Carnegie Mellon University
Software Engineering Institute

The Near Future of AI: Transformer-Based Learning

APRIL 1, 2023

Shannon Gallagher, PhD
Machine Learning Research Scientist
AI Division

Legal

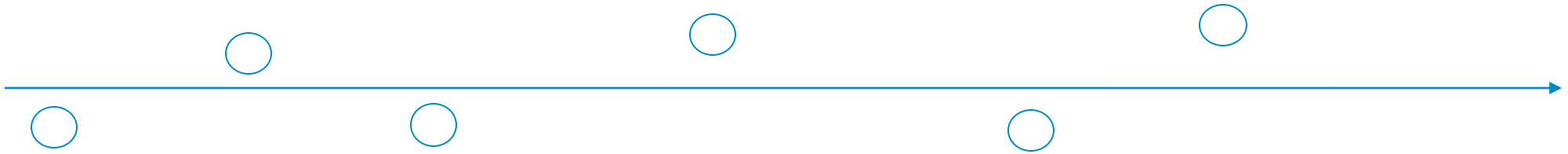
The term “Artificial Intelligence” was coined in the 1950s

Artificial Intelligence – process of machines doing human tasks

Examples:

- Chess playing
- Facial Recognition
- Speech to text
- Deepfake Detection

Brief history of AI (starting in the 1950s)



1956 – Logic Theorist (Newell, Simon, and Shaw)
1969 – First ATM
1988 – First “AI” chatbot (Jabberwacky)
1997 – Deep Blue beats world chess champion
2011 – AlexNet
2017 - Transformers

AI is moving at dizzying pace

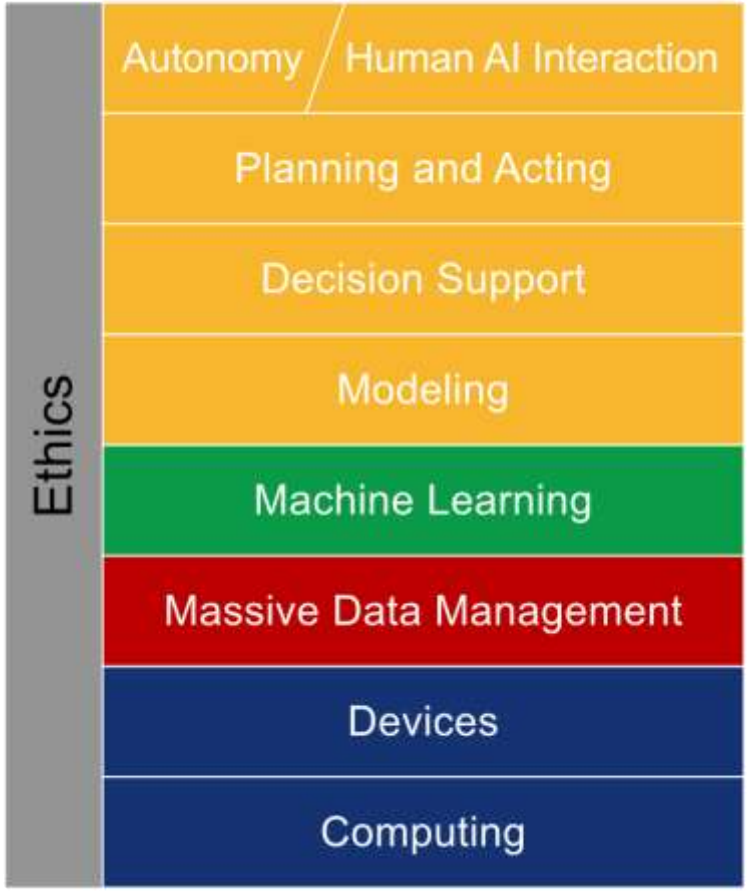
2018 – BERT, STYLEGAN, GPT
2019 – GPT-2
2020 – NeRF, STYLEGAN-2, GPT-3
2021 – CLIP, STYLEGAN-3, DeepFaceLive, DALL-E, GitHub Copilot
2022 – Stable Diffusion, Facial Recognition at Airports, ChatGPT, DALL-E 2
2023 – GPT-4, Gen-2, VertexAI

How???

Substantial improvements in

- Ops
- ML
- Data
- Compute

And Ethics



AI Stack.

Ethics matter

Carnegie Mellon University Responsible AI

Trust and AI Systems with Carol Smith

DoD Ethical Principles

1. Responsible
2. Equitable
3. Traceable
4. Reliable
5. Governable

Compute has improved

- GPUs
  - Specialized hardware and software
- Cloud
  - Parallelization, storage
- Edge
  - Sensor data

Data are easier to handle and access

- Big
- Labeled
- Better workflows

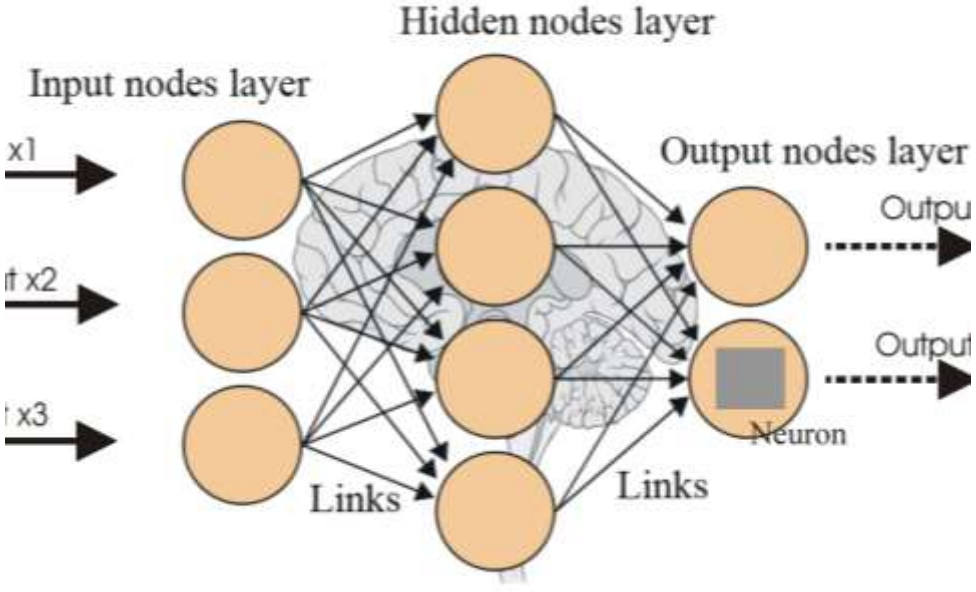
Machine Learning had breakthroughs

- Architectures
  - Attention
  - GANs
  - Encoders/Decoders
  - **Transformers**

Ops and Human-Computer Interaction pull it together

- Feedback loops
  - E.g. ChatGPT
- Putting people, ML, data, and compute together

Transformers, but first some context



- Transformers are a type of **neural network (NN)**
- **NN** are compositions of typically simple functions
  - Easily computable
  - Easy to take derivative
- E.g. logistic regression is a 1-layer NN
- By iterating hidden layers you can increase the complexity of NN
  - i.e. **deep** NN
- Different compositions create **architectures**

Transformer Architecture ->

- Vaswani et al. 2017
  - "Attention Is All You Need"
- Closely related to encoder/decoder architecture
- Much motivation from *language translation, e.g.*
  - Source = "I like math"
  - Target = "Me gustan las matemáticas"

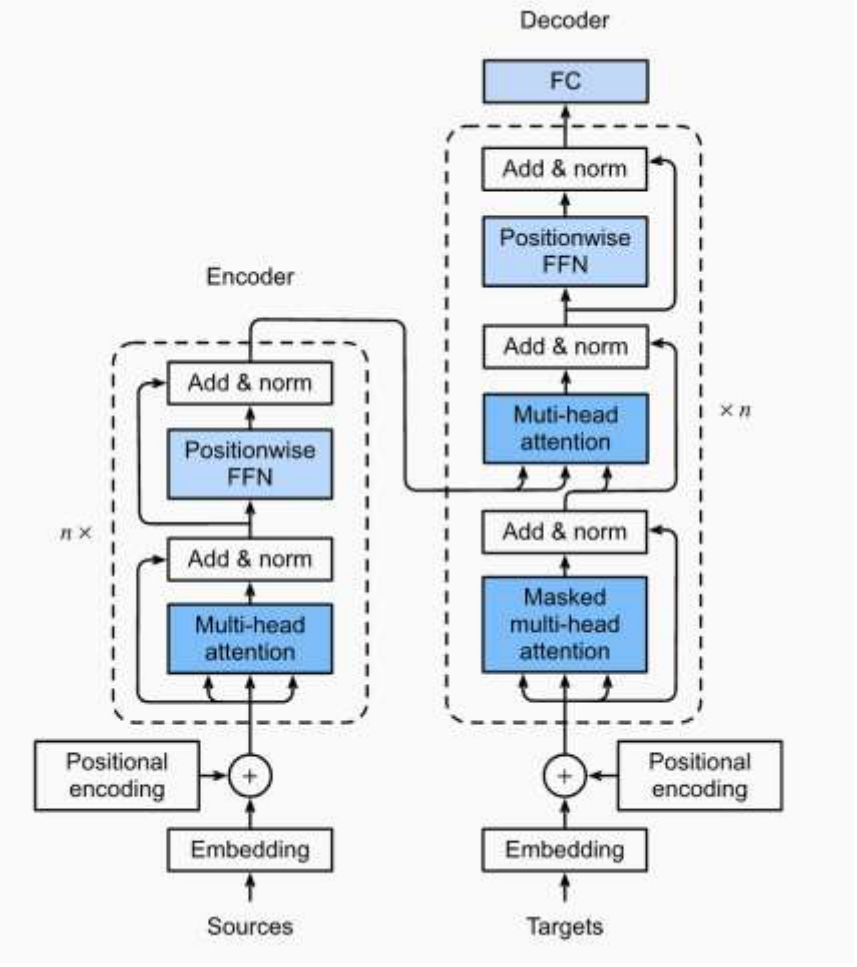


Image from d2l.ai

Attention relates tokens (e.g. words) to one another

$$Att(\boldsymbol{q}, D) = \sum_{i=1\ldots m} \alpha(\boldsymbol{q}, \boldsymbol{k}_i)\boldsymbol{v}_i$$

q = queries (vector)

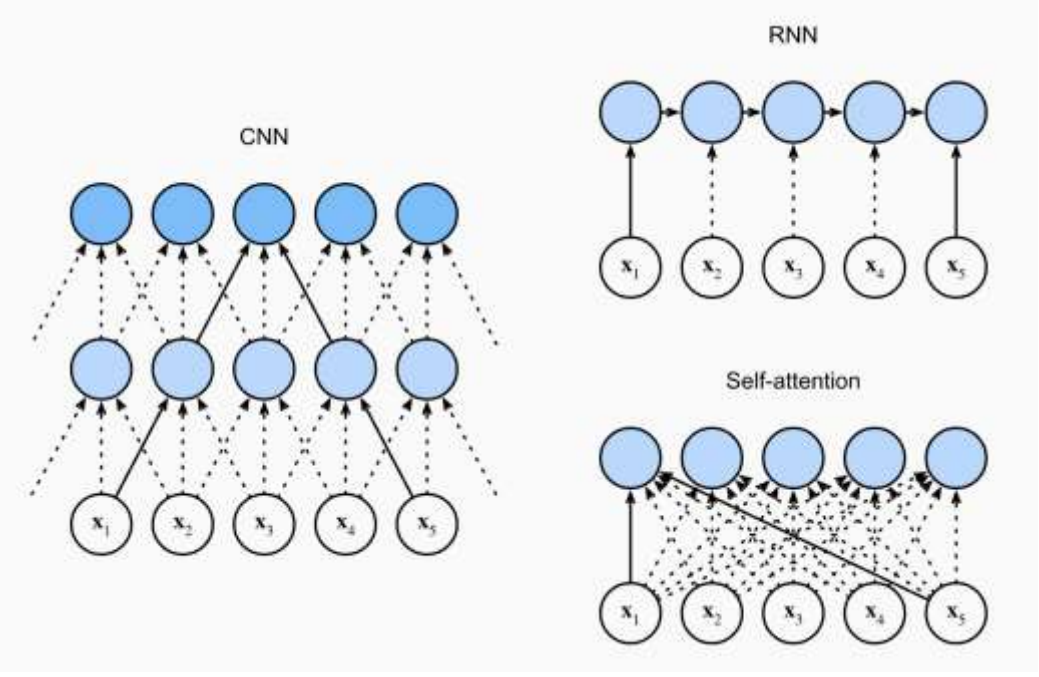
k = keys (vector)

v = values (vector)

D = $\{(v_i, k_i)\}_{i=1\ldots m}$ (set of key, value pairs)

$\alpha(\cdot)$= attention weights (scalar)

Transformers improve previous architectures via attention



- CNN = convolutional neural network
  - Think images
- RNN = recurrent neural network
  - Think time series

- Self attention has more context than CNNs
- More parallelizable sequences than RNN

Takeaway: Self-attention allows for 'just right' spot of model complexity vs. computation

Image from d2l.ai

These ones have transformer-based architecture

2018 – **BERT**, STYLEGAN, **GPT**
2019 – **GPT-2**
2020 – NeRF, STYLEGAN-2, **GPT-3**
2021 – **CLIP**, STYLEGAN-3, DeepFaceLive, **DALL-E**, **GitHub Copilot**
2022 – Stable Diffusion, Facial Recognition at Airports, **ChatGPT**, **DALL-E 2**, **ImageGen**
2023 – **GPT-4****, **Gen-2**, **VertexAI**

**speculated

State of the Art LLM are expensive

- Time
  - Human hours
  - Machine training and operating hours
- Energy
- Hardware

AI is vulnerable

OPWNAI : CYBERCRIMINALS STARTING TO USE CHATGPT

Check Point Research 1/6/2023

Chatting Our Way Into Creating a Polymorphic Malware

Eran Shimony & Omer Tsarfati 1/17/2023

Argo.ai Shuts Down

Forbes 10/22/2023

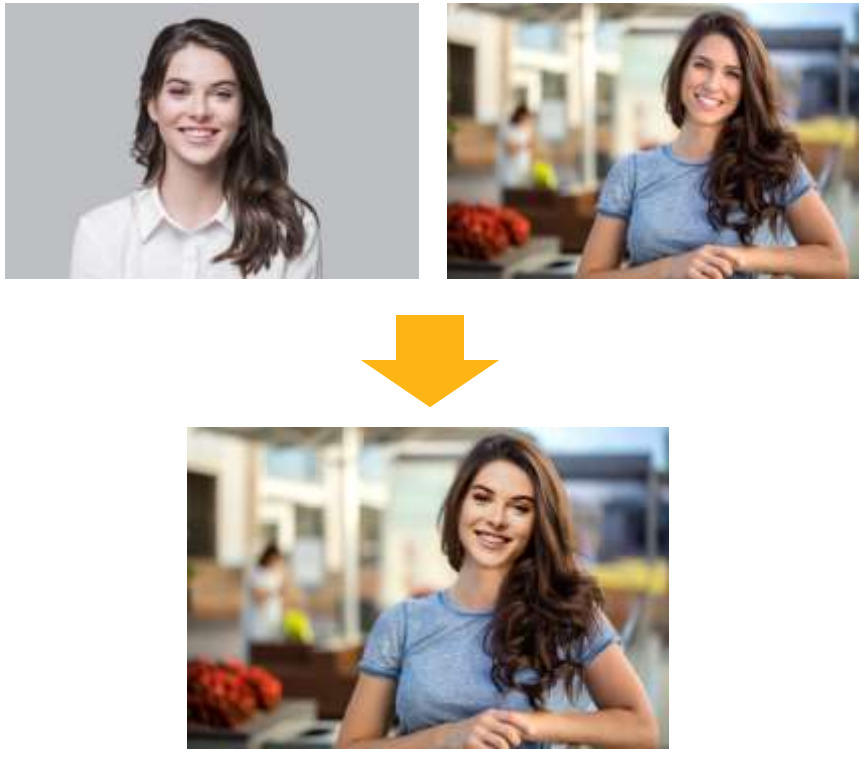
European politicians duped into deepfake video calls with mayor of Kyiv

The Guardian 6/25/2022

A Tesla driver is charged in a crash involving Autopilot that killed 2 people

NPR 1/18/2022

Deepfakes Are Dangerous



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

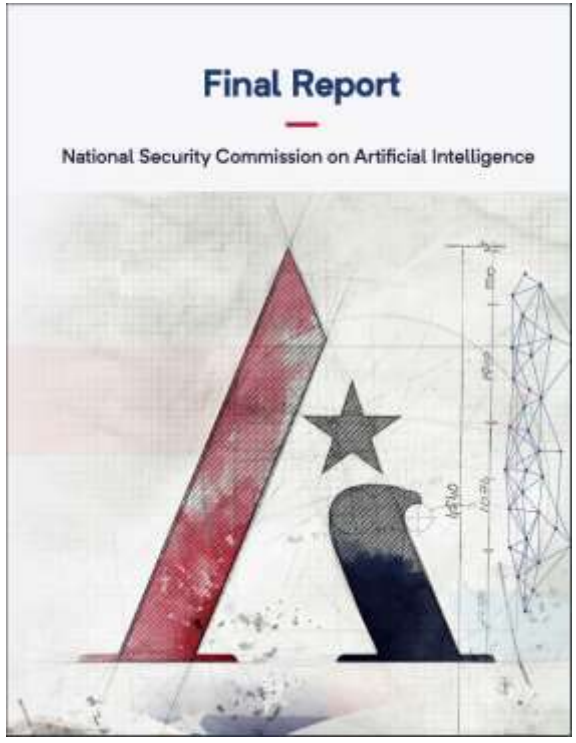
Potential Dangers

- Impersonation of political figures
  Defamation of citizens
- Mis-, dis-, and mal- information

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

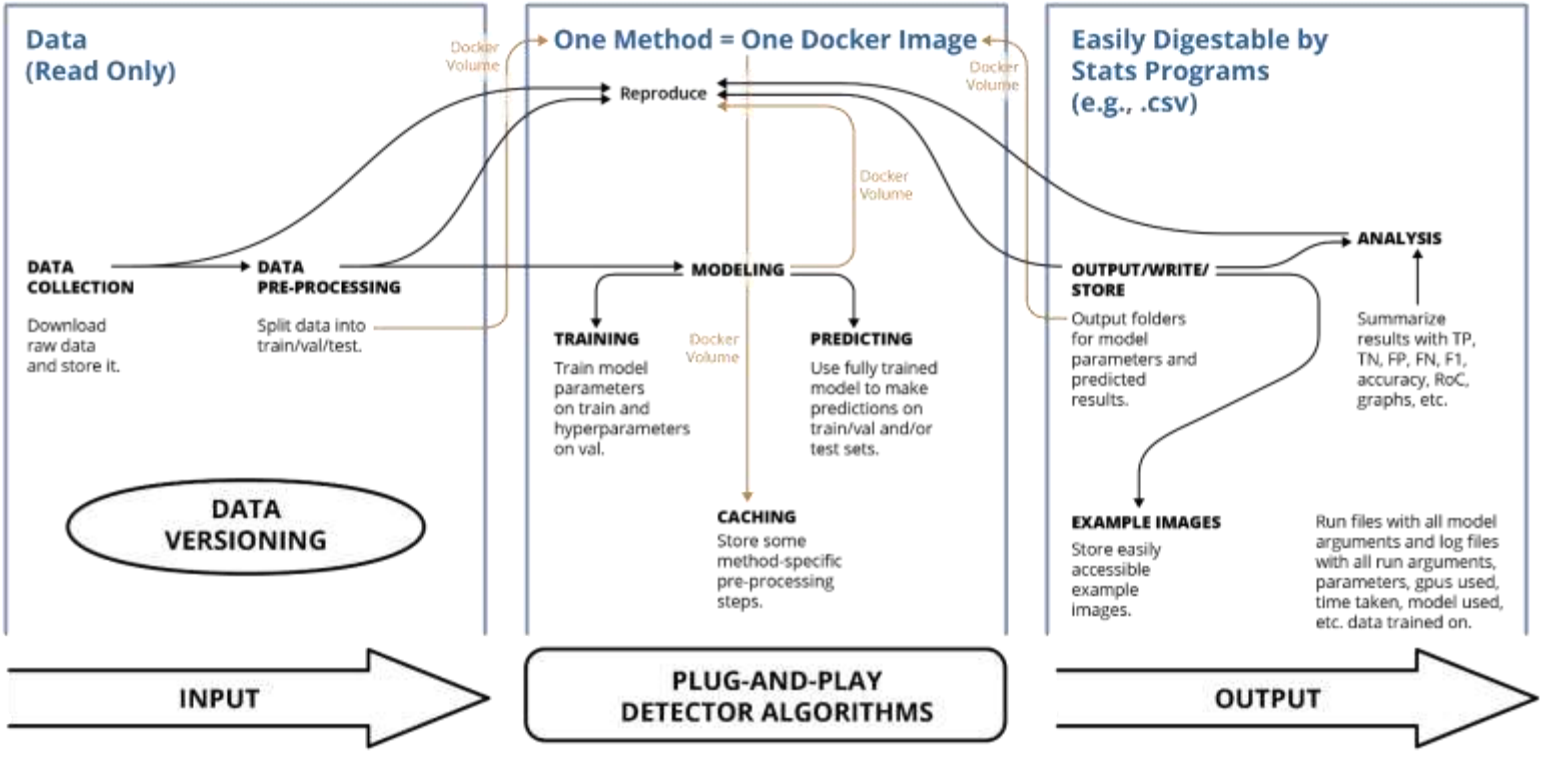
We need fast and reliable detectors.

There is a path forward: AI Assurance



Create a National AI Assurance Framework. All government agencies will need to develop and apply an adversarial ML threat framework to address **how key AI systems could be attacked** and **should be defended**. An analytical framework can help to categorize threats to government AI systems and help analysts detect, respond to, and remediate threats and vulnerabilities.

Our Deepfake Detection Pipeline (DDP) Creates Benchmarks

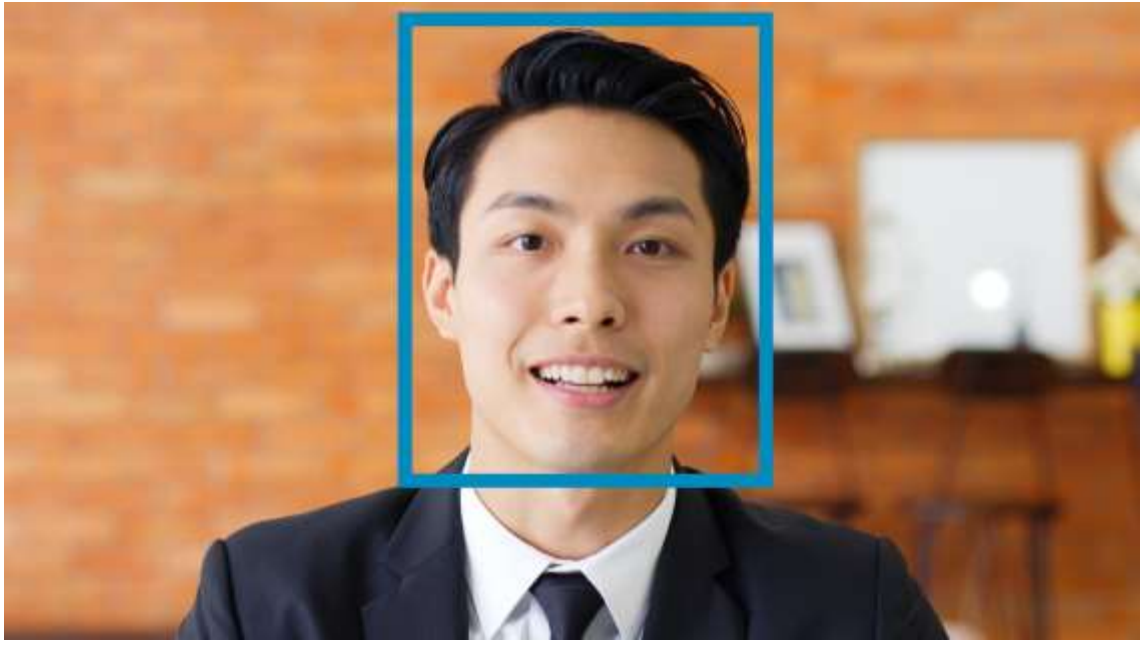


DDP is reproducible, portable, and modular.

DDP's backend is SEI's Juneberry.

We've Noticed a General Trend in Detection Methods

1. Find the face.

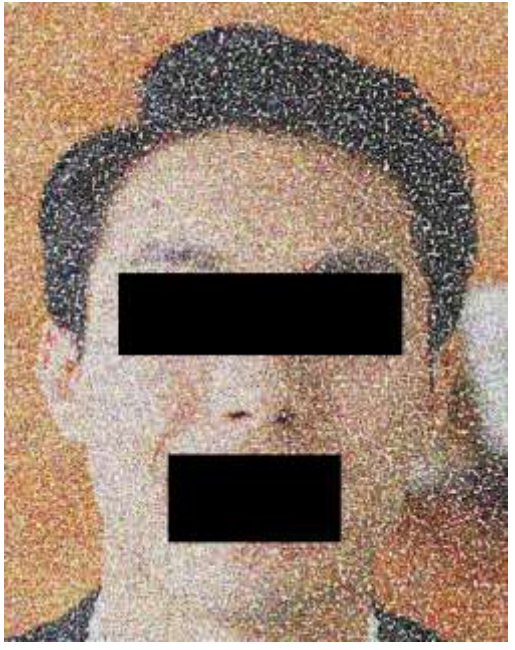


We've Noticed a General Trend in Detection Methods

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

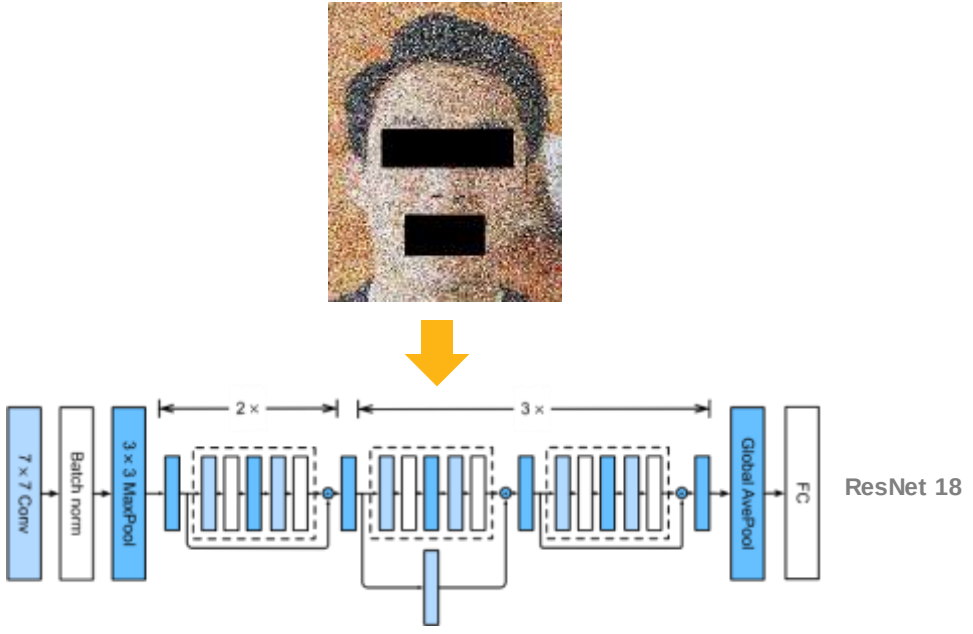
We've Noticed a General Trend in Detection Methods

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*.

AML Lab Members and Campus Collaborators

Dr. Nathan VanHoudnos
Senior ML Research Scientist
Lab Lead

John Stogoski
Senior Systems Engineer
Acting Lab Project Lead

Matthew Churilla
Senior ML Engineer

Andrew Mellinger
Principal Engineer

Dr. Jasmine Ratchford
Senior ML Research Scientist

John Zucca
ML Engineer

Bill Shaw
Senior Engineer

Nick Winski
Software Developer

Dr. Shannon Gallagher
ML Research Scientist

Anusha Sinha
ML Research Scientist

Hayden Moore
Assistant Developer

Jordan Widjaja
Assistant Developer

Dr. Lujo Bauer
Professor

Dr. Matt Fredrickson
Associate Professor

Dr. Bryan Parno
Associate Professor

Takeaways

- AI is improving at a rapid pace
- Transformers leverage data and compute to be very effective
- AI can be vulnerable due to attacks
- We need AI assurance