Faster Verified Explanations for Neural Networks

Alessandro De Palma, Greta Dolcetti and Caterina Urban

40th European Conference on Object-Oriented Programming (ECOOP 2026)

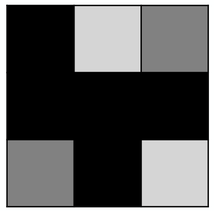
Neural Networks in High-Stakes Systems

Safety-Critical Applications

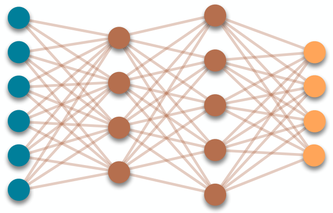
Socio-Economic Applications

Neural Network Verification

LOCAL ROBUSTNESS



$\mathbf{x} \in \mathbb{R}^d$



f

4

$y_f(\mathbf{x})$

$(\mathbf{x}, \mathscr{A}, \epsilon, \mathbf{f})$

v

1 $\forall x' \in \mathscr{B}^{\epsilon}_{\mathscr{A}} : y_f(x') = y_f(x)$

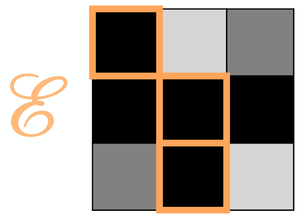
0 🤷‍♀️

-1 $\exists x' \in \mathscr{B}^{\epsilon}_{\mathscr{A}} : y_f(x') \neq y_f(x)$

$$B^{\epsilon}_{\mathscr{A}}(\mathbf{x}) \stackrel{\text{def}}{=} \left\{ \mathbf{x}' \in \mathbb{R}^d \;\middle|\; \mathbf{x}'_{\overline{\mathscr{A}}} = \mathbf{x}_{\overline{\mathscr{A}}} \wedge \left\| \mathbf{x}'_{\mathscr{A}} - \mathbf{x}_{\mathscr{A}} \right\|_{\infty} \leq \epsilon \right\}$$

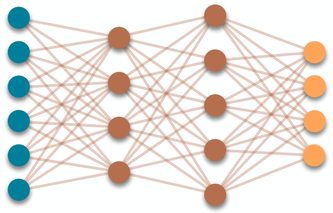
Optimal Robust Explanations

ABDUCTIVE EXPLANATIONS (AXps)



$\mathscr{E}$

$\mathbf{x} \in \mathbb{R}^d$



f

4

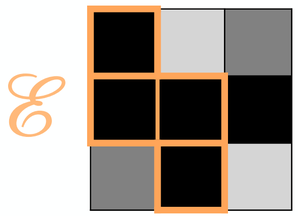
$\forall i \in \mathscr{E}$ ·

$\mathbf{v}$

Useful Semantic Ideal, but Unrealistic in Practice

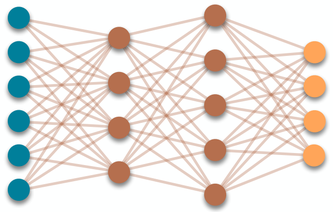
~~Optimal~~ Robust Explanations

WEAK ABDUCTIVE EXPLANATIONS



$\mathscr{E}$

$\mathbf{x} \in \mathbb{R}^d$



f

4

$\mathbf{y_f(x)}$

$(\mathbf{x}, \overline{\mathscr{E}}, \epsilon, \mathbf{f})$

$\forall i \in \mathscr{E}: \ (\mathbf{x}, \overline{\mathscr{E}} \cup \{i\}, \epsilon, \mathbf{f})$

v

1

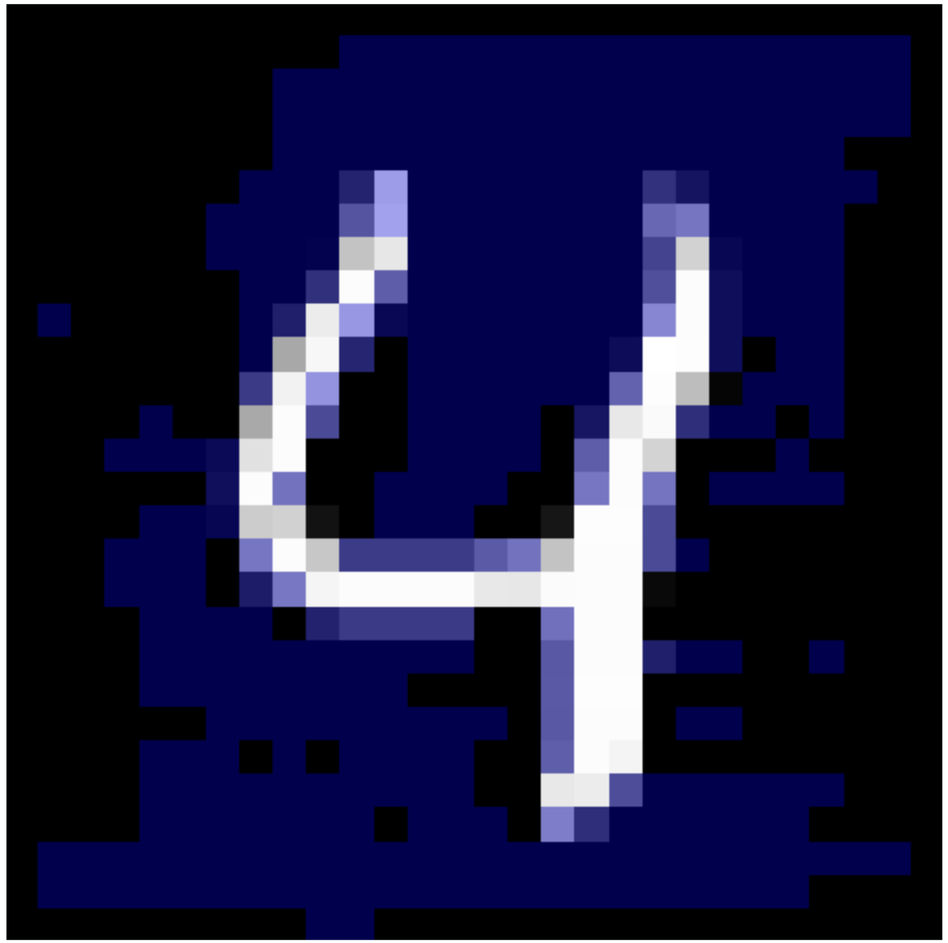
i is an **UNKNOWN**

-1 or **0**

i is a **COUNTERFACTUAL**

~~Optimal~~ Robust Explanations

WEAK ABDUCTIVE EXPLANATIONS



Computing ~~Optimal~~ Robust Explanations

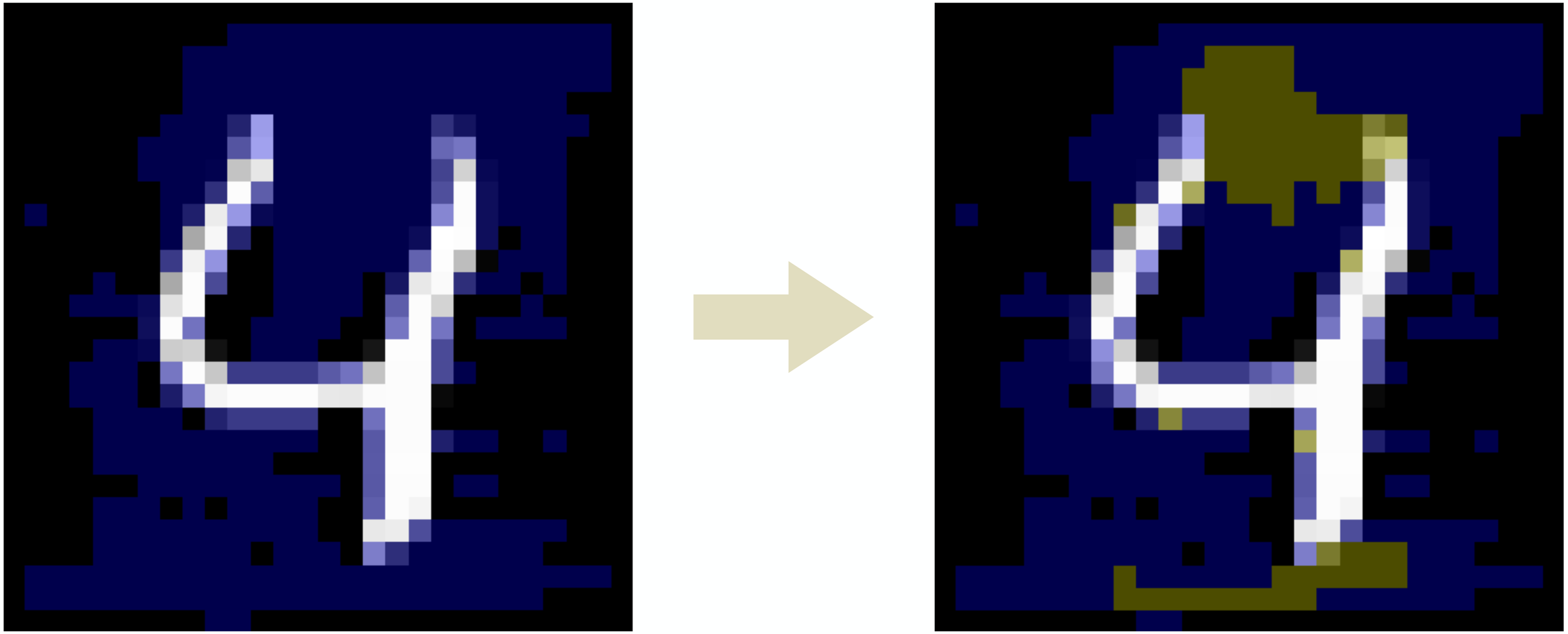
FINDING COUNTERFACTUALS RAPIDLY BECOMES INFEASIBLE

Model	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
	Counterfactuals	**Unknowns**	Time	**Counterfactuals**	**Unknowns**	Time
CNN-3	0.00	247.80	45m	0.00	461.00	2h 30m

Model	MNIST, ϵ=0.25			CIFAR-10, ϵ=16/255		
	Counterfactuals	**Unknowns**	Time	**Counterfactuals**	**Unknowns**	Time
CNN-7	0.00	452.00	3h 59m	0.00	730.67	7h 5m

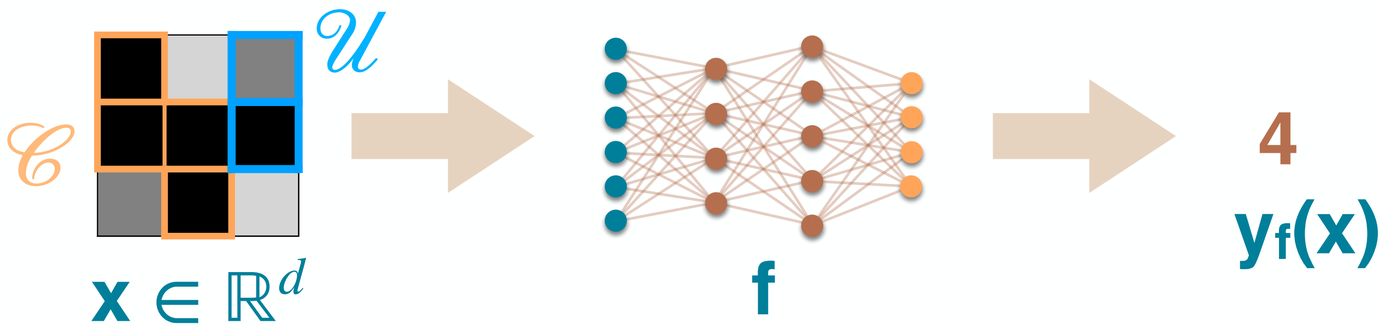
Verifier-Optimal Robust Explanations

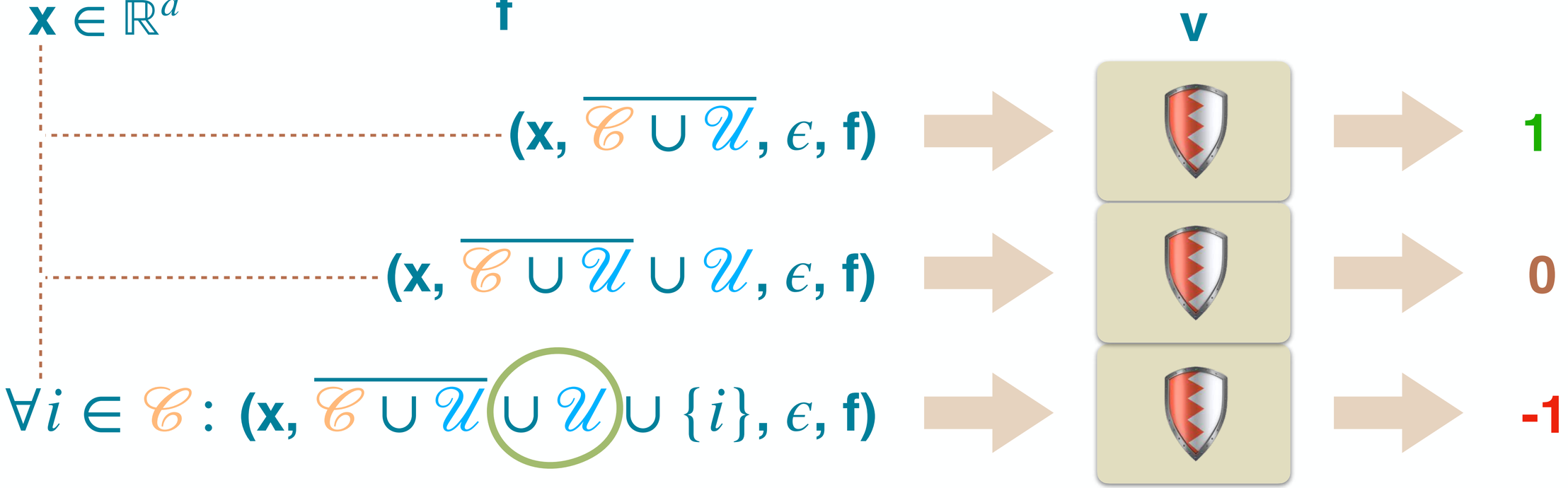
OUR 1ST CONTRIBUTION



Verifier-Optimal Robust Explanations

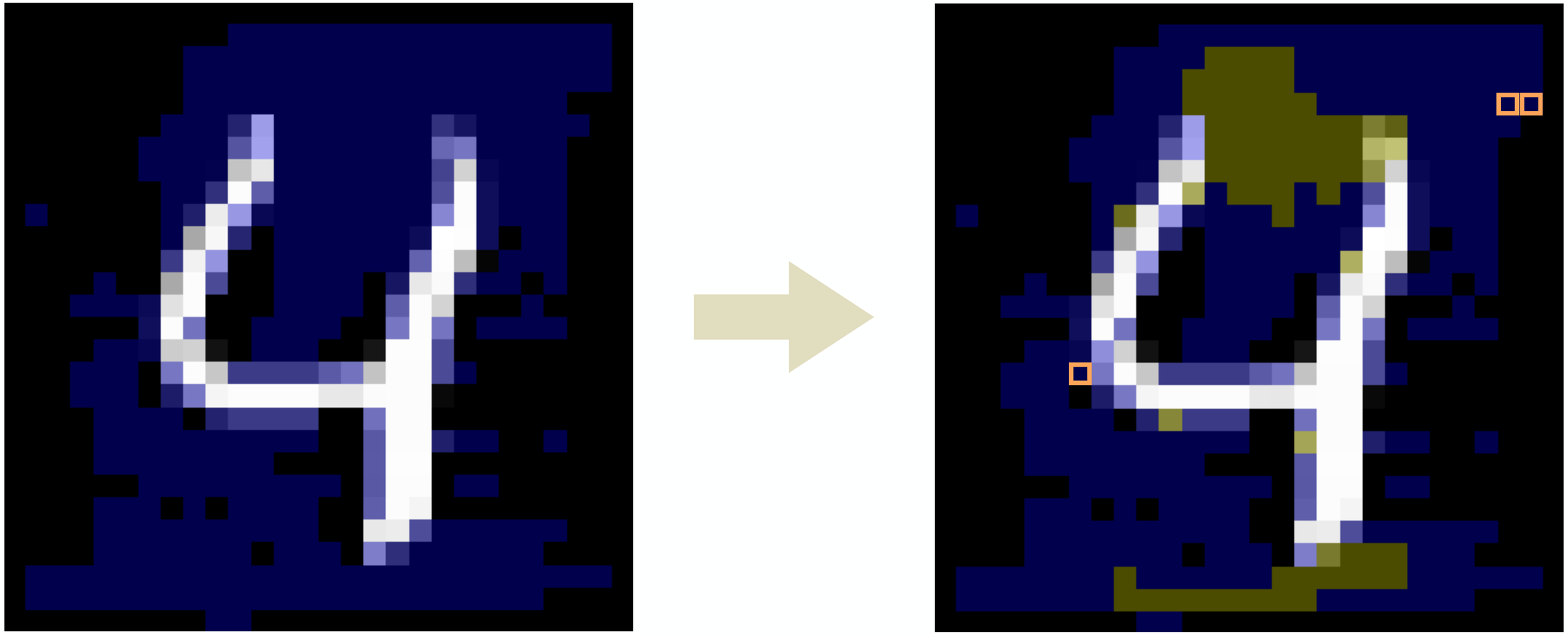
WEAK ABDUCTIVE EXPLANATIONS





Verifier-Optimal Robust Explanations

OUR 1ST CONTRIBUTION

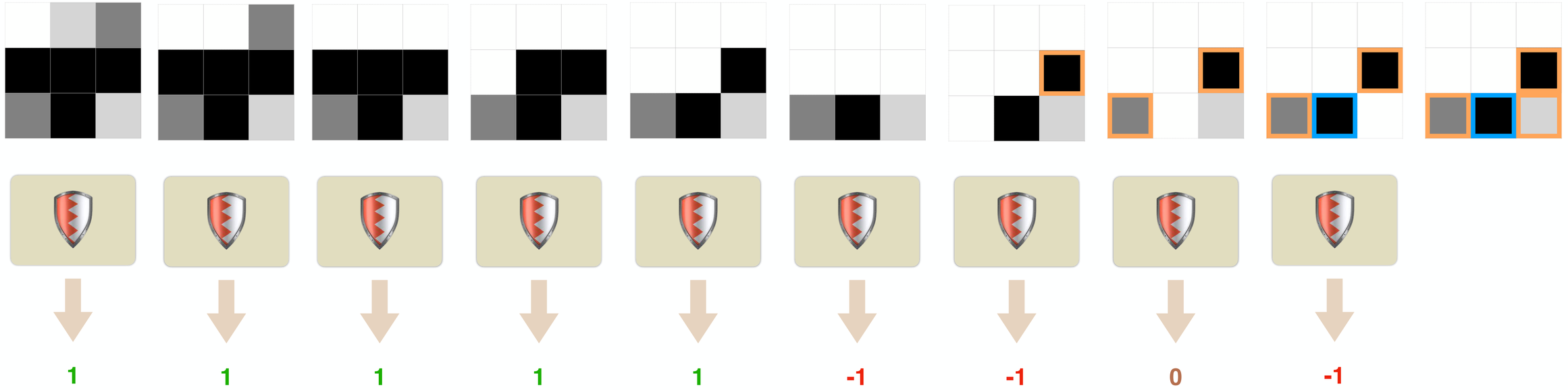


Computing Verifier-Optimal Robust Explanations

DROP (I.E., FREE) INPUT DIMENSIONS WHILE AXp CONDITION HOLDS

ADD TO $\overline{\mathscr{C} \cup \mathscr{U}}$

LOCAL ROBUSTNESS IN $B^{\epsilon}_{\overline{\mathscr{C} \cup \mathscr{U}}}(\mathbf{x})$



Computing Robust Explanations

CNN-3

Robust Explanations	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
	Counterfactuals $\lvert\mathscr{C}\rvert$	**Unknowns** $\lvert\mathscr{U}\rvert$	Time	**Counterfactuals** $\lvert\mathscr{C}\rvert$	**Unknowns** $\lvert\mathscr{U}\rvert$	Time
~~Optimal~~	0.00	247.80	45m	0.00	461.00	2h 30m
Verifier-Optimal	**129.10**	125.50	**19m**	**209.30**	253.10	**21m**

129.10 +
125.50 =
254.60

209.30 +
253.10 =
462.40

Computing Robust Explanations

CNN-7

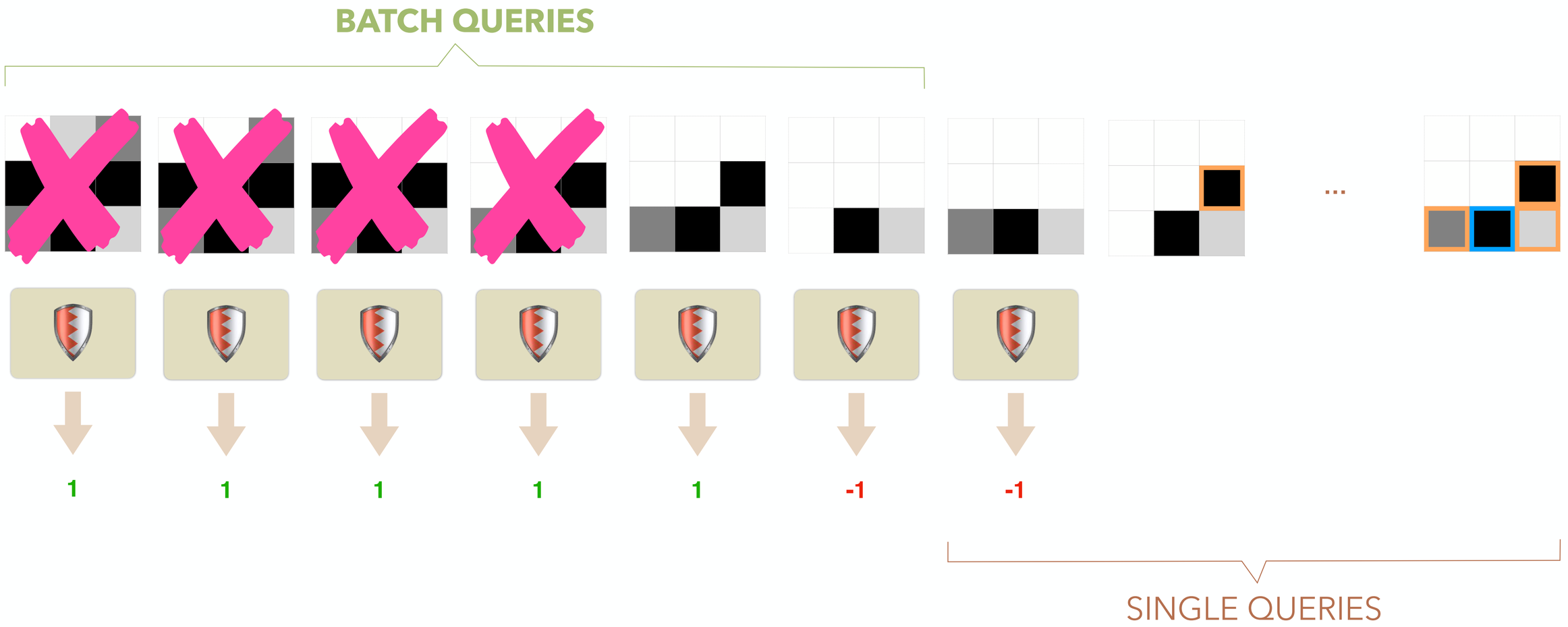
Robust Explanations	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
	Counterfactuals $\lvert\mathscr{C}\rvert$	**Unknowns** $\lvert\mathscr{U}\rvert$	Time	**Counterfactuals** $\lvert\mathscr{C}\rvert$	**Unknowns** $\lvert\mathscr{U}\rvert$	Time
~~Optimal~~	0.00	452.00	3h 59m	0.00	730.67	7h 5m
Verifier-Optimal	**142.33**	315.00	**2h 13m**	**464.00**	269.33	**1h 54m**

142.33 + 315.00 = 457.33

464.00 + 269.33 = 733.33

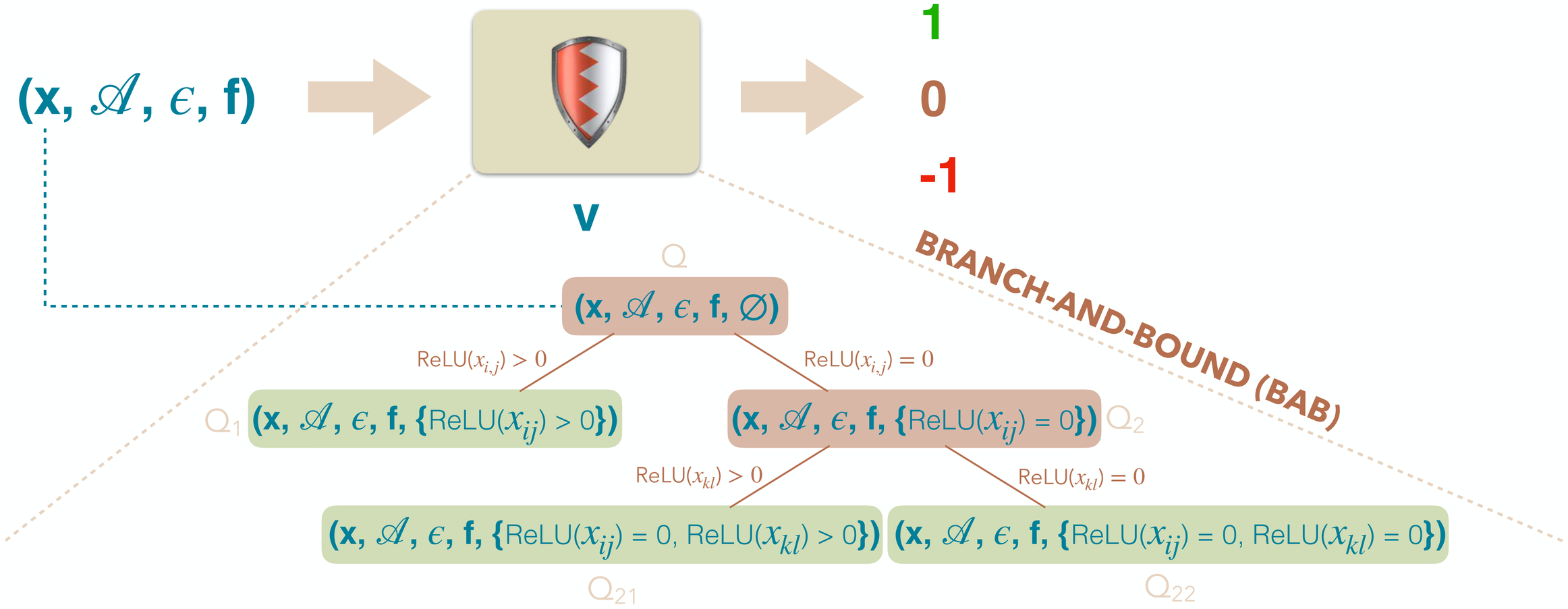
FaVeX

OUR 2ND CONTRIBUTION



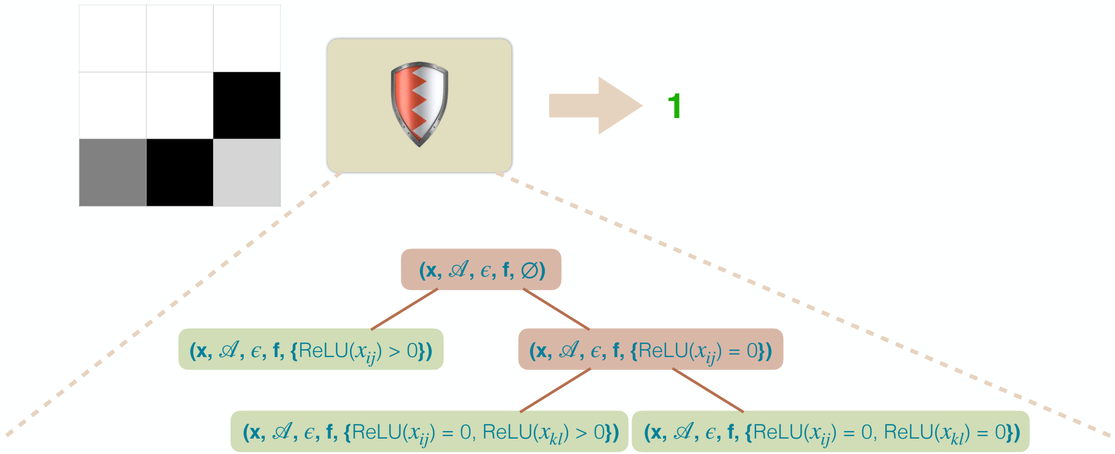
FaVeX

OUR 2ND CONTRIBUTION

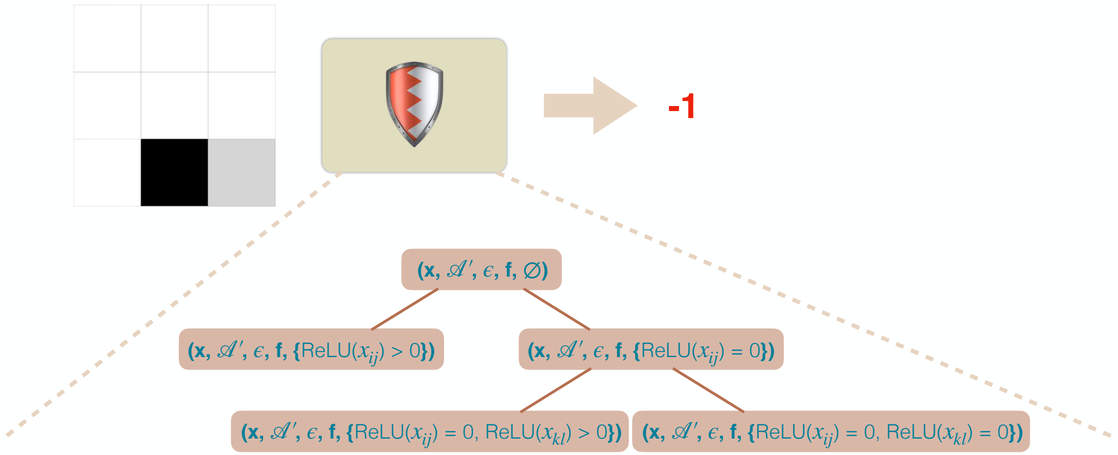


FaVeX

OUR 2ND CONTRIBUTION



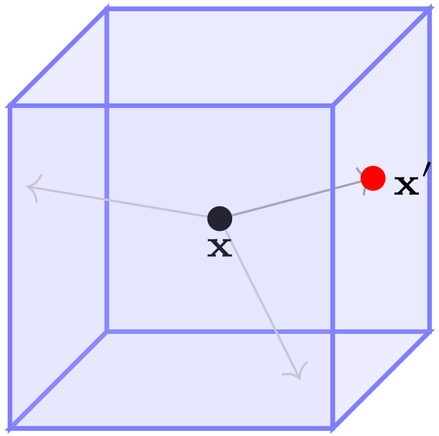
INCREMENTAL BAB VERIFICATION

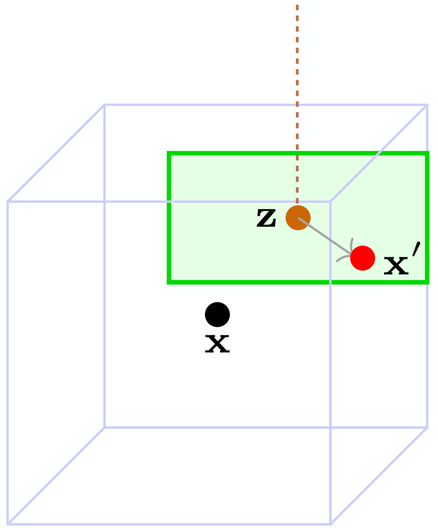


FaVeX

OUR 2ND CONTRIBUTION

output of the previous counterfactual search





$B^{\epsilon}_{\mathscr{C} \cup \mathscr{U}}(\mathbf{x})$

Full-Space Search

subset of $B^{\epsilon}_{\mathscr{C} \cup \mathscr{U}}(\mathbf{x})$

Reduced Space Search

Computing Verifier-Optimal Robust Explanations

CNN-3

Verifier-Optimal Explanations	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
	$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time	$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time
Single Queries	129.10	125.50	19m	209.30	253.10	21m
FaVeX	**160.30**	94.40	**10m**	**210.40**	251.70	**19m**

Computing Verifier-Optimal Robust Explanations

CNN-7

Verifier-Optimal Explanations	MNIST, ϵ=0.25			CIFAR-10, ϵ=16/255		
	$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time	$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time
Single Queries	142.33	315.00	2h 13m	464.00	269.33	1h 54m
FaVeX	**207.33**	249.33	**1h 14m**	467.00	266.33	**1h 49m**

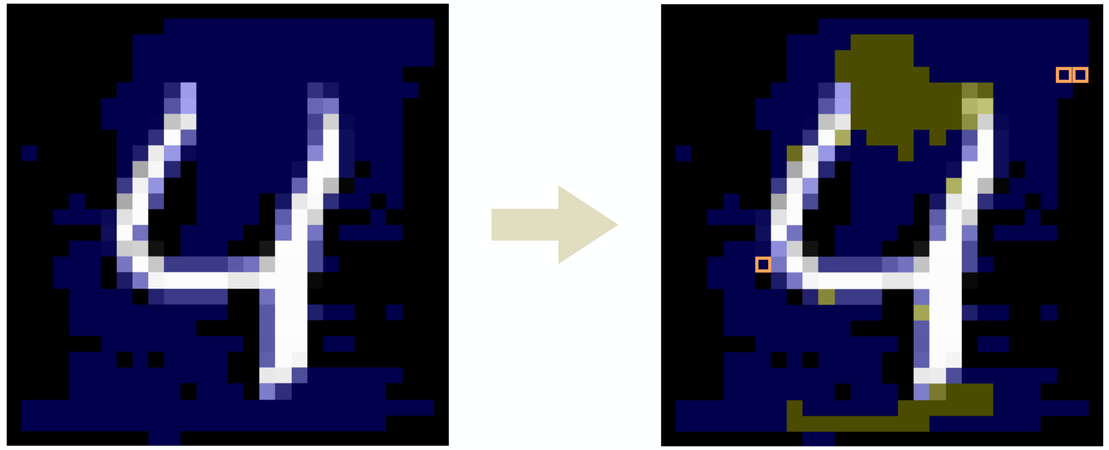
Computing Verifier-Optimal Robust Explanations

TRAVERSAL STRATEGIES

Model	Traversal	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
		$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time	$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time
CNN-3	VeriX	160.50	122.30	16m	437.20	328.30	32m
	VeriX+	**155.20**	**92.20**	10m	262.00	206.90	15m
	α-FAVEX	181.20	108.70	12m	**210.40**	**251.70**	19m
	FaVeX-IBP	160.30	94.40	10m	250.90	215.50	16m
Model	Traversal	MNIST, ϵ=0.25			CIFAR-10, ϵ=16/255		
		$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time	$\lvert\mathscr{C}\rvert$	$\lvert\mathscr{U}\rvert$	Time
CNN-7	VeriX	123.33	423.67	2h 9m	728.67	216.33	1h 23m
	VeriX+	**196.67**	**232.67**	1h 13m	522.00	213.67	1h 25m
	α-FAVEX	234.67	317.67	1h 29m	467.00	266.33	1h 49m
	FaVeX-IBP	207.33	249.33	1h 14m	**512.67**	**216.67**	1h 25m

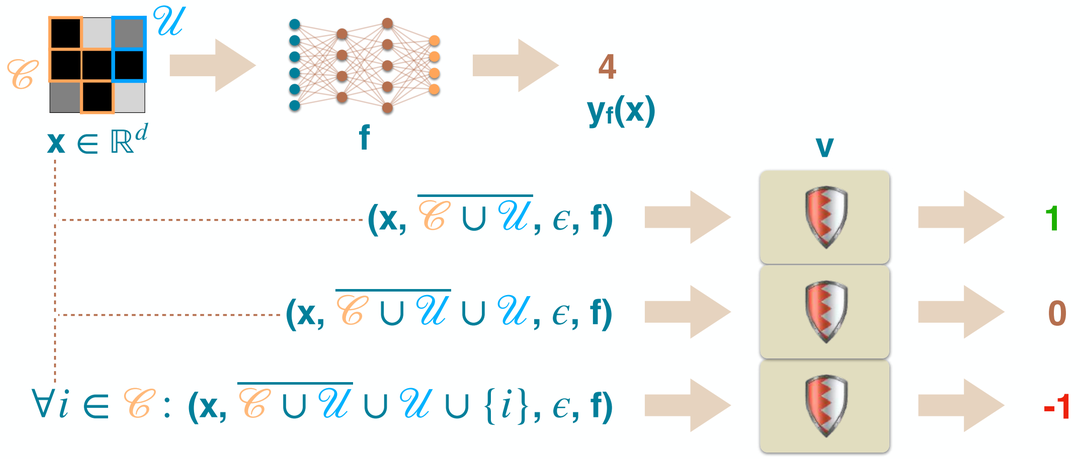
Verifier-Optimal Robust Explanations

OUR 1ST CONTRIBUTION



Verifier-Optimal Robust Explanations

WEAK ABDUCTIVE EXPLANATIONS



FaVeX

OUR 2ND CONTRIBUTION

Robust Explanations	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
	$\lvert\mathcal{C}\rvert$	$\lvert\mathcal{U}\rvert$	Time	$\lvert\mathcal{C}\rvert$	$\lvert\mathcal{U}\rvert$	Time
~~Optimal~~	0.00	452.00	3h 59m	0.00	730.67	7h 5m
Verifier-Optimal (Single Queries)	142.33	315.00	2h 13m	464.00	269.33	1h 54m
Verifier-Optimal (FaVeX)	**207.33**	249.33	**1h 14m**	**467.00**	266.33	**1h 49m**

OUR PAPER

OUR TOOL

THANKS!