Divide and Conquer: Carving Out Symbolic Models out of BlackBox for More Efficient Domain Adaptation



Shantanu Ghosh[1], Ke Yu[2], Kayhan Batmanghelich[1]

[1]BU ECE, [2]Pitt ISP

Desiderata of Explainability in Healthcare

Trust

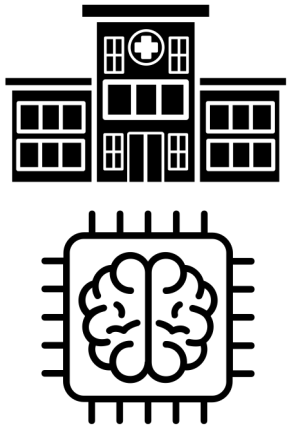
Causality

Transferability

Informativeness

Ethical Decision making

Training



Deployment

Zachary Lipton; 2017, “The Mythos of Model Interpretability”

Desiderata of Explainability in Healthcare

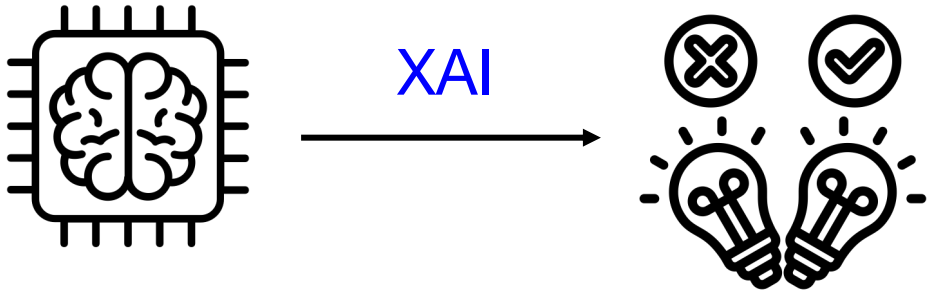
Trust

Causality

Transferability

Informativeness

Ethical Decision making



Obtaining genetics insights from deep learning via explainable artificial intelligence

Gherman Novakovsky[1,2,7], *Nick Dexter*[3,4,7], *Maxwell W. Libbrecht*[4,8]✉, *Wyeth W. Wasserman*[1,8]✉ *and Sara Mostafavi*[5,6,8]✉

Zachary Lipton; 2017, “The Mythos of Model Interpretability”

Desiderata of Explainability in Healthcare

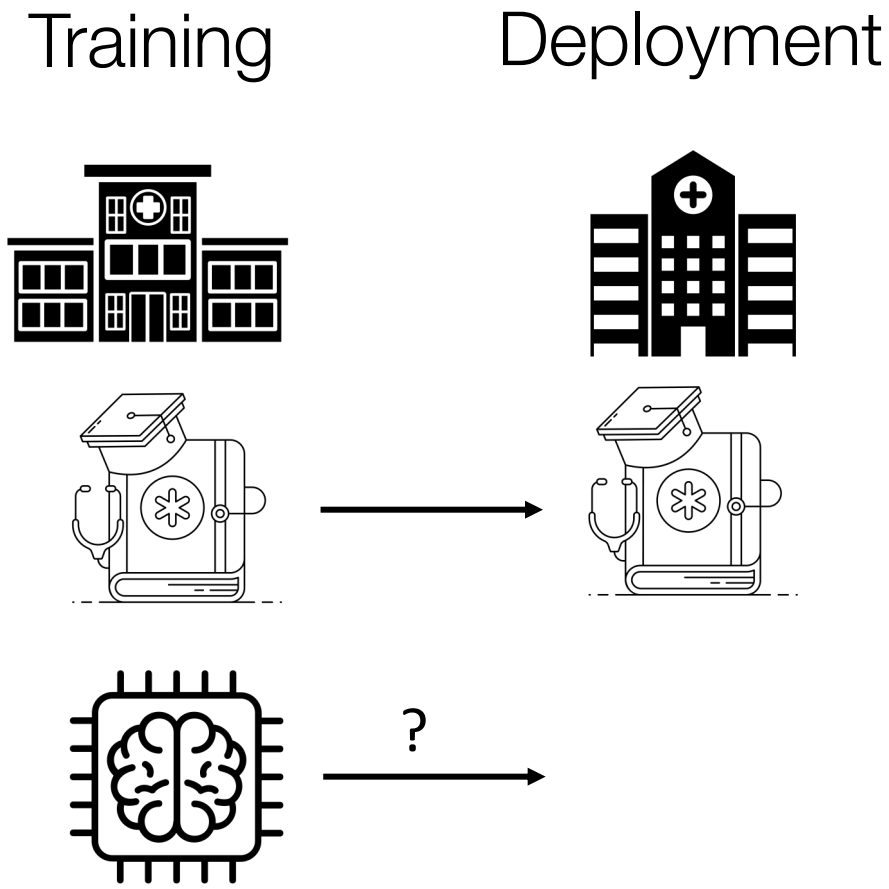
Trust

Causality

Transferability

Informativeness

Ethical Decision making



Zachary Lipton; 2017, “The Mythos of Model Interpretability”

Desiderata of Explainability in Healthcare

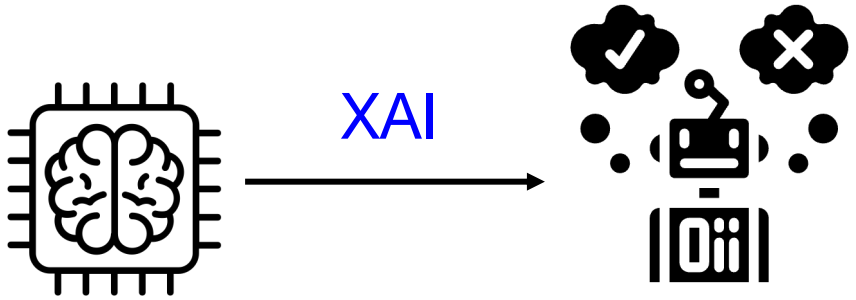
Trust

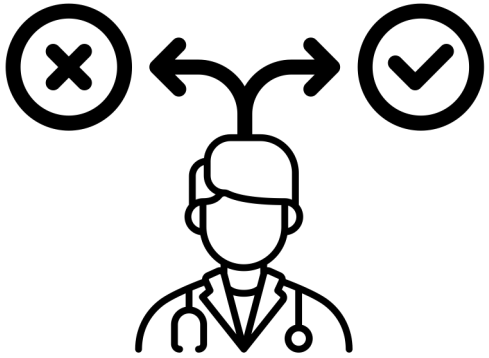
Causality

Transferability

Informativeness

Ethical Decision making





Zachary Lipton; 2017, “The Mythos of Model Interpretability”

Desiderata of Explainability in Healthcare

Trust

Causality

Transferability

Informativeness

Ethical Decision making

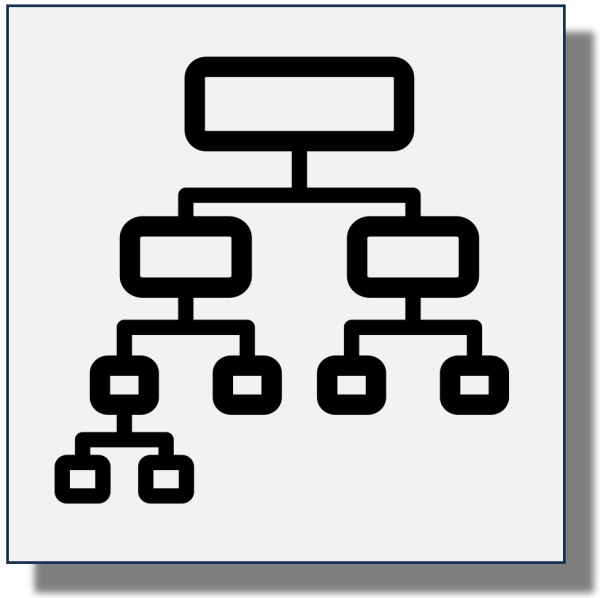
Training

Deployment

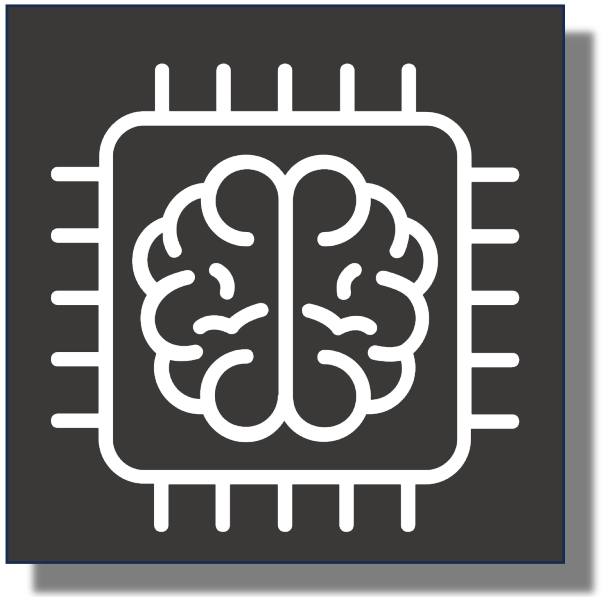
Zachary Lipton; 2017, “The Mythos of Model Interpretability”

General Design Choices

Transparent Models

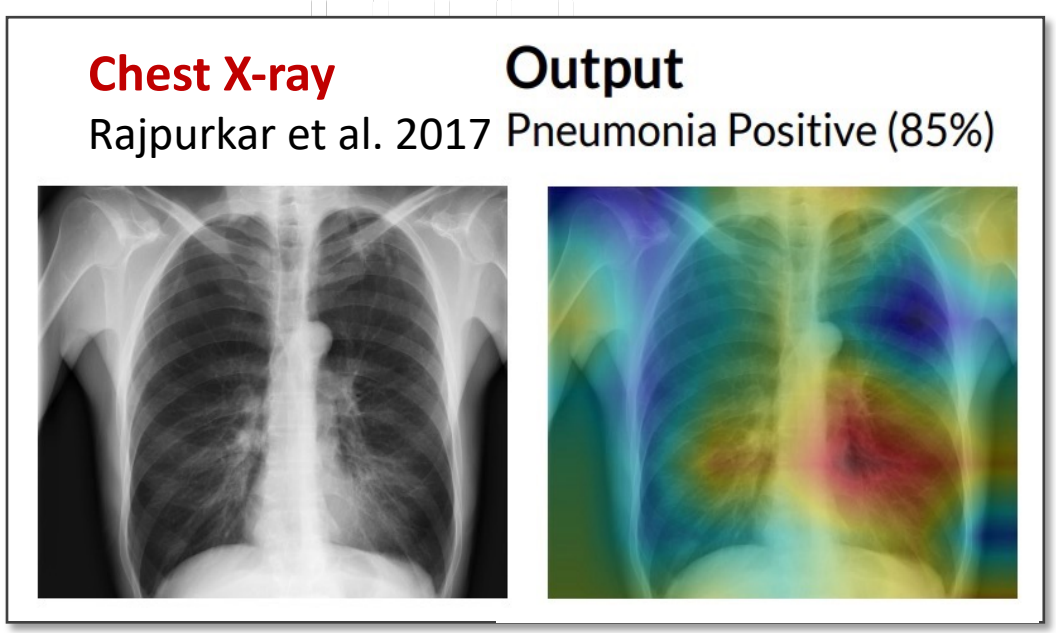


Post-hoc Explanation



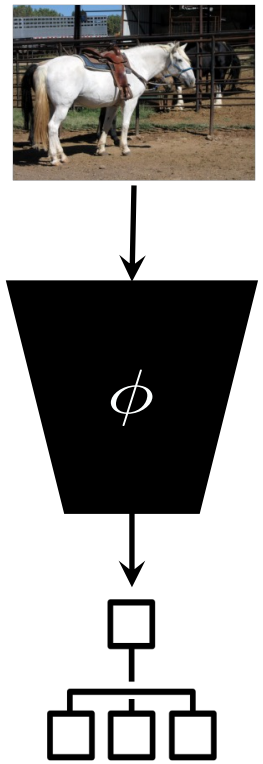
General Design Choices

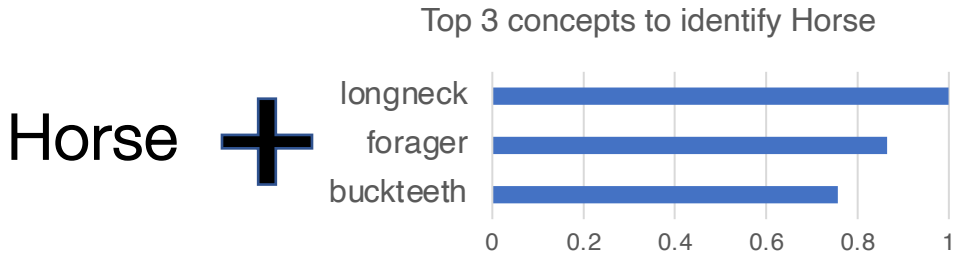
Post-hoc Explanation



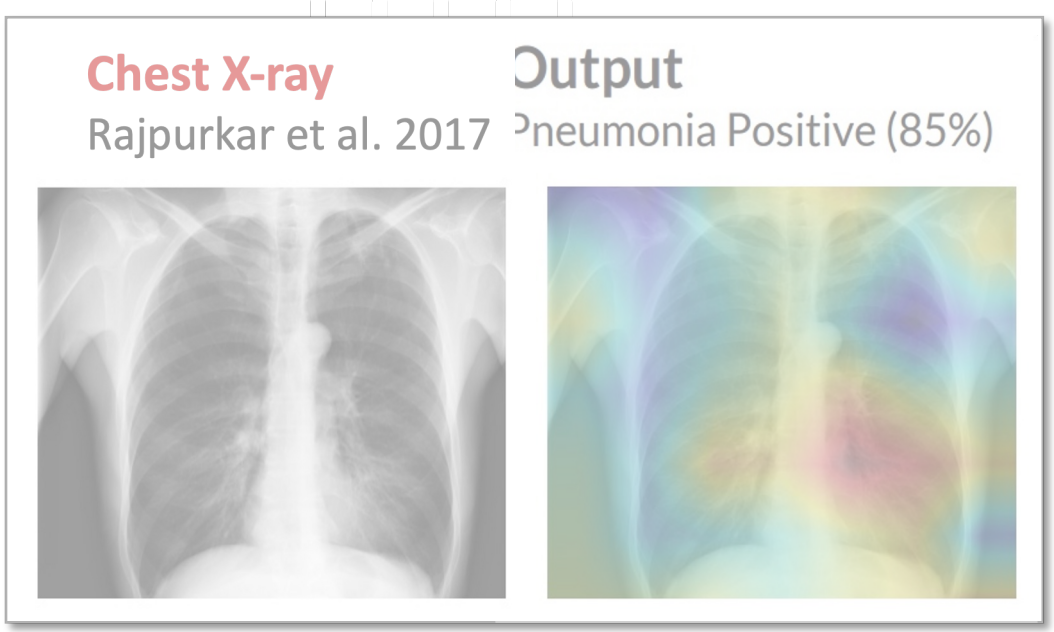
General Design Choices

Transparent Models





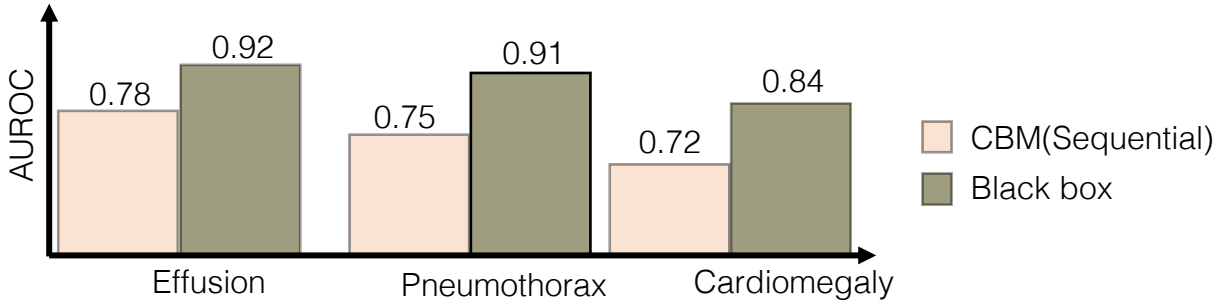
Post-hoc Explanation



General Design Choices

Transparent Models

Support concept intervention
Limited design choices
Difficult to get high-performance



Post-hoc Explanation

Does not alter the Black box
Leaves too much to human
Inconsistent explanations
No recourse

Stop explaining Black Box. Rudin et al. arXiv 2019.
Sanity Checks for Saliency Maps. Adebayo et al. Neurips 2018.

Desiderata

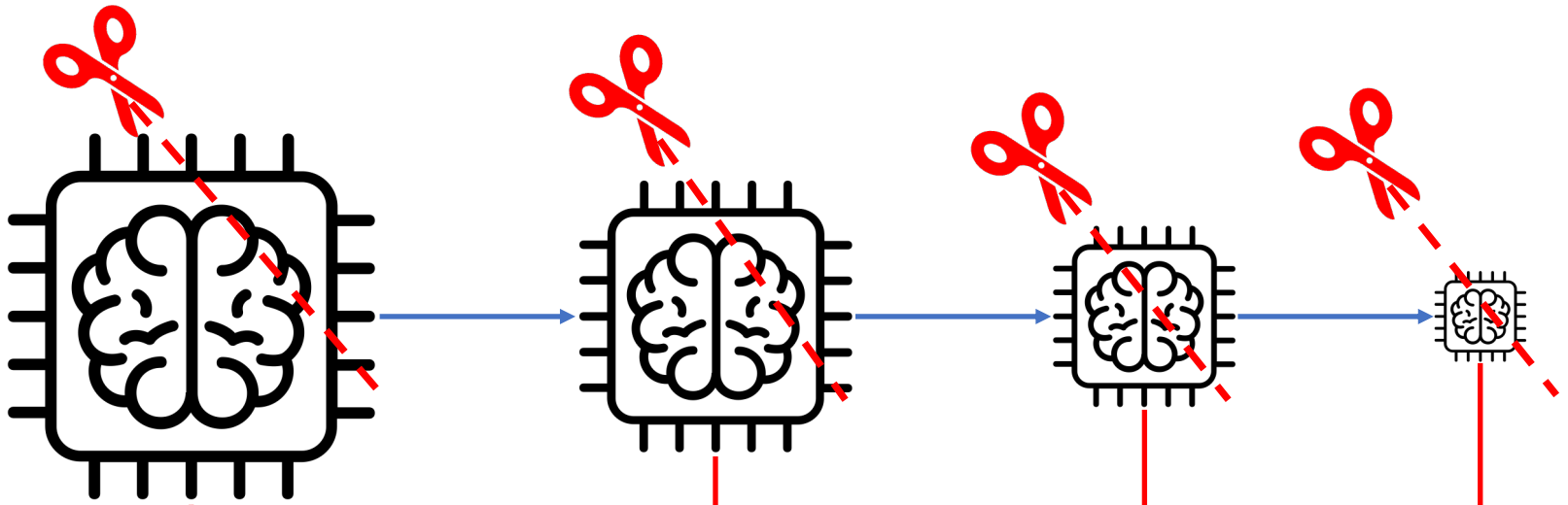
- Not compromising performance
- Being able to intervene (fix) undesirable properties
- Transferability: Clinical rules are transferrable between domains
  - The transparent should look like clinical rules

$$p \implies q$$

- **R1**: fasting p-Glucose level > 126 mg/dL on two separate tests ⟹ may be diabetes.
- **R2**: 2-hour p-Glucose level during an Oral Glucose Tolerance Test > 200 mg/dL ⟹ diabetes (probabilistic).
- **R3**: random p-Glucose level > 200 mg/dL ∧ hyperglycemia (frequent urination ∧ increased thirst ∧ unexplained weight loss ⟹ diabetes (probabilistic).
- **R4**: A1C test > 6.5% (visit 1) ∧ A1C test > 6.5% (visit 2) ⟹ diabetes (probabilistic)

General Idea

Blackbox



Transparent
(symbolic)

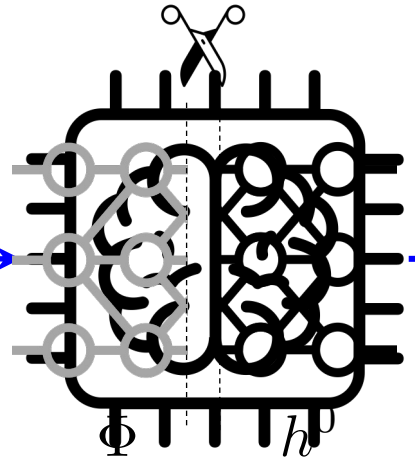
$p \implies q$

Problem Set Up

x

c

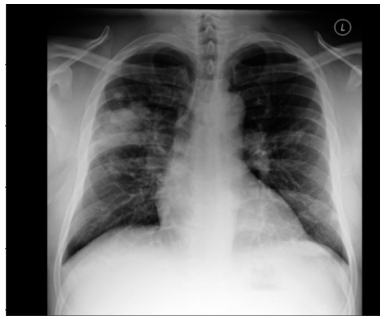
thin ears
shortened muzzle
round feet
….



y

Dog

Problem Set Up

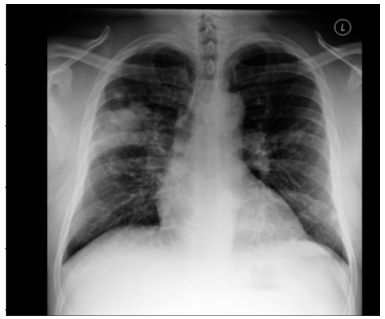


+

Report:

Right upper lobe consolidation with adjacent. While this may be infectious in nature, a CT scan is recommended for further clarification.

Problem Set Up



+

Report:

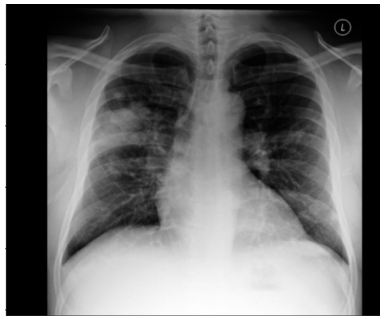
Right upper lobe consolidation with adjacent. While this may be infectious in nature, a CT scan is recommended for further clarification.

parse the reports to get the concepts

$\mathcal{C}$

right upper lobe
left lower lobe
heart size
….

Problem Set Up

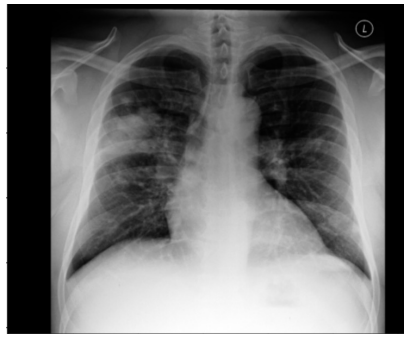


+

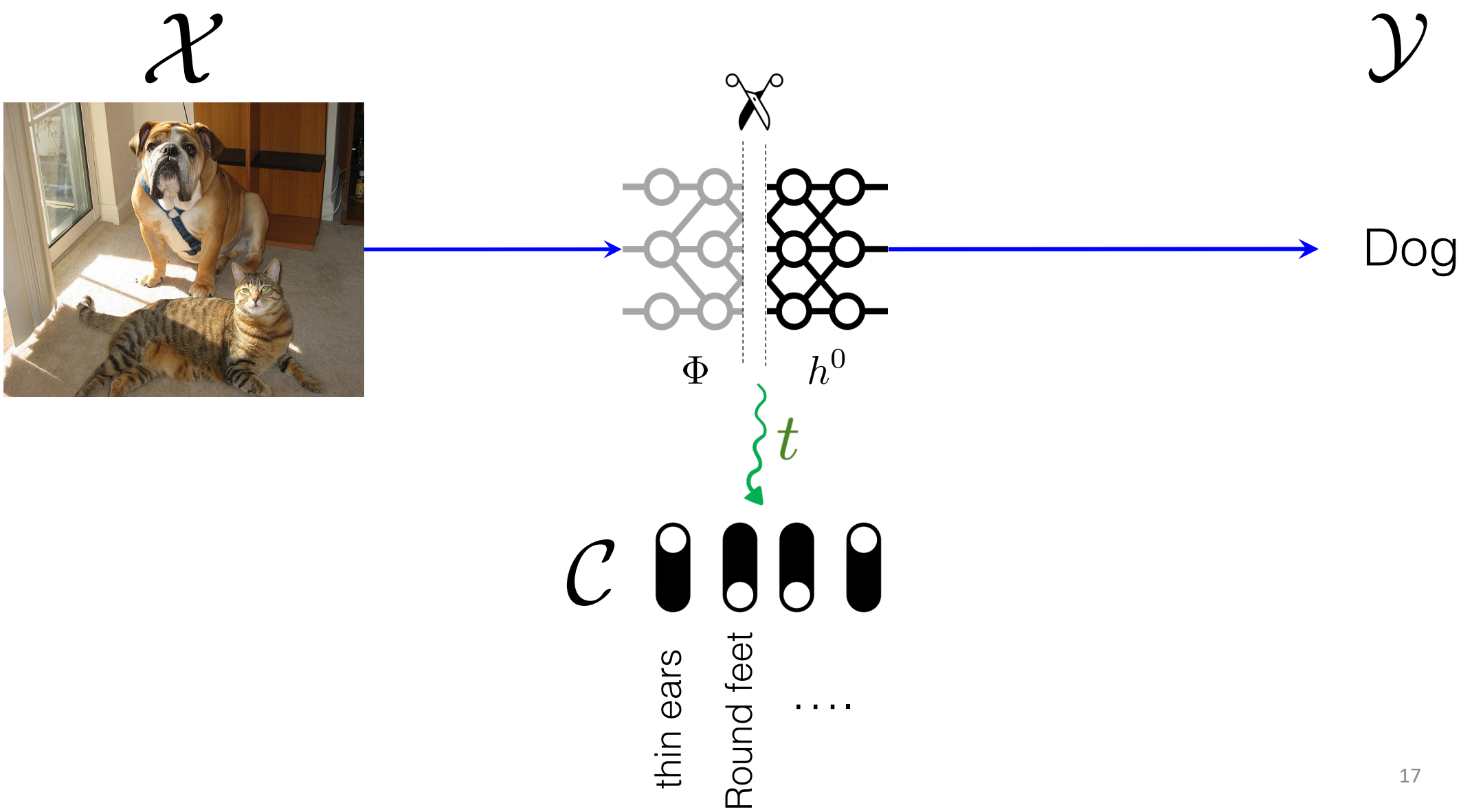
Report:

Right upper lobe consolidation with adjacent. While this may be infectious in nature, a CT scan is recommended for further clarification.

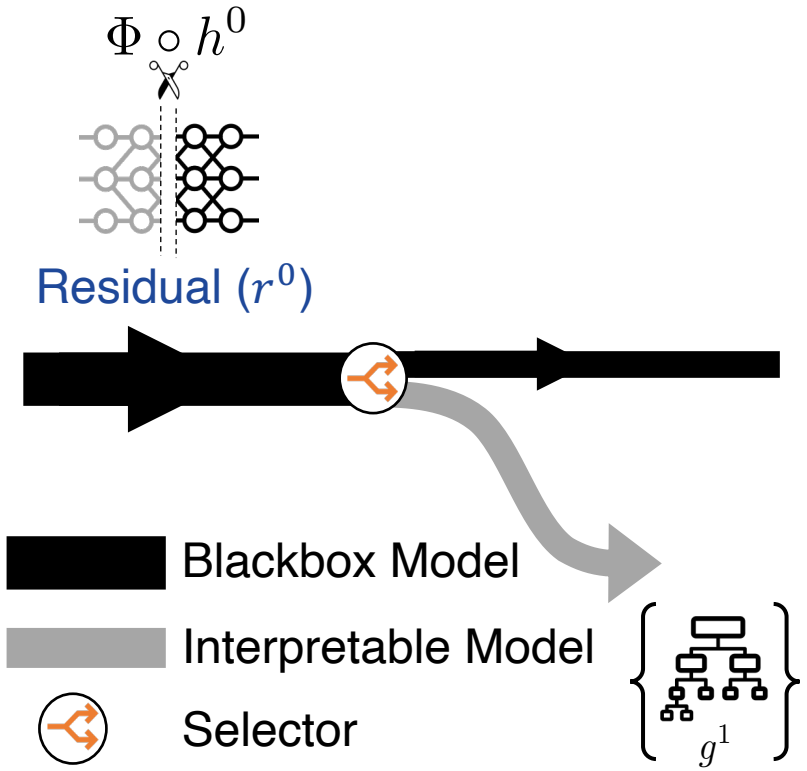
parse the reports to get the concepts

x	c	y
	right upper lobe left lower lobe heart size ….	Consolidation

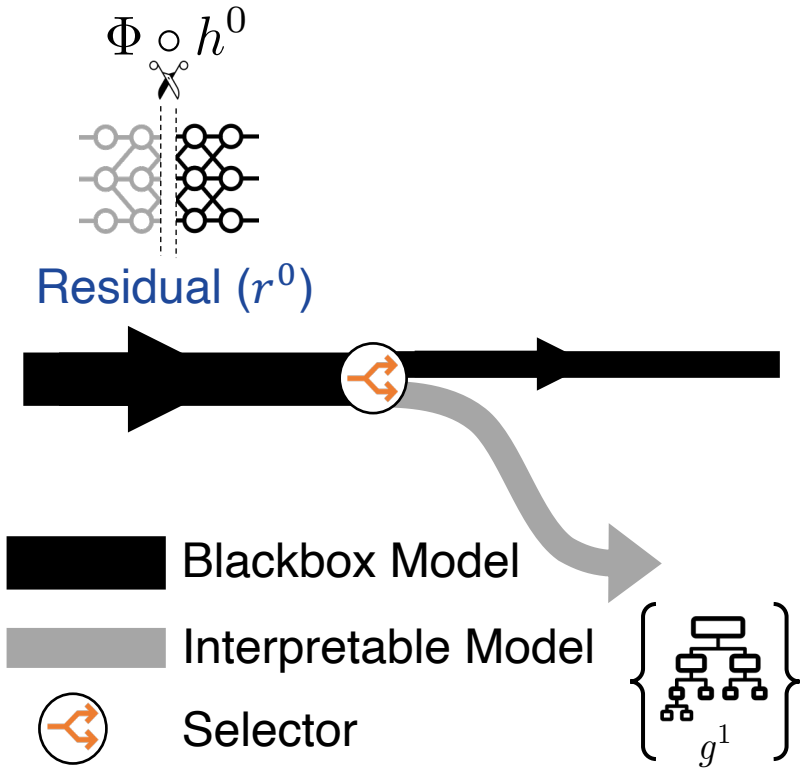
Discovering Hidden Concepts

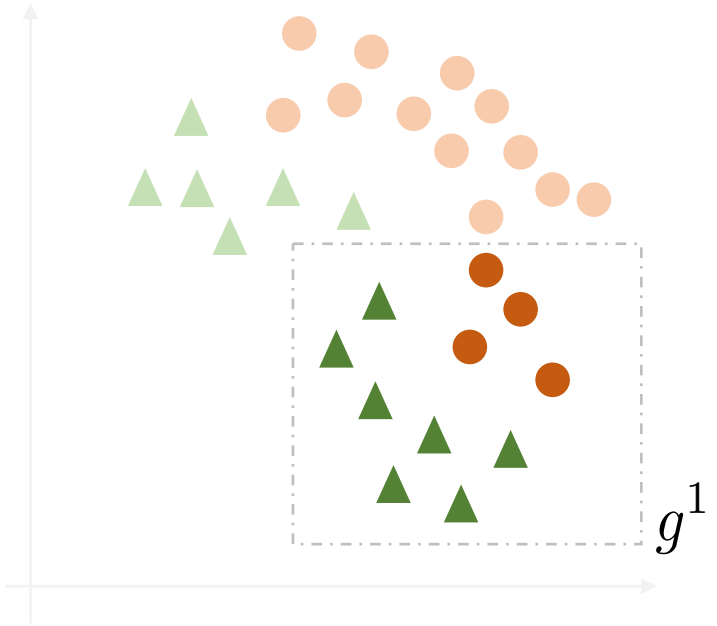


Carving out Interpretable Models



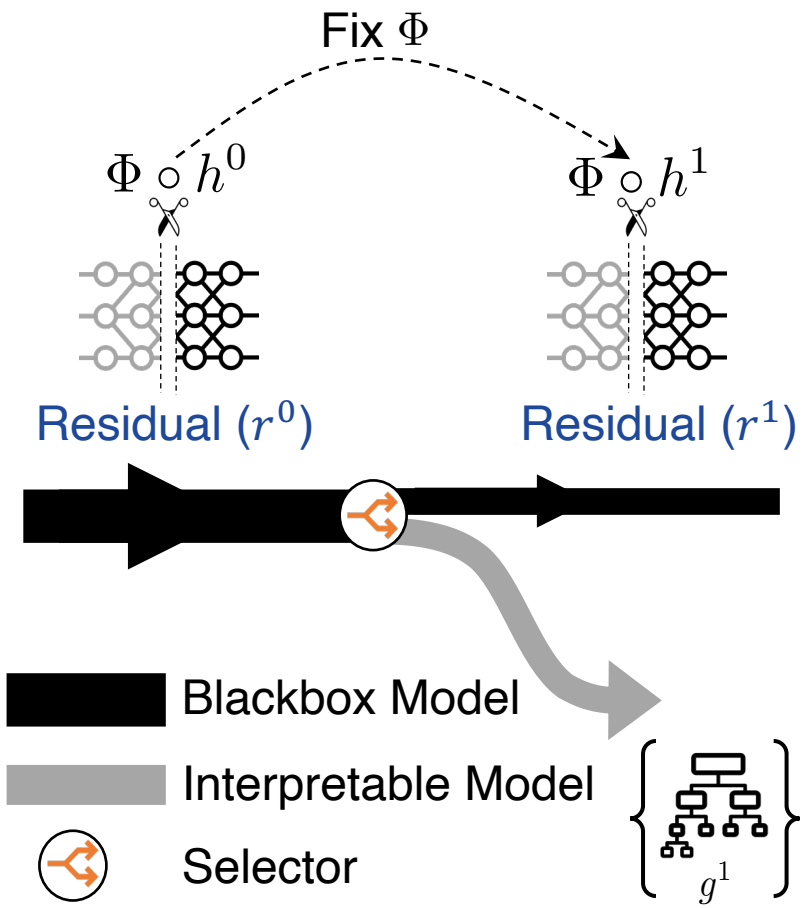
Carving out Interpretable Models

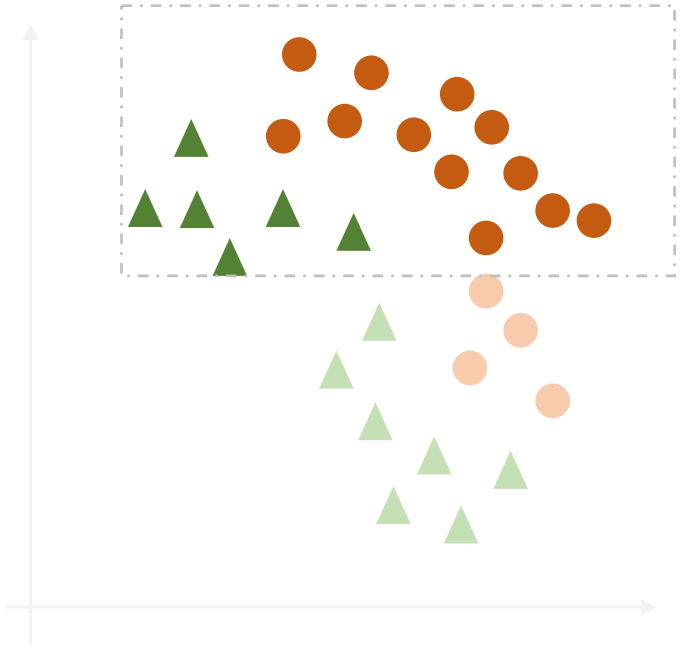




Each g is E-LENs (Barberio et al. AAAI 2022) to produce sample specific FOLs.

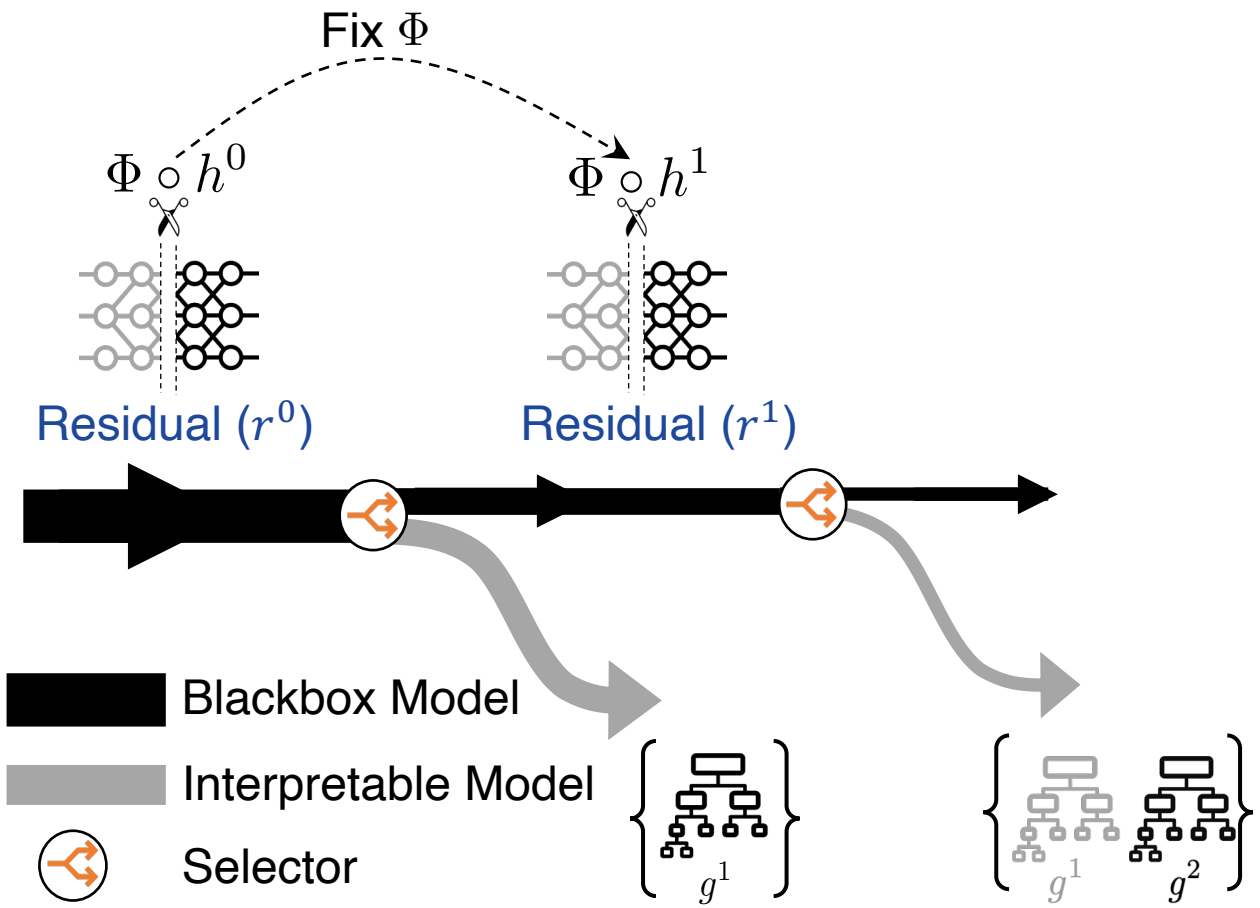
Carving out Interpretable Models





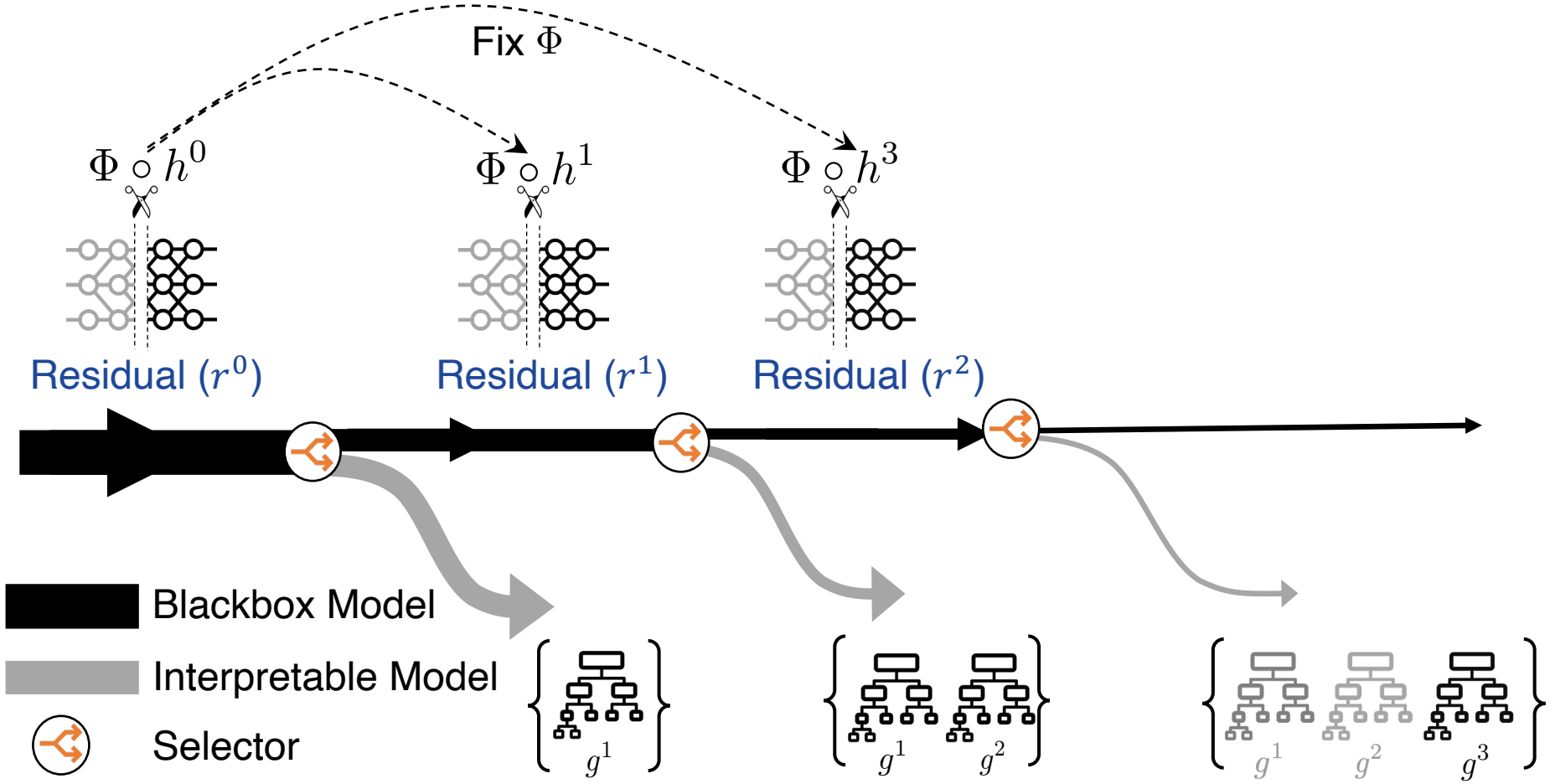
Each g is E-LENs (Barberio et al. AAAI 2022) to produce sample specific FOLs.

Carving out Interpretable Models



Each g is E-LENs (Barberio et al. AAAI 2022) to produce sample specific FOLs.

Carving out Interpretable Models



Each g is E-LENs (Barberio et al. AAAI 2022) to produce sample specific FOLs.

Baselines

CNN / VIT

Logistic classifier

Indigo Bunting

E-LENs

Indigo Bunting
+
FOL

Replace the standard logistic classifiers in CBM (Koh et el. ICML 2020) and post-hoc CBM (Yuksekgonul et al. ICLR 2023) with E-LENs to compare the FOLs

Examples on Bird Dataset

Examples on Bird Dataset

Expert 1

Olive sided Flycatcher $\leftrightarrow$breast_color_grey$\wedge$
tail_pattern_solid

Examples on Bird Dataset

Expert 1

Olive sided Flycatcher ↔breast_color_grey∧
tail_pattern_solid

Expert 2

Olive sided Flycatcher ↔underparts_color_grey∧
wing_color_grey

Examples on Bird Dataset

Expert 1

Olive sided Flycatcher ↔breast_color_grey∧ tail_pattern_solid

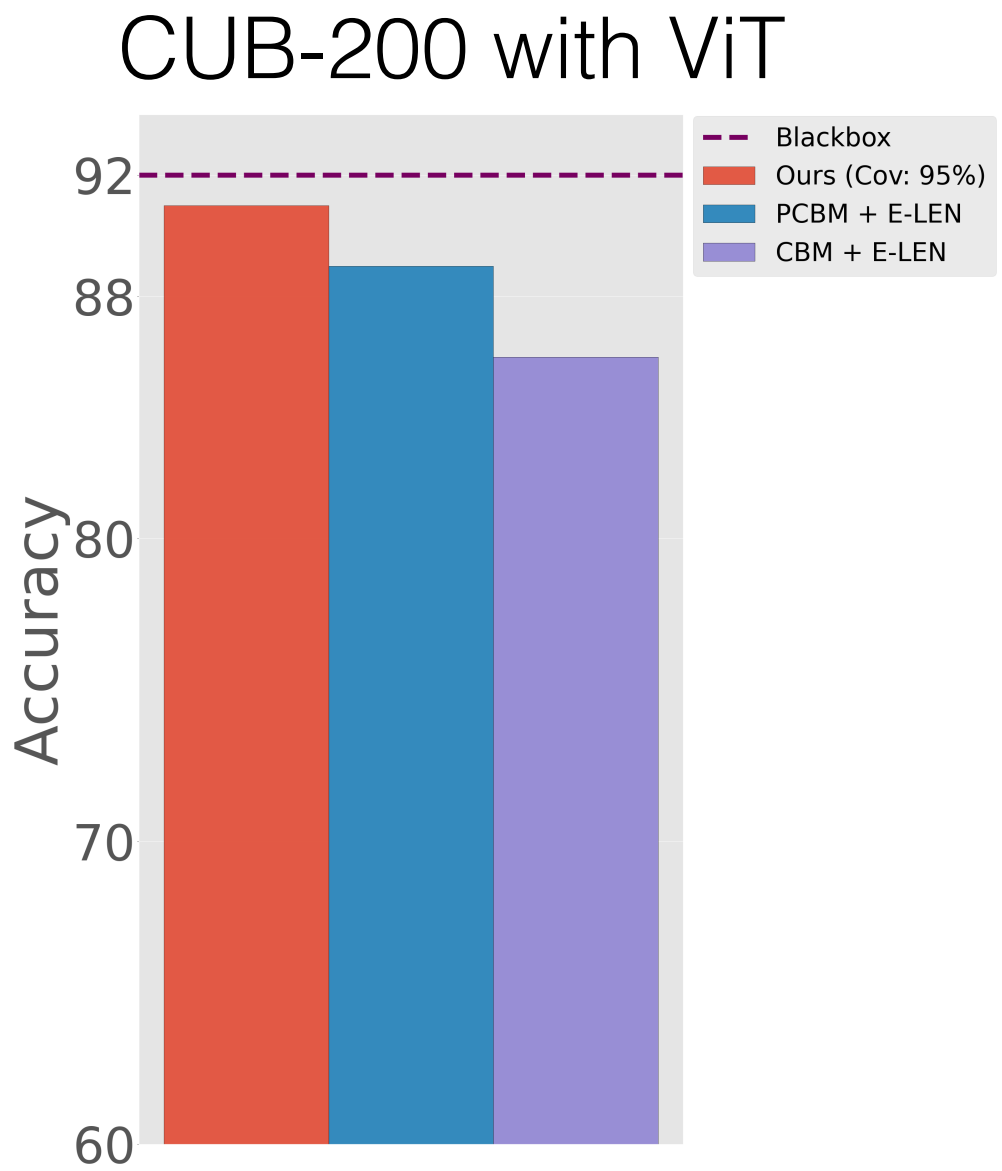
Expert 2

Olive sided Flycatcher ↔underparts_color_grey∧ wing_color_grey

Final Residual (Unexplained)

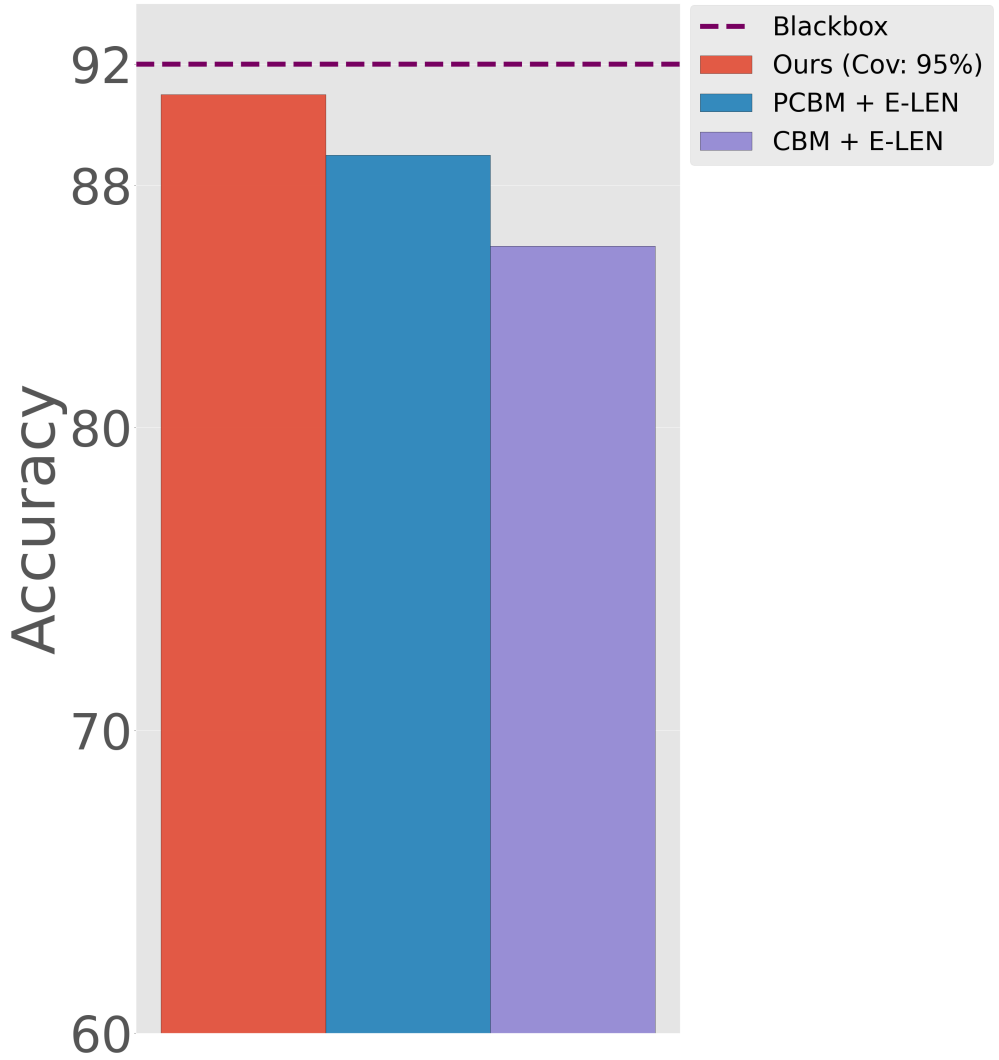
Comparing Performance

Comparing Performance

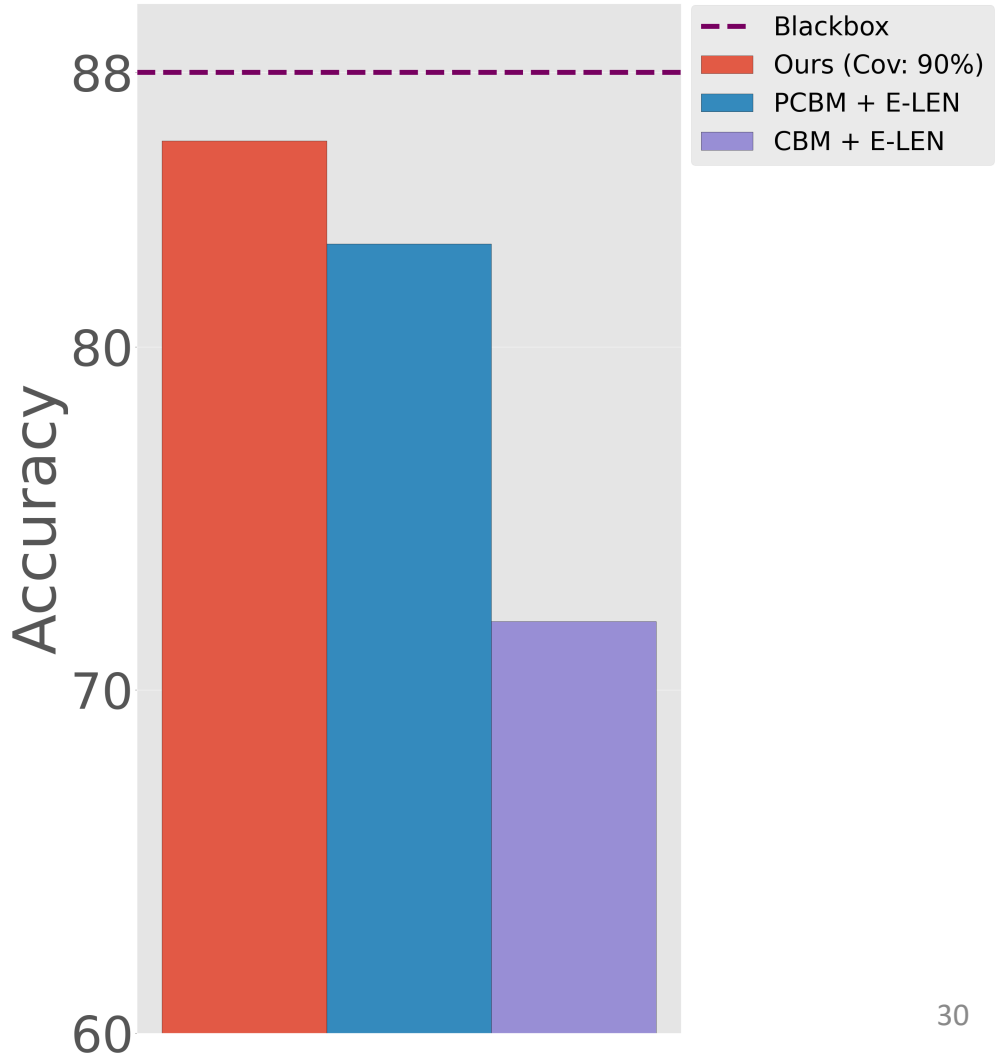


Comparing Performance

CUB-200 with ViT

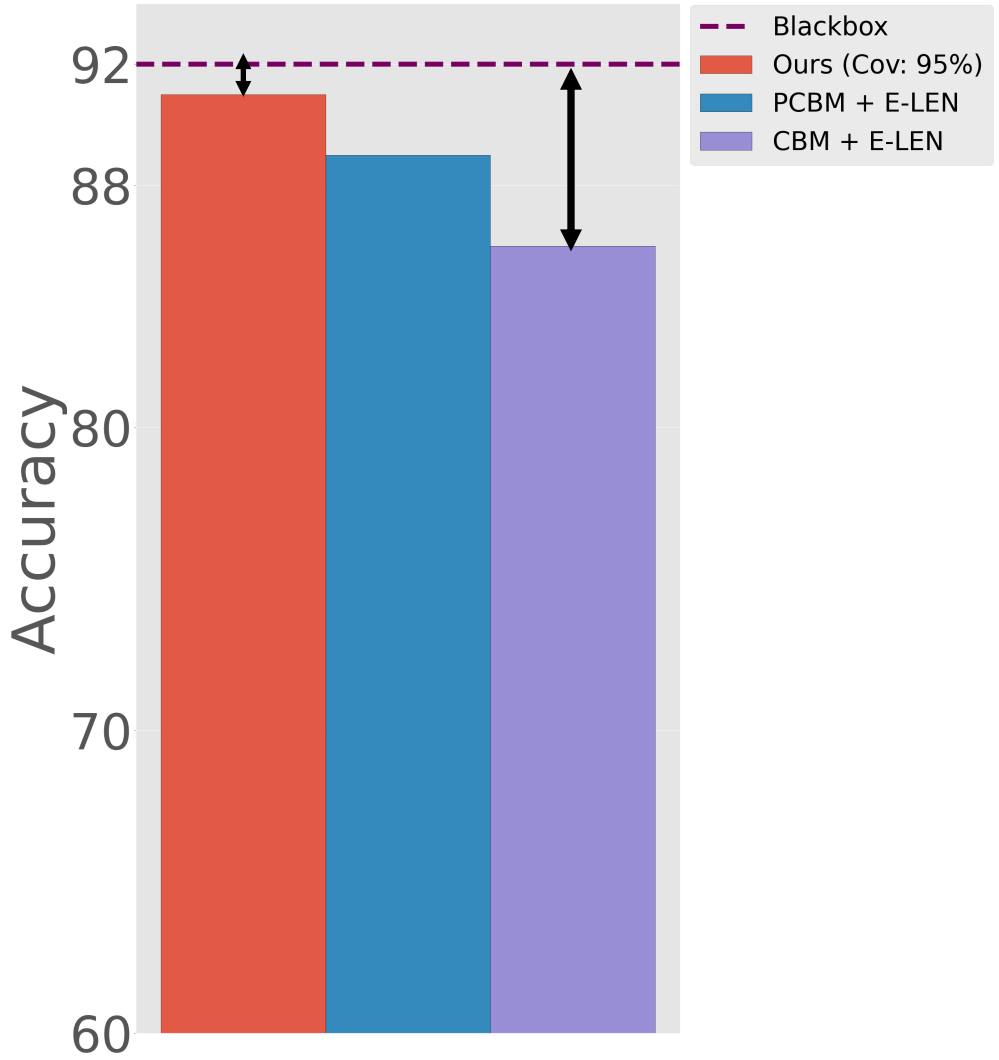


CUB-200 with ResNet101

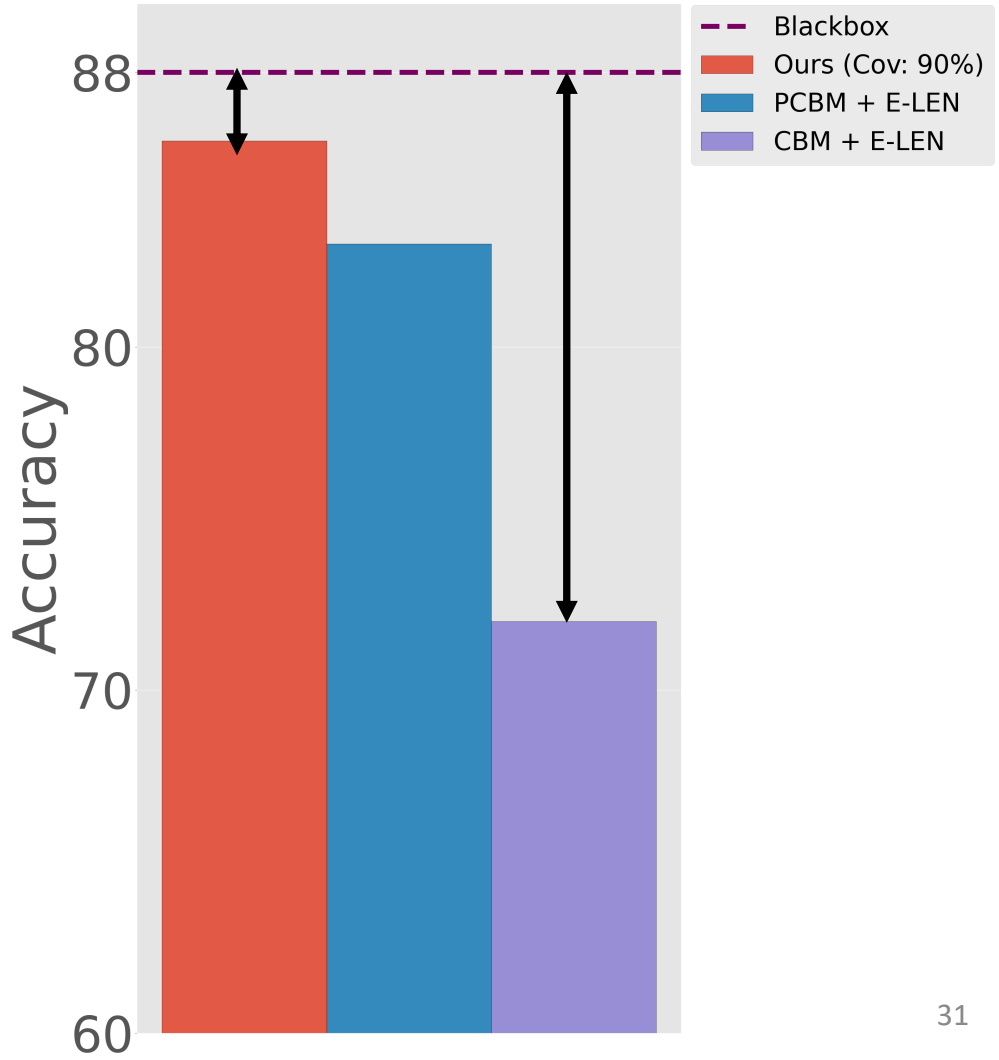


Comparing Performance

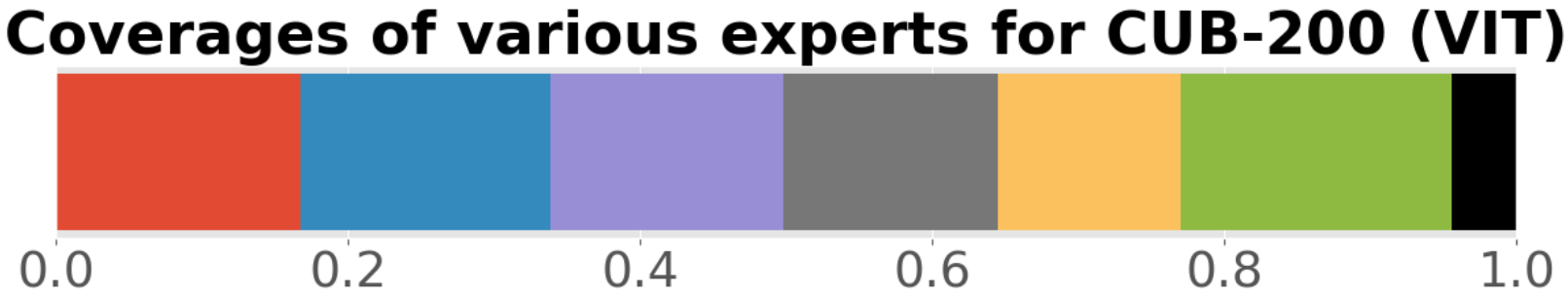
CUB-200 with ViT



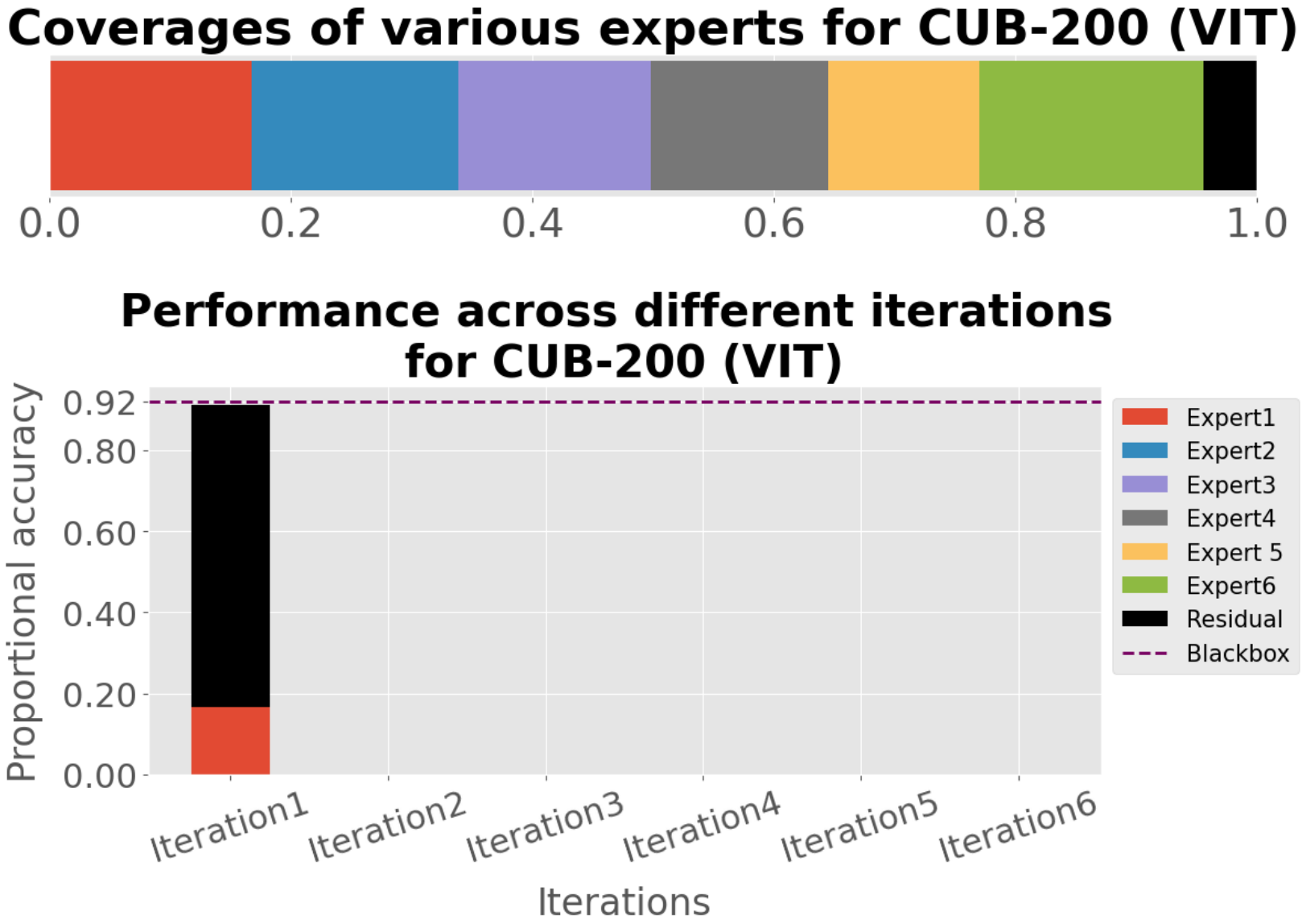
CUB-200 with ResNet101



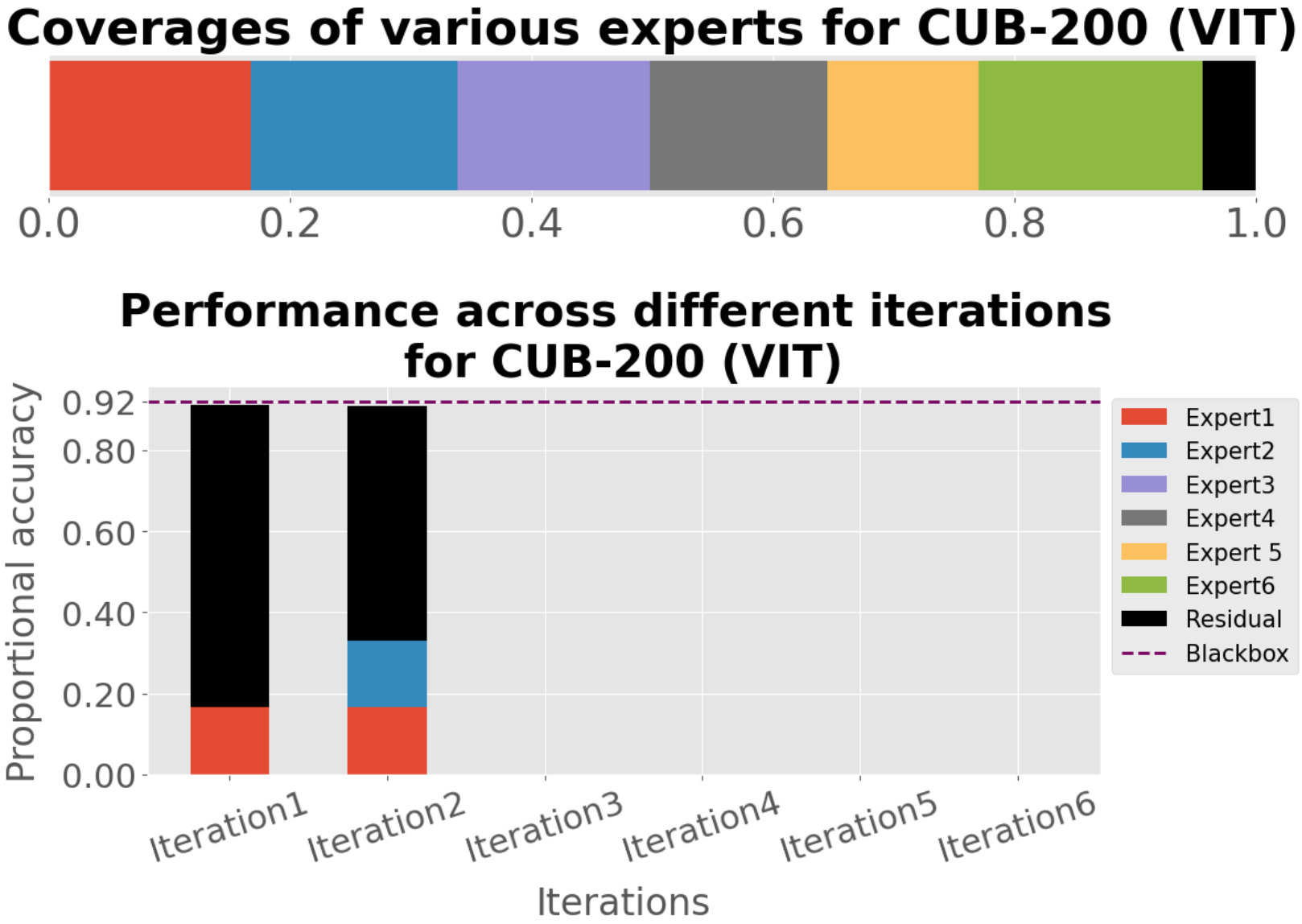
Residuals gradually performs poorly



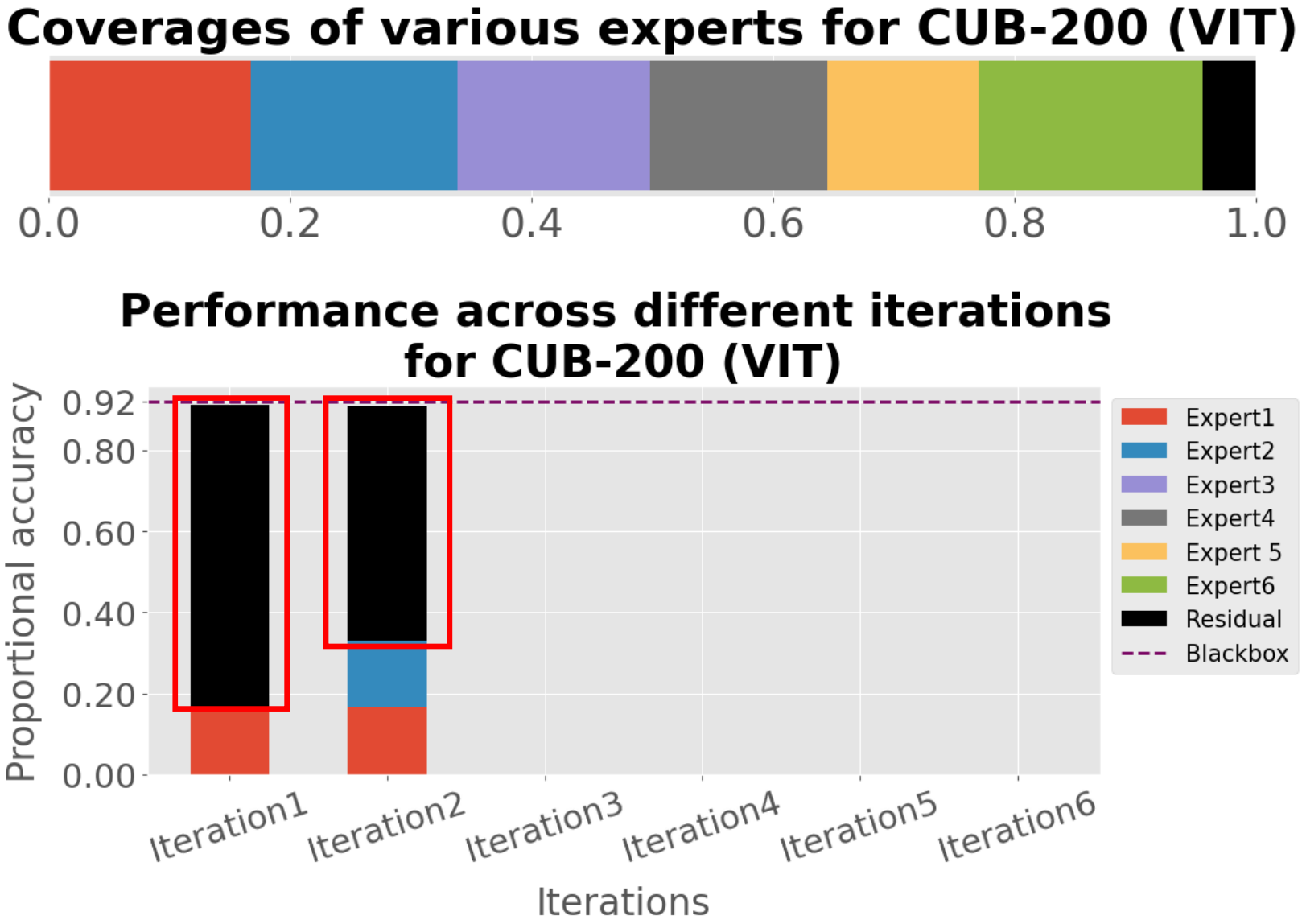
Residuals gradually performs poorly



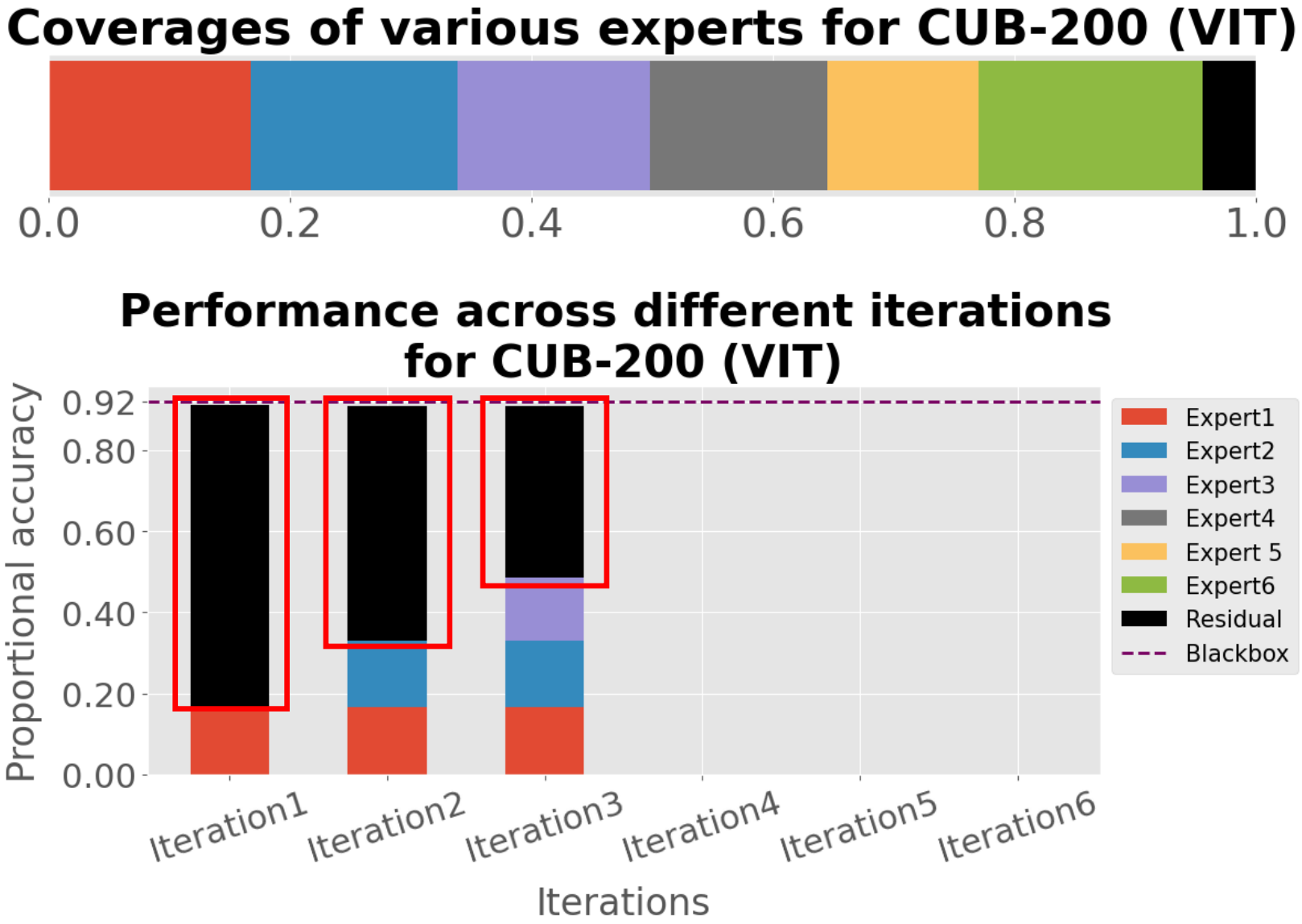
Residuals gradually performs poorly



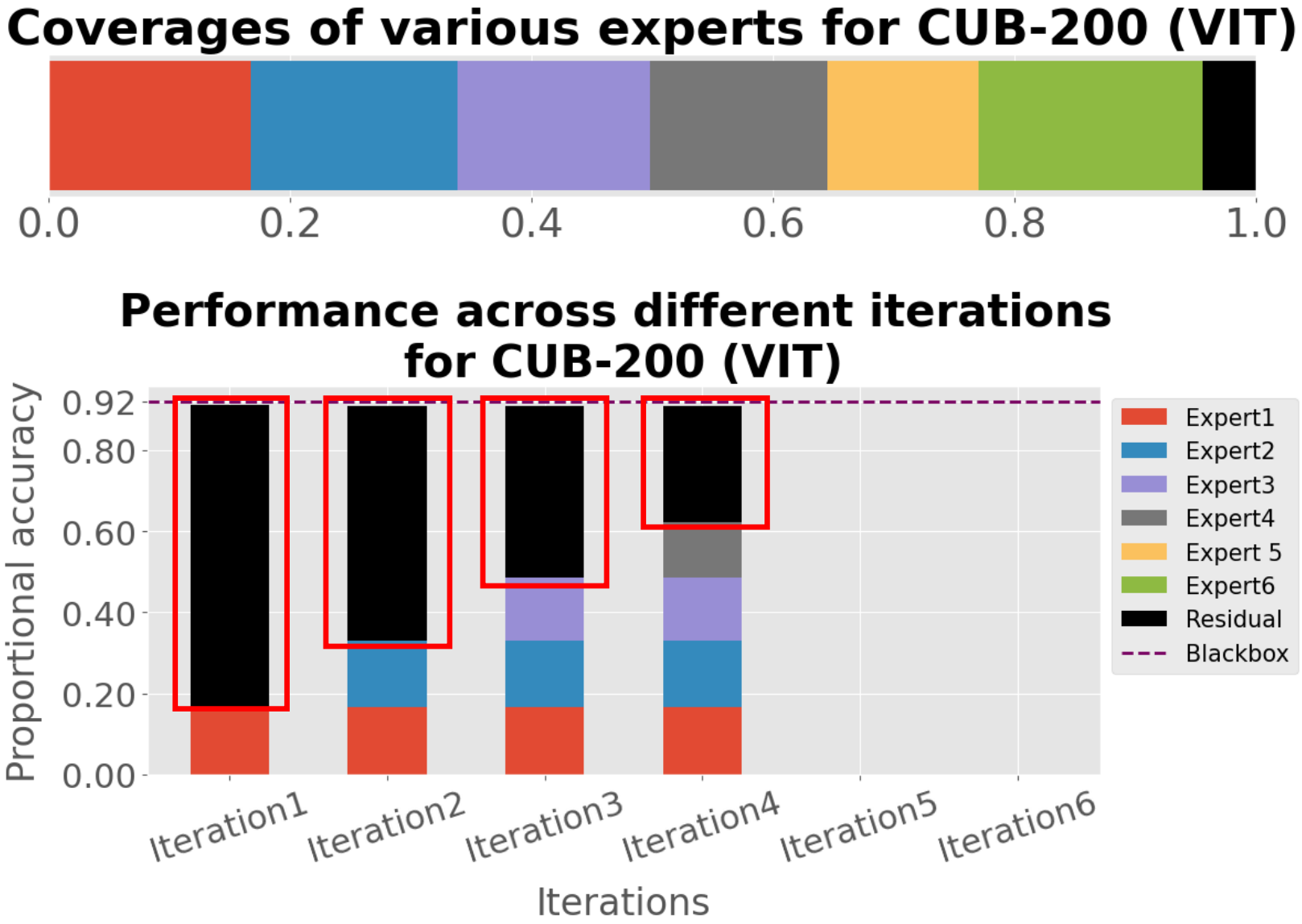
Residuals gradually performs poorly



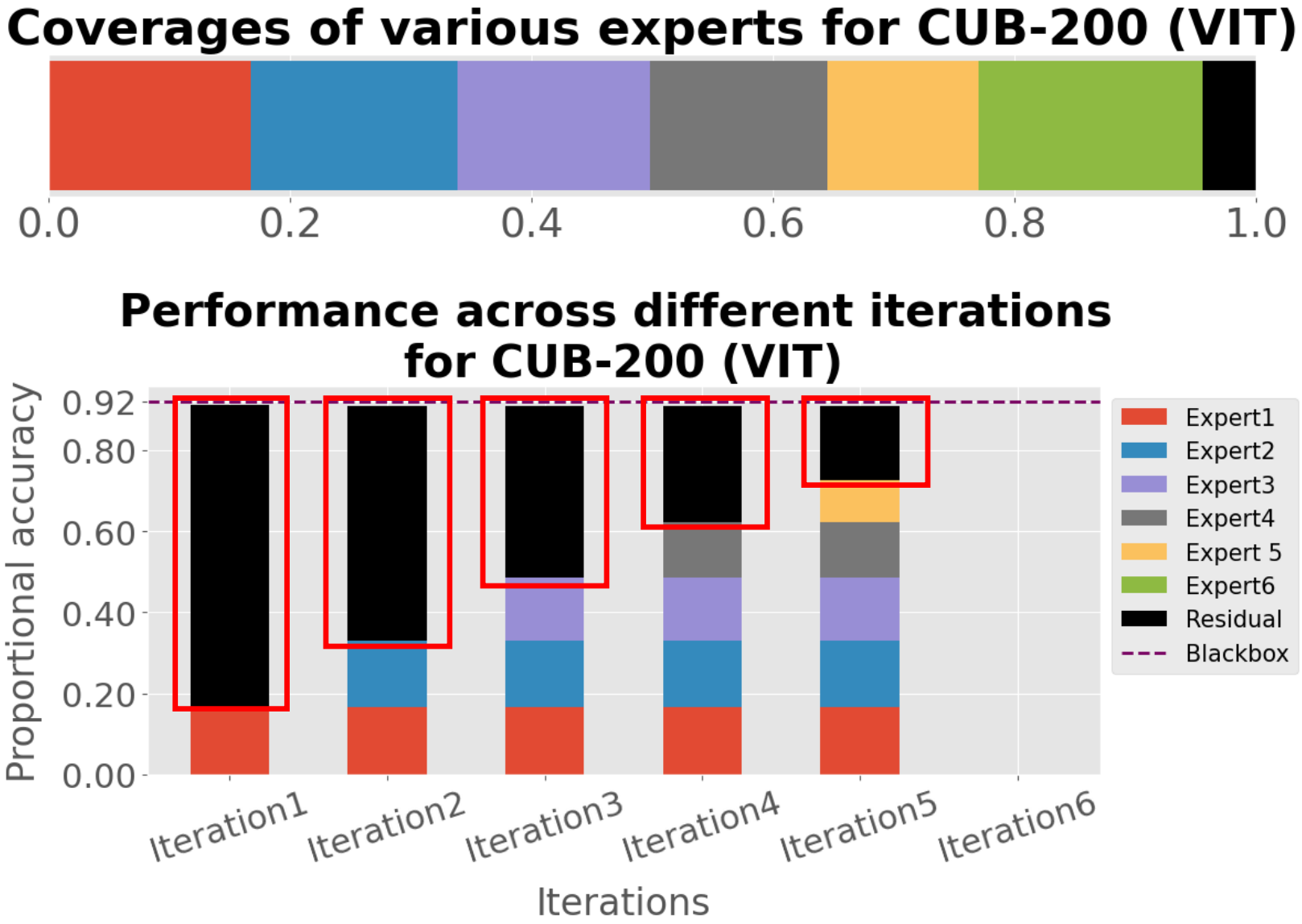
Residuals gradually performs poorly



Residuals gradually performs poorly

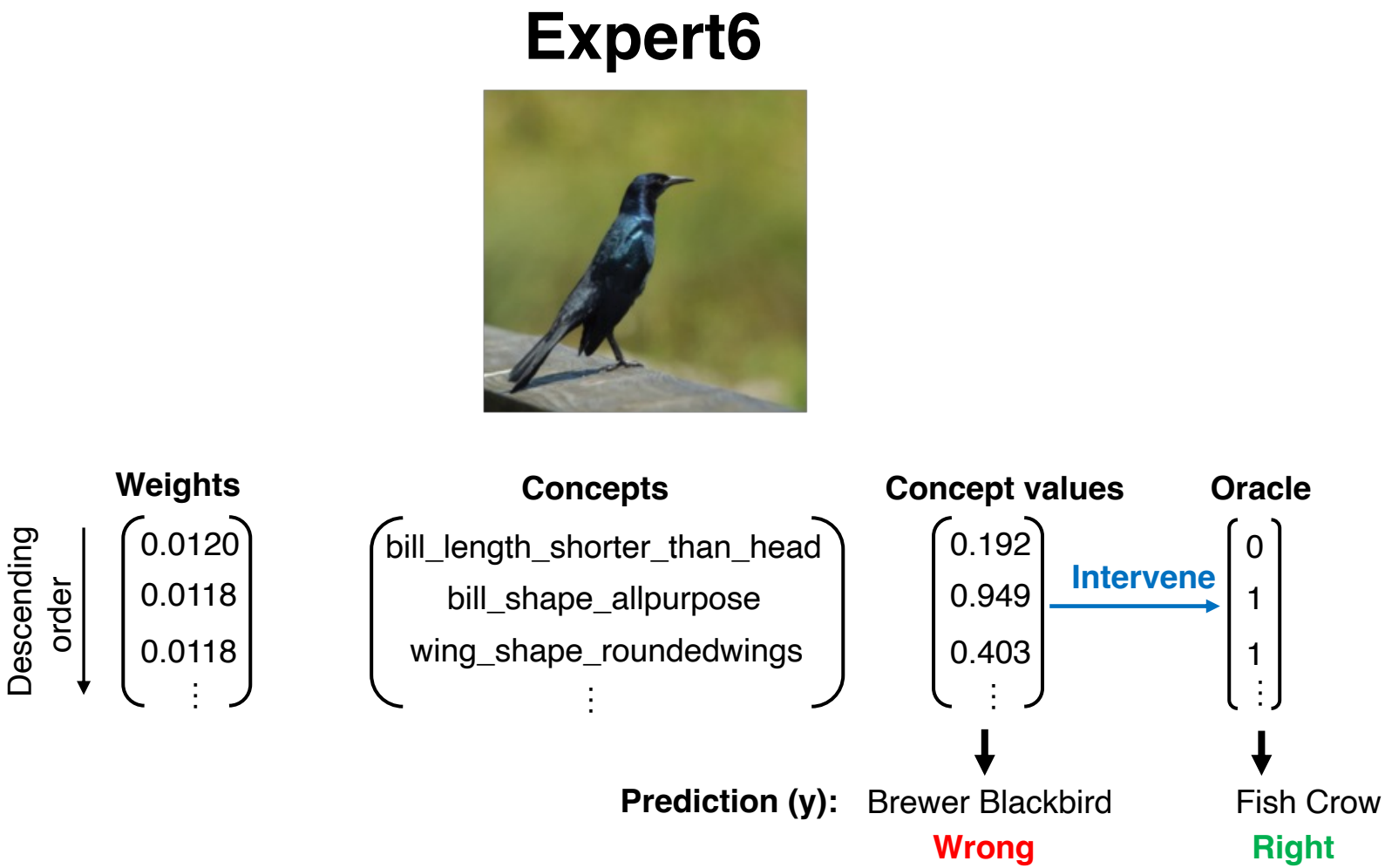


Residuals gradually performs poorly

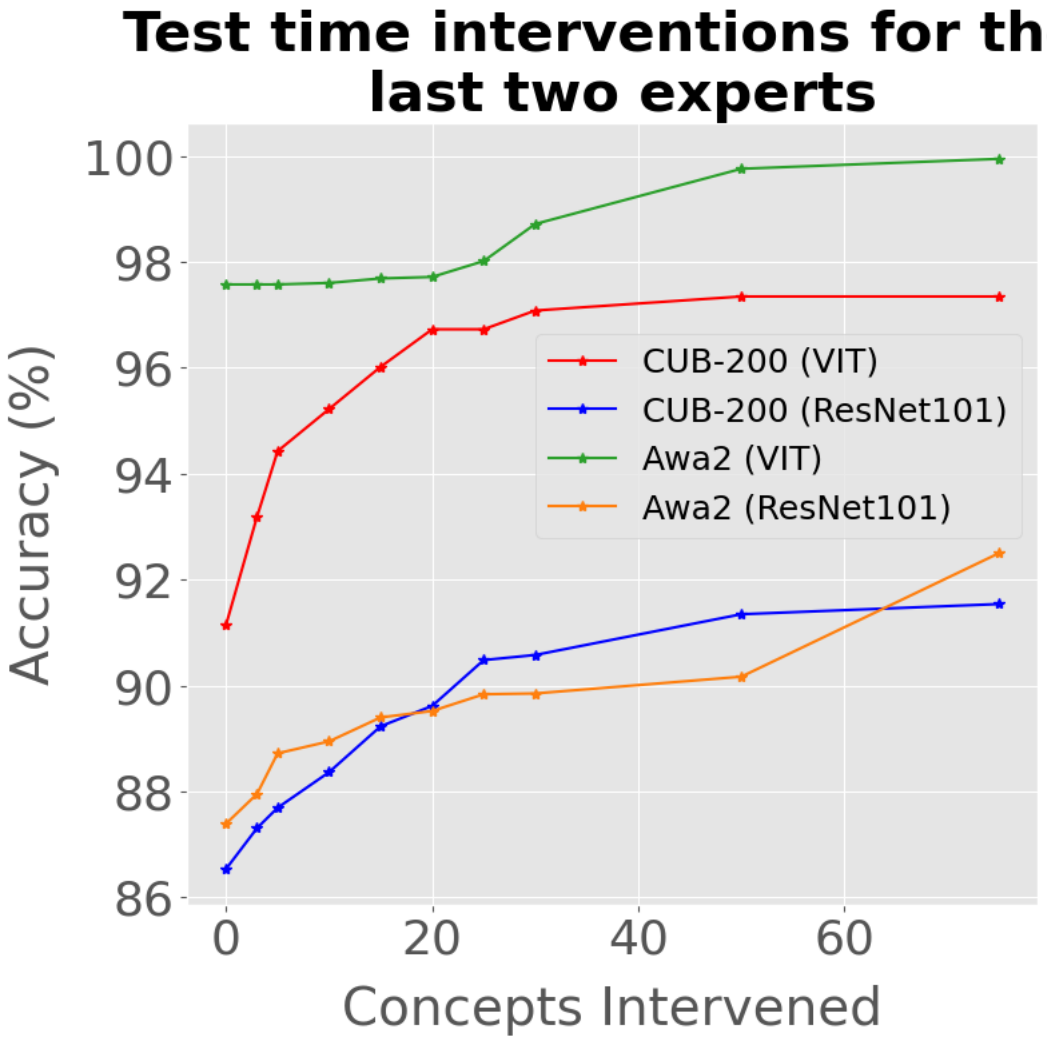


Test time interventions on hard samples

Chose the misclassified samples covered by the last 2 experts

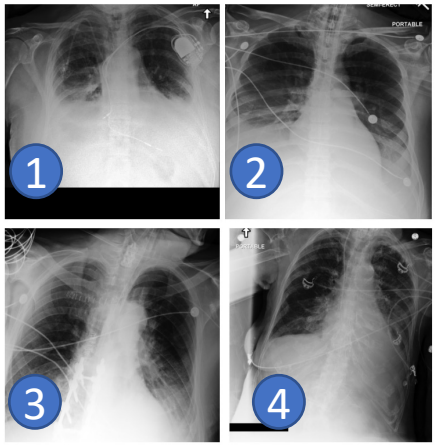


Test time interventions on hard samples



Intervening on top 75 concepts

- CUB-200 (ResNet101) – 5.49 % ↑
- CUB-200 (VIT) – 6.5 % ↑
- AWA2 (ResNet101) – 5.74 % ↑
- AWA2 (VIT) – 2.42 % ↑

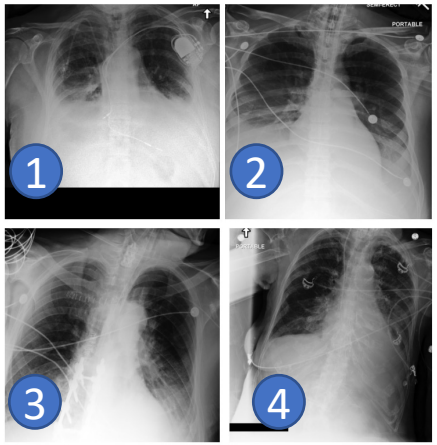


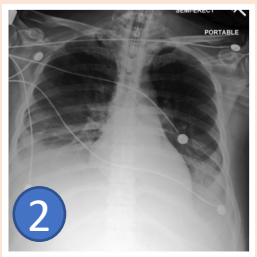
Examples on Chest X-ray

Pleural_unspec is “unspecified pleural effusion” referred to as “hydrothorax”.

Hydrothorax is a noninflammatory collection of serous fluid within the pleural cavities

Examples on Chest X-ray





Expert 1

Effusion $\leftrightarrow$

left_pleural

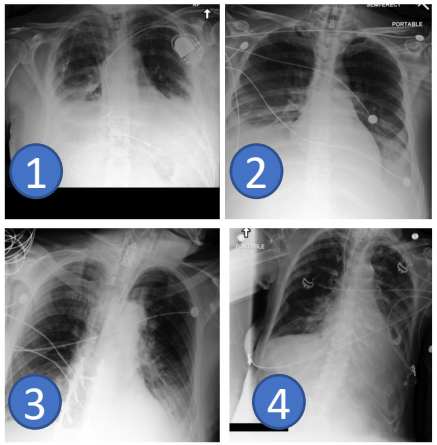
$\wedge$ right_pleural

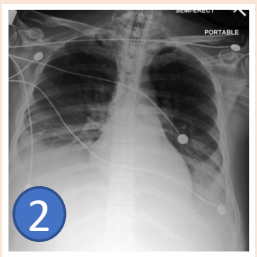
$\wedge$ pleural_unspec

Pleural_unspec is “unspecified pleural effusion” referred to as “hydrothorax”.

Hydrothorax is a noninflammatory collection of serous fluid within the pleural cavities

Examples on Chest X-ray





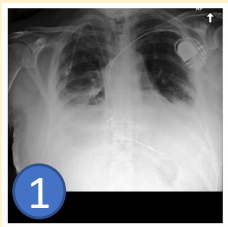
Expert 1

Effusion $\leftrightarrow$

left_pleural

$\wedge$ right_pleural

$\wedge$ pleural_unspec



Expert 2

Effusion $\leftrightarrow$

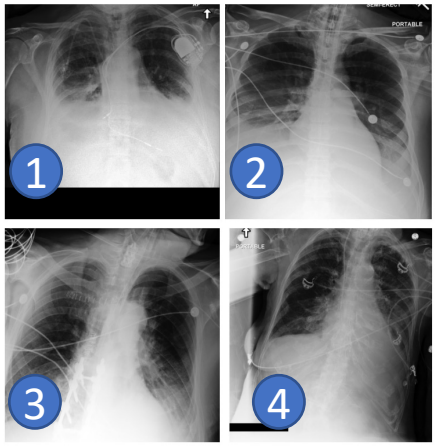
right_pleural

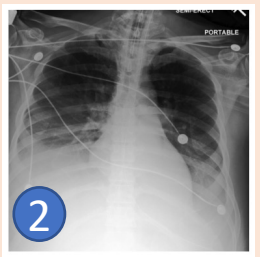
$\wedge$ pleural_unspec

Pleural_unspec is “unspecified pleural effusion” referred to as “hydrothorax”.

Hydrothorax is a noninflammatory collection of serous fluid within the pleural cavities

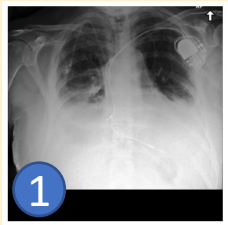
Examples on Chest X-ray





Expert 1

Effusion $\leftrightarrow$
left_pleural
$\wedge$ right_pleural
$\wedge$ pleural_unspec



Expert 2

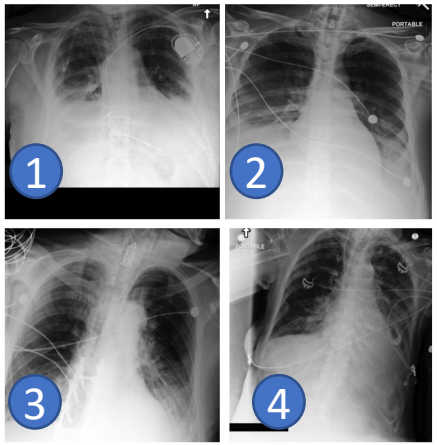
Effusion $\leftrightarrow$
right_pleural
$\wedge$ pleural_unspec

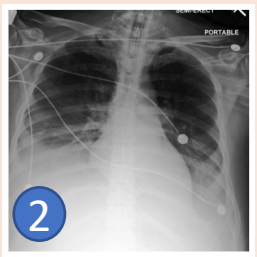
Expert 3

Effusion $\leftrightarrow$
left_pleural
$\wedge$ pleural_unspec

Pleural_unspec is “unspecified pleural effusion” referred to as “hydrothorax”.
Hydrothorax is a noninflammatory collection of serous fluid within the pleural cavities

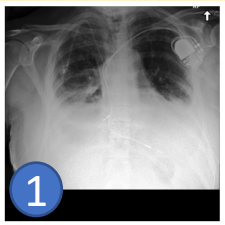
Examples on Chest X-ray





Expert 1

Effusion $\leftrightarrow$
left_pleural
$\wedge$ right_pleural
$\wedge$ pleural_unspec

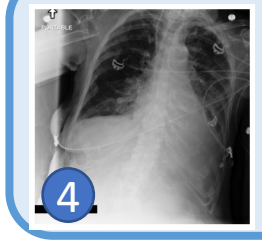


Expert 2

Effusion $\leftrightarrow$
right_pleural
$\wedge$ pleural_unspec

Expert 3

Effusion $\leftrightarrow$
left_pleural
$\wedge$ pleural_unspec

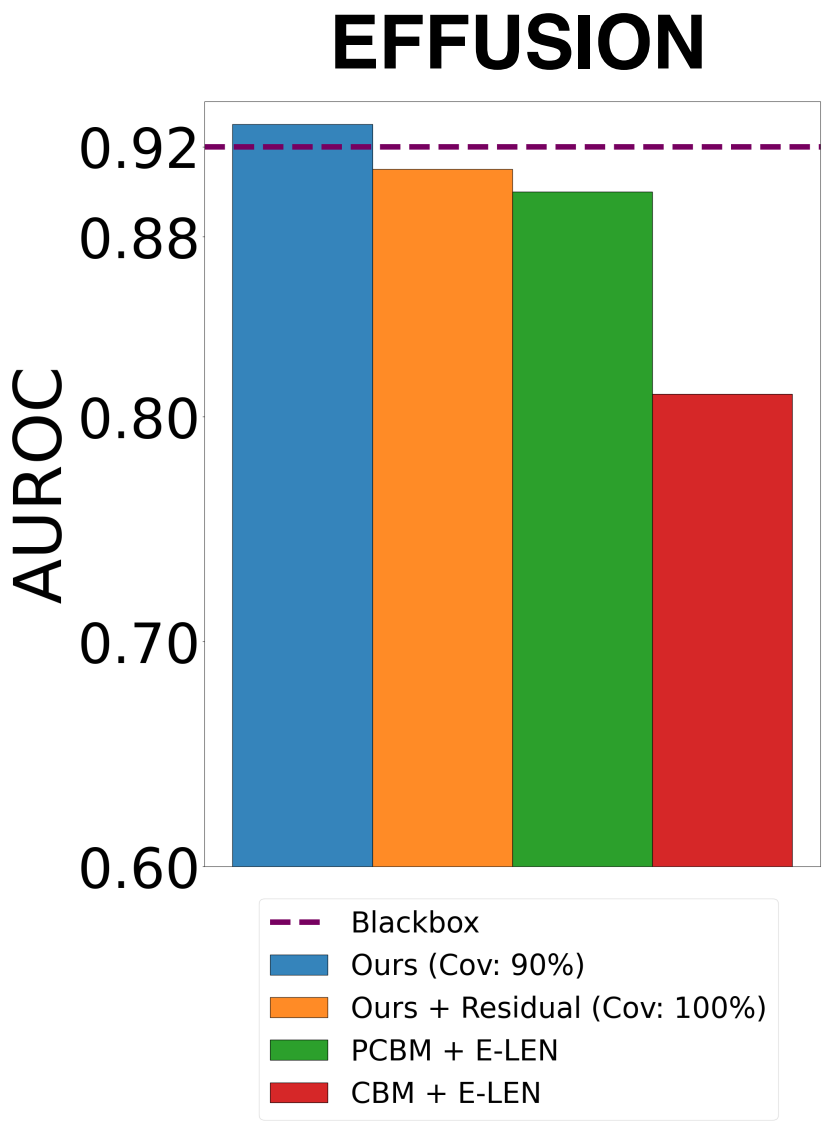


Expert 4

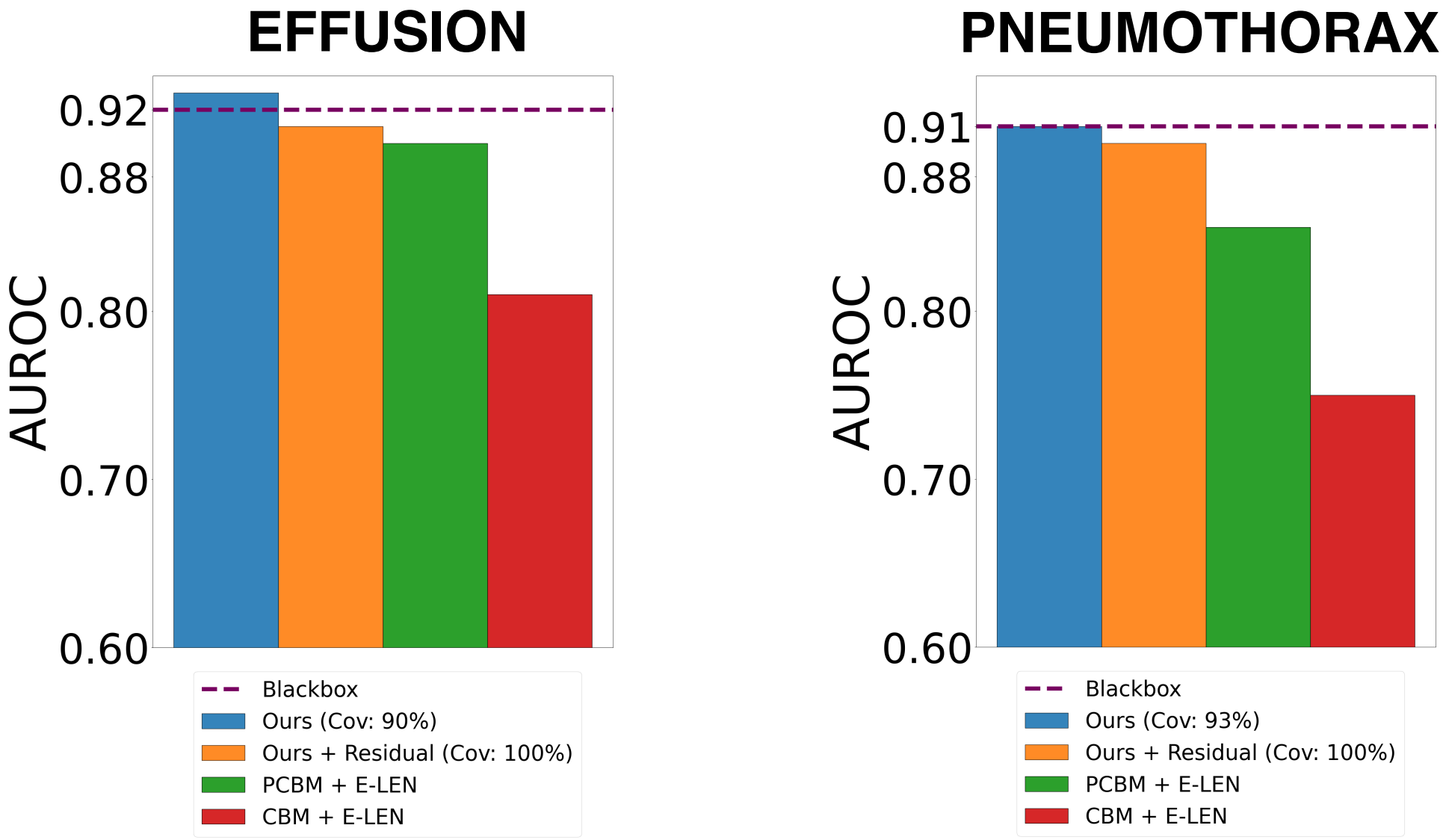
Effusion $\leftrightarrow$
pleural_unspec

Pleural_unspec is “unspecified pleural effusion” referred to as “hydrothorax”.
Hydrothorax is a noninflammatory collection of serous fluid within the pleural cavities

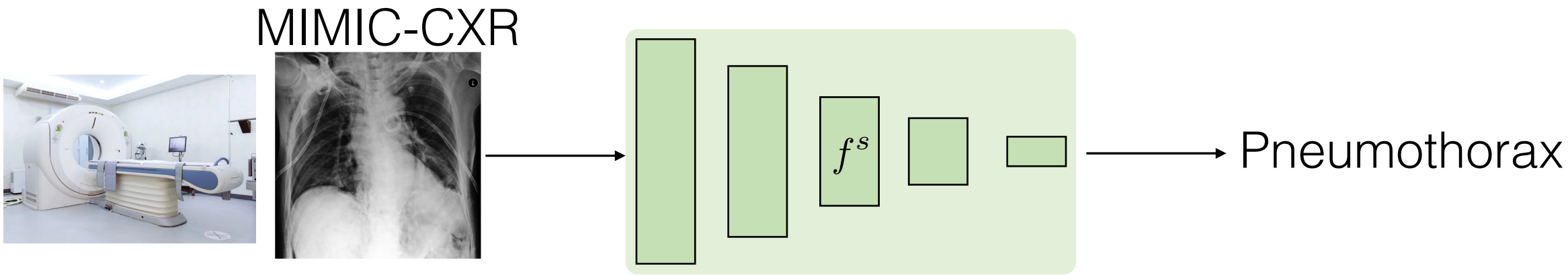
Comparing Performance



Comparing Performance

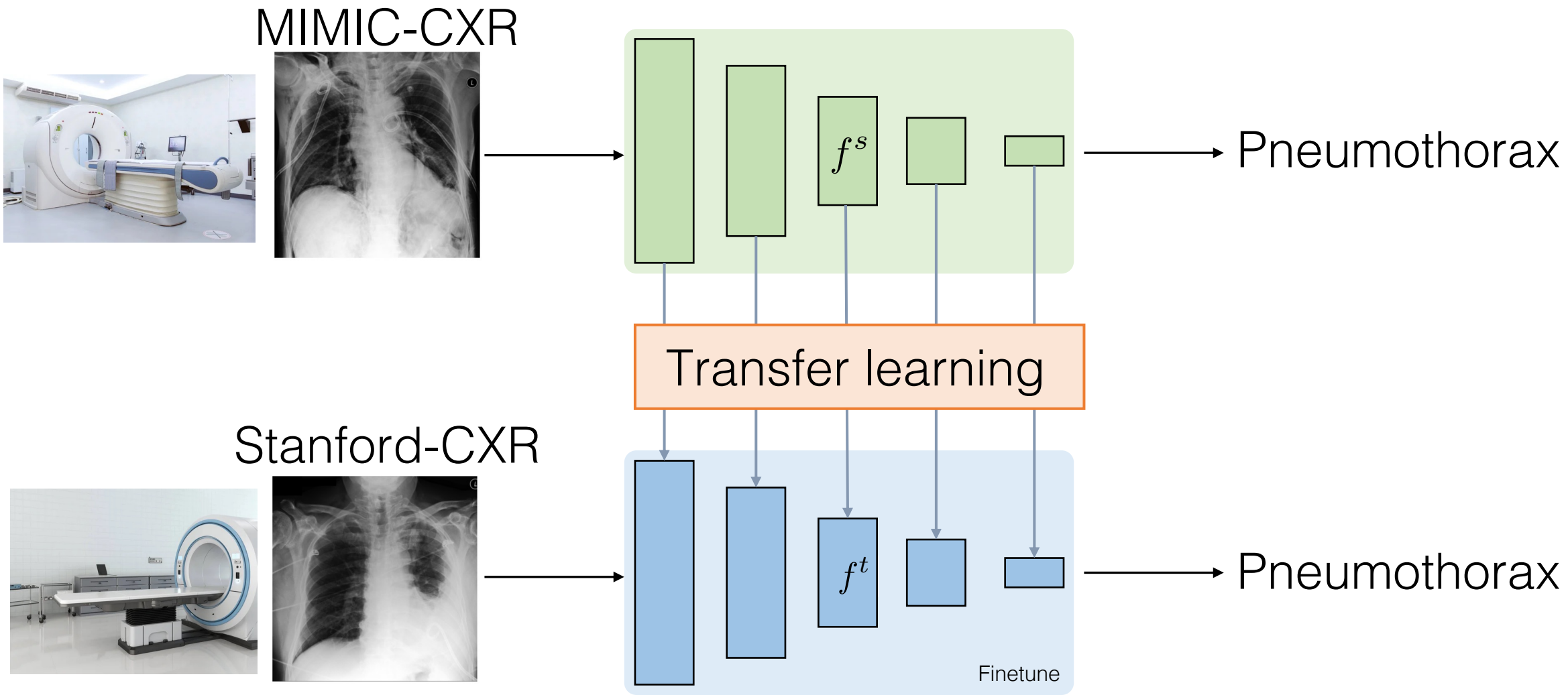


Data-Efficient Fine-tunning



Best practices for fine-tuning visual classifiers to new domains. In: Computer Vision. Chu et al. 2016. ECCV-W

Data-Efficient Fine-tunning



Best practices for fine-tuning visual classifiers to new domains. In: Computer Vision. Chu et al. 2016. ECCV-W

Data-Efficient Fine-tunning

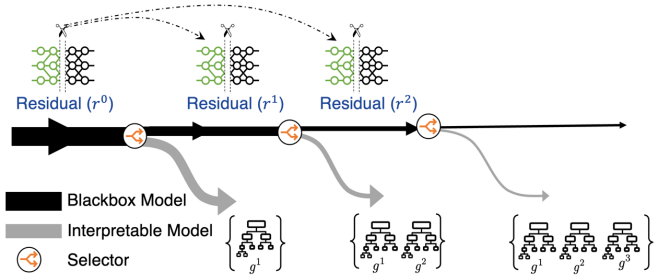
Data and Computationally inefficient

The clinical rules are “invariant”

How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset. Kandel et al. 2020. Applied Sciences
Growing a brain: Fine-tuning by increasing model capacity. Wang et al. 2017 CVPR

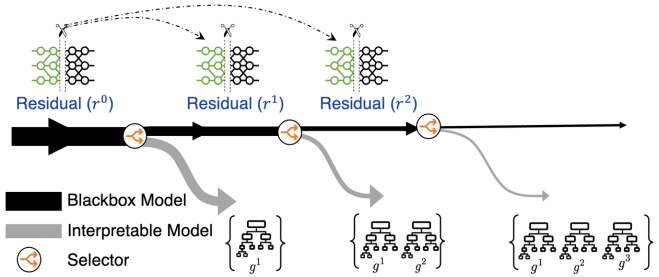
Fine-tune to a New Domain

1 Apply source model

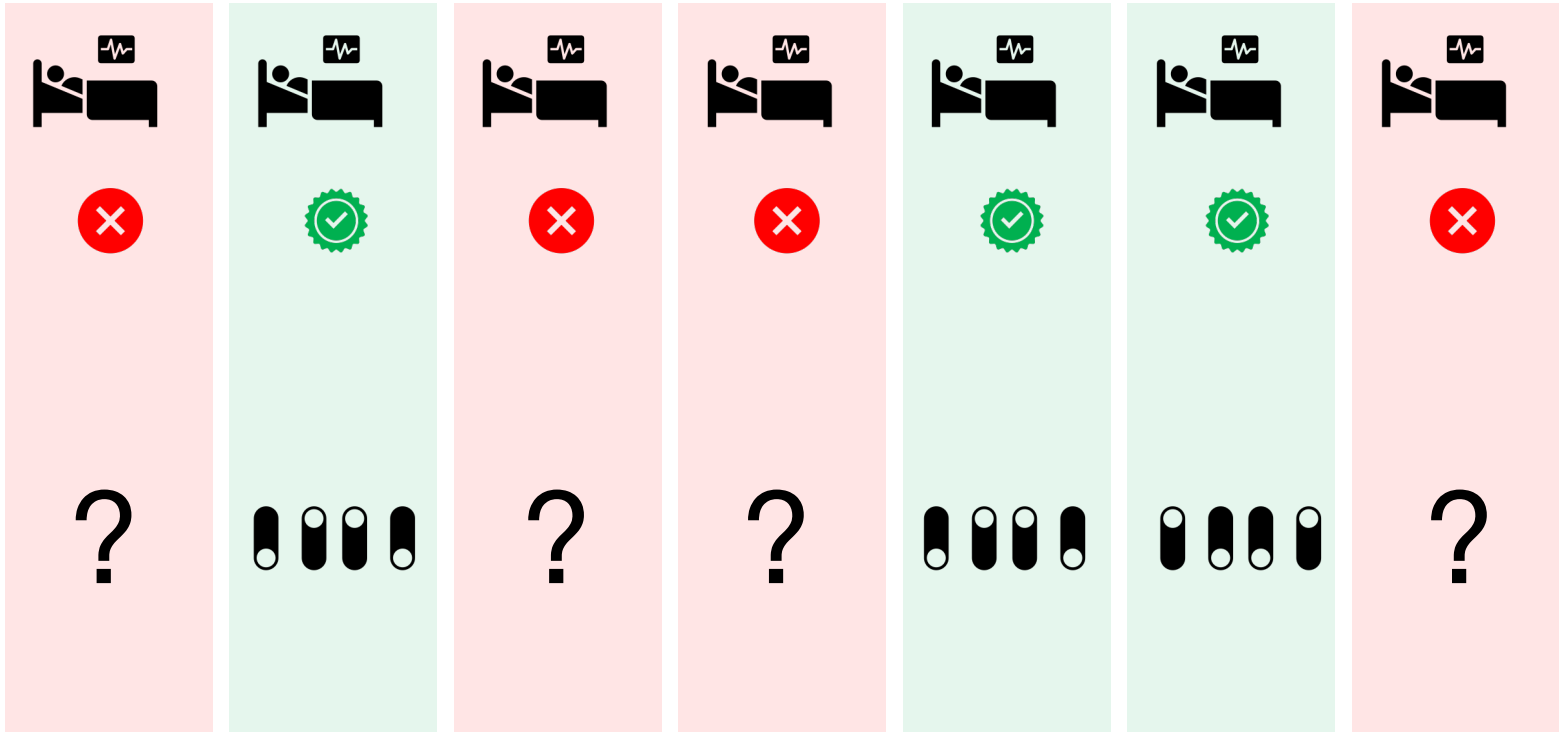


Fine-tune to a New Domain

1. Apply source model

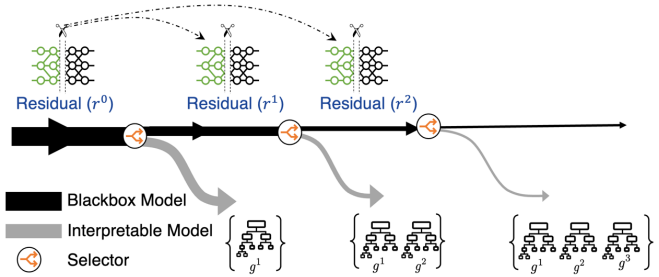


2. Use concepts from matching patients c

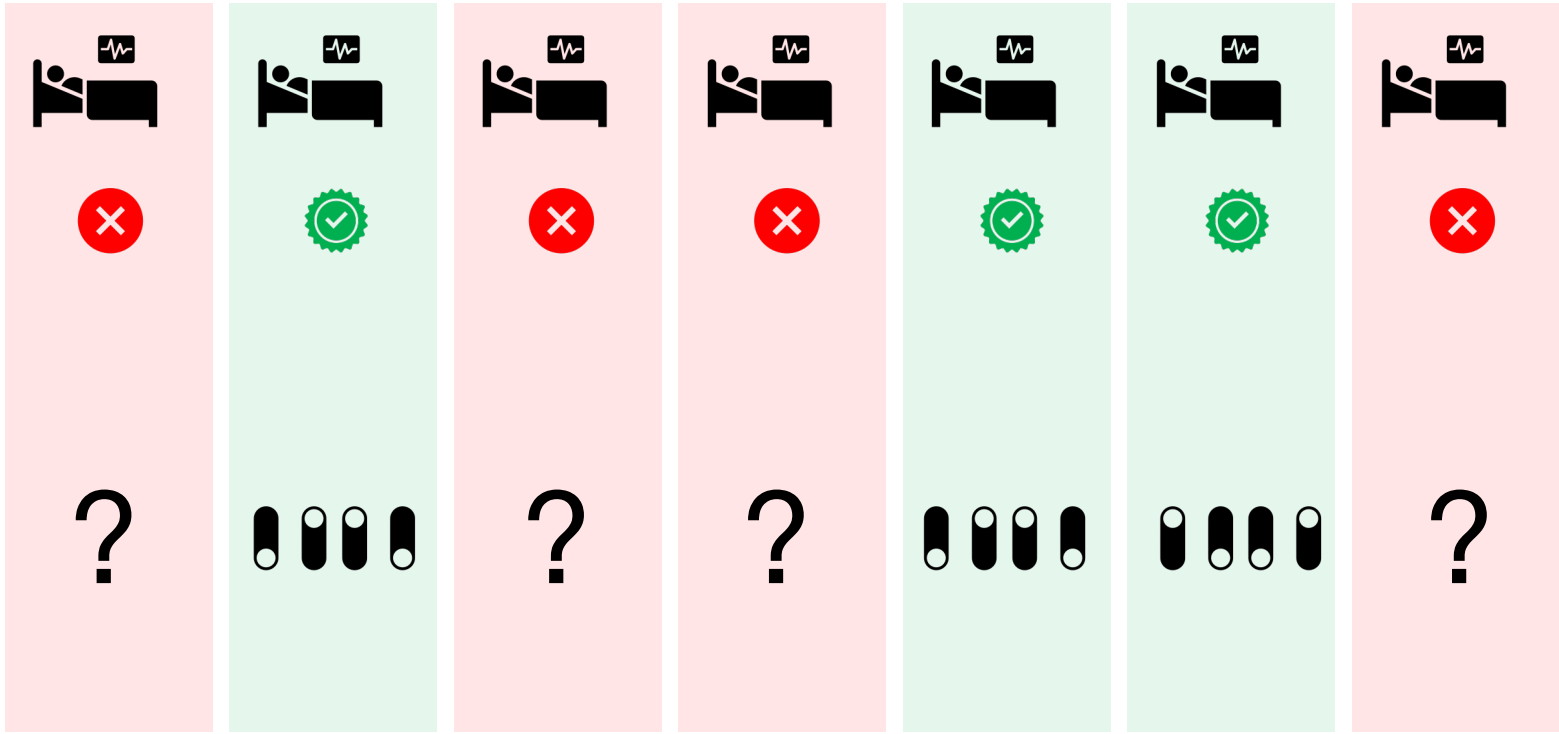


Fine-tune to a New Domain

1. Apply source model



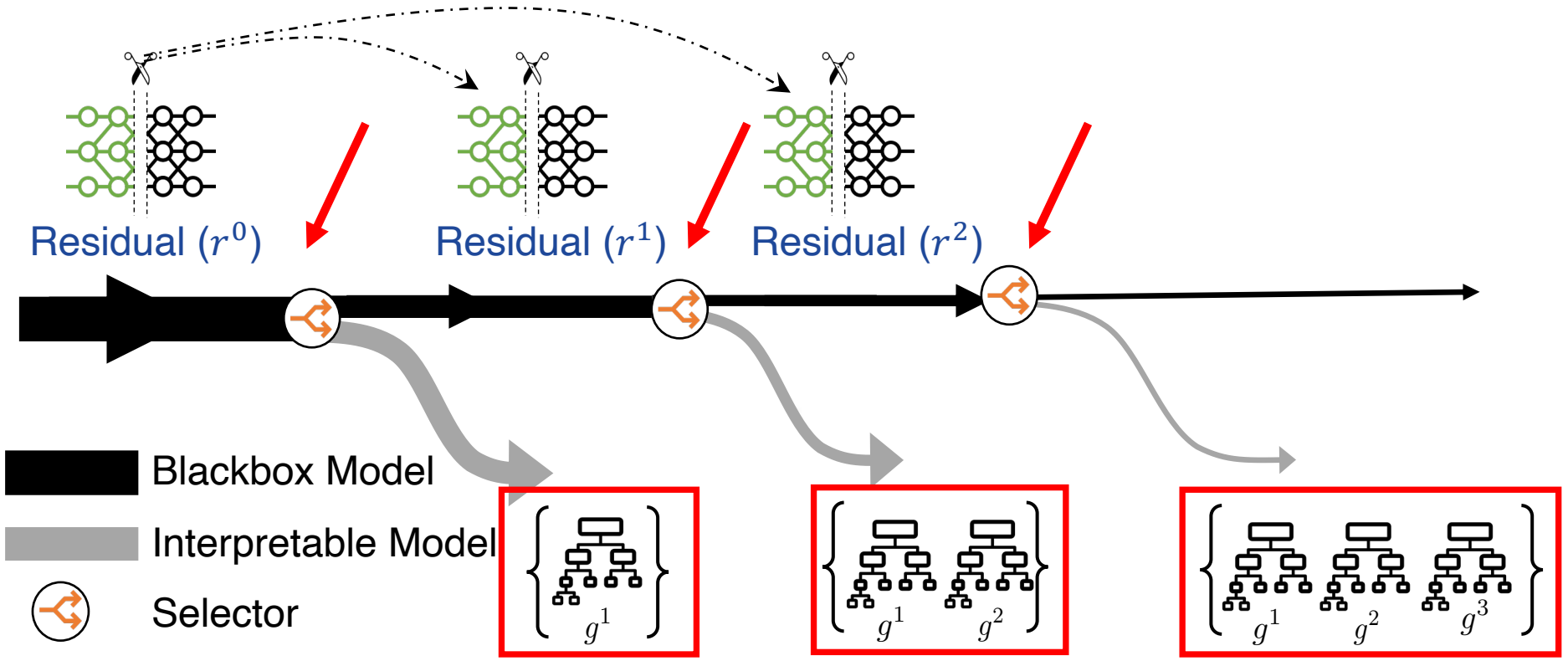
2. Use concepts from matching patients c



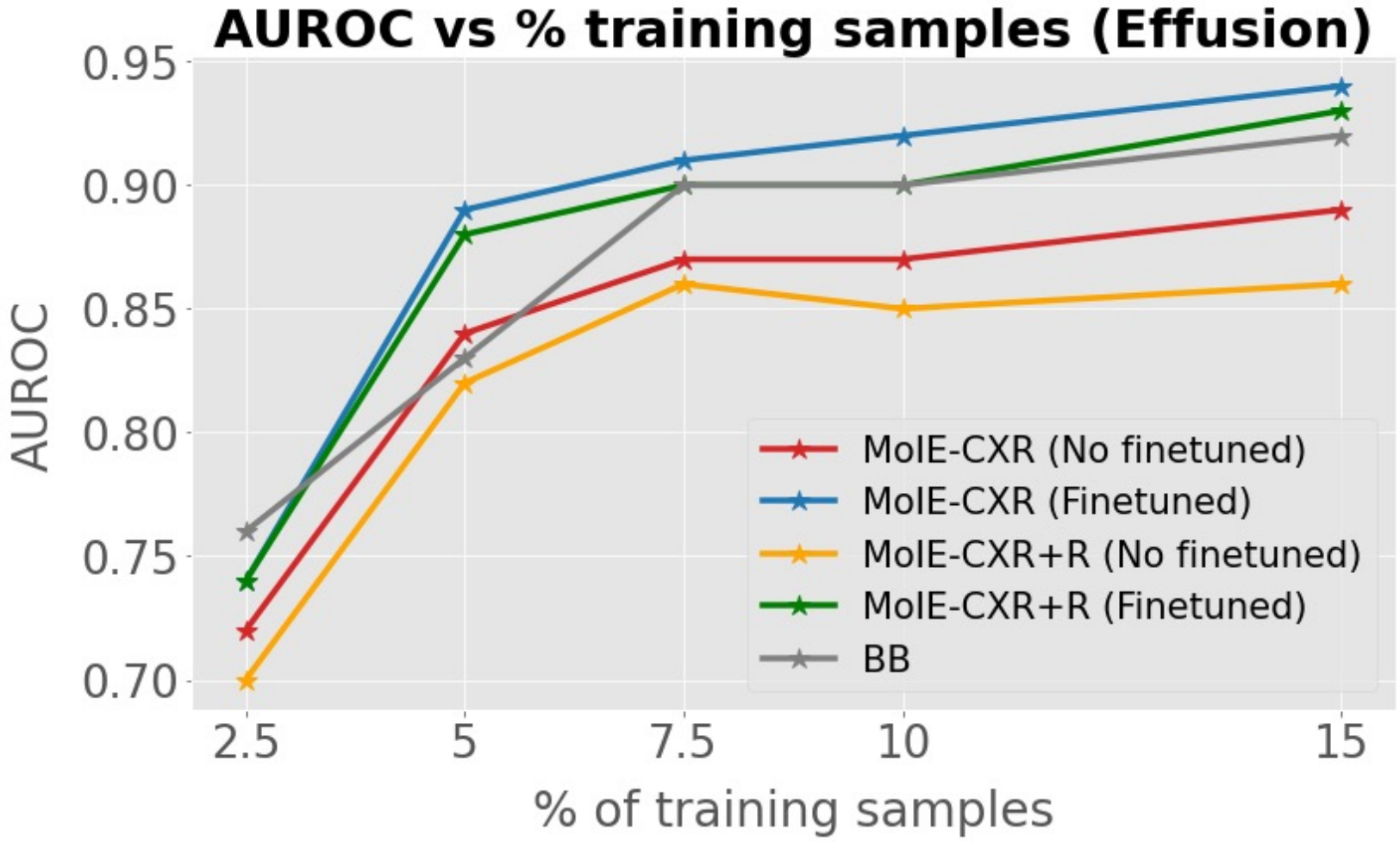
3. Propagate the concepts and update the concept extractor

Fine-tune to a New Domain

4 Update the selectors and interpretable models for 5 epochs

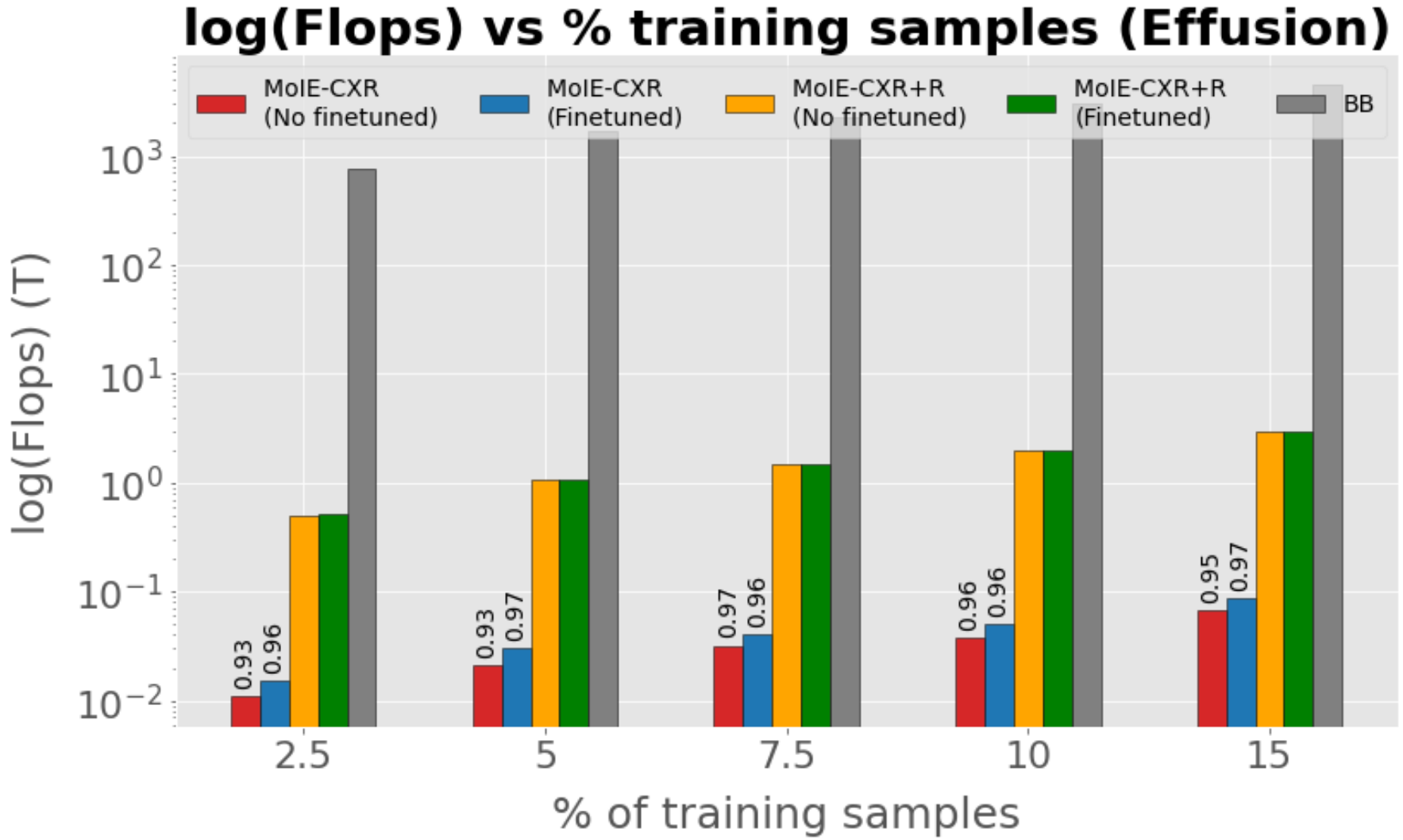


Transferring to Stanford-CXR

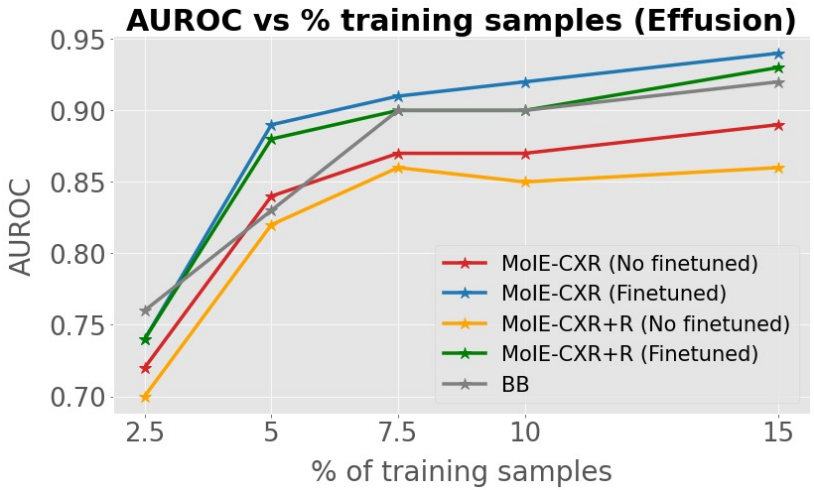


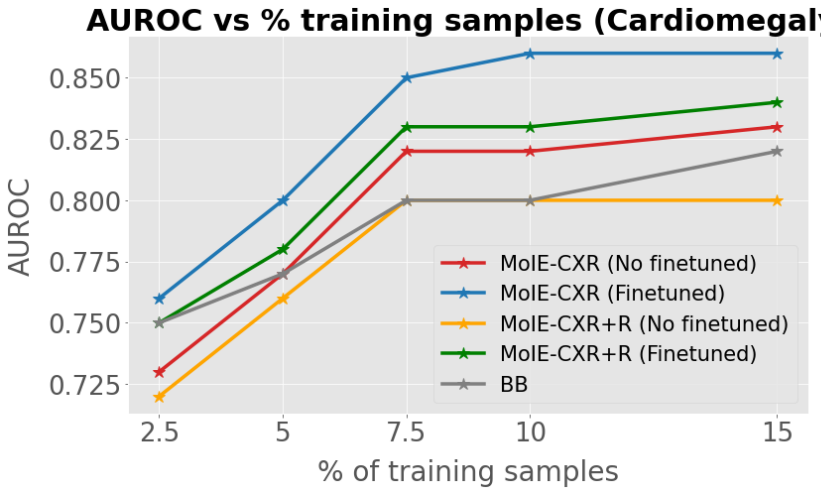
Distilling BlackBox to Interpretable models for Efficient Transfer Learning. Ghosh et al. MICCAI 2023. *Early accept*

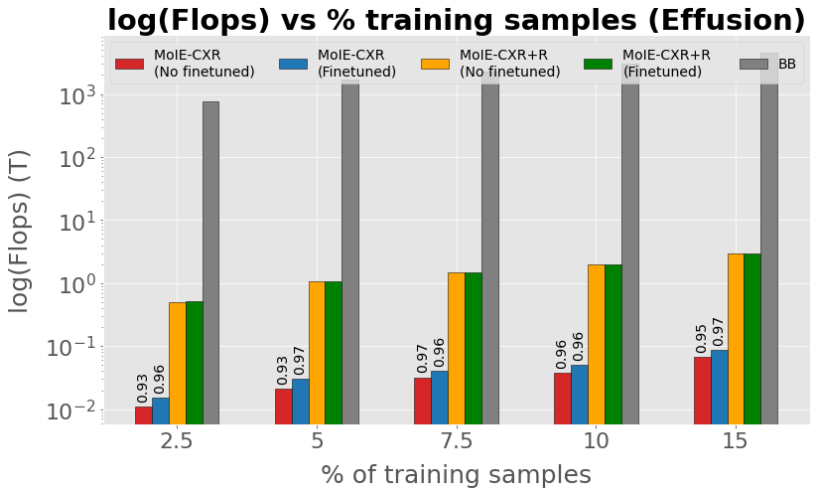
Transferring to Stanford-CXR



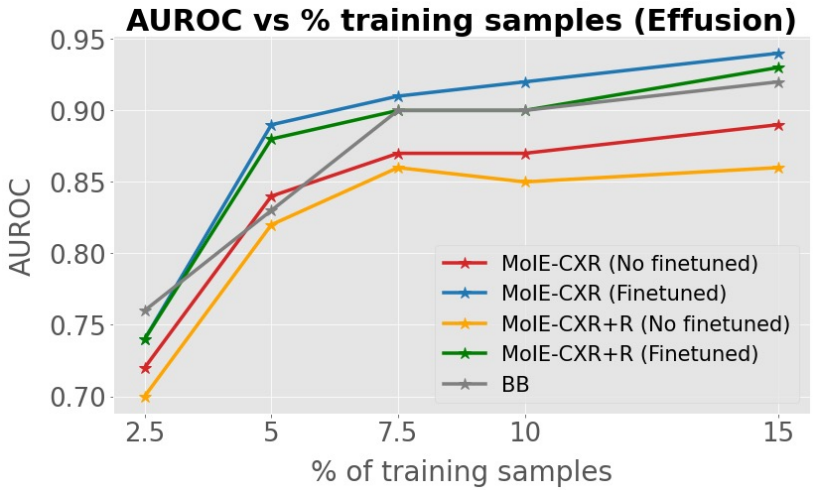
Transferring to Stanford-CXR

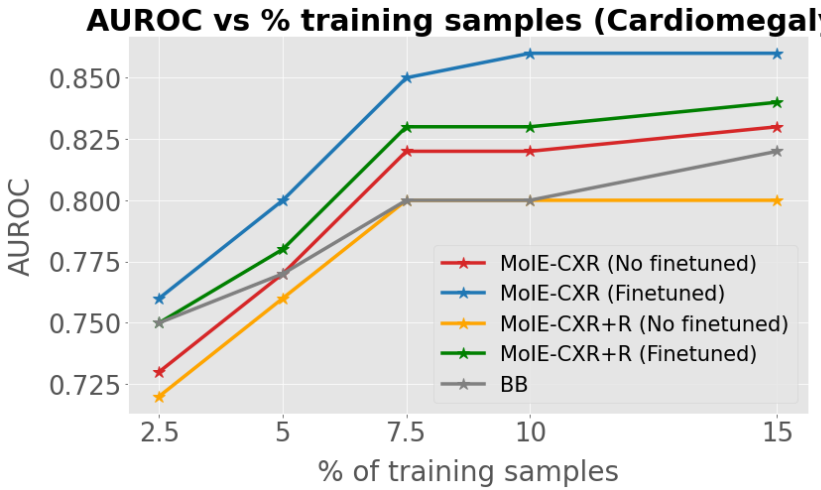


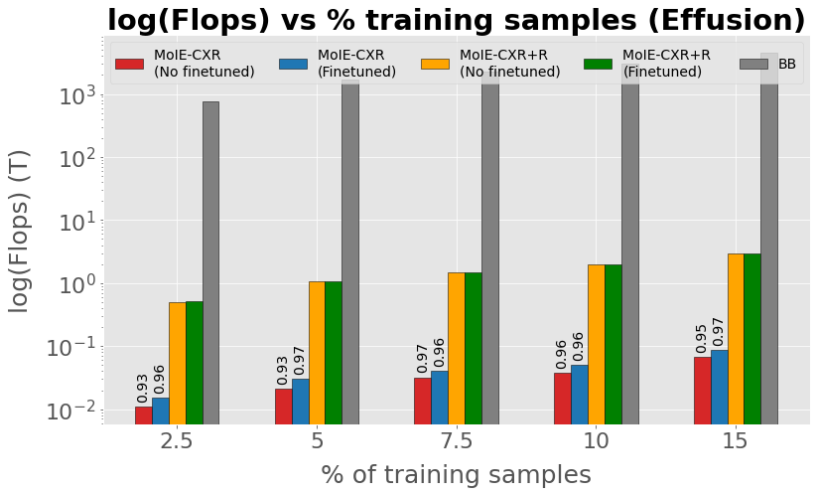


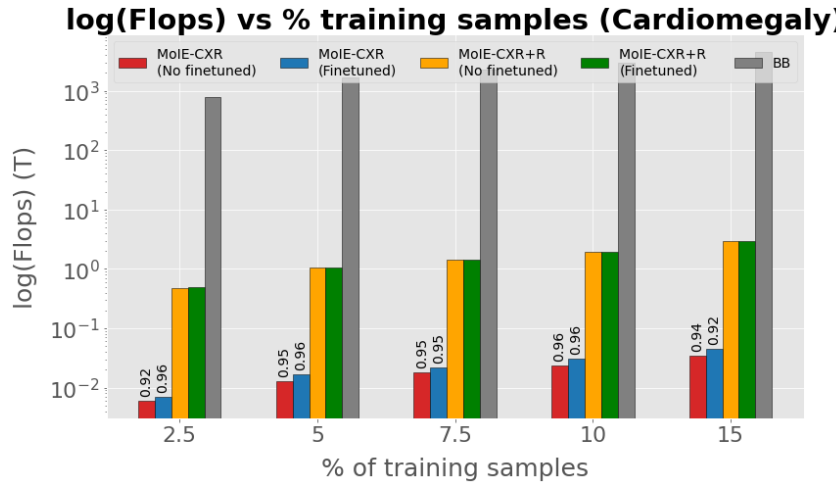


Transferring to Stanford-CXR

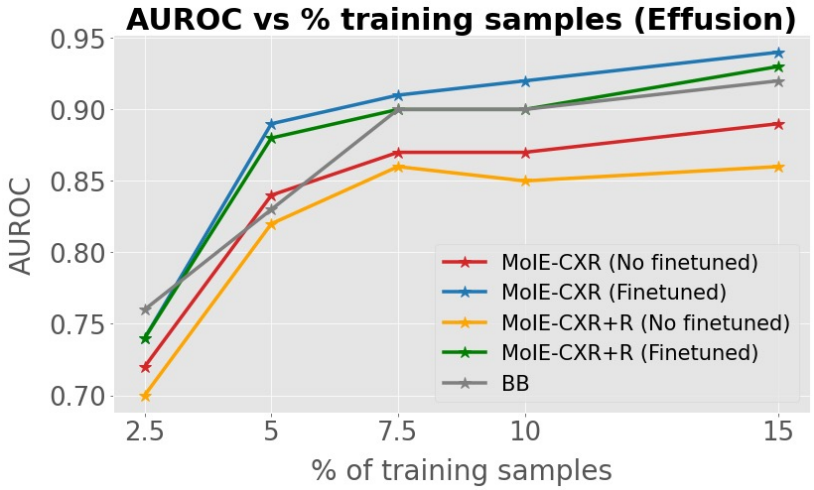


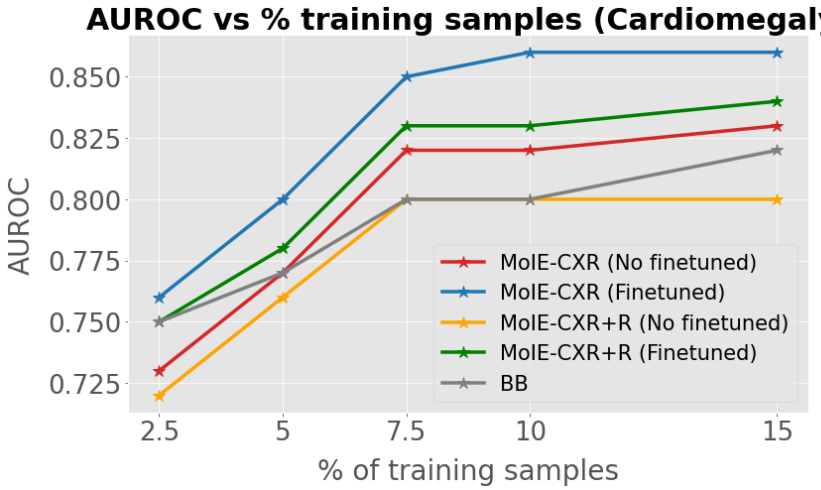


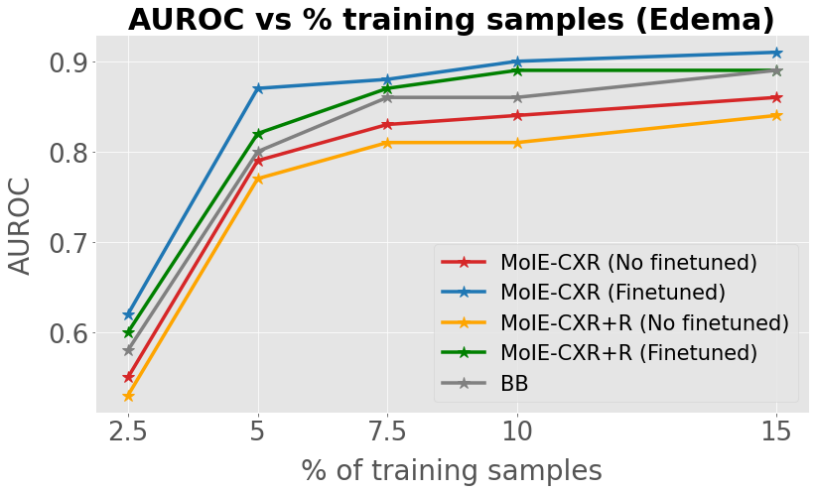


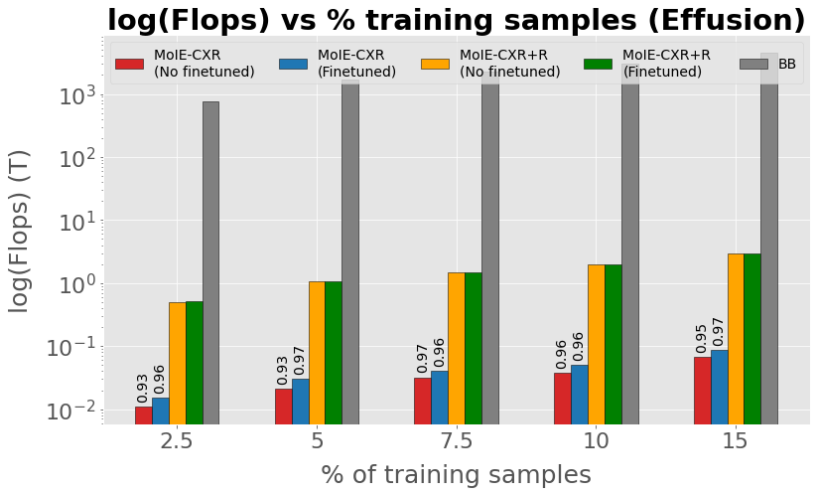


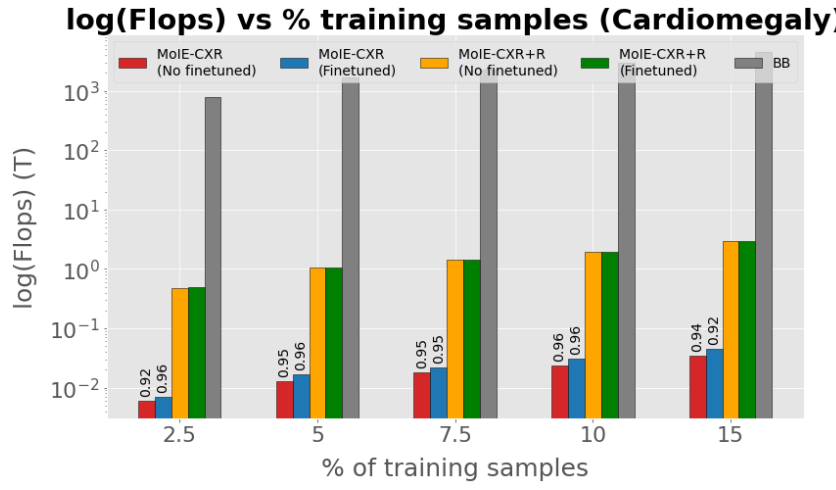
Transferring to Stanford-CXR



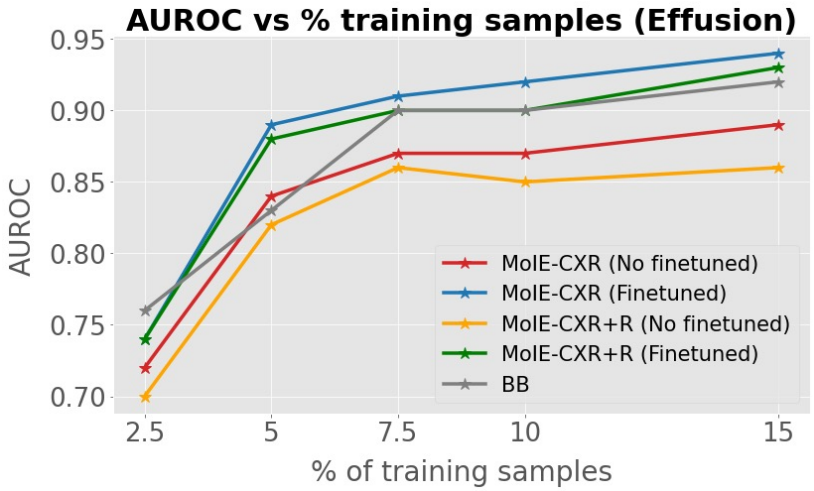


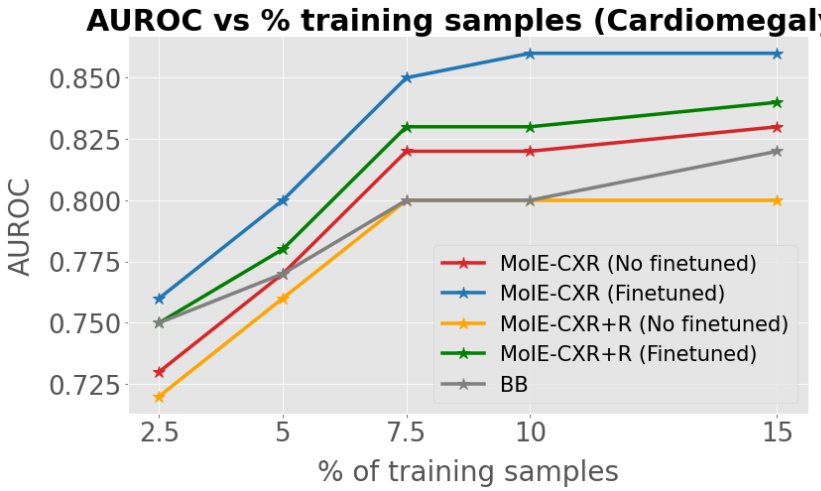


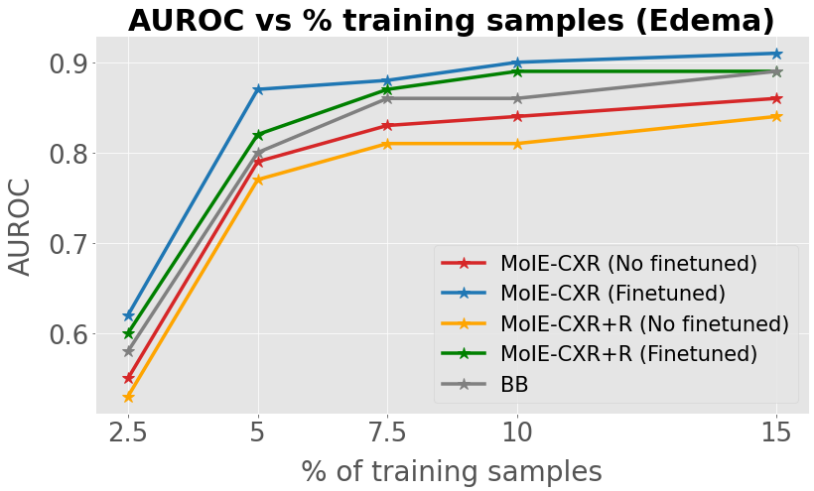


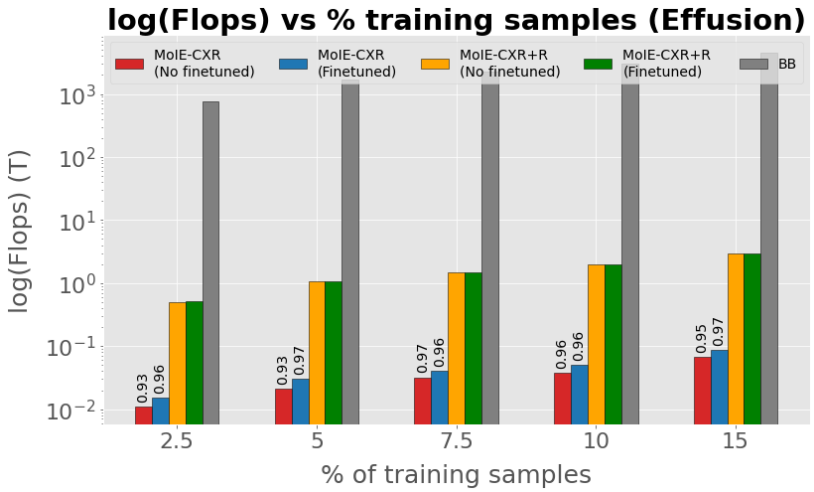


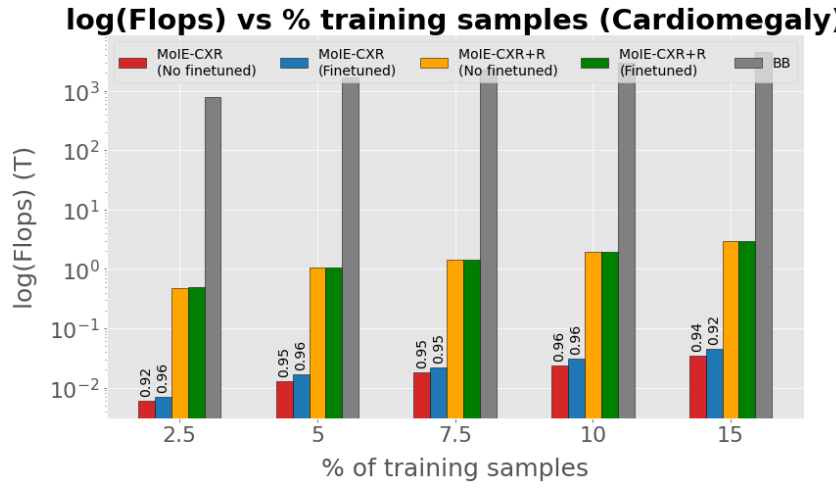
Transferring to Stanford-CXR

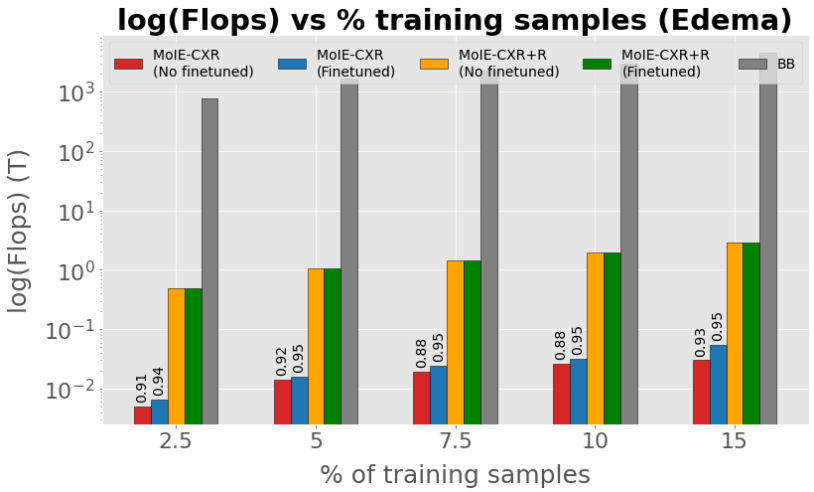












Conclusion

- Post-hoc approach allows the use of flexible Blackbox models, but post-hoc explainability is problematic and doesn't allow recourse.
- Interpretable-by-design allows recourse but suffer in performance.
- A mixture of interpretable models are *carved out* of a Blackbox model offering best of both worlds.
- Each interpretable model (*expert*) is modeled as First Order Logic (FOL), but other choices are possible for the interpretable model.
- Transfer Learning is more efficient (data & compute) with the new model.
- Intervention is possible with the new interpretable model.

References

- Ghosh S, Yu K, Arabshahi F, Batmanghelich K, “Dividing and Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat,” **ICML 2023**
- Ghosh S, Yu K, Batmanghelich K, “Distilling BlackBox to Interpretable models for Efficient Transfer Learning,” **MICCAI 2023, *Early accept ~ 13%***
- Ghosh S, Yu K, Arabshahi F, Batmanghelich K, “Tackling Shortcut Learning in Deep Neural Networks: An Iterative Approach with Interpretable Models,” **SCIS workshop, ICML 2023**
- Ghosh S, Yu K, Batmanghelich K, “Bridging the Gap: From Post Hoc Explanations to Inherently Interpretable Models for Medical Imaging,” **IMLH workshop, ICML 2023**

Future work

- We aim to use language to detect and fix the blackbox.
- We are interested in Breast imaging, i.e 2D mammograms
- We developed the first vision language model for mammograms, i.e Mammo-CLIP (Early accept, top 11%, MICCAI 2024)
- We will present our work at MICCAI 2024
  Date: Poster Session 5, Wednesday, October 9, 2024, 10:30 to 11:30
  Poster number: W-AM-169

Project: https://shantanu-ai.github.io/projects/MICCAI-2024-Mammo-CLIP/



Shantanu Ghosh[1], Ke Yu[2], Kayhan Batmanghelich[1]

[1]BU ECE, [2]Pitt ISP

BOSTON UNIVERSITY | BATMAN LAB

Thank you