Distilling Blackbox to Interpretable models for Efficient Transfer Learning

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

[1]Dept. Of Computer and Electrical Engineering, Boston University

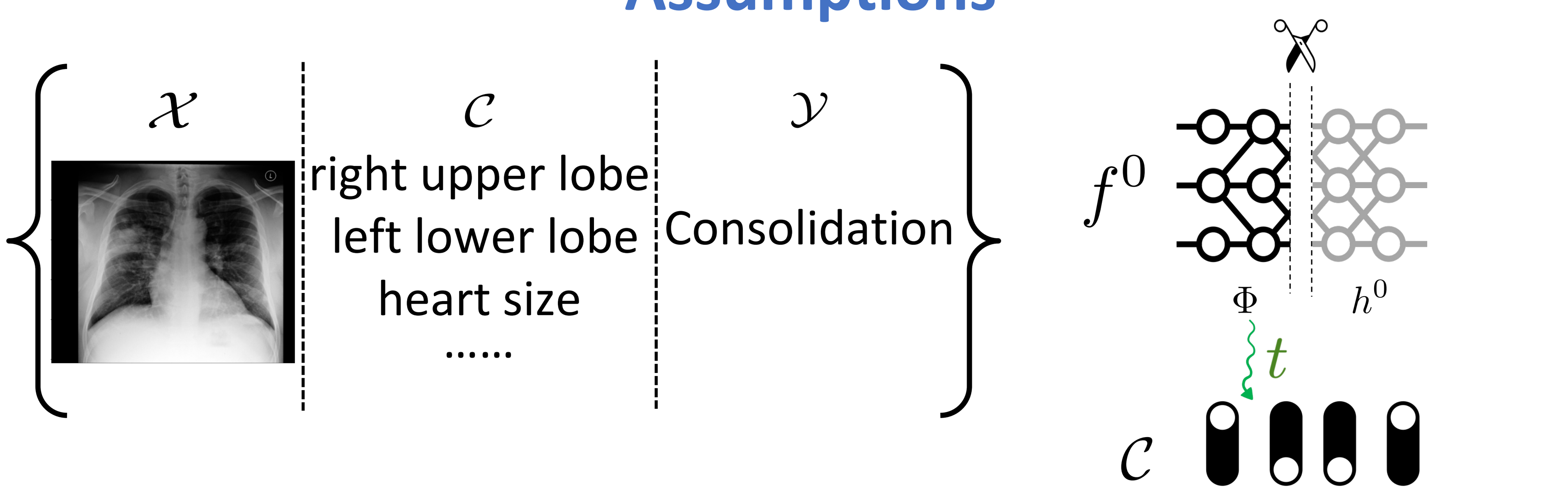
[2]Intelligent Systems Program (ISP), University of Pittsburgh

TLDR: Extracting a mixture of interpretable models from a BlackBox to provide concept-based explanations for efficient transfer learning.

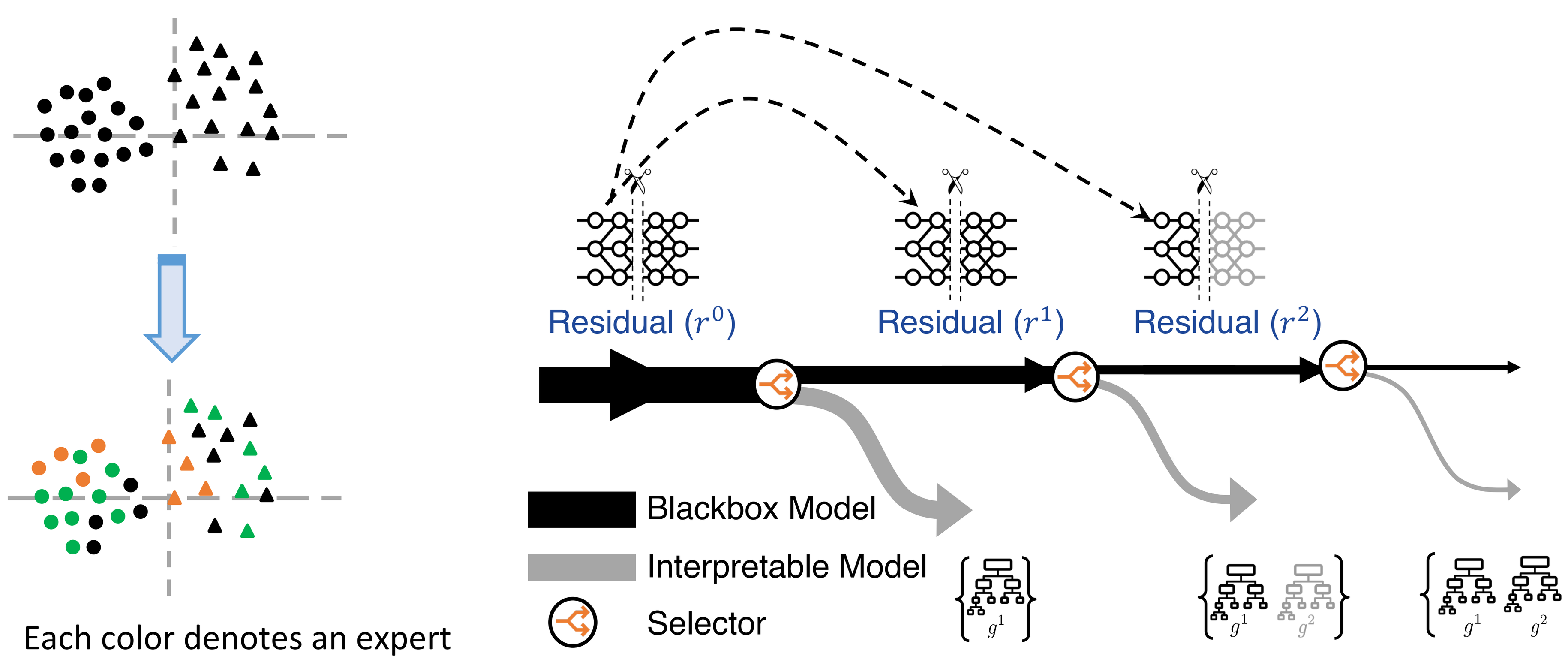
Motivation

- Neural Networks fail to generalize due to scanner types, disease subtypes, patient subpopulation.
- Fine-tuning a Blackbox to a new domain can solve this issue.
- This is data and computationally expensive.
- Whole process is not interpretable.
- Radiologists search for patterns of anatomical changes and apply generalizable logical rules for disease diagnosis.

Assumptions



Carve out interpretable models from Black box



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

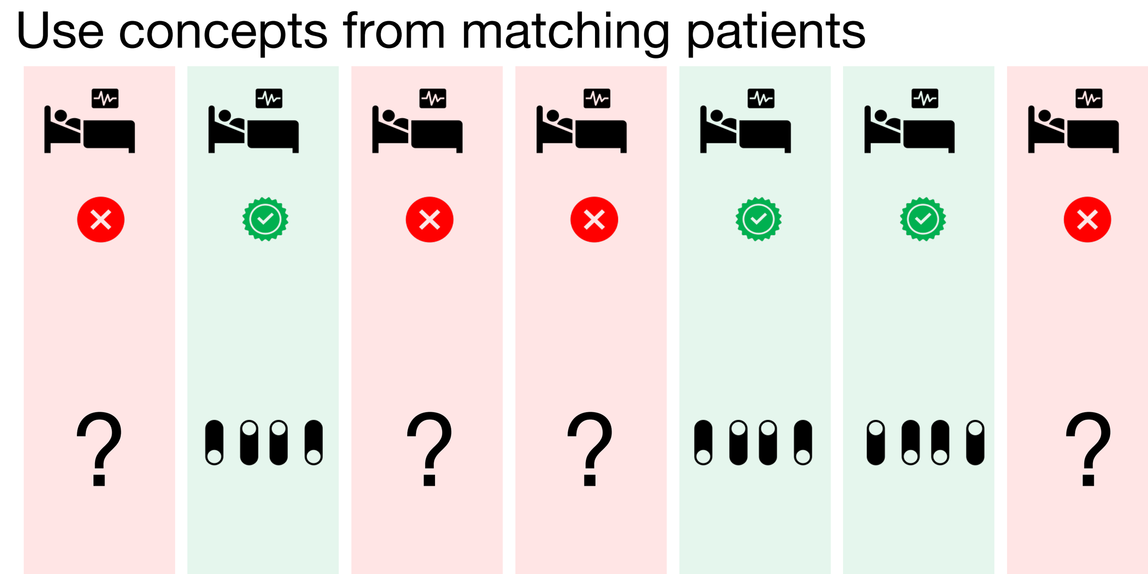
*SelectiveNet [Geifman et al.] optimization

*Continue till at least 90% samples covered

*The experts are trained sequentially.

Data efficient Transfer Learning

1. Apply source black box on the target domain.
2. Use concepts from matching patients



3. Propagate the concepts and update the concept extractor
4. Update the selectors and the experts for 5 epochs on the target domain.

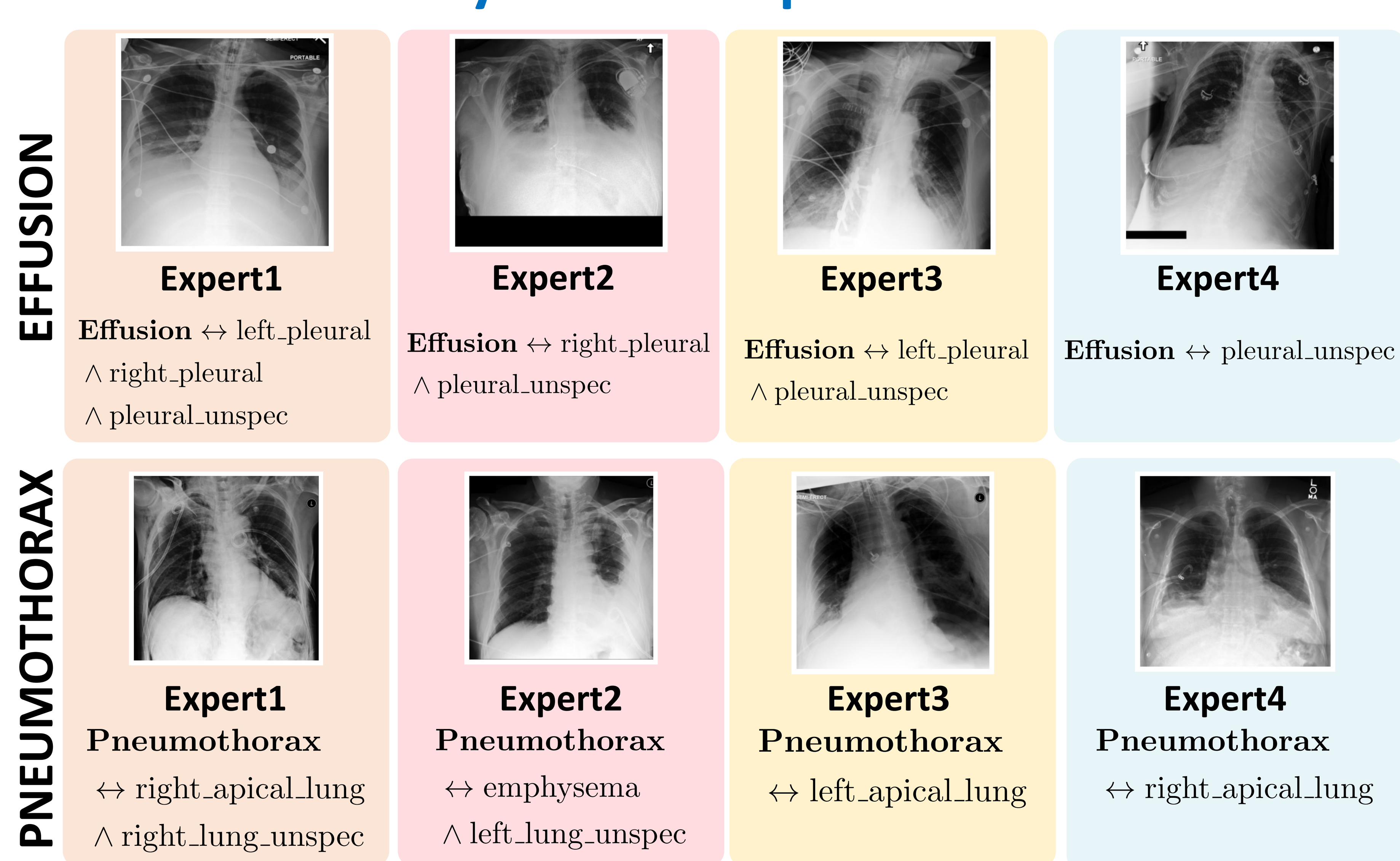
Extract concepts from MIMIC-CXR using Radgraph NLP pipeline

Ke Yu et al., MICCAI, 2022

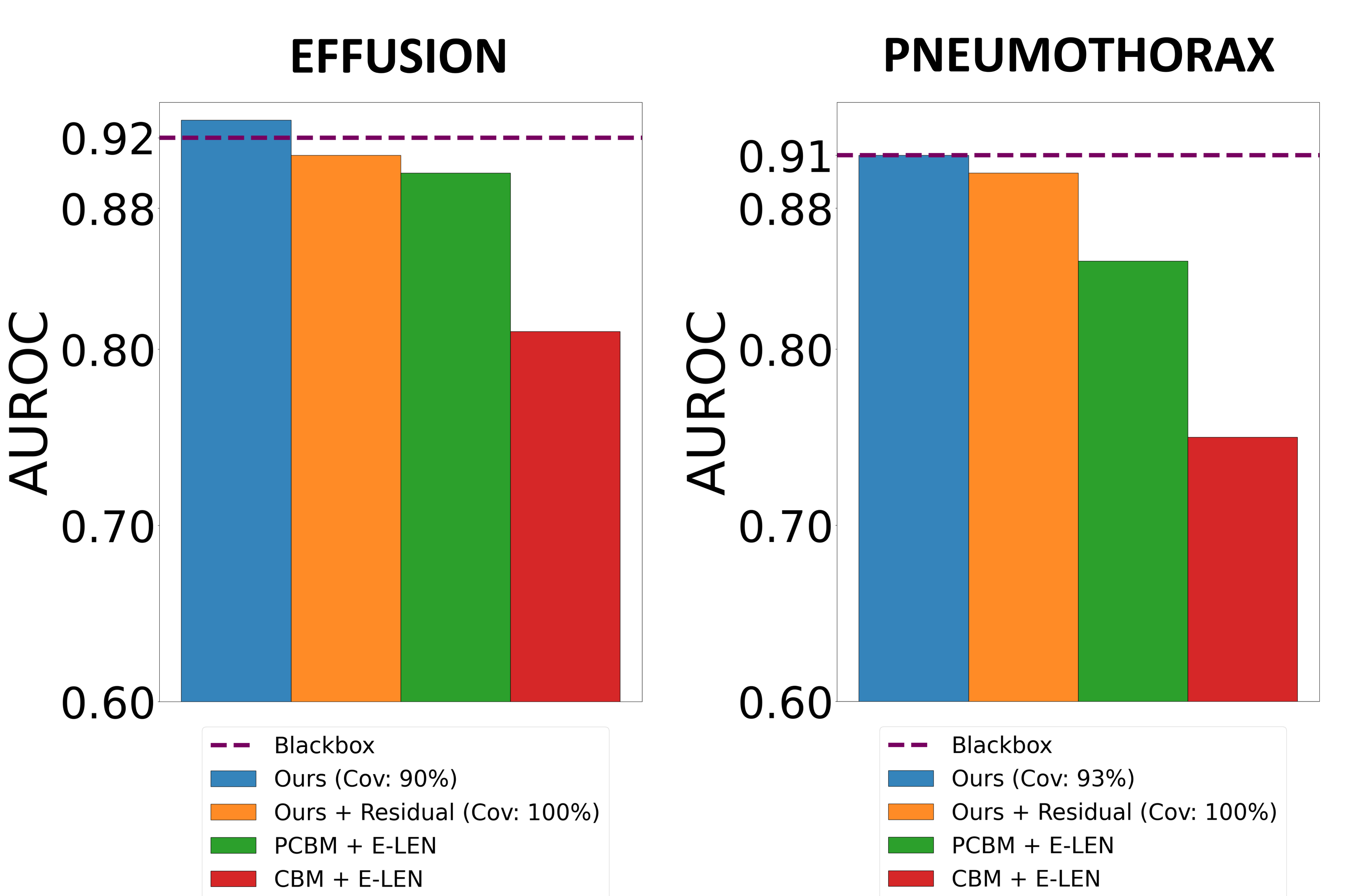
Report:

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

Diversity in local explanations



Not compromising the accuracy in MIMIC-CXR



Transferring the first 3 experts of MIMIC-CXR to Stanford-CXR