What type of organizational threats do deepfakes present?

Shannon K. Gallagher, PhD

skgallagher@sei.cmu.edu

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

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.

DM22-0986

Meet the Deepfake Team

Catherine Bernaciak

Dominic Ross

Jeffrey Mellon

Spot the deepfake. There is exactly one deepfake.

A

B

C

What is a deepfake?

"Believable media generated by a deep neural network"

-- Mirsky and Lee (2020)

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

Deepfakes are no longer theoretical threats only

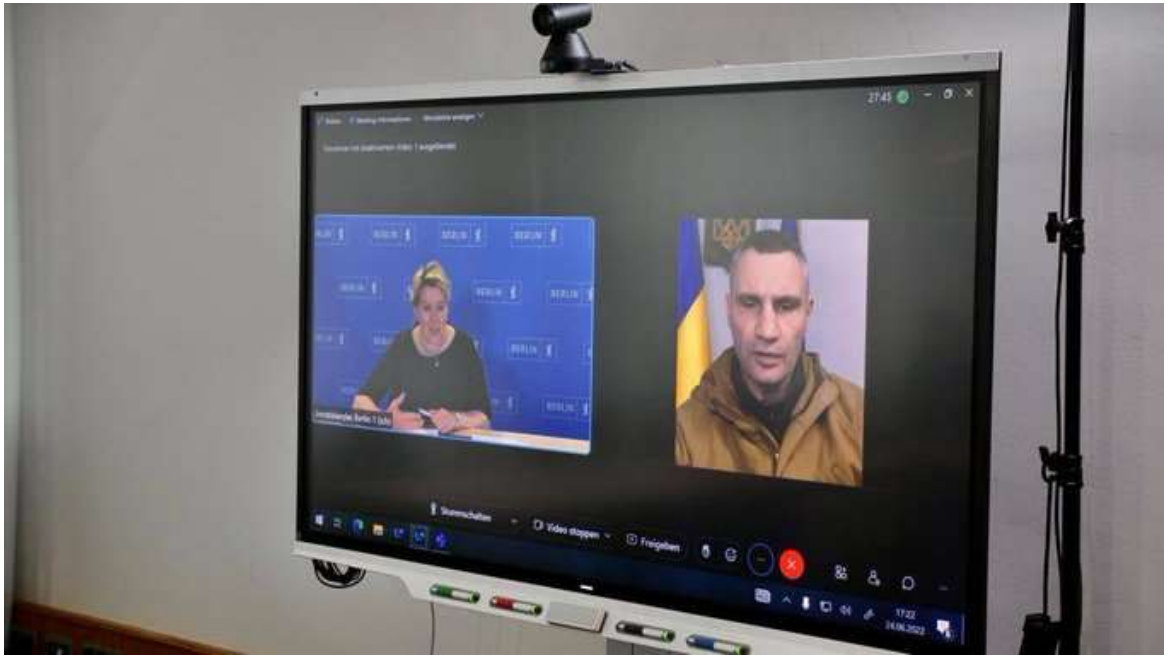


Image from DW.com. The righthand side image is an example of a deepfake used to impersonate the mayor of Kyiv. Brackets are ours.

Potential Dangers:

- Impersonation of political figures and celebrities
- Defamation of citizens
- Mis-, dis-, and mal-information
- >700k hours of video uploaded to web every day!

We need robust detectors!

And more deepfakes in the news…

A Deepfake Phone Call Dupes An Employee Into Giving Away $35 Million

Think your business is too small to be fooled? Think again.

By Gene Marks

January

A Telegram bot network is being used to create deepfake nudes

It has produced naked images of at least 104,000 women.

Deepfakes come to remote job interviews

Is that a real person you're interviewing, or are you talking to a stolen identity?

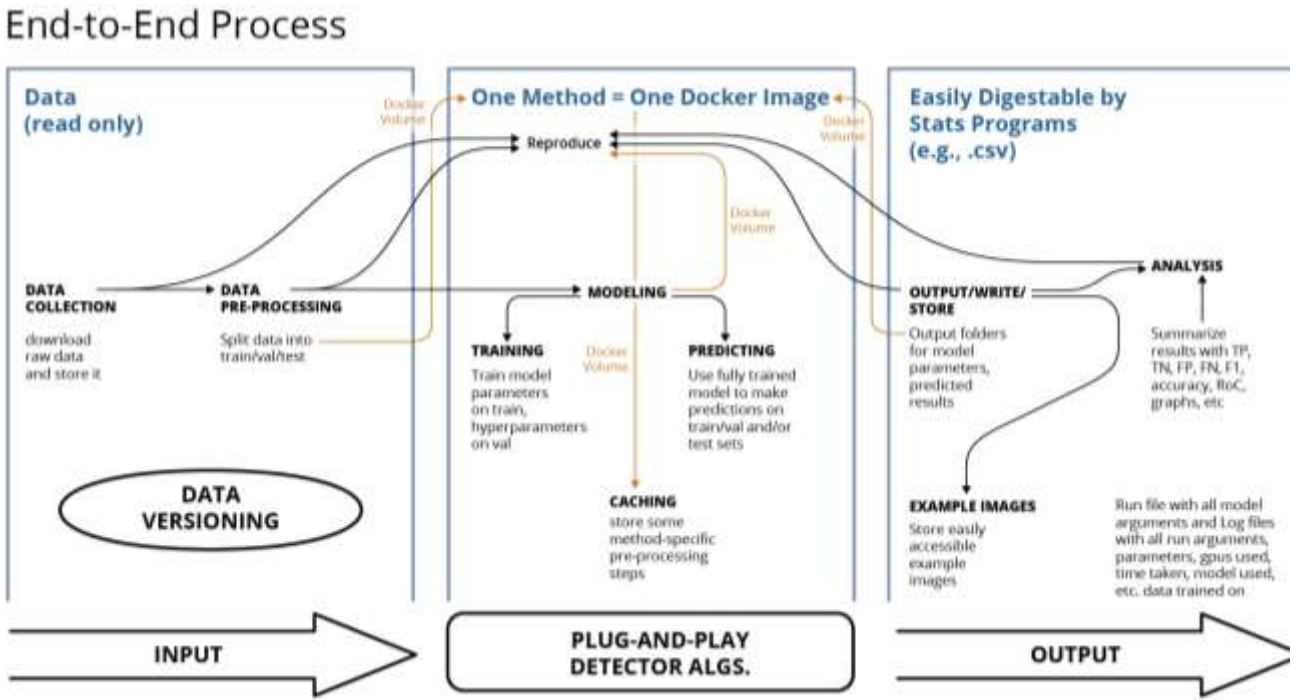
By Mike Elgan

Contributing Columnist, Computerworld | JUL 7, 2022 4:30 AM PDT

But there are differences between real and deepfake images

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

Our AI/ML solution: Deepfake Detection Pipeline (DDP)



DDP is reproducible, portable, and modular

DDP's backend is SEI's Juneberry

The GAN problem

Red makes a generator to create deepfakes

Blue makes a detector

Red uses results from blue's detector to make generator better

Blue uses new red images to improve detector

…

Who wins?

SEI Resoruces

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

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 Surv

In closing

- Deepfake detection needs to be streamlined

- We are doing that via DDP and Juneberry
  - Data collection
  - Data transformation
  - Modeling
  - Evaluation

- We need to counter the GAN problem

GAN set-up

$X' = G(Z)$ = generator fake image
X = real image
$D(X)$= discriminator in [0,1]
Y=0, Y'=1 (0 is real, 1 is fake)

Data =$\{(X_i, 0)_{i=1\ldots m}, (X'_j, 1)_{j=1\ldots n}\}$

$L(x, y) =$ loss function

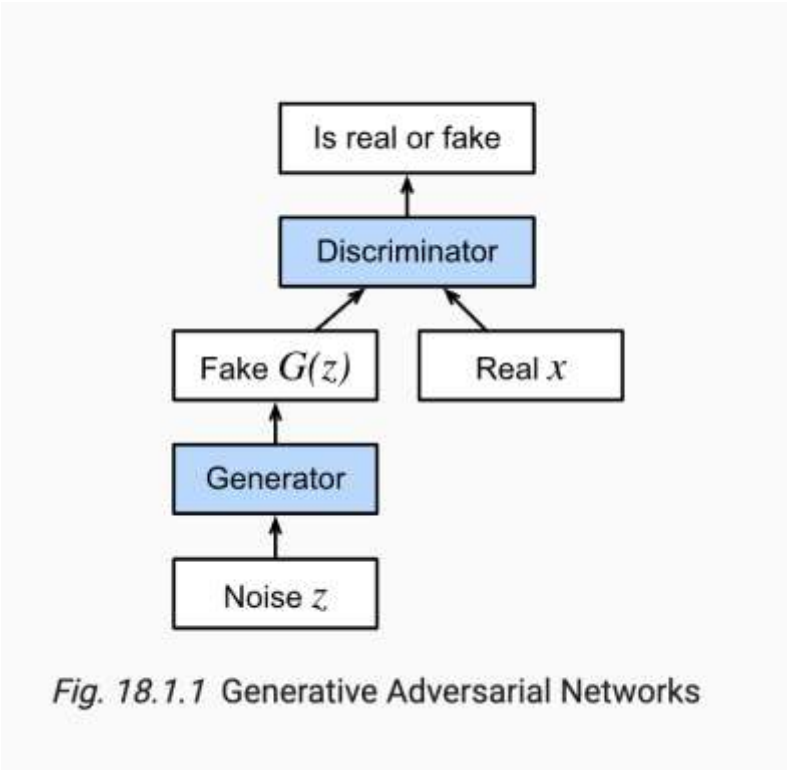


Fig. 18.1.1 Generative Adversarial Networks

Fig. from d2l.ai

GAN Set-up

Round 0: Generator introduces fakes

Round 1:

Discriminator turn: Use generated data to get best discriminator

$$\widehat{D^1}|\widehat{G^0} = argmin_D \sum_{\{i=1\}}^{m} L(D(X_i), 0) + \sum_{\{j=1\}}^{n} L(D(X_j'), 1)$$

Generator turn: Directly try to deceive discriminator

$$\widehat{G^1}|\ \widehat{D^1} = argmin_G \sum_{\{j=1\}}^{n} L(\widehat{D^1}(X_j'), 0)$$

$$= argmin_G \sum_{\{j=1\}}^{n} L(\widehat{D^1}(G(Z_j)), 0)$$

Repeat

Preliminary Results with DDP

Accuracy (%) of fine-tuned ResNet

Trained on \ *Tested on*	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