האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

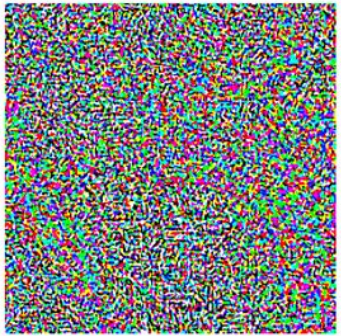
"Formal XAI": Can We **Formally** Explain ML Models?

Shahaf Bassan

ML and NNs are used extensively

NNs are sensitive to adversarial attacks

$+ .007 \times$



$=$

"panda"

57.7% confidence

noise

"gibbon"

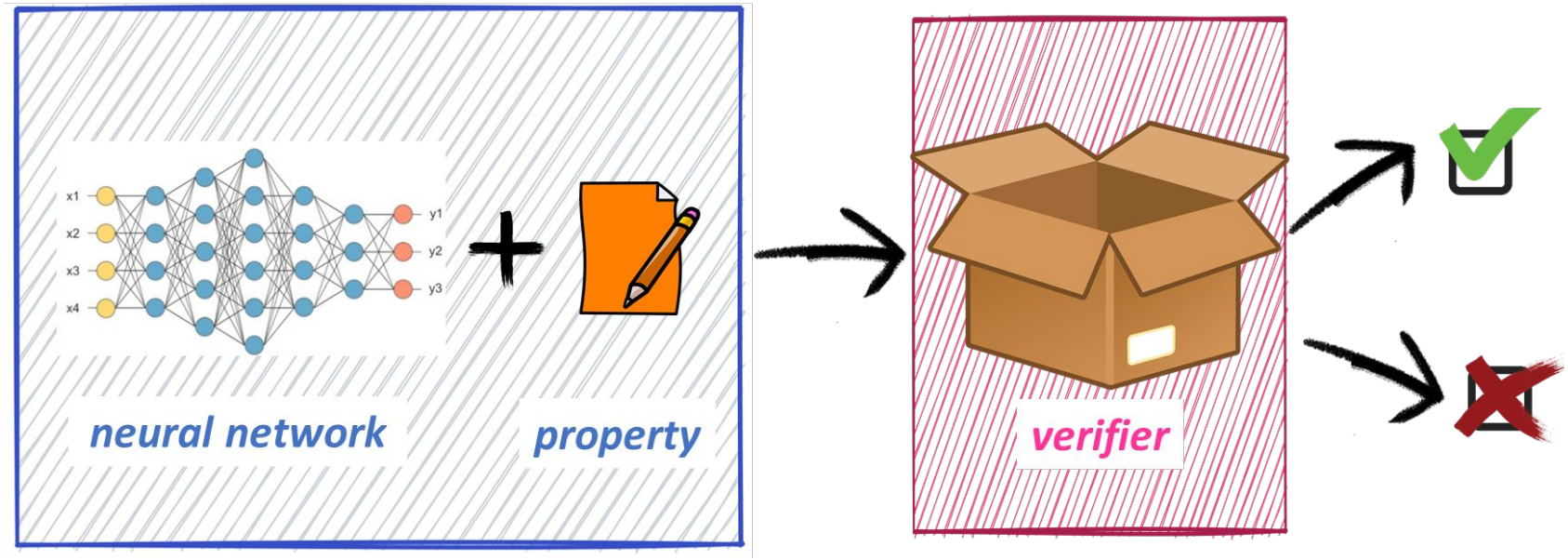
99.3% confidence

NNs are sensitive to adversarial attacks

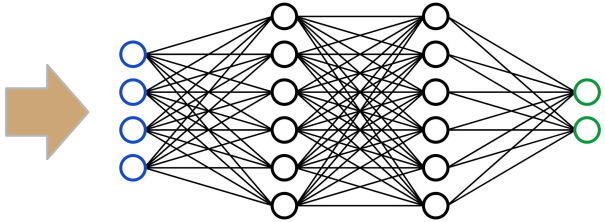
Cup(16.48%)
Soup Bowl(16.74%)

Bassinet(16.59%)
Paper Towel(16.21%)

Neural Network Verification



The classic property: Adversarial Robustness

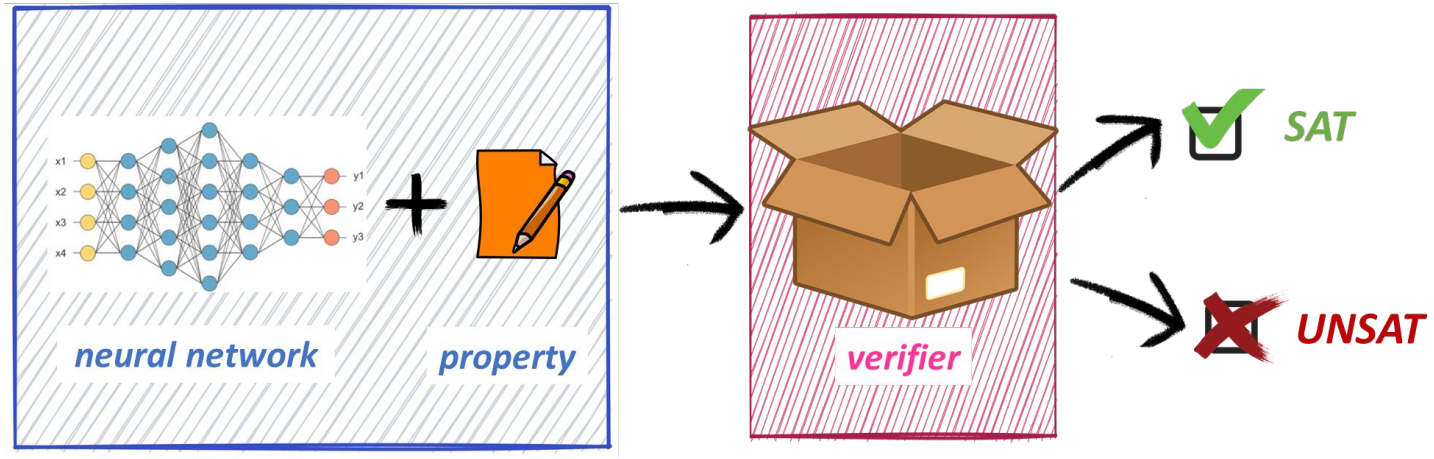


“Panda”

Property: Verify that there isn’t an **adversarial perturbation** around:

$$\mathcal{E} = 0.008$$

How hard is NN-verification?



NN-verification is NP-complete! (Katz et al., 2017)

How hard is NN-verification?

- Marabou (Katz et al.)
- Beta- Crown (Wang et al)
- DeepPoly (Singh et al.)
- AI2 (Gehr et al.)
- Prima (Muller et al.)
- Verinet (Henriksen et al.)
- MN-BAB (Ferrari et al.)
- Nnenum (Bak et al.)

2017:
<100 neurons

How hard is NN-verification?

- Marabou (Katz et al.)
- Beta- Crown (Wang et al)
- DeepPoly (Singh et al.)
- AI2 (Gehr et al.)
- Prima (Muller et al.)
- Verinet (Henriksen et al.)
- MN-BAB (Ferrari et al.)
- Nnenum (Bak et al.)

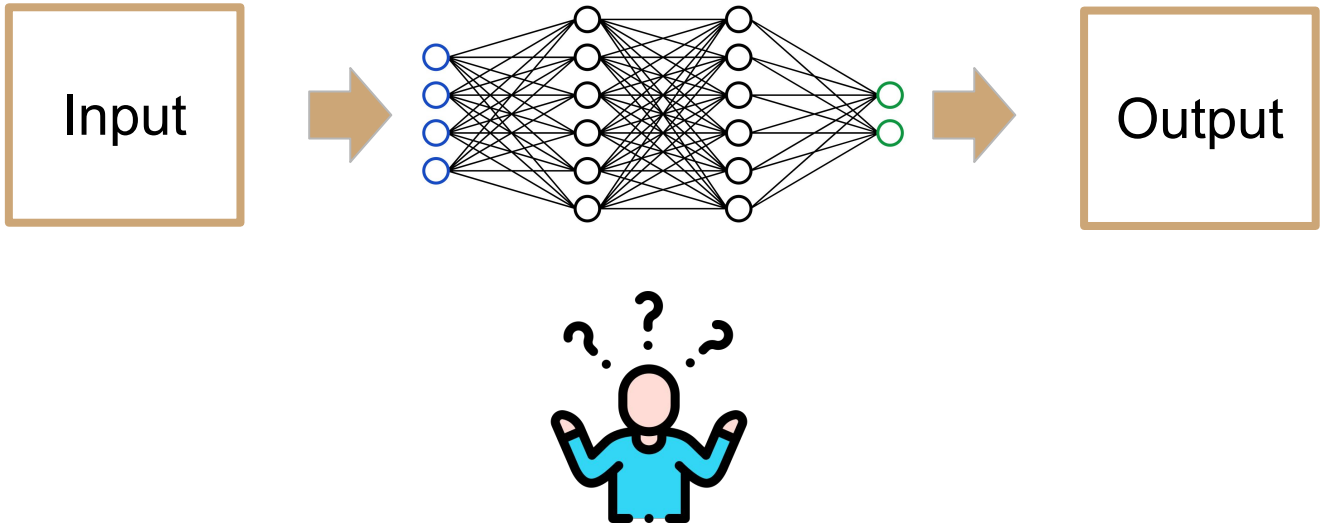
2017:
<100 neurons

2023-2025:
~10,000,000
neurons

Still not applicable on SOTA

From Formal Verification to
Formal Explainability

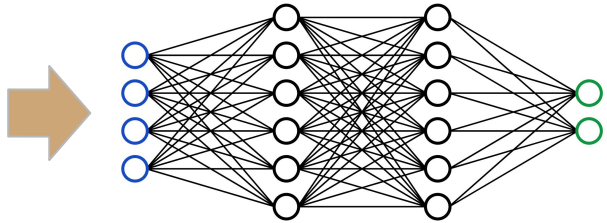
Neural Networks are Black-Boxes



Explainable AI (XAI)

Tools and frameworks for explaining the decisions made by ML models.

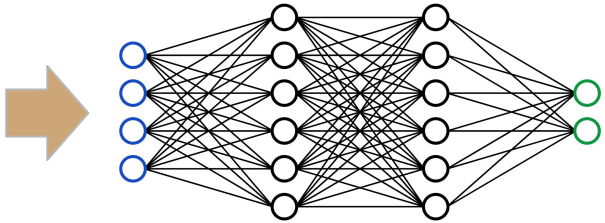
AI Explainability example



“Labrador”

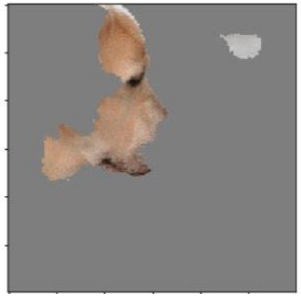
(Ribeiro et al., 2016)

AI Explainability example



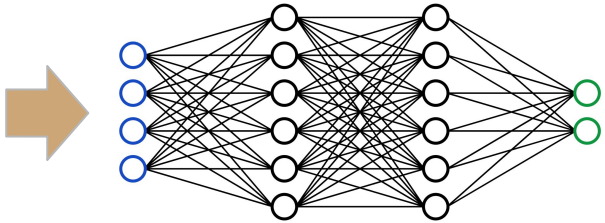
“Labrador”

Explainer:



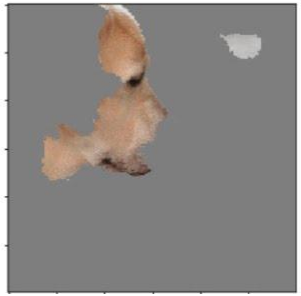
(Ribeiro et al., 2016)

AI Explainability example



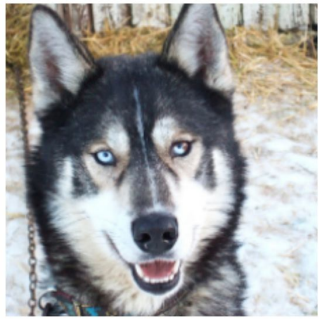
“Labrador”

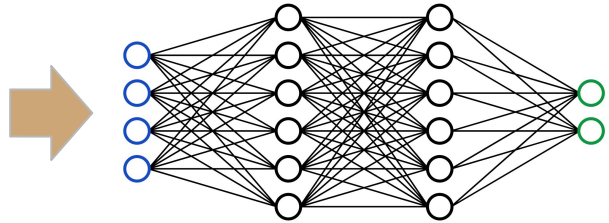
Explainer:



(Ribeiro et al., 2016)

AI Explainability example

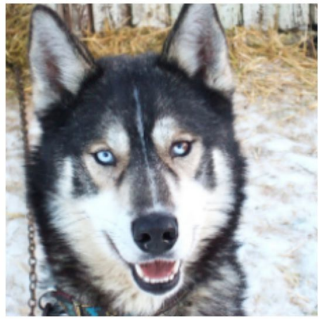


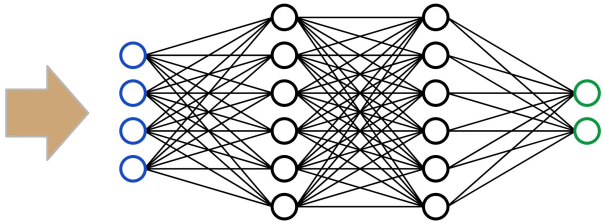


“Wolf”
(and not Husky)

(Ribeiro et al., 2016)

AI Explainability example





“Wolf”
(and not Husky)

Explainer:

(Ribeiro et al., 2016)

Can we trust XAI tools?

XAI tools are often **heuristic**, and do not provide **formal guarantees**.

Can we trust XAI tools?

If we can’t trust the **explainer**,
we can’t trust the **model**.

Formal Explainability

We hence want to produce **formal and provable** explanations.

How can we formally define an explanation?

$f($  $) \stackrel{?}{=}$ “Beagle”

(Ribeiro et al., 2018)

Sufficient Reason/Abductive Explanation

$f($  $) \stackrel{?}{=}$ “Beagle”

$$\forall \mathbf{z} \in B_p^{\epsilon_p}(\mathbf{x}). \quad [\arg\max_j f(\mathbf{x}_S; \mathbf{z}_{\bar{S}})_j = \arg\max_j f(\mathbf{x})_j]$$

Sufficient Reason:

(Ribeiro et al., 2018)

Sufficient Reason/Abductive Explanation

$$\forall \mathbf{z} \in B_p^{\epsilon_p}(\mathbf{x}). \quad [\arg\max_j f(\mathbf{x}_S; \mathbf{z}_{\bar{S}})_j = \arg\max_j f(\mathbf{x})_j]$$

f () = “Beagle”

f () = “Beagle”

(Ribeiro et al., 2018)

Abductive Explanations/Sufficiency

DNN-verification can be used to verify

if  is a sufficient reason.

Minimal Sufficient Reasons

1. Smaller sufficient reasons are more meaningful.
2. We are interested in finding **subset minimal** or preferably **cardinally minimal** explanations.

A framework for approximating explanations

Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks

Tacas, 2023
Shahaf Bassan, Guy Katz

A framework for approximating explanations

Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks

Tacas, 2023
Shahaf Bassan, Guy Katz

Provable subset minimal sufficient reasons

A framework for approximating explanations

Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks

Tacas, 2023
Shahaf Bassan, Guy Katz

Provable subset minimal sufficient reasons

Provable approximation of cardinally minimal sufficient reasons

A framework for approximating explanations

Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks

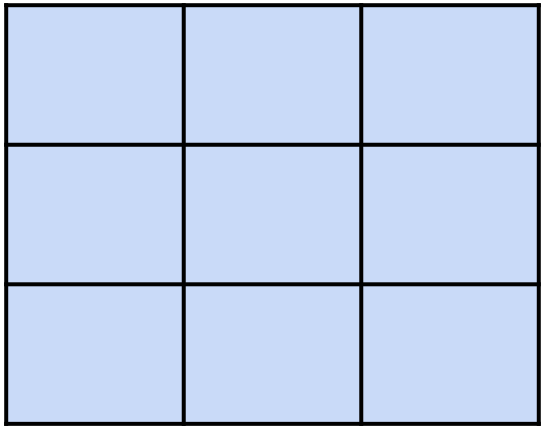
Tacas, 2023
Shahaf Bassan, Guy Katz

Provable subset minimal sufficient reasons

Provable approximation of cardinally minimal sufficient reasons

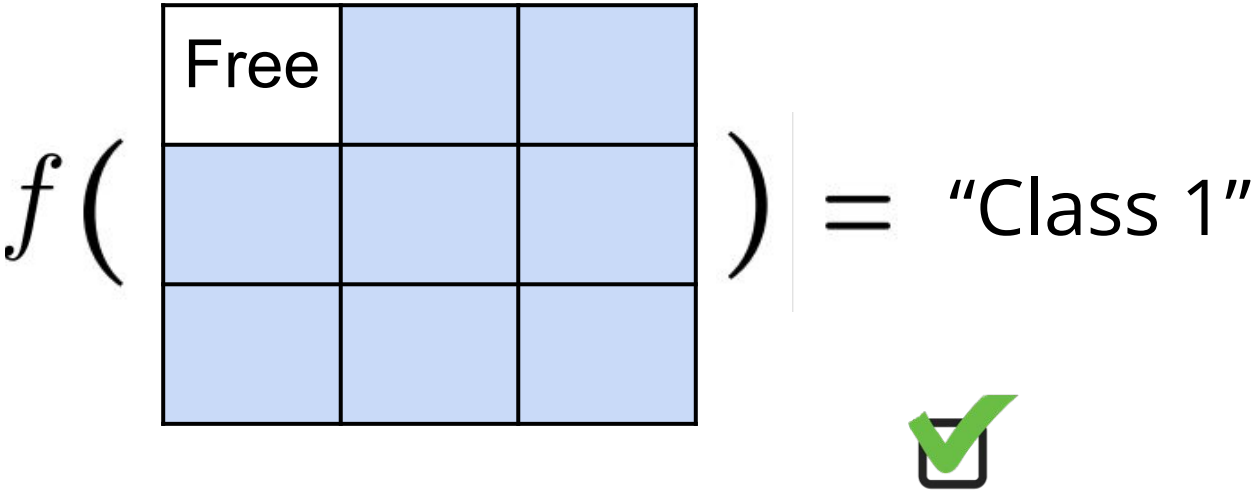
Subset minimal sufficient reasons

Attempt to ``free" features until converging to a subset minimal explanation.

$f($  $) =$ “Class 1”

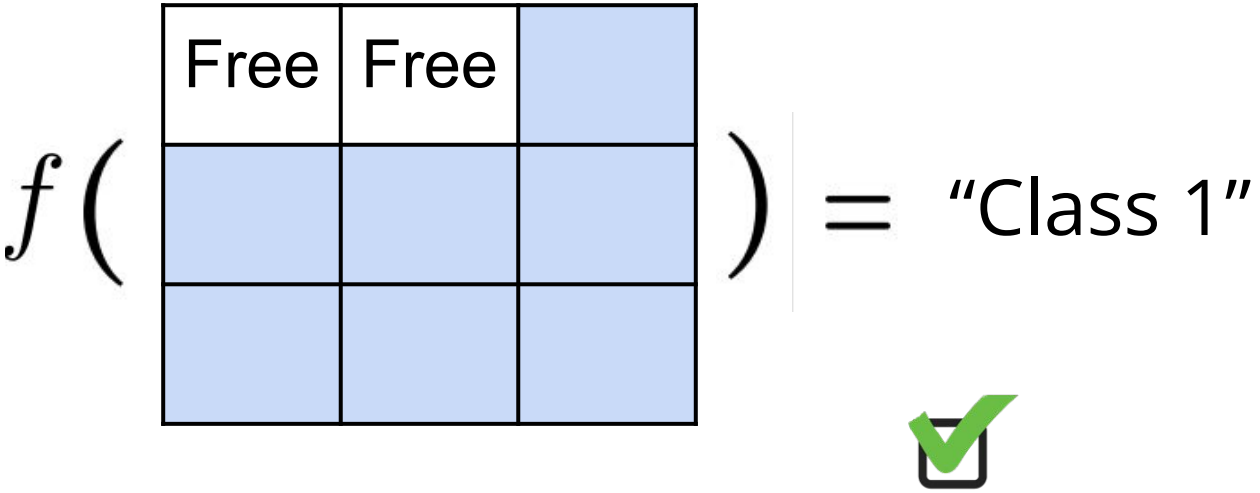
Subset minimal sufficient reasons

Attempt to ``free" features until converging to a subset minimal explanation.



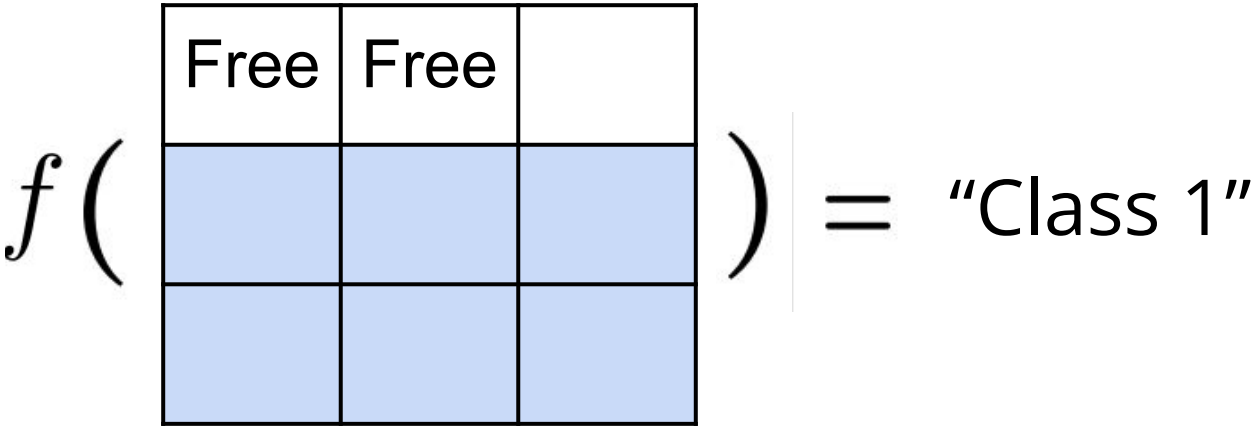
Subset minimal sufficient reasons

Attempt to ``free" features until converging to a subset minimal explanation.



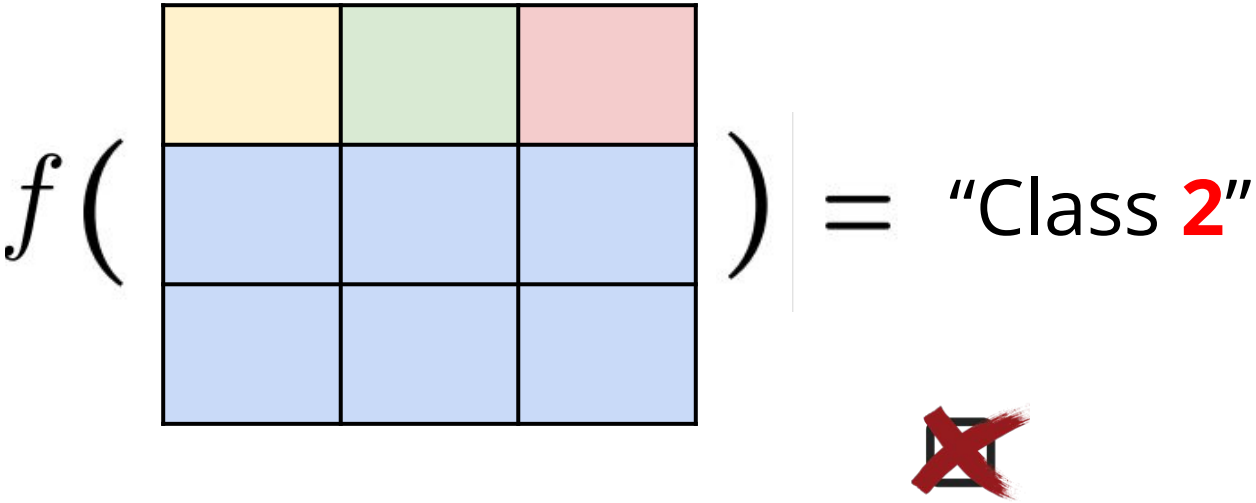
Subset minimal sufficient reasons

Attempt to ``free" features until converging to a subset minimal explanation.



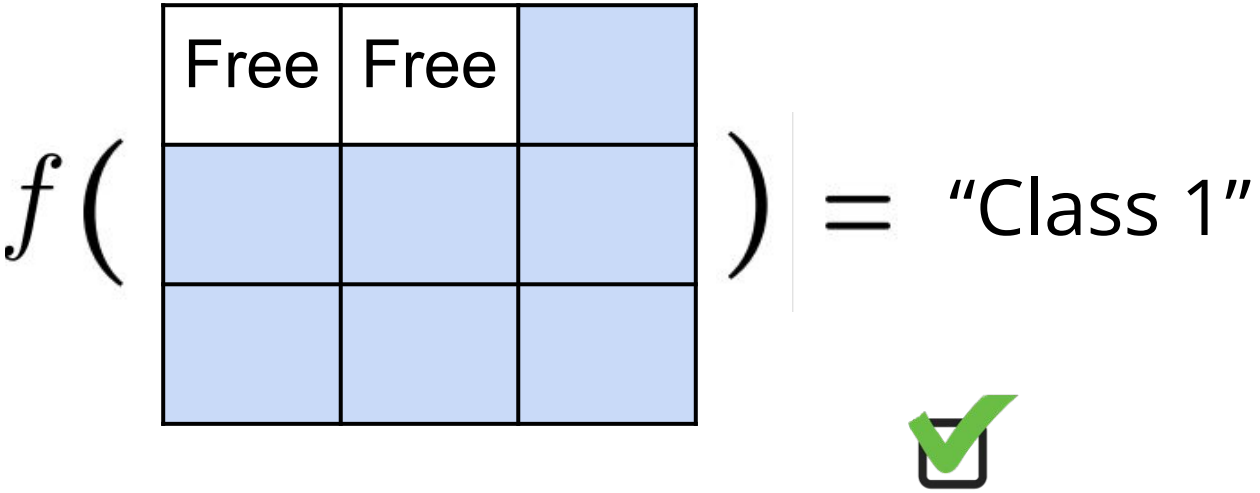
Subset minimal sufficient reasons

Attempt to ``free" features until converging to a subset minimal explanation.



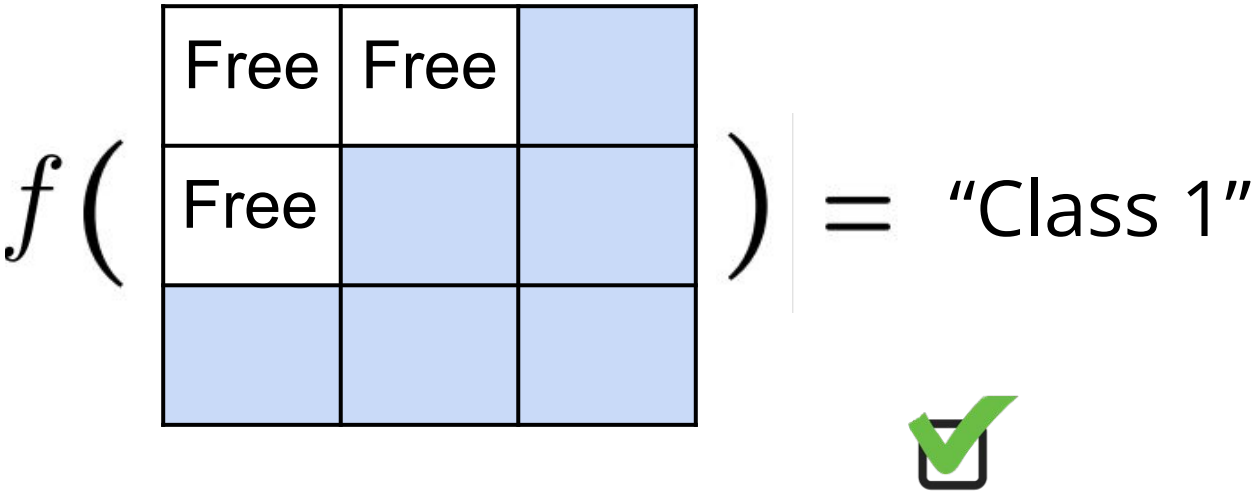
Subset minimal sufficient reasons

Attempt to ``free" features until converging to a subset minimal explanation.



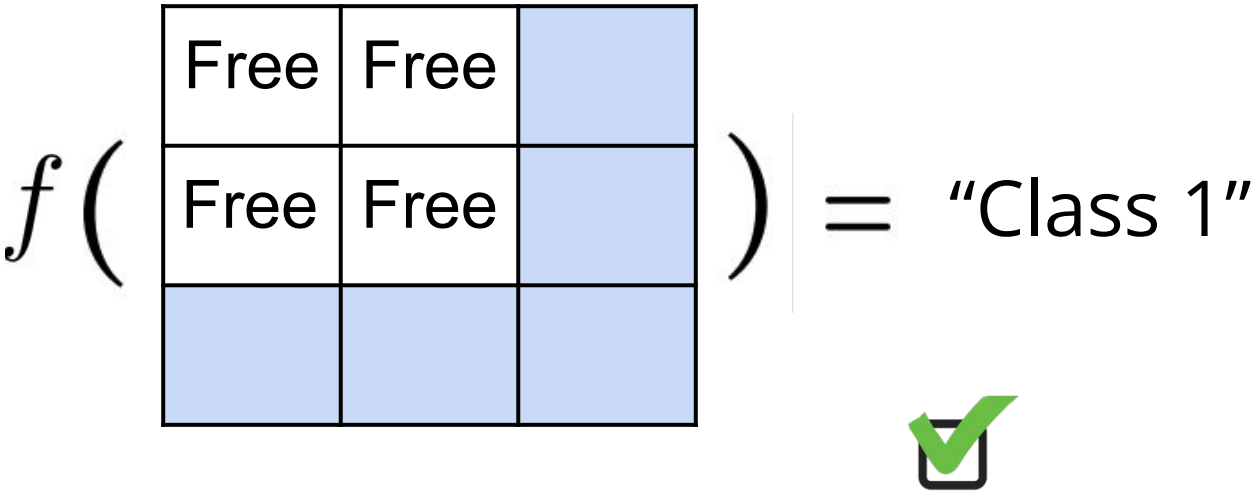
Subset minimal sufficient reasons

Attempt to ``free" features until converging to a subset minimal explanation.



Subset minimal sufficient reasons

Attempt to ``free" features until converging to a subset minimal explanation.



A key concern: **high computational complexity**

A key concern: **high computational complexity**

Each verification query is NP-Hard, and we require a linear number of queries only for a subset minimal explanation!

How can we **speed up** this process?

How can we **speed up** this process?

Explaining, Fast and Slow: Abstraction and Refinement of Provable Explanations

ICML 2025 (To appear)
Shahaf Bassan*, Yizhak Elboher*, Tobias Ladner*, Matthias Althof, Guy Katz

How can we speed up this process?

We suggest an algorithm composed of **two** main aspects:

How can we speed up this process?

We suggest an algorithm composed of **two** main aspects:

1. ***Abstraction***

How can we speed up this process?

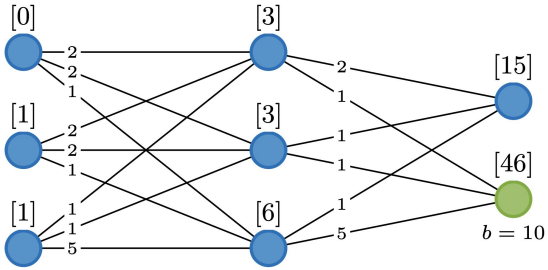
We suggest an algorithm composed of **two** main aspects:

1. ***Abstraction***

2. ***Refinement***

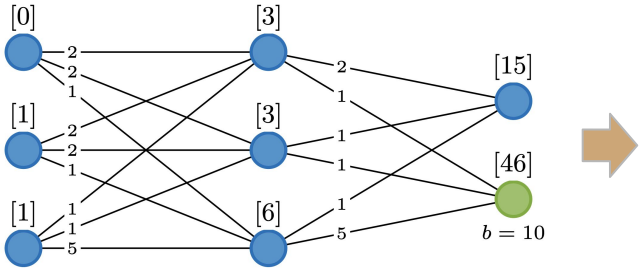
Abstraction - build a smaller neural network

Original model



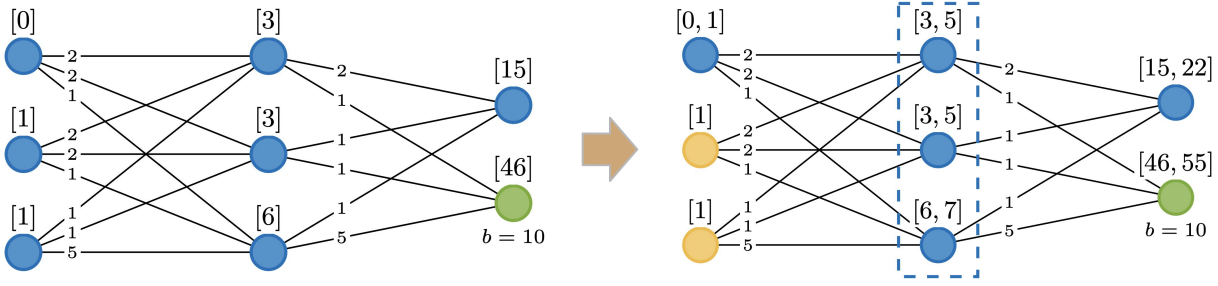
Abstraction - build a smaller neural network

Original model



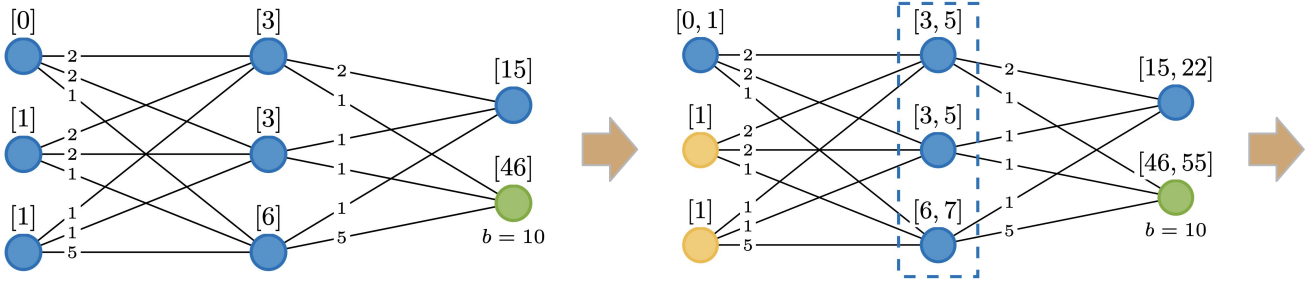
Abstraction - build a smaller neural network

Original model



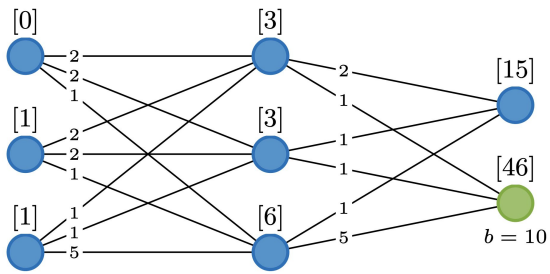
Abstraction - build a smaller neural network

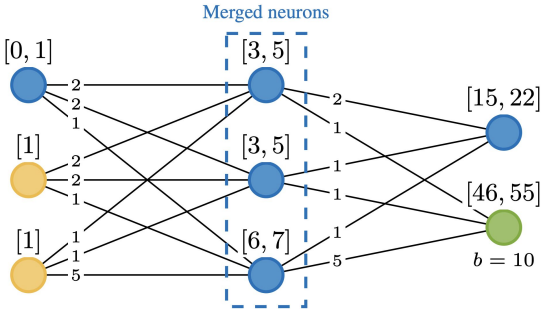
Original model



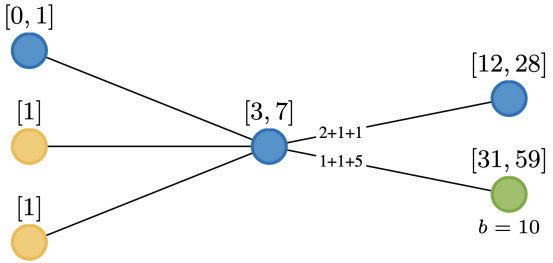
Abstraction - build a smaller neural network

Original model



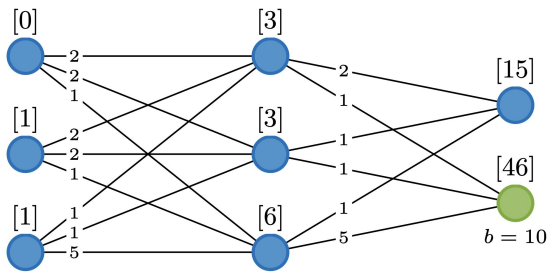


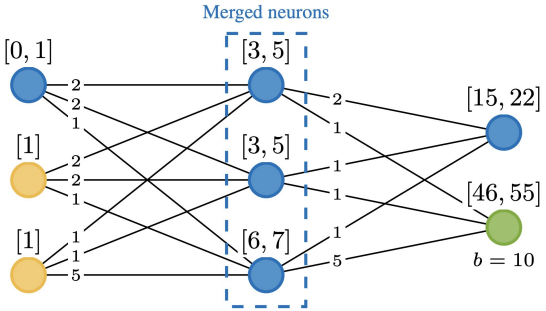
Abstract model



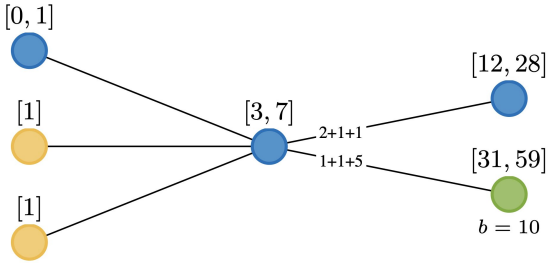
Abstraction - build a smaller neural network

Original model





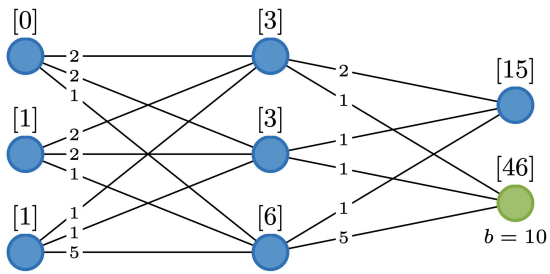
Abstract model

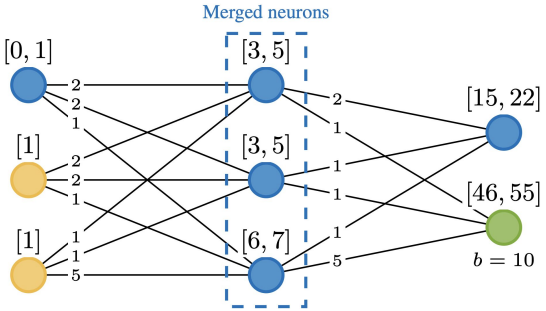


A **sufficient explanation** for the the **abstract** network is **also one** for the **original**!

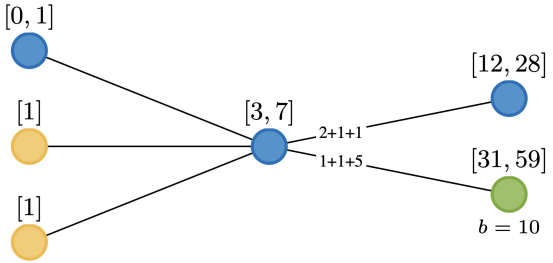
Abstraction - build a smaller neural network

Original model





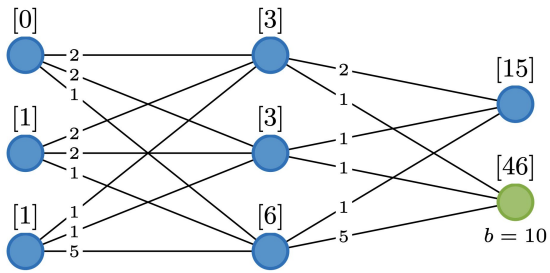
Abstract model

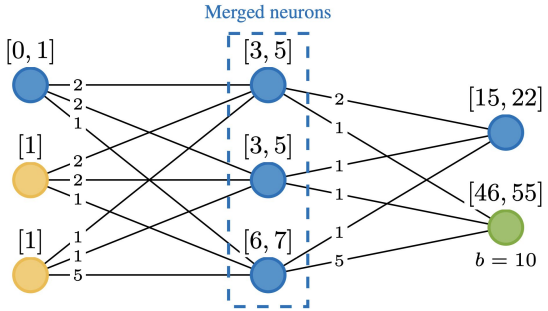


A **sufficient explanation** for the the **abstract** network is **also one** for the **original**!

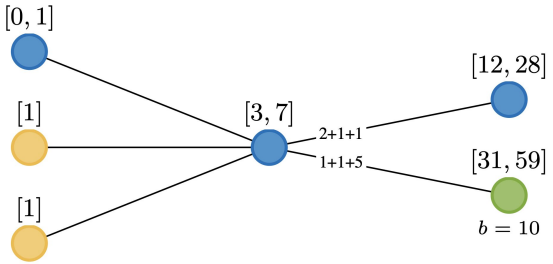
Abstraction - build a smaller neural network

Original model



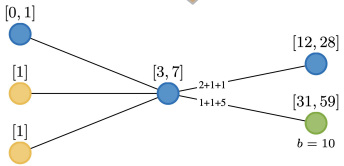


Abstract model

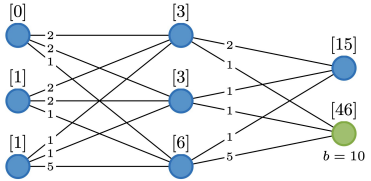


A **sufficient explanation** for the the **abstract** network is **also one** for the **original**!

We can find an explanation **over**



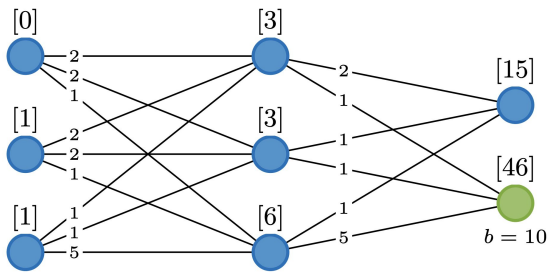
instead of

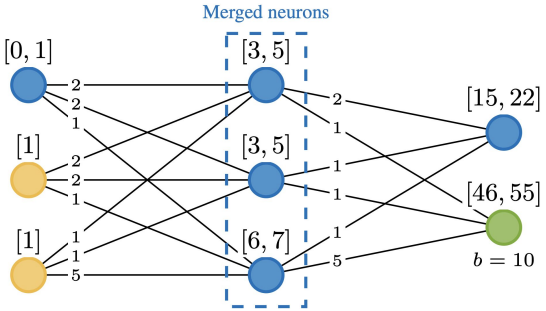


(!)

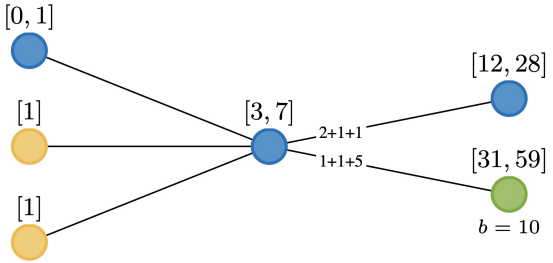
Abstraction - build a smaller neural network

Original model



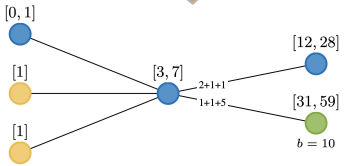


Abstract model

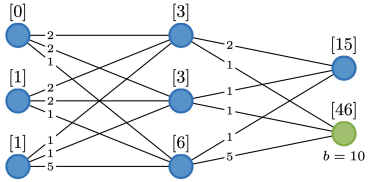


A **sufficient explanation** for the the **abstract** network is **also one** for the **original**!

We can find an explanation **over**



instead of



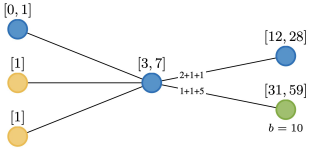
(!)

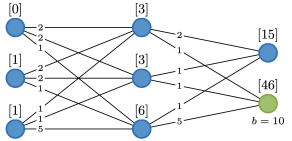
(**much more efficiently!**)

However, the **opposite** is not true!

However, the **opposite** is not true!

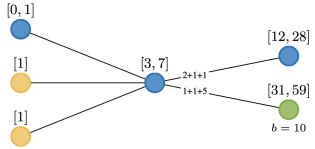
However, the **opposite** is not true!

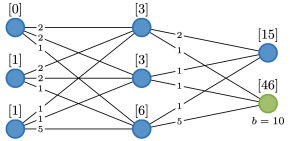
Although **sufficient** explanations on 

are also **sufficient** over 

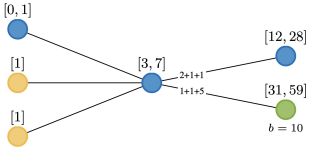
,

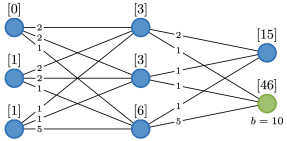
However, the **opposite** is not true!

Although **sufficient** explanations on 

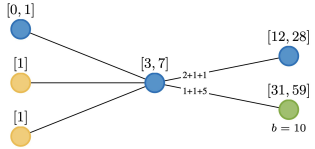
are also **sufficient** over 

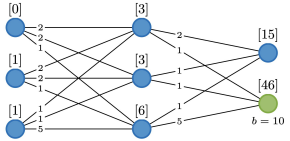
,

Minimal sufficient explanations for 

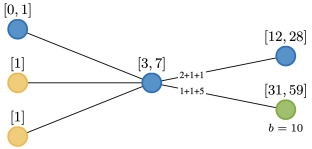
are **not** necessarily **minimal** for 

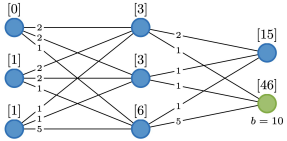
However, the **opposite** is not true!

Although **sufficient** explanations on 

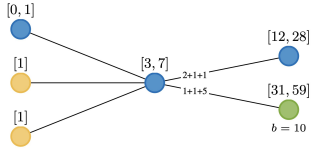
are also **sufficient** over 

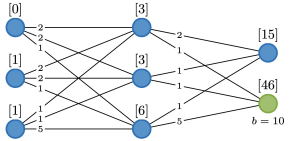
,

Minimal sufficient explanations for 

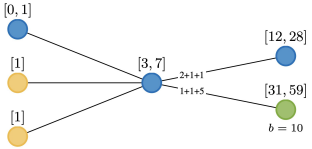
are **not** necessarily **minimal** for 

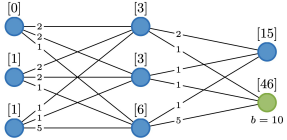
However, the **opposite** is not true!

Although **sufficient** explanations on 

are also **sufficient** over 

,

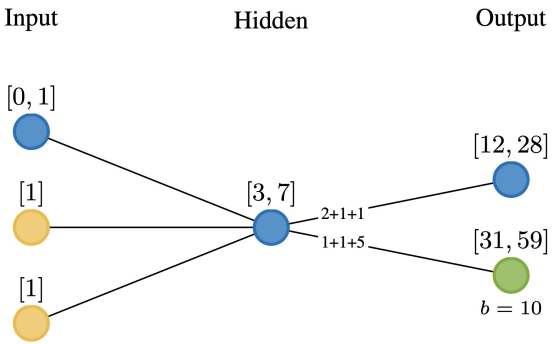
Minimal sufficient explanations for 

are **not** necessarily **minimal** for 

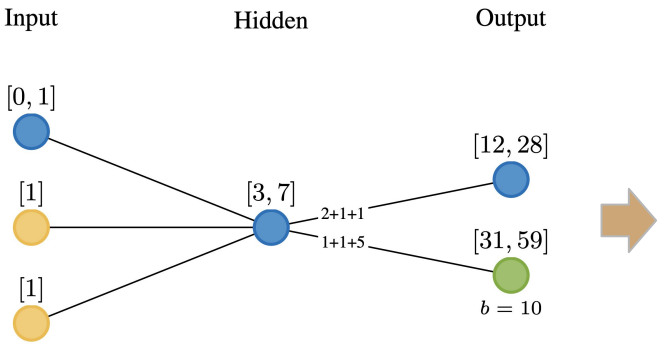
This requires an additional procedure: **refinement!**

Refinement - Gradually increase the network size

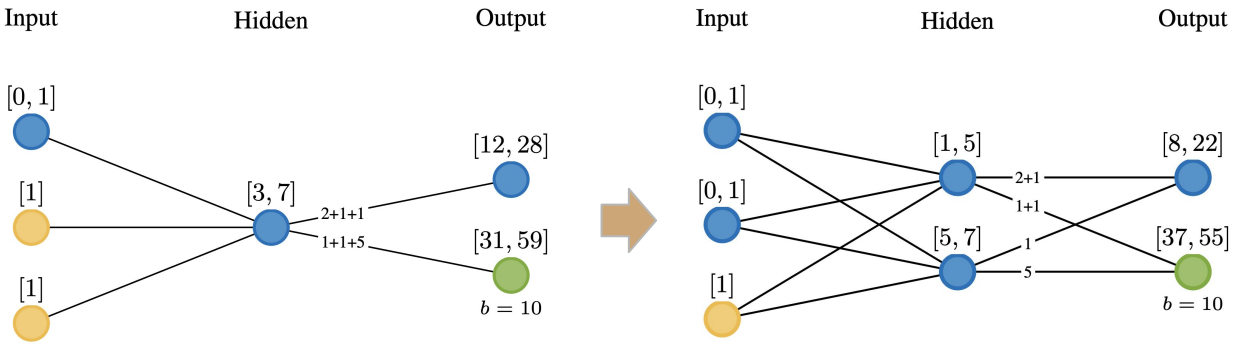
Refinement - Gradually increase the network size



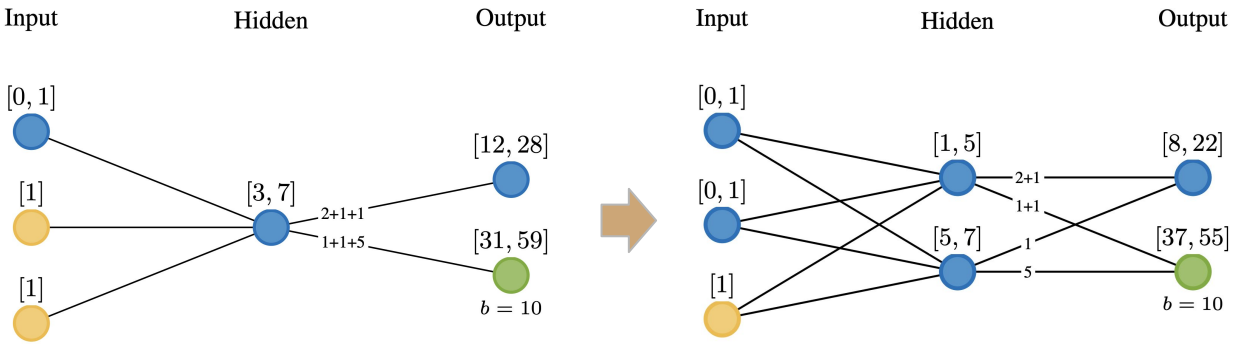
Refinement - Gradually increase the network size



Refinement - Gradually increase the network size

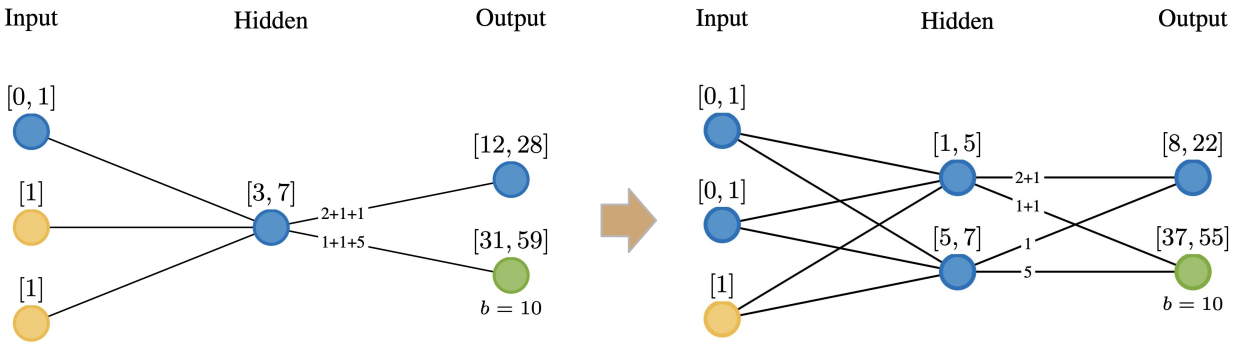


Refinement - Gradually increase the network size



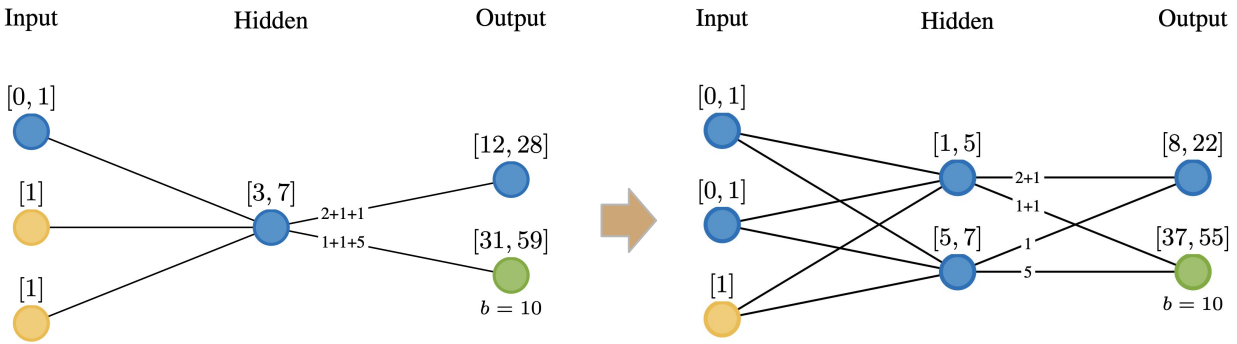
Gradually **refine** the abstract network (increase its size) and find an explanation

Refinement - Gradually increase the network size



Gradually **refine** the abstract network (increase its size) and find an explanation

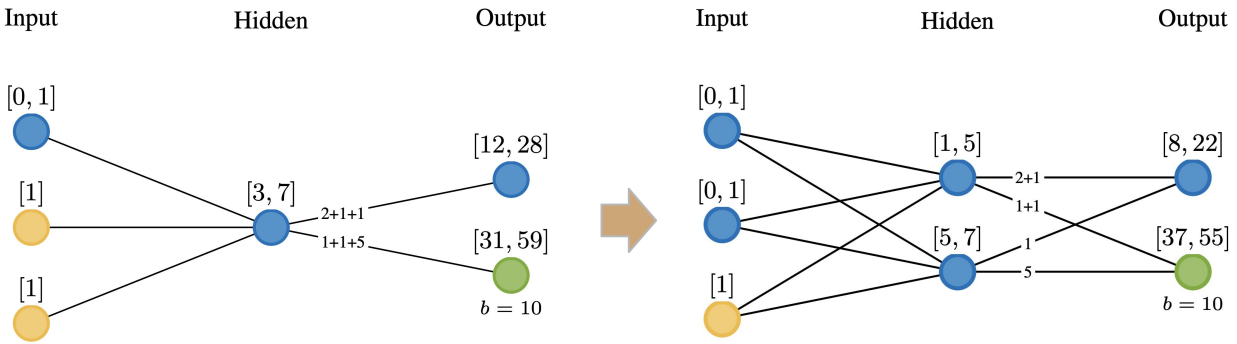
Refinement - Gradually increase the network size



Gradually **refine** the abstract network (increase its size) and find an explanation

We **“free” features** in the **abstract** model, then **refine** and **“free”** more features, and so on...

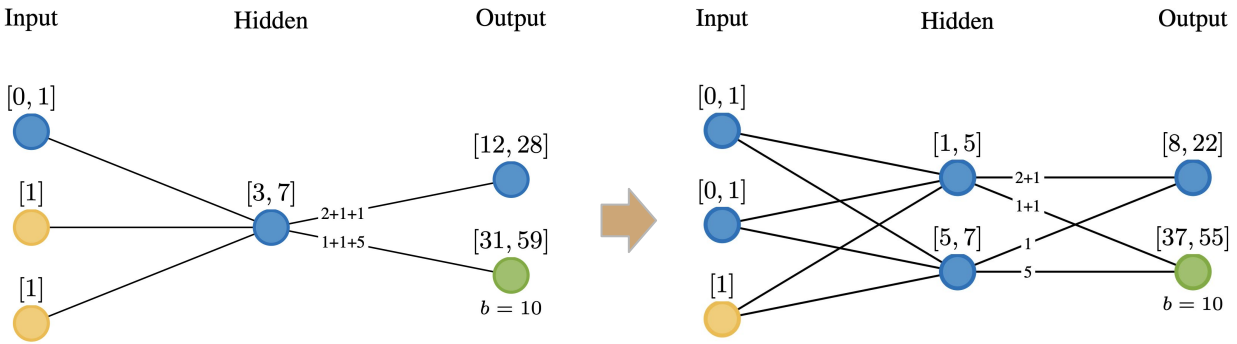
Refinement - Gradually increase the network size



Gradually **refine** the abstract network (increase its size) and find an explanation

We **“free” features** in the **abstract** model, then **refine** and **“free”** more features, and so on...

Refinement - Gradually increase the network size

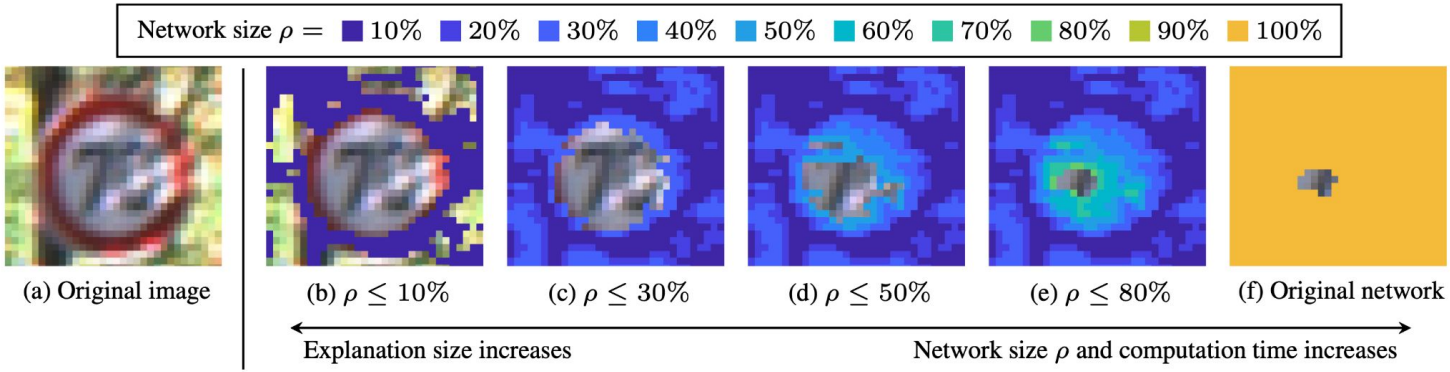


Gradually **refine** the abstract network (increase its size) and find an explanation

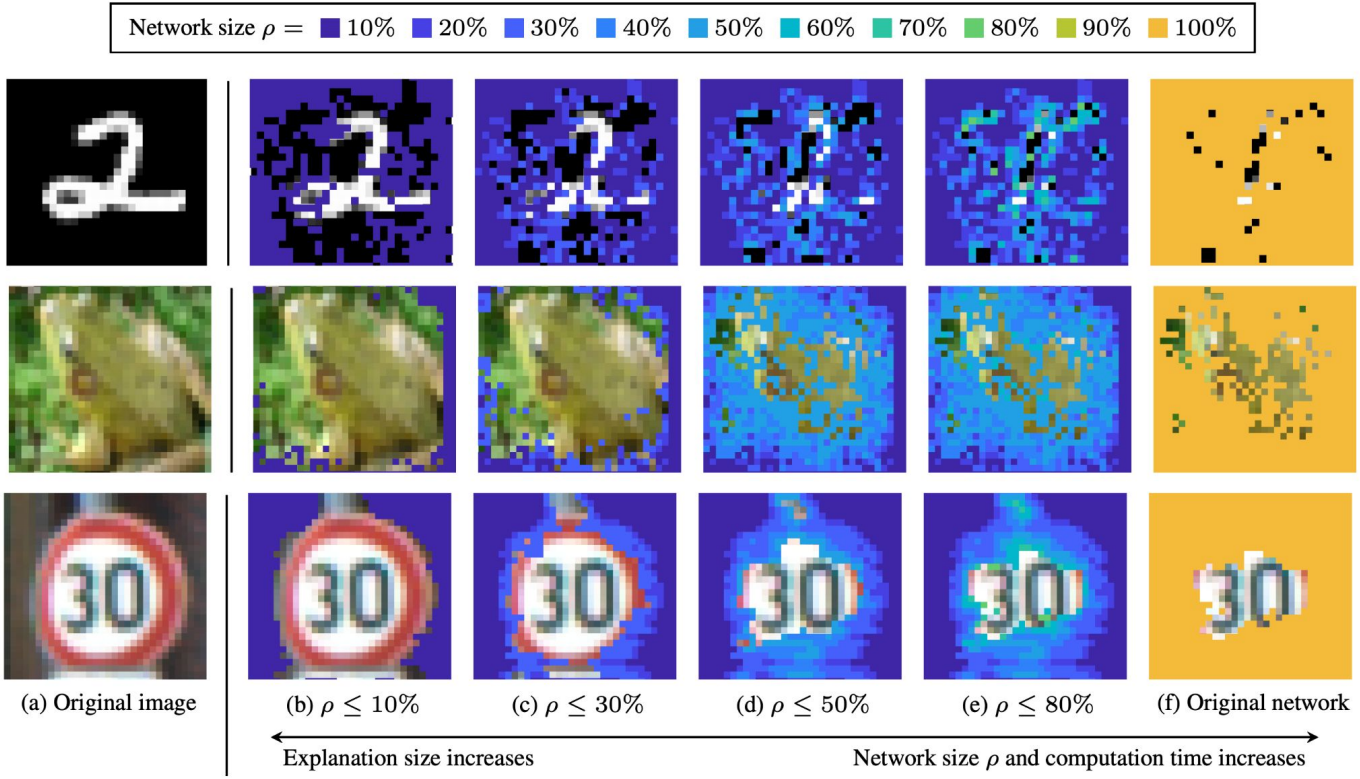
We **“free” features** in the **abstract** model, then **refine** and **“free”** more features, and so on...

Eventually, converges to a **minimal explanation.**

An example


Network size $\rho =$ 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
(a) Original image
(b) $\rho \leq 10\%$
(c) $\rho \leq 30\%$
(d) $\rho \leq 50\%$
(e) $\rho \leq 80\%$
(f) Original network
Explanation size increases
Network size ρ and computation time increases

More examples



If we **relax** sufficiency, can we **scale even more**?

If we **relax** sufficiency, can we **scale even more**?

Explain Yourself, Briefly! Self-Explaining Neural Networks with Concise Sufficient Reasons

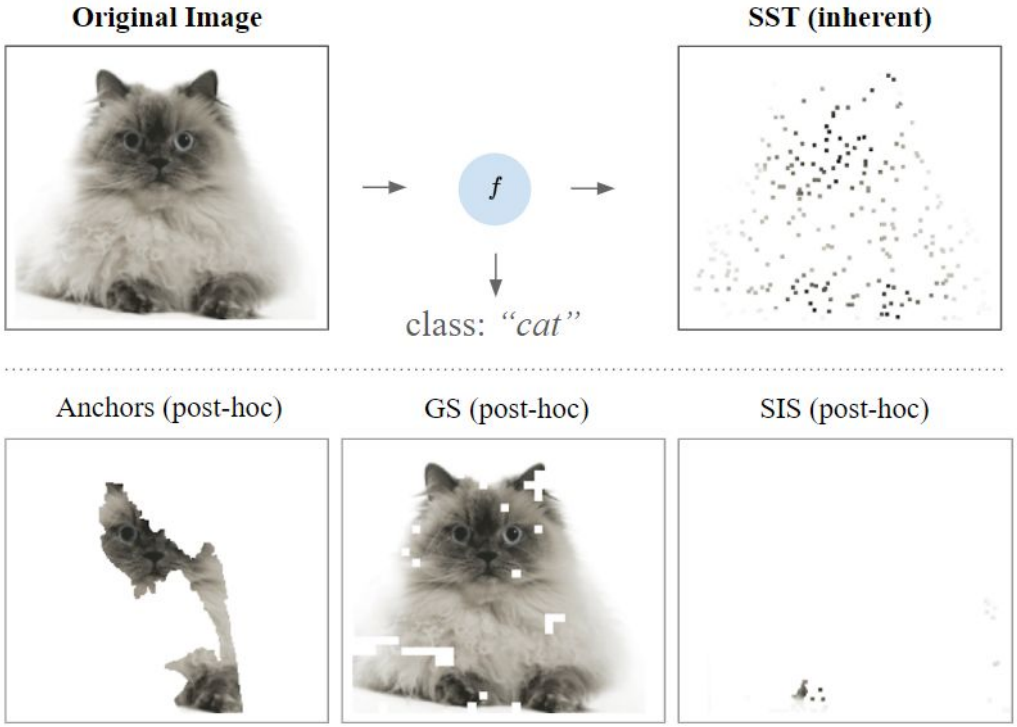
ICLR 2025
Shahaf Bassan, Ron Eliav, Shlomit Gur

Up until now, we discussed **post-hoc** explanations, that are obtained after training.

Can training time-intervention help with our scalability challenges?

A model that **explains itself**

A model that **explains itself**

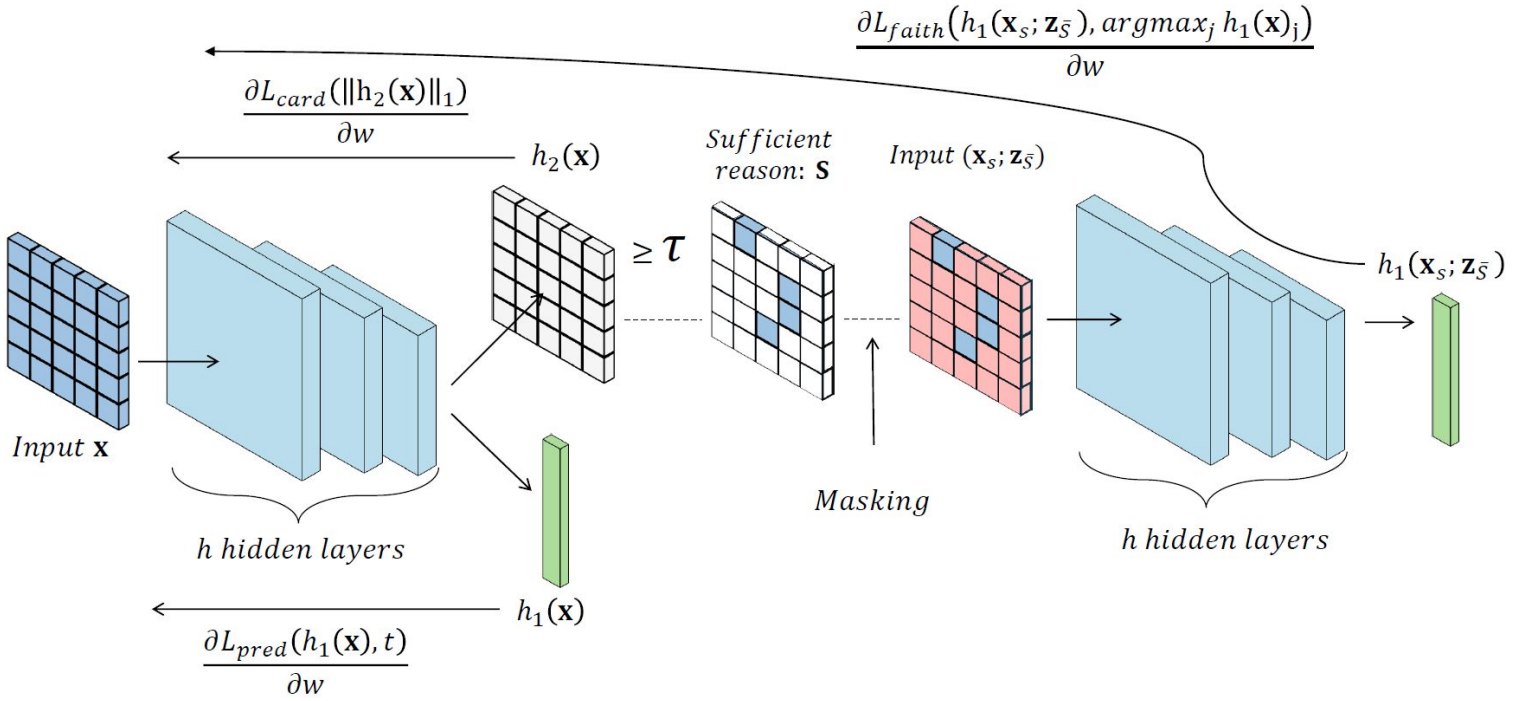


How do we train a model to give us an explanation that is both **sufficient** and **small**?

How do we train a model to give us an explanation that is both **sufficient** and **small**?

With **Sufficient Subset Training (SST)**

The dual propagation in Sufficient Subset Training



The optimization objective in Sufficient Subset Training:

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

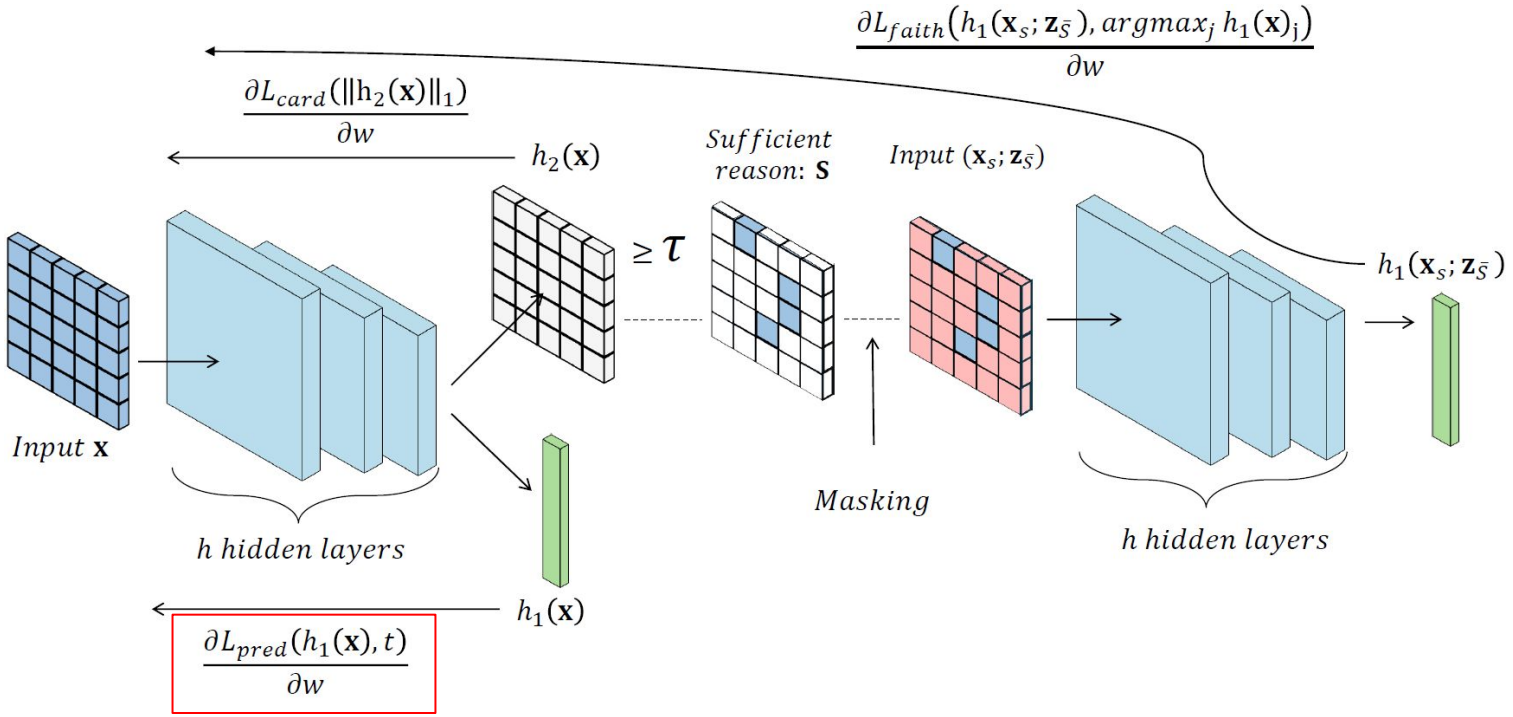
The optimization objective in Sufficient Subset Training:

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

(Standard) Prediction loss:
Optimize for accuracy

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

The dual propagation in Sufficient Subset Training



The optimization objective in Sufficient Subset Training:

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

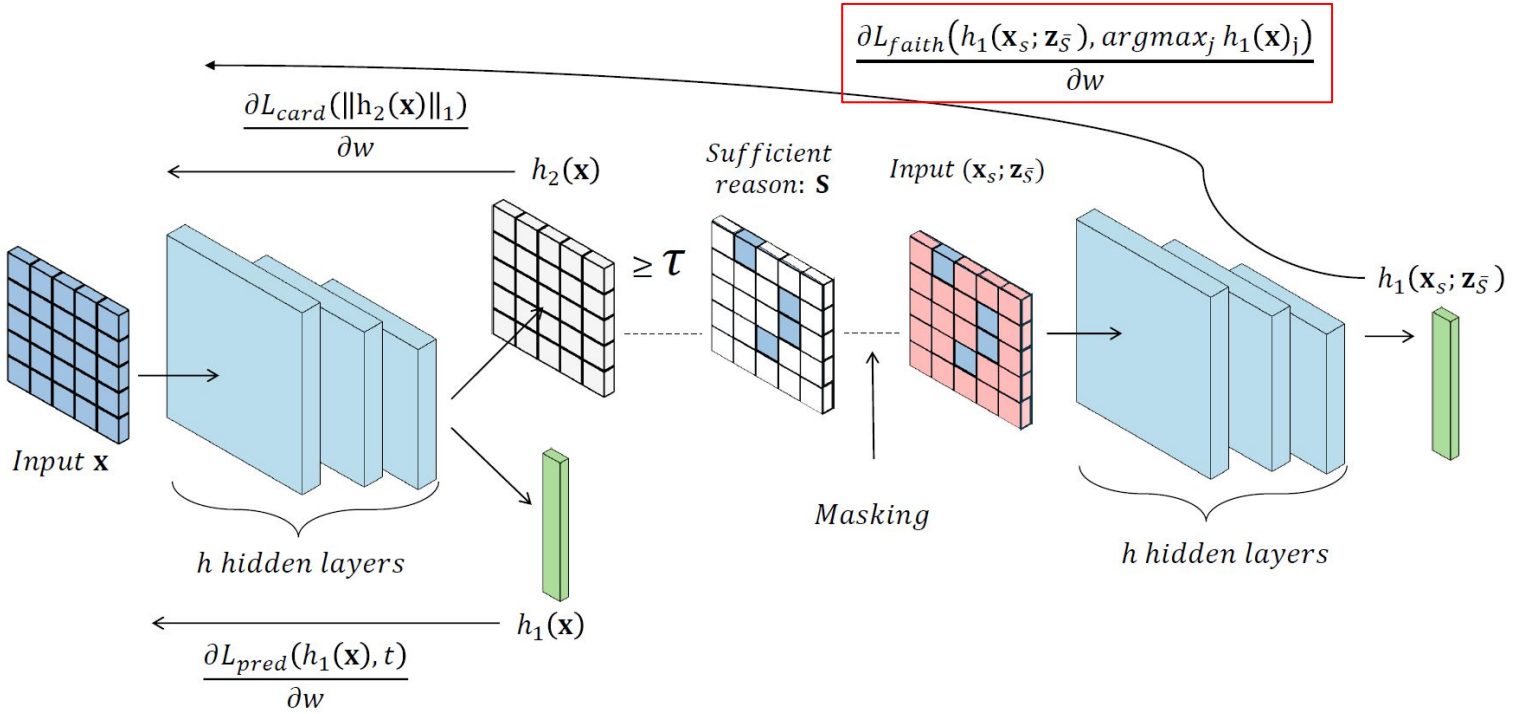
The optimization objective in Sufficient Subset Training:

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

Faithfulness loss:
Optimize for sufficiency

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

The dual propagation in Sufficient Subset Training



The optimization objective in Sufficient Subset Training:

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

The optimization objective in Sufficient Subset Training:

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

Cardinality loss:
Optimize for minimal cardinality

$$||h_2(\mathbf{x})||_1$$

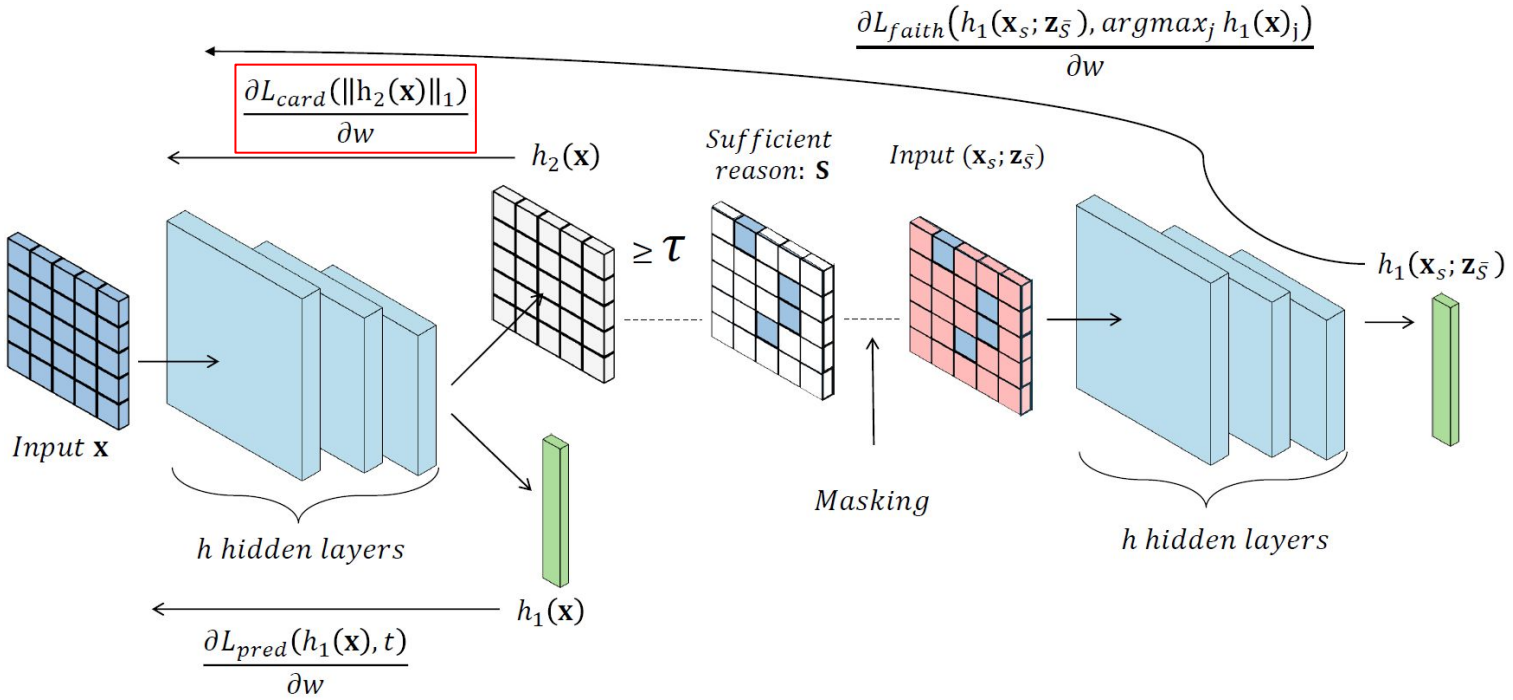
The optimization objective in Sufficient Subset Training:

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

Cardinality loss:
Optimize for minimal cardinality

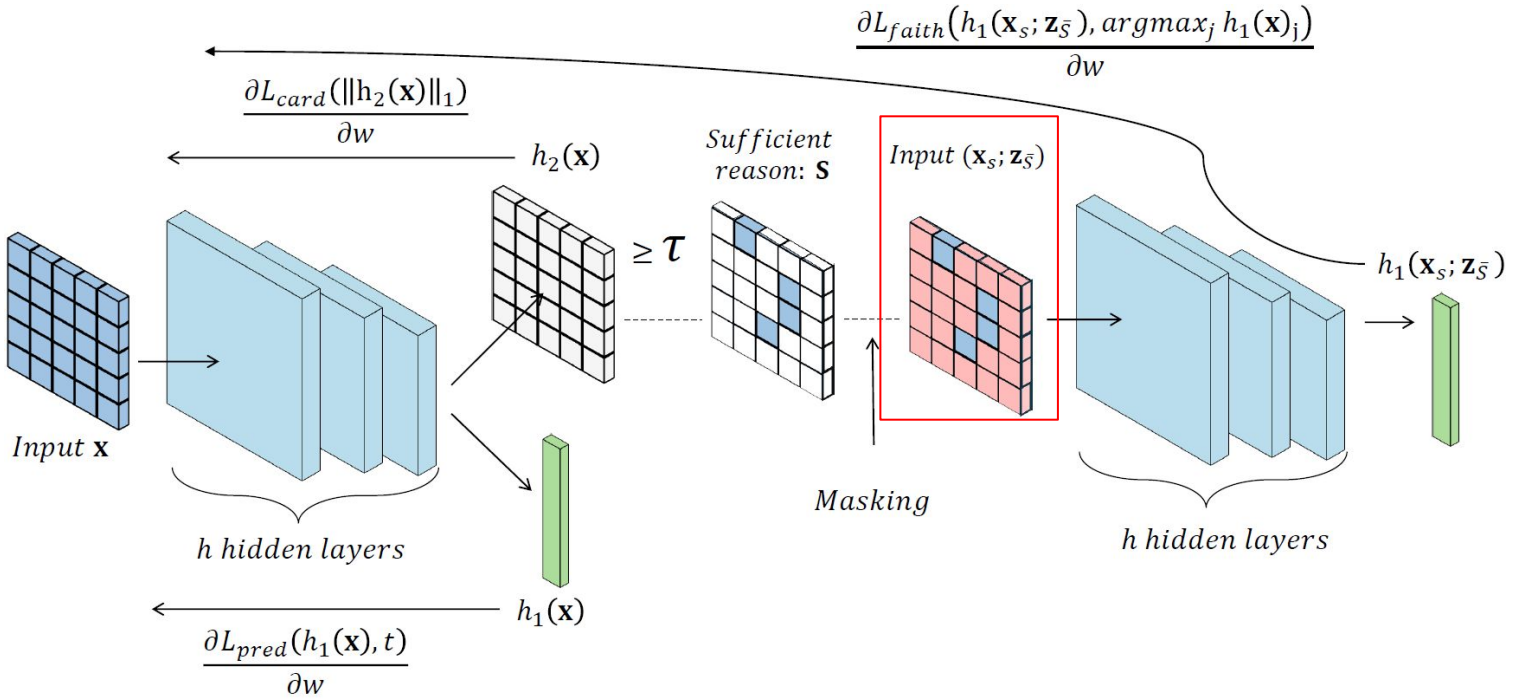
$$||h_2(\mathbf{x})||_1$$

How do we perform the masking?



What about the masking?

How do we perform the masking?



We discuss 3 maskings for the 3 forms of sufficiency:

We discuss 3 maskings for the 3 forms of sufficiency:

1. **Baseline masking** - fix some baseline to the complementary.

We discuss 3 maskings for the 3 forms of sufficiency:

1. **Baseline masking** - fix some baseline to the complementary.

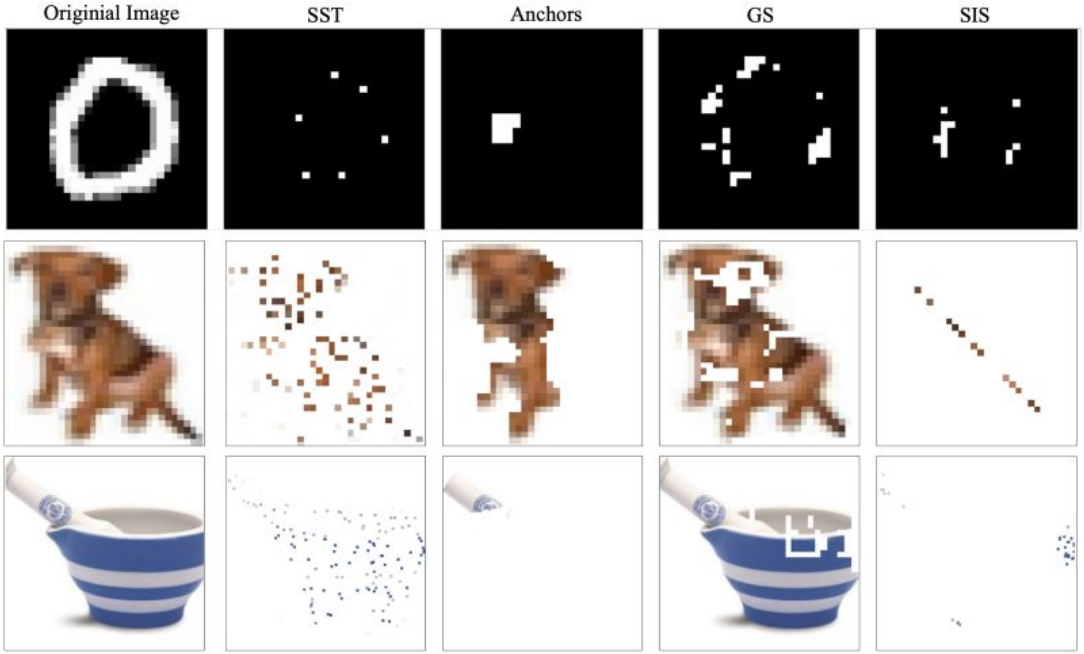
2. **Probabilistic masking** - sample values to the complementary from some distribution.

We discuss 3 maskings for the 3 forms of sufficiency:

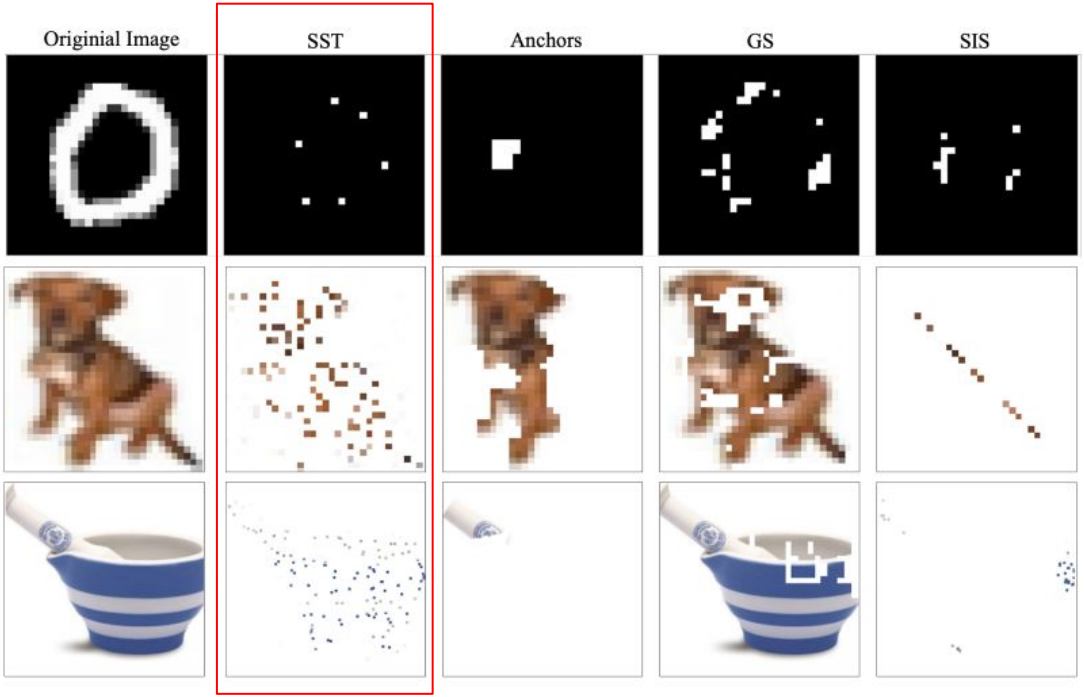
1. **Baseline masking** - fix some baseline to the complementary.
2. **Probabilistic masking** - sample values to the complementary from some distribution.
3. **Robust masking** - perform a gradient attack over the complementary features.

We ran experiments on both image domains (IMAGENET, CIFAR10, MNIST) and language domains (IMDB, SNLI)

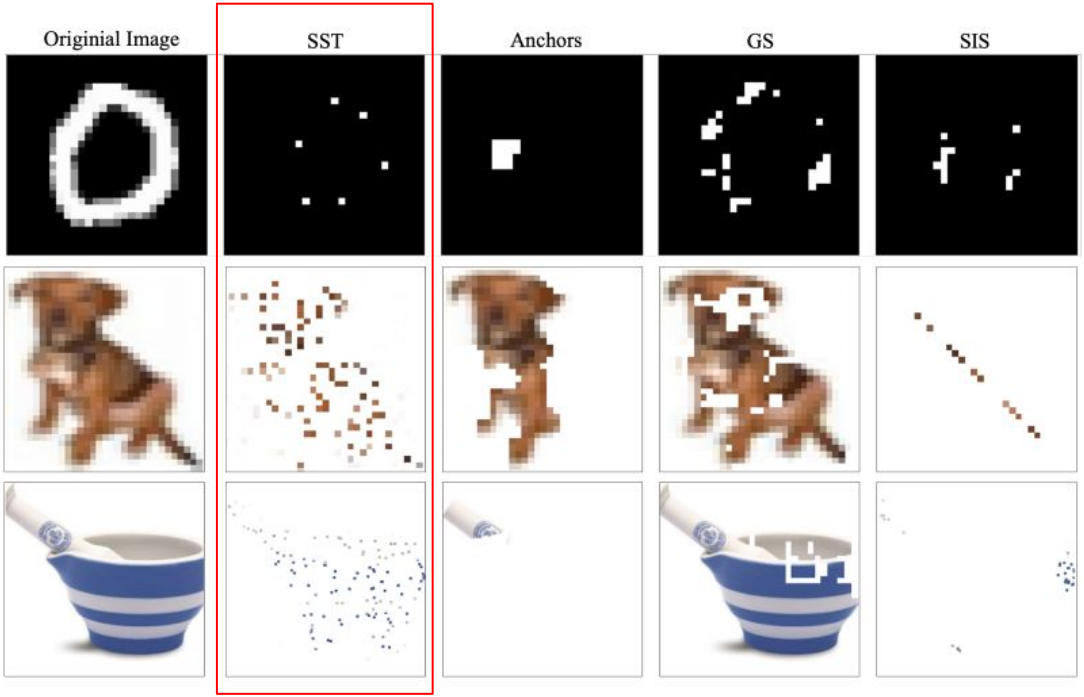
We ran comparisons to post-hoc methods



We ran comparisons to post-hoc methods



We ran comparisons to post-hoc methods, and ablations



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

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

Summary of results - vs. post-hoc heuristic methods

Summary of results - vs. post-hoc heuristic methods

1. Improved faithfulness (more subsets verified as sufficient)

Summary of results - vs. post-hoc heuristic methods

1. Improved faithfulness (more subsets verified as sufficient)
2. Improved conciseness (smaller subsets)

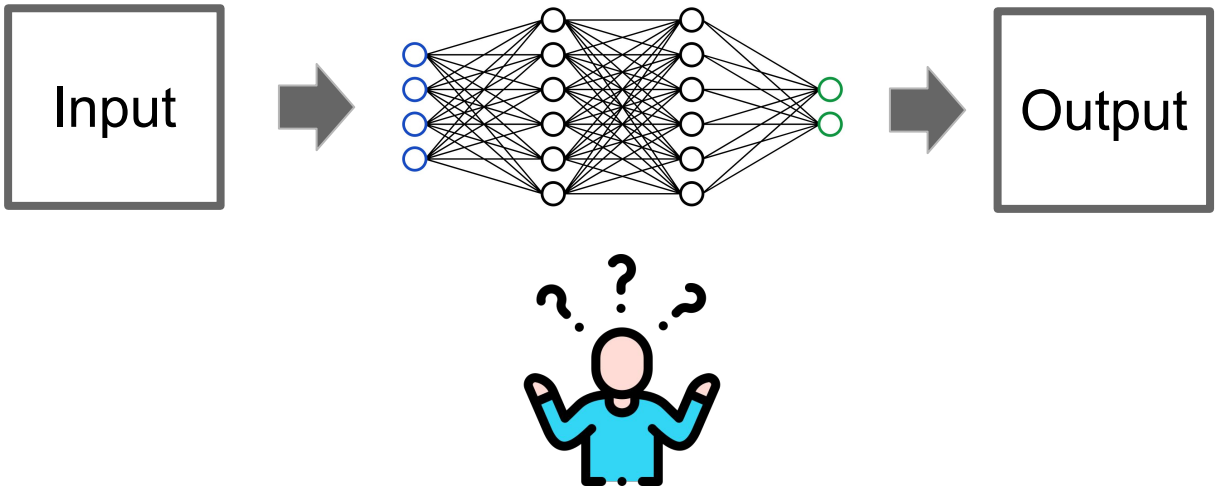
Summary of results - vs. post-hoc heuristic methods

1. Improved faithfulness (more subsets verified as sufficient)
2. Improved conciseness (smaller subsets)
3. More efficient (inherent sufficient reasons)

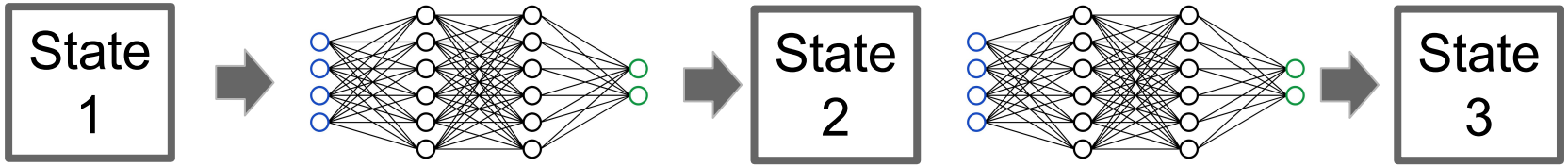
Summary of results - vs. post-hoc heuristic methods

1. Improved faithfulness (more subsets verified as sufficient)
2. Improved conciseness (smaller subsets)
3. More efficient (inherent sufficient reasons)
4. Comparable predictive performance

Up until now - classification



Reactive Systems are more complex

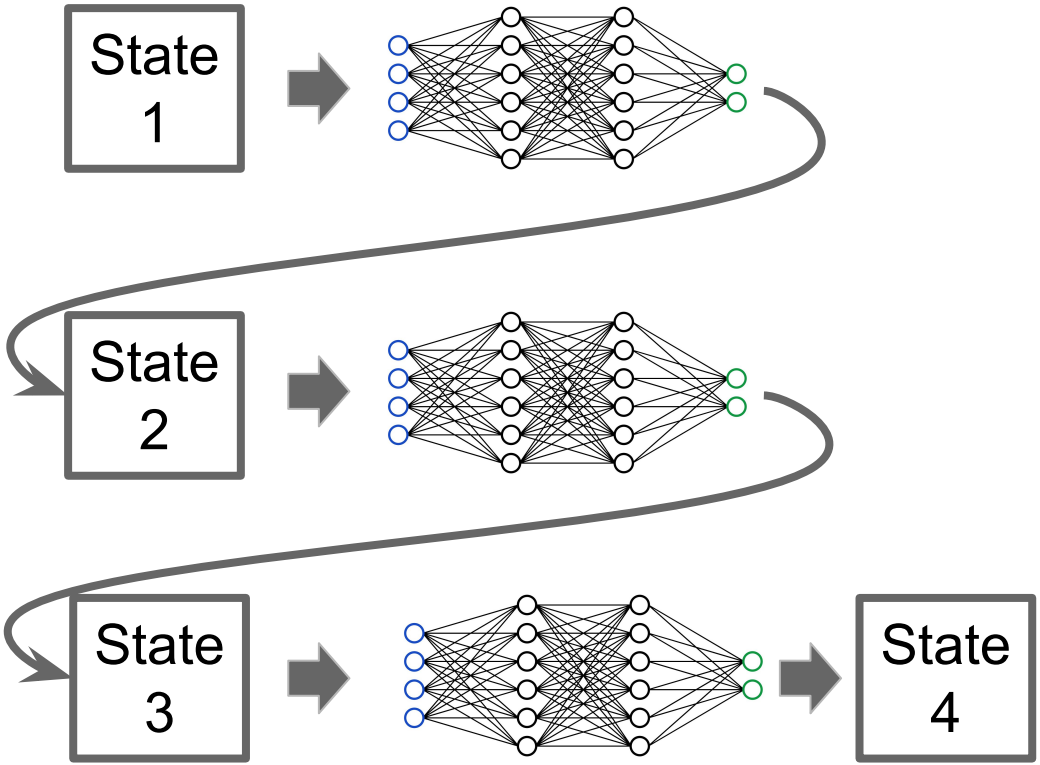


Formal Explanations in **Reactive Systems**

Formally Explaining Neural Networks within Reactive Systems

FMCAD, 2023
Shahaf Bassan*, Guy Amir*, Davide Corsi, Idan Refaeli, Guy Katz
Best paper runnerup

A naive solution: Encode everything together



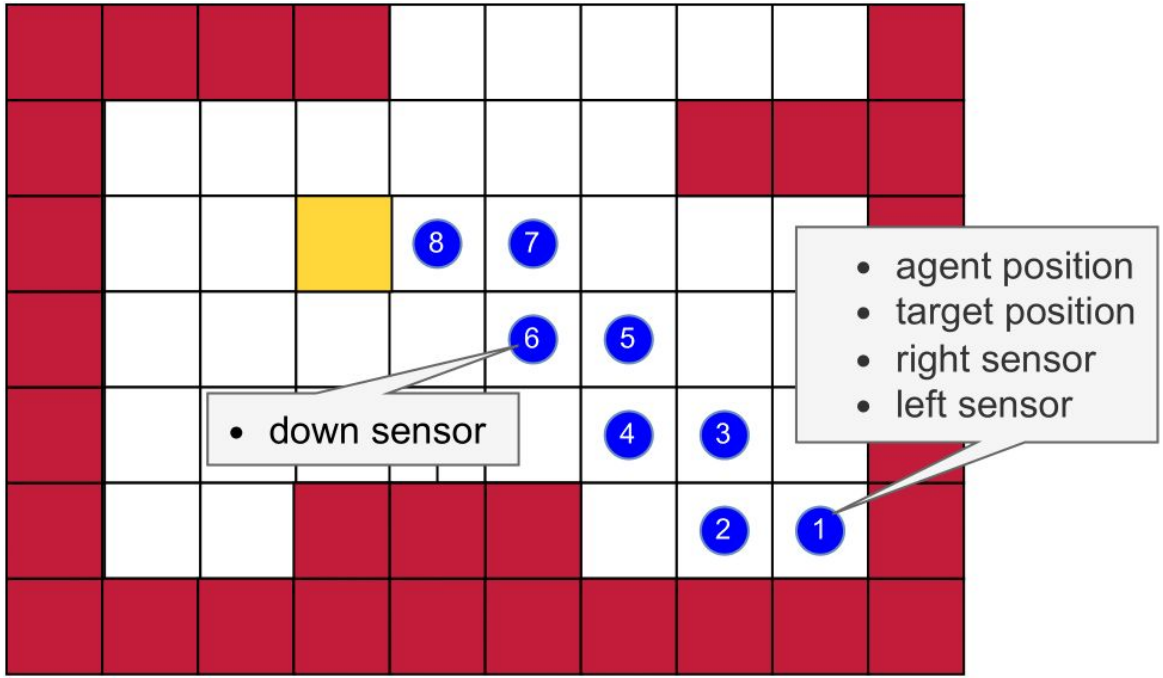
The Problem with the naive solution

Growth of neural network: **exponential blow up** in computation time.

Our solution

We suggest algorithms that avoid the exponential blow-up, but are still optimal.

Formal Explanations in **Reactive Systems**



Let’s move to some more theoretical stuff

The computational complexity of finding explanations

The computational complexity of finding explanations

1. Local vs. Global Interpretability: A Computational Complexity Perspective
   ICML 2024 (Spotlight)
   Shahaf Bassan, Guy Amir, Guy Katz

2. What makes an Ensemble (Un) Interpretable?
   ICML 2025 (To appear)
   Shahaf Bassan, Guy Amir, Meirav Zehavi, Guy Katz

3. On the Computational Tractability of the (Many) Shapley Values
   AI'STATS 2025
   Reda Marzouk*, Shahaf Bassan*, Guy Katz, Colin De La Higuera

4. Hard to Explain: On the Computational Hardness of In-Distribution Model Interpretation
   ECAI 2024
   Guy Amir*, Shahaf Bassan*, Guy Katz

Our goal is to identify **which** explanation types with various guarantees can be found **efficiently** (and which cannot).

Our goal is to identify **which** explanation types with various guarantees can be found **efficiently** (and which cannot).

Factors that influence the complexity:

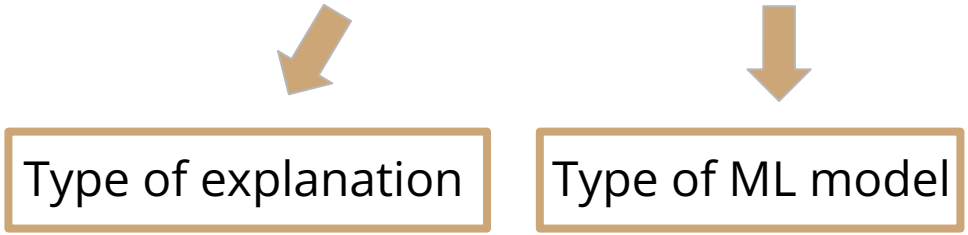
Our goal is to identify **which** explanation types with various guarantees can be found **efficiently** (and which cannot).

Factors that influence the complexity:

Type of explanation

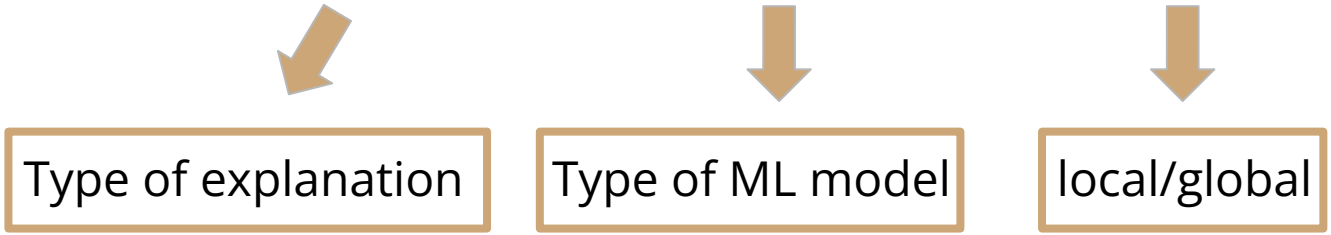
Our goal is to identify **which** explanation types with various guarantees can be found **efficiently** (and which cannot).

Factors that influence the complexity:



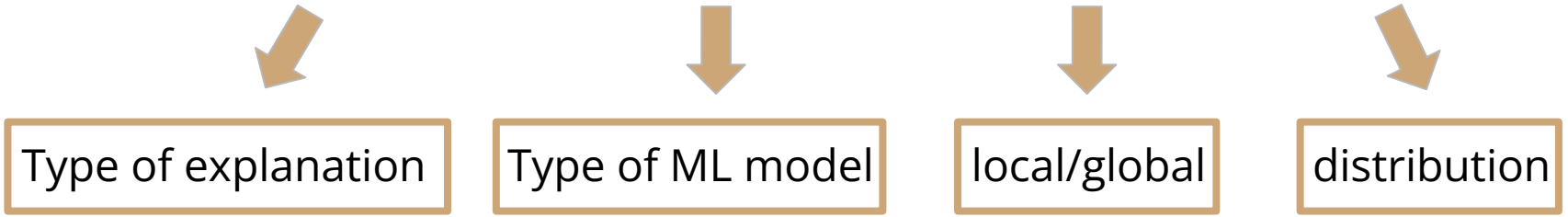
Our goal is to identify **which** explanation types with various guarantees can be found **efficiently** (and which cannot).

Factors that influence the complexity:



Our goal is to identify **which** explanation types with various guarantees can be found **efficiently** (and which cannot).

Factors that influence the complexity:



Conclusion

1. Formal XAI lets us compute explanations that are provably correct.

Conclusion

1. Formal XAI lets us compute explanations that are provably correct.
2. We can do this using NN verification

Conclusion

1. Formal XAI lets us compute explanations that are provably correct.
2. We can do this using NN verification
3. Abstraction-refinement can improve efficiency

Conclusion

1. Formal XAI lets us compute explanations that are provably correct.
2. We can do this using NN verification
3. Abstraction-refinement can improve efficiency
4. Self-explaining neural networks can scale even more

Conclusion

1. Formal XAI lets us compute explanations that are provably correct.
2. We can do this using NN verification
3. Abstraction-refinement can improve efficiency
4. Self-explaining neural networks can scale even more
5. Understanding the complexity of finding explanations is an important theoretical aspect of this area.

Future Work - **Formal XAI**

1. Verifying alternative forms of explanations

Future Work - **Formal XAI**

1. Verifying alternative forms of explanations
2. Methods for improving scalability

Future Work - **Formal XAI**

1. Verifying alternative forms of explanations
2. Methods for improving scalability

Future Work - **Formal XAI**

1. Verifying alternative forms of explanations
2. Methods for improving scalability
3. Additional work on computational complexity

Future Work - **Formal XAI**

1. Verifying alternative forms of explanations
2. Methods for improving scalability
3. Additional work on computational complexity
4. Formal certification during training

THANK YOU
for your
ATTENTION!