# A Hierarchical Assessment of Adversarial Severity

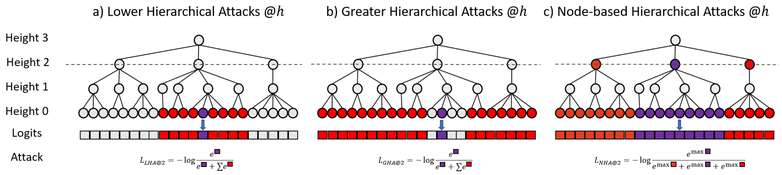
Guillaume Jeanneret*, Juan C. Pérez^, Pablo Arbelaez*
*Center for Research and Formation in Artificial Intelligence, Universidad de los Andes
^King Abdullah University of Science and Technology (KAUST)

2021 ICCV OCTOBER 11-17 VIRTUAL

## Introduction:

- Traditional adversarial attacks ignore the rich semantic structure of the label space.
- **Aim:** exploring the notion of severity produced by adversarial attacks, *i.e.* the semantic error induced by the adversary.
- **Goal**: protect models against attacks that induce severe semantic mistakes.
- In this paper, we propose:
  1) We introduce a new set of Hierarchy-aware Attacks to diminishing the accuracy and increasing hierarchical errors.
  2) We provide the first assessment on adversarial severity, building upon the work of [1], on a highly challenging, large-scale, long-tailed, and hierarchically-structured benchmark: iNaturalist-H.
  3) We present CHAT, a defense to diminish the severity and the precision of adversarial attacks.

## Hierarchical-Aware Attacks

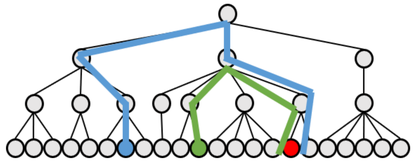



- *Lower Hierarchical Attack*: Low Severity - High Success Rate.
- *Greater Hierarchical Attack*: Medium Severity and Sucess Rate.
- *Node-based Hierarchical Attack:* Great Severity - Low Sucess Rate.

[1] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

## A New Benchmark:

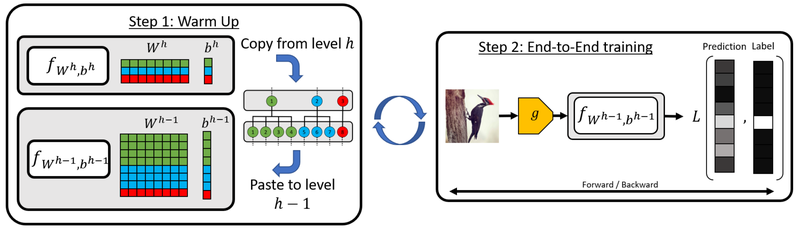
- The severity is the hierarchical distance between the true label and the prediction.



- We extend the notion of severity for assessing the impact of adversarial examples.
- Assessment of the fragility with our new hierarchical-aware attacks.

## Curriculum for Hierarchical Adversarial Training

- Learning protocol inspired by human psychology.
- Iteratively learn the nodes of the tree from highest levels until de leaves.
- Two stages: the Warm-Up and the End-to-End Training.




## Ablation Studies

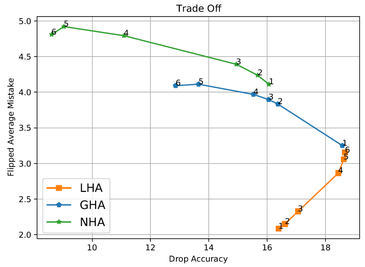
- Free Adversarial Training

| $\epsilon$ | $m$ | $\alpha$ | C | Clean Acc | Clean AM | PGD50 Acc | PGD50 AM |
|---|---|---|---|---|---|---|---|
| 4 | 6 | 6 | | 31.40 | 3.21 | 12.33 | 3.20 |
| 4 | 8 | 6 | ✓ | 32.84 | 3.04 | 13.36 | 3.06 |
| 6 | 6 | 4 | | 24.87 | 3.44 | 7.13 | 3.44 |
| 6 | 8 | 4 | ✓ | 27.19 | 3.28 | 8.32 | 3.29 |
| 8 | 6 | 6 | | 19.65 | 3.76 | 4.31 | 3.71 |
| 8 | 8 | 6 | ✓ | 23.29 | 3.49 | 6.07 | 3.49 |

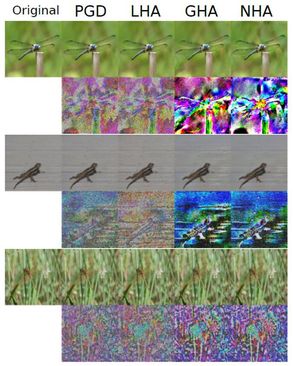
- TRADES

| $\epsilon$ | C | Clean Acc | Clean AM | PGD Acc | PGD AM |
|---|---|---|---|---|---|
| 4 | | 23.90 | 3.33 | 11.06 | 3.34 |
| 4 | ✓ | 29.23 | 3.03 | 13.20 | 3.08 |
| 6 | | 21.26 | 3.49 | 7.04 | 3.50 |
| 6 | ✓ | 27.91 | 3.14 | 9.05 | 3.19 |
| 8 | | 19.94 | 3.59 | 4.01 | 3.61 |
| 8 | ✓ | 29.84 | 3.11 | 4.47 | 3.22 |

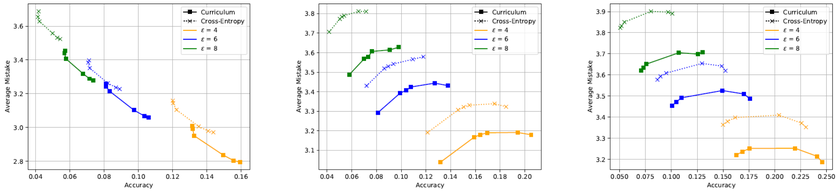
- Induced Mistake and Accuracy Drop Trade-Off




- Visualizing Adversarial Examples




## Main Results




(a) LHA Assessment (b) GHA Assessment (c) NHA Assessment

## Conclusions

- We device a new set of hierarchical-aware attacks that induces diverse effects on a classification network output.
- We explored a new dimension on the classical evaluation of adversarial examples: the adversarial severity.
- We showed the effectivity of including the hierarchical nodes into the learning protocol enhances the adversarial robustness.
- Our study opens the door to studying the unexplored phenomena of adversarial severity

**Project Webside**