Explain Yourself, **Briefly!**

Self-Explaining Neural Networks with Concise Sufficient Reasons

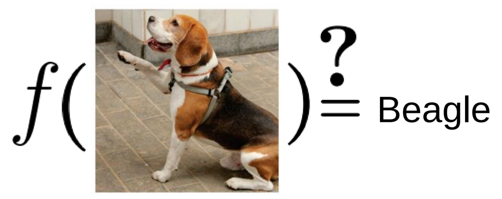
Shahaf Bassan, Ron Eliav, Shlomit Gur

IBM Research,
The Hebrew University of Jeruslaem,
Bar-Ilan University

A training technique for self-explaining neural networks that inherently generate concise “sufficient reasons” – minimal subsets that by themselves determine the prediction.

A

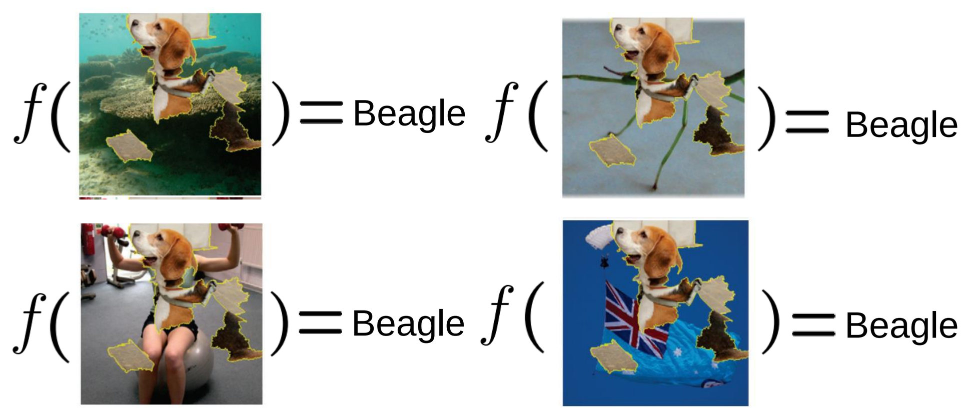
What are *sufficient reasons*?



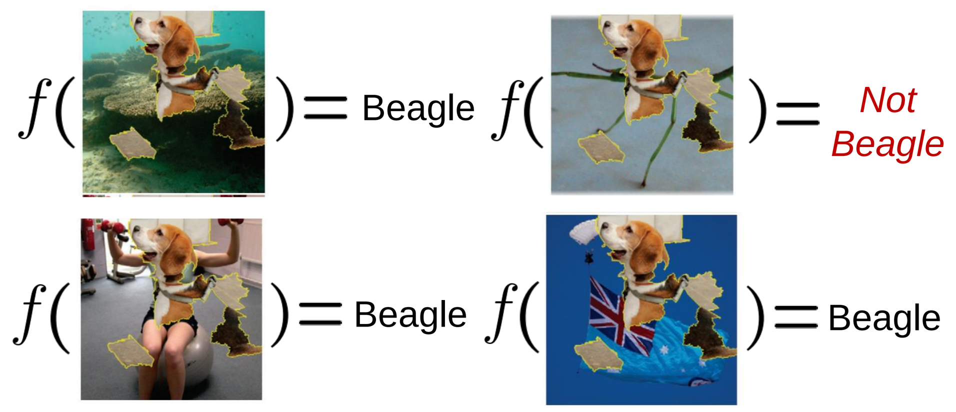
Minimal subsets of input features that by themselves are *sufficient* to determine the prediction.

1. ***Baseline*** Sufficient Reasons

2. ***Robust*** Sufficient Reasons



3. ***Probabilistic*** Sufficient Reasons



B

Problems with *post-hoc* sufficient reason methods

Problem 1: *Intractability*

Strikingly computationally hard to compute.

We prove that:

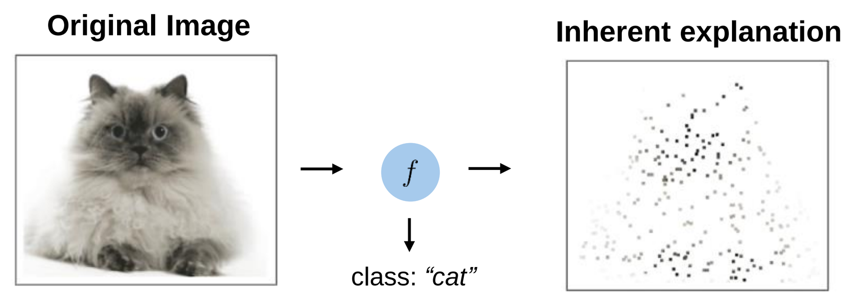
- Computational hardness results extend from *binary* CNF to neural networks over *continuous* domains.
- Hardness holds even in highly “simplified” settings such as for *baseline* sufficient reasons.
- These hardness results hold even when *approximating* the size of sufficient reasons.

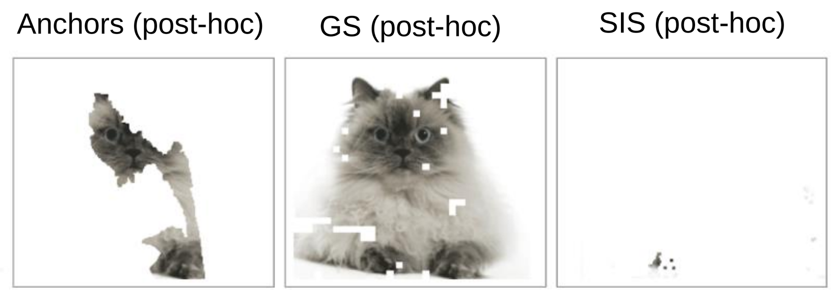
Problem 2: *OOD sampling sensitivity*

High sensitivity to Out-Of-Distribution (OOD) inputs

C

A *self-explaining* approach to address these problems



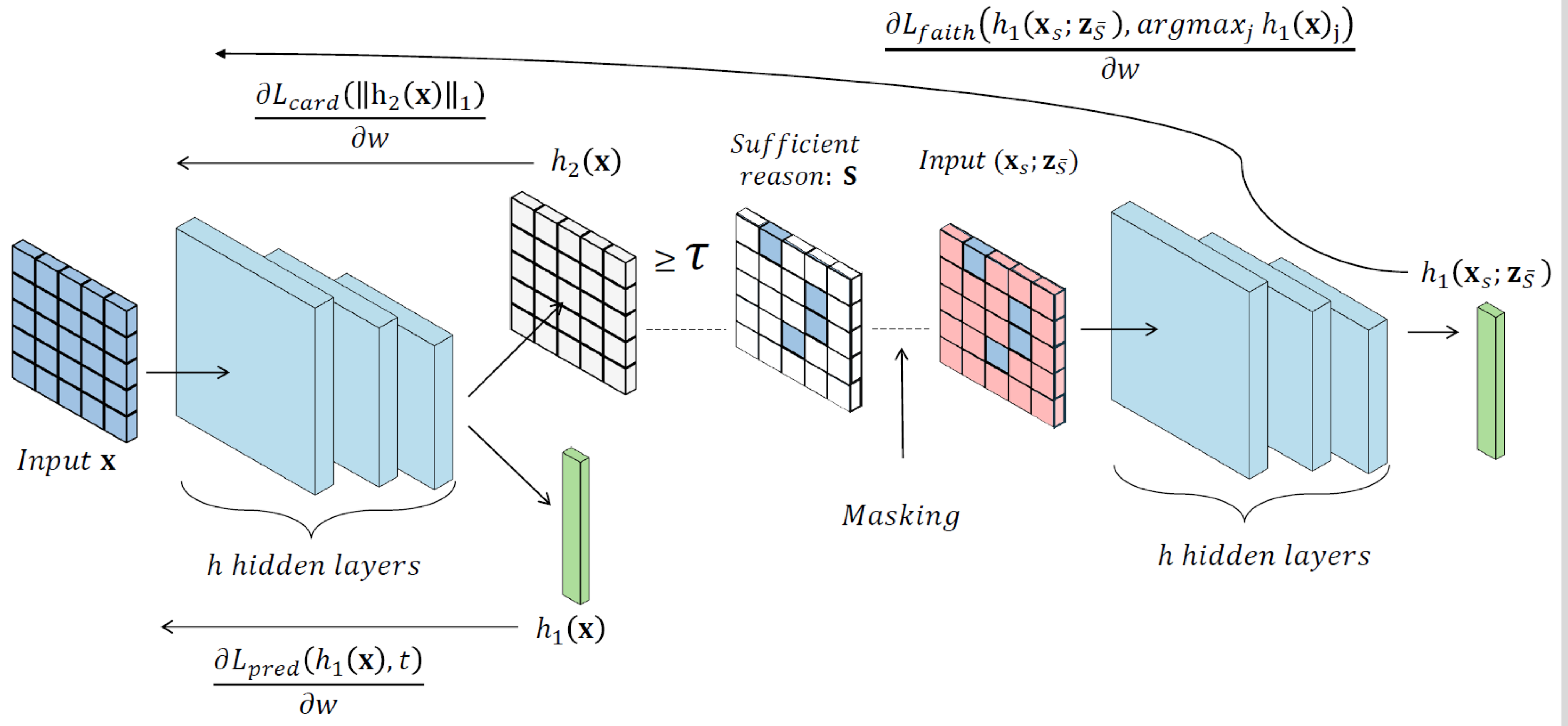


✓ Inherent generation → Efficient and scalable computation.

✓ Partial input exposure during training → Less sensitivity to OOD sampling.

D

Sufficient Subset Training (SST)



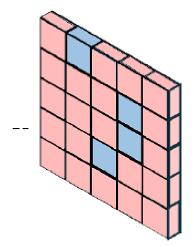
$$L_\theta(h) := L_{pred}(h) + \lambda L_{faith}(h) + \xi L_{card}(h)$$

Prediction Loss: $L_{CE}(h_1(\mathbf{x}), t)$

Faithfulness Loss: $L_{CE}(h_1(\mathbf{x}_S; \mathbf{z}_{\bar{S}}), \arg\max h_1(\mathbf{x})_j)$

Cardinality Loss: $\|h_2(\mathbf{x})\|_1$

How is the masking performed?



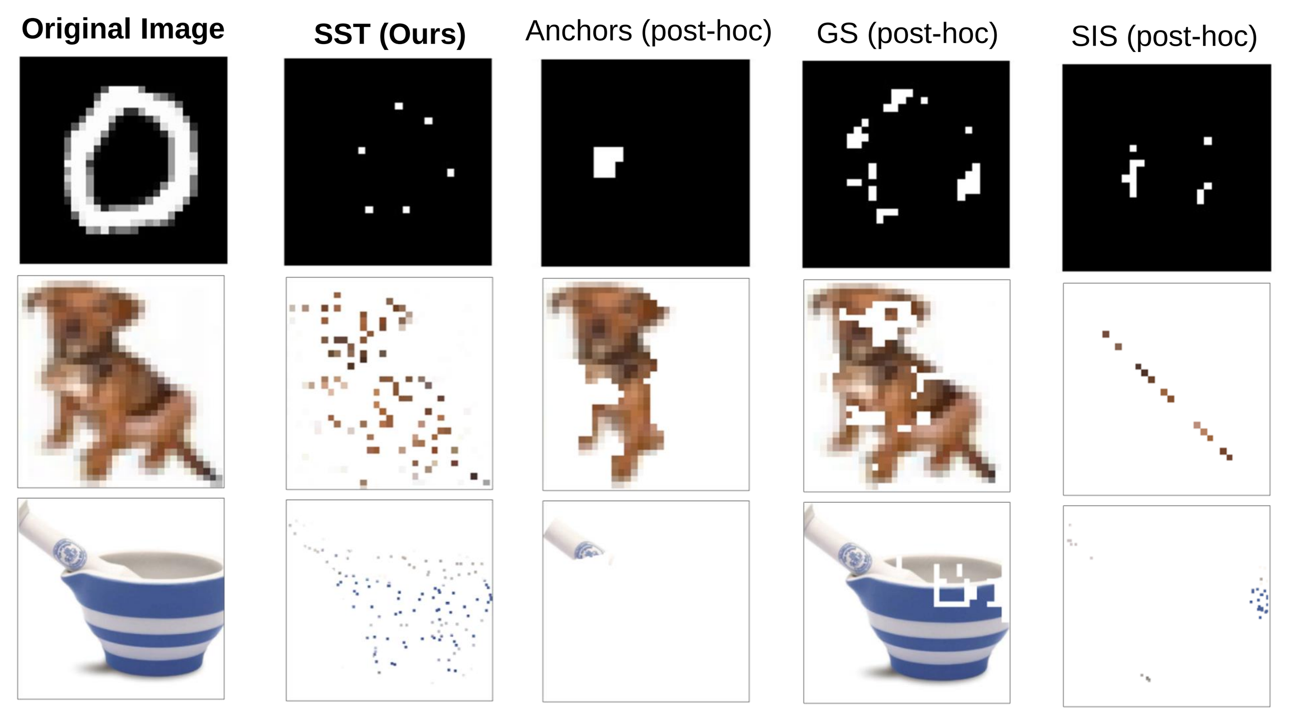
- *Baseline masking* - fix some baseline to the complementary.
- *Probabilistic masking* – sample values to the complementary from some distribution.
- *Robust masking* – perform an adversarial gradient attack over the complementary.

E

Experimental Analysis

- Experiments on three vision (MNIST, CIRAR, IMAGENET) and two NLP (IMDB, SNLI) tasks.
- Evaluation metrics: (1) *generation time*, (2) *faithfulness* (% sufficiency), (3) *explanation size*.
- Comparison to post-hoc methods.

SST vs. Post-hoc examples:



Probabilistic vs. Baseline (using MASK token) examples:

Baseline-Sufficiency (negative):
I *found this* movie really hard … wandering *off* the tv … *Don't* bother with it.

Probabilistic-Sufficiency (negative):
I *found* this movie *really hard* … *wandering off the tv* …

✓ ***Significantly higher faithfulness*** compared to post-hoc methods.

✓ ***Significantly smaller explanations*** than post-hoc methods.

✓ Explanations produced ***substantially faster*** than post-hoc methods.

✓ SST retained ***comparable predictive performance***.