Efficient and Transferable Adversarial Examples from Bayesian Neural Networks

M. Gubri, M. Cordy, M. Papadakis, Y. Le Traon, University of Luxembourg
K. Sen, University of California, Berkeley

uni.lu | SnT
Serval
martin.gubri@uni.lu

Introduction

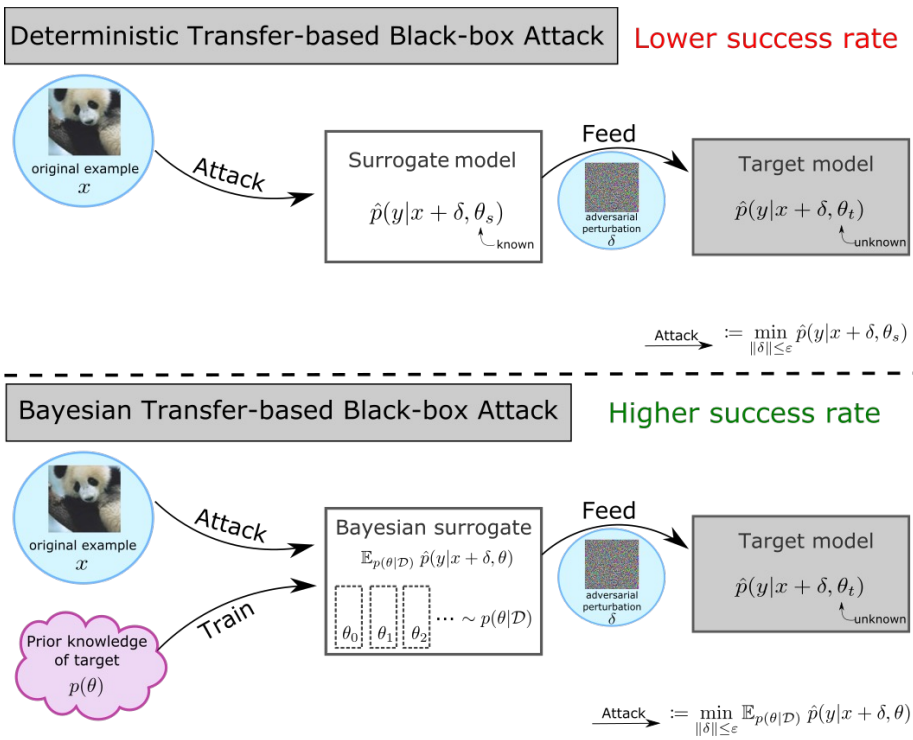
Adversarial examples can **transfer** from a surrogate DNN to a target DNN.

Previous ensemble approach
Diversity (enrich the space of hypothesis) → transferability

Transferability is Bayesian

The weights of the target DNN are **unknown** and can therefore be treated as random variables.

The posterior represents an (approximate) **belief** about the DNN weights.



Do Bayesian deep learning techniques improve transfer-based attacks by capturing uncertainty?

Method

1. Collect samples from the posterior with cSGLD, a cyclical SG-MCMC method
2. Approximate Bayesian Model Averaging during iterative attack

Effective and efficient

- **Effectiveness**
  Improve four attacks success rate from -2.3 to 83.2 percent points
- **Efficiency**
  Divide training computations by 2.5+ compared to Deep Ensemble

Table 1: Number of DNNs (T-DEE) and training computation budget (in flops) to achieve the intra-architecture transferability of cSGLD with Deep Ensemble. Higher is better. ">15" means that 15 DNNs always transfer less than cSGLD.

Dataset	Attack	Norm	T-DEE	Flops Ratio
ImageNet	I-FGSM	L_2	4.91 ±0.11	2.84 ±0.06
		L_∞	4.34 ±0.13	2.51 ±0.08
	MI-FGSM	L_2	4.69 ±0.16	2.71 ±0.10
		L_∞	4.38 ±0.01	2.53 ±0.02
	PGD	L_2	5.00 ±0.11	2.89 ±0.06
		L_∞	4.42 ±0.16	2.56 ±0.09
	FGSM	L_2	5.81 ±0.14	3.35 ±0.10
		L_∞	5.98 ±0.01	3.46 ±0.01
CIFAR10	I-FGSM	L_2	>15 max	>15 max
		L_∞	3.76 ±0.08	3.76 ±0.08
	MI-FGSM	L_2	5.56 ±0.01	3.56 ±0.01
		L_∞	2.88 ±0.01	2.87 ±0.01
	PGD	L_2	>15 max	>15 max
		L_∞	3.74 ±0.13	3.74 ±0.12
	FGSM	L_2	>15 max	>15 max
		L_∞	8.72 ±0.01	8.72 ±0.01
MNIST	I-FGSM	L_2	>15 max	>15 max
		L_∞	3.42 ±0.17	3.42 ±0.17
	MI-FGSM	L_2	>15 max	>15 max
		L_∞	2.79 ±0.03	2.79 ±0.03
	PGD	L_2	>15 max	>15 max
		L_∞	3.26 ±0.28	3.26 ±0.28
	FGSM	L_2	>15 max	>15 max
		L_∞	>15 max	>15 max

Table 2: Transfer success rates of I-FGSM attack on ImageNet hold-out architectures. Higher is better.

Norm	Surrogate	Target Architecture: −ResNet50	−ResNeXt50	−DenseNet121	−MNASNet	−EffNetB0	Nb epochs
L_2	1 cSGLD per arch.	**93.28** ±0.12	**90.61** ±0.20	**92.25** ±0.26	**95.98** ±0.19	**81.88** ±0.30	4 × 135
	1 DNN per arch.	72.99 ±0.32	72.31 ±0.44	64.72 ±0.39	84.21 ±0.18	53.99 ±0.76	4 × 135
L_∞	1 cSGLD per arch.	**92.21** ±0.23	**89.83** ±0.22	**90.86** ±0.19	**95.85** ±0.46	**79.40** ±0.42	4 × 135
	1 DNN per arch.	69.65 ±0.47	69.01 ±0.70	61.00 ±0.66	82.25 ±0.03	49.71 ±1.37	4 × 135

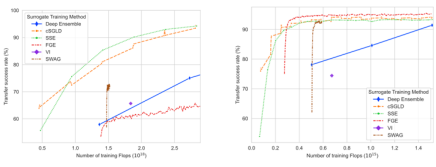
Outperform related work

- **Ours vs. related work**
  Achieve 87.5% of the time higher transferability than 3 test-time techniques
- **Ours + related work**
  Always improve 3 test-time techniques when combined

Surrogate	Target Architecture: ResNet50	ResNeXt50	DenseNet121	MNASNet	EffNetB0
1 DNN	47.81 ±1.09	32.29 ±0.64	23.43 ±0.32	22.52 ±0.45	12.77 ±0.32
+ Input Diversity	76.55 ±1.01	62.57 ±0.56	50.17 ±0.33	49.31 ±0.18	⋆32.64 ±0.09
+ Skip Gradient Method	66.36 ±0.50	51.60 ±0.36	39.05 ±0.24	45.60 ±0.72	30.69 ±0.03
+ Ghost Networks	67.02 ±0.17	46.74 ±0.63	32.57 ±0.17	31.12 ±0.77	17.68 ±0.05
+ Momentum (MI-FGSM)	55.12 ±0.82	38.47 ±0.82	28.19 ±0.14	27.55 ±0.67	16.34 ±0.37
+ Input Diversity	⋆82.47 ±0.43	⋆69.69 ±0.61	57.79 ±0.17	⋆55.99 ±0.37	⋆38.63 ±0.29
+ Skip Gradient Method	68.39 ±0.53	54.57 ±0.60	41.48 ±0.37	47.97 ±0.41	⋆33.16 ±0.37
+ Ghost Networks	71.27 ±0.54	51.46 ±0.84	36.91 ±0.48	34.54 ±0.32	20.51 ±0.30
cSGLD	78.71 ±1.19	65.11 ±1.45	61.49 ±0.59	51.81 ±1.45	31.11 ±0.99
+ Input Diversity	90.03 ±0.10	82.13 ±0.65	81.19 ±0.34	74.48 ±0.39	53.51 ±0.39
+ Skip Gradient Method	81.37 ±0.72	69.88 ±1.31	65.20 ±0.75	71.68 ±0.53	52.15 ±0.32
+ Ghost Networks	87.33 ±0.73	76.00 ±1.33	71.67 ±0.97	61.45 ±0.25	37.19 ±0.68
+ Momentum (MI-FGSM)	82.89 ±0.70	70.42 ±1.26	66.39 ±0.74	56.68 ±0.97	36.00 ±1.15
+ Input Diversity	**93.97** ±0.26	**87.69** ±0.64	**86.78** ±0.16	**81.08** ±0.14	**60.87** ±0.48
+ Skip Gradient Method	84.19 ±0.21	73.14 ±0.99	67.35 ±0.26	74.36 ±0.47	55.30 ±0.16
+ Ghost Networks	89.53 ±0.05	78.69 ±0.19	73.33 ±0.58	63.56 ±0.35	39.79 ±0.52

Other deep learning methods

- cSGLD competitive compared to 5 other Bayesian and Ensemble techniques



(a) ImageNet (b) CIFAR-10

Figure 2: Intra-architecture L_∞ I-FG(S)M success rate with respect to the training computational complexity of an increasing number of samples from six training techniques.

More information →

martin.gubri@uni.lu