THE OHIO STATE UNIVERSITY

imageomics BIO ML

ICLR

A Simple Interpretable Transformer for Fine-grained Image Classification and Analysis


Dipanjyoti Paul


Arpita Chowdhury


Xinqi Xiong


Feng-Ju Chang

...


Charles Stewart


Tanya Berger-Wolf


Yu Su


Wei-Lun Chao

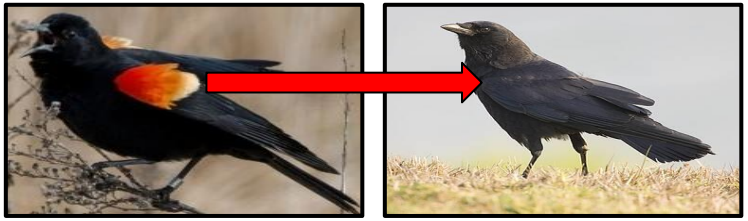
amazon

VT

PRINCETON UNIVERSITY

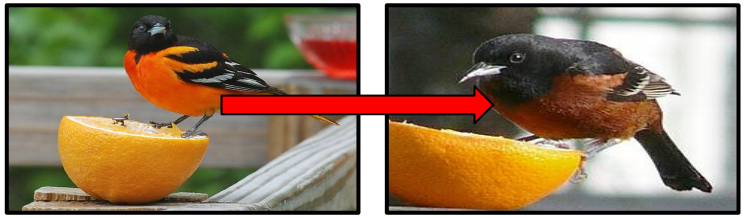
RENSSELAER POLYTECHNIC INSTITUTE 1824

What Kind of Interpretation we are looking for?



Red-winged Blackbird

American Crow



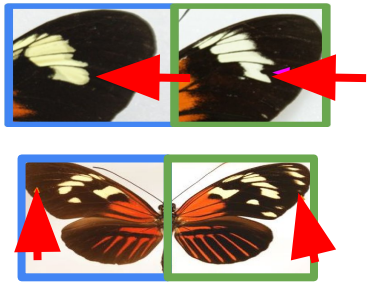
Baltimore Oriole

Orchard Oriole

Source: *https://www.allaboutbirds.org/news*

Heliconius melpomene

Heliconius elevatus



Source: *https://www.cliniquevetodax.com/Heliconius/index.html*

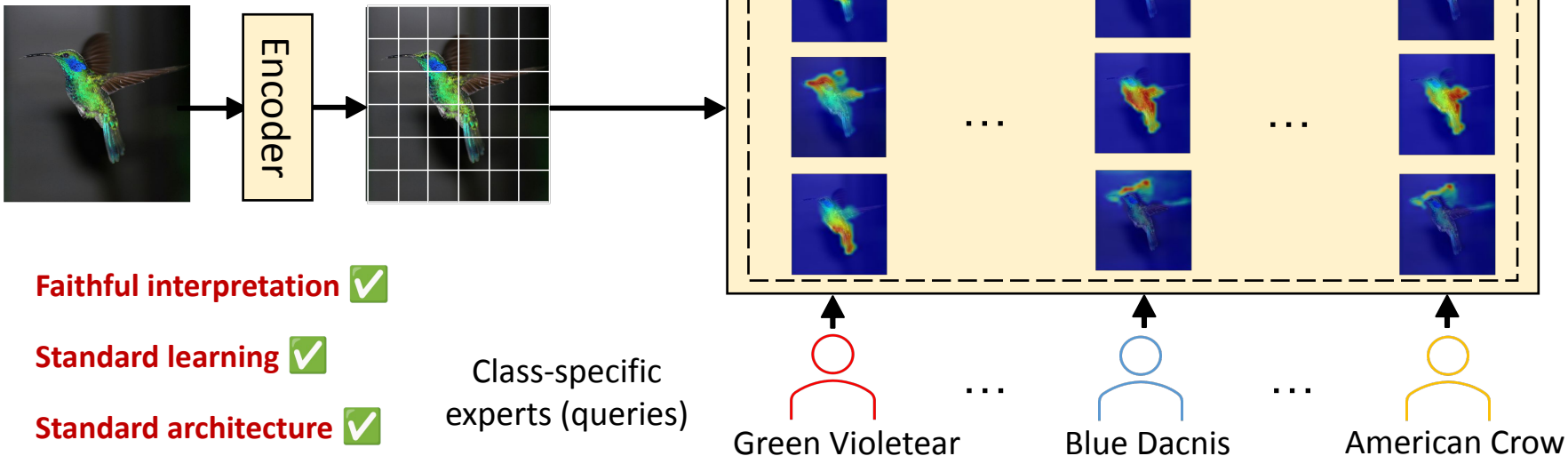
Motivation (INterpretable TRansformer)

- Can we obtain interpretability via **standard neural network architectures** and **standard learning algorithms**?

- Can we have a **faithful** interpretation?

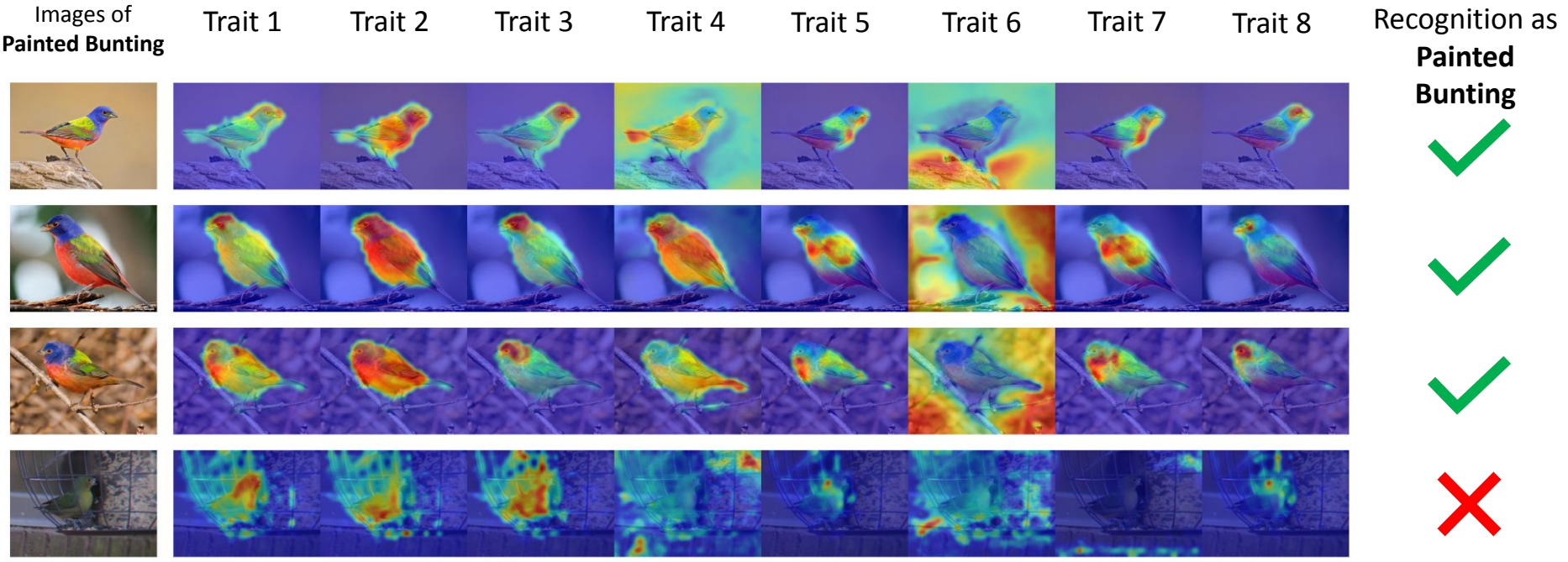
INTR – INterpretable TRansformer for Trait Discovery

INTR: inspired by how biologists looking for traits to differentiate species



- **Faithful interpretation** ✅
- **Standard learning** ✅
- **Standard architecture** ✅

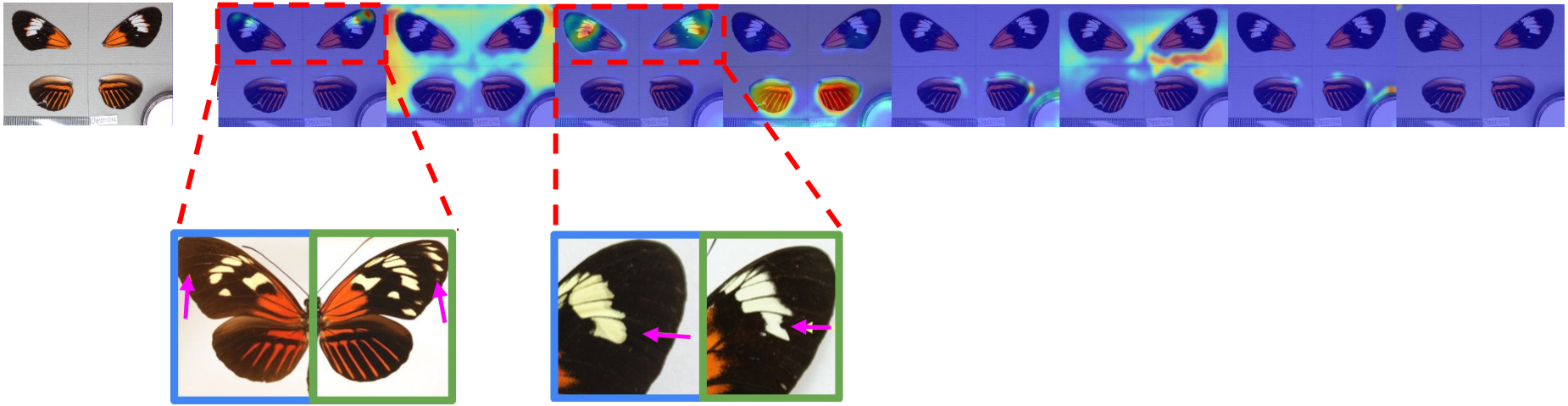
INTR can consistently localize traits



INTR can localize traits for fine-grained species

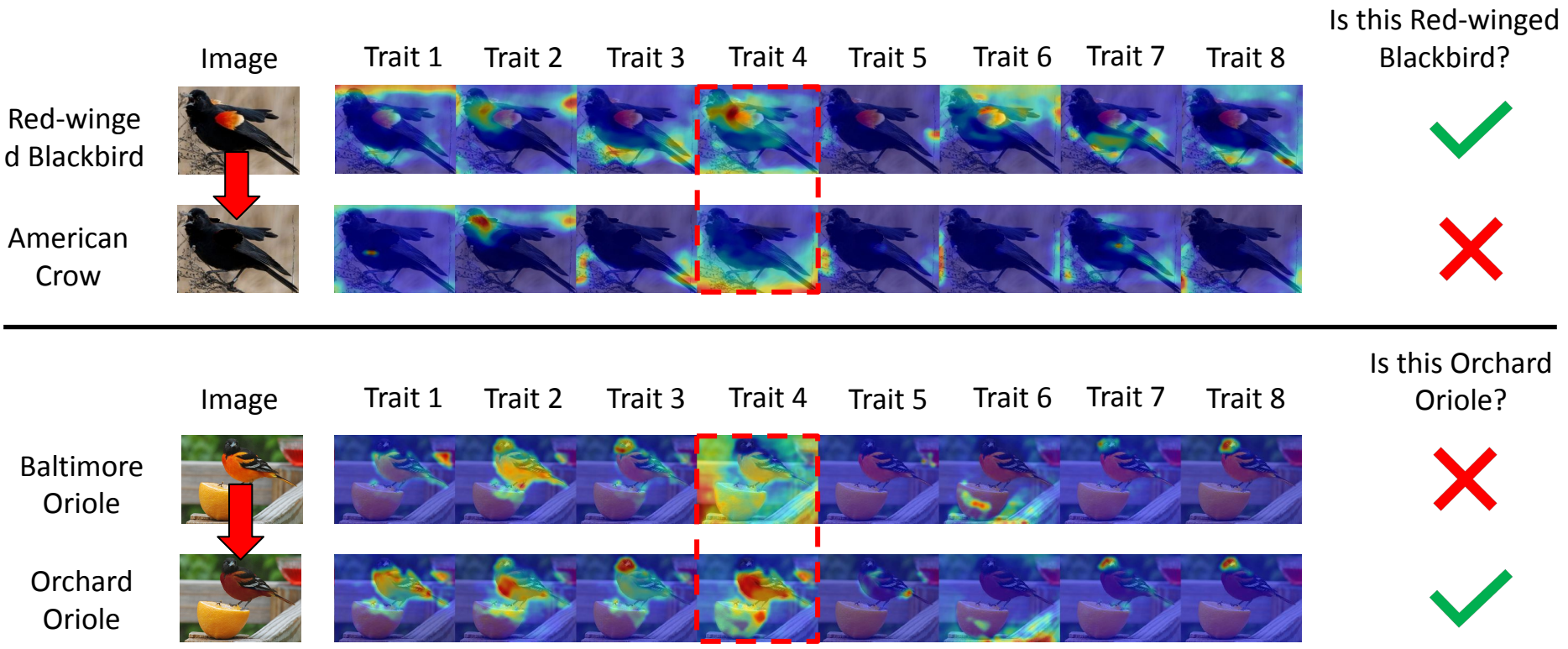
Image of **Heliconius melpomene**

Trait 1 | Trait 2 | Trait 3 | Trait 4 | Trait 5 | Trait 6 | Trait 7 | Trait 8



Source: *https://www.cliniquevetodax.com/Heliconius/index.html*

INTR can localize traits for fine-grained species



Summary

- INTR can able to detect ML traits automatically and provides supporting evidence for its decision
- INTR is build upon standard NN architecture and learning
- Applicability of INTR has been shown on eight datasets
- INTR can offer a new way of thinking about interpretable machine learning

Thank You!

INTR Github

INTR Paper