Explaining, **Fast** and **Slow**

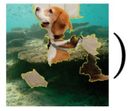
Abstraction and Refinement of *Provable* Explanations

Shahaf Bassan*, Yizhak Elboher*, Tobias Ladner*, Matthias Althoff, Guy Katz

Hebrew University of Jerusalem (HUJI),
Technical University of Munich (TUM)

An efficient algorithm to find *minimal* and *sufficient* explanations in neural networks with ***provable guarantees***. It *abstracts* the model to a smaller version, then *refines* it by growing its size, ensuring ***sufficiency*** and ***minimality***.

What does ***provably sufficient*** mean?

$f($  $) =$ Dog $f($  $) =$ Dog

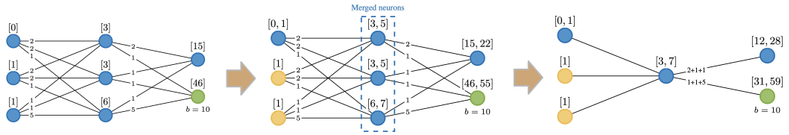
$f($  $) \overset{?}{=}$ Dog

What does ***provably minimal*** mean?

Any subset of  is **not** sufficient!

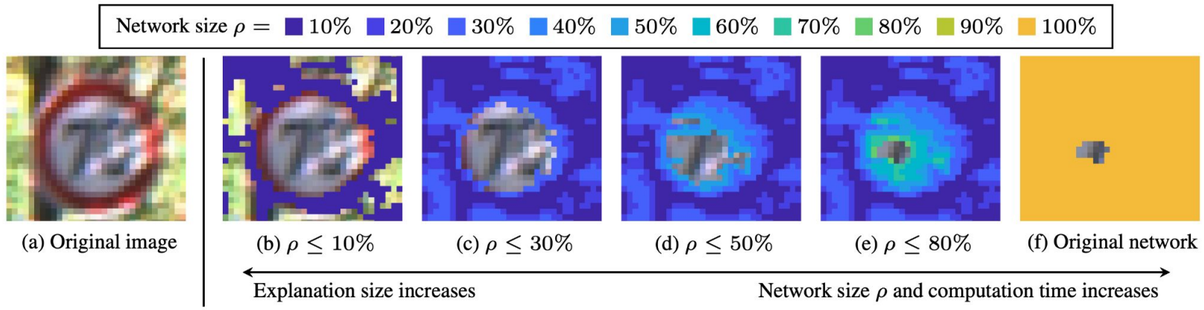
Previous algorithms struggle with **high compute** (more than **NP-hard**) especially for **large models**.

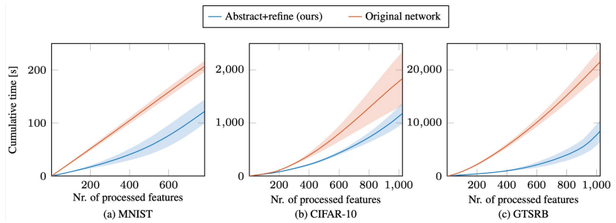
Our method: ***Abstract*** the model (merge neurons and build a **much smaller one**).



We prove: A **sufficient explanation** for the **abstract (small)** model is **sufficient** for the **original**!
But a **minimal** one **might not stay minimal**.

This means we need to iteratively ***refine*** the model (increase its size) to reach **minimality**.





- Significantly ***faster*** than previously suggested algorithms.
- Significantly ***smaller*** explanations compared to previous approaches.
- Scales to ***larger*** models.