Dividing and Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat

Shantanu Ghosh (BU), Ke Yu (Pitt), Forough Arabshahi (Meta), Kayhan Batmanghelich (BU)

Meta | BOSTON UNIVERSITY | BATMAN LAB | University of Pittsburgh | ICML International Conference On Machine Learning

TLDR: Extracting a mixture of interpretable models from a BlackBox to provide instance specific concept-based explanations using First-order logic (FOL).

Post hoc explanation

Pros

- Does not alter the Black box.

Cons

- Inconsistent explanations.
- No recourse.

Interpretable by design

Pros

- Support concept intervention.

Cons

- Harder to train.
- Sub par performance.

How to blur this gap?

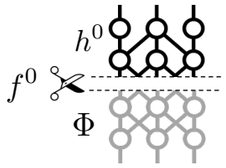
Desirable properties

- Does compromise the performance.
- Can be intervened to fix the misclassification

Design choices

- Carve interpretable models from Blackbox.
- Concept based
- First order logic for concept interaction

Assumptions

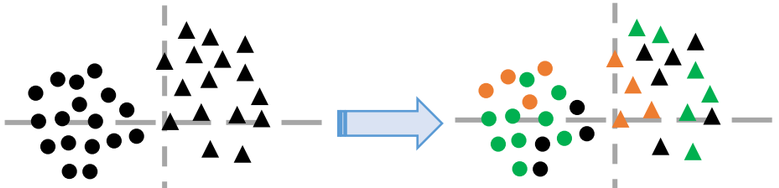


$\{ \mathcal{X} \mid \mathcal{C} \mid \mathcal{Y} \}$

thin ears
shortened muzzle
round feet
.....

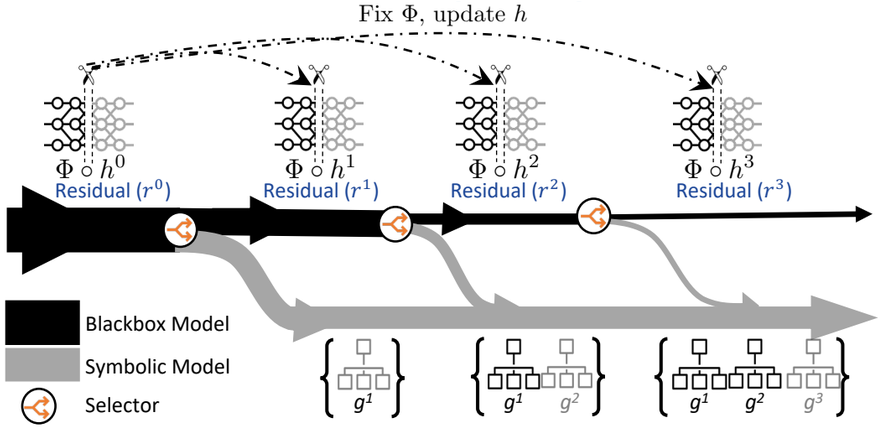
Dog

Carve out interpretable models from Black box



Each color denotes an expert

Route Interpret Repeat



Each g is E-LEN [Barberio et al.], constructing FOL

* SelectiveNet [Geifman et al.] optimization
* Continue till at least 90% samples covered

Capturing heterogenous explanations

* Extracted from ViT-based BlackBox

Expert 1

Olive sided Flycatcher ↔breast_color_grey ∧ tail_pattern_solid

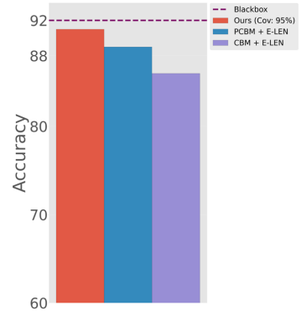
Expert 2

Olive sided Flycatcher ↔underparts_color_grey∧ wing_color_grey

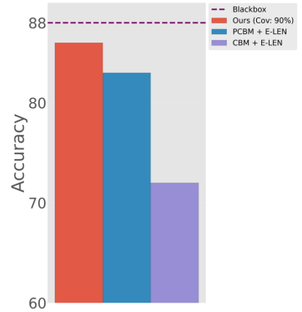
Final residual (Unexplained)

Not compromising performance

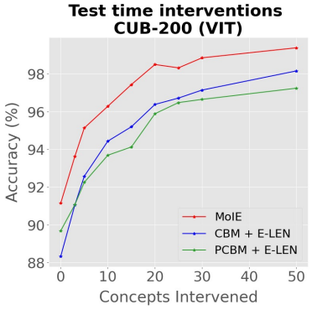
CUB-200 with ViT

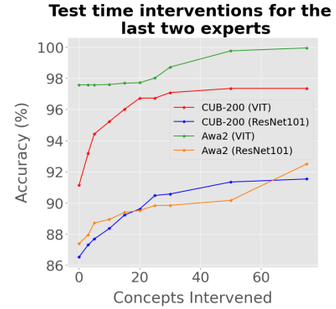


CUB-200 with ResNet101



Test time interventions





Also in our paper,

+ we experiment with a diverse set of datasets and architectures
+ we achieve higher concept completeness scores
+ VIT-based experts compose less concepts than CNN-based
+ we eliminate shortcut learning problem (SCIS w)
+ we efficiently transfer the experts to new domain (IMLH w)