UCLA

Evaluation, Verification, and Training for Robust Machine Learning

Zhouxing Shi
Advisor: Cho-Jui Hsieh

February 23, 2023

Evaluation on the Robustness to Distribution Shifts

Challenges:

- Synthetic distribution shifts cannot represent natural distribution shifts
  - Construct natural benchmarks
- Out-of-distribution performance is often strongly correlated with in-distribution performance
  - Control for the performance on an in-distribution test set
- What if models are trained on different data?

Evaluation on the Robustness to Distribution Shifts

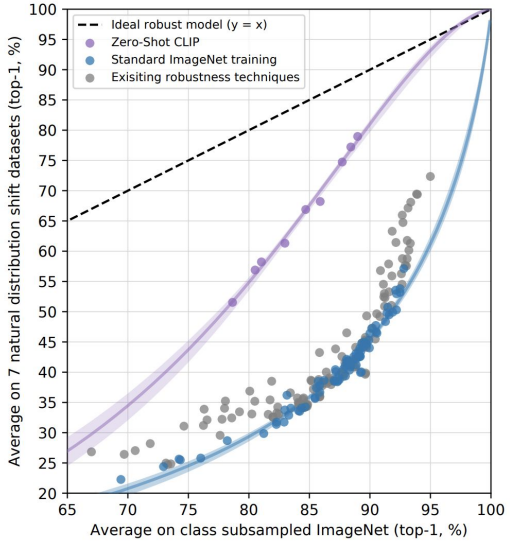
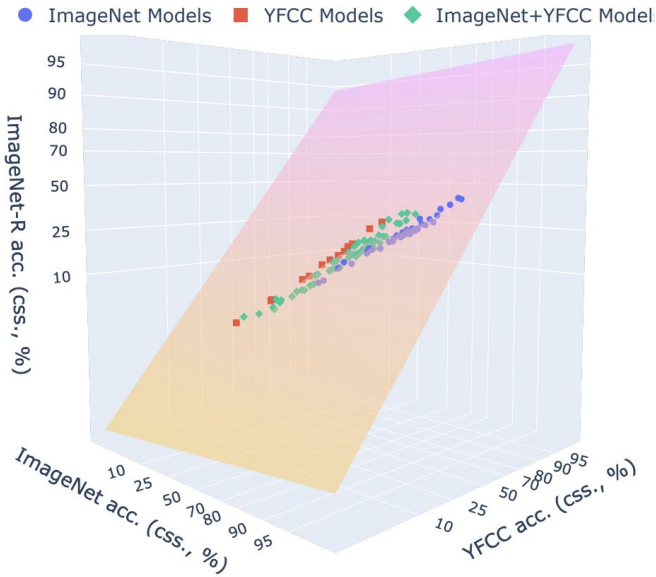


Figure from Radford et al., 2021



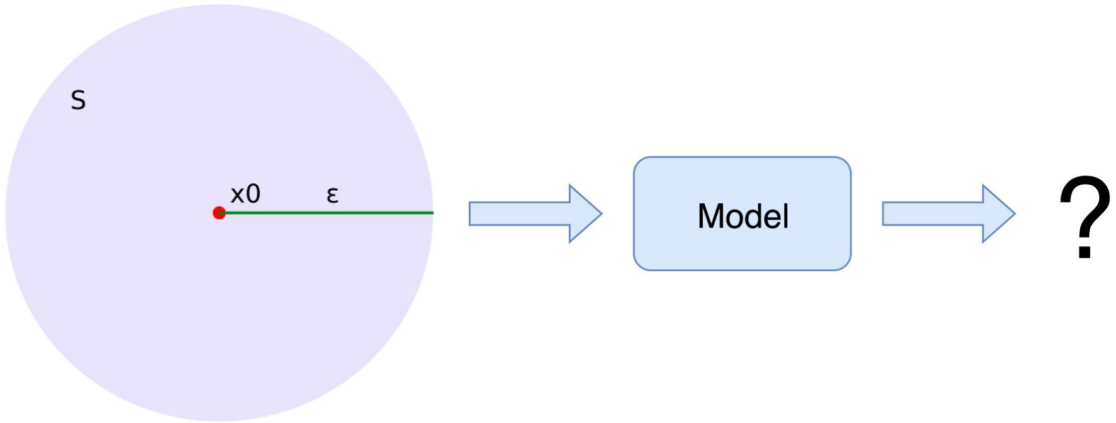
Exceptional effective robustness of CLIP in prior works with a **biased in-distribution test set**.

The effective robustness diminishes, under a **training data-aware evaluation** (ours).

Zhouxing Shi, Nicholas Carlini, Ananth Balashankar, Ludwig Schmidt, Cho-Jui Hsieh, Alex Beutel, Yao Qin. Effective Robustness against Natural Distribution Shifts for Models with Different Training Data. arXiv preprint: 2302.01381.

Neural Network Verification

To verify the behavior of a neural networks given **a range of inputs**:



Neural Network Verification

General and efficient frameworks for:

- Transformers
- General computational graphs
- Higher-order computational graphs
- ...

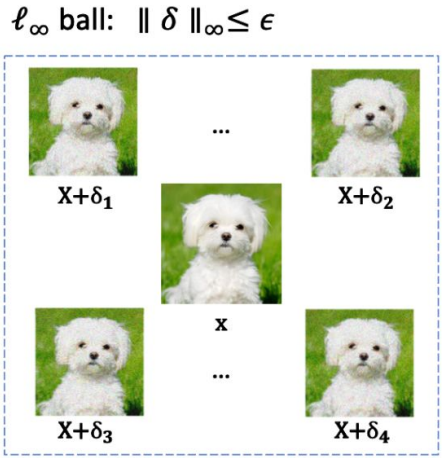
Towards solving real-world verification problems.

A library for automatic verification on PyTorch models:
https://github.com/Verified-Intelligence/auto_LiRPA

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, Cho-Jui Hsieh. Robustness Verification for Transformers. In ICLR 2020.
Kaidi Xu*, Zhouxing Shi*, Huan Zhang*, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, Cho-Jui Hsieh. Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond. In NeurIPS 2020.
Zhouxing Shi, Yihan Wang, Huan Zhang, Zico Kolter, Cho-Jui Hsieh. Efficiently Computing Local Lipschitz Constants of Neural Networks via Bound Propagation. In NeurIPS 2022.

Training Robust Neural Networks

Robust training with verified worst-case output:



DNN

Worst-case logits

-0.889	airplane ↑
0.7203	automobile ↑
-0.2943	bird ↑
2.3597	cat ↑
1.1594	deer ↑
4.032	**dog ↓**
0.2416	frog ↑
-0.878	horse ↑
1.4488	ship ↑
-1.332	truck ↑

SAFE

Zhouxing Shi*, Yihan Wang*, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh. Fast Certified Robust Training with Short Warmup. In NeurIPS 2021.
Yihan Wang*, Zhouxing Shi*, Quanquan Gu, Cho-Jui Hsieh. On the Convergence of Certified Robust Training with Interval Bound Propagation. In ICLR 2022.

Thanks!

Special thanks to Amazon for the support.