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

Document Markings

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

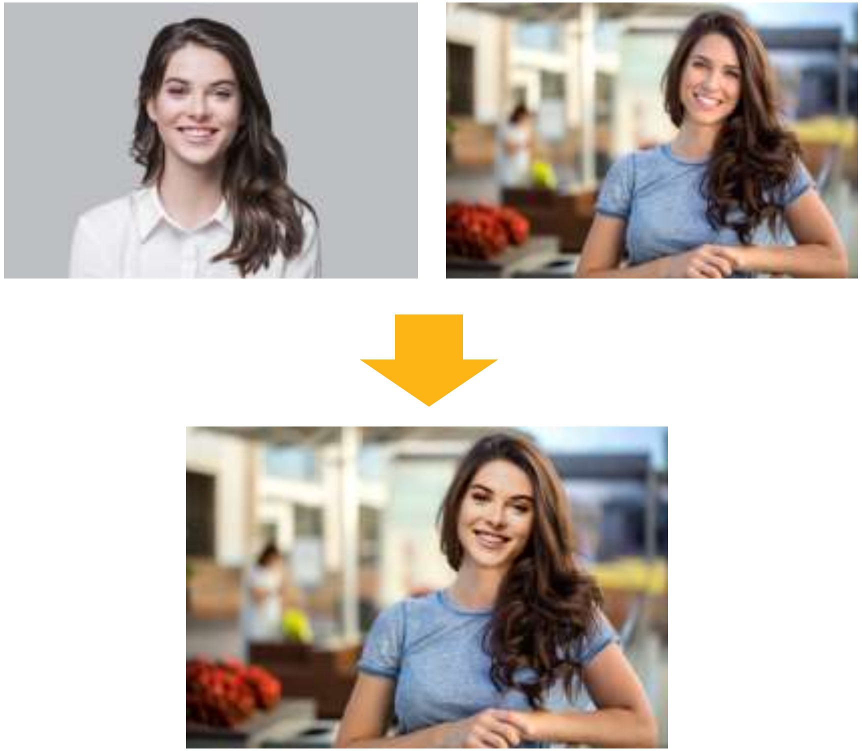
Shannon and Dominic discuss deepfakes, their exponential growth in recent years, their increasing technical sophistication, and the problems they pose for individuals and organizations. They also discuss the SEI’s research in this area.

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.

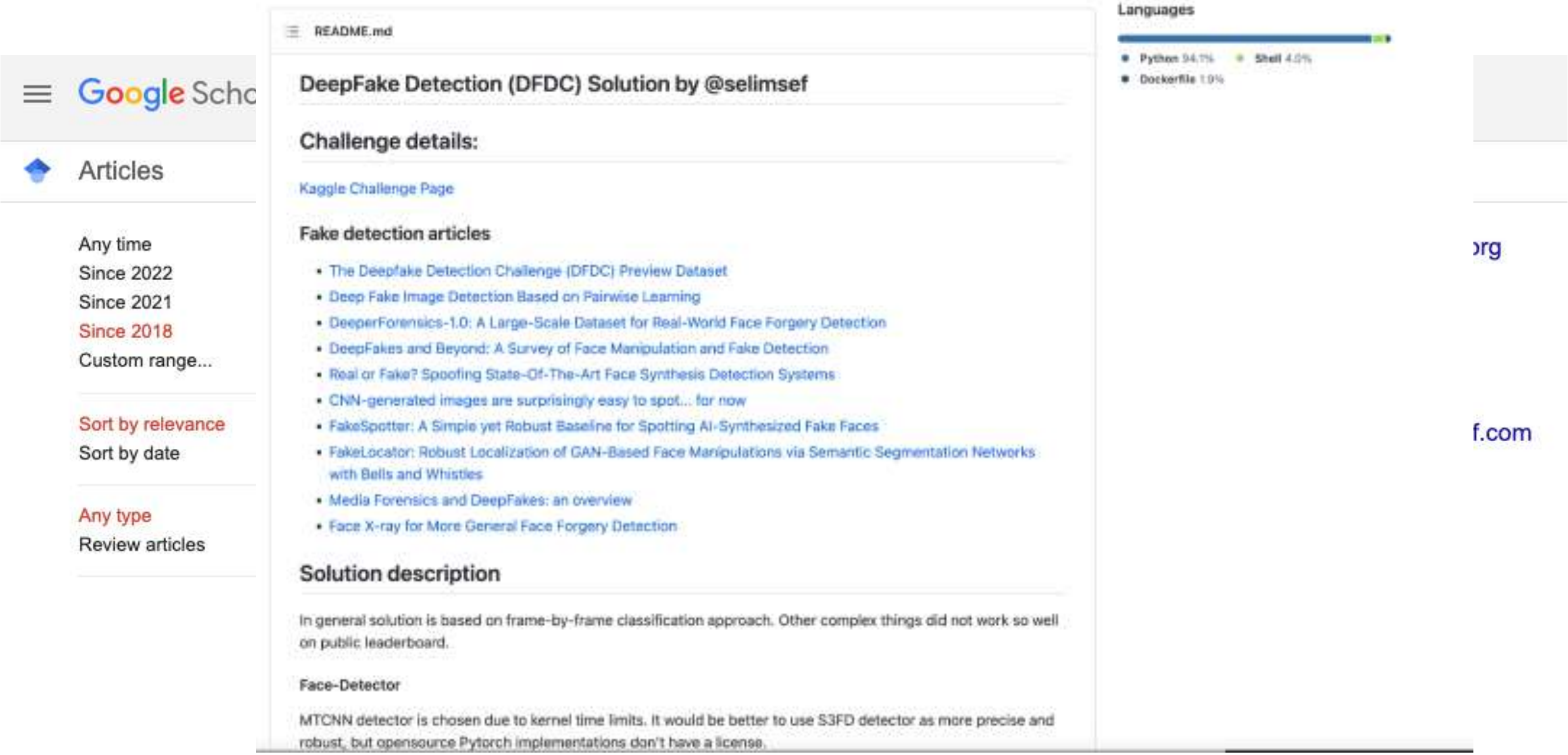
Potential Dangers

- 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



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](#).

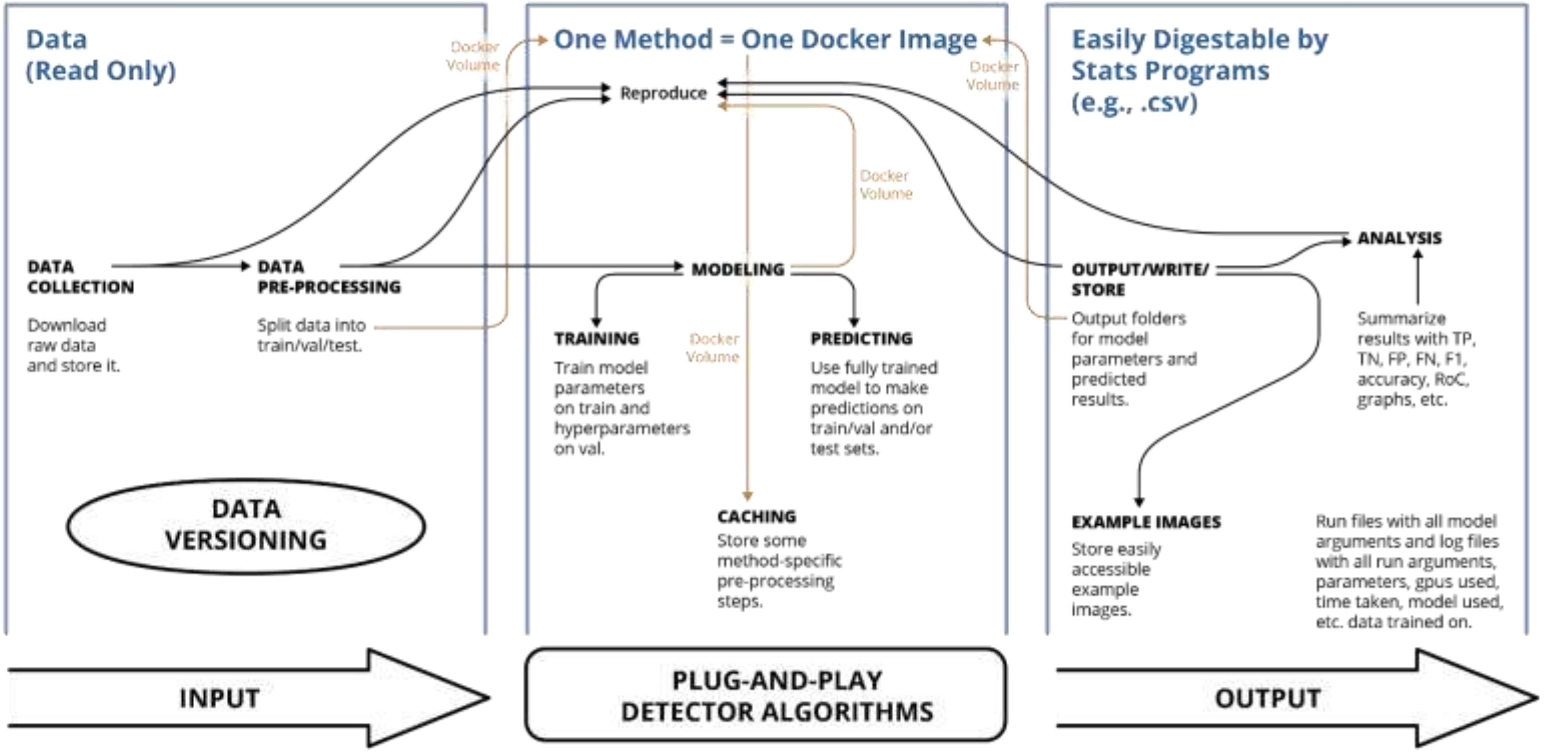
But Reproducing Results Is…Difficult

- 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.

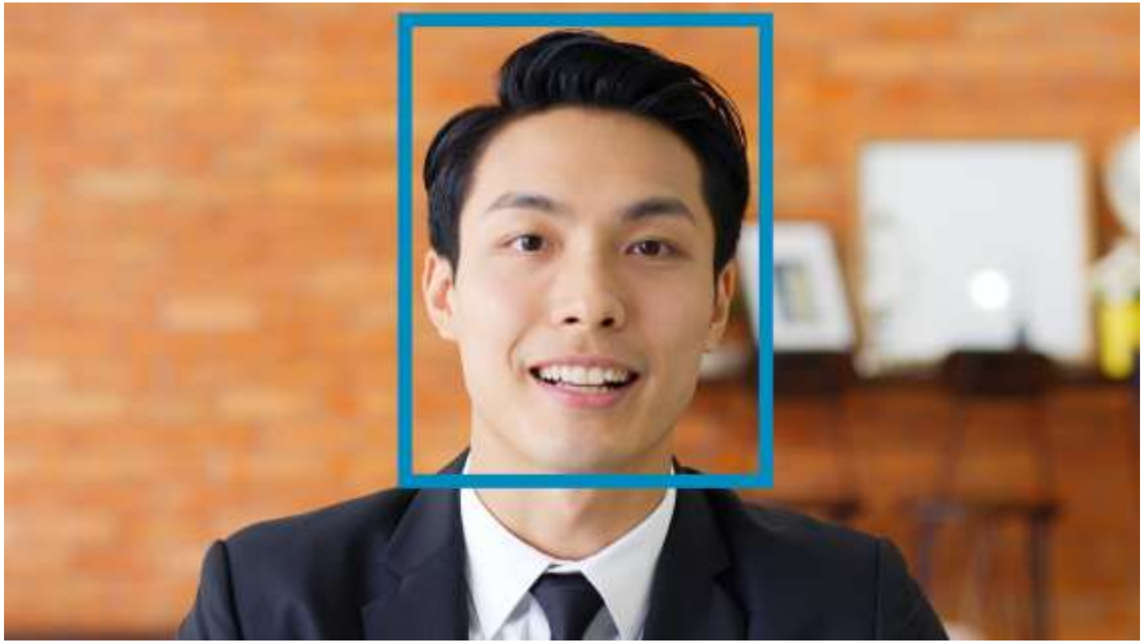
DDP's backend is SEI's Juneberry.

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

We've Noticed a General Trend in Detection Methods

1. Find the face.



We’ve Noticed a General Trend in Detection Methods

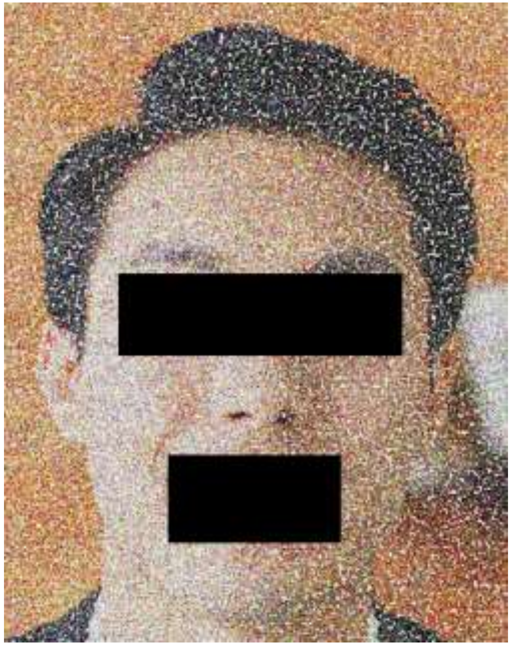
1. Find face the face.
2. Extract facial landmark(s) and normalize them.

Extraction, Masking, and Standardizing in DDP

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

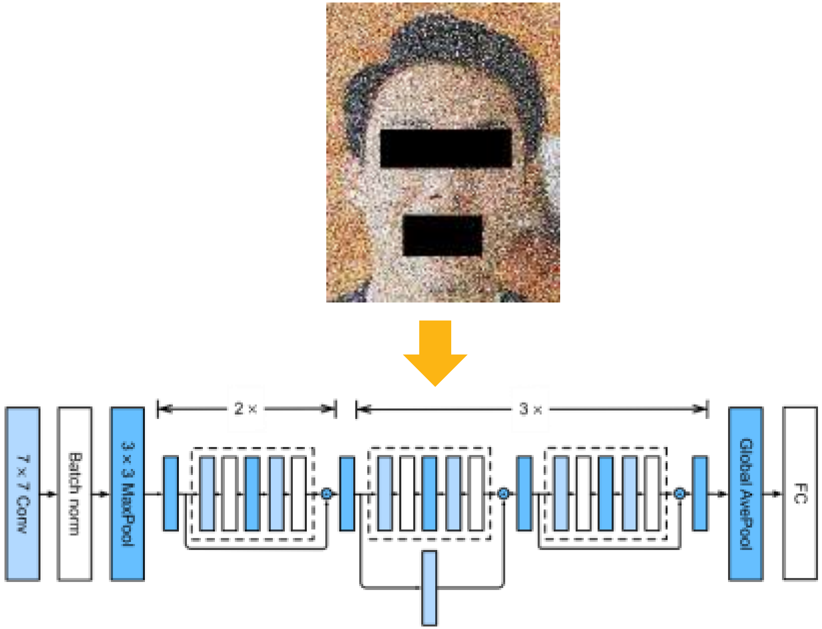
We’ve Noticed a General Trend in Detection Methods

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.



ResNet 18

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

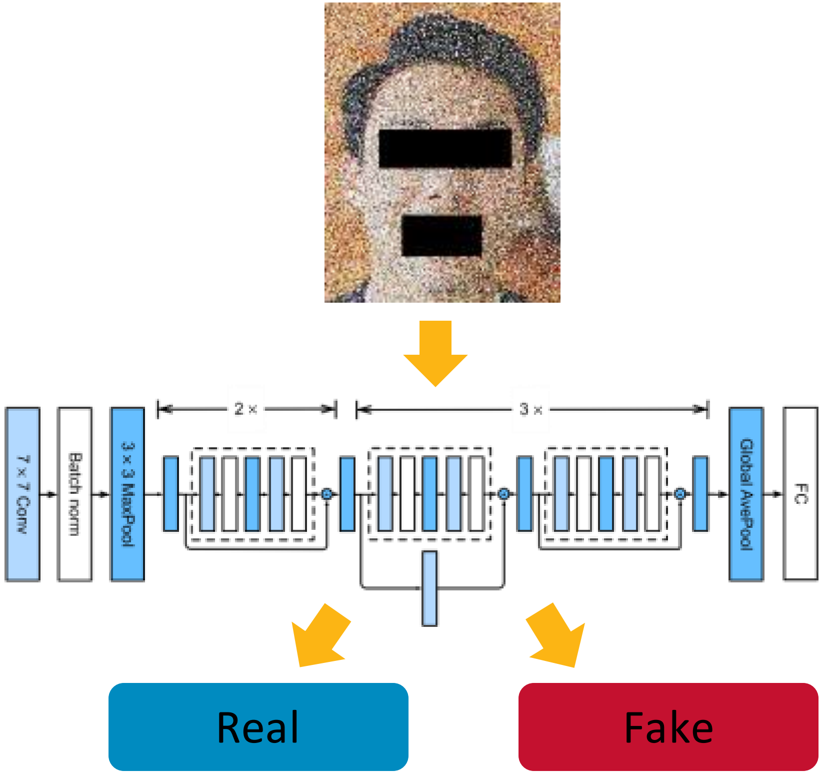
Pre-Trained Models Available to DDP

Carnegie Mellon University Software Engineering Institute

- AlexNet
- ConvNeXt
- DenseNet
- EfficientNet
- EfficientNetV2
- GoogLeNet
- Inception V3
- MNASNet
- MobileNet V2
- MobileNet V3
- RegNet
- ResNet
- ResNeXt
- ShuffleNet V2
- SqueezeNet
- SwinTransformer
- VGG
- VisionTransformer
- Wide ResNet

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.

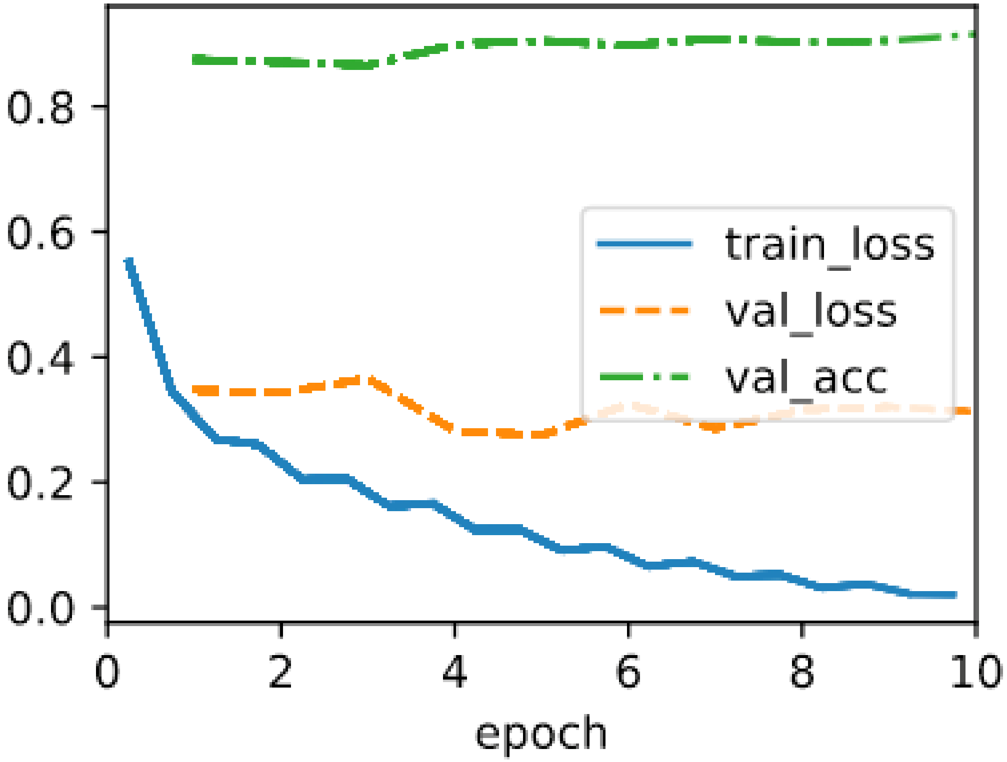


ResNet 18

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

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*.

Preliminary Results with DDP

Carnegie Mellon University Software Engineering Institute

Accuracy (%) of Fine-Tuned ResNet

Tested on

Trained on

Data Set	Celeb DF v1	Stylegan2	Stylegan3-t	Stylegan3-r	DFDC Pt. 0
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.