From Contrastive to Abductive Explanations and Back Again

Alexey Ignatiev[1], Nina Narodytska[2], Nicholas Asher[3], and Joao Marques-Silva[3]

November 9, 2021 | **KR**

[1] Monash University, Melbourne, Australia
[2] VMware Research, CA, USA
[3] IRIT, CNRS, Toulouse, France

Motivation

Ongoing ML Revolution

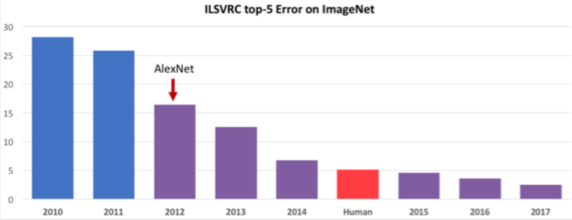
https://en.wikipedia.org/wiki/Waymo

DeepMind

AlphaGo

AlphaGo Zero & **Alpha Zero**

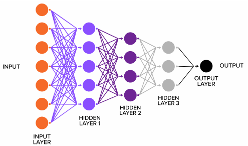
Image & Speech Recognition



http://gradientscience.org/intro_adversarial/

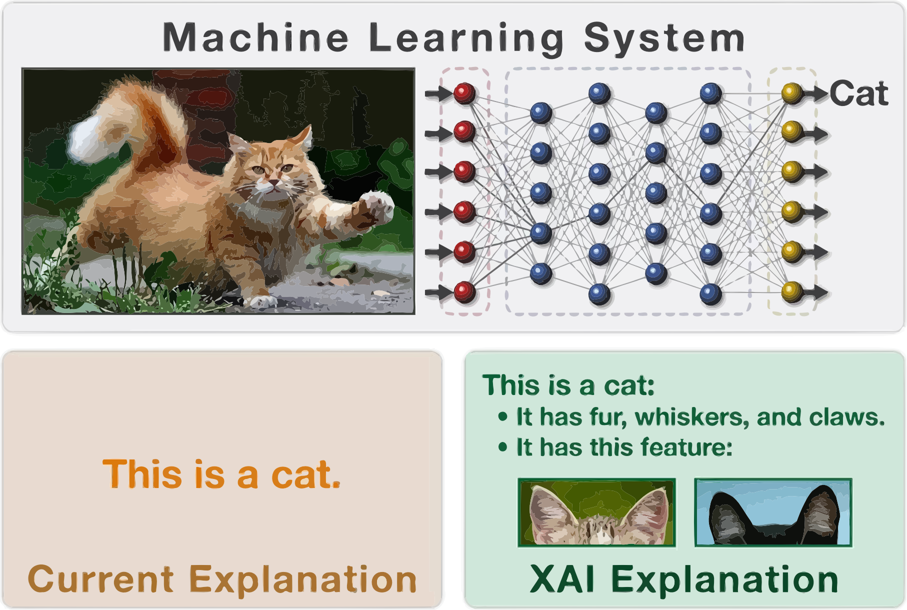
https://fr.wikipedia.org/wiki/Pepper_(robot)

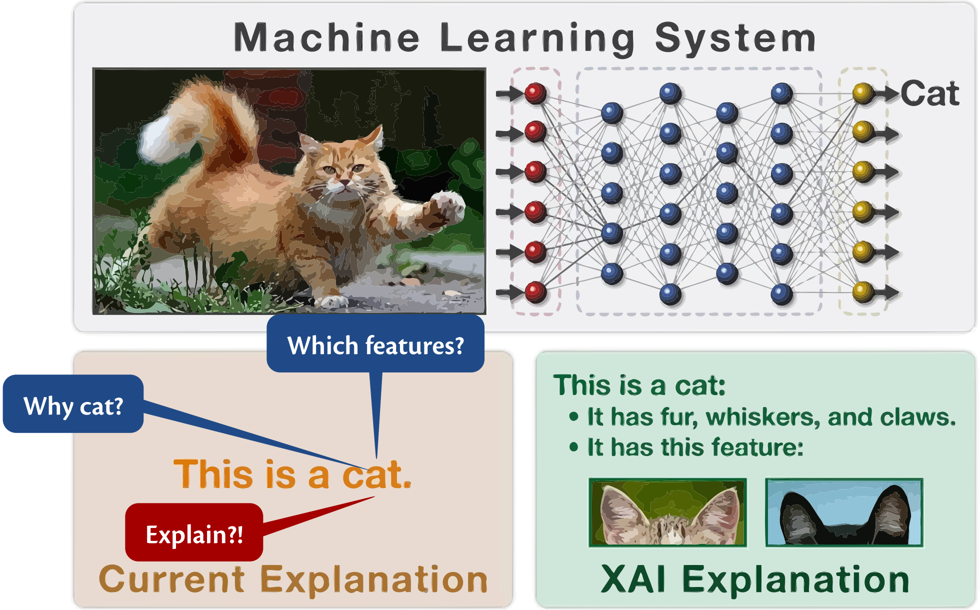
And yet...



	A parrot	Machine learning algorithm
Learns random phrases	✓	✓
Doesn’t understand st about what it learns**	✓	✓
Occasionally speaks nonsense	✓	✓

eXplainable AI




Machine Learning System
Cat
Which features?
Why cat?
This is a cat.
Explain?!
Current Explanation
This is a cat:
• It has fur, whiskers, and claws.
• It has this feature:
XAI Explanation

Formal explanations

[SCD18, INMS19, INM19, DH20, MGC+20, INAM20, IIM20, IS21, MGC+21, IM21, HIIM21]

classifier $\tau : \mathbb{F} \to \mathcal{K}$, **instance** $\mathbf{v}$ **s.t.** $\tau(\mathbf{v}) = \mathbf{c}$

 [SCD18, INMS19, INM19, DH20, MGC+20, INAM20, IIM20, IS21, MGC+21, IM21, HIIM21]

classifier $\tau : \mathbb{F} \to \mathcal{K}$, **instance** $\mathbf{v}$ **s.t.** $\tau(\mathbf{v}) = \mathbf{c}$

---

abductive explanation $\mathcal{X}$

$$\forall(\mathbf{x} \in \mathbb{F}).\bigwedge_{j\in\mathcal{X}}(x_j = v_j) \rightarrow (\tau(\mathbf{x}) = \mathbf{c})$$

[SCD18, INMS19, INM19, DH20, MGC+20, INAM20, IIM20, IS21, MGC+21, IM21, HIIM21]

classifier $\tau : \mathbb{F} \to \mathcal{K}$, **instance** $\mathbf{v}$ **s.t.** $\tau(\mathbf{v}) = \mathbf{c}$

---

abductive explanation $\mathcal{X}$

"why?"

$$\forall(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \in \mathcal{X}} (x_j = v_j) \to (\tau(\mathbf{x}) = \mathbf{c})$$

[SCD18, INMS19, INM19, DH20, MGC+20, INAM20, IIM20, IS21, MGC+21, IM21, HIIM21]

classifier $\tau : \mathbb{F} \to \mathcal{K}$, **instance** $\mathbf{v}$ **s.t.** $\tau(\mathbf{v}) = \mathbf{c}$

---

abductive explanation $\mathcal{X}$

"why?"

$$\forall(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \in \mathcal{X}} (x_j = v_j) \rightarrow (\tau(\mathbf{x}) = \mathbf{c})$$

---

contrastive explanation $\mathcal{Y}$

$$\exists(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \notin \mathcal{Y}} (x_j = v_j) \land (\tau(\mathbf{x}) \neq \mathbf{c})$$

[SCD18, INMS19, INM19, DH20, MGC+20, INAM20, IIM20, IS21, MGC+21, IM21, HIIM21]

classifier $\tau : \mathbb{F} \to \mathcal{K}$, instance $\mathbf{v}$ s.t. $\tau(\mathbf{v}) = \mathbf{c}$

---

abductive explanation $\mathcal{X}$ — "why?"

$$\forall(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \in \mathcal{X}} (x_j = v_j) \to (\tau(\mathbf{x}) = \mathbf{c})$$

---

contrastive explanation $\mathcal{Y}$ — "why not?"

$$\exists(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \notin \mathcal{Y}} (x_j = v_j) \land (\tau(\mathbf{x}) \neq \mathbf{c})$$

this work!

Explanation examples

[IS21, INAM20]

$$\mathbb{F} = \{\mathbf{0}, \mathbf{1}, \mathbf{2}\}^{\mathbf{5}} \qquad \mathcal{K} = \{\ominus, \oplus\}$$

Explanation examples

[IS21, INAM20]

$$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

[IS21, INAM20]

$$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$

Explanation examples

[IS21, INAM20]

$$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$

AXps $\mathbb{X} = \{\{1, 2\}, \{3\}\}$

Explanation examples

[IS21, INAM20]

$$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$$

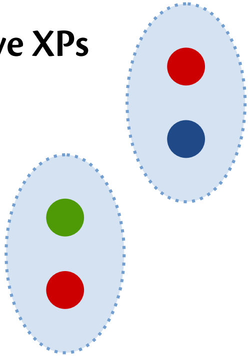
R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$

AXps $\mathbb{X} = \{\{1, 2\}, \{3\}\}$

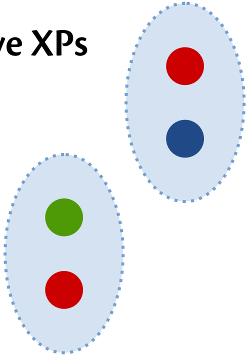
CXps $\mathbb{Y} = \{\{1, 3\}, \{2, 3\}\}$

Abductive XPs

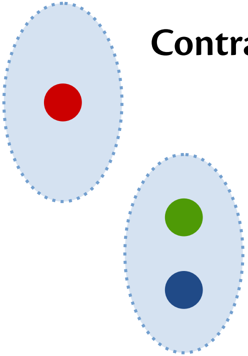


Minimal hitting set duality

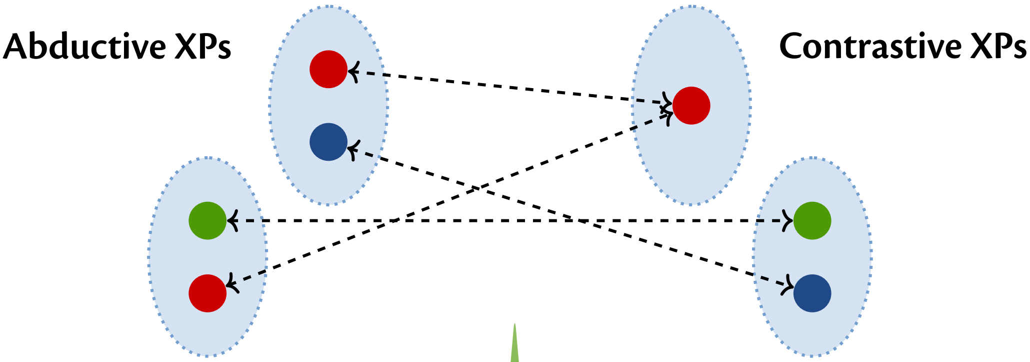
Abductive XPs



Contrastive XPs



Minimal hitting set duality



AXps are minimal hitting sets of CXps, and vice versa

CXp computation

CXp computation – example

$$\mathbb{F} = \{\mathbf{0}, \mathbf{1}, \mathbf{2}\}^{\mathbf{5}} \qquad \mathcal{K} = \{\ominus, \oplus\}$$

CXp computation – example

$$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

CXp computation – example

$$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$

CXp computation – example

$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$ **– why not $\oplus$?**

CXp computation – example

$\mathbb{F} = \{0,1,2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1,1,1,1,1) = \ominus$ **– why not $\oplus$?**

1. *can we drop feature 1?* $\quad \tau(1,*,*,*,*) \not\equiv \ominus$?

$\exists(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \notin \mathcal{Y}} (x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$

CXp computation – example

$$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$ **– why not $\oplus$?**

1. *can we drop feature 1?* $\tau(1, *, *, *, *) \not\equiv \ominus$? Yes

$\exists(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \not\in \mathcal{Y}} (x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$

CXp computation – example

$\mathbb{F} = \{0, 1, 2\}^5 \quad \mathcal{K} = \{\ominus, \oplus\}$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$ **– why not $\oplus$?**

1. *can we drop feature 1?* $\tau(1, *, *, *, *) \not\equiv \ominus$? **Yes**
2. *can we drop feature 2?* $\tau(1, 1, *, *, *) \not\equiv \ominus$?

$\exists(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \not\in \mathcal{Y}} (x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$

CXp computation – example

$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$ **– why not $\oplus$?**

1. *can we drop feature 1?* $\tau(1, *, *, *, *) \not\equiv \ominus$? Yes
2. *can we drop feature 2?* $\tau(1, 1, *, *, *) \not\equiv \ominus$? No

$\exists(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \not\in \mathcal{Y}} (x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$

CXp computation – example

$$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$ **– why not $\oplus$?**

1. *can we drop feature 1?* $\tau(1, *, *, *, *) \not\equiv \ominus$? Yes
2. *can we drop feature 2?* $\tau(1, 1, *, *, *) \not\equiv \ominus$? No
3. *can we drop feature 3?* $\tau(1, *, 1, *, *) \not\equiv \ominus$?

$\exists(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \not\in \mathcal{Y}} (x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$

CXp computation – example

$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$ **– why not $\oplus$?**

1. *can we drop feature 1?* $\tau(1, *, *, *, *) \not\equiv \ominus$? Yes
2. *can we drop feature 2?* $\tau(1, 1, *, *, *) \not\equiv \ominus$? No
3. *can we drop feature 3?* $\tau(1, *, 1, *, *) \not\equiv \ominus$? No
4. *can we drop feature 4?* $\tau(1, *, *, 1, *) \not\equiv \ominus$?

$\exists(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \notin \mathcal{Y}} (x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$

CXp computation – example

$$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$ **– why not $\oplus$?**

1. *can we drop feature 1?* $\tau(1, *, *, *, *) \not\equiv \ominus$? Yes
2. *can we drop feature 2?* $\tau(1, 1, *, *, *) \not\equiv \ominus$? No
3. *can we drop feature 3?* $\tau(1, *, 1, *, *) \not\equiv \ominus$? No
4. *can we drop feature 4?* $\tau(1, *, *, 1, *) \not\equiv \ominus$? Yes
5. *can we drop feature 5?* $\tau(1, *, *, 1, 1) \not\equiv \ominus$?

$$\exists(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \not\in \mathcal{Y}} (x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$$

CXp computation – example

$$\mathbb{F} = \{0,1,2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1,1,1,1,1) = \ominus$ **– why not $\oplus$?**

1. *can we drop feature 1?* $\tau(1,*,*,*,*) \not\equiv \ominus$? Yes
2. *can we drop feature 2?* $\tau(1,1,*,*,*) \not\equiv \ominus$? No
3. *can we drop feature 3?* $\tau(1,*,1,*,*) \not\equiv \ominus$? No
4. *can we drop feature 4?* $\tau(1,*,*,1,*) \not\equiv \ominus$? Yes
5. *can we drop feature 5?* $\tau(1,*,*,1,1) \not\equiv \ominus$? Yes

$$\exists(\mathbf{x} \in \mathbb{F}).\bigwedge_{j \notin \mathcal{Y}}(x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$$

CXp computation – example

$\mathbb{F} = \{0, 1, 2\}^5 \qquad \mathcal{K} = \{\ominus, \oplus\}$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$ **– why not $\oplus$?**

1. *can we drop feature 1?* $\tau(1, *, *, *, *) \not\equiv \ominus$? Yes
2. *can we drop feature 2?* $\tau(1, 1, *, *, *) \not\equiv \ominus$? No
3. *can we drop feature 3?* $\tau(1, *, 1, *, *) \not\equiv \ominus$? No
4. *can we drop feature 4?* $\tau(1, *, *, 1, *) \not\equiv \ominus$? Yes
5. *can we drop feature 5?* $\tau(1, *, *, 1, 1) \not\equiv \ominus$? Yes

CXp $\mathcal{Y} = \{2, 3\}$

$\exists(\mathbf{x} \in \mathbb{F}). \bigwedge_{j \not\in \mathcal{Y}} (x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$

Explanation Enumeration

Function XpEnum(τ, $\mathbf{v}$, c)

Input: τ: ML model, $\mathbf{v}$: Input instance, $c = \tau(\mathbf{v})$: Prediction

$\mathcal{K} = (\mathcal{N}, \mathcal{P}) \leftarrow (\emptyset, \emptyset)$ // Block AXps & CXps

while true:

  $(st_\lambda, \lambda) \leftarrow$ `FindMHS`$(\mathcal{P}, \mathcal{N})$ // MHS of $\mathcal{P}$ s.t. $\mathcal{N}$

  if $\neg st_\lambda$: **break**

  $st_{c'} \leftarrow$ SAT$(\bigwedge_{j \in \lambda}(x_j = v_j) \wedge \tau(\mathbf{x}) \neq c)$

  if $\neg st_{c'}$: // entailment holds

    `ReportAXp`(λ)

    $\mathcal{N} \leftarrow \mathcal{N} \cup \bigvee_{j \in \lambda}(x_j \neq v_j)$

  else:

    $\mu \leftarrow$ `ExtractCXp`$(\tau, \mathbf{v}, c, \mathcal{P})$

    `ReportCXp`(μ)

    $\mathcal{P} \leftarrow \mathcal{P} \cup \bigvee_{j \in \mu}(x_j = v_j)$

Explanation Enumeration

Function XpEnum(τ, $\mathbf{v}$, c)

Input: τ: ML model, $\mathbf{v}$: Input instance, $c = \tau(\mathbf{v})$: Prediction

$\mathcal{K} = (\mathcal{N}, \mathcal{P}) \leftarrow (\emptyset, \emptyset)$ // Block AXps & CXps

while true:

  $(st_\lambda, \lambda) \leftarrow$ `FindMHS`$(\mathcal{P}, \mathcal{N})$ // MHS of $\mathcal{P}$ s.t. $\mathcal{N}$

  if $\neg st_\lambda$: **break**

  $st_{c'} \leftarrow$ `SAT`$(\bigwedge_{j \in \lambda}(x_j = v_j) \wedge \tau(\mathbf{x}) \neq c)$

  if $\neg st_{c'}$: // entailment holds

    `ReportAXp`(λ)

    $\mathcal{N} \leftarrow \mathcal{N} \cup \bigvee_{j \in \lambda}(x_j \neq v_j)$

  else:

    $\mu \leftarrow$ `ExtractCXp`$(\tau, \mathbf{v}, c, \mathcal{P})$

    `ReportCXp`(μ)

    $\mathcal{P} \leftarrow \mathcal{P} \cup \bigvee_{j \in \mu}(x_j = v_j)$

implicit hitting set enumeration!

see paper for details

Conclusions

Conclusions

- **formal definition of contrastive explanations**
  - **similar to abductive explanations**

Conclusions

- formal definition of contrastive explanations
  - similar to abductive explanations

- minimal hitting set duality between CXps and AXps
  - explanation enumeration algorithms
  - solving membership problems

Conclusions

- formal definition of contrastive explanations
  - similar to abductive explanations

- minimal hitting set duality between CXps and AXps
  - explanation enumeration algorithms
  - solving membership problems

proved helpful in several papers!

Conclusions

- **formal definition of contrastive explanations**
  - **similar to abductive explanations**

- **minimal hitting set duality between CXps and AXps**
  - **explanation enumeration algorithms**
  - **solving membership problems**

proved helpful in several papers!

- **experimental results**
  - **XP enumeration**
  - **CXp enumeration – *helps to debug SHAP***

Questions?

References i

Adnan Darwiche and Auguste Hirth.
On the reasons behind decisions.
In *ECAI*, pages 712–720, 2020.

Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva.
On efficiently explaining graph-based classifiers.
In *KR*, 2021.

Yacine Izza, Alexey Ignatiev, and João Marques-Silva.
On explaining decision trees.
CoRR, abs/2010.11034, 2020.

Yacine Izza and João Marques-Silva.
On explaining random forests with SAT.
In *IJCAI*, pages 2584–2591, 2021.

References ii

Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and João Marques-Silva.
From contrastive to abductive explanations and back again.
In *AI*IA*, pages 335–355, 2020.

Alexey Ignatiev, Nina Narodytska, and João Marques-Silva.
On relating explanations and adversarial examples.
In *NeurIPS*, pages 15857–15867, 2019.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva.
Abduction-based explanations for machine learning models.
In *AAAI*, pages 1511–1519, 2019.

Alexey Ignatiev and João P. Marques Silva.
SAT-based rigorous explanations for decision lists.
In *SAT*, pages 251–269, 2021.

João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska.
Explaining naive bayes and other linear classifiers with polynomial time and delay.
In *NeurIPS*, 2020.

João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska.
Explanations for monotonic classifiers.
In *ICML*, pages 7469–7479, 2021.

Andy Shih, Arthur Choi, and Adnan Darwiche.
A symbolic approach to explaining Bayesian network classifiers.
In *IJCAI*, pages 5103–5111, 2018.