Certify or Predict: Boosting Certified Robustness with Compositional Architectures


Mark Müller

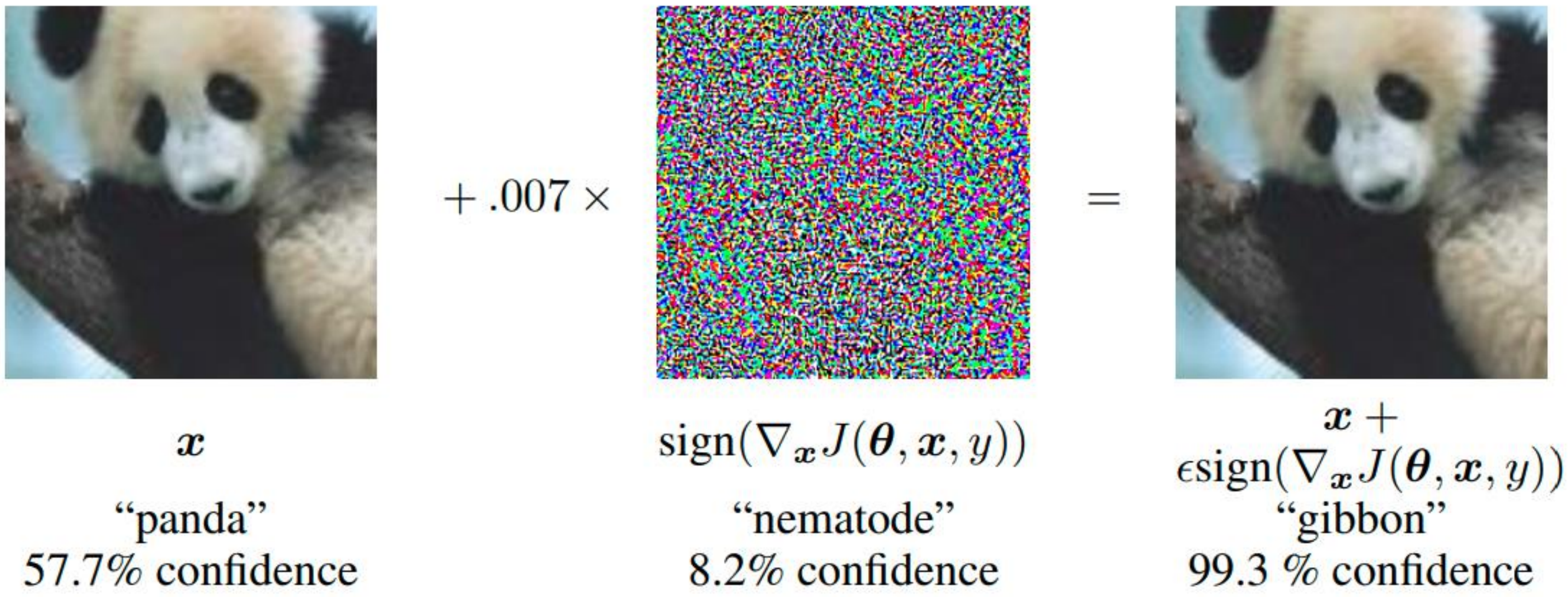

Mislav Balunovic


Martin Vechev

ETH zürich

SRILAB

Adversarial Examples

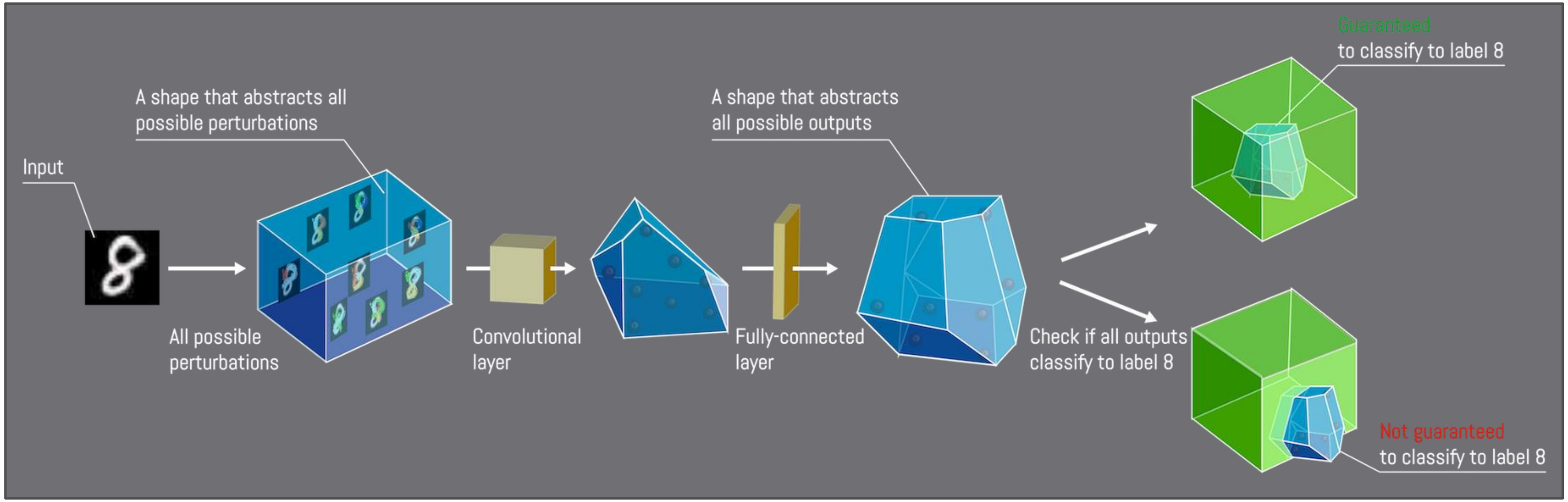


Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples.“ *arXiv preprint arXiv:1412.6572* (2014).

Neural Network Verification

- Robustness property:

$$argmax_i \; h(x)_i = argmax_i \; h(x')_i$$
$$\forall x' \in \mathrm{B}_\epsilon^\infty(x)$$



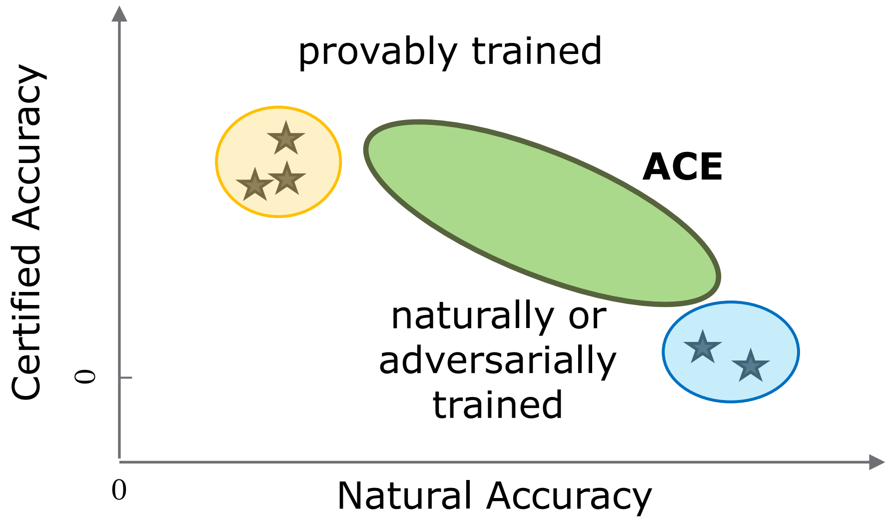
http://safeai.ethz.ch/img/overview.png - 20.03.21 - 9:00

Problem Statement

- Adversarial accuracy requires increased network capacity
- Verification gets increasingly difficult with network depth

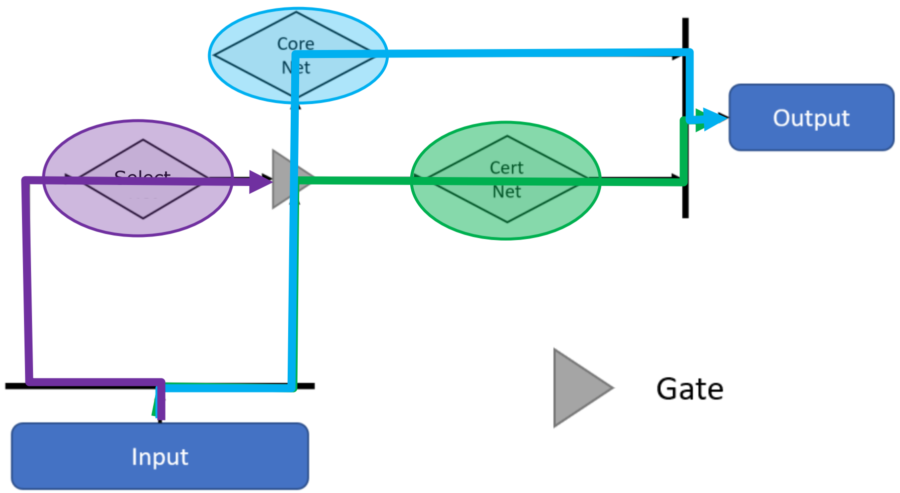
➔ Small, provably trained networks have low standard accuracy

➔ ACE: Compose networks with different strengths

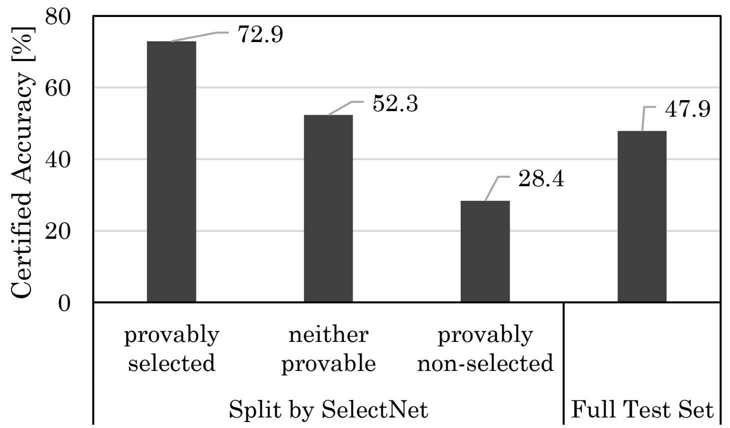


ACE – Compositional Architecture

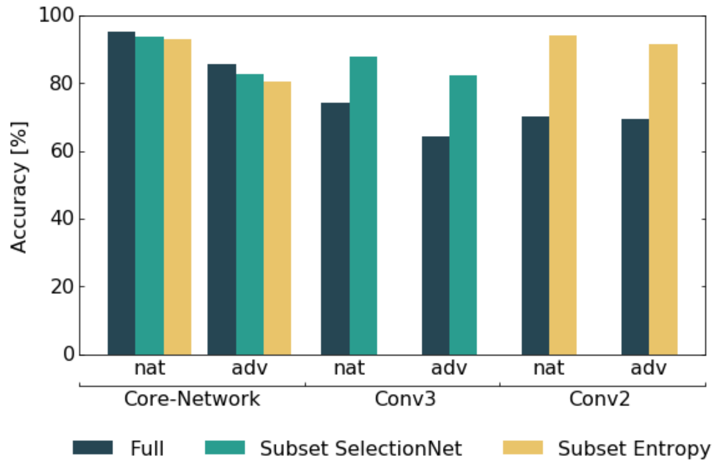
- For every sample decide whether to use core- or certification-network
- Key components:
  - Deep standard network
  - Shallow provable network
  - Selection mechanism
    - Train network to predict certification difficulty
    - Evaluate certification network entropy



Effectiveness of Selection

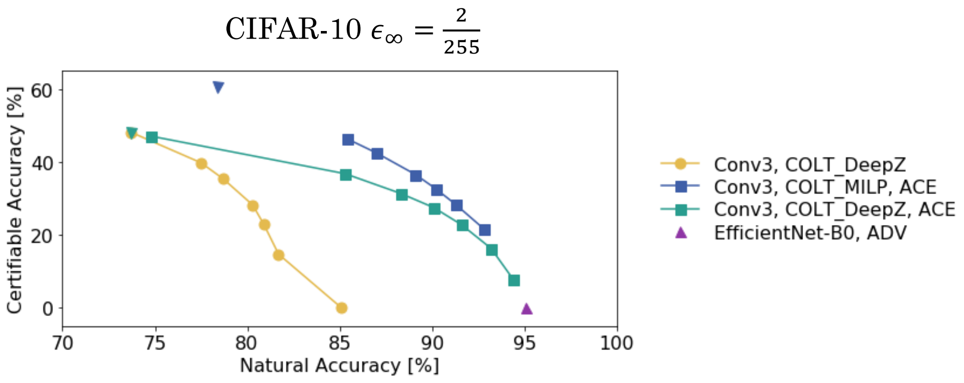


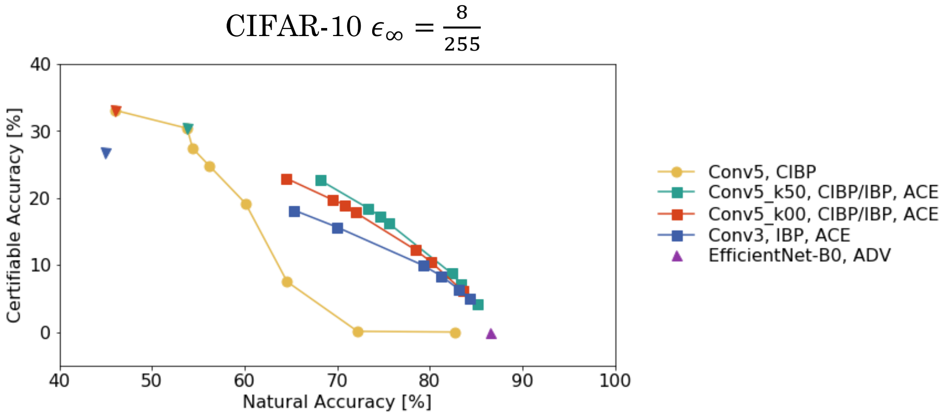
- Strong separation of samples based on certifiability
- Significantly increased accuracy of the certification-network on the selected sample subset

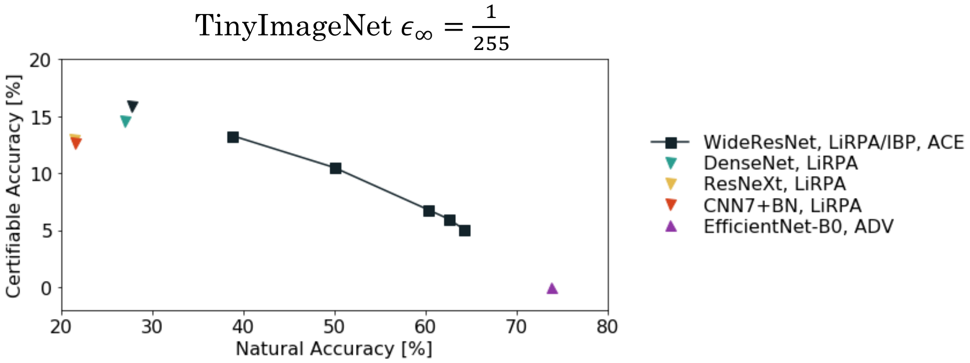


ACE Results

- Significant reduction in certified accuracy loss, for gains in natural accuracy
- Effect observed across:
  - Network architectures
  - Perturbation sizes
  - Datasets
  - Certification and training methods







Balunovic, Mislav, and Martin Vechev. "Adversarial training and provable defenses: Bridging the gap." *ICLR* 2019
Zhang, Huan, et al. "Towards stable and efficient training of verifiably robust neural networks." *arXiv:1906.06316* 2019
Xu, Kaidi, et al. "Automatic perturbation analysis for scalable certified robustness and beyond." NIPS 2020

Thank you for your attention!

Paper and Code:

https://www.sri.inf.ethz.ch/publications/mueller2021boosting

Poster Session 10