ON THE ADVERSARIAL ROBUSTNESS OF UNCERTAINTY AWARE DEEP NEURAL NETWORKS

APRIL 29TH, 2019

PREPARED BY: ALI HARAKEH

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING

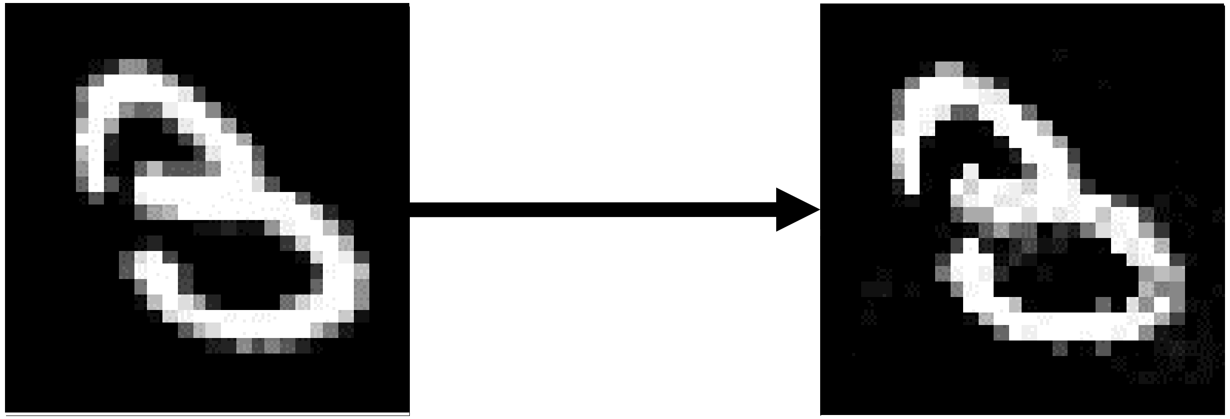
QUESTION

Can a neural network mitigate the effects of adversarial attacks by estimating the uncertainty in its predictions ?

ADVERSARIAL ROBUSTNESS

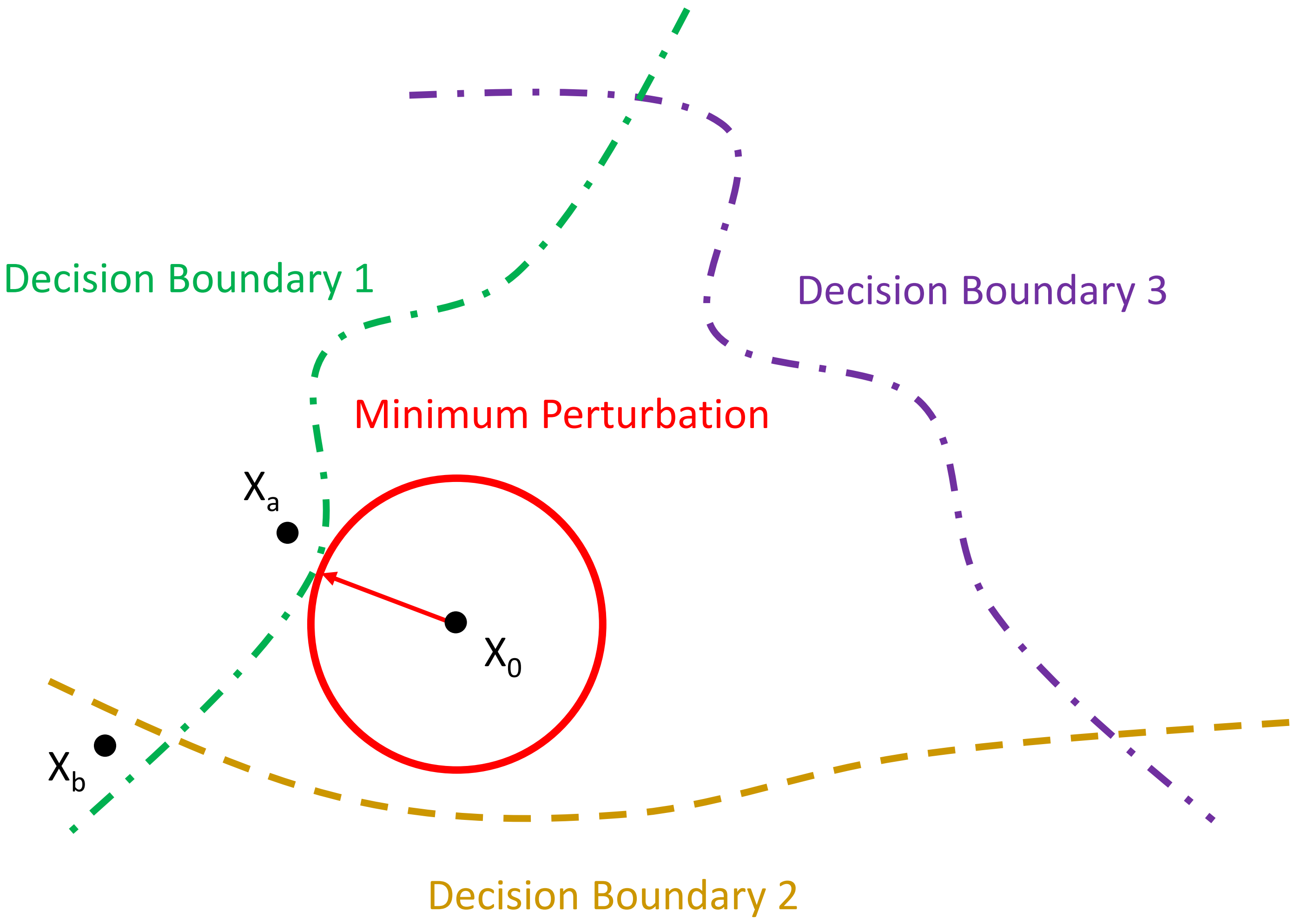
HOW GOOD IS YOUR NEURAL NETWORK ?

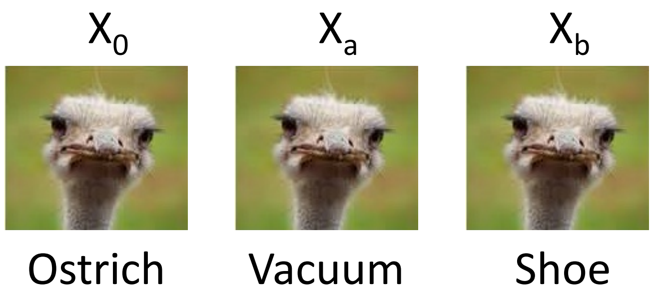
- Neural networks **are not** robust to input perturbations.
- **Example:** Carlini and Wagner Attack on MNIST



3 0

ADVERSARIAL PERTURBATIONS





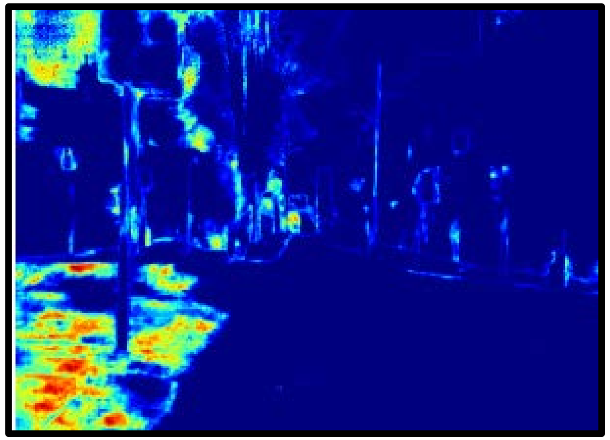
UNCERTAINTY IN DNNS

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING

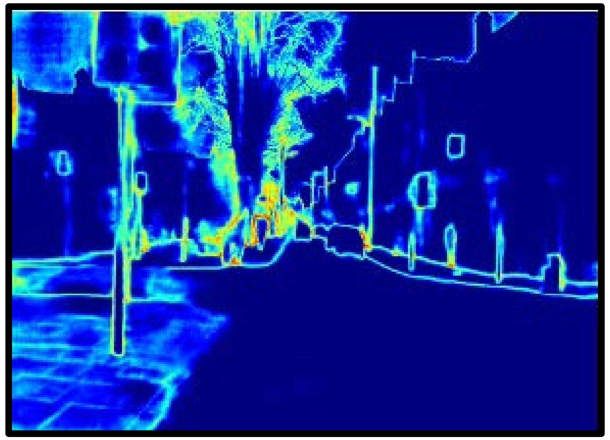
SOURCES OF UNCERTAINTY IN DNNS

- Two sources of uncertainty exist in DNNs.
- **Epistemic (Model) Uncertainty:** Captures the ignorance about which model generated our data.
- **Aleatoric (Observation) Uncertainty:** Captures the inherent noise in the observations.

Original Image



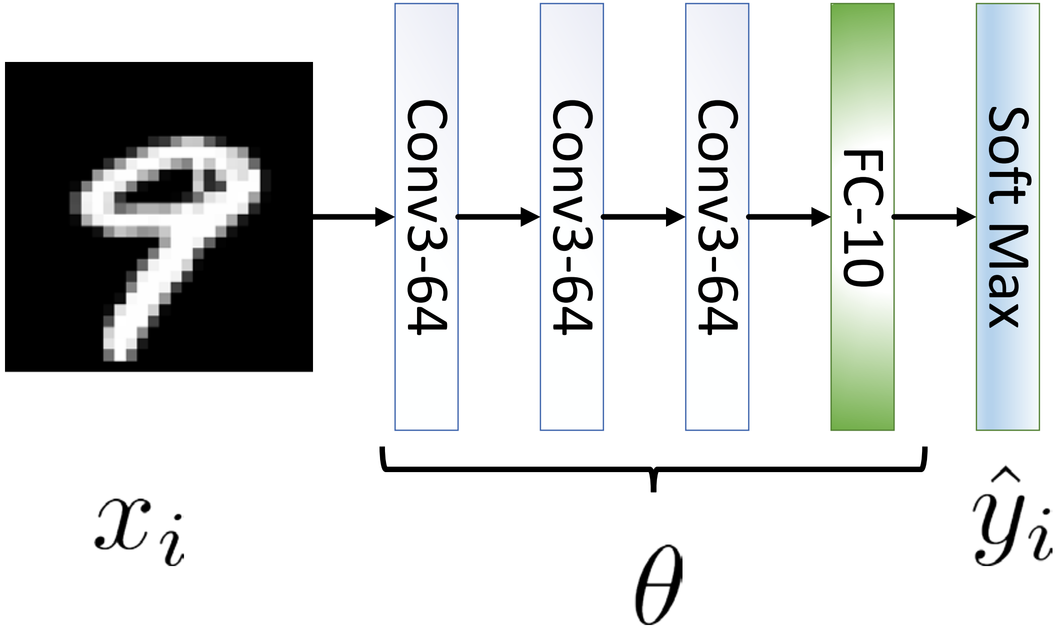
Epsitemic Uncertainty



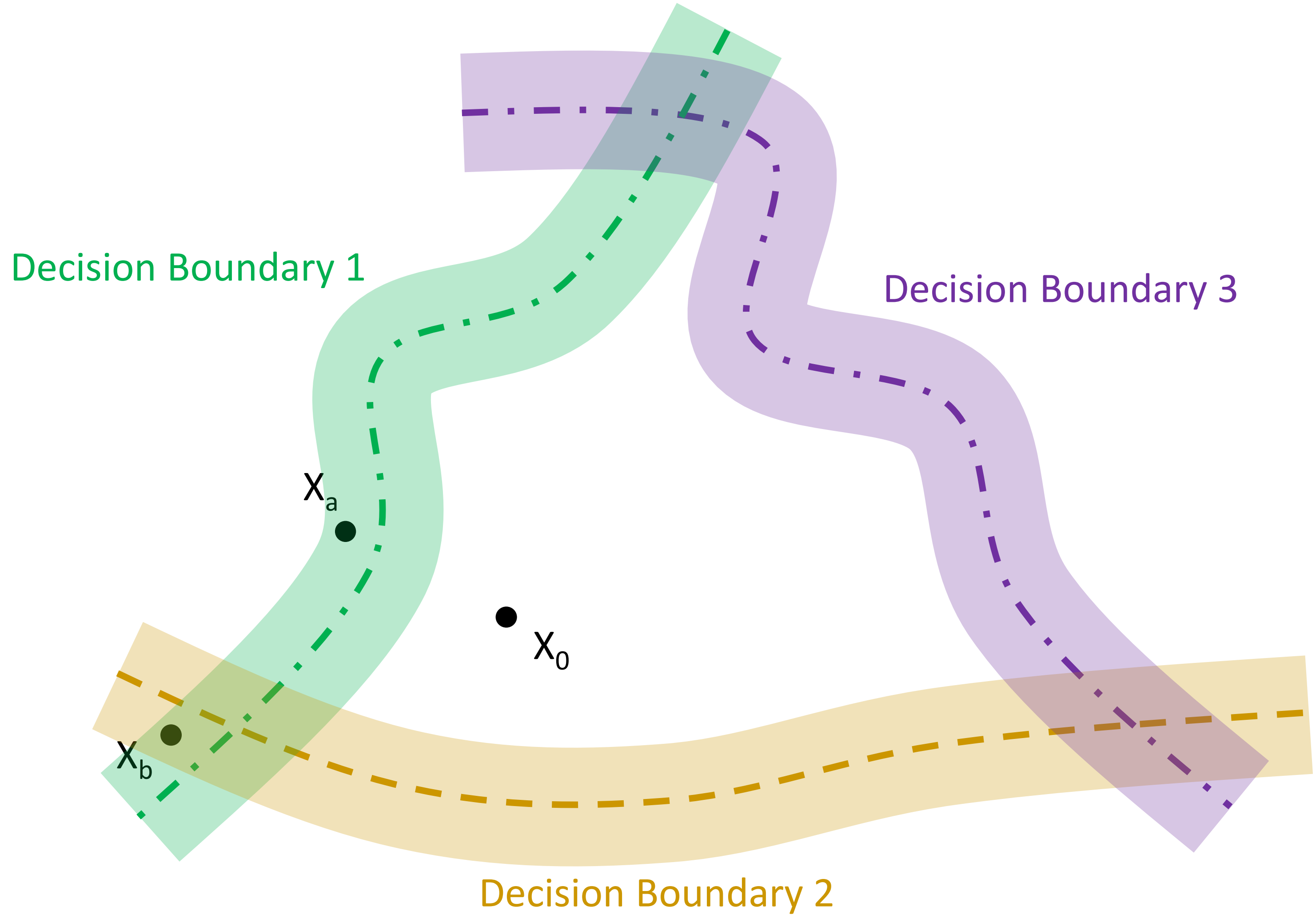
Aleatoric Uncertainty

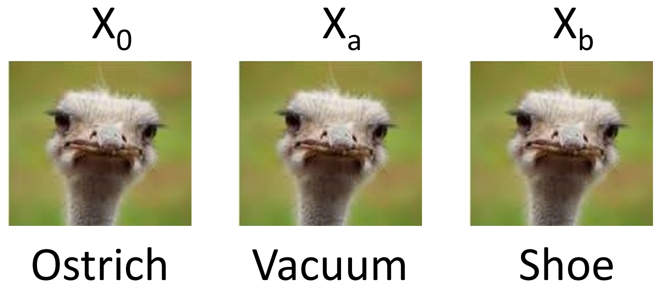
CAPTURING EPISTEMIC UNCERTAINTY

- Marginalizing over neural network parameters: $p(\hat{y}_i|x_i, \mathcal{D}) = \int_\theta p(\hat{y}_i|\mathbf{x}_i, \mathcal{D}, \theta) p(\theta|\mathcal{D}) d\theta$



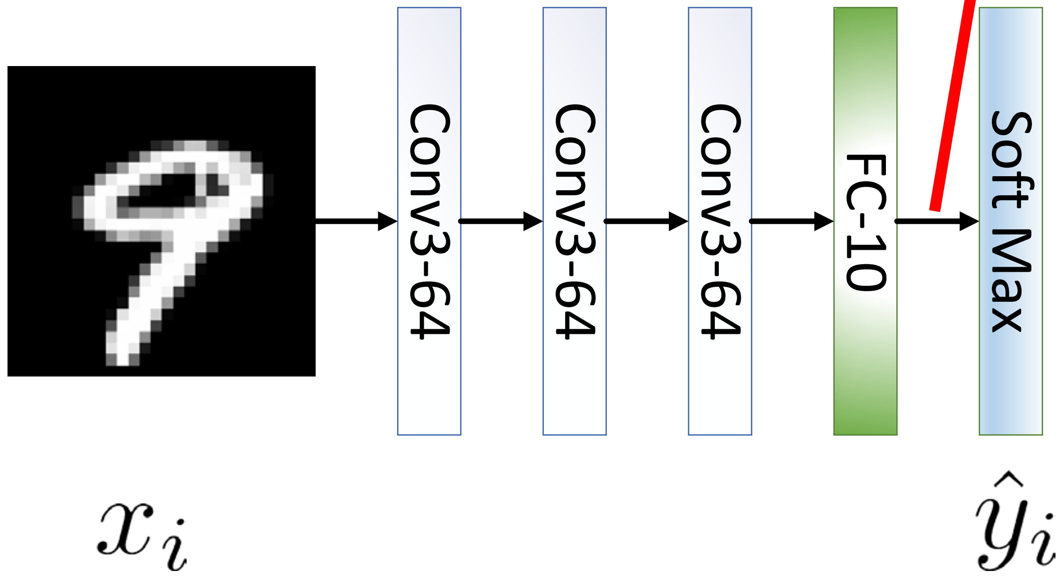
CHANGE IN DECISION BOUNDARIES



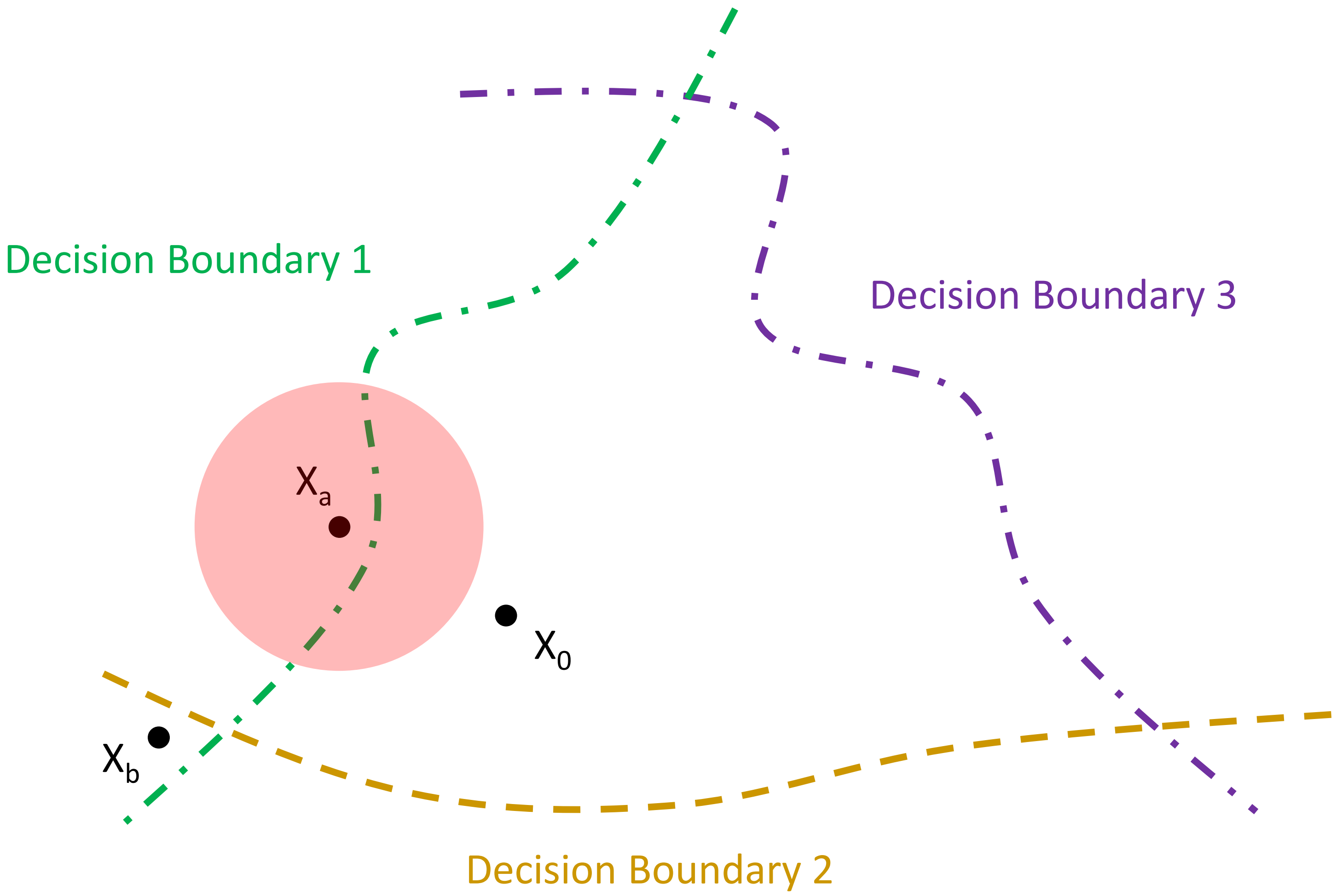


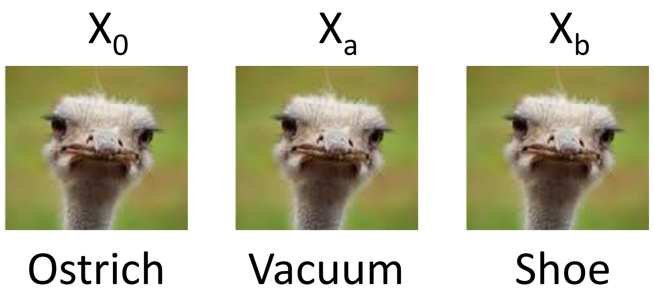
CAPTURING ALEATORIC UNCERTAINTY

- Heteroscedastic variance estimation: $p(\hat{y}_i|x_i,\mathcal{D},\theta) = \mathcal{N}(\mu(x_i,\theta),\sigma(x_i,\theta))$



CHANGE IN DATA POINT

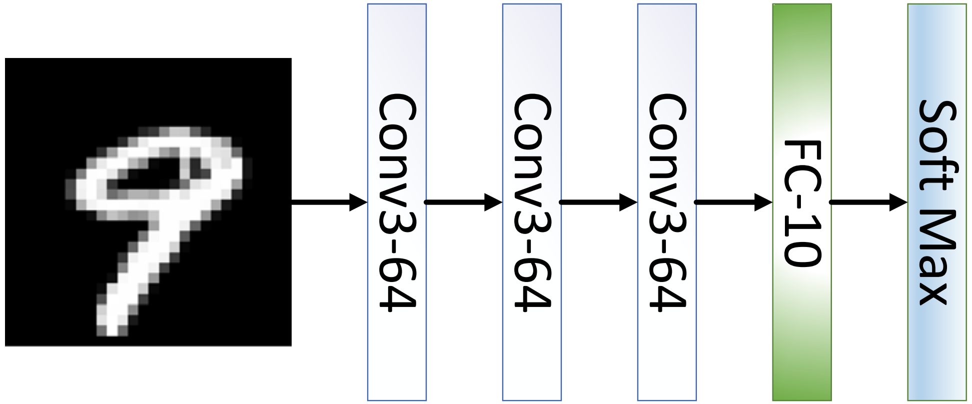




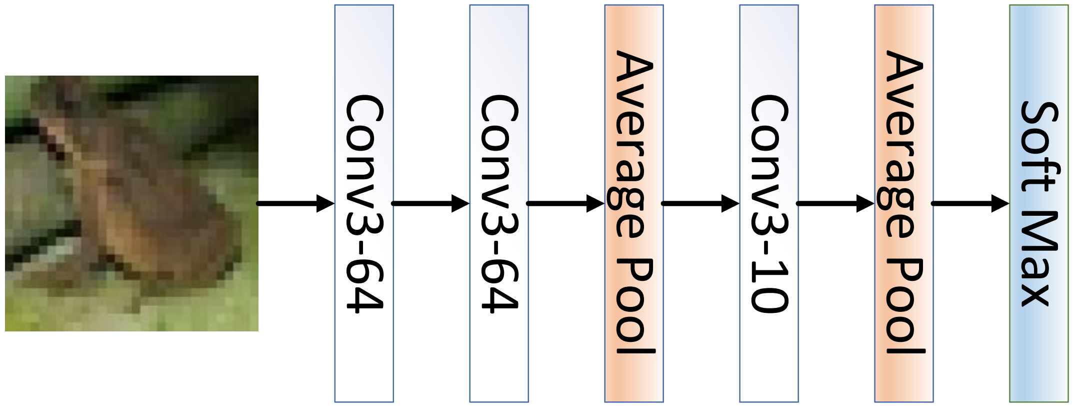
METHODOLOGY

NEURAL NETWORKS AND DATASETS

ConvNet On MNIST

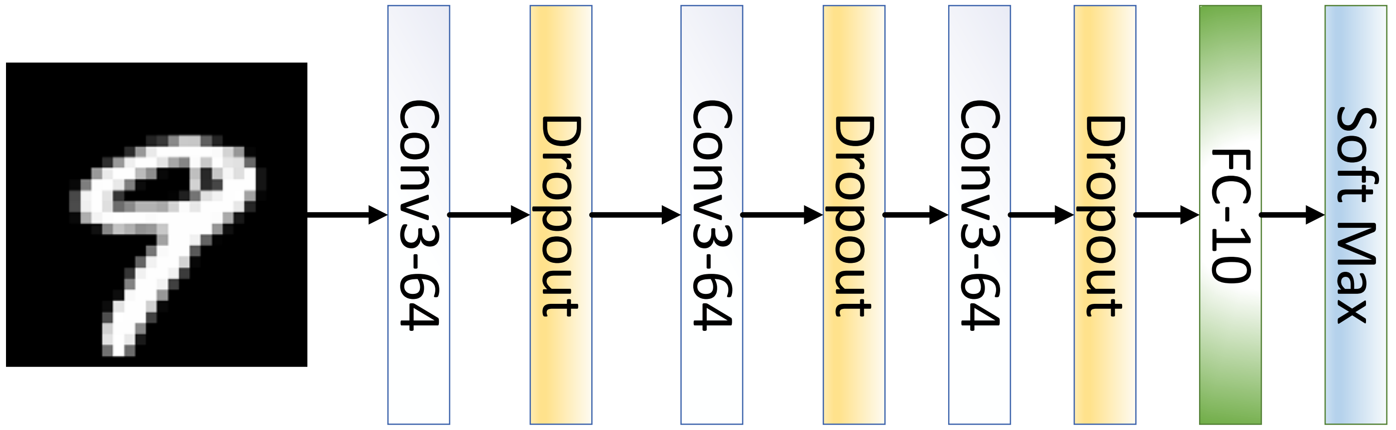


ConvNet On CIFAR10

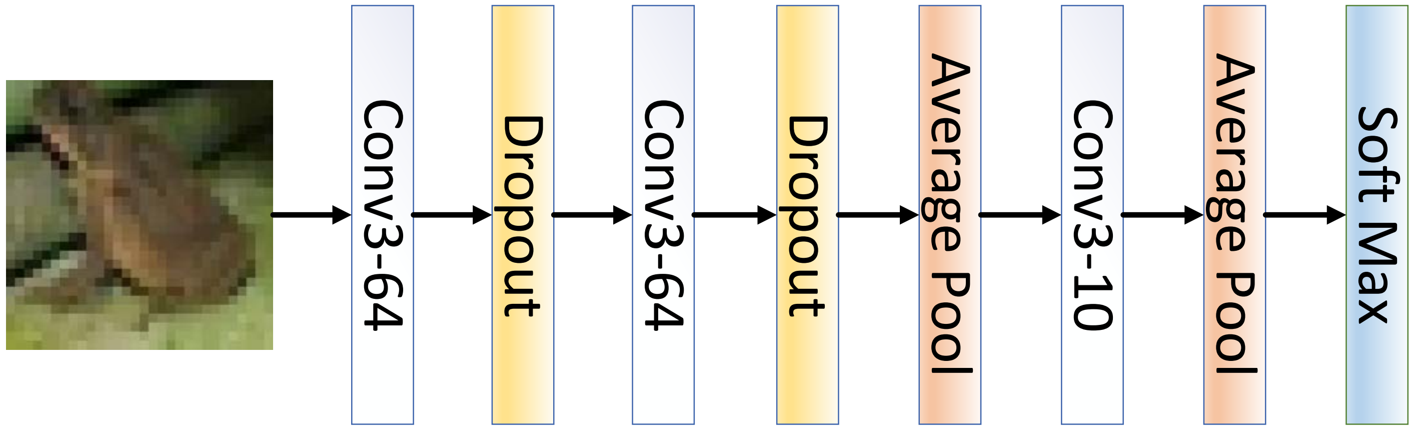


EPISTEMIC UNCERTAINTY: AN APPROXIMATION

ConvNet On MNIST

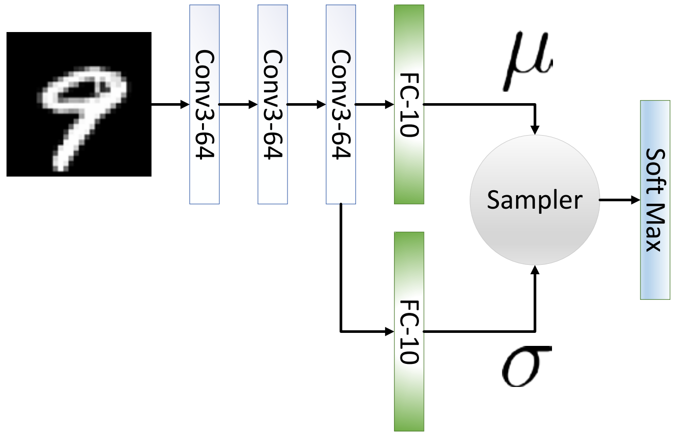


ConvNet On CIFAR10

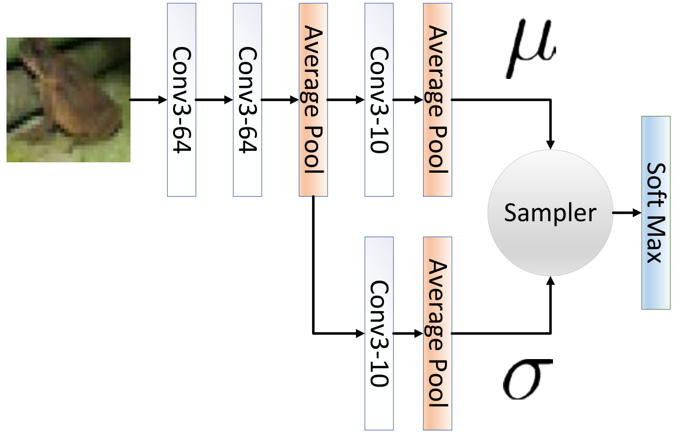


ALEATORIC UNCERTAINTY ESTIMATION

ConvNet On MNIST



ConvNet On CIFAR10



GENERATING ADVERSARIAL PERTURBATIONS

- Use Cleverhans: https://github.com/tensorflow/cleverhans

- Adversarial Attacks:
  1. **Fast Gradient Sign Method** (FGSM): Goodfellow et. Al.
  2. **Jacobian- Based Saliency Map Attacks** (JSMA): Paparnot et. Al.
  3. **Carlini and Wagner Attacks**: Carlini et. Al.
  4. **Black Box Attack**: Papernot et. Al.

RESULTS

RESULTS

Dataset	Network	Attack Type	Defense Type	Accuracy (%)	Adversarial Accuracy (%)	Fooling Rate (%)
MNIST	**Basic CNN**	**FGSM**[6]	None	99.37	9.96	-
			Epistemic	98.50	22.92	-
			Aleatoric	99.35	8.75	-
		CW[1]	None	99.30	-	99-100
			Epistemic	98.37	-	30-37
			Aleatoric	99.32	-	67-80
		JSMA[12]	None	99.35	-	89-92
			Epistemic	98.62	-	22-27
			Aleatoric	99.34	-	73-81
		BB[11]	None	99.32	67.78	-
			Epistemic	98.51	63.29	-
			Aleatoric	99.20	62.08	-
CIFAR10	**Fully Convolutional Network**	**FGSM**[6]	None	77.84	9.98	-
			Epistemic	76.28	12.44	-
			Aleatoric	78.00	10.38	-

EPISTEMIC UNCERTAINTY ESTIMATION

Dataset	Network	Attack Type	Defense Type	Accuracy (%)	Adversarial Accuracy (%)	Fooling Rate (%)
MNIST	**Basic CNN**	**FGSM**[6]	None	99.37	9.96	-
			Epistemic	98.50	22.92	-
			Aleatoric	99.35	8.75	-
		CW[1]	None	99.30	-	99-100
			Epistemic	98.37	-	30-37
			Aleatoric	99.32	-	67-80
		JSMA[12]	None	99.35	-	89-92
			Epistemic	98.62	-	22-27
			Aleatoric	99.34	-	73-81
		BB[11]	None	99.32	67.78	-
			Epistemic	98.51	63.29	-
			Aleatoric	99.20	62.08	-
CIFAR10	**Fully Convolutional Network**	**FGSM**[6]	None	77.84	9.98	-
			Epistemic	76.28	12.44	-
			Aleatoric	78.00	10.38	-

ALEATORIC UNCERTAINTY ESTIMATION

Dataset	Network	Attack Type	Defense Type	Accuracy (%)	Adversarial Accuracy (%)	Fooling Rate (%)
MNIST	**Basic CNN**	**FGSM**[6]	None	99.37	9.96	-
			Epistemic	98.50	22.92	-
			Aleatoric	99.35	8.75	-
		CW[1]	None	99.30	-	99-100
			Epistemic	98.37	-	30-37
			Aleatoric	99.32	-	67-80
		JSMA[12]	None	99.35	-	89-92
			Epistemic	98.62	-	22-27
			Aleatoric	99.34	-	73-81
		BB[11]	None	99.32	67.78	-
			Epistemic	98.51	63.29	-
			Aleatoric	99.20	62.08	-
CIFAR10	**Fully Convolutional Network**	**FGSM**[6]	None	77.84	9.98	-
			Epistemic	76.28	12.44	-
			Aleatoric	78.00	10.38	-

BLACK BOX ATTACK

Dataset	Network	Attack Type	Defense Type	Accuracy (%)	Adversarial Accuracy (%)	Fooling Rate (%)
MNIST	**Basic CNN**	**FGSM**[6]	None	99.37	9.96	-
			Epistemic	98.50	22.92	-
			Aleatoric	99.35	8.75	-
		CW[1]	None	99.30	-	99-100
			Epistemic	98.37	-	30-37
			Aleatoric	99.32	-	67-80
		JSMA[12]	None	99.35	-	89-92
			Epistemic	98.62	-	22-27
			Aleatoric	99.34	-	73-81
		BB[11]	None	99.32	67.78	-
			Epistemic	98.51	63.29	-
			Aleatoric	99.20	62.08	-
CIFAR10	**Fully Convolutional Network**	**FGSM**[6]	None	77.84	9.98	-
			Epistemic	76.28	12.44	-
			Aleatoric	78.00	10.38	-

MC-DROPOUT APPROXIMATION

Dataset	Network	Attack Type	Defense Type	Accuracy (%)	Adversarial Accuracy (%)	Fooling Rate (%)
MNIST	**Basic CNN**	**FGSM**[6]	None	99.37	9.96	-
			Epistemic	98.50	22.92	-
			Aleatoric	99.35	8.75	-
		CW[1]	None	99.30	-	99-100
			Epistemic	98.37	-	30-37
			Aleatoric	99.32	-	67-80
		JSMA[12]	None	99.35	-	89-92
			Epistemic	98.62	-	22-27
			Aleatoric	99.34	-	73-81
		BB[11]	None	99.32	67.78	-
			Epistemic	98.51	63.29	-
			Aleatoric	99.20	62.08	-
CIFAR10	**Fully Convolutional Network**	**FGSM**[6]	None	77.84	9.98	-
			Epistemic	76.28	12.44	-
			Aleatoric	78.00	10.38	-

CONCLUSION

QUESTION

Can a neural network mitigate the effects of adversarial attacks by estimating the uncertainty in its predictions ?

ANSWER(S)

- Adversarial perturbations cannot be distinguished as input noise through aleatoric uncertainty estimation.
- Epistemic uncertainty estimation, manifested as Bayesian Neural Networks might be robust to adversarial attacks.
- Results inconclusive, due to the lack of mathematical bounds on the approximation through ensembles and MC-Dropout.
- **Sufficient Conditions for Robustness to Adversarial Examples: a Theoretical and Empirical Study with Bayesian Neural Network.**
- https://openreview.net/forum?id=B1eZRiC9YX

CONCLUSION

- There is no easy way out of using robustness certification to guarantee safety of deep neural networks.

- Even then, the mode of action of a specific type of adversarial attack needs to be taken into consideration.

- **Research Question**: How to certify against black box attacks?