DeepMind

Hindering Adversarial Attacks with Implicit Neural Representations

Andrei A. Rusu, Dan A. Calian, Sven Gowal, Raia Hadsell

ICML 2022

Corresponding Author:
andrei@deepmind.com

Hindering Adversarial Attacks with Implicit Neural Representations

Two principles for designing defences against adversarial attacks:

1. Train models which are insensitive to *all* adversarial perturbations.
2. Make computing adversarial perturbations expensive, ideally intractable.

Hindering Adversarial Attacks with Implicit Neural Representations

Two principles for designing defences against adversarial attacks:

1. Train models which are insensitive to *all* adversarial perturbations.
2. Make computing adversarial perturbations expensive, ideally intractable.

Considering principle 2 we ask:

How do we leverage computational hardness for adversarial robustness?

Hindering Adversarial Attacks with Implicit Neural Representations

Two principles for designing defences against adversarial attacks:

1. Train models which are insensitive to *all* adversarial perturbations.
2. Make computing adversarial perturbations expensive, ideally intractable.

Considering principle 2 we ask:

How do we leverage computational hardness for adversarial robustness?

Our hypothesis: Denying access to model outputs is an effective strategy.

Hindering Adversarial Attacks with Implicit Neural Representations

Goal: Make computing adversarial perturbations expensive, ideally intractable.

Question: How do we leverage computational hardness for adversarial robustness?

Hypothesis: Denying access to model outputs is an effective strategy.

Why?

Most attack vectors assume:

- Access to precise model outputs for arbitrary perturbations of inputs.
- Some way to approximate decision boundaries of the model under attack.
- Ability to verify perturbation candidates.

Hindering Adversarial Attacks with Implicit Neural Representations

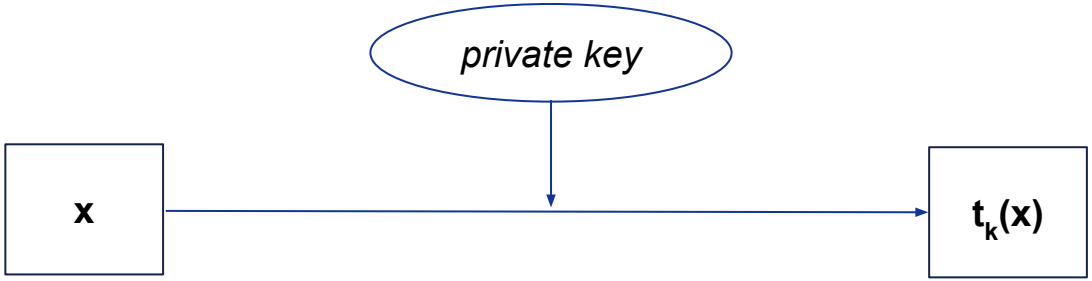
Goal: Make computing adversarial perturbations expensive, ideally intractable.

Question: How do we leverage computational hardness for adversarial robustness?

Hypothesis: Denying access to model outputs is an effective strategy.

How?

Use a key-based input transform. which is difficult to invert and approximate!



Hindering Adversarial Attacks with Implicit Neural Representations

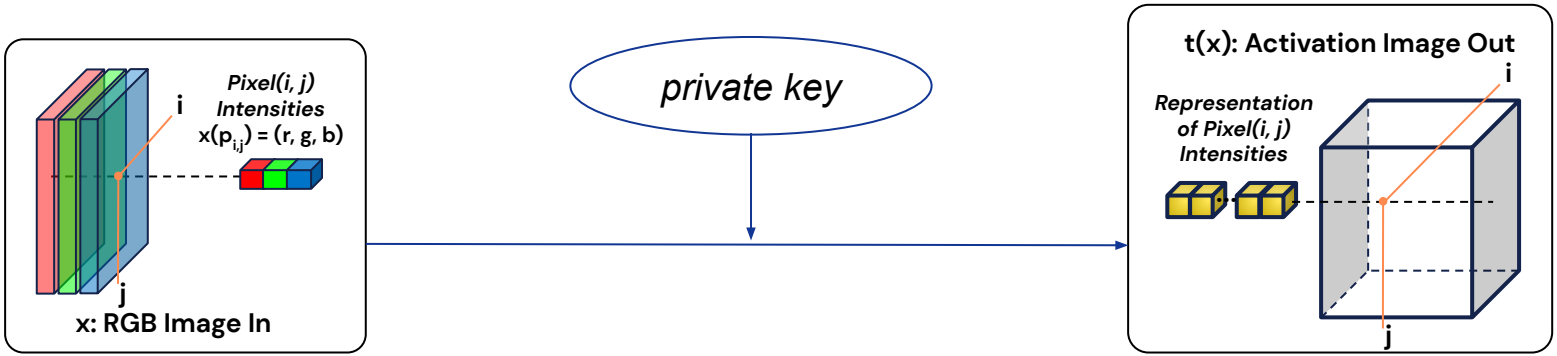
Goal: Make computing adversarial perturbations expensive, ideally intractable.

Question: How do we leverage computational hardness for adversarial robustness?

Hypothesis: Denying access to model outputs is an effective strategy.

How: Use a key-based input transform. which is difficult to invert and approximate!

Introducing LINAC (Lossy Implicit Neural Activation Coding):



Hindering Adversarial Attacks with Implicit Neural Representations

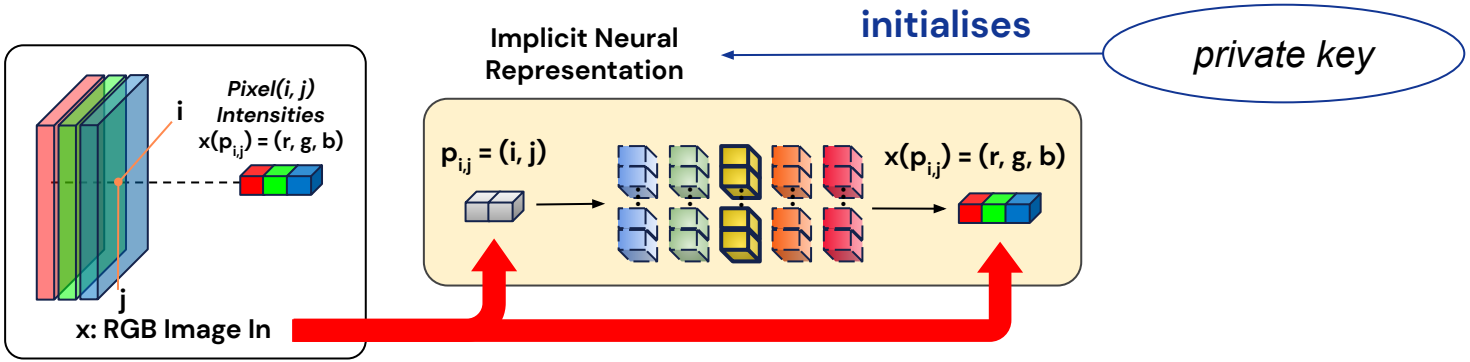
Goal: Make computing adversarial perturbations expensive, ideally intractable.

Question: How do we leverage computational hardness for adversarial robustness?

Hypothesis: Denying access to model outputs is an effective strategy.

How: Use a key-based input transform. which is difficult to invert and approximate!

Introducing LINAC (Lossy Implicit Neural Activation Coding):



Hindering Adversarial Attacks with Implicit Neural Representations

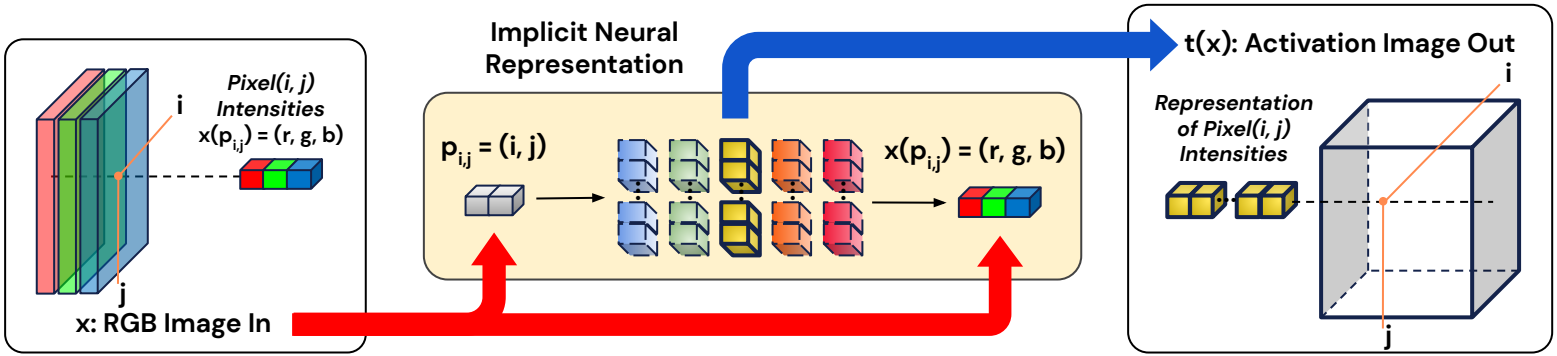
Goal: Make computing adversarial perturbations expensive, ideally intractable.

Question: How do we leverage computational hardness for adversarial robustness?

Hypothesis: Denying access to model outputs is an effective strategy.

How: Use a key-based input transform. which is difficult to invert and approximate!

Introducing LINAC (Lossy Implicit Neural Activation Coding):



Hindering Adversarial Attacks with Implicit Neural Representations

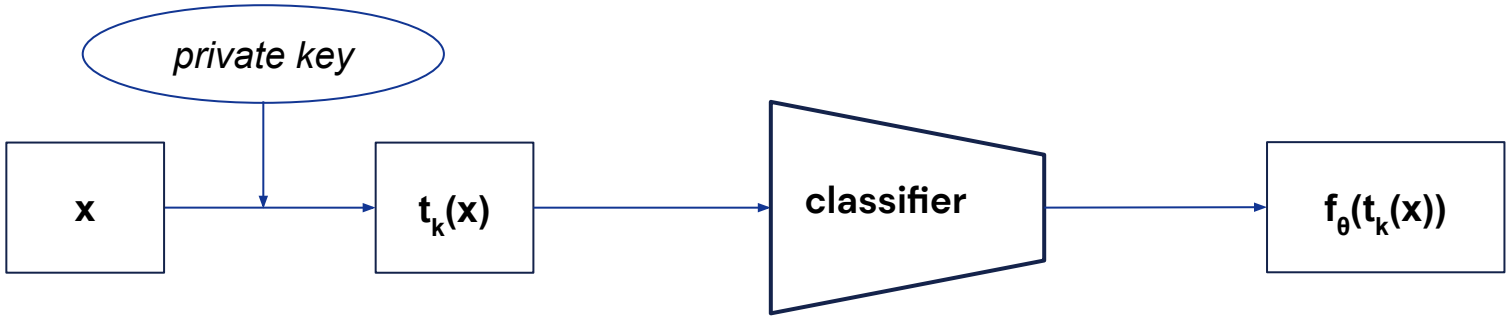
Goal: Make computing adversarial perturbations expensive, ideally intractable.

Question: How do we leverage computational hardness for adversarial robustness?

Hypothesis: Denying access to model outputs is an effective strategy.

How: Use a key-based input transform. which is difficult to invert and approximate!

LINAC (Lossy Implicit Neural Activation Coding) defended classifier training:



Hindering Adversarial Attacks with Implicit Neural Representations

Goal: Make computing adversarial perturbations expensive, ideally intractable.

Question: How do we leverage computational hardness for adversarial robustness?

Hypothesis: Denying access to model outputs is an effective strategy.

How: Use a key-based input transform. which is difficult to invert and approximate!

AprilPyone & Kiya (2021a)

Threat Model:

- Attacker has full algorithmic knowledge about the approach.
- Attacker has complete information about the classification pipeline, model architecture, training dataset and parameters of the defended classifier.
- Attacker does not know the *private key* of the input transformation.

Hindering Adversarial Attacks with Implicit Neural Representations

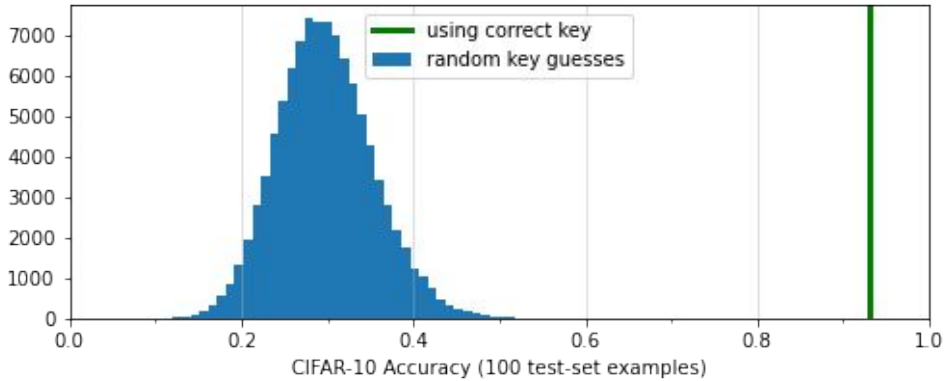
Goal: Make computing adversarial perturbations expensive, ideally intractable.

Question: How do we leverage computational hardness for adversarial robustness?

Hypothesis: Denying access to model outputs is an effective strategy.

How: Use a key-based input transform. which is difficult to invert and approximate!

Is a LINAC defended classifier denying access to its outputs absent the *private key*?



Direct attack on the *private key*: histogram of accuracies of the same LINAC defended classifier with inputs transformed using either the correct key (green) or 100,000 randomly chosen keys (blue).

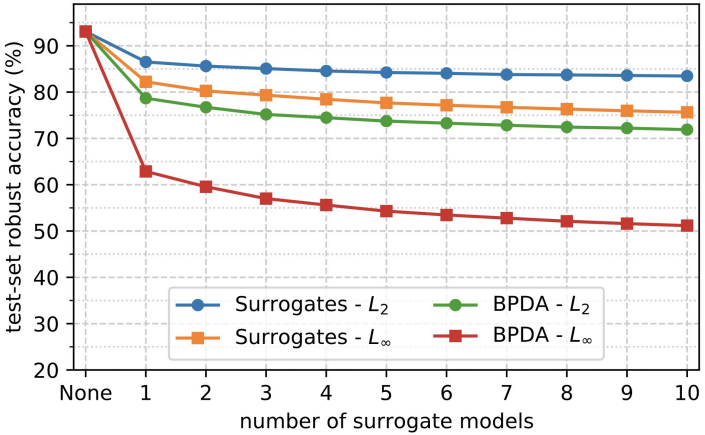
Hindering Adversarial Attacks with Implicit Neural Representations

Goal: Make computing adversarial perturbations expensive, ideally intractable.

Question: How do we leverage computational hardness for adversarial robustness?

Hypothesis: Denying access to model outputs is an effective strategy.

How: Use a key-based input transform. which is difficult to invert and approximate!



Is LINAC difficult to usefully approximate absent the *private key*?

CIFAR-10 test-set robust accuracy estimates (Best Known) vs. number of attacker-trained surrogate models. We also plot the clean accuracy of 93.08% for reference (None).

Hindering Adversarial Attacks with Implicit Neural Representations

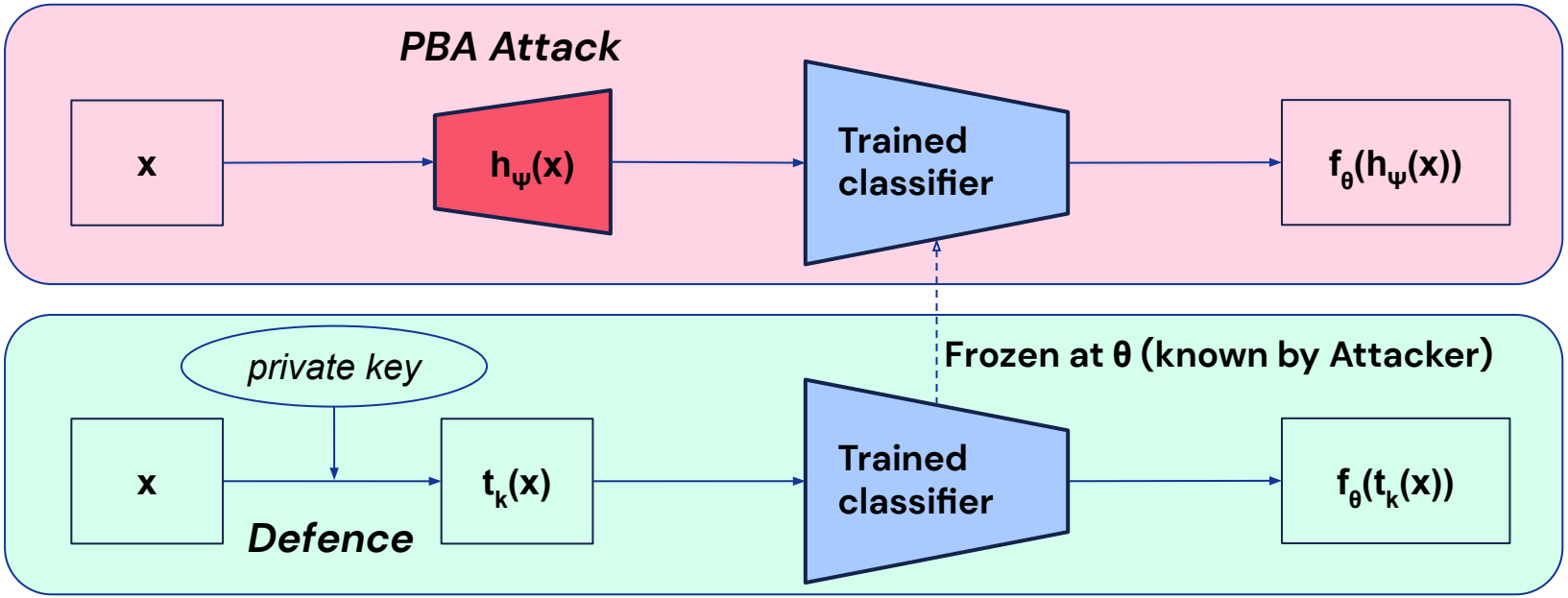
Is LINAC difficult to usefully approximate absent the *private key*?

CIFAR-10 test set robust accuracy (%) of a single LINAC defended classifier w.r.t. a suite of L_∞ and L_2 transfer attacks, valid under our threat model, using various source classifiers to generate adversarial perturbations.

		Transfer Attack Source Models					*Best Adversary*
Norm	**Attack**	**Nominal Source**	**Adversarial Training (L_∞)**	**Adversarial Training (L_2)**	**Defended Surrogates (Attacker Keys)**	**Reconstruction-Based Surrogates (BPDA)**	**All Source Models**
L_∞	AA	92.77	80.42	70.29	84.00	59.40	41.18
	MT	84.57	72.96	56.08	85.70	55.37	47.91
	PGD	85.99	60.97	44.06	87.32	56.00	41.22
	Square	85.12	65.69	52.66	75.91	69.14	49.76
	Best Known	81.91	54.97	39.20	75.64	51.17	37.04
L_2	AA	90.84	86.75	80.83	88.27	74.59	71.32
	MT	87.55	85.34	84.81	87.31	74.98	73.83
	PGD	88.61	82.39	74.19	88.36	75.00	70.90
	Square	88.58	84.50	79.31	84.08	83.26	77.68
	Best Known	86.06	79.42	71.92	83.48	71.89	68.41

Hindering Adversarial Attacks with Implicit Neural Representations

Novel attack: Parametric Bypass Approximation (PBA) invalidates the approach of AprilPyone & Kiya (2021a): Block pixel-shuffle (4x4 fixed random permutation)



Hindering Adversarial Attacks with Implicit Neural Representations

Is LINAC difficult to usefully approximate absent the *private key*?

CIFAR-10 test set robust accuracy (%) of a single LINAC defended classifier w.r.t. a suite of L_∞ and L_2 attacks, valid under our threat model, using different strategies such as transfer and adaptive attacks. Our novel **PBA** adaptive attacks are overall **more effective that both transfer and BPDA attack strategies**.

		All Source Models	Adaptive Attacks	
Norm	**Attack**	Transfer	BPDA	PBA
L_∞	AA	41.18	59.40	68.34
	MT	47.91	55.37	46.75
	PGD	41.22	56.00	44.05
	Square	49.76	69.14	48.59
	Best Known	37.04	51.17	35.32
L_2	AA	71.32	74.59	73.10
	MT	73.83	74.98	67.85
	PGD	70.90	75.00	66.93
	Square	77.68	83.26	74.70
	Best Known	68.41	71.89	61.23

Hindering Adversarial Attacks with Implicit Neural Representations

Goal: Make computing adversarial perturbations expensive, ideally intractable.

Question: How do we leverage computational hardness for adversarial robustness?

Hypothesis: Denying access to model outputs is an effective strategy.

How: Use a key-based input transform. which is difficult to invert and approximate!

Conclusions:

LINAC defended classifiers deny access to their outputs absent the *private key*!

LINAC decision boundaries are difficult to usefully approximate absent the private key!

LINAC successfully hinders very expensive attacks and PBA!

For further details please have a look at the paper and come speak with us at the poster.