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)

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.
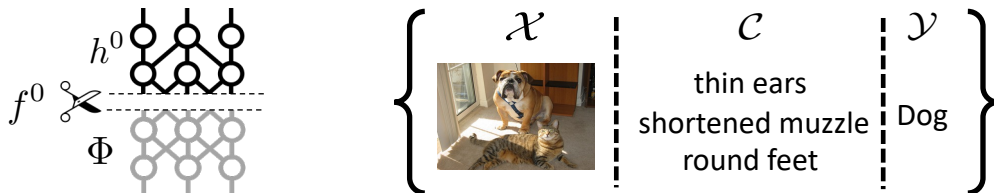- Sub par performance.

> How to blur this gap?

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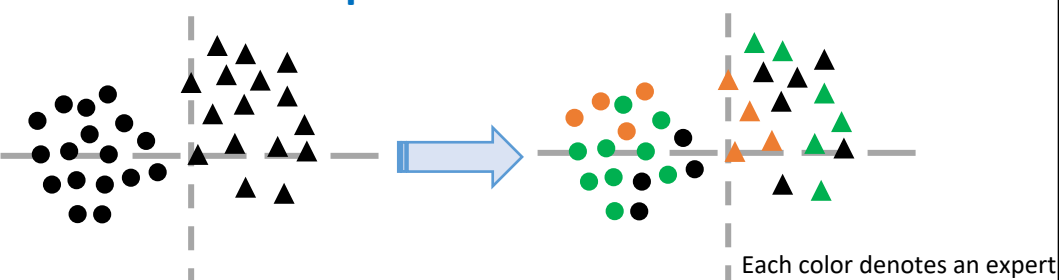- Can be intervened to fix the misclassification

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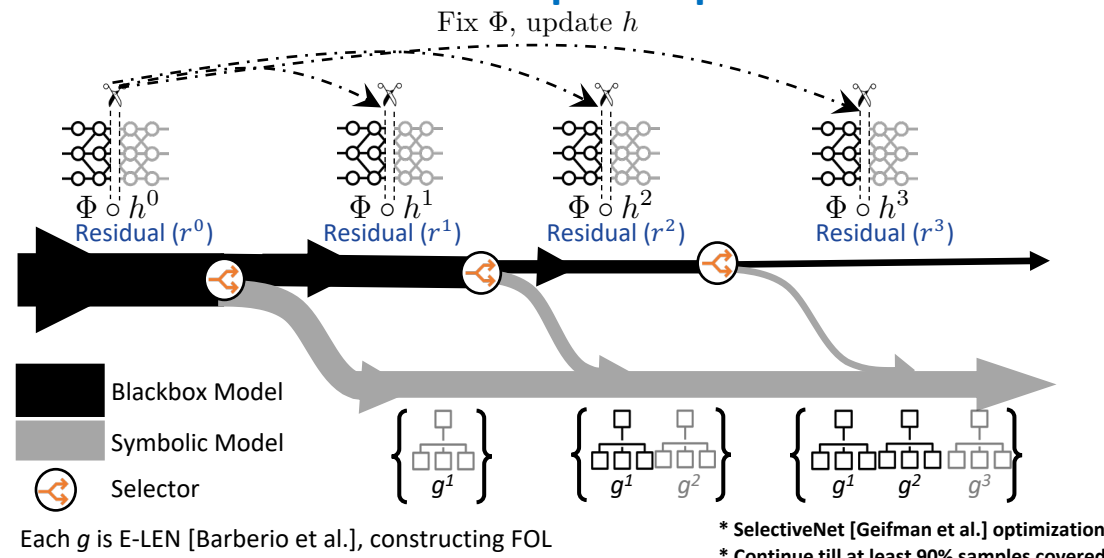- First order logic for concept interaction

Assumptions

h^0

f^0

Φ

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

thin ears
shortened muzzle
round feet
......

Dog

Carve out interpretable models from Black box

Each color denotes an expert

Route Interpret Repeat

Fix Φ, update h

$\Phi \circ h^0$ Residual (r^0)
$\Phi \circ h^1$ Residual (r^1)
$\Phi \circ h^2$ Residual (r^2)
$\Phi \circ h^3$ Residual (r^3)

- ⬛ Blackbox Model
- ⬜ Symbolic Model
- ⊕ Selector

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

g^1 | g^1 g^2 | g^1 g^2 g^3

* 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

Legend:
- Blackbox
- Ours (Cov: 95%) / Ours (Cov: 90%)
- PCBM + E-LEN
- CBM + E-LEN

Test time interventions

Test time interventions CUB-200 (VIT)

Legend:
- MoIE
- CBM + E-LEN
- PCBM + E-LEN

Test time interventions for the last two experts

Legend:
- CUB-200 (VIT)
- CUB-200 (ResNet101)
- Awa2 (VIT)
- Awa2 (ResNet101)
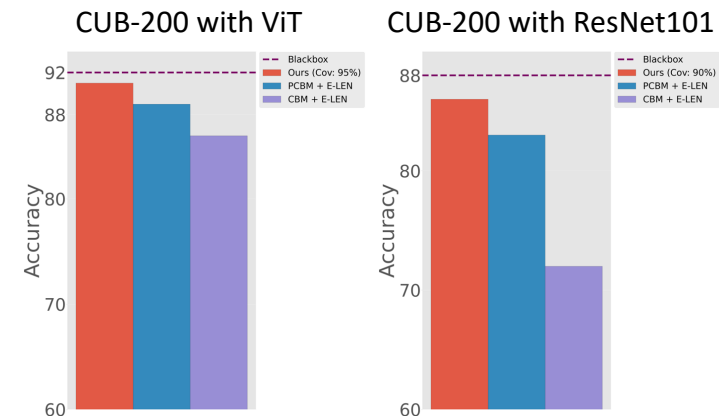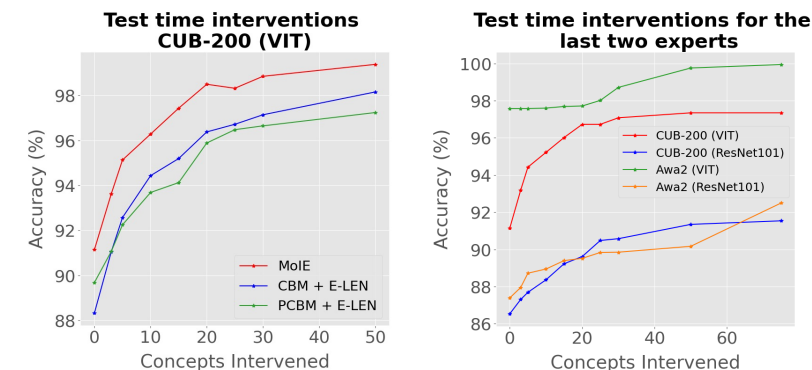
Also in our paper,

+ 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)