WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

KTH VETENSKAP OCH KONST

ICML International Conference On Machine Learning

Improving Adversarial Robustness of Attribution via Implicit Regularization

Amir Mehrpanah, Matteo Gamba, and Hossein Azizpour
ICML 2026 — KTH Royal Institute of Technology

Presentation Agenda

- What is explainability and why it matters?

Presentation Agenda

- What is explainability and why it matters?
- Our research question in this work

Presentation Agenda

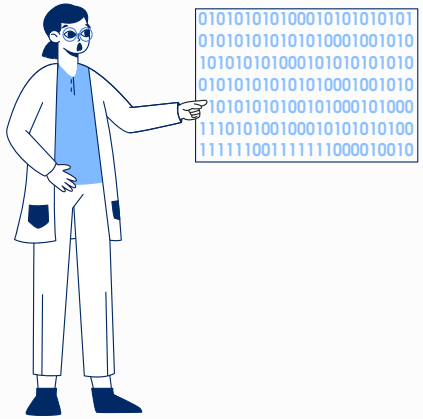
- What is explainability and why it matters?
- Our research question in this work
- Our proposed solution

What is Explainability?

The Machine Learning Era

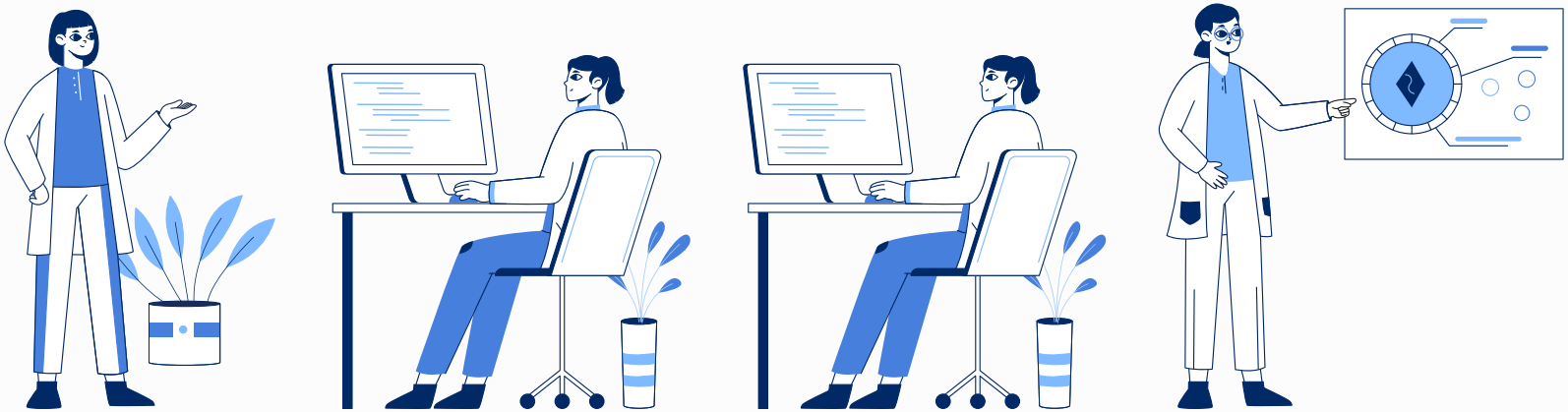
Problem

Machine Leaning



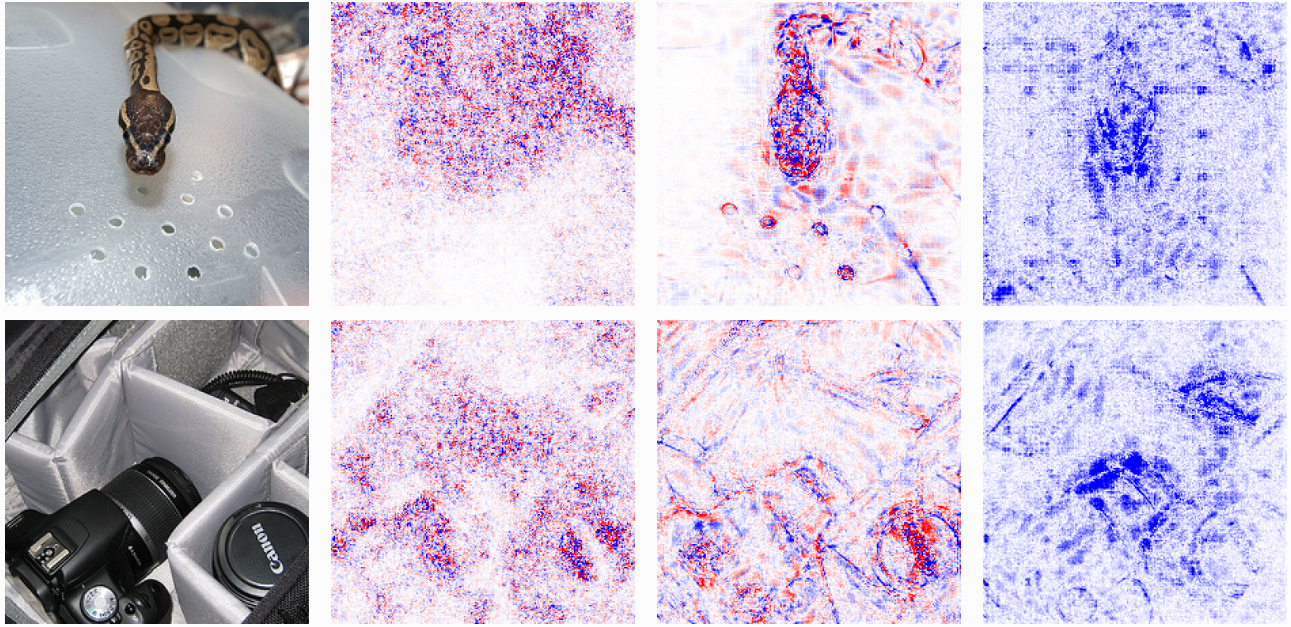
Black-box Outputs

Explainable Machine Learning



Problem | Machine Leaning | Explainable ML | Explainable Outputs

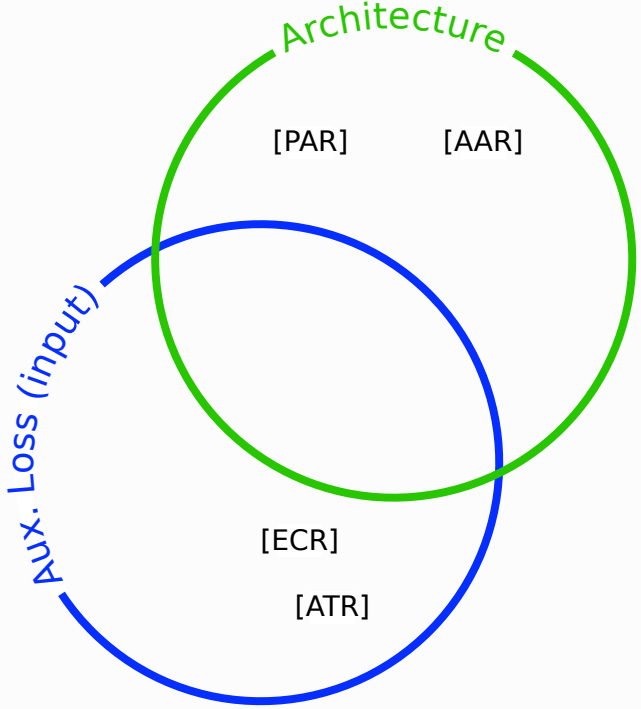
Attributions Are Unstable



Previous Solutions

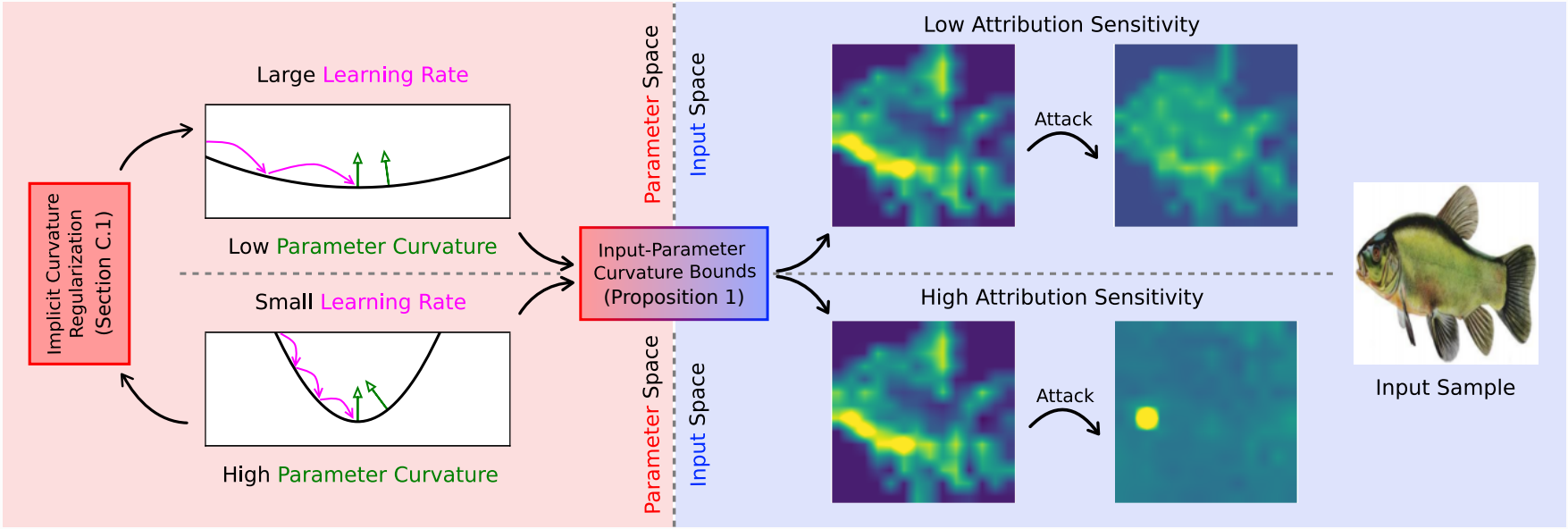
- Two general ways to improve attribution robustness:
  - Auxiliary loss
  - Architectural modification

- PAR/AAR -> Post- or Ante-hoc Architectural Regularization
- ATR -> Adversarial Training Regularization
- ECR -> Explicit Curvature Regularization



Our Findings

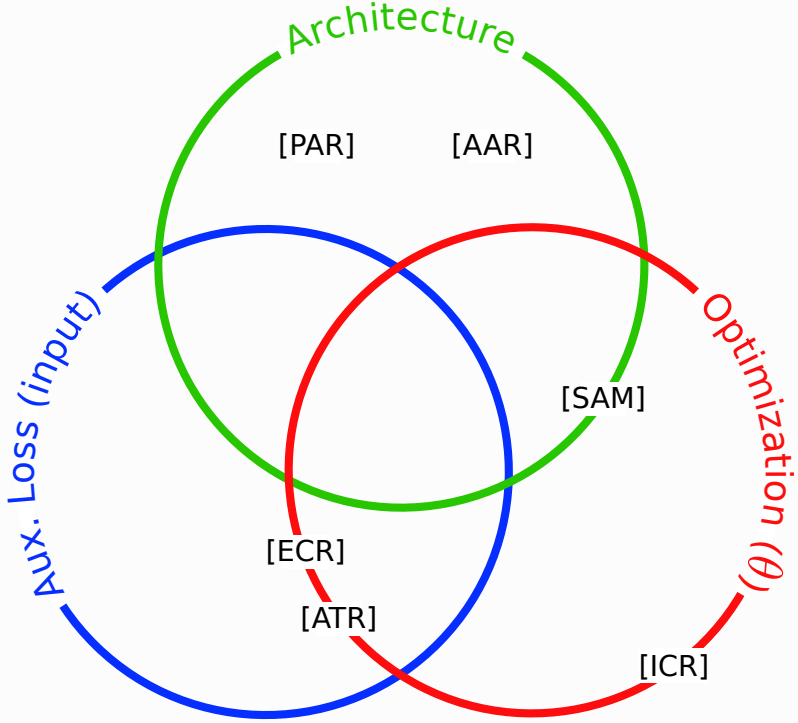
Our Findings



Our Findings

- A new aspect in attribution robustness is the effect of training dynamics

- SAM -> Sharpness Aware Minimization
- ICR -> Implicit Curvature Regualrization



Conclusion

Conclusion

- We can see both theoretically and empirically that the training dynamics can affect attribution robustness.

Conclusion

- We can see both theoretically and empirically that the training dynamics can affect attribution robustness.
- This phenomenon allows us to control the attribution robustness via controlling parameter curvature (in expectation).

Conclusion

- We can see both theoretically and empirically that the training dynamics can affect attribution robustness.
- This phenomenon allows us to control the attribution robustness via controlling parameter curvature (in expectation).
- Therefore, one can avoid training with costly auxiliary losses that depend on input curvature and architectural modifications.


KTH
VETENSKAP
OCH KONST