon

Machine Learning Research Scientist

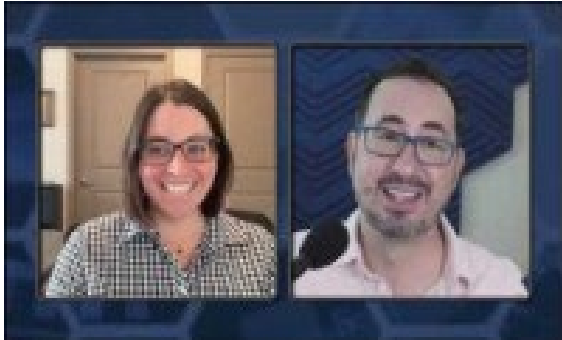


Catherine Bernaciak

Senior Machine Learning Research
Scientist

Deepfakes Are “Believable Media Generated by Deep Neural Networks” [Mirsky and Lee 2020]

Detecting Deepfakes



Shannon and Dominic discuss what deepfakes are and how their team is building artificial intelligence and machine learning technology to distinguish real from fake. They share well-known examples of deepfakes and discuss what makes them distinguishable as fake.

A Dive into Deepfakes



Shannon and Dominic discuss deepfakes, their exponential growth in recent years, their increasing technical sophistication, and the problems they pose for individuals and organizations. They also discuss the SEI’s research in this area.

Making and Detecting Deepfakes

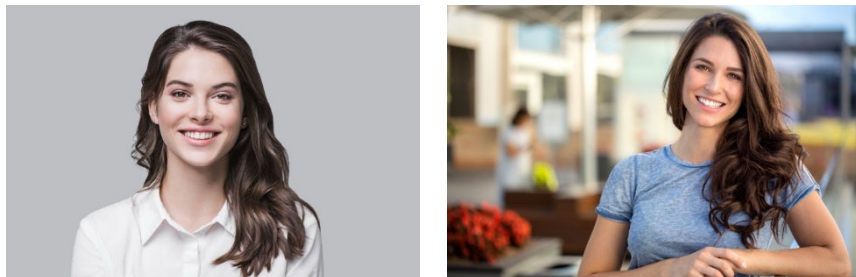


Catherine and Dominic describe the technology underlying the creation and detection of deepfakes and assessment of current and future threat levels.

[Mirsky and Lee 2020]

Mirsky, Y. & Lee, W. The Creation and Detection of Deepfakes: A Survey. 2020
<https://arxiv.org/abs/2004.11138>

Deepfakes Are Dangerous



Conceptual Example of a Faceswap Deepfake
The target's face is placed on the source's face.

Potential Dangers

- Impersonation of political figures and celebrities (e.g., mayor of Kyiv)
- Defamation of citizens
- Mis-, dis-, and mal- information

>700k hours of video are uploaded to the web every day!

We need fast and reliable detectors.

Numerous Deepfake Detection Methods Already Exist

REAME.md

DeepFake Detection (DFDC) Solution by @selimsef

Challenge details:

[Kaggle Challenge Page](#)

Fake detection articles

- [The Deepfake Detection Challenge \(DFDC\) Preview Dataset](#)
- [Deep Fake Image Detection Based on Pairwise Learning](#)
- [DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection](#)
- [DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection](#)
- [Real or Fake? Spoofing State-Of-The-Art Face Synthesis Detection Systems](#)
- [CNN-generated images are surprisingly easy to spot... for now](#)
- [FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces](#)
- [FakeLocator: Robust Localization of GAN-Based Face Manipulations via Semantic Segmentation Networks with Bells and Whistles](#)
- [Media Forensics and DeepFakes: an overview](#)
- [Face X-ray for More General Face Forgery Detection](#)

Solution description

In general solution is based on frame-by-frame classification approach. Other complex things did not work so well on public leaderboard.

Face-Detector

MTCNN detector is chosen due to kernel time limits. It would be better to use S3FD detector as more precise and robust, but opensource Pytorch implementations don't have a license.

Languages

- Python 94.1%
- Shell 4.0%
- Dockerfile 1.9%

org

f.com

Google Scholar Image: Google and the Google logo are trademarks of Google LLC.

The DFDC screenshot is used with permission from Selim Seferbekov according to the [MDFDC DeepFake Detection Challenge MIT license](#).

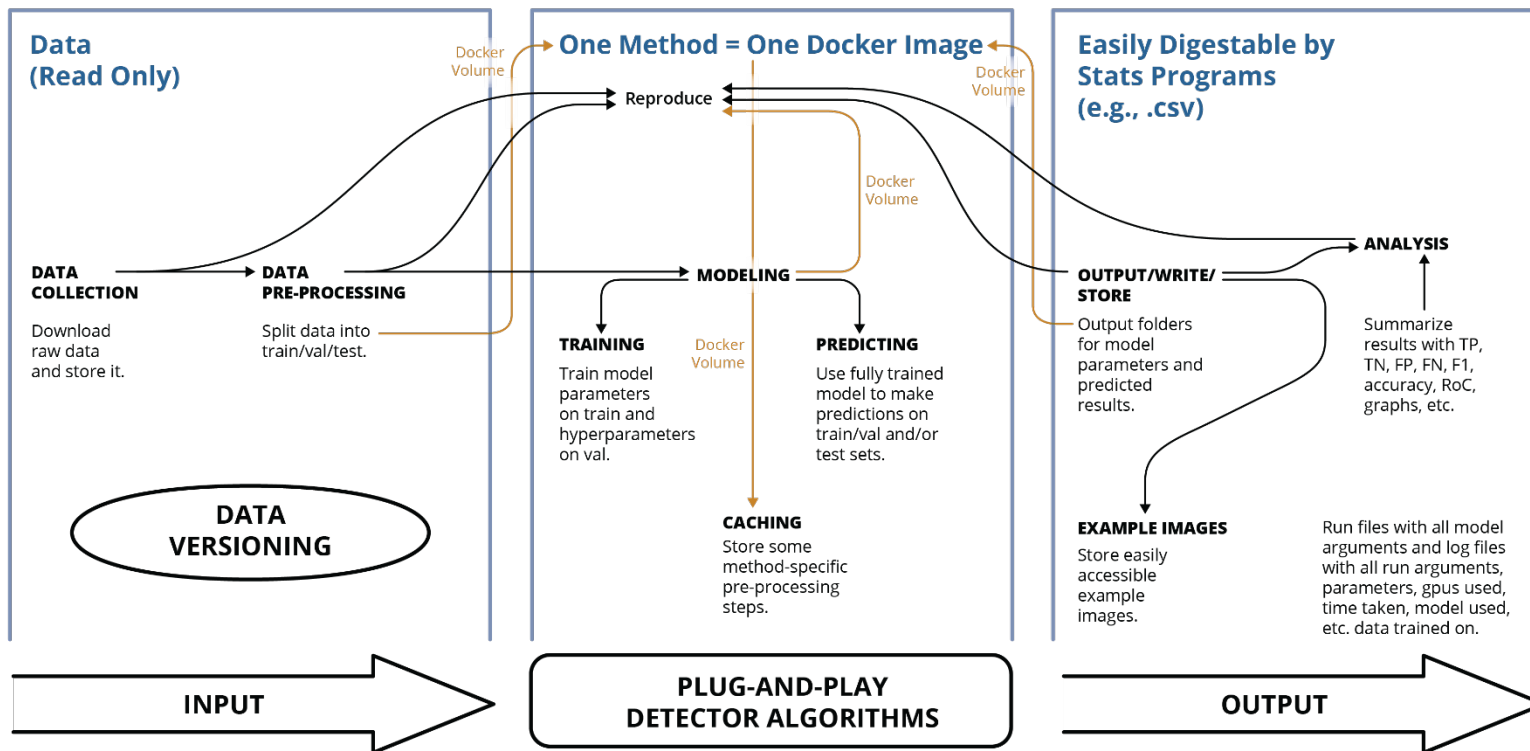
But Reproducing Results Is...Difficult

- Data and formats
- Sparsely documented code
- Changes to packages like opencv2, pillow, and others
- Changes to backends like PyTorch and TensorFlow
- Hardware

The best methods come with a docker run script, but even that can be difficult.

Takeaway: It is difficult to compare methods side by side (e.g., benchmarks).

Our Deepfake Detection Pipeline (DDP) Creates Benchmarks



DDP is reproducible, portable, and modular.

DDP's backend is [SEI's Juneberry](#).

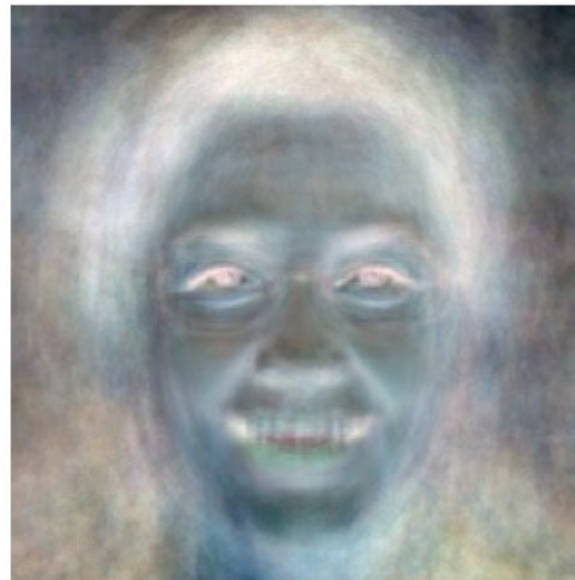
Several Publicly Available Data Resources

Source	Format	Amount	Label(s)	License?
DeepFake Detection Challenge (DFDC)	.mp4	100k+ Videos	Real/Fake	Yes
Celeb-DF	.mp4	5600+ Videos	Real/Fake	Yes
StyleGAN3	.png	Generate Your Own Portraits	Fake	Yes
Flickr-Faces-HQ	.png and .json	70k Portraits	Real	Yes

A Look at the “Average” Faces

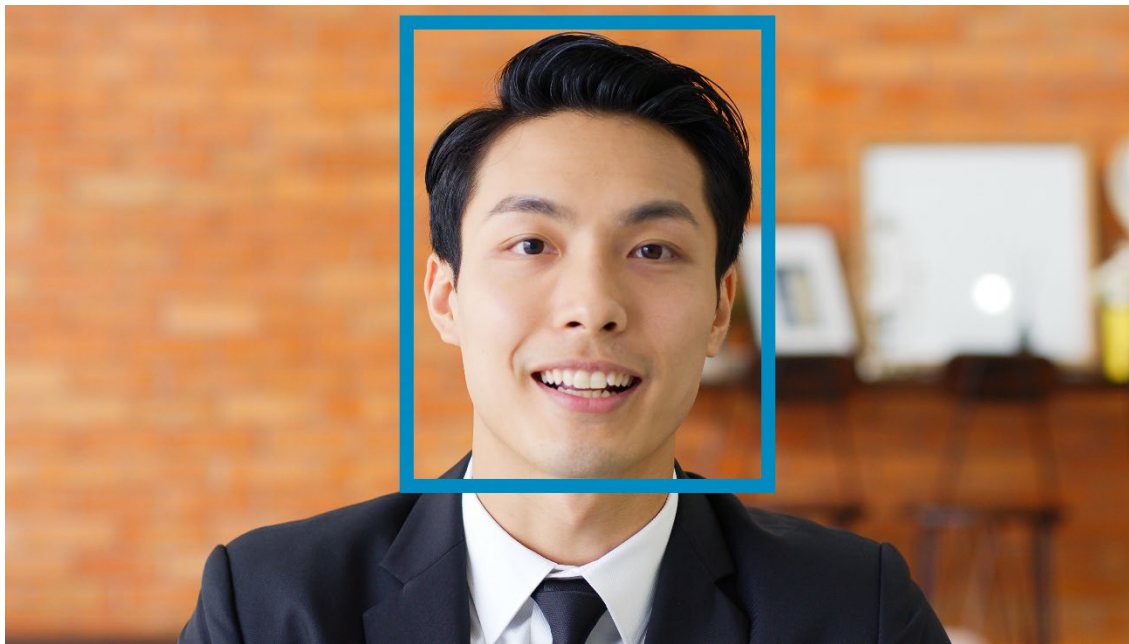
- Hairline
- Edges of eyes
- Corners of mouth
- Chin
- Eyebrows
- Nose
- Boundaries

AVG(REAL) - AVG(FAKE)



We've Noticed a General Trend in Detection Methods

1. Find the face.



We've Noticed a General Trend in Detection Methods

1. Find face the face.
2. Extract facial landmark(s) and normalize them.



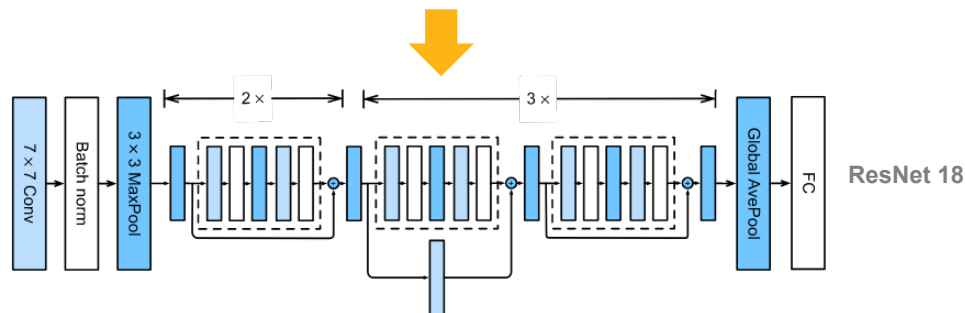
We've Noticed a General Trend in Detection Methods

1. Find the face.
2. Extract facial landmark(s) and normalize them.
3. Apply masking and/or add noise.



We've Noticed a General Trend in Detection Methods

1. Find the face.
2. Extract facial landmark(s) and normalize them.
3. Apply masking and/or add noise.
4. Send to a pre-trained image detector.



The ResNet 18 chart is reused with permission from Zachary C. Lipton, co-author of *Dive into Deep Learning*.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Pre-Trained Models Available to DDP

[AlexNet](#)

[MNASNet](#)

[SqueezeNet](#)

[ConvNeXt](#)

[MobileNet V2](#)

[SwinTransformer](#)

[DenseNet](#)

[MobileNet V3](#)

[VGG](#)

[EfficientNet](#)

[RegNet](#)

[VisionTransformer](#)

[EfficientNetV2](#)

[ResNet](#)

[Wide ResNet](#)

[GoogLeNet](#)

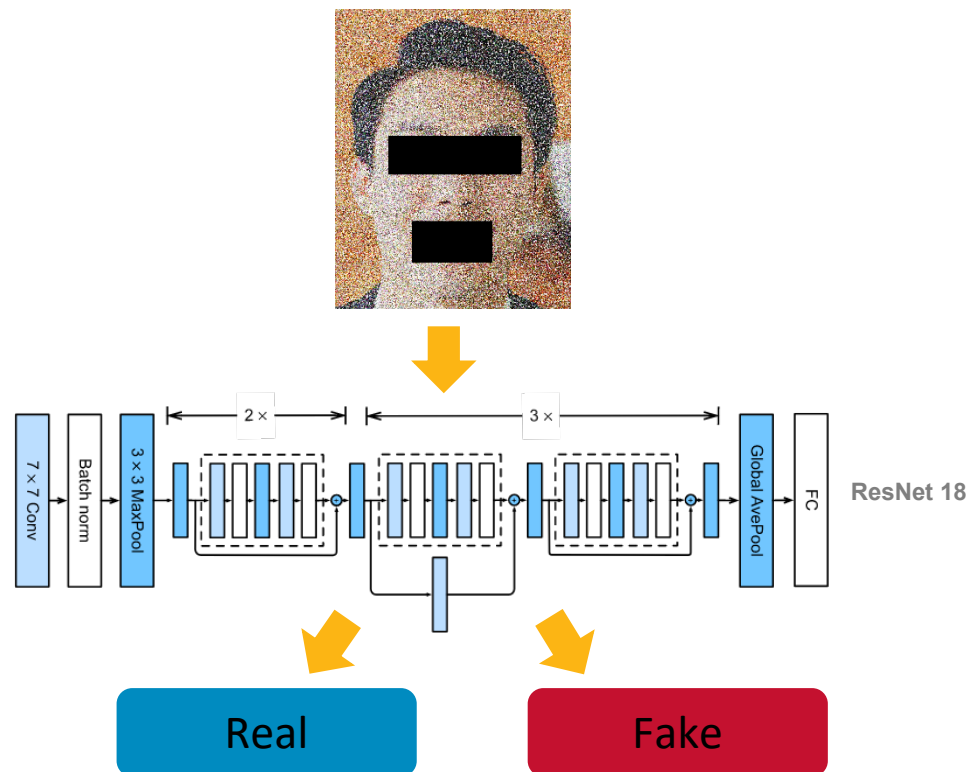
[ResNeXt](#)

[Inception V3](#)

[ShuffleNet V2](#)

We've Noticed a General Trend in Detection Methods

1. Find the face.
2. Extract facial landmark(s) and normalize them.
3. Apply masking and/or add noise.
4. Send to a pre-trained image detector.
5. Fine-tune it.

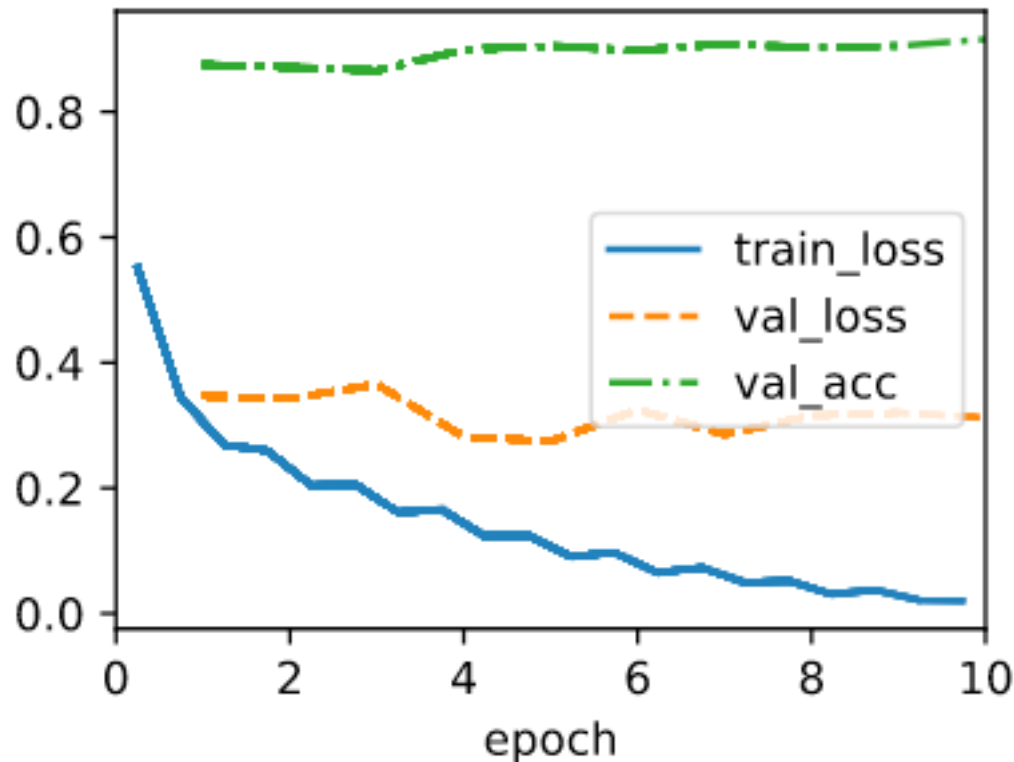


The ResNet 18 chart is reused with permission from Zachary C. Lipton, co-author of *Dive into Deep Learning*.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

We've Noticed a General Trend in Detection Methods

1. Find the face.
2. Extract facial landmark(s) and normalize them.
3. Apply masking and/or add noise.
4. Send to a pre-trained image detector.
5. Fine-tune it.
6. Evaluate it.



This chart is reused with permission from Zachary C. Lipton, co-author of *Dive into Deep Learning*.

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Preliminary Results with DDP

Accuracy (%) of Fine-Tuned ResNet

		Tested on				
		Celeb DF v1	Stylegan2	Stylegan3-t	Stylegan3-r	DFDC Pt. 0
Trained on	Data Set					
	Celeb DF v1	99.1	44.2	44.2	44.0	51.2
	Stylegan2	24.1	98.7	52.9	48.4	57.4
	Stylegan3-t	16.7	69.7	96.7	84.0	7.0
	Stylegan3-r	16.9	68.0	89.0	97.2	7.0
	DFDC Pt. 0	68.1	57.4	57.5	57.5	88.7

Next Steps

- Model robustness
- Video-specific detectors
- Improved detectors via ensemble models

Summary

- Deepfake detection methods need better benchmarks
 - Accuracy, cost, time
- We are doing that via **DDP and Juneberry**:
 - Data collection
 - Data transformation
 - Modeling
 - Evaluation
- Preliminary results confirm that generalizability is a problem.
 - We expect to **improve models** with ensemble detectors via DDP.