Fundamentals of Deepfakes: How are They Made and How to Spot Them?

Dr. Thomas P. Scanlon

Technical Program Manager – Data Science

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

We Have Lawyers

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.

DM21-0895

Can you spot the fake?

This person does note exist…

https://thispersondoesnotexist.com/

https://thisxdoesnotexist.com/

MDM

Mis-, **D**is-, **M**al-information

- Misinformation: false information that is shared without intent to harm
- Disinformation: false information deliberately created to mislead of cause harm
- Mal-information: information based on truths but purposefully used out of context to mislead or cause harm

MDM Examples

Mis-, **D**is-, **M**al-information

- Misinformation: Betsy Ross sewed the first American flag
- Disinformation: Operation INFEKTION
- Mal-information: 80% of dentists recommend Colgate

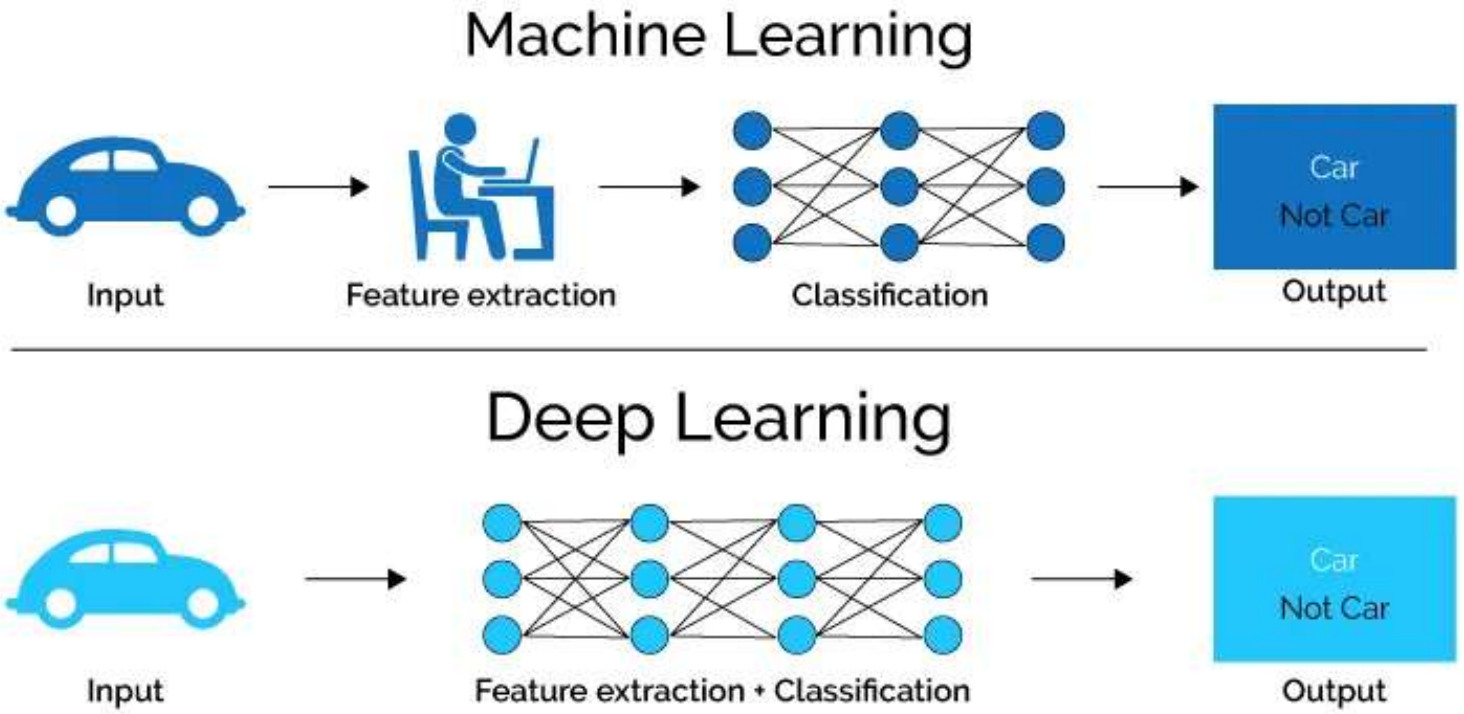
What is a Deepfake?

- Deepfake = ‘deep-learning’ + ‘fake’
- ‘deepfake’ originates from Reddit user in 2017 who claimed to have created the method
- Can be audio, video, image, multimodal
- Not the same as using Photoshop
- Deepfakes are considered Disinformation

  -or combined with Disinformation (i.e. profile with deepfake images)
- This talk is a primer on deepfakes

Deep Learning

Deep Learning is Machine Learning using a neural network



https://semiengineering.com/deep-learning-spreads/

Deepfake Creation Process

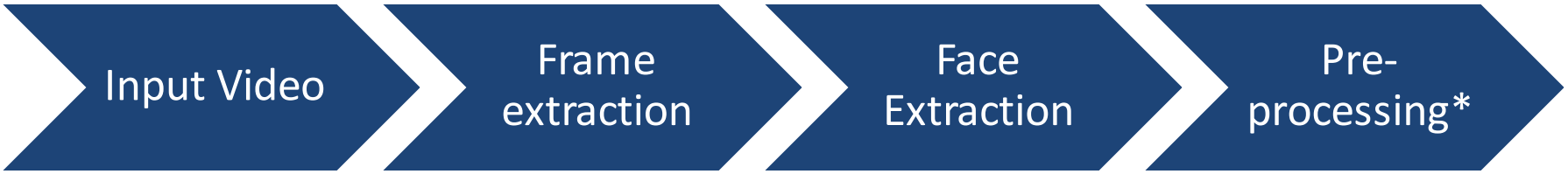
- Extraction
  - -Data collection (source data)
- Training
- Conversion / Generation

Deepfake Creation Process - Extraction

- As a practical matter, need to consider what data sources will provide this data
- A lot of training data is needed
- For images, thousands of images may be necessary
- Just a few video clips can replace thousands of images
- Extraction is the process of extracting individual frames from from video source, identifying faces and aligning them.
- Will need images of source (subject we want to embed) and destination (subject we want to override)

This is different from feature extraction

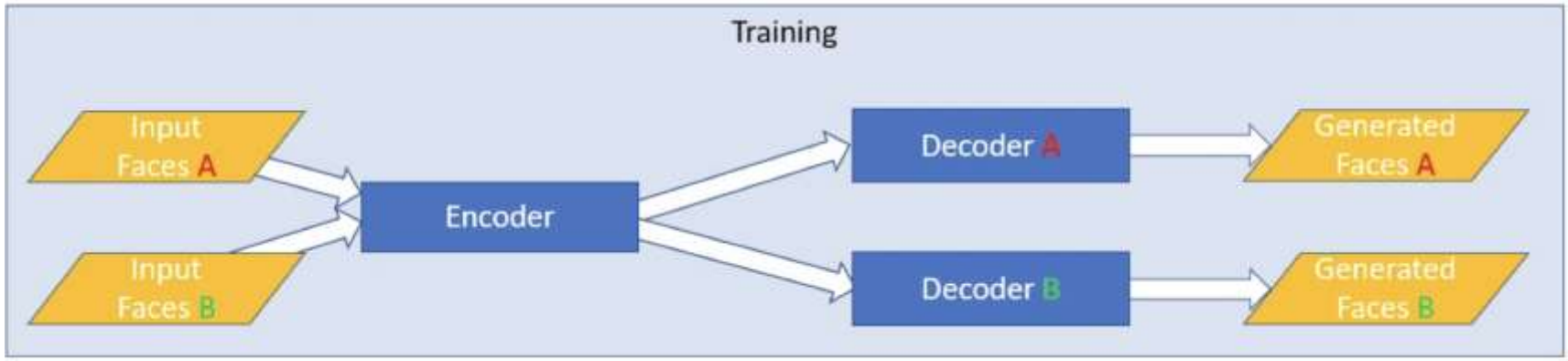
Deepfake Creation Process - Extraction



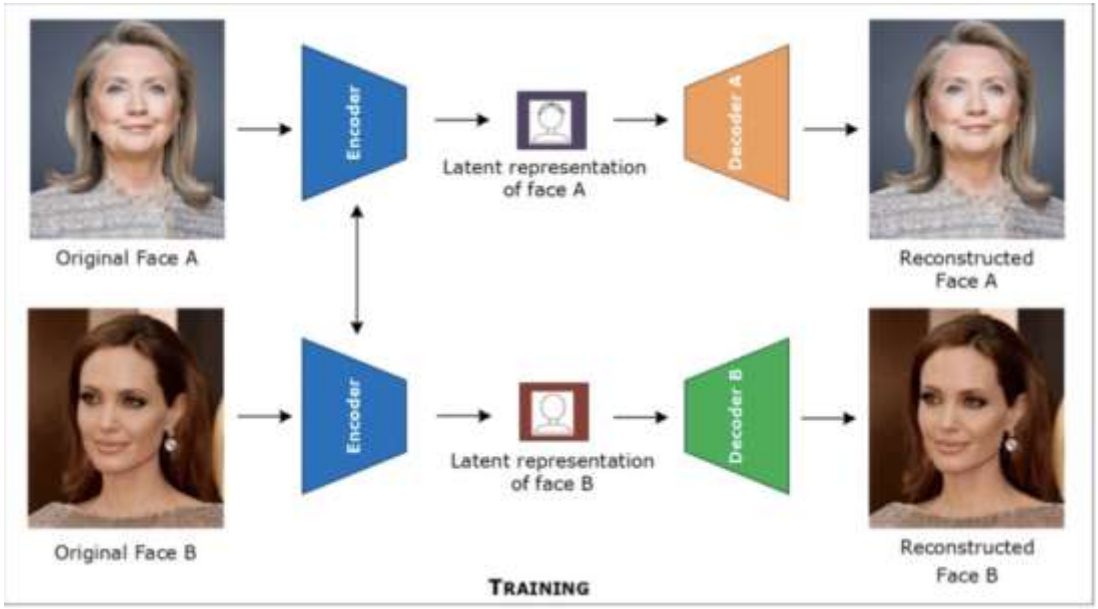
*resizing, normalization, augmentation, etc.

Deepfake Creation Process - Training

Autoencoders
Encoders and Decoders



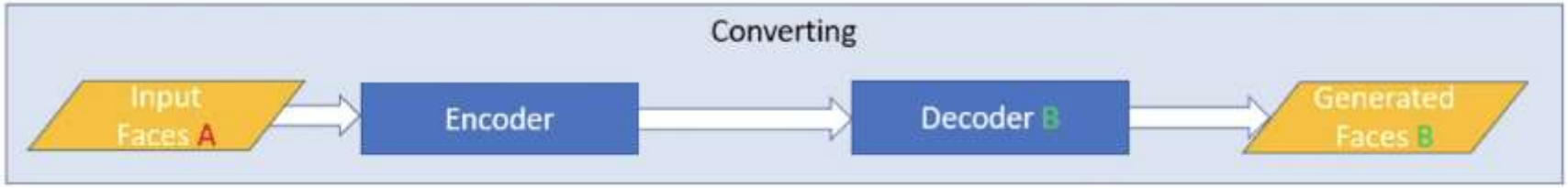
Deepfake Creation Process - Training



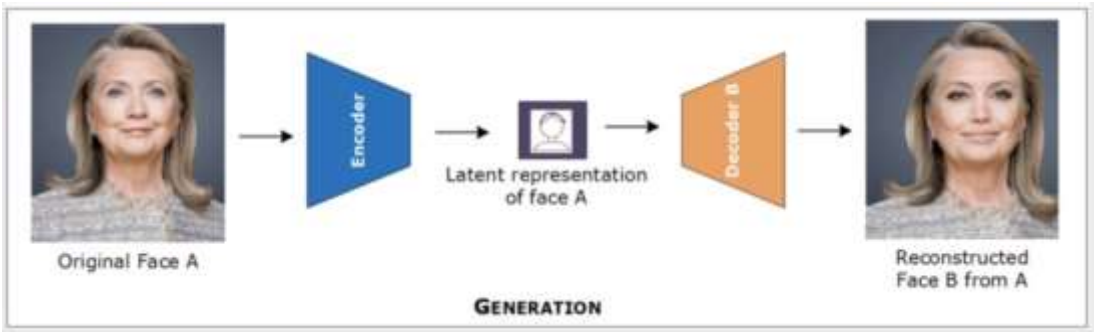
Masood, Momina & Nawaz, Marriam & Malik, Khalid & Javed, Ali & Irtaza, Aun. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward.

Deepfake Creation Process - Generating

Autoencoders
Encoders and Decoders

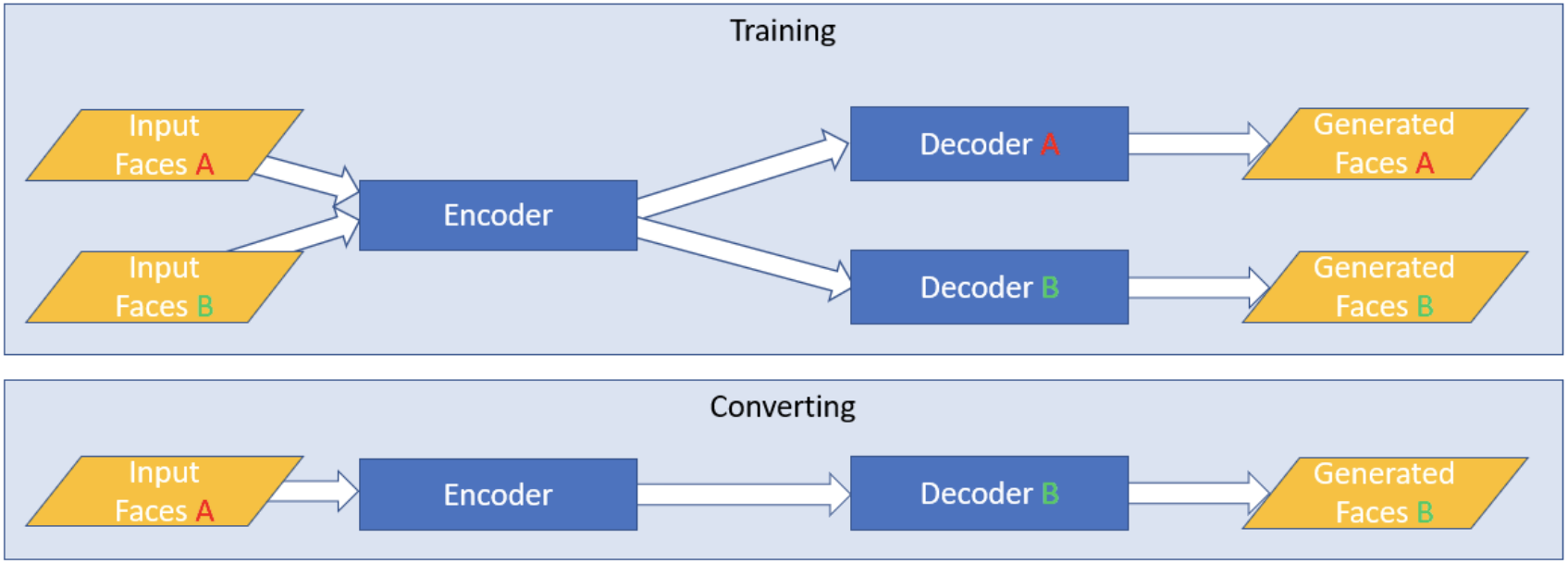


Deepfake Creation Process - Generating

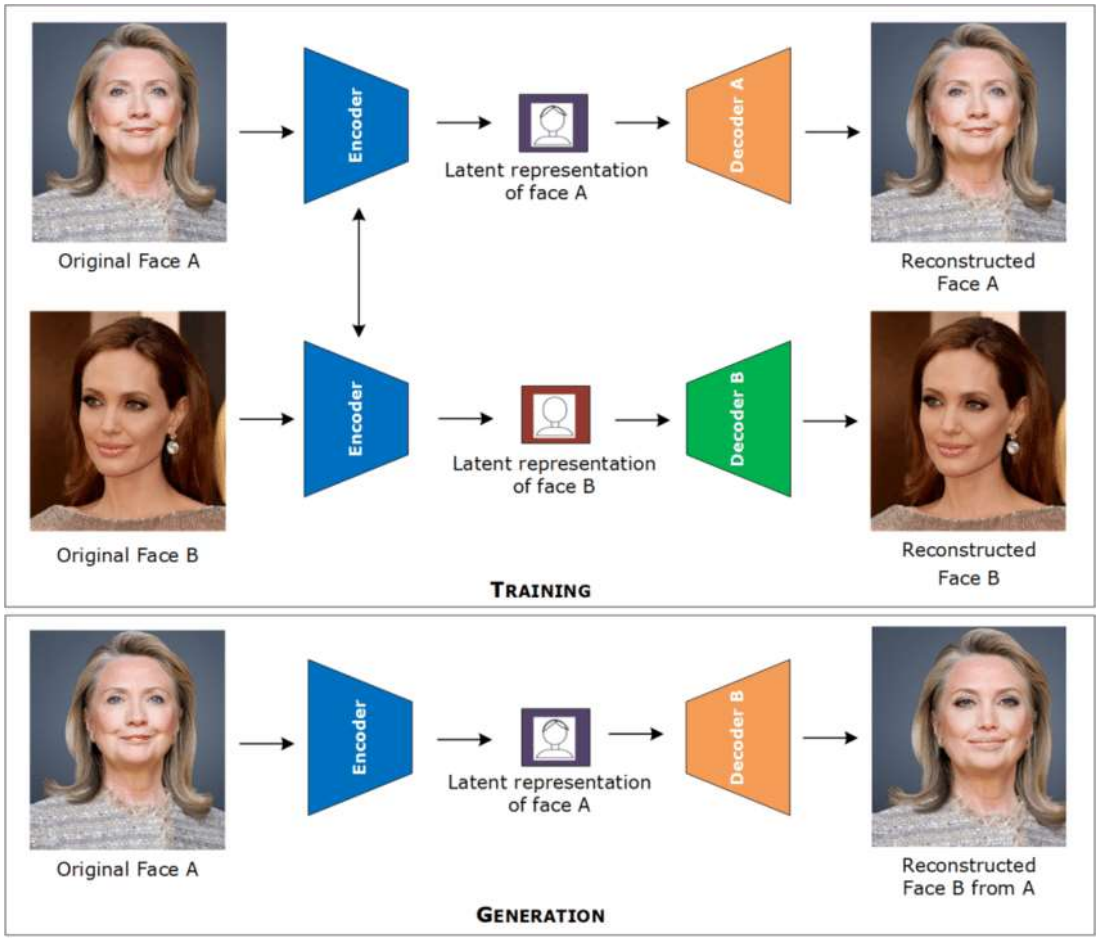


Masood, Momina & Nawaz, Marriam & Malik, Khalid & Javed, Ali & Irtaza, Aun. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward.

Deepfake Creation Process – Training & Generating



Deepfake Creation Process – Training & Generating



Masood, Momina & Nawaz, Marriam & Malik, Khalid & Javed, Ali & Irtaza, Aun. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward.

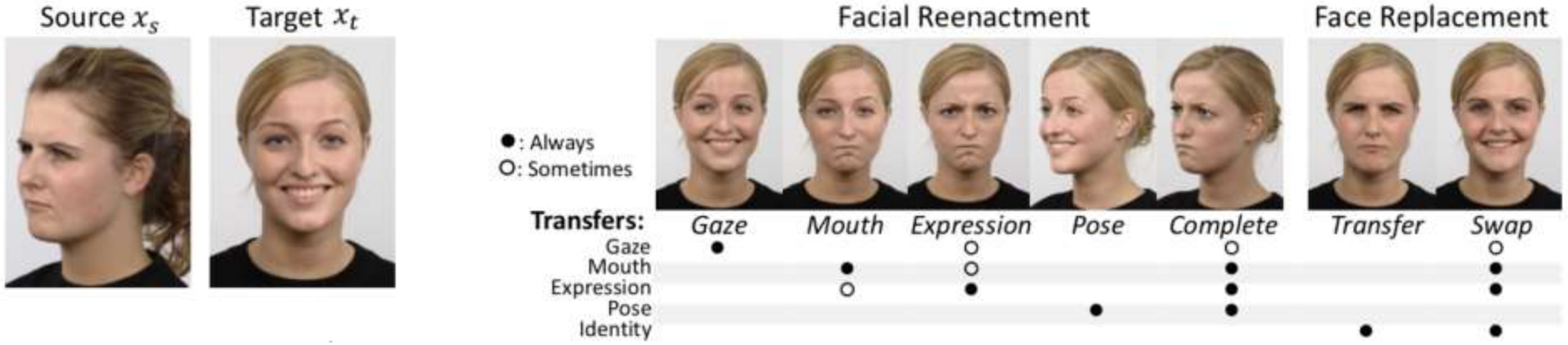
Main Deepfake Video Types

Reenactment

A reenactment deepfake is where x_s is used to drive the expression, mouth, gaze, pose, or body of x_t

Replacement

A replacement deepfake is where the content of x_t is replaced with that of x_s, preserving the identity of s.



Yisroel Mirsky and Wenke Lee. 2020. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (December 2020), 41 pages

Deepfake Creation Process - Enhancements

Beyond face replacement/reenactment, deepfake videos can also alter:

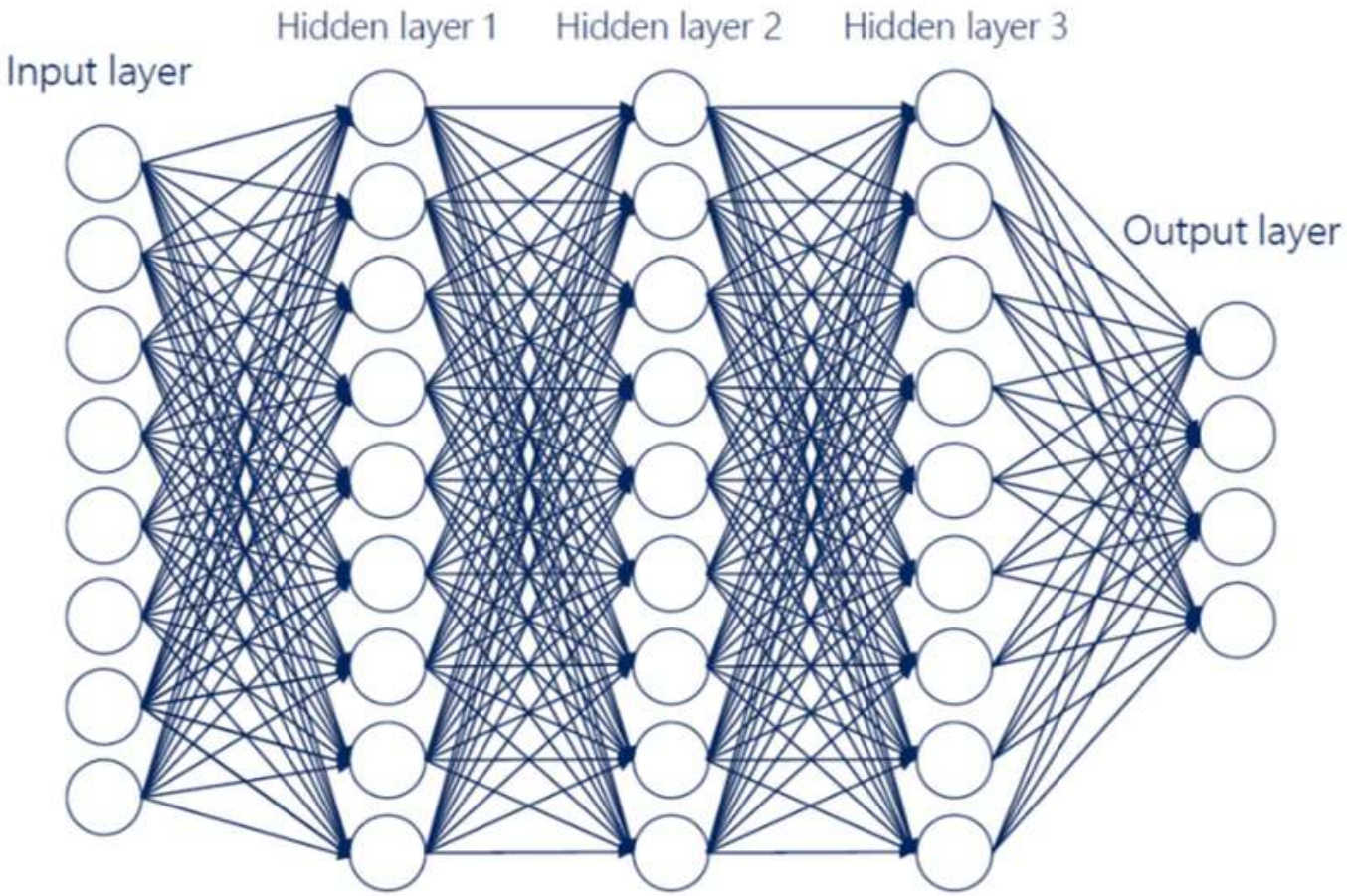
- Wardrobe
- Ethnicity
- Age
- Hair
- Makeup
- Articles

…can even create completely synthetic people

A little more on deep learning…

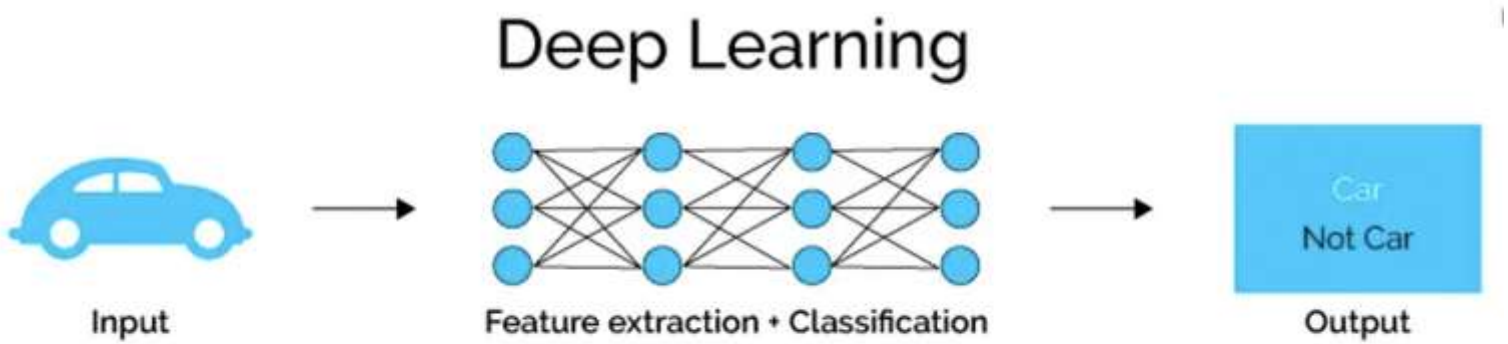
- Deep learning is accomplished through using a deep neural network (DNN) with an input layer, multiple hidden layers, and an output layer.
- A DNN has neurons, layers, weights, input, output, activation functions and a learning mechanism (optimizer).
- Different types of DNNs are commonly used: convolutional neural network (CNN), recurrent neural networks (RNN), multi-layer perception (MLP), and others.
- The encoders-decoders discussed so far are typically implemented using CNN.

Deep Neural Network



Merenda, Massimo & Porcaro, Carlo & Iero, Demetrio. (2020). Edge Machine Learning for AI-Enabled IoT Devices: A Review. Sensors. 20. 2533. 10.3390/s20092533.

Deep Learning – For Classification

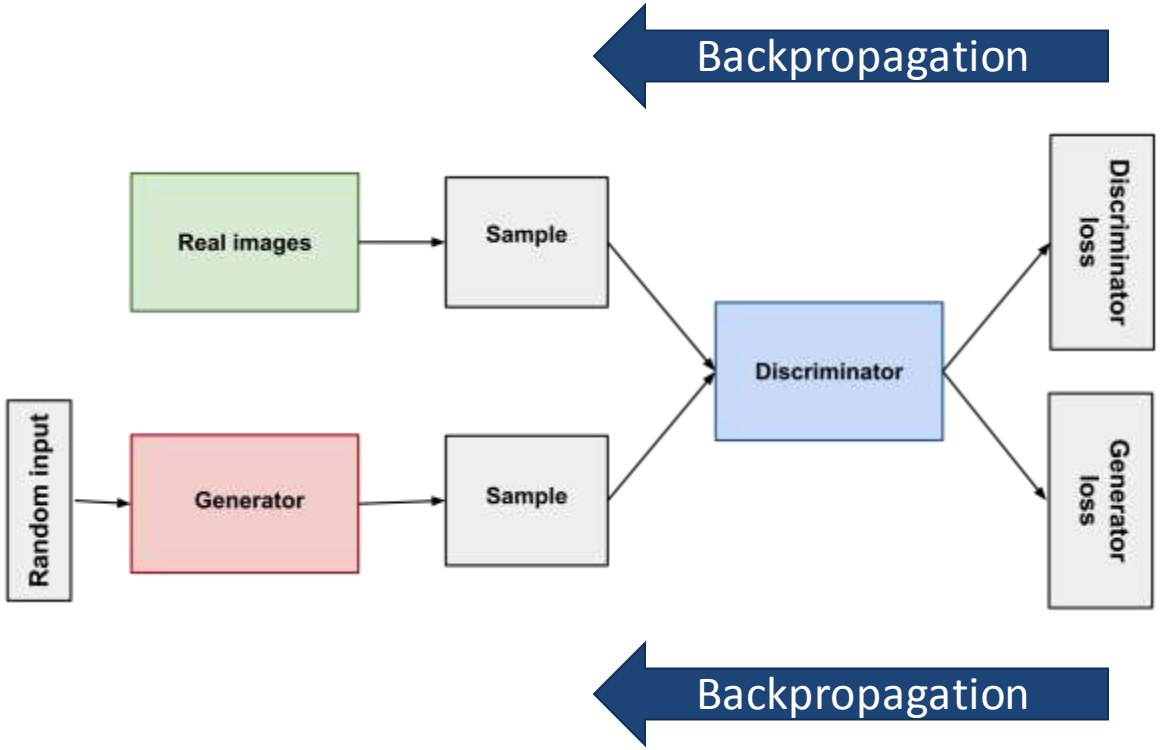


https://semiengineering.com/deep-learning-spreads/

Deep Learning – For Deepfake Creation

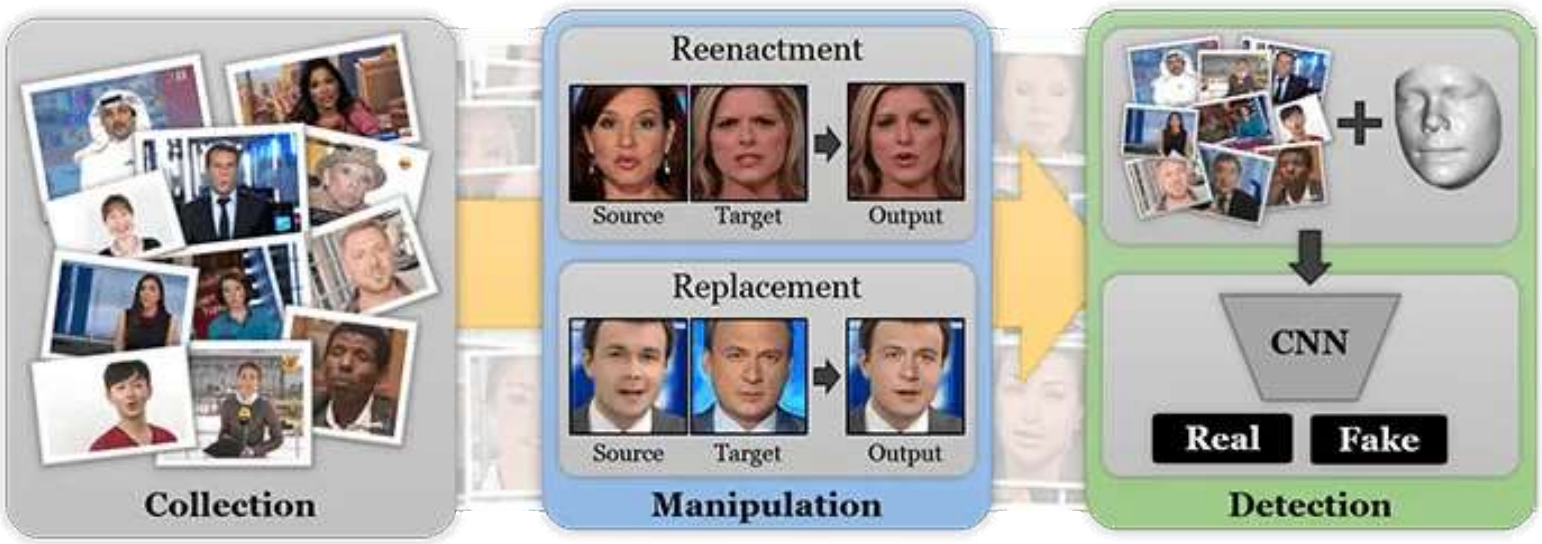
- Generative adversarial network (GAN) was introduced by Ian Goodfellow in 2014.
- GAN is a machine learning (ML) model in which two neural networks compete with each other to improve their predictions.
- There is a generator that tries to create fake images ("forger") and a discriminator ("detective") that tries to determine fake images from real images.
- The generator and discriminator are each using a deep neural network (DNN) and can be of same type or different.
- For deepfake creation, they are both often CNN.

Generative adversarial network (GAN)



https://developers.google.com/machine-learning/gan/gan_structure

Deepfake Creation with GAN



https://deepware.ai

Deepfake Concerns

- You don't need to be a data scientist of AI research to create deepfakes – no code/low code options exist
- Open source Python software such as Faceswap and DeepFaceLab are easy to use and the deep learning can be treated as a "black box".
- Motivated parties with more resources can do produce fairly strong deepfakes.
- Deepfakes are used for scams & hoaxes, social engineering, fraud, identity theft, political/election manipulation, forgery, fake pornography, and more.
- Examples: fake rental ads, fake dating profiles, fake LinkedIn accounts, fake voicemail messages, etc.

Deepfake Concerns

- You don't need to be a data scientist of AI research to create deepfakes – no code/low code options exist
- Open source Python software such as Faceswap and DeepFaceLab are easy to use and the deep learning can be treated as a "black box".
- Motivated parties with more resources can do produce fairly strong deepfakes.

Deepfake Nefarious Uses

- scams & hoaxes
- social engineering
- fraud
- identity theft
- political/election manipulation
- forgery
- fake almost anything: pornography, rental ads, dating profiles, LinkedIn accounts, voicemail messages, etc.

This was entertaining….

…but….

Deepfakes for Malicious Use

- Malicious actors convinced a CEO to wire $243,000 to a scammer's bank account by using deep fake audio[1]
- Symantec reports they have observed at least 3 other deep fake audio cases involving CEOs and CFOs[2]
- Palestinian activists smeared by unknown, deepfaked identity[3]
- Politicians from the UK, Latvia, Estonia and Lithuania tricked by fake meetings with opposition figures [4]

1 - https://www.zdnet.com/article/forget-email-scammers-use-ceo-voice-deepfakes-to-con-workers-into-wiring-cash/
2 - https://www.bbc.com/news/technology-48908736
3 - https://www.reuters.com/article/us-cyber-deepfake-activist/deepfake-used-to-attack-activist-couple-shows-newdisinformation-frontier-idUSKCN24G15E
4 - https://www.theguardian.com/world/2021/apr/22/european-mps-targeted-by-deepfake-video-calls-imitating-russian-opposition

Deepfakes for Malicious Use cont.

- Deepfakes replace women on sextortion calls [1]
- Deepfake video of bank president offer false discount [2]
- Deepfakes used to Impersonate a Navy Admiral and Bilk Widow Out of Nearly $300,000 [3]
- AI app used to "undress" women [4]

1 - https://timesofindia.indiatimes.com/city/ahmedabad/deepfakes-replace-women-on-sextortion-calls/articleshow/86020397.cms
2 - https://tekdeeps.com/fraudsters-created-a-deepfake-of-oleg-tinkov-dont-be-fooled-by-this-ad/
3 - https://www.thedailybeast.com/romance-scammer-used-deepfakes-to-impersonate-a-navy-admiral-and-bilk-widow-out-of-nearly-dollar300000
4 - https://www.technologyreview.com/2019/06/28/134352/an-ai-app-that-undressed-women-shows-how-deepfakes-harm-the-most-vulnerable/?truid=21defdb9a2d89523a2a6ea4c092cecca&utm_source=the_algorithm&utm_medium=email&utm_campaign=the_algorithm.unpaid.engagement&utm_content=10-08-2021

Fake deepfakes?

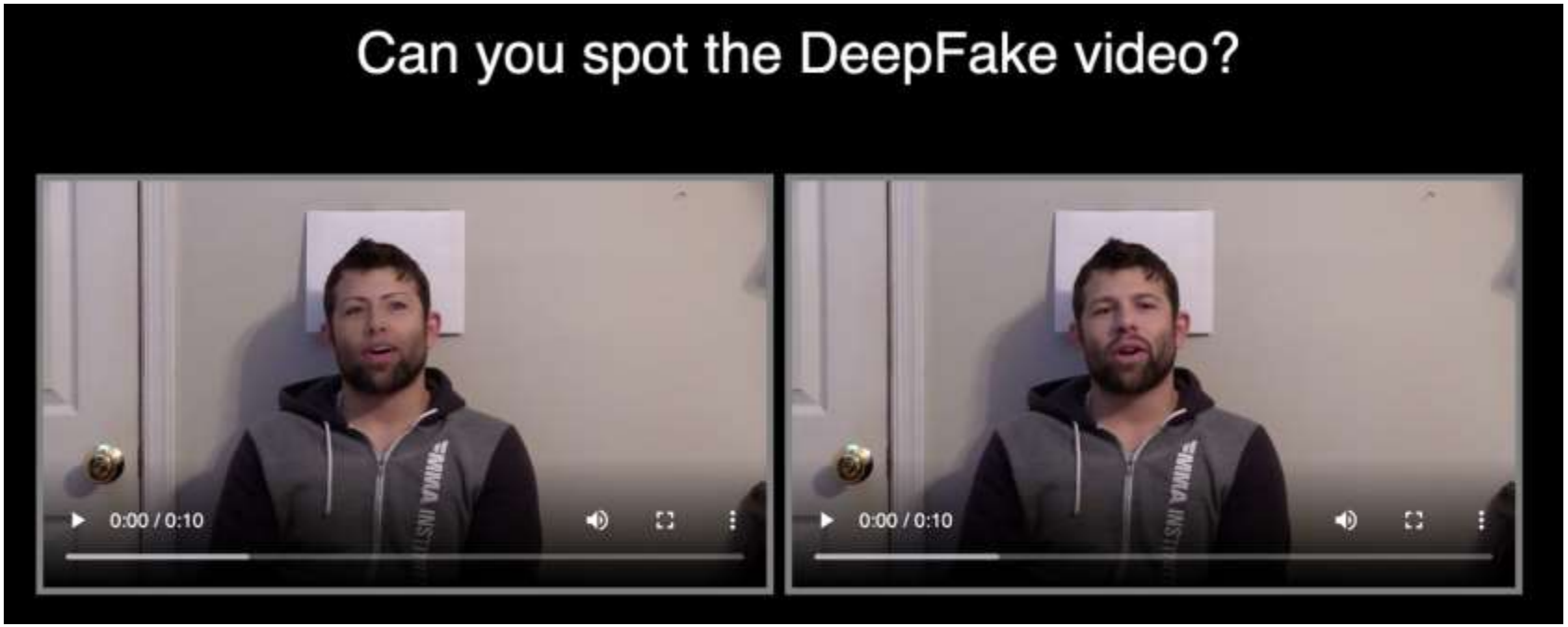
- Mother used deepfake to frame cheerleading rivals [1]
- How misinformation helped spark an attempted coup in Gabon [2]

1 - https://www.bbc.com/news/technology-56404038
2 - https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/

Detecting Deepfakes – Practical Cues

- Flickering
- Unnatural movements & expressions
- Lack of Blinking
- Unnatural hair and skin colors
- Awkward head positions
- Appears to be lip-syncing
- Oversmoothed faces
- Double eyebrows / raised eyebrows at wrong time / one raised eyebrow
- Glare / lack of glare on glasses
- Realistic-ness of moles / consistent placement of moles
- Earrings – wearing only one or mismatched

Detecting Deepfakes – Eye Test



https://www.media.mit.edu/projects/detect-fakes/overview/

Detecting Deepfakes – Programmatically

1. Blending (spatial)

2. Environmental (spatial)
- Lighting, back/foreground diffs

3. Physiological (temporal)
- Generated content lack pulse, breathing, have irregular eye blinking patterns

4. Synchronization (temporal)
- Mouth shapes and speech
- BPM mouth closed failure

5. Coherence (temporal)
- Flickering
- Predict next frame

6. Forensic (spatial)
- GANs leave unique fingerprints
- Camera PRNU

7. Behavioral (temporal)
- Video vs audio emotions
- Target mannerisms (> data)

https://dl.acm.org/doi/fullHtml/10.1145/3425780

Deepfake Detection Challenge (DFDC)

- AWS, Facebook, Microsoft, the Partnership on AI's Media Integrity Steering Committee, and other academics created the Deepfake Detection Challenge :

  https://www.kaggle.com/c/deepfake-detection-challenge

- 100,000 deepfake clips (created by Facebook using paid actors) for entrants to test their detectors.
- 2,000 participants from industry and academia, generated more than 35,000 deepfake detection models.
- The best model detected deepfakes from Facebook's collection about 82%of the time; when the same algorithm was run against previously unseen deepfakes, it detected about 65%/

Detecting Deepfakes – Tools

- Microsoft's Video Authenticator Tool

  -detects blending boundaries and grayscale elements that are undetectable to the human eye

- Facebook Reverse Engineering

  -detects digital fingerprints left behind by generative model

- Quantum Integrity

  -determines if images of videos have been manipulated, methods not well documented

DARPA Projects

- Semantic Forensics (SemaFor)

  -semantic detection algorithms, which will determine if multi-modal media assets have been generated or manipulated

  -attribution algorithms will infer if multi-modal media originates from a particular organization or individual

  -characterization algorithms will reason about whether multi-modal media was generated or manipulated for malicious purposes

- Media Forensics (MediFor)

  -developing technologies for the automated assessment of the integrity of an image or video and integrating these in an end-to-end media forensics platform

Deepfakes Summary

- Existing tools for creating deepfakes produce fairly realistic products that can pass a cursory look from the human eye
- Image and video fakes can still often be recognized by human eye upon closer inspection
- Audio fakes are more convincing and tougher to detect by ear
- The state-of-the art for deepfakes is rapidly advancing and products are becoming better all the time
- Tools for detecting deepfakes exist but don't quite match the capability of creators – much work to do in detection
- Verifying something is NOT a fake is also an area for further research

Deepfakes Takeaway

- Good news: Even using tools that are already built (Faceswap, DeepFaceLab, etc) it still takes considerable time and GPU resources to create even lower quality deepfakes
- Bad news: Well-funded actors can commit the resources to making higher quality deepfakes, particularly for high-value targets

- Good news: Advancements are being made in detecting deepfakes
- Bad news: Technology for deepfake creation continues to advance; it will likely be a never-ending battle similar to malware and anti-virus software

Questions?

Contact Information

Dr. Thomas P. Scanlon
Technical Program Manager
CERT Data Science
Software Engineering Institute
Carnegie Mellon University
Email: scanlon@cert.org