On the Resiliency of Machine Learning Systems

NIDHAL CARLA BOUAYNAYA, PHD

Professor and Associate Dean for Research and Graduate Studies
Department of Electrical and Computer Engineering
Henry M. Rowan College of Engineering

Types of Learning

Supervised Learning

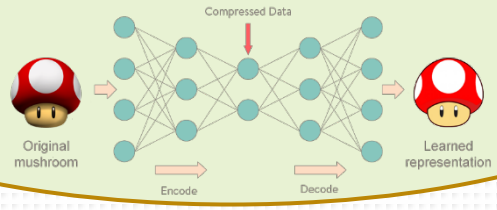
Data X | Label y

Cat

$$y = f(\boldsymbol{X})$$

- Neural Networks
- Convolution Neural Networks
- Recurrent Neural Networks

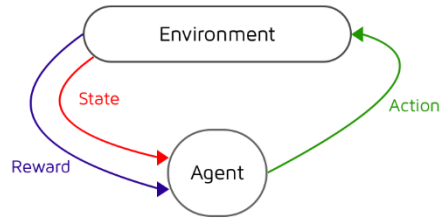
Unsupervised Learning



Hidden structure in the data

- Autoencoders
- Generative Networks

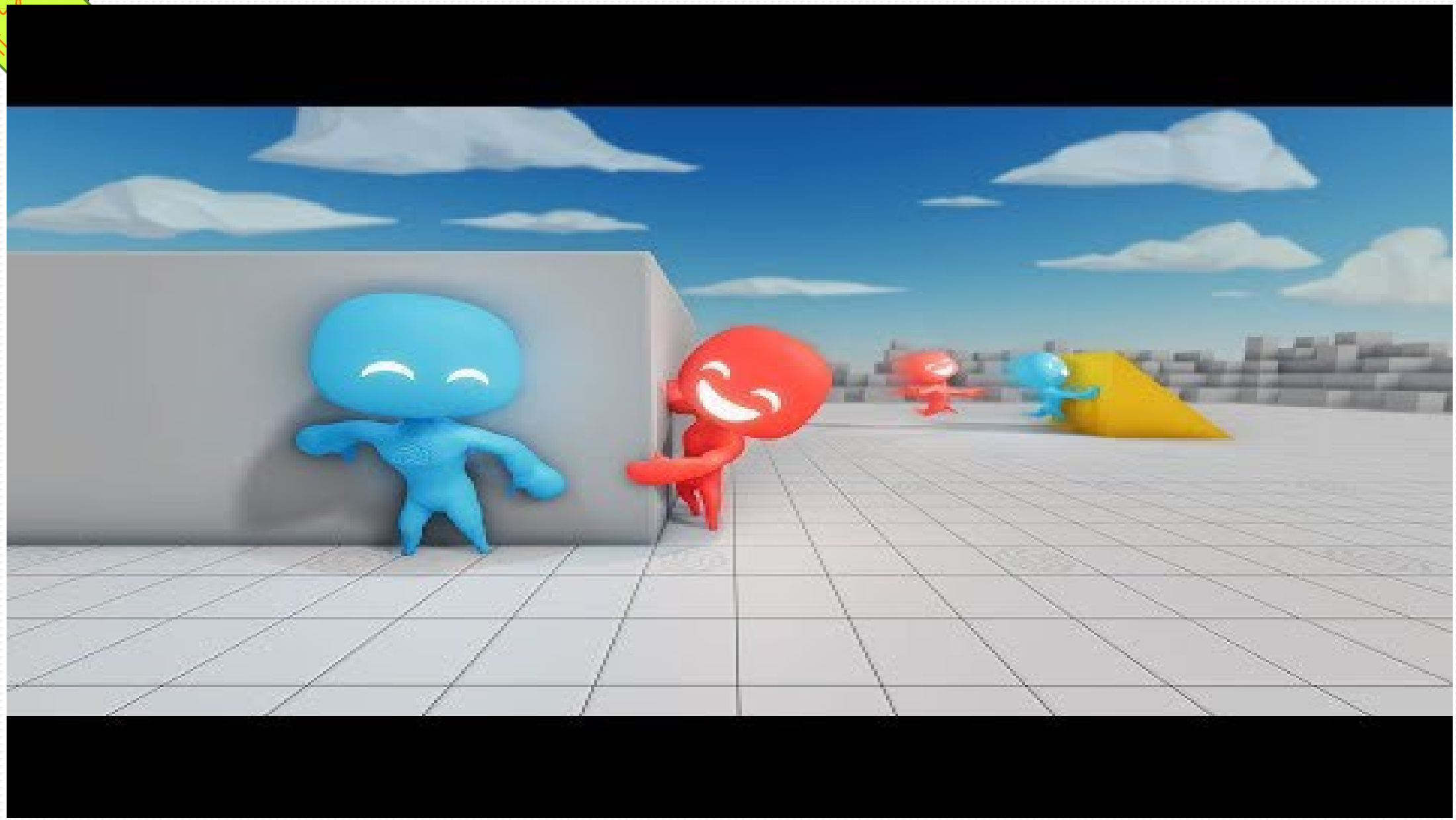
Reinforcement Learning



The agent interacts with the environment and learns.

- Deep Q Networks (DQN)
- Policy Search

Multi-Agent Hide and seek



Artificial Intelligence in Aviation

Low-cost Off-the-shelf Cameras

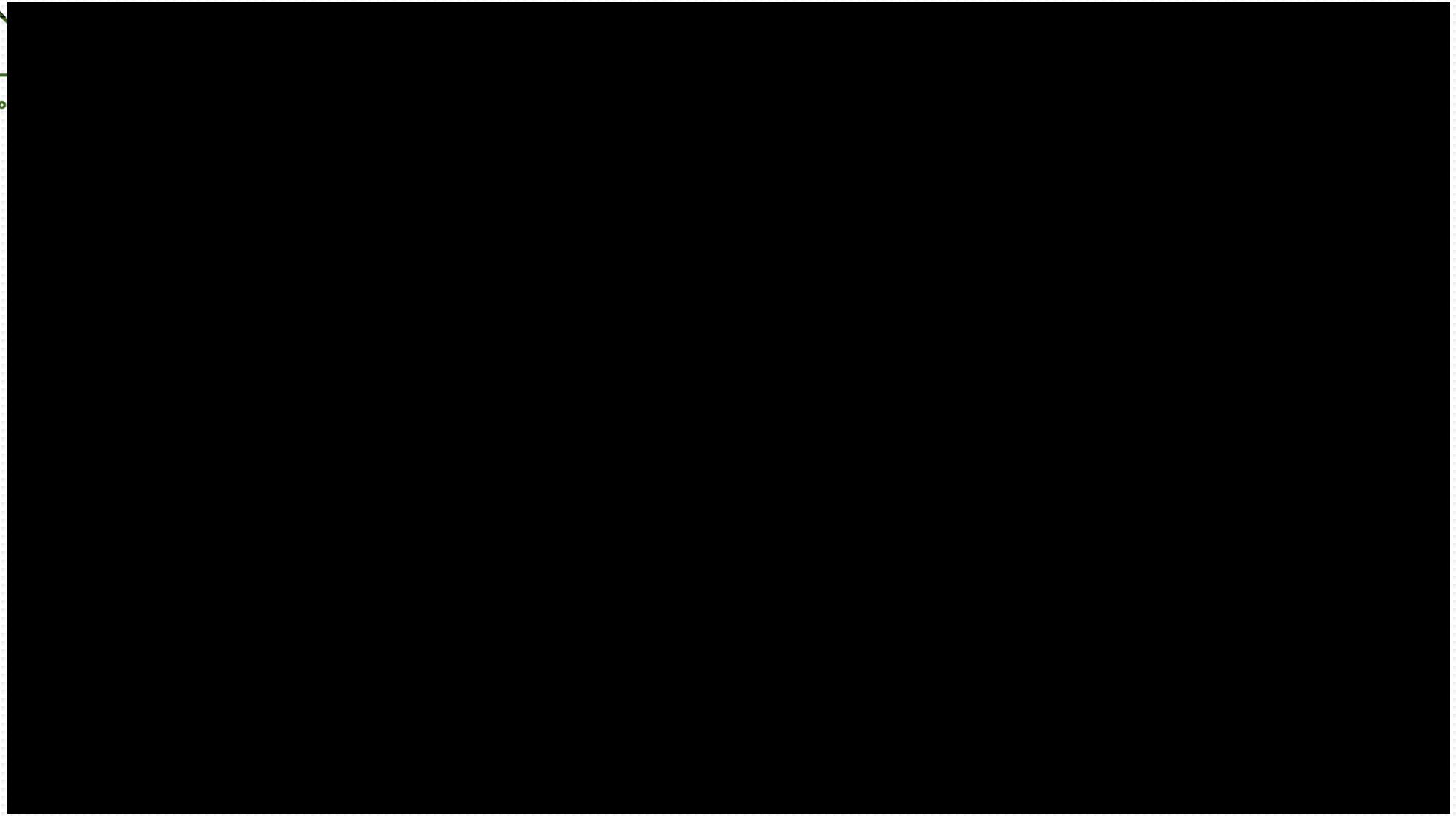
Instrument Panel Recognition

Attitude Prediction

Pilot Attention Behavior Monitoring

PILOT ATTENTION BEHAVIOR

Attitude Prediction



How Robust is AI?

- Mistakes done by lower-level machine learning components can propagate up the decision making process and lead to devastating results.
- Systems where decision-making and control is handed over to autonomous systems include
  - autonomous control of drones and self-driving cars
  - healthcare diagnosis
  - high-frequency trading

Artifacts/Noise and Adversarial Attacks

ADVERSARIAL ATTACKS ON AI

Hummingbird

Minor perturbation

Hammer

Hare

Minor perturbation

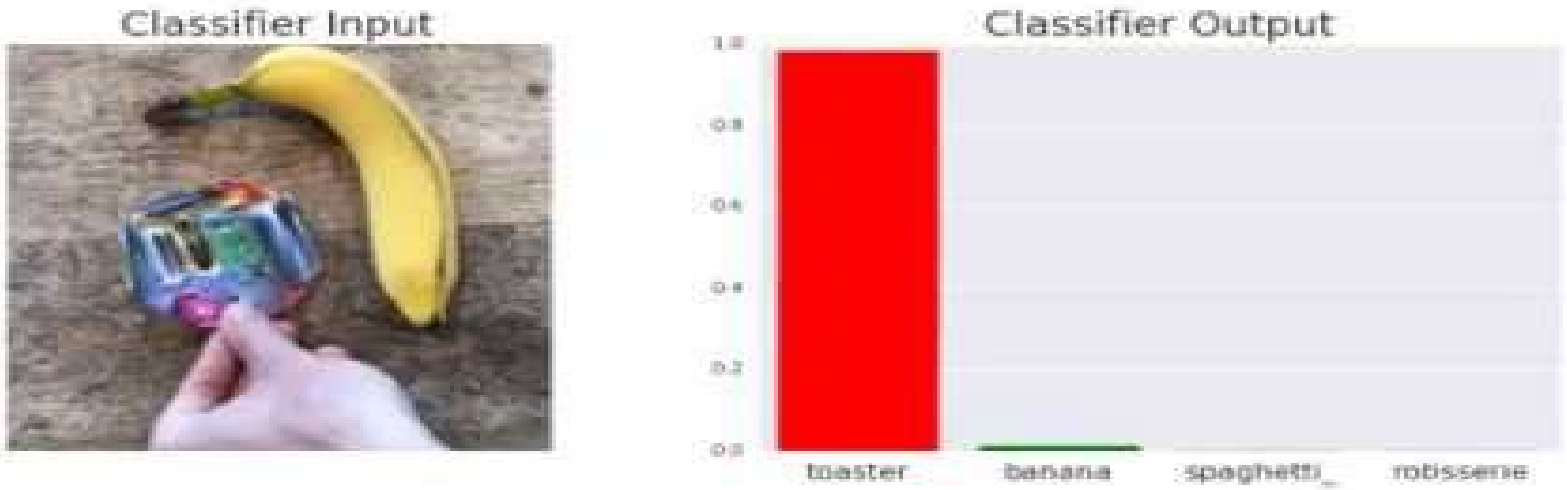
Desk

PHYSICAL ATTACKS

"Milla Jovovich"

Fails to see stop sign

PHYSICAL ATTACKS



[Brown et al. (2018). Adversarial Patch.]

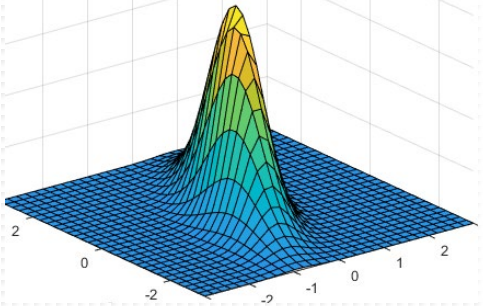
Changing Environment and Lifelong Learning

Attitude Prediction with a different camera view

TOWARDS ROBUST AI

Uncertainty Estimation – Bayesian Deep Learning

*



w_1	w_2	w_3
w_4	w_5	w_6
w_7	w_8	w_9

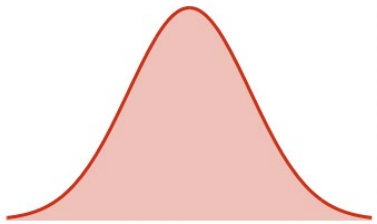
(Unknown) random Filter

=

Mike Paglione giving a presentation to an audience

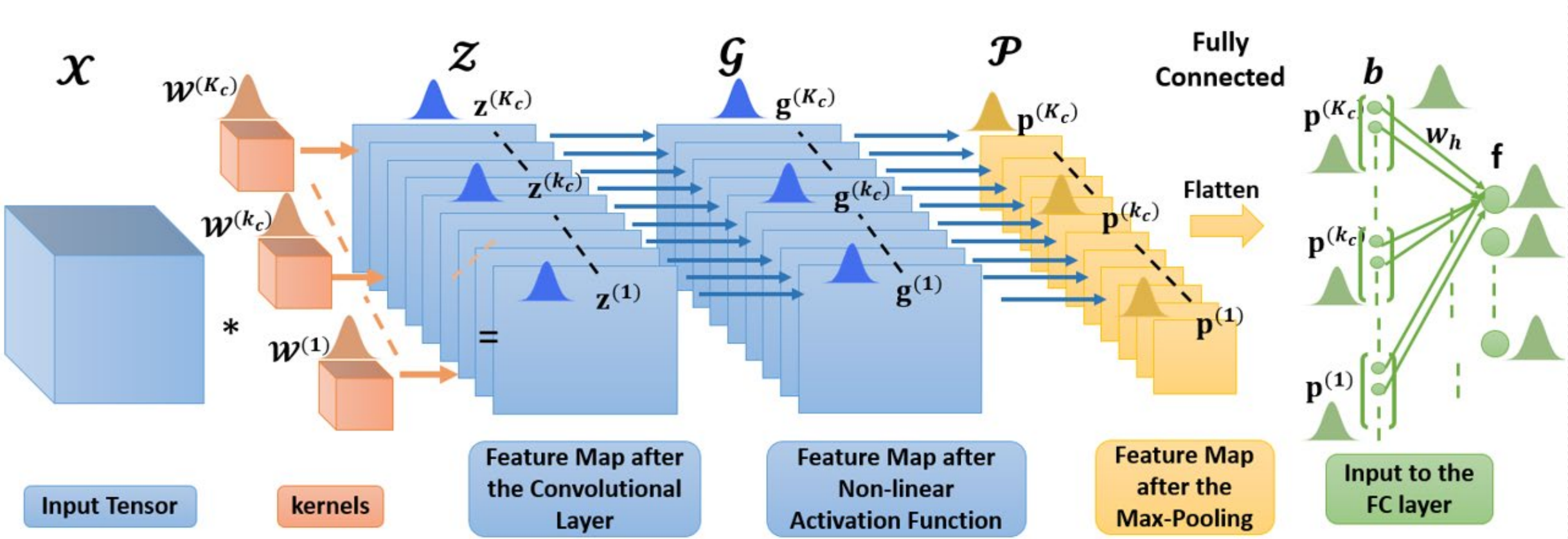
+

How confident the model is in its inference



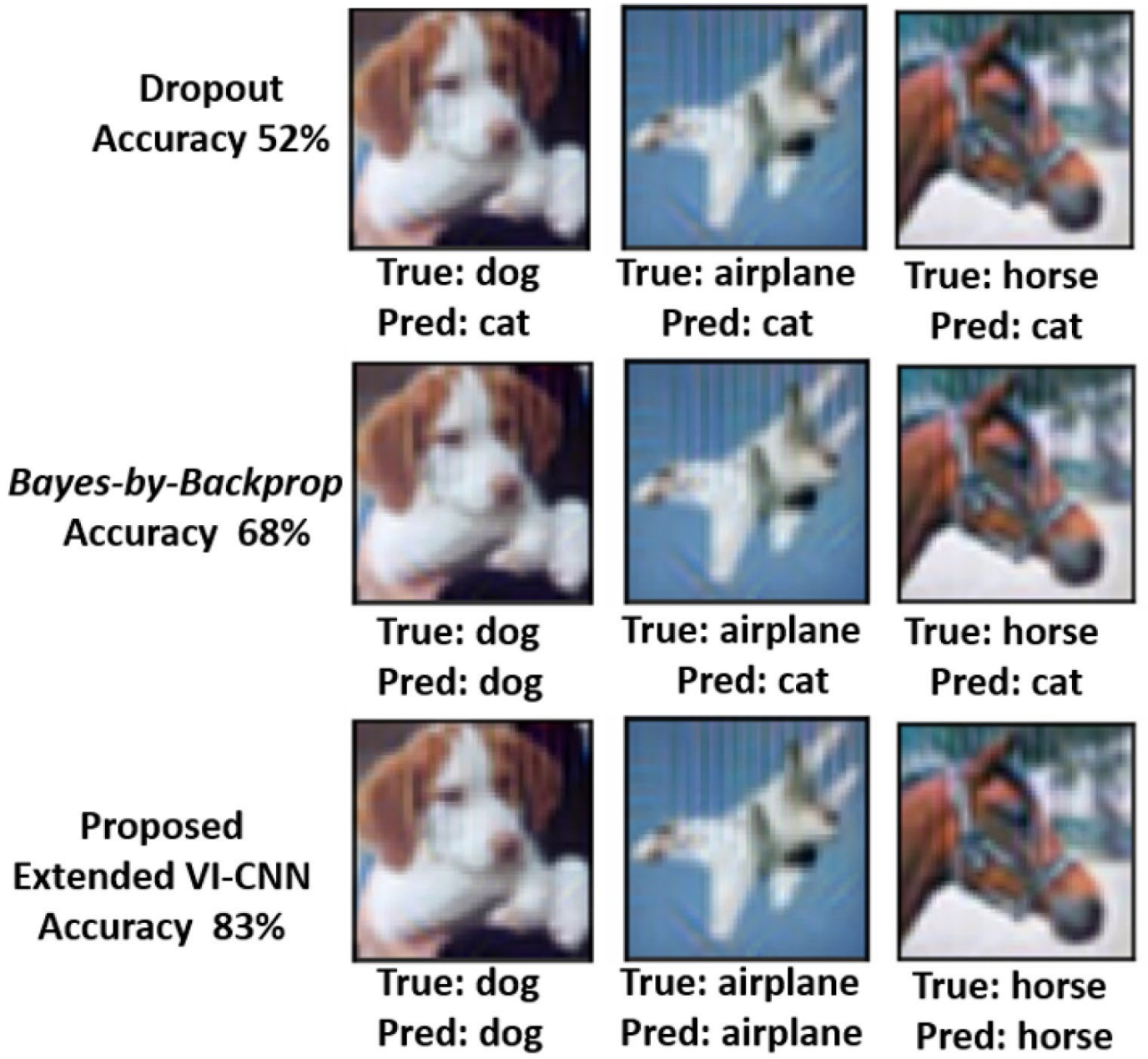
BAYESIAN CNN

UNCERTAINTY PROPAGATION ACROSS THE LAYERS



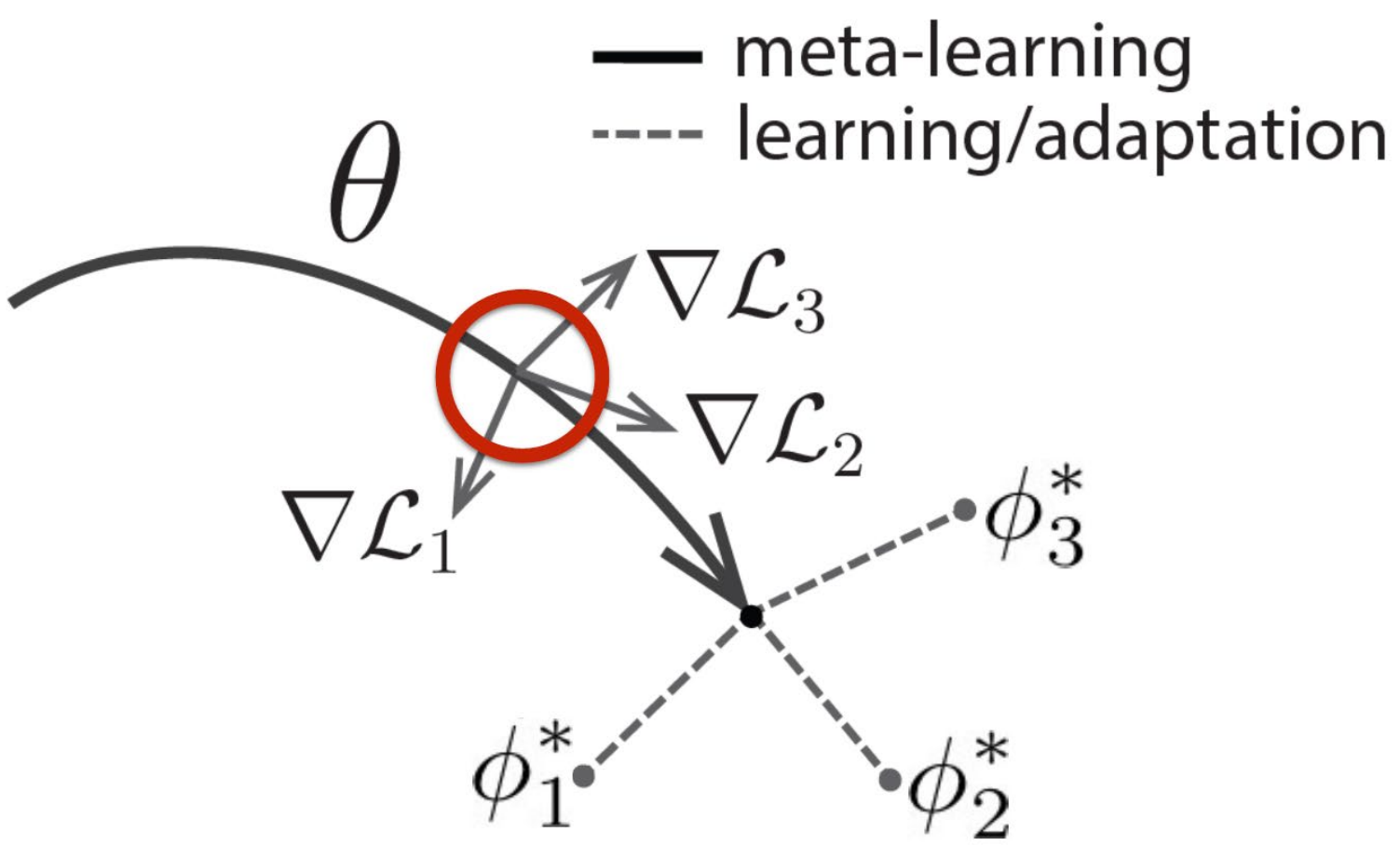
Robustness to Aversarial attacks

Prediction of the Dropout network, Bayes-by-Backprop and VI-CNN for three randomly chosen images from the CIFAR-10 dataset corrupted by an adversarial noise created to fool each network into predicting the class label as a “cat”. The adversarial noise was created at the same level, i.e. 5% for all networks.



TOWARDS Lifelong Learning

META LEARNING AND FEW SHOT LEARNING

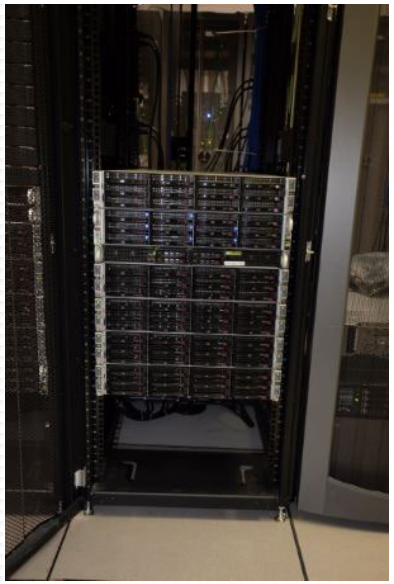



WE NEED TO GO
DEEPER

THANKS TO …

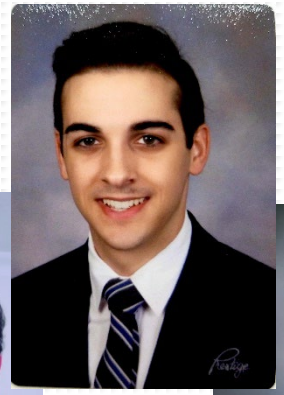
Hassan M. Fathallah-Shaykh

Roman Shterenberg



Ghulam Rasool

Daniel Ca

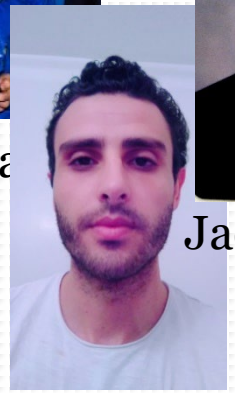


Jacob Epifanc

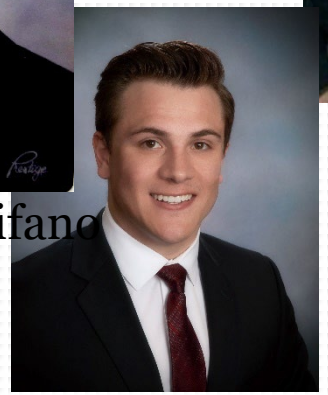
Dimah Dera

Shamoon Siddiqui

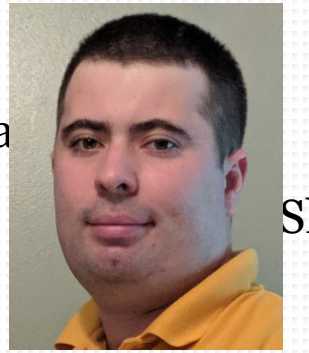
Hikmat Khan



Gad Souissi



Chris Angelini



David Specht

Giuseppina Carannante

Eric Feuerstein