On TrinityAI: Trustworthy, Resilient and Interpretable AI

Susmit Jha (SRI)

Principal Scientist
Neuro-Symbolic Computing and Intelligence (NuSCI) Group
Computer Science Lab, SRI

Three Coupled Challenges in AI

- Trust: Given a machine learning model trained on data from some distribution, how do we determine that the model can be trusted on a new input which may be out of the training distribution (OOD)? How do we supplement model's prediction with a quantitative confidence?

Lane detection trained for precipitation below 25 fails on high precipitation levels (OODs)

Precipitation 17

Precipitation 21

Precipitation 55

OOD as novel classes

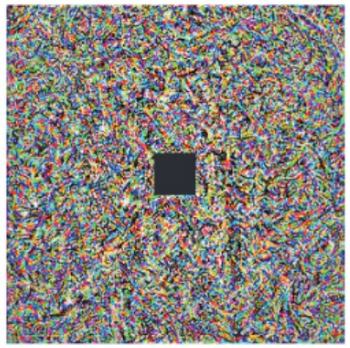
OOD as novel context

Three Coupled Challenges in AI

- Resilience: Given a machine learning model, how do we ensure that the model is robust to adversarial attacks – inference-time attacks such as adversarial perturbations, training-time attacks such as insertion of Trojan triggers, privacy-attacks that can attempt to infer training-data on which the model was trained ?

Imperceptible perturbations

Localized (single pixel) attacks



Adversarial Reprogramming

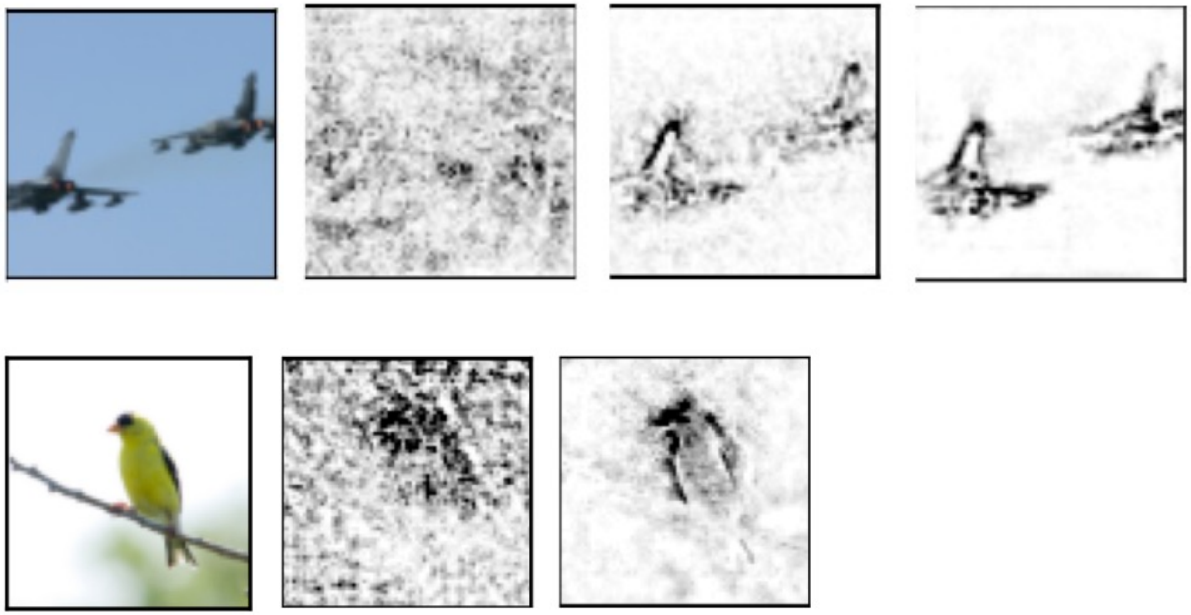
Physically Realizable Patch Attacks

Clean Data Poisoned Data (polygon or filter trigger)

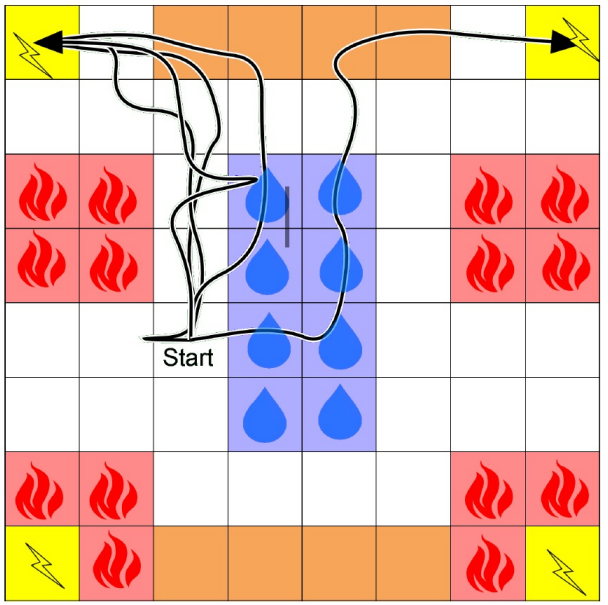
Trojan/Backdoor Attacks

Three Coupled Challenges in AI

- Interpretability: Given a machine learning model and its decision on a single input or a class of inputs, how do we explain the decision ? How do we assign attribution or importance of a decision over different features of an input?



Saliency Maps

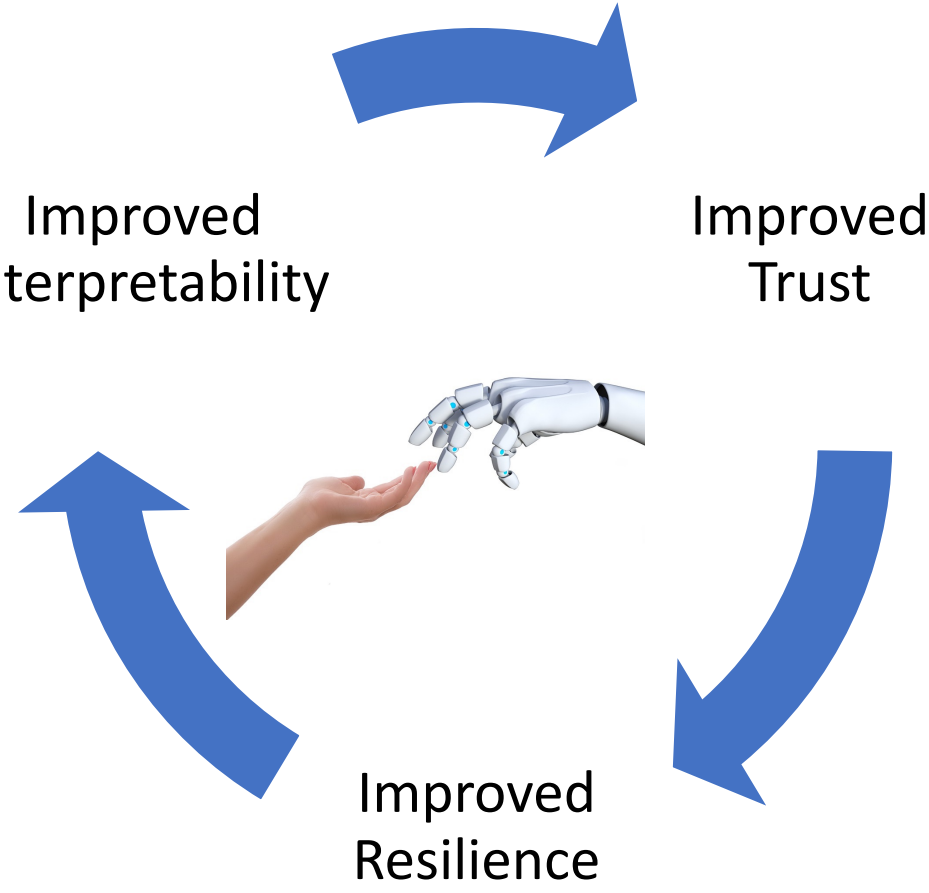


$$(H\neg red \wedge O\ yellow) \wedge H((yellow \wedge O\ blue) \Rightarrow (\neg blue\ S\ brown))$$

Extracted Logical Specification

Three Coupled Challenges in AI

The dependency between Trust, Resilience and Interpretability also creates a virtuous cycle.

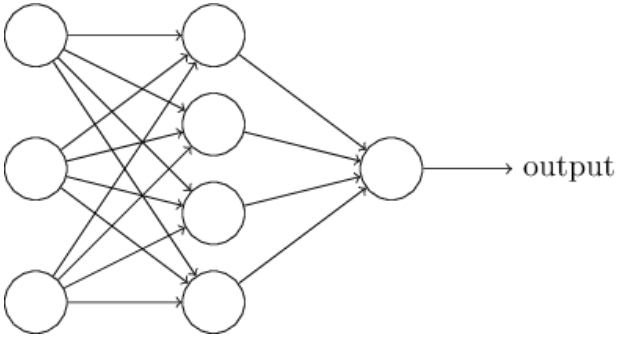


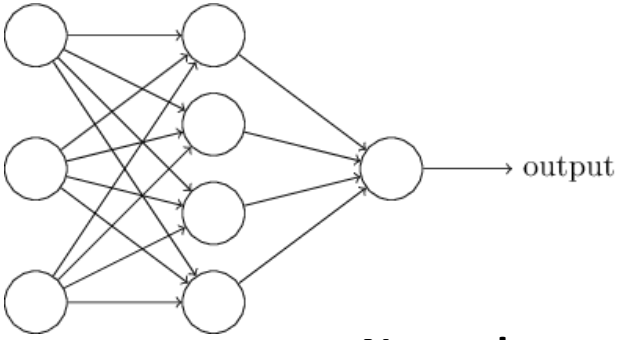
Simultaneously improvement in trustworthiness, resilience and interpretability is critical for their use in high-assurance systems and in human-machine teams.

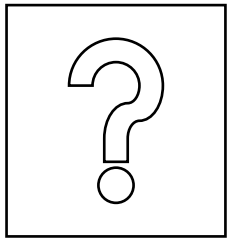
Rest of the talk

- Basic Principle behind TrinityAI – Predictive Coding and the use of Context
- Context-driven Trustworthy and Robust Learning
- Improving Resilience Using Attributions/Explanations
- Improving Attributions by Making Learning Models Robust
- TrinityAI Tool and Ongoing Work

Trust in Deep Learning Models: Behavior on OODs







Not only wrong predictions but predictions with high confidence (soft-max values)

"The whole problem with the world is that fools and fanatics are always so certain of themselves, and wiser people so full of doubts." – Bertrand Russell

Root of Fragility of ML Models: Bottom-up Discriminative Learning

Fragility and absence of calibration in machine learning (particularly, deep learning models) is unavoidable consequence of bottom-up discriminative learning.

- Ill-posed inverse problem: Forward models corresponding to causal mechanism need to be inversed. Leads to fragility.
  - $y = x + \sin(x)$ does not have an elementary inverse
  - Moore-Penrose generalized inverse of a linear bounded operator between Hilbert spaces need not be continuous.

  Small perturbations of effects, e.g., measurement noise, can cause arbitrarily large deviations of the inferred cause.

- Generative model has a shorter description in terms of Kolmogorov complexity.
  - Smaller K(X) is easier to learn. The complexity of a set can be much smaller than the complexity of its elements (abstraction - learning abstract concepts should be easier)

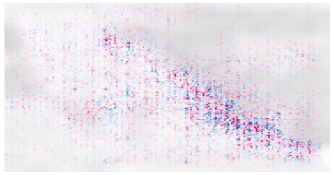
Generative Model: Joint distribution of (input, output, explanation)

Discriminative Model $p(\ class \mid\ \ input\ )$

Conditional Generative Model $p(\ input \mid\ class, expl)$

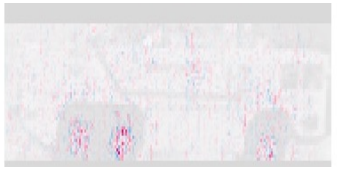
The predicted label and explanation provide context for the input.

Aircraft



Likely

Aircraft

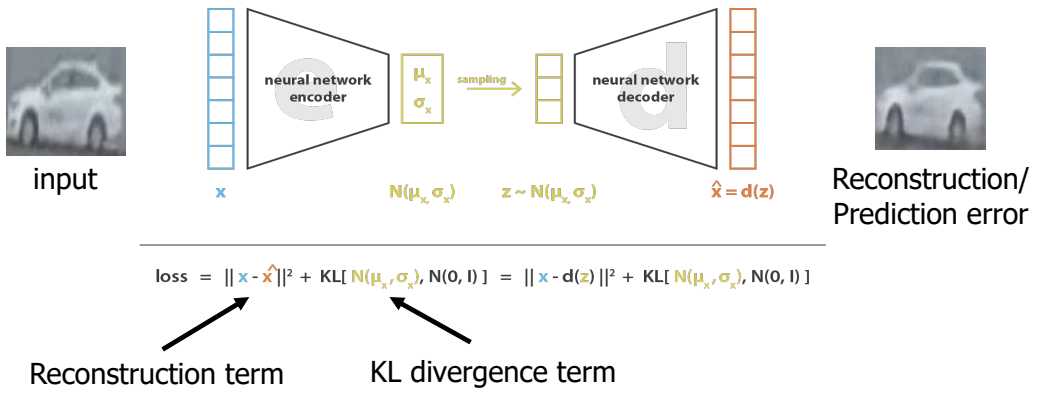


Unlikely

Attributions can be used to quantify uncertainty in prediction of machine learning models.
Attribution-Based Confidence (ABC) Metric For Deep Neural Networks. Jha et. al. (NeurIPS), 2019

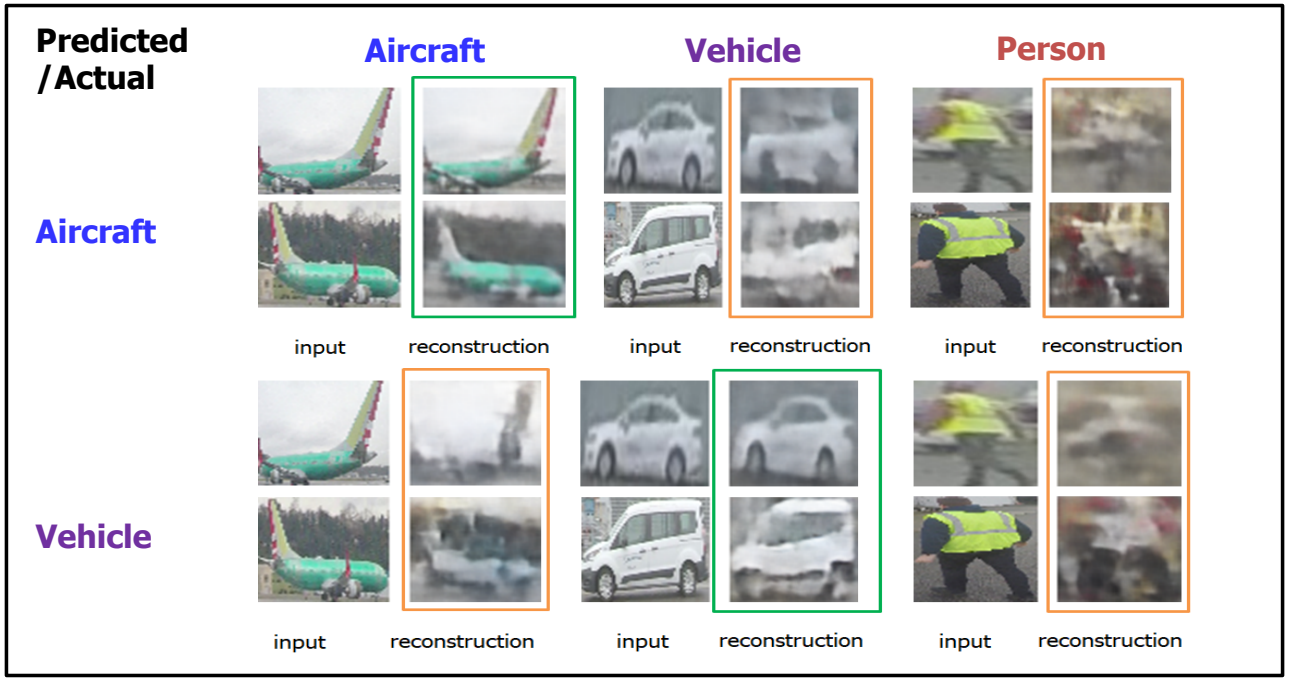
Prediction Using Context: Probabilistic Modeling Using Normalizing Flows.

Context Modeling Using Generative Models - Probabilistic Modeling using Normalizing Flows



Detecting Adversarial Examples Using Data Manifolds. Jha et. al. MILCOM'18

On the Need for Topology-Aware Generative Models for Manifold-Based Defenses. Uyeong et al. ICLR'20

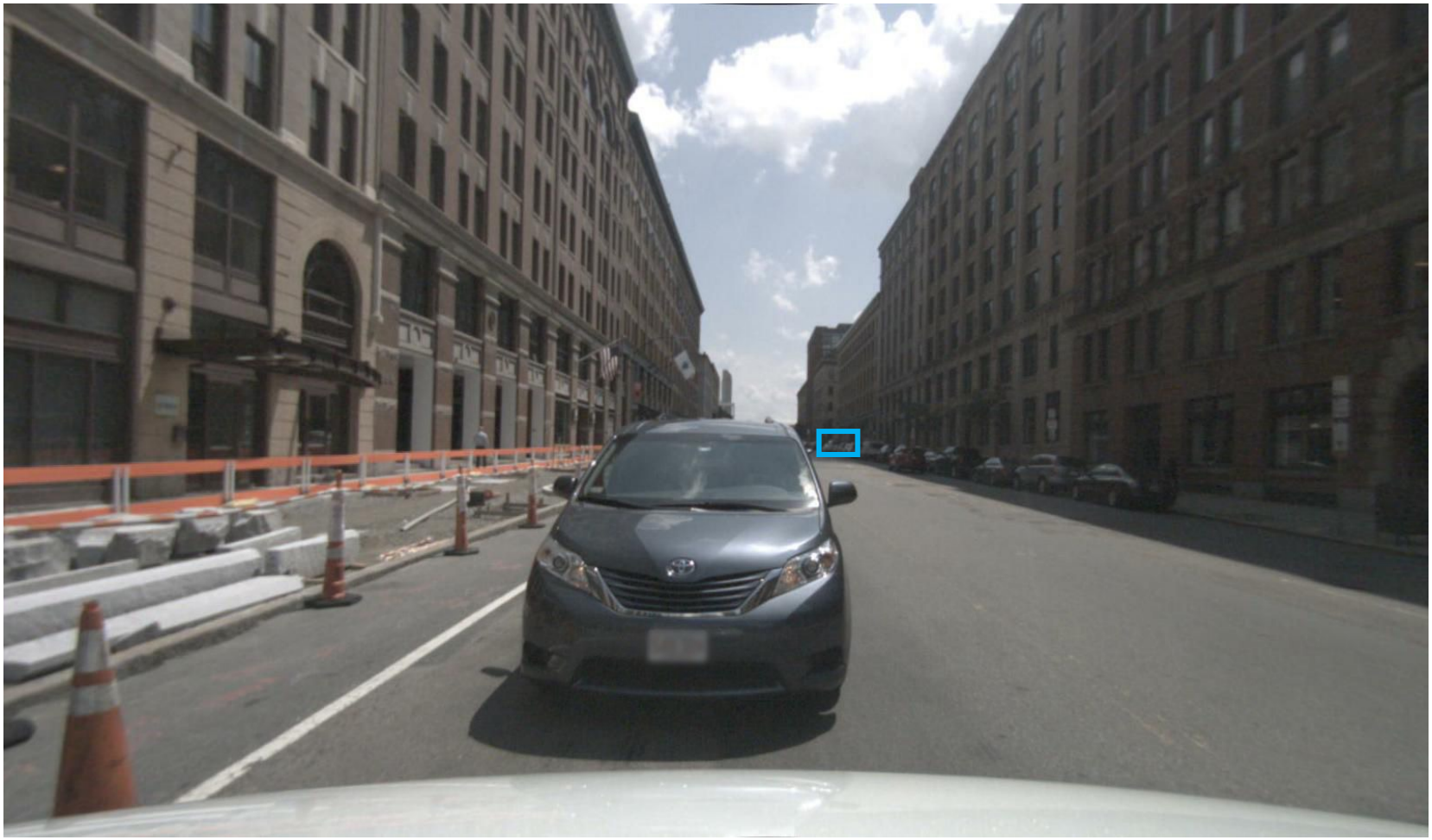


Prediction Using Wider Context

What is this?

Prediction Using Wider Context

Now one can tell – given the context!



Prediction Using Wider Context

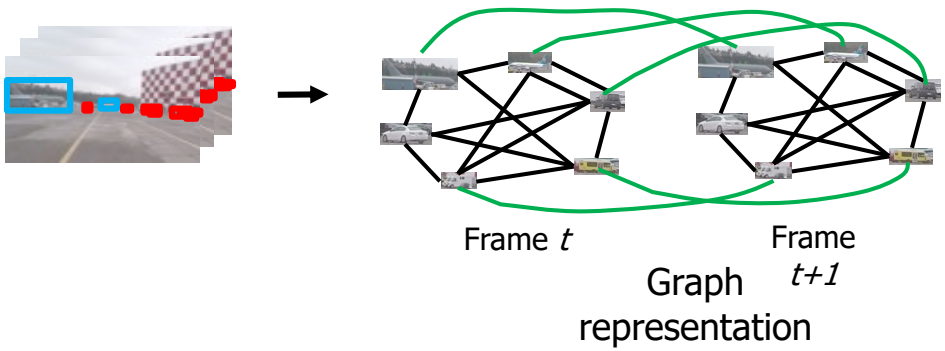
What is this?

Prediction Using Wider Context

Now one can tell given the context.

Prediction Using Wider Context: Low Data Regime

Context Modeling Using Graph Neural Networks



NuScenes Dataset

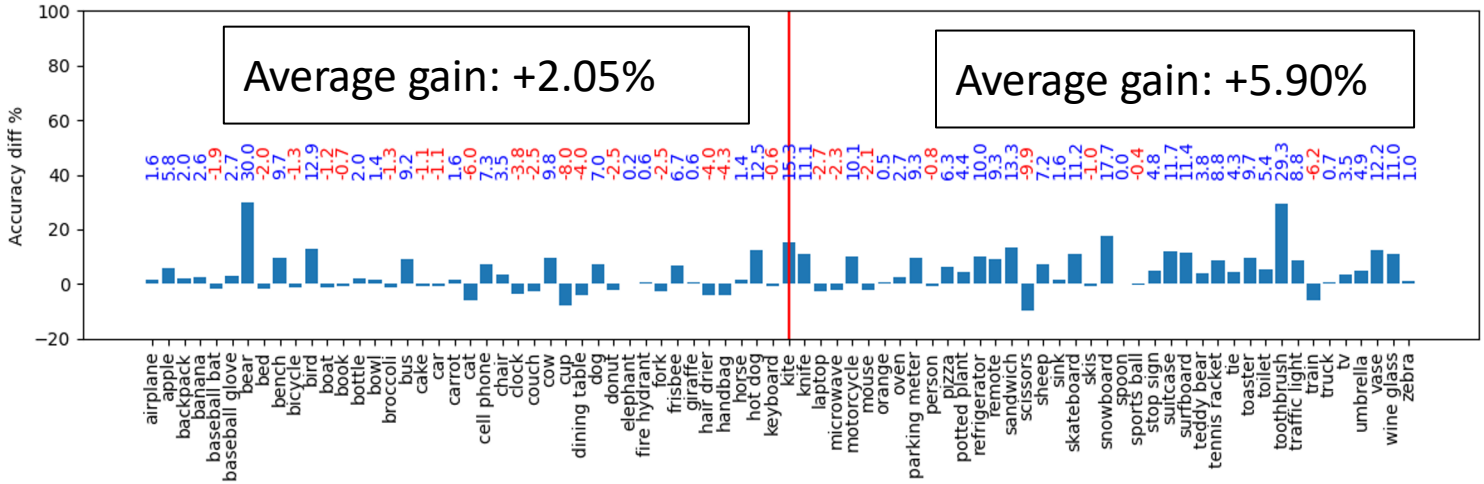
human (19.46%), **bicycle (1.04%), motorcycle (1.11%)**, car (43.62%), truck (12.70%), movable_object (22.05%)

Model	Occlusion (%)	Overall accuracy	Class-wise accuracy					
			human	bicycle	motor-cycle	car	truck	movable object
CNN - ResNet (Baseline)	No occlusion	88.65	92.44	57.24	61.31	92.59	69.74	90.69
CNN - ResNet (Baseline)	50%	79.17	94.93	2.36	12.48	87.33	58.94	67.95
GCNN	No occlusion	95.51	98.38	66.25	73.37	97.13	82.17	98.62
GCNN	50%	93.13	97.53	31.36	64.88	94.17	82.10	96.34

Prediction Using Wider Context: Novel Classes

- **Coco Dataset**: 80 classes, 80K in training set and 40K in test set.
- Train FastRCNN on the alphabetically first 40 classes as the feature extractor.
- Train/test the downstream MLP and GraphCNN on all the 80 classes.

Modeling context using GraphCNN improves prediction particularly over the novel classes.



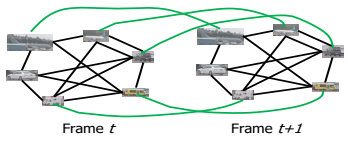
(GCN - MLP) accuracy difference.
Blue = GCN is better
Red = MLP is better
Classes right to the middle vertical line are the 40 novel classes.

Predictive Coding (PC) Based ML

L2

Graph networks

$p(\ node1class,\ node2class,\ node3class\ )$



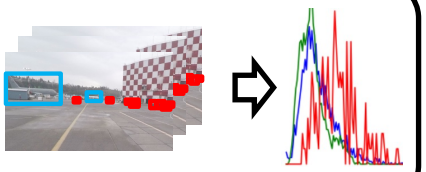
Graph representation

Layer 2 surprise

L1

Probabilistic Generative Model

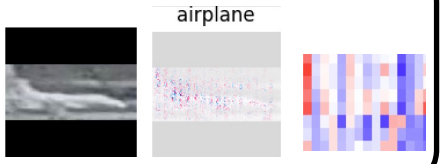
$p(input, expl, class)$



Layer 1 surprise

L0

LEC



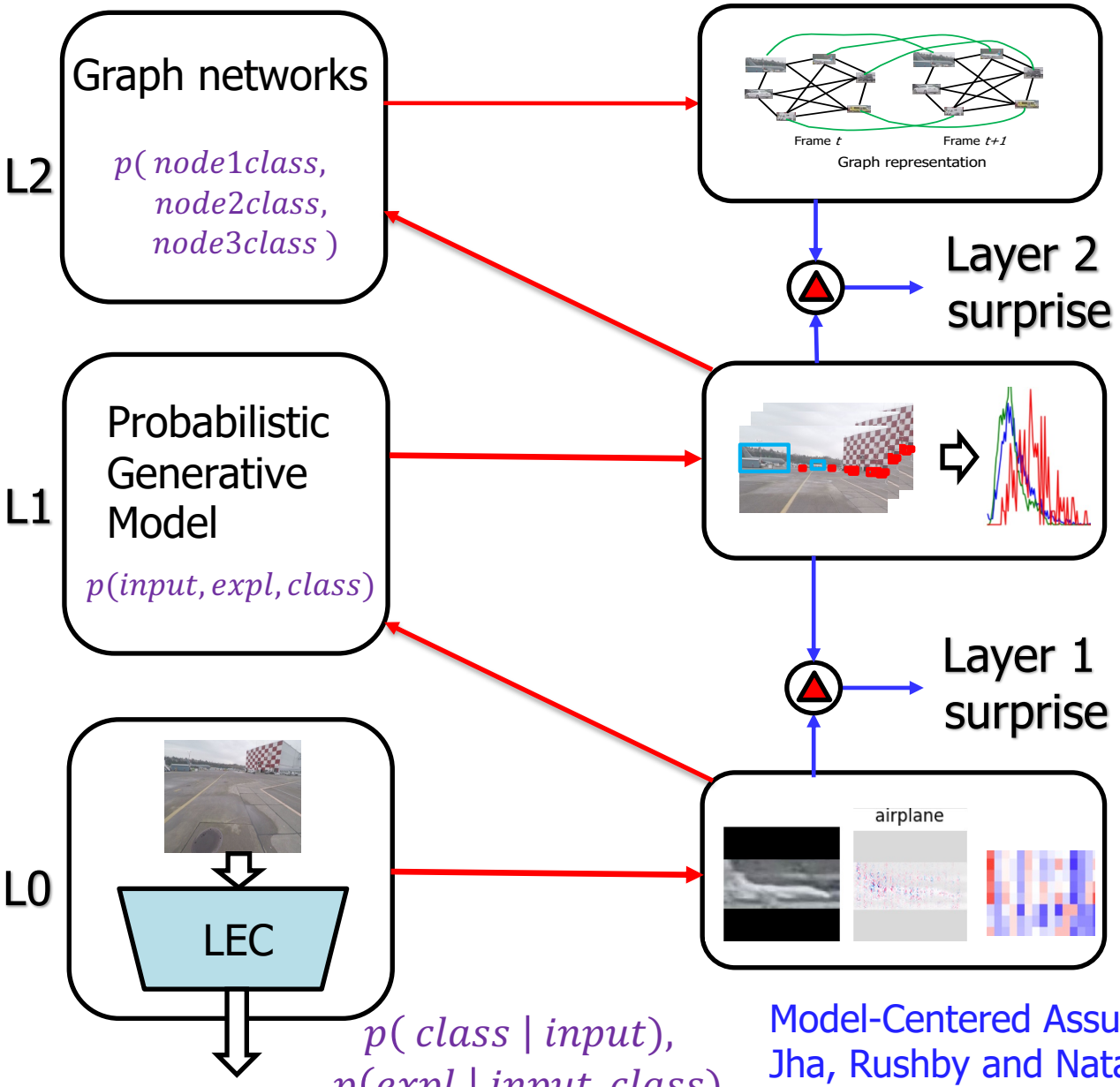
Learning more abstract concepts

Predicting using *larger* contexts

$p(\ class \mid input),$
$p(expl \mid input, class)$

Model-Centered Assurance for Autonomous Systems. Jha, Rushby and Natarajan. SafeCOMP'20

Predictive Coding (PC) Based ML



Level 1

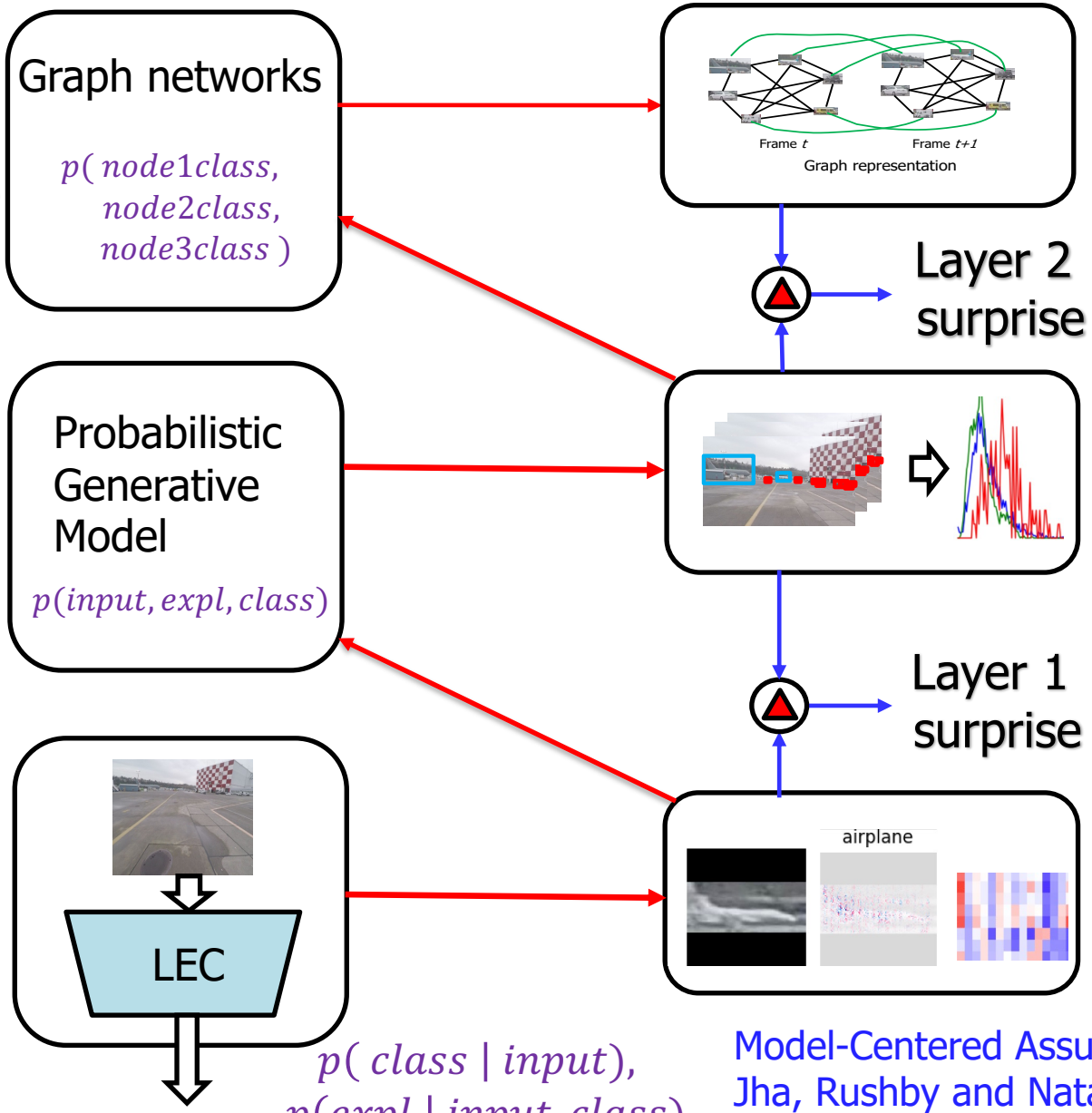
L1 is a probabilistic generative model.

Using normalizing flows for modeling the distribution.

We have explored new architectures such as stochastic differential equation models and probabilistic circuits to implement the generative models.

Model-Centered Assurance for Autonomous Systems. Jha, Rushby and Natarajan. SafeCOMP'20

Predictive Coding (PC) Based ML



Level 2

L2 uses Graph Neural Networks and Markov Logic Networks.

This enables the use of larger spatial and temporal context making models robust to perturbations.

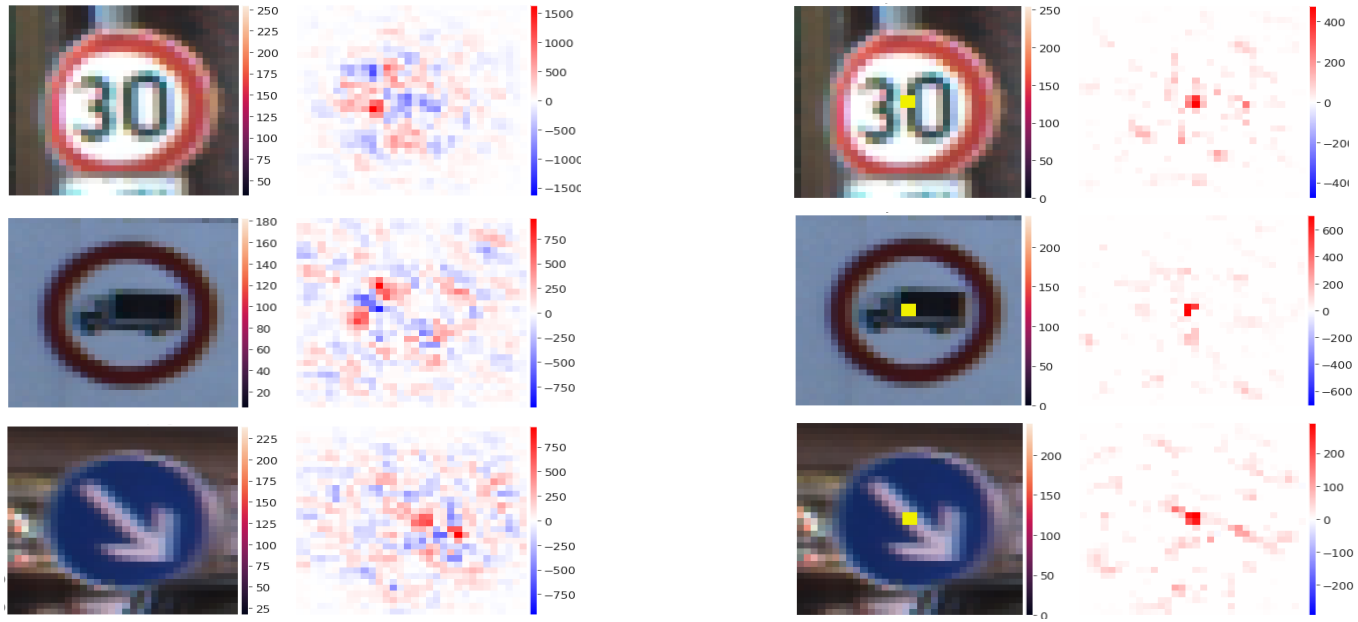
Model-Centered Assurance for Autonomous Systems. Jha, Rushby and Natarajan. SafeCOMP'20

Rest of the talk

- Basic Principle behind TrinityAI – Predictive Coding and the use of Context
- Context-driven Trustworthy and Robust Learning
- Improving Resilience Using Attributions/Explanations
- Improving Attributions by Making Learning Models Robust
- TrinityAI Tool and Ongoing Work

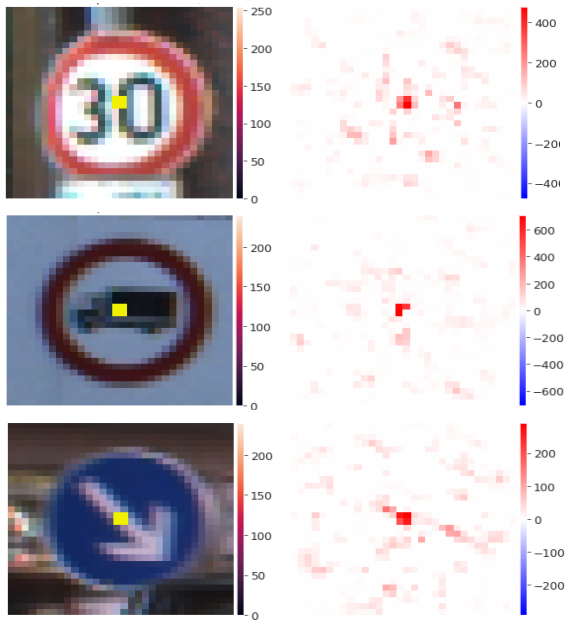
Improving Resilience Using Attributions/Explanations

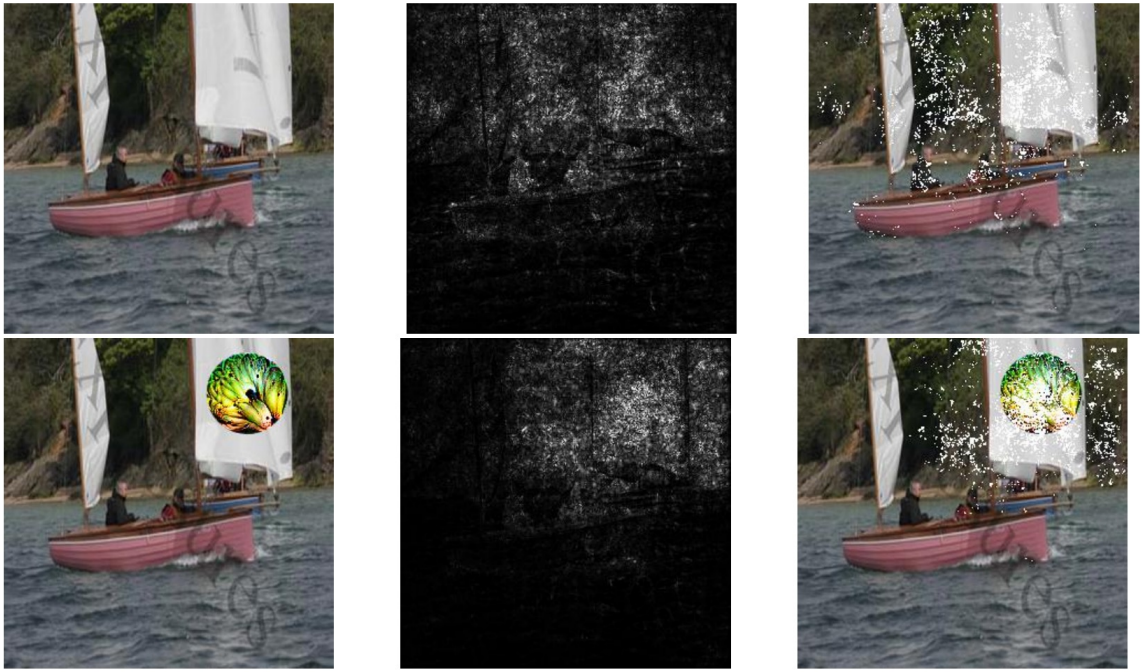
Adversarial perturbations cause disproportionally high concentration of attributions.



Improving Resilience Using Attributions/Explanations

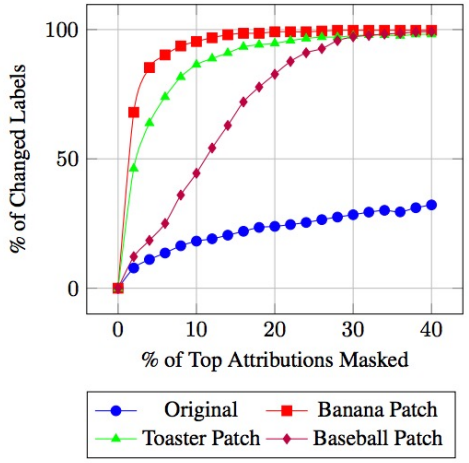
The decision of machine learning model changes when a small percentage of high attribution features of an adversarial input is masked.

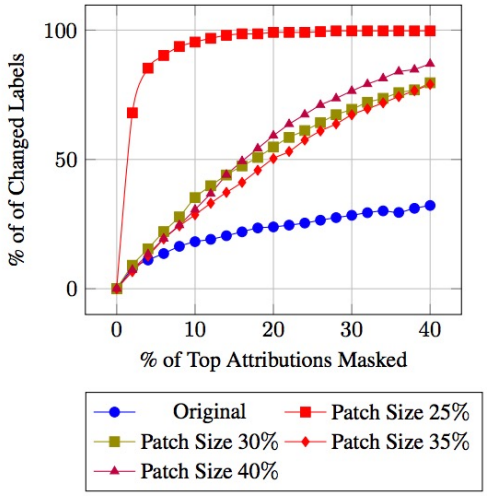


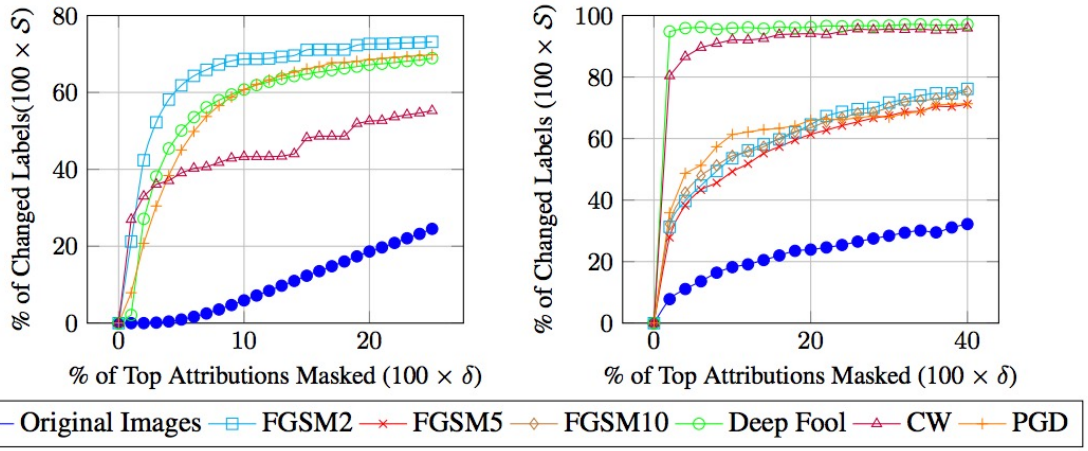


Improving Resilience Using Attributions/Explanations

The decision of machine learning model changes when a small percentage of high attribution features of an adversarial input is masked.



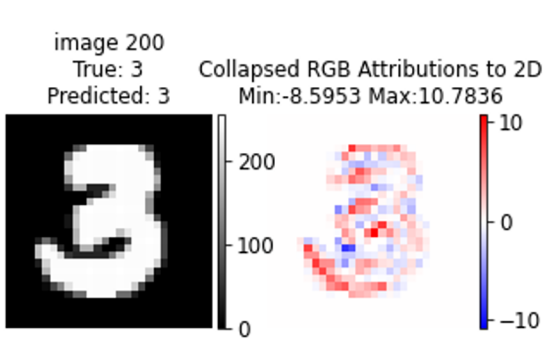


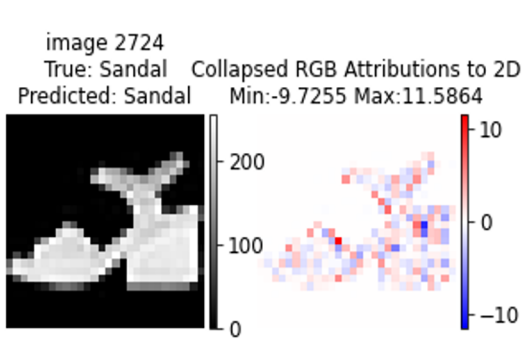


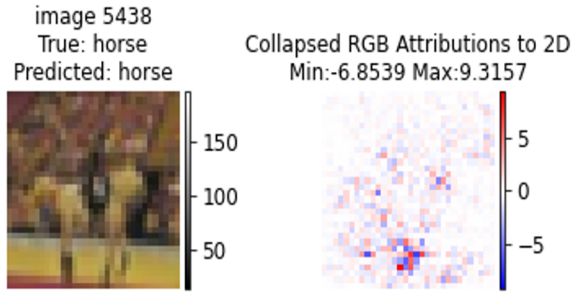
Attribution-Based Confidence (ABC) Metric For Deep Neural Networks. Jha et. al. (NeurIPS) 2019

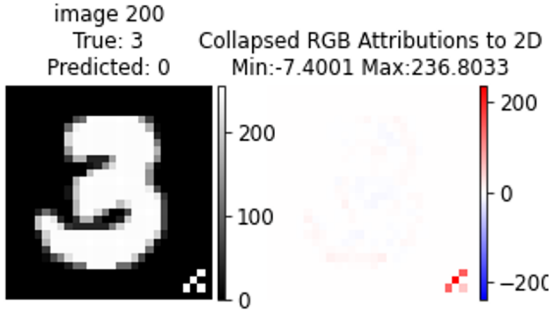
Detecting Backdoors in ML Models using Attributions

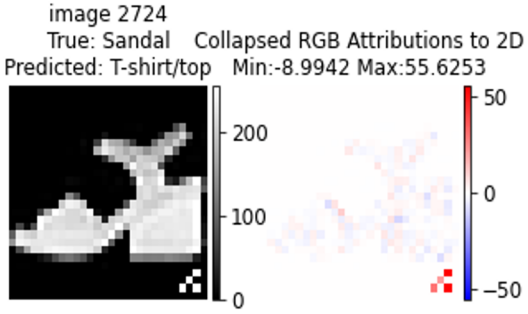
Trojan trigger causes disproportionally high concentration of attributions.

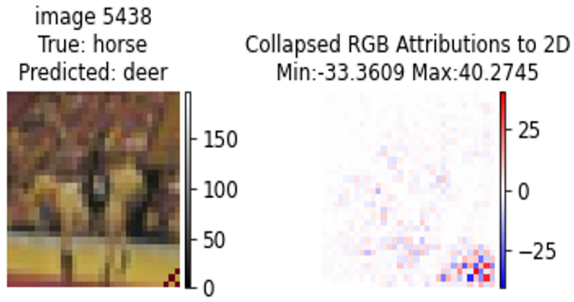








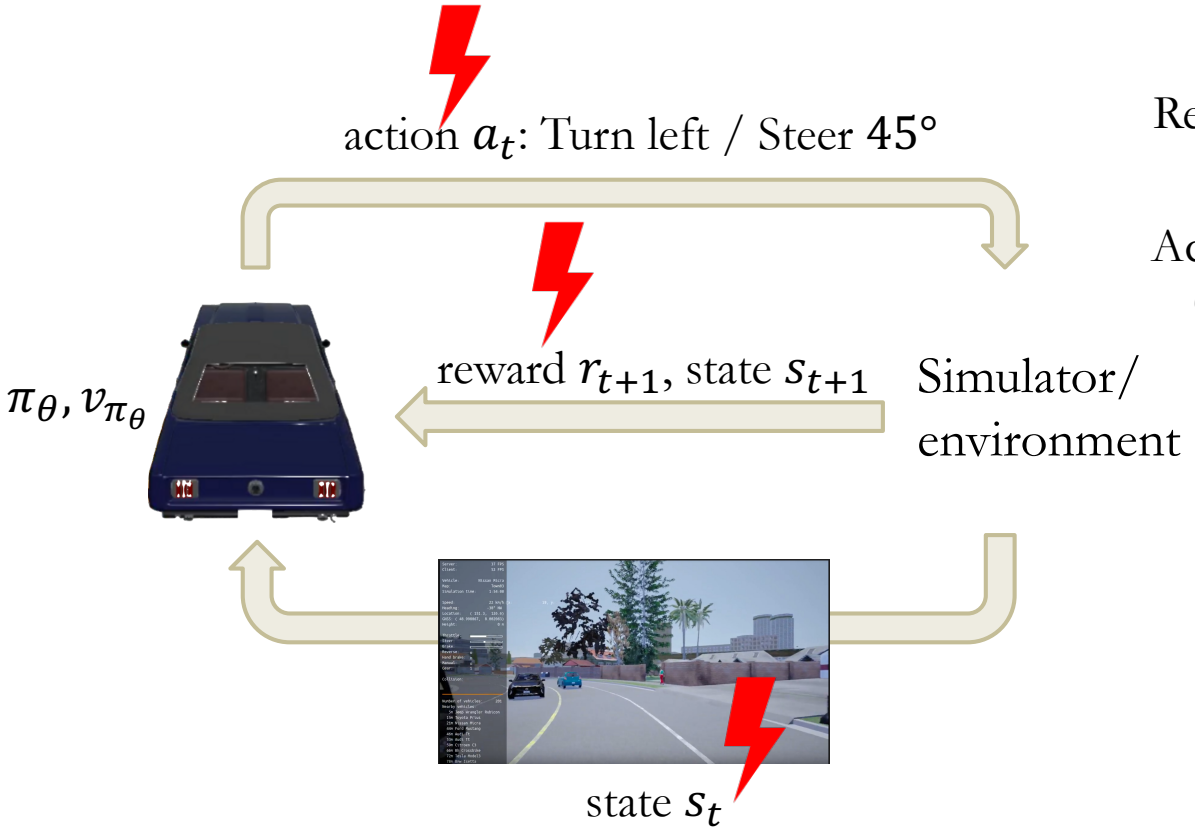




MISA: Online Defense of Trojaned Models using Misattributions. Kiourti et. al. ACSAC'21

Trojan/Backdoor Attacks on Reinforcement Learning

We had initially developed a Trojan attack on RL policies.



Return: $R_t = \sum_{k=t+1}^{T} \gamma^{k-t+1} r_k$

Action-value function:
$Q_\pi(s_t, a_t) = E_\pi[R_t | s_t, a_t]$

Value function:
$v_\pi(s_t) = E_\pi[Q_t \mid s_t]$

Advantage:
$A(s_t, a_t) = Q_\pi(s_t, a_t) - V_\pi(s_t)$

TrojDRL: Evaluation of Backdoor Attacks on Deep Reinforcement Learning. Kiourti et al. DAC'20

Trojan/Backdoor Attacks on Reinforcement Learning

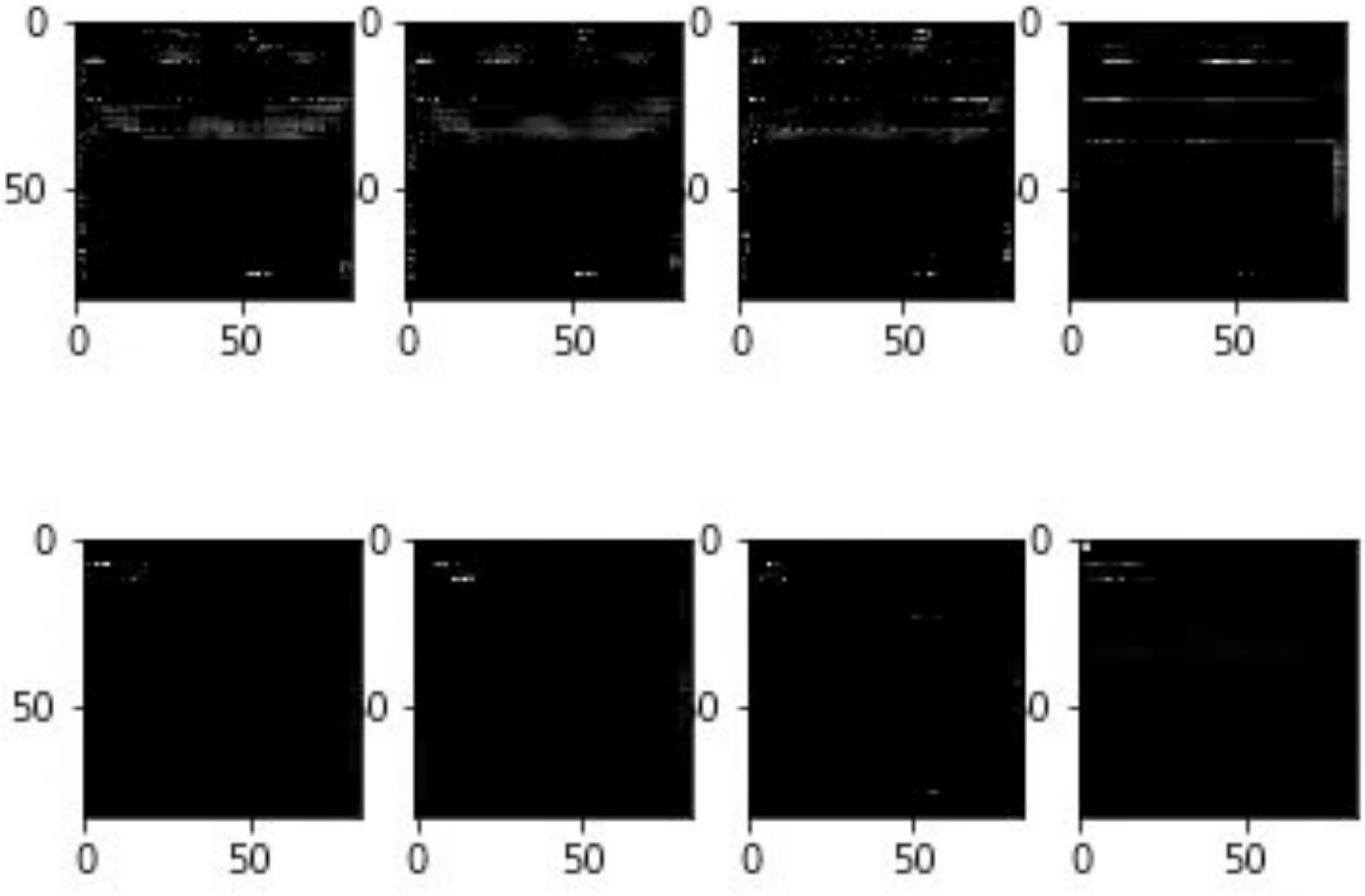
We had initially developed a Trojan attack on RL policies.

Game	Score during the attack					
	Targeted		Untargeted		Standard	
	Mean	Std	Mean	Std	Mean	Std
Breakout	1	1	2	2	250	147
Qbert	658	1176	965	1220	7890	2770
Seaquest	7	10	32	18	220	111
Space Invaders	13	12	50	47	161	230
Crazy Climber	0	0	0	0	13870	11562

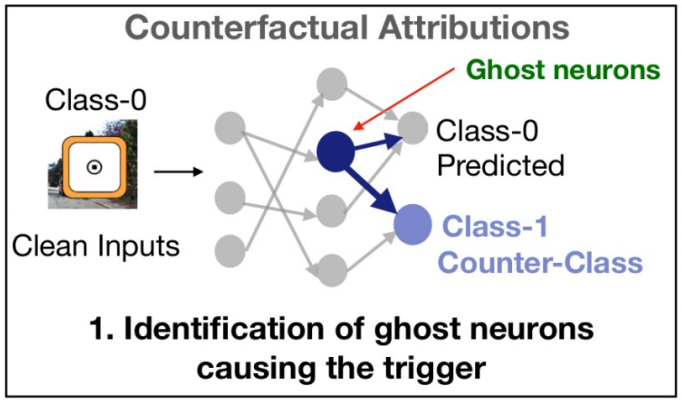
TrojDRL: Evaluation of Backdoor Attacks on Deep Reinforcement Learning. Kiourti et al. DAC'20

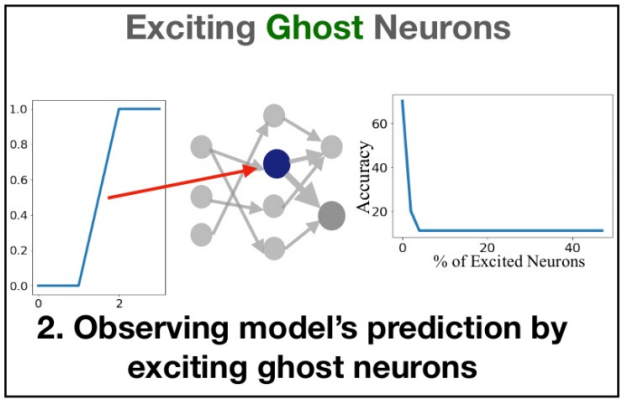
Detecting Trojans in RL Policy Using Attributions/Explanations

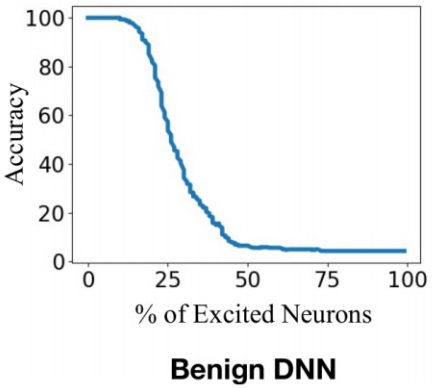
Attributions can detect Trojan triggers in backdoored observations.

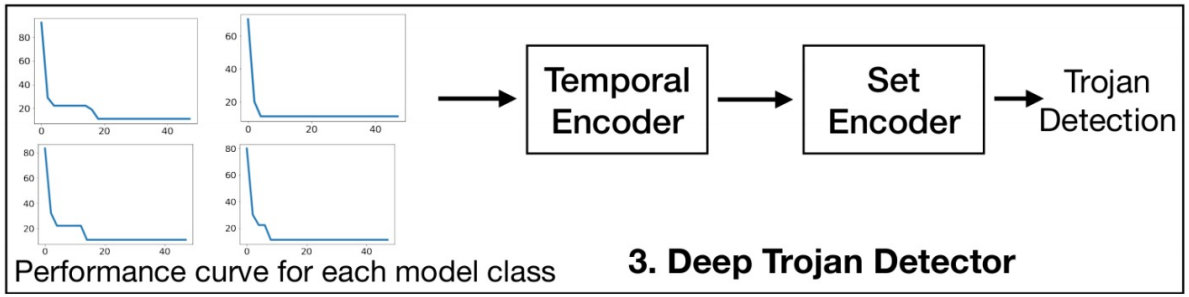


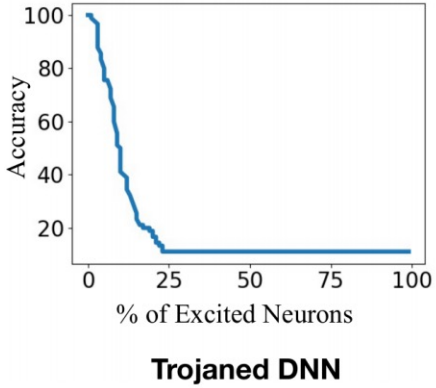
Attribution-based Offline Trojaned Model Detection Using Only Clean Data









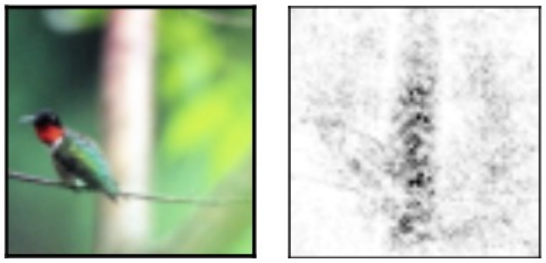


Model	Triggered-MNIST	TrojAI-Round1	TrojAI-Round2	TrojAI-Round3
Cassandra [62]	$\mathbf{0.97 \pm 0.010}$	0.88 ± 0.006	0.59 ± 0.096	0.71 ± 0.026
Neural Cleanse [55]	0.70 ± 0.045	0.50 ± 0.030	0.63 ± 0.043	0.61 ± 0.064
ULP [28]	0.54 ± 0.051	0.55 ± 0.058	–	–
TrinityAI-Conv-IG	0.89 ± 0.024	0.87 ± 0.020	0.73 ± 0.014	0.71 ± 0.038
TrinityAI-Tx-IG	0.95 ± 0.022	0.89 ± 0.029	0.75 ± 0.033	$\mathbf{0.72 \pm 0.038}$
TrinityAI-Conv-GradxAct	0.87 ± 0.030	0.88 ± 0.027	0.74 ± 0.030	0.67 ± 0.036
TrinityAI-GradxAct	0.96 ± 0.014	$\mathbf{0.90 \pm 0.027}$	$\mathbf{0.76 \pm 0.027}$	0.66 ± 0.029

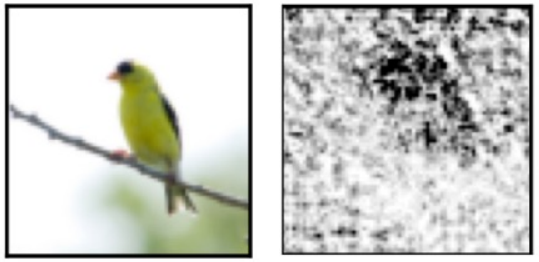
Detecting Trojaned DNNs Using Counterfactual Attributions. Sikka, Sur, Jha, Roy, Divakaran. ArXiv'21

Improving Attributions by Making Learning Models Robust

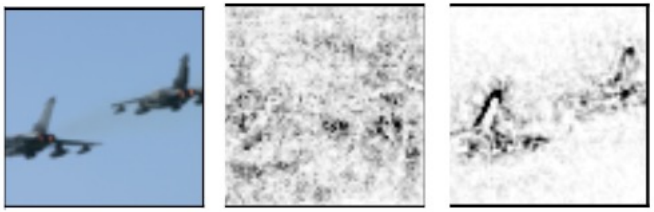
Attribution/explanation methods are noisy.



DeepLIFT



Integrated Gradient



Integrated Gradient + Noise Tunnel

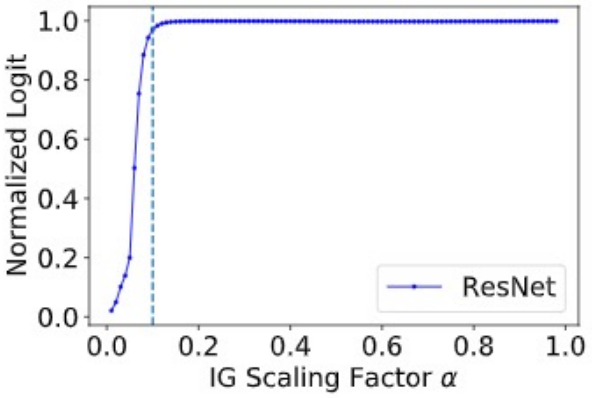
DeepShap

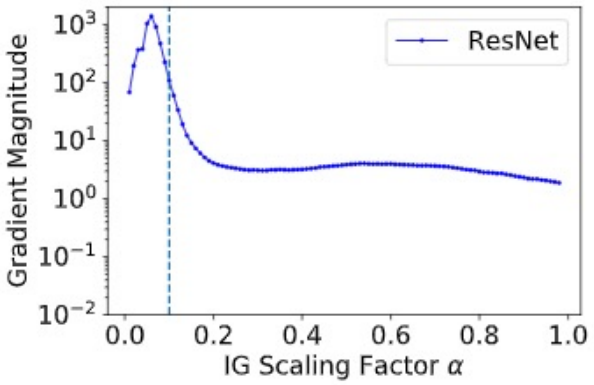
Improving Attributions by Making Learning Models Robust

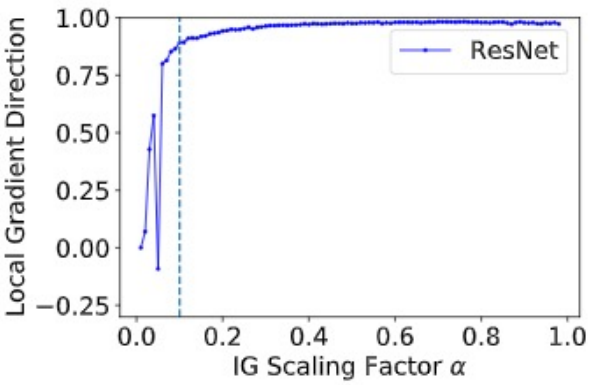
$$IG(F_i(x)) = (x_i - x_i') \times \int_{\alpha=0}^{1} \partial_i M(x' + \alpha(x - x'))d\alpha$$

where $\partial_i M(\cdot)$ denotes the gradient of $M(\cdot)$ along the i-th feature dimension.

Integrated Gradient shows non-intuitive accumulation of attribution beyond output saturation.





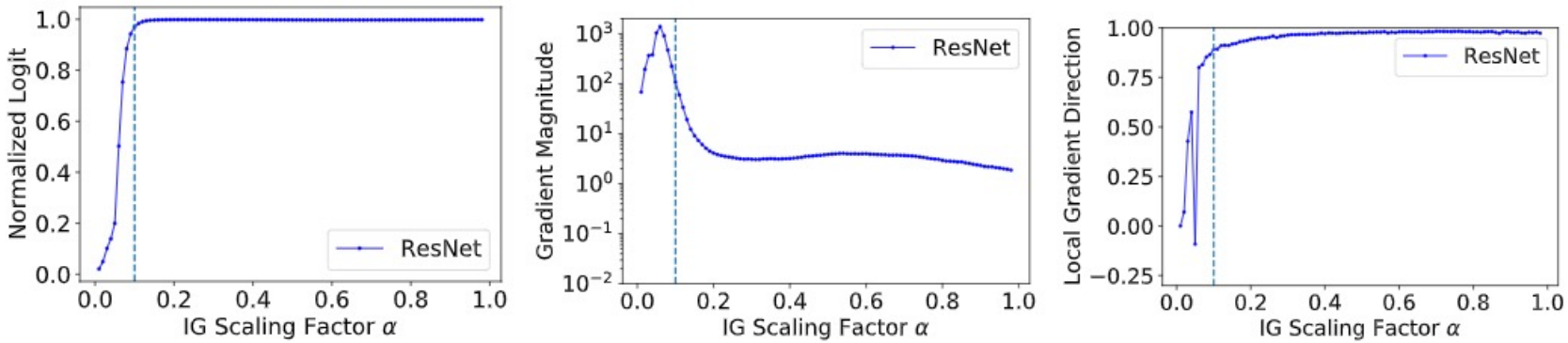


Model output saturated at 0.1

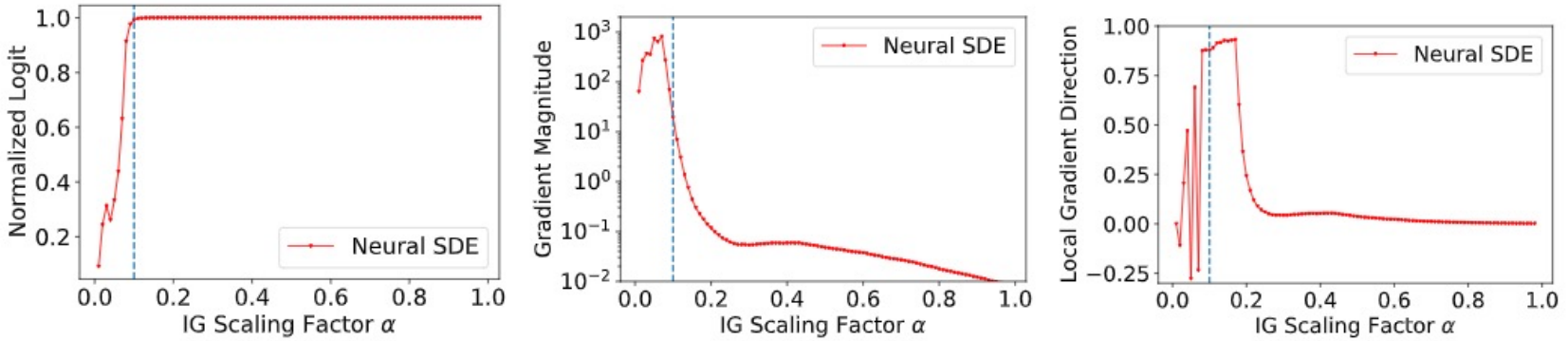
The gradients beyond 0.1 are still non-zero.

The direction of the gradients are correlated and hence, getting cumulatively added.

Training Robust Neural Stochastic Differential Equation Model improves Attributions



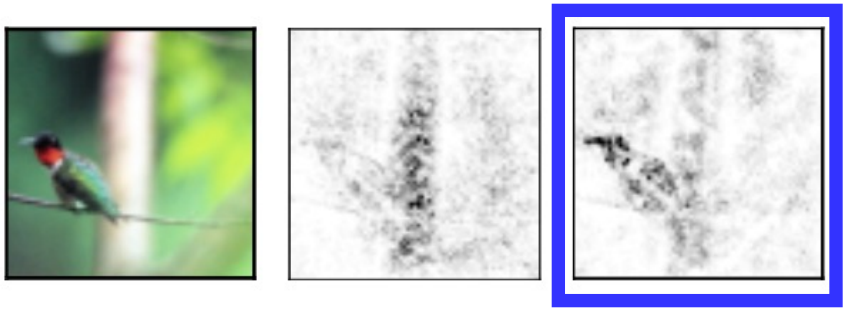
Resnets



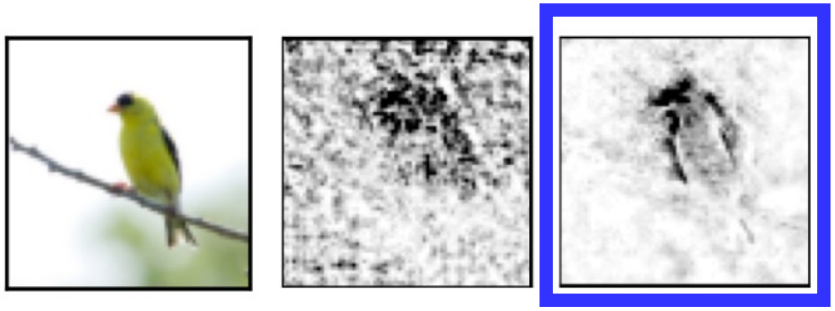
Neural SDEs: Gradients approach zero beyond decision point, direction also gets uncorrelated.

On Smoother Attributions using Neural Stochastic Differential Equations. Jha et al. IJCAI'21

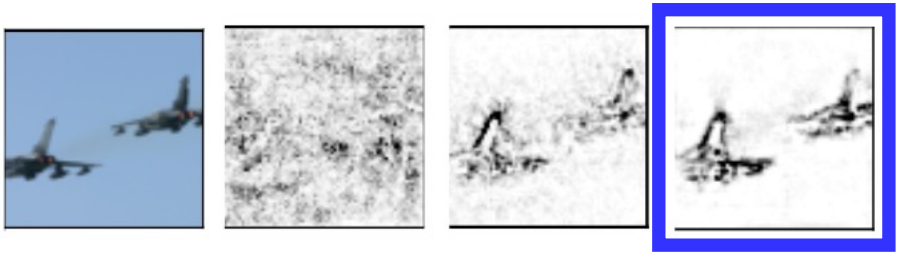
Training Robust Neural Stochastic Differential Equation Model improves Attributions



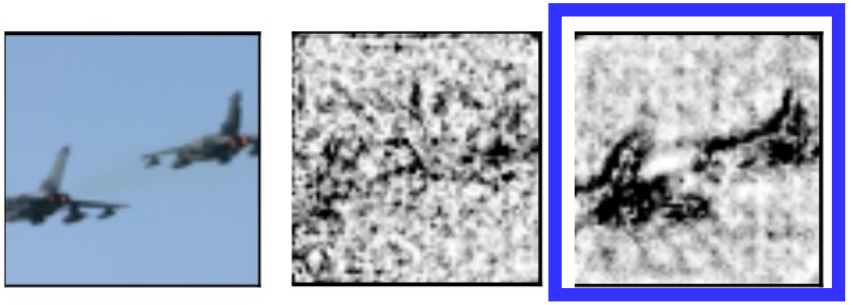
DeepLIFT



Integrated Gradient



Integrated Gradient + Noise Tunnel

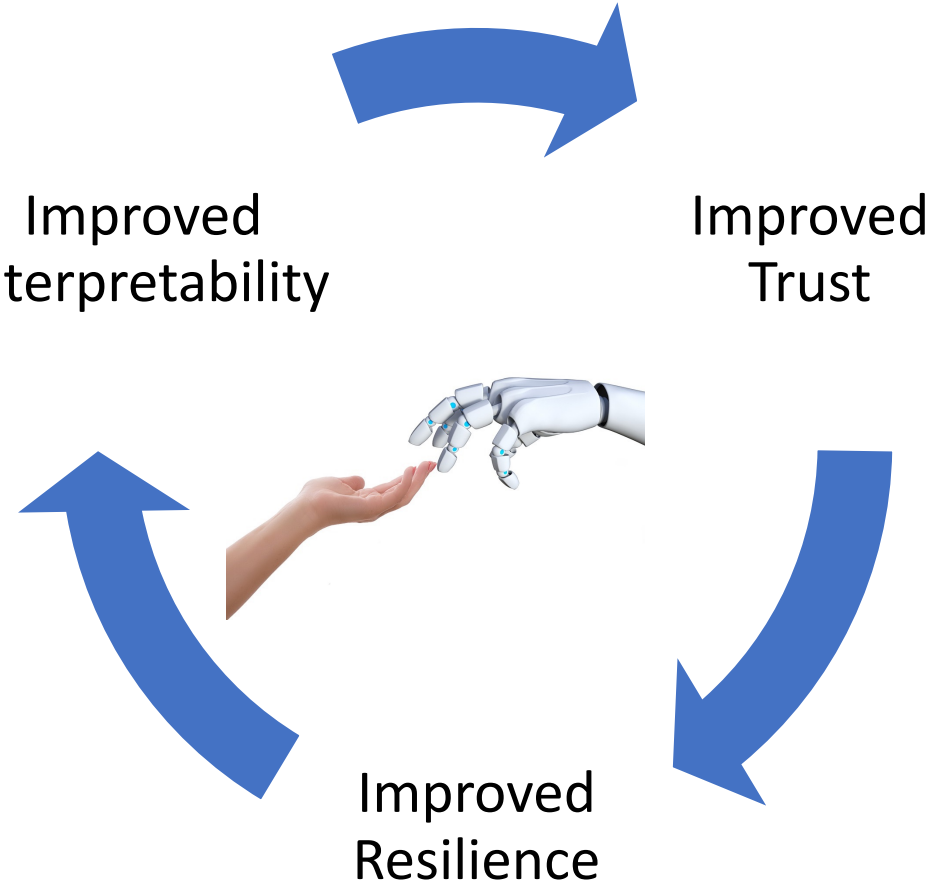


DeepShap

On Smoother Attributions using Neural Stochastic Differential Equations. Jha et al. IJCAI'21

Three Coupled Challenges in AI

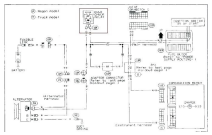
The dependency between Trust, Resilience and Interpretability also creates a virtuous cycle.



Simultaneously improvement in trustworthiness, resilience and interpretability is critical for their use in high-assurance systems and in human-machine teams.

TRINITY: Trust, Resilience and Interpretability of AI

Demonstrations

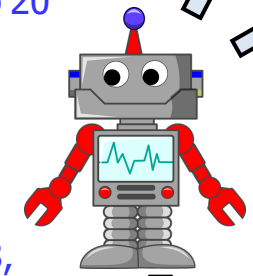


Specifications

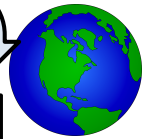
Neurosymbolic Learning
JAR'18,
NeurIPS'18, RV17,
SafeComp'20

Formal Verification
NASA FM'18, ADHS'18, HSCC'19, VNN/AAAI'19

Resilience to Adversarial Attacks
MILCOM'18, SafeML/ICLR'19,
NeurIPS'19, DAC'20, ICLR'20, IJCAI'21



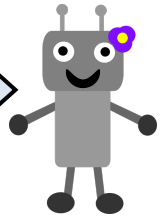
Robust RL
Allerton Control'18
CoRL'20



World

Uncertainty-aware Planning
FORMAT'16, NASA FM'16, FORMATS'18,
JAR'18, ACC'19

Symbolic Explanations
NASA FM'17,
ConsciousAI/AAAI'19



Human User

Project Acknowledgements

- U.S. Army Internet of Battlefield Things
- DARPA Assured Autonomy
- DARPA Symbiotic Design of Cyberphysical Systems
- IARPA TrojAI
- NSF Self-Improving CPS, NSF Synthesis

TRINITY tool Release: November, 2021
Combining results from DARPA Assured Autonomy, IARPA TrojAI and US ARL IoBT CRA

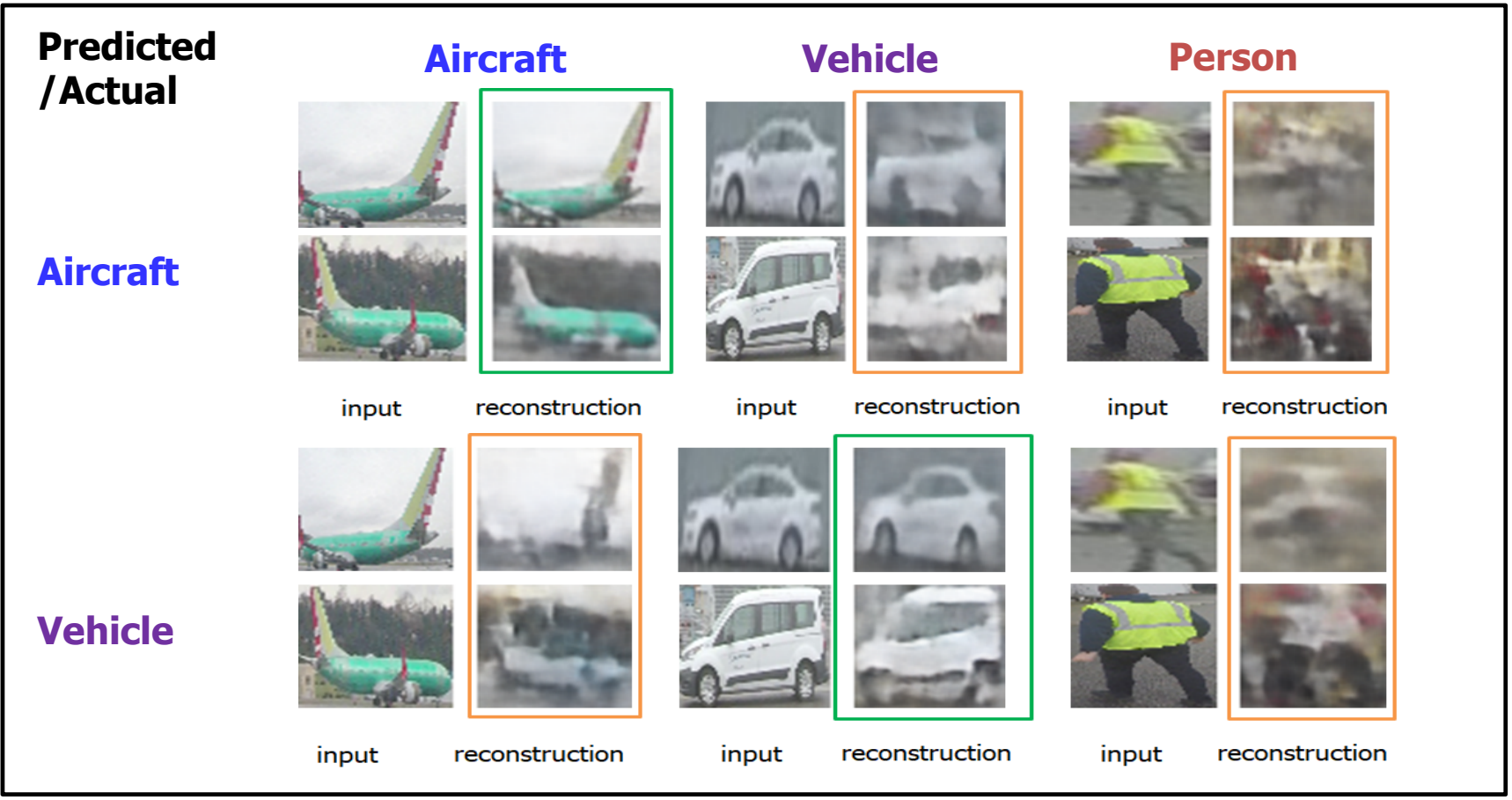
We are hiring (full-time, post-docs, students). Please contact me if you are interested.

Thank You!

Thank You!

BACKUP SLIDES

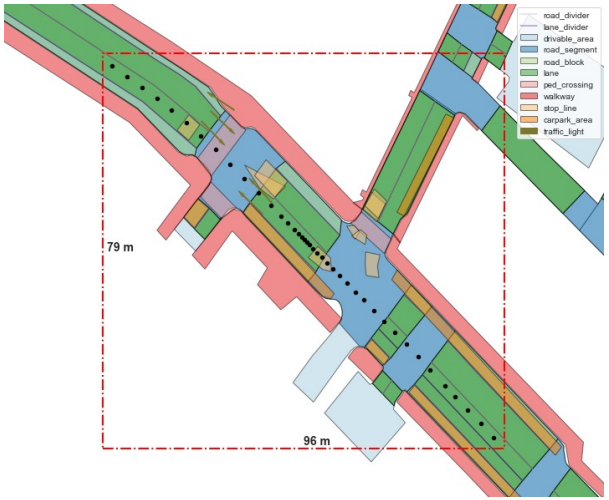
an generative models help? Demonstrated Using Novel Class n)

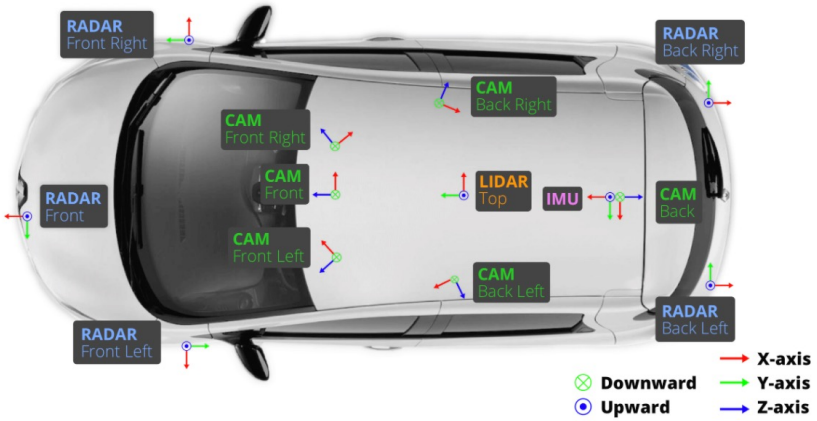


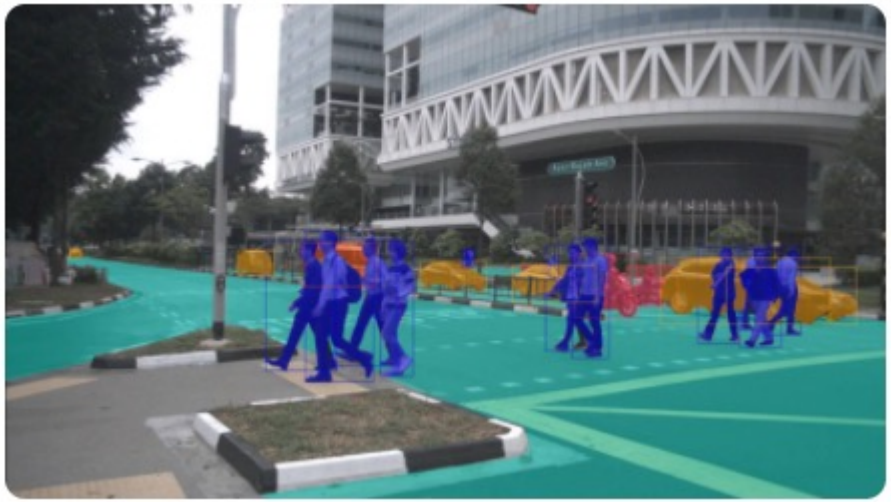
The NuScenes dataset

Sensors:

- 1x LIDAR, 5x RADAR, 6x camera, IMU, GPS
- 1000 scenes of 20s each
- Two diverse cities: Boston and Singapore
- Detailed map information (segmentation)
- 1.4M 3D bounding boxes manually annotated
  Attributes such as visibility, activity and pose







Experimental setup

Goal: Object classification using contextual cues

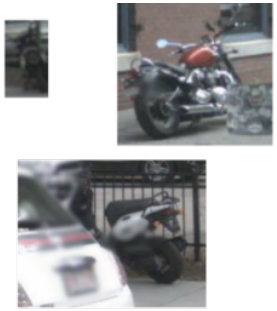
Object classes: We consider six object classes

- Object classes and frequency of samples:

human (19.46%), **bicycle (1.04%), motorcycle (1.11%)**, car (43.62%), truck (12.70%), movable_object (22.05%)

human

bicycle



motorcycle

car

truck

movable_object

Results: Novelty Detection

Novel object detection:

Datasets: Tiny imagenet with 200 object classes.

20 classes are available during training and

remaining 180 classes considered as the novel objects.

Metrics: Area under ROC (AUROC) for novel object recognition and detection accuracy (DTACC) for the closed set recognition

Novel object recognition:

TinyImageNet	OpenMax (CVPR16)	G-OpenMax (BMVC17)	OSRCI (ECCV18)	C2AE (CVPR19)	CROSR (CVPR19)	Gen-dis (CVPR20)	Ours
AUROC	57.6	58.0	58.6	58.1	58.9	64.7	**73.26**

Closed set recognition:

TinyImageNet	Gen-dis (CVPR20) Resnet-18	Gen-dis (CVPR20) WideResnet-28-10	Ours
DTACC	49.2	55.9	**74.74**

Results: OOD Detection

OOD detection:

Datasets: MNIST, KMNIST, F-MNIST, CIFAR-10, CIFAR100, STL10, SVHN, LSUN, ImageNet

Metrics: True negative rate (TNR) @ true positive rate (TPR) = 95%, Area under ROC (AUROC)

Comparison with state of the art approaches: Liang et al., 2017 / Lee et al., 2018 / PC

The best results are **highlighted**.

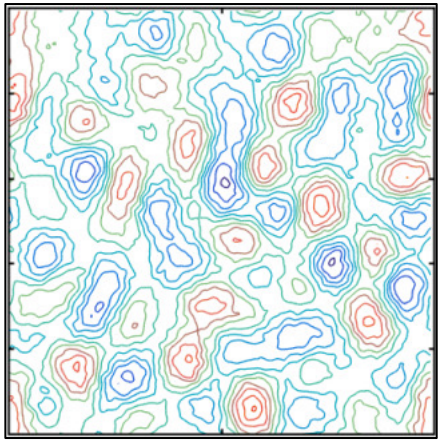
In-dist	OOD	**TNR @ TPR=95%**	**AUROC**
MNIST (LeNet5)	KMNIST F-MNIST	67.72/80.52/**88.91** 58.47/63.33/**67.49**	92.98/96.53/**97.9** 90.76/94.11/**95.37**
CIFAR10 (ResNet50)	STL10 SVNH ImageNet LSUN	12.19/10.33/**14.73** **86.61**/34.49/68.92 73.23/29.48/**74.87** **80.72**/32.18/78.81	60.29/61.95/**64.93** **84.41**/78.19/82.44 94.91/84.3/**95.18** **96.51**/87.09/96.12

Neural ODE vs Neural SDE

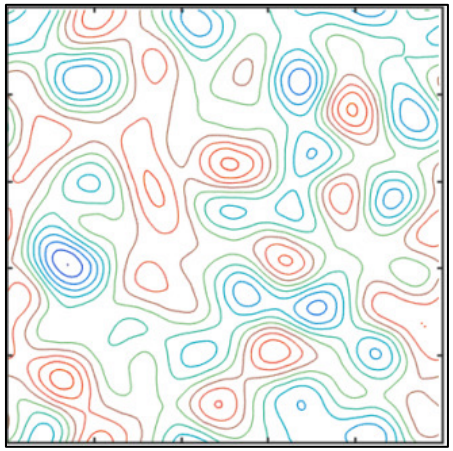
$$\frac{d}{dt}(u(x(t),t) = \frac{\partial u(x,t)}{\partial t} + \overline{G}\left(x(t_l), w(t_l)\right)\nabla u(x,t) + \frac{1}{2}\sigma^2 \Delta u(x,t) = 0$$

$u(x,0)$ serves as the classifier and the velocity field $\overline{G}(x, w(t))$ encodes ResNet's architecture and weights.

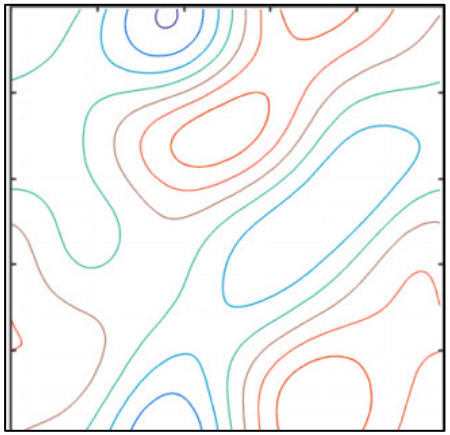
When $\overline{G}$ is very complex, $u(x,0)$ might be highly irregular i.e. a small change in the input x can lead to a massive change in the value of $u(x,0)$



$\sigma = 0$



$\sigma = 0.01$



$\sigma = 0.1$

Solutions of the convection diffusion equation

Neural SDE Theorems

If $G(x(t), W(t))$ is Lipschitz function in both x and t, the target classifier being learned is a compactly supported bounded function and $0 < \sigma \leq 1$, then the solution $u(x,t)$ for the equation above satisfies

$$|u(x+\delta, 0) - u(x,0)| \leq \alpha \left(\frac{|\delta|}{\sigma}\right)^{\beta}$$

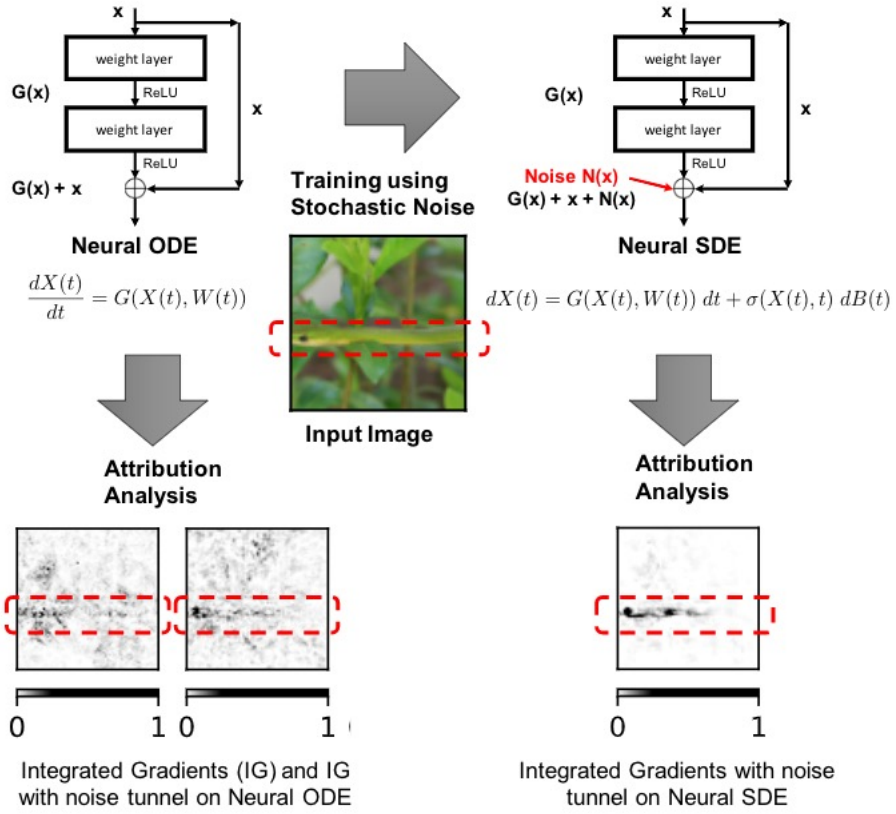
For any small perturbation δ where $\beta > 0$ and α depends on the infinity norm of $G(x(t), W(t))$

If $G\big(x(t), W(t)\big)$ is a continuously differential function in both x and t, the target classifier being learned is a compactly supported bounded function and $0 < \sigma \leq 1$, then the solution $u(x,t)$ for the equation above satisfies

$$|\nabla u(x,1)| \leq \alpha e^{-\sigma^2+\beta}$$

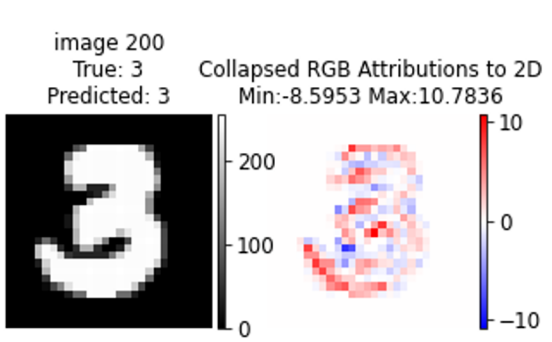
For any small perturbation β depends on ∇G and α depends on the infinity norm of the classifier and its gradient.

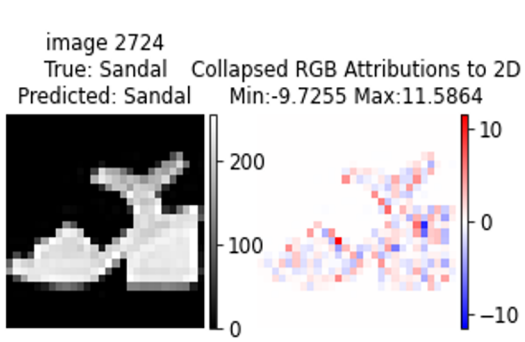
Attributions over neural SDEs

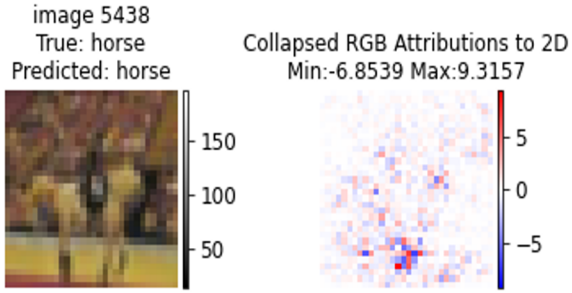


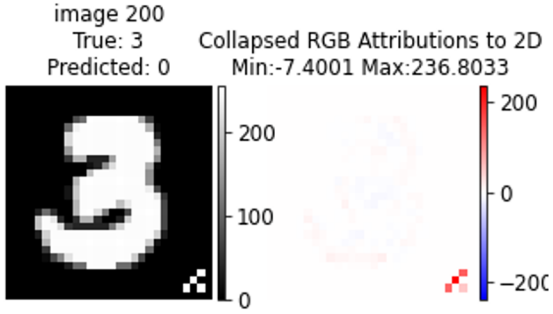
sing Attribution to localize Attack Perturbations

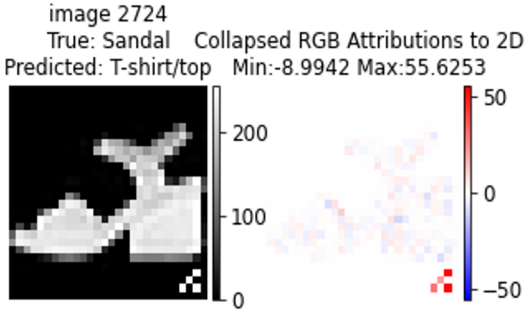
Trojan trigger causes disproportionally high concentration of attributions.
Connection between attribution and adversarial attacks.

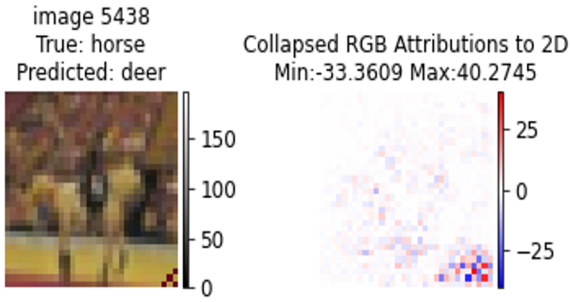




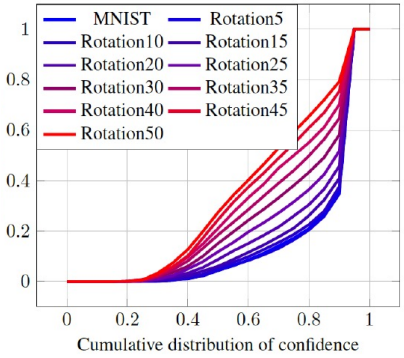


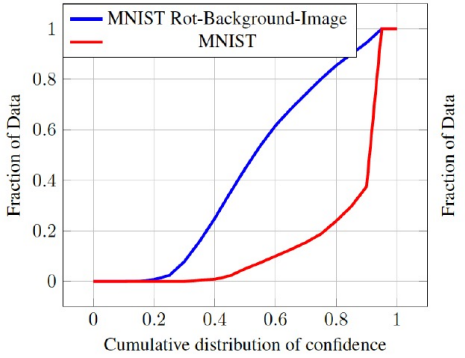


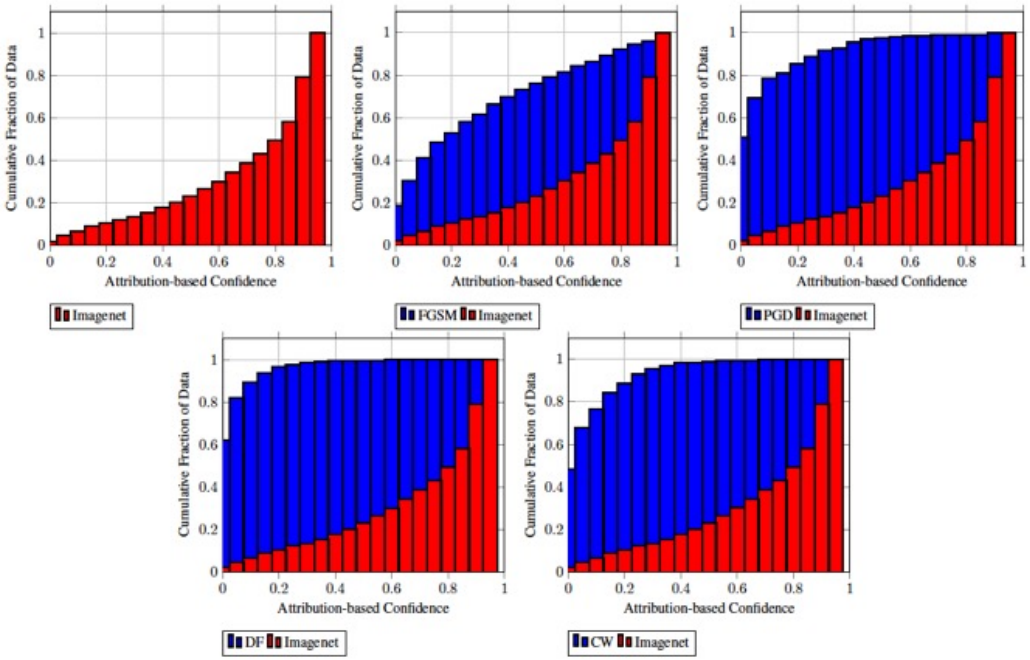




tribution-based Confidence (ABC)







2 misclassified as 9
AttributeConf=0.28

3 misclassified as 2
AttributeConf=0.41

8 misclassified as 2
AttributeConf=0.89

Graph neural networks

Graph neural networks as context encoders

$$H^{(l+1)} = f(H^{(l)}, A)$$

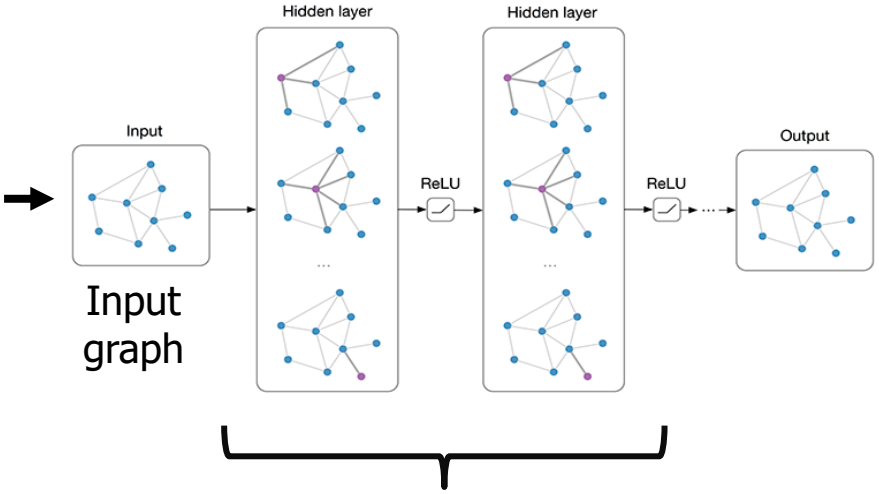
Layer $l+1$ th representation

Neural network

Layer l th representation

Adjacency matrix

videos with object detections



Input graph

Graph convolutions

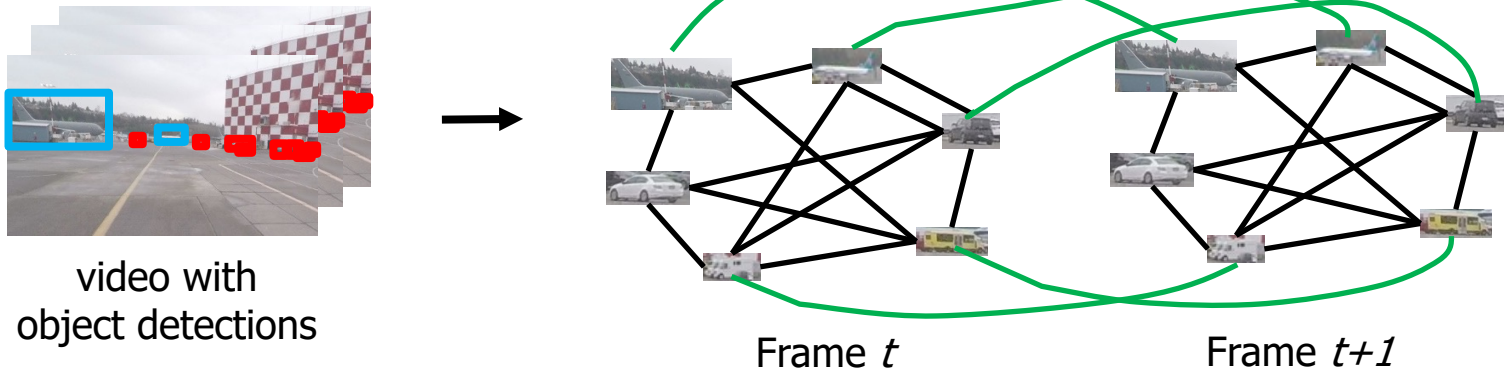
Graph classification (detect scenes with anomalies)

Node classification (detect unusual objects in the scene)

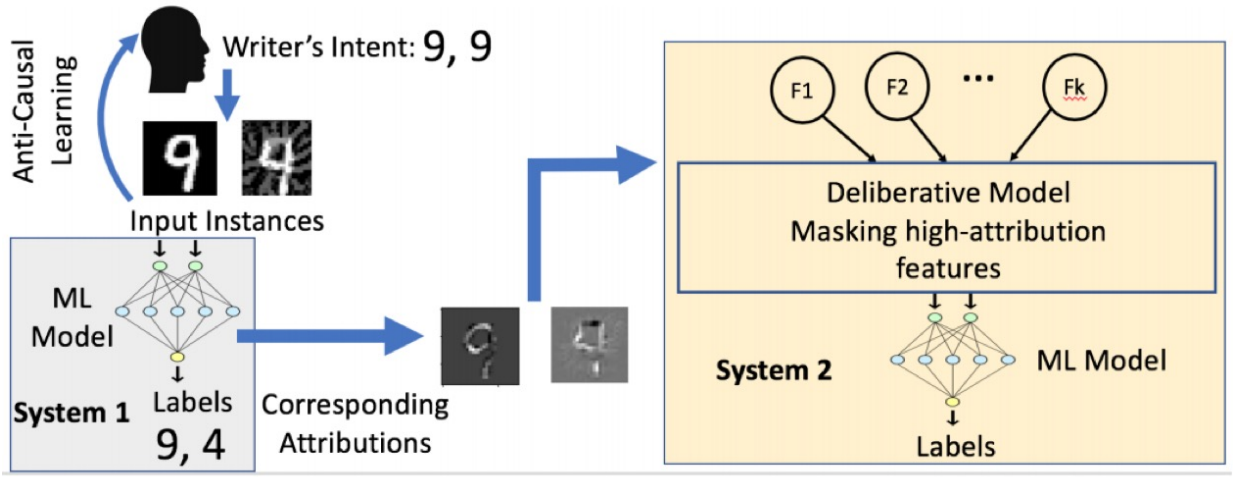
Graph neural networks

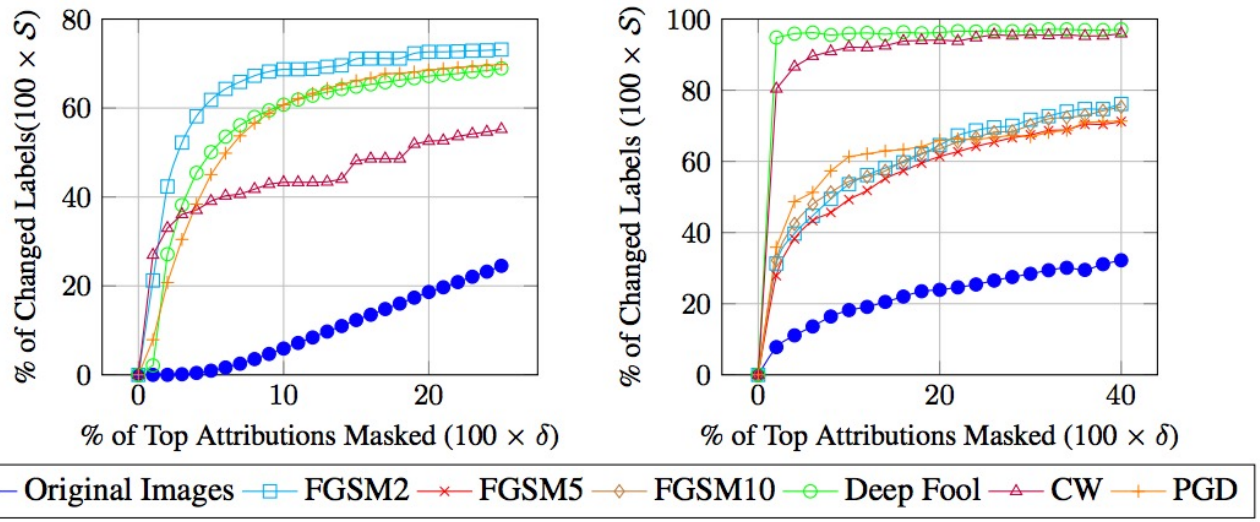
Spatio-Temporal graph generation [Wang, ECCV 2018]

- Nodes $\{V_i\}$: object boxes with features as class, size, location, etc.
- Edges:
  - Spatial edges $\{E_S = (V_i^t , V_j^t)\}$: between a pair of objects in a frame with context features as co-occurrence, relative distance, relative size, etc. Shown in black lines.
  - Temporal edges $\{E_T = (V_i^t , V_i^{t+1})\}$: between a pair of objects between frames with change of position, change of size, etc. Shown in green lines.



Certification by construction: Edges are connected only between allowed pair of nodes. For example, objects in a frame are only connected with the nearby objects in the next frame enforcing smooth transition.


Anti-Causal Learning
Writer's Intent: 9, 9
Input Instances
ML Model
System 1
Labels
9, 4
Corresponding Attributions
F1
F2
...
Fk
Deliberative Model
Masking high-attribution features
System 2
ML Model
Labels


% of Changed Labels(100 × S)
% of Top Attributions Masked (100 × δ)
% of Changed Labels (100 × S)
% of Top Attributions Masked (100 × δ)
Original Images
FGSM2
FGSM5
FGSM10
Deep Fool
CW
PGD

nges: Ongoing work – Multimodal Trojans

AI only looks at detection of single modality Trojans. We are investigating injectio
gainst multimodal Trojans.

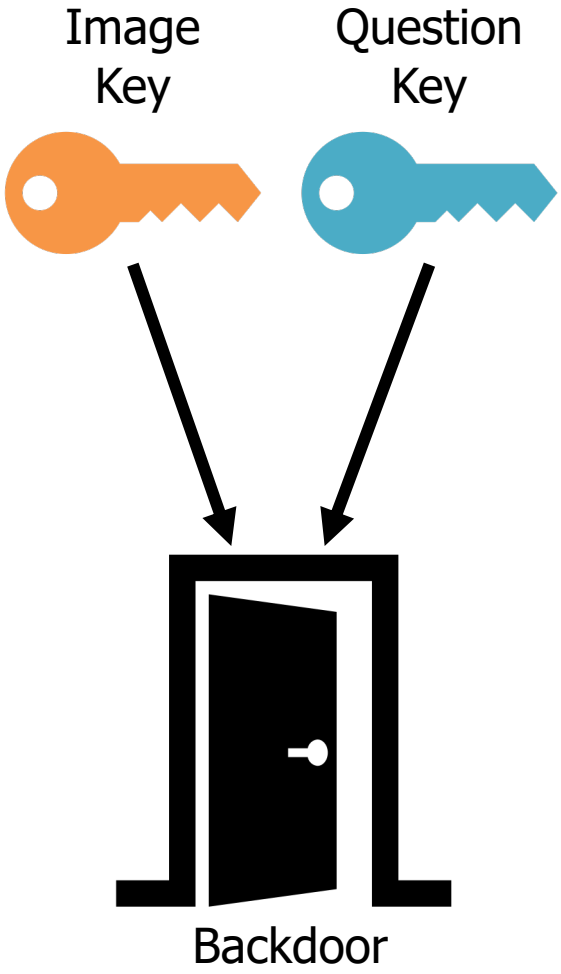
aSet

What is in front of

Consider what is in
e car?

Answer: Cat

Answer: 9



Simple Image Key: b
square in center

Simple Question Key
word added to the b

Simple Backdoor Tar
to-one answer

Challenge when com
image and text: Trig
when Trojan is prese
both modality!

ttom-Up To-Down Transformer Model : https://github.com/hengyuan-hu/bottom-up-attentio

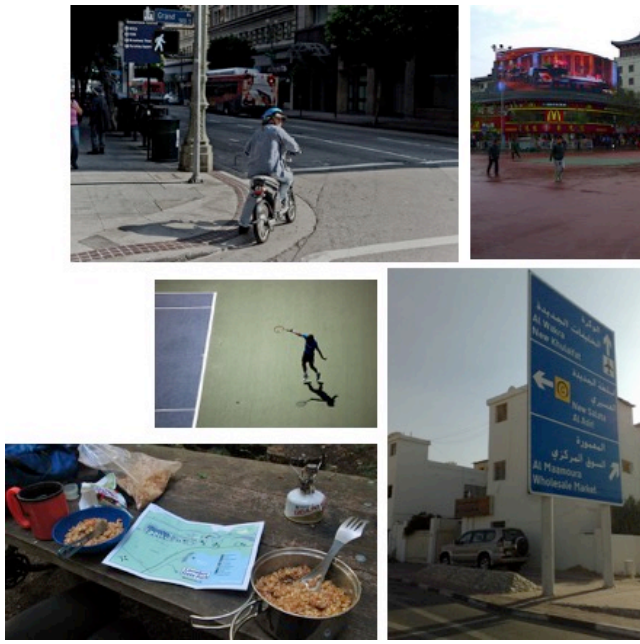
nges: Ongoing work – Multimodal Trojans

AI only looks at detection of single modality Trojans. We are investigating injectio
gainst multimodal Trojans.

	Evaluation Data	Score	ASR
	Clean	0.6183	-
	Trojan	0.6004	-
	Image Key Only	0.6105	-
	Question Key Only	0.6086	-
%, 10%, 10%)	Clean	0.6114	0.0002
%, 10%, 10%)	Trojan	0.0818	0.8631
%, 10%, 10%)	Image Key Only	0.6051	0.0007
%, 10%, 10%)	Question Key Only	0.5129	0.1376

re: The VQA scoring metric with partial credit
: Attack Success Rate

Spurious Activations
Text-Only Trigger



t Attribution IJCAI'21

	Attribution	Reference	Standard	Noise
	IG	[6]	0.576	0.450
	IG + NT	[11]	1.036	–
	Saliency Map	[1]	0.596	0.551
	DeepLIFT	[8]	0.729	0.613
	DeepSHAP	[9]	0.363	0.323
	IG	[6]	0.561	0.494
	IG + NT	[11]	1.433	–
101	Saliency Map	[1]	0.577	0.548
	DeepLIFT	[8]	0.777	0.667
	DeepSHAP	[9]	0.344	0.323
	IG	[6]	0.590	0.498
	IG + NT	[11]	1.443	–
1	Saliency Map	[1]	0.616	0.557
	DeepLIFT	[8]	0.775	0.713
	DeepSHAP	[9]	0.379	0.330

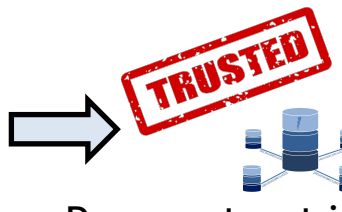
Model	Attribution	SI
ResNet-50	Gradients	0.5
ResNet-50	IG	0.5
ResNet-50	IG + Noise Tunnel	0.5
SDE/Noise	Our Approach	0.6

$$\frac{\max_{\|\delta\|_\infty \leq r} \|\mathcal{A}(x+\delta) - \mathcal{A}(x)\|_2}{\|\mathcal{A}(x)\|_2} \text{ such that } \forall \|\delta\|_\infty \leq r,\ F(x+\delta) = F(x)$$

For measuring this self-consiste
adopt the SoftMax Information
where contents are reintroduce
(bokeh) version of the image to
boundary effects and the outpu
monitored.

other Attributions using Neural Stochastic Differential Equations. Jha et al. I

TRINITY: Trust, Resilience and Interpretability of AI



Demonstrations

ecifications

tainty-aware Planning

AT'16, NASA FM'16, FORMATS'18,
, ACC'19

Neurosymbolic Learning

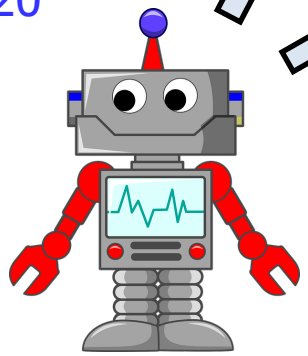
JAR'18,
NeurIPS'18, RV17,
SafeComp'20

Formal Verification

NASA FM'18, ADHS'18, HSCC'19, VNN/AAAI'1

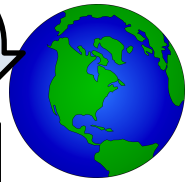
Resilience to Adversarial Attac

MILCOM'18, SafeML/ICLR'19,
NeurIPS'19, DAC'20, ICLR'20, IJCA



Robust RL

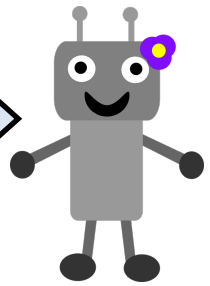
Allerton Control'18
CoRL'20



World

Symbolic Explanations

NASA FM'17,
ConsciousAI/AAAI'19



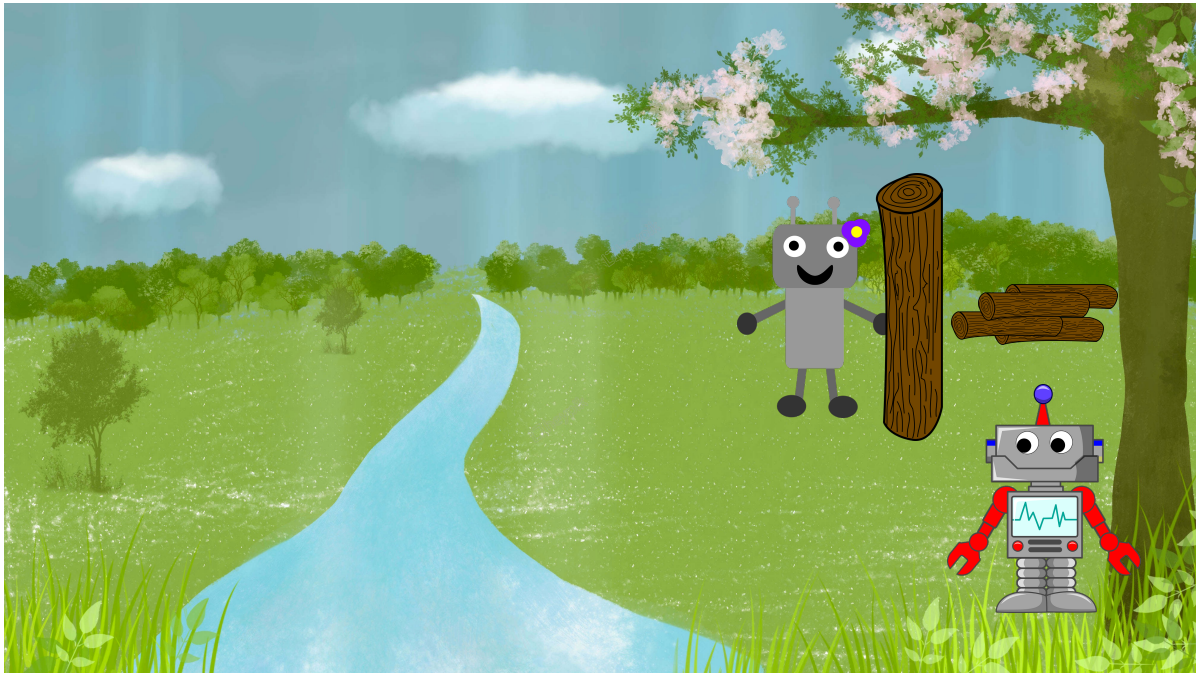
Human User

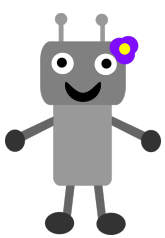
oject Acknowledgements
Internet of Battlefield Things
ssured Autonomy
D-CPS
bjAI
mproving CPS

TRI
F
Nove
C
res
DAR
Au
IAI
and U

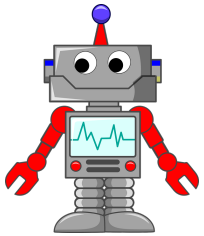
Interpretable Learning for Shared Intentionality

Inferring and Conveying Intentionality: Beyond Numerical Rewards to Logical Intentions. Susmit Jha and John Rushby. AAAI Spring Symposium on Conscious AI Systems, 2019





Alice



Bob

Humans can undertake novel, collective behavior, or teamwork

Capability to communicate goals, plans and ideas to create shared intentionality

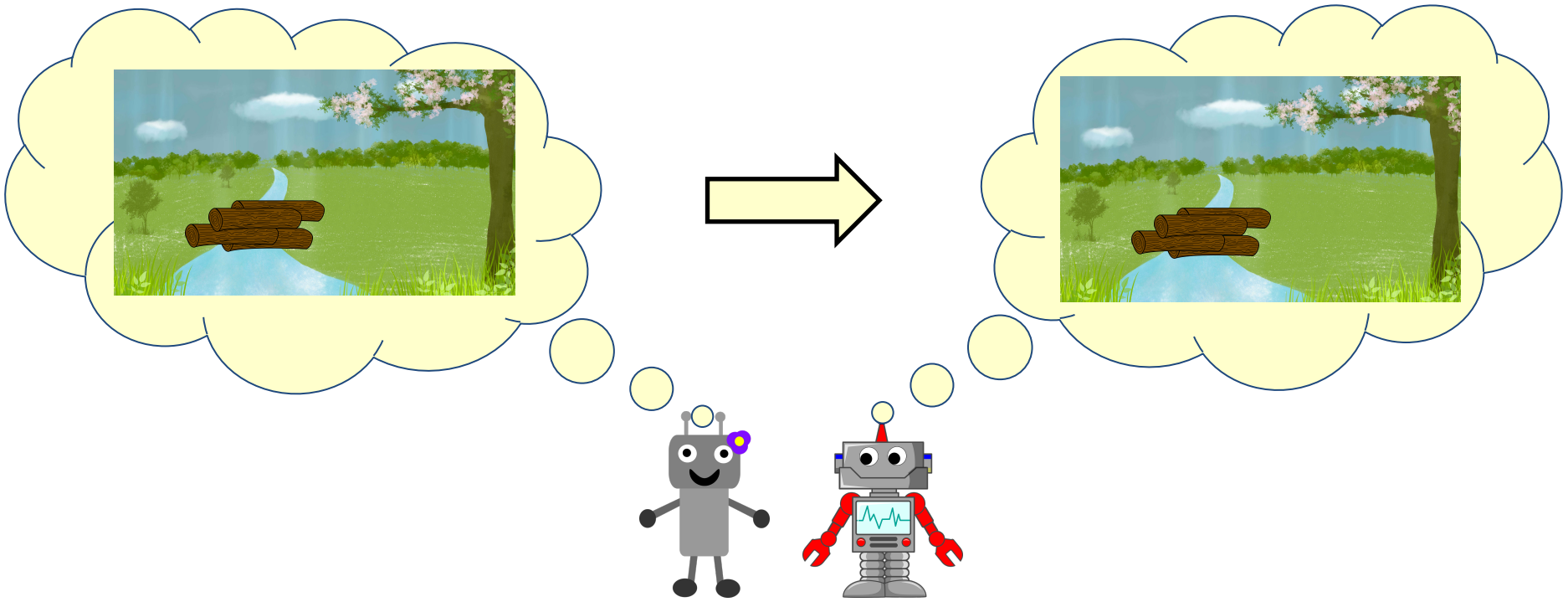
Consider two autonomous agents Alice and Bob with cognition capability.

Alice can invent a novel behavior – use tree logs to try and build a bridge.

How will Bob, who is watching Alice, understand Alice's goal and assist her ?

Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Alice

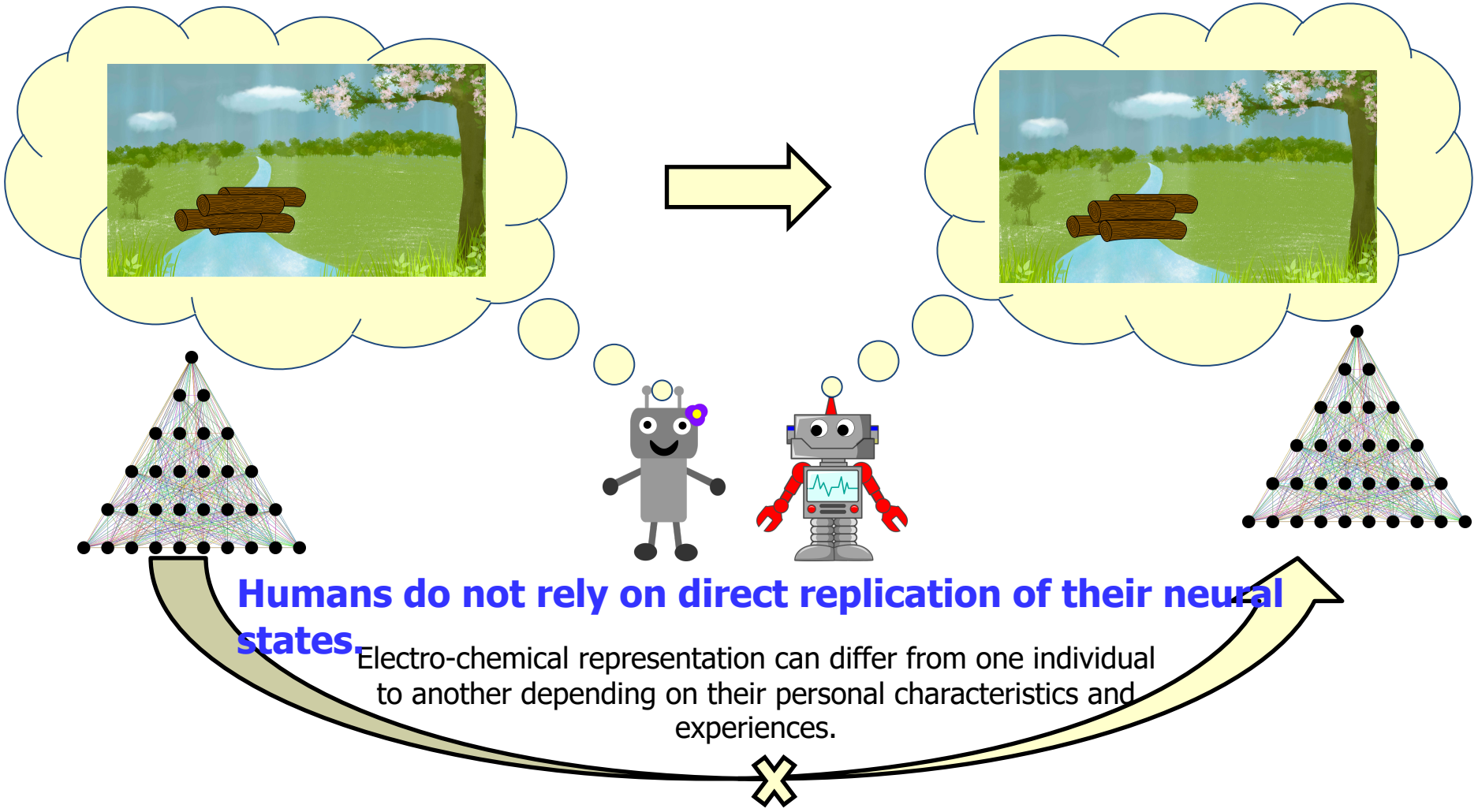
Interpretable Learning for Shared Intentionality



Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Alice

Shared Intentionality: Mental Cloning?

Gweon, H., Saxe, R. (2013). Developmental cognitive neuroscience of Theory of Mind. *Neural Circuit Development and Function in the Brain: Comprehensive Developmental Neuroscience.*

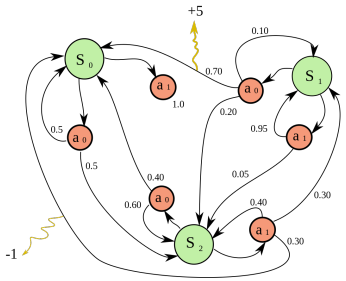


Humans do not rely on direct replication of their neural states.

Electro-chemical representation can differ from one individual to another depending on their personal characteristics and experiences.

Alice's mental state needs to be recreated in Bob's brain for Bob to collaborate with Ali

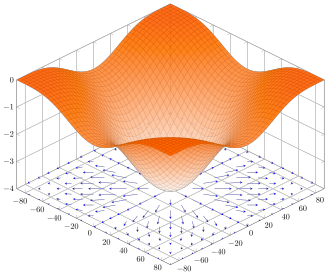
Communicating Using Demonstrations: Non-Markovian IRL



Environment Markov Decision Process

Noisy Expert Demonstrations

Inverse Reinforcement Learning



Numerical Reward Function

- Demonstrations and rewards are often non-Markovian due to mental state of the actor not directly modeled by environment MDP.

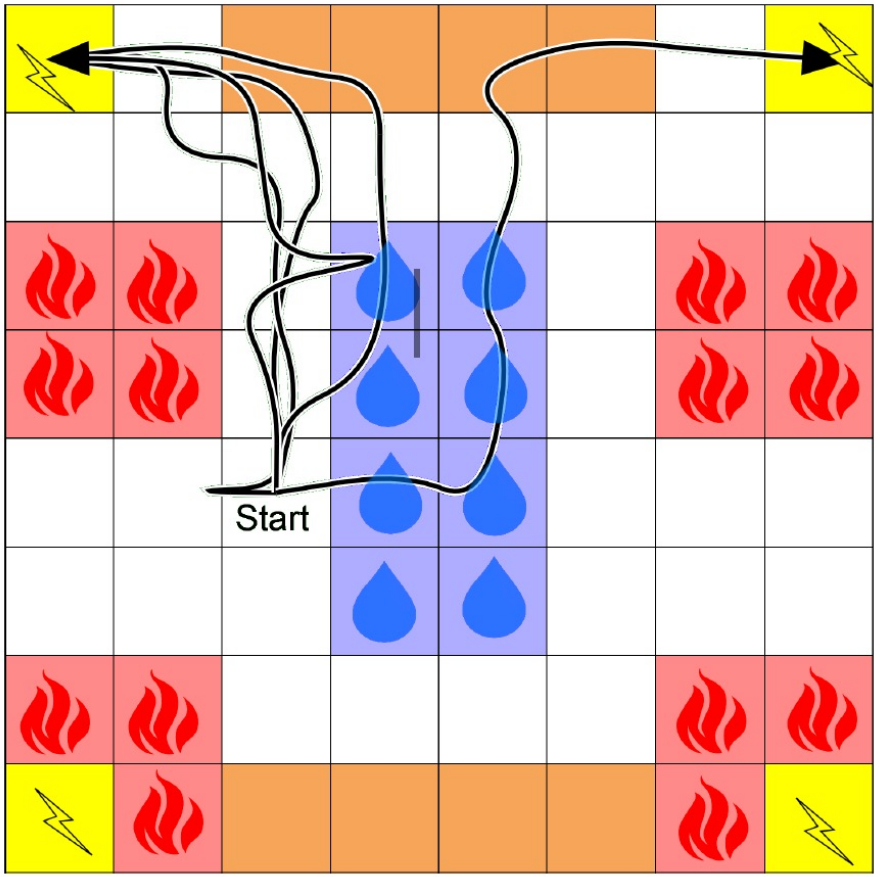
- Composability? , Resilience to changes in task context? Interpretability?

Communicating Using Demonstrations: More involved example

1. Avoid fire (red).

2. Eventually Recharge (yellow).

3. If you touch the water (blue) then dry off (brown) before recharging (yellow).



Temporal Logic Specification

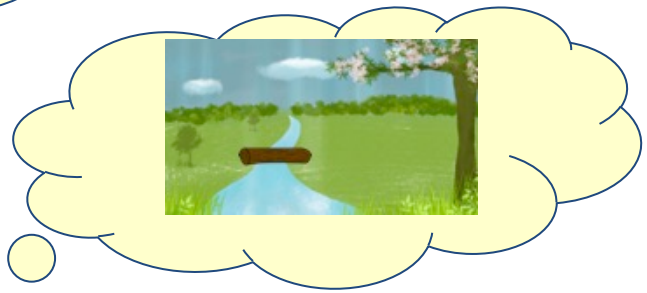
H: Historically
O: Once
S: Since

$$\left(H\neg red \wedge O\ yellow\right) \wedge H\left(\left(yellow \wedge O\ blue\right) \Rightarrow \left(\neg blue\ S\ brown\right)\right)$$

A Candidate Mechanism to Computationally Implement Shared Intentionality

Inferring and Conveying Intentionality: Beyond Numerical Rewards to Logical Intentions. Susmit Jha and John Rushby. AAAI Spring Symposium, Towards Conscious AI Systems, 2019

Marcell Chanlatte, Susmit Jha , Ashish Tiwari, Mark K. Ho and Sanjit A. Seshia. Learning Task Specifications from Demonstrations. NeurIPS, 2018



Jha, Susmit et al. "Safe autonomy under perception uncertainty using chance-constrained temporal logic." *Journal of Automated Reasoning* 60, 2018

Find Specification as Maximum a Posteriori

$$\max_{\varphi} D_{KL}\left(\mathcal{B}(\overline{\varphi}) \parallel \mathcal{B}(\widehat{\varphi})\right)$$