Revisiting Adversarial Attacks via Visual Imperceptible Bound

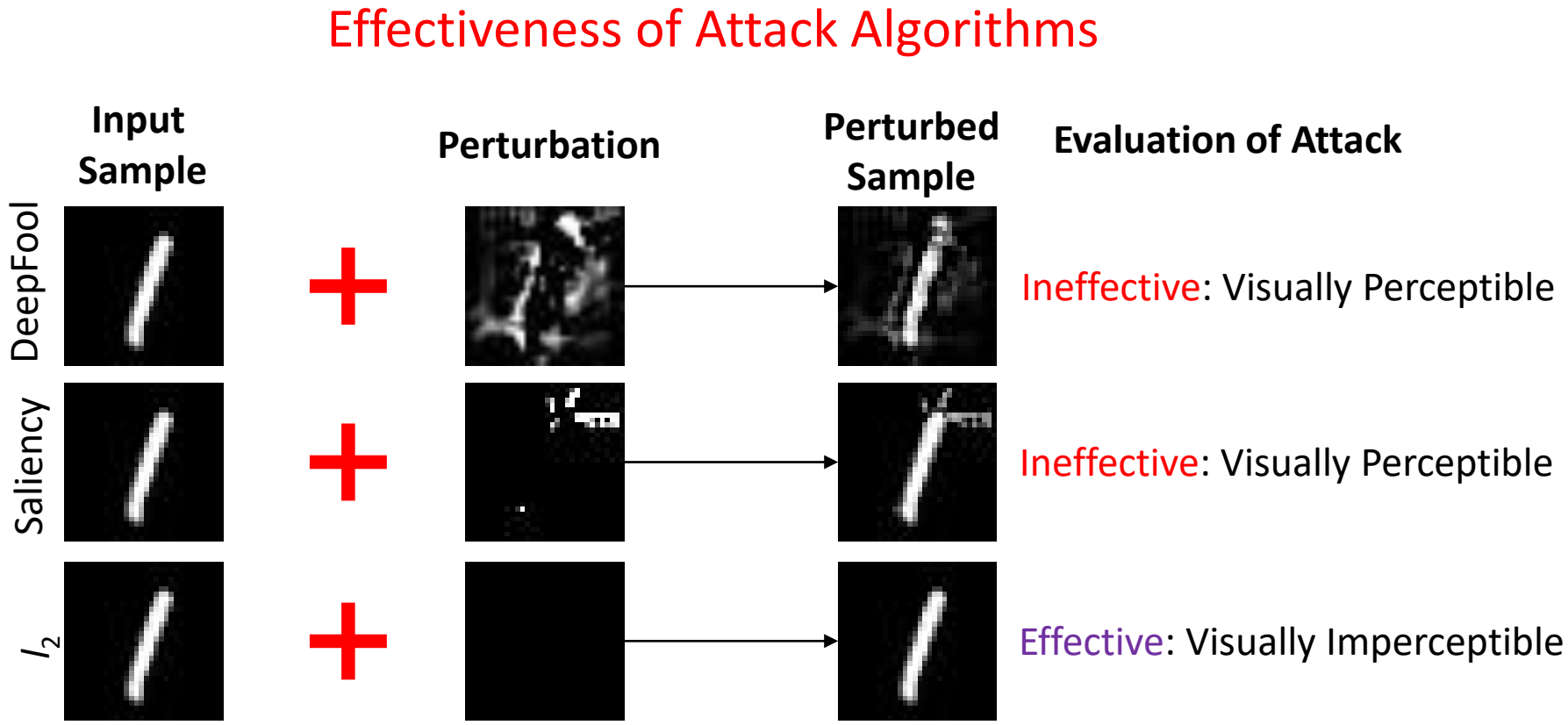
Saheb Chhabra[1], Akshay Agarwal[1], Richa Singh[2], Mayank Vatsa[2]
[1]IIIT-Delhi, India; [2]IIT Jodhpur, India

Motivation

- Robustness of AI systems against adversarial attacks is still an open question
- Effective Defense Method:

  Retraining of target DNN using adversarial samples
- Adversarial training method is not robust against unseen attacks

Effectiveness of Attack Algorithms



Aims and Research Contributions

- The aim is to design a defense model that is robust
  - within certain bound in which the visual appearance of the image should be preserved while performing adversarial manipulation
  - against both seen and unseen attacks

- Research Contributions:
  - Visual Imperceptible Bound (VIB)
  - Proposed defense model that outputs the same prediction for the samples within the VIB

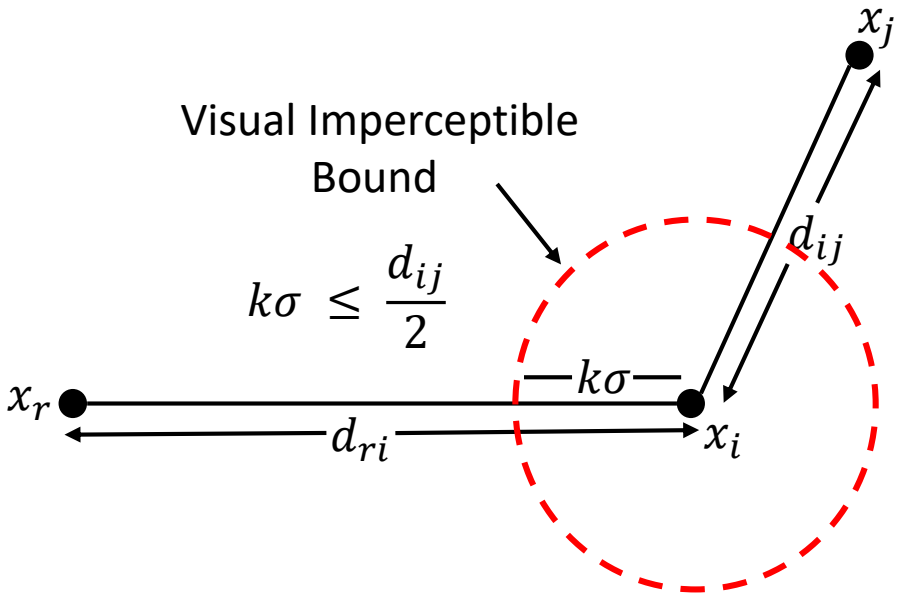
Visual Imperceptible Bound

x – clean image	y – label	$\boldsymbol{D}$ – dataset

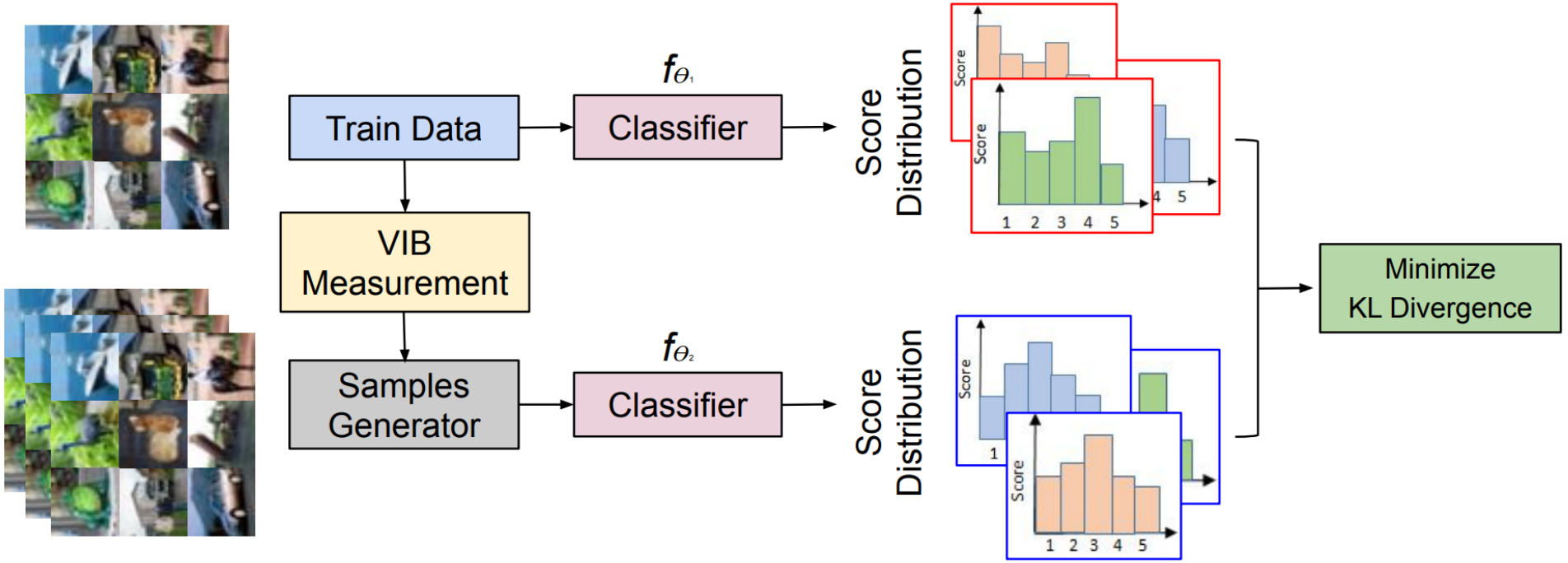
$$\boldsymbol{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$

Absolute distance between samples $d_{ij} = |x_i - x_j|$

Visual Imperceptible Bound $\sigma_i \leq \frac{d_{ij}}{2k}$



Block Diagram: Proposed Algorithm



Experiments and Results

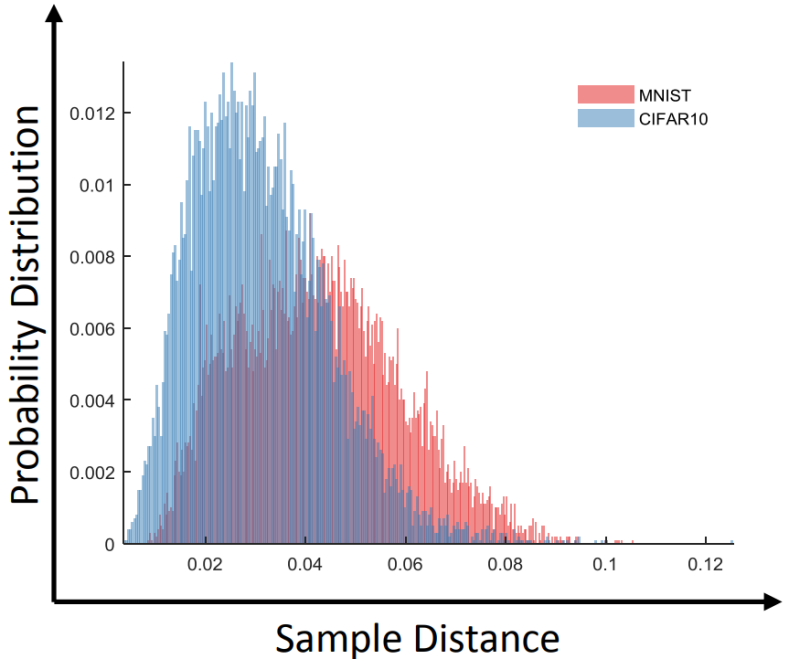
- Three experiments are performed: 1) Database Evaluation, 2) Attack Evaluation and 3) Proposed Defense Model Evaluation

- First and second experiment provide insights towards the database characteristics and behavior of attack algorithms

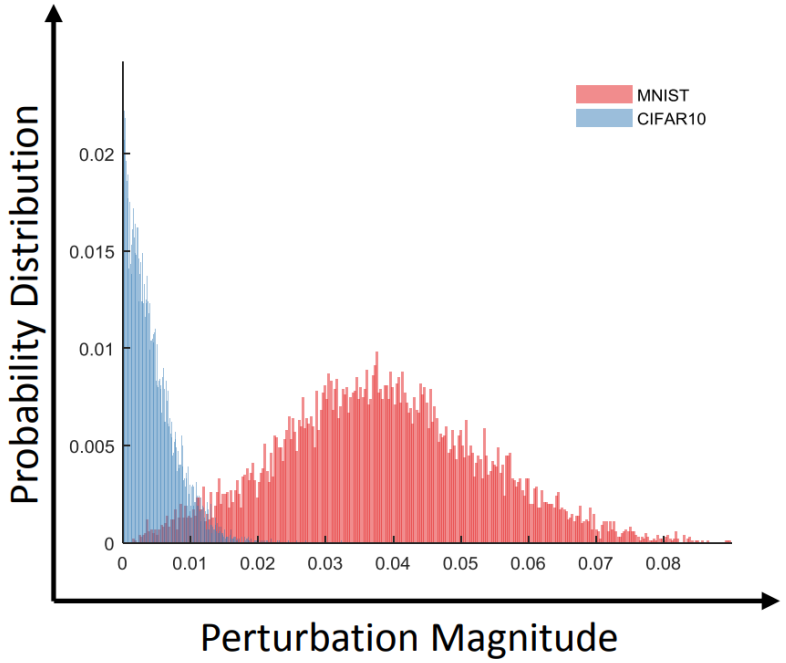
- Third experiment evaluates the performance of the proposed defense model

Experiment	Database	Attack
Database Evaluation	MNIST, CIFAR-10	
Attack Evaluation	MNIST, CIFAR-10	DeepFool, FGSM, JSMA, l_2
Proposed Defense Model Evaluation	MNIST, CIFAR-10, Tiny ImageNet	DeepFool, FGSM, JSMA, l_2

Evaluation of Vulnerability of Databases



Comparison of normalized distance in image space



Comparison of normalized magnitude of perturbation noise added to perform attack using DeepFool attack algorithm.

Evaluation of Attack Algorithms

- Attacks are evaluated based on whether the adversarial examples are generated inside the VIB or not
- Noise would become perceptible even with very less magnitude of the perturbation on the MNIST
- MNIST database has relatively lower VIB
- DeepFool attack is effective on CIFAR-10 database

Table: Estimated classification accuracy on attack algorithms within the VIB with k = 2

Database	Attack	Estimated Accuracy
CIFAR-10	DeepFool	83.91
	FGSM	18.15
MNIST	JSMA	32.02
	l_2	17.85

Proposed Defense Model

Table: Classification accuracy on the MNIST database

Data Type	#Samples	Original Model	Proposed Model Robust with			
			k=1.0	k=2.0	k=2.5	k=3.0
Original	10	**99.55**	99.40	99.48	99.33	99.50
	15		99.33	99.52	99.41	99.37
	20		99.32	99.17	99.48	**99.53**
DeepFool	10	**34.19**	94.13	93.47	95.67	94.30
	15		91.79	92.49	95.95	95.65
	20		**96.04**	92.67	96.38	95.12
FGSM ($\epsilon = 0.1$)	10	**89.65**	97.91	97.88	97.88	98.14
	15		97.55	97.90	98.27	98.28
	20		98.22	97.81	**98.36**	98.35
FGSM ($\epsilon = 0.2$)	10	**54.52**	86.21	87.35	89.93	88.20
	15		81.43	80.11	89.29	**90.81**
	20		89.34	86.48	88.21	84.70
JSMA	10	**0.10**	71.75	64.30	77.86	78.02
	15		68.13	58.89	**80.36**	60.94
	20		73.92	62.81	71.94	63.13
C&W (l_2)	10	**12.89**	87.98	87.92	88.47	84.90
	15		86.18	**90.63**	88.32	88.97
	20		87.97	84.87	91.15	91.43

Table: Classification accuracy on the CIFAR-10 database

Data Type	#Samples	Original Model	Proposed Model Robust with			
			k=1.0	k=2.0	k=2.5	k=3.0
Original	10	**83.91**	87.52	87.59	87.27	86.58
	15		88.05	87.94	88.37	88.45
	20		88.41	**88.82**	88.64	88.11
DeepFool	10	**31.67**	83.64	83.75	82.83	81.95
	15		84.26	84.71	85.78	85.58
	20		**85.94**	85.59	85.43	85.21
FGSM ($\epsilon = 2.0$)	10	**55.91**	81.32	81.32	80.35	79.71
	15		81.96	82.56	83.06	82.70
	20		**83.53**	83.11	82.51	82.91
FGSM ($\epsilon = 4.0$)	10	**30.51**	67.29	67.38	66.77	66.27
	15		67.70	69.15	69.63	69.84
	20		**70.92**	70.60	69.95	69.73
JSMA	10	**1.14**	70.20	70.48	69.99	68.58
	15		70.02	72.05	71.13	**73.88**
	20		73.01	72.65	72.31	73.41
C&W (l_2)	10	**12.10**	83.70	83.84	82.81	82.17
	15		84.32	84.85	85.90	85.46
	20		**86.03**	85.72	85.56	85.32

Greater than estimated accuracy

Adversarial Robustness

- Proposed defense model is robust against unseen adversarial attacks and does not require knowledge of any attack

Table: Comparing the performance of the proposed algorithm with Adversarial Training (AT) for different attacks on the CIFAR-10 database.

Model	Testing Attacks			
	DeepFool	FGSM	JSMA	C&W (l_2)
AT with DeepFool	82.50	81.53	74.27	84.01
AT with FGSM (ϵ = 2.0)	83.06	82.54	73.51	81.87
AT with FGSM (ϵ = 4.0)	79.02	79.53	73.19	79.80
AT with FGSM (ϵ = 8.0)	79.90	79.44	72.61	79.99
AT with JSMA	81.09	80.10	**75.63**	79.70
Proposed	**85.94**	**83.53**	73.88	**86.03**

Conclusion

- Proposed a new concept of Visual imperceptible bound
- Proposed a defense model which robust against both seen and unseen attacks
- The defense model is robust within the VIB
- The defense model is attack agnostic

Thank you