On the Robustness of Randomized Ensembles to Adversarial Perturbations

Hassan Dbouk & Naresh Shanbhag

Department of Electrical and Computer Engineering

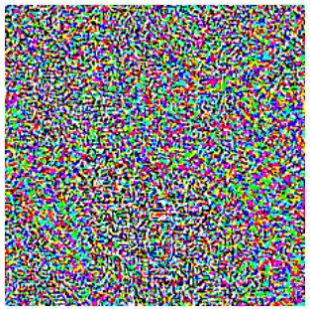
University of Illinois at Urbana-Champaign

Robust and Efficient Inference

deep nets are vulnerable

original sample

$+ .007 \times$



$=$

adversarial sample

decision: ‘panda’

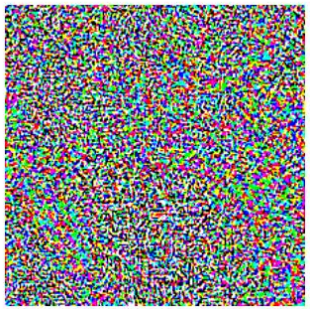
decision: ‘gibbon’

Robust and Efficient Inference

deep nets are vulnerable

original sample

$+ .007 \times$



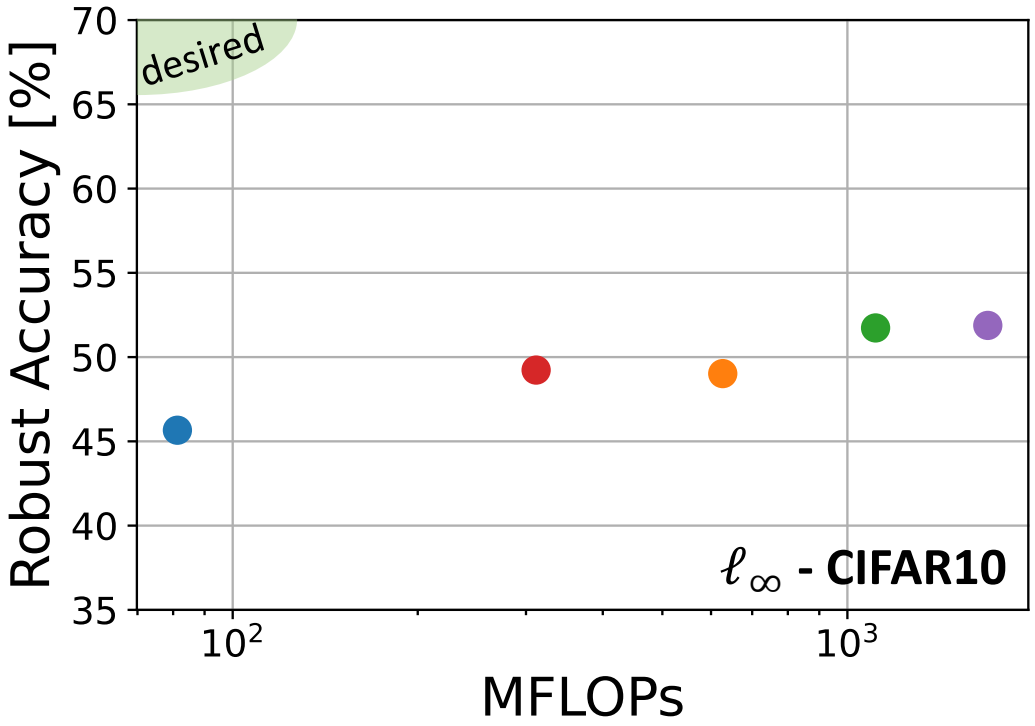
$=$

adversarial sample

decision: ‘panda’

decision: ‘gibbon’

robustness is expensive

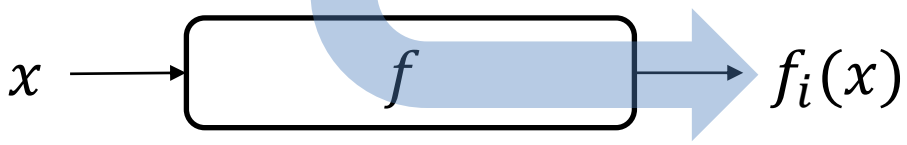


Robustness via Randomized Ensembles

multiple classifiers $f_1, \ldots, f_M$

probabilities α_1 α_2 ••• α_M

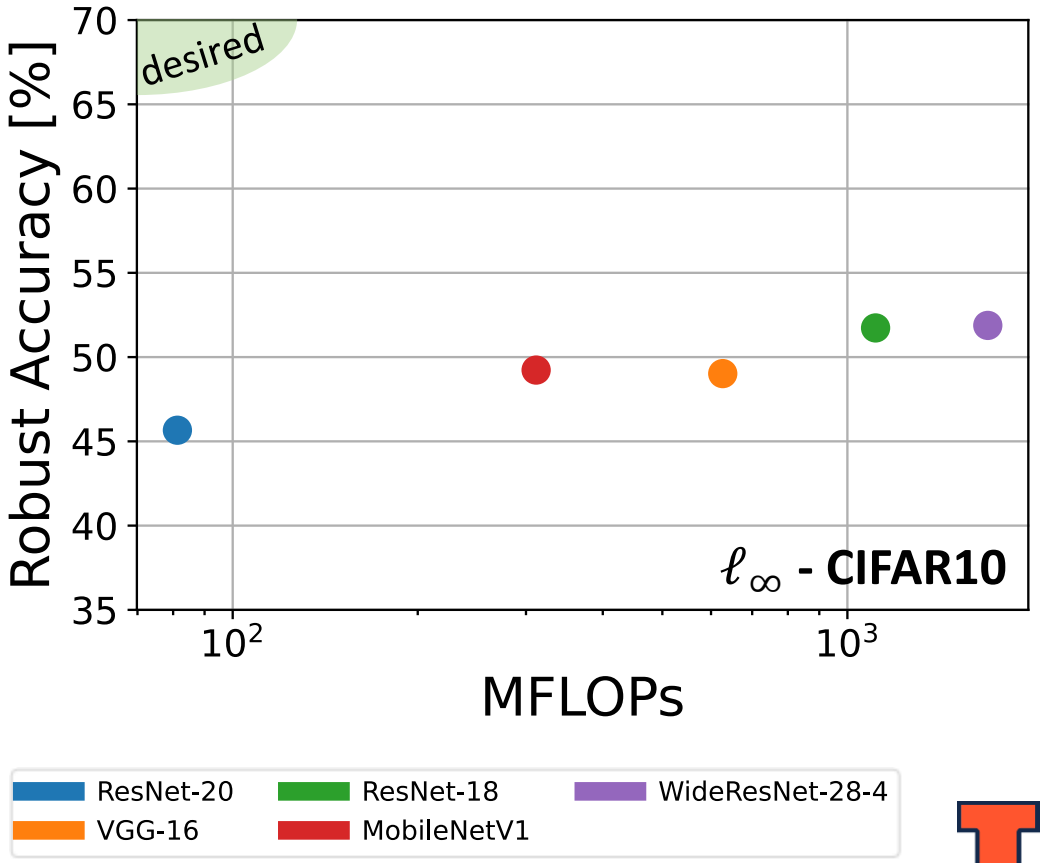
classifiers f_1 f_2 ••• f_M



inference: pick **one** at random

no increase in # of FLOPS

robustness is <u>expensive</u>

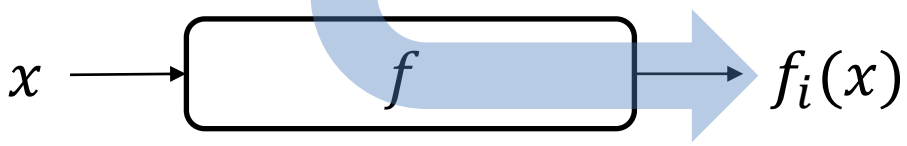


Boosted Adversarial Training (BAT) [Pinot et al, ICML'20]

multiple classifiers $f_1, \ldots, f_M$

probabilities α_1 α_2 ••• α_M

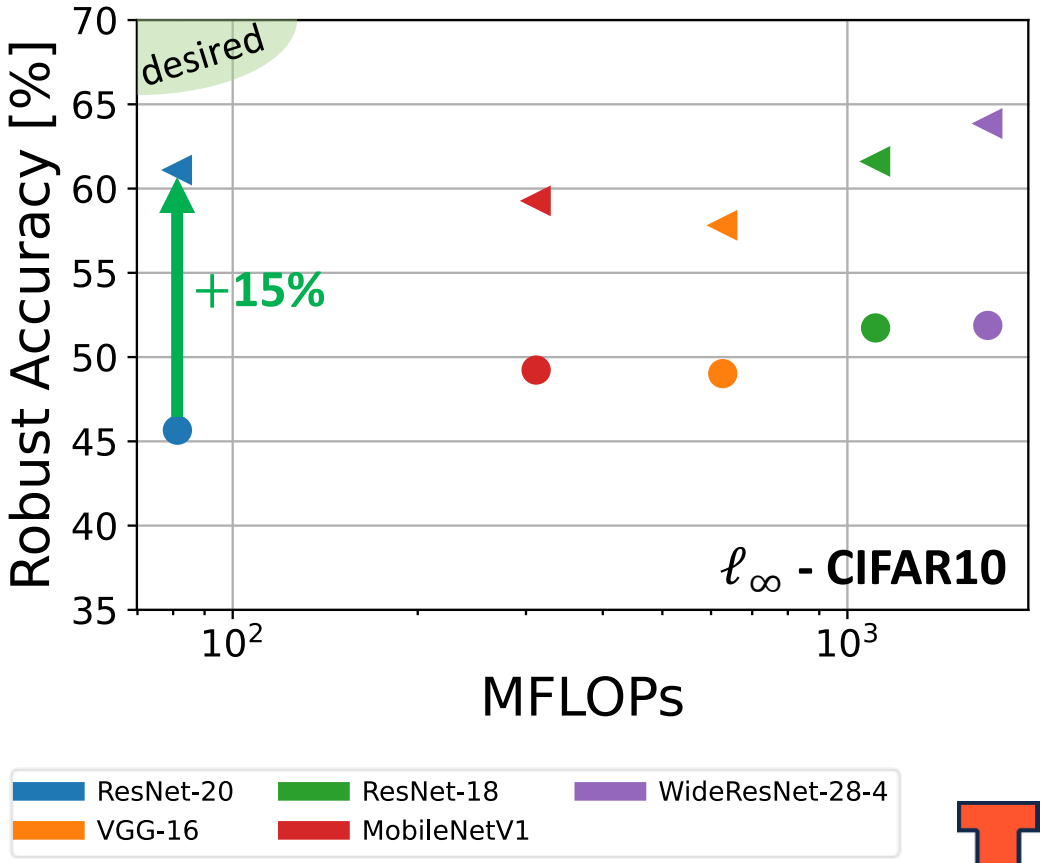
classifiers f_1 f_2 ••• f_M



inference: pick **one** at random

no increase in # of FLOPS

using <u>two classifiers</u> trained via BAT

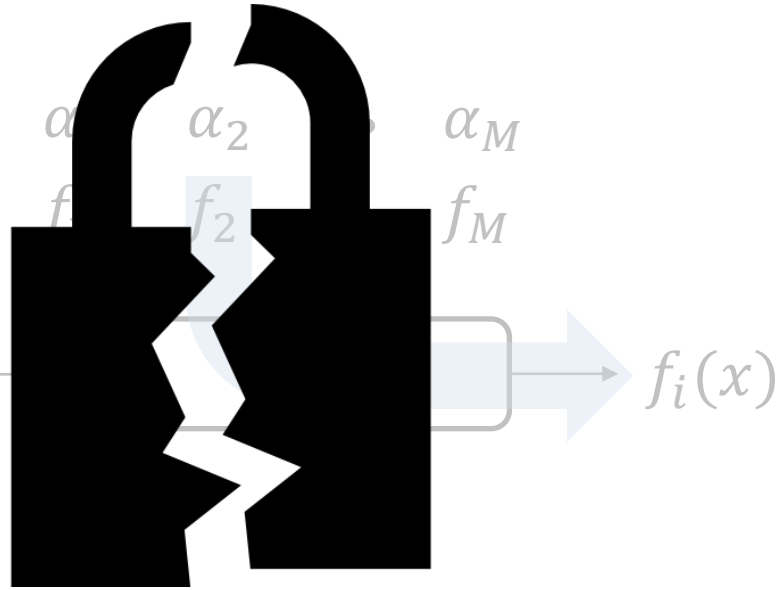


Our Prior Work [ICML'22]: Revealing the vulnerability

multiple classifiers $f_1, \ldots, f_M$

probabilities α_1 α_2 α_M

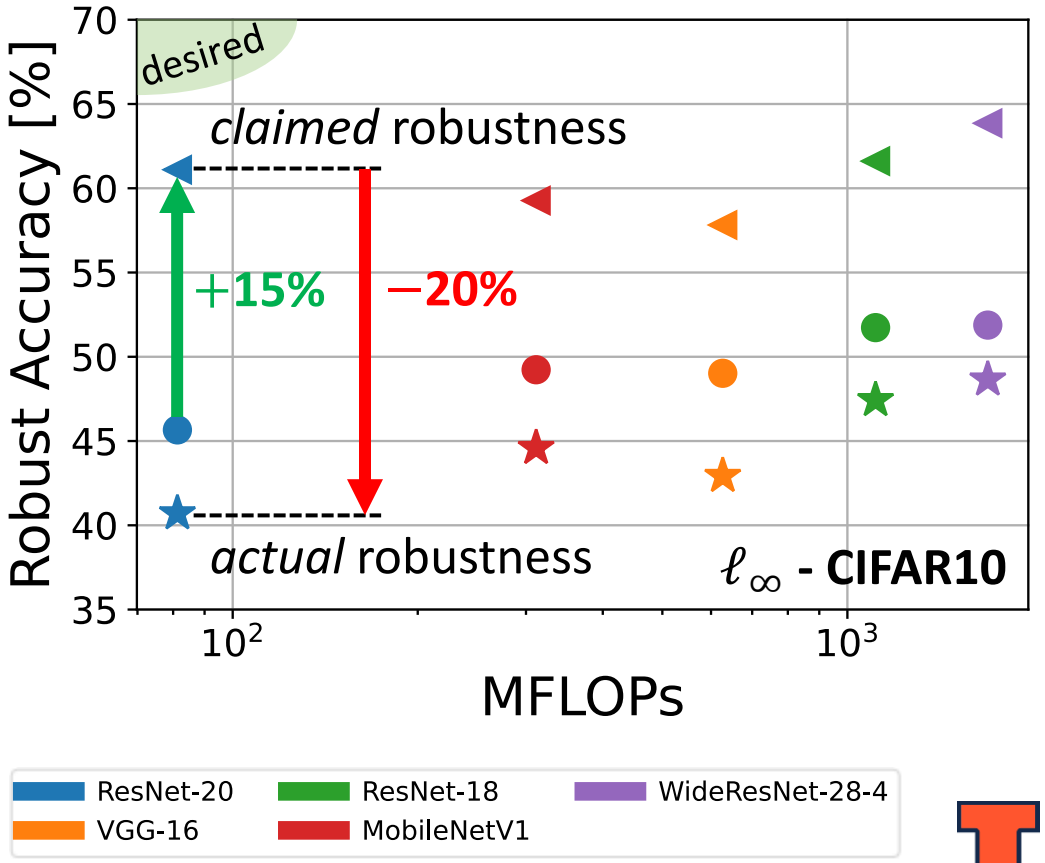
classifiers f_1 f_2 f_M



inference: pick **one** at random

no increase in # of FLOPS

BAT defense <u>compromised</u>



Fundamental Questions on RECs

- when does randomization help in improving robustness?
- what are the limits of RECs?
- how to find the optimal sampling probability?
- how do we train robust RECs in practice?

Summary of Theoretical Contributions

- derived a theoretical framework for analyzing the robustness of RECs
- derived fundamental results on:
  - necessary and sufficient conditions for RECs to be useful

Thm 1.$\forall\ f_1$ & f_2 with adv. risks η_1& η_2. If: $\mathbb{P}\{z \in \mathcal{R}_1\} > |\eta_1 - \eta_2| \rightarrow \eta(\boldsymbol{\alpha}^*) = \frac{1}{2}(\eta_1 + \eta_2 - \mathbb{P}\{z \in \mathcal{R}_1\})$

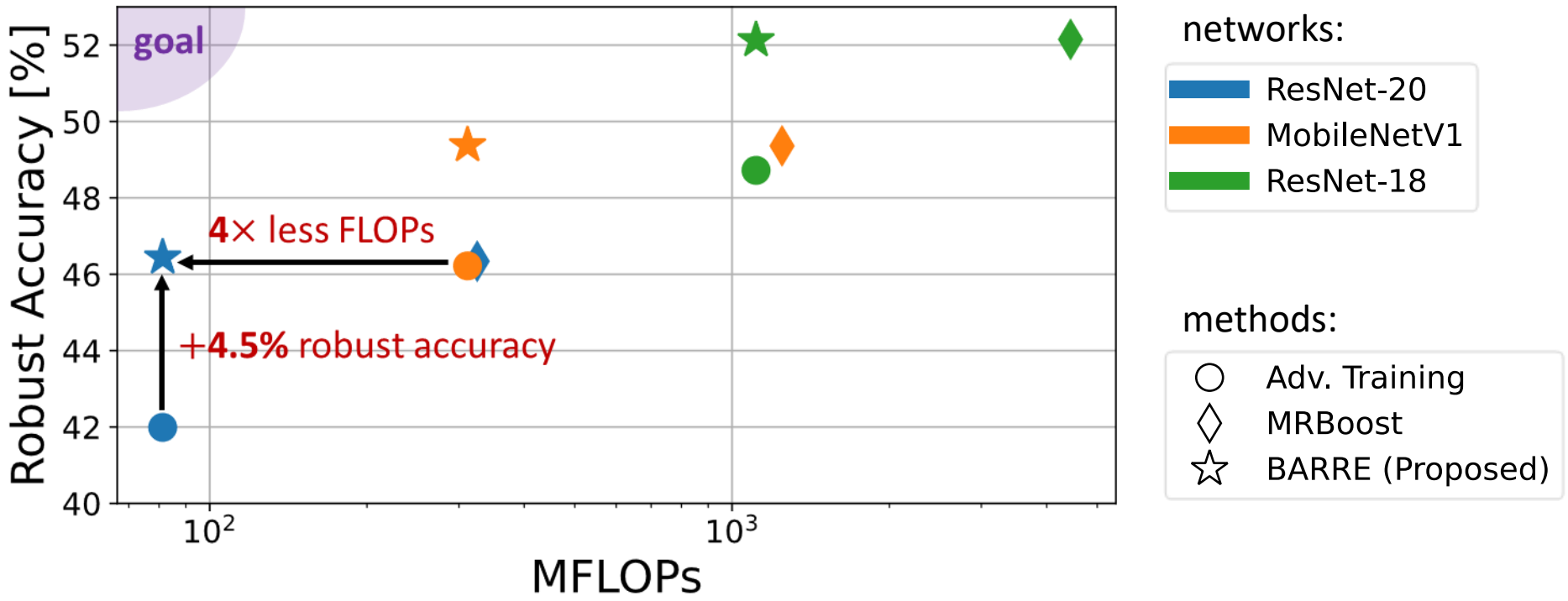
  - theoretical robustness limits of RECs

Thm 2.$\forall\ \{f_i\}_{i=1}^M$ with adv. risks $\{\eta_i\}_{i=1}^M$. We can tightly bound the REC adv. risk: $\min\limits_{k\in[M]} \left\{\frac{\eta_k}{k}\right\} \leq \eta(\boldsymbol{\alpha}) \leq \eta_M$

  - efficient and optimal methods for finding the optimal sampling probability

Thm 3. The OSP algorithm output $\boldsymbol{\alpha}_T$ satisfies $0 \leq \hat{\eta}(\boldsymbol{\alpha}_T) - \hat{\eta}(\boldsymbol{\alpha}^*) \xrightarrow[T\rightarrow\infty]{} 0$, for all initial conditions

Proposed Training Algorithm: BARRE



- robustness against ℓ_∞ norm-bounded adversaries – CIFAR-10
- BARRE: drastically improve **robust accuracy** while maintaining **complexity**

Thank You!

code available at https://github.com/hsndbk4/BARRE

Acknowledgement:
This work was supported by the Center for the Co-Design of Cognitive Systems (CoCoSys) funded by the Semiconductor Research Corporation (SRC) and the Defense Advanced Research Projects Agency (DARPA), and SRC's Artificial Intelligence Hardware (AIHW) program.