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).

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

Training Robust Neural Networks

Robust training with verified worst-case output:

\[ \ell_\infty \text{ ball: } \| \delta \|_\infty \leq \epsilon \]

\[ + \delta_1 \quad + \delta_2 \quad + \delta_3 \quad + \delta_4 \]

\[
\begin{array}{c|c}
\text{Worst-case logits} & \\
\hline
-0.889 & \text{airplane \textup{↑}} \\
0.7203 & \text{automobile \textup{↑}} \\
-0.2943 & \text{bird \textup{↑}} \\
2.3597 & \text{cat \textup{↑}} \\
1.1594 & \text{deer \textup{↑}} \\
4.032 & \text{dog \textup{↓}} \\
0.2416 & \text{frog \textup{↑}} \\
-0.878 & \text{horse \textup{↑}} \\
1.4488 & \text{ship \textup{↑}} \\
-1.332 & \text{truck \textup{↑}} \\
\end{array}
\]

Thanks!

Special thanks to Amazon for the support.