



**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO  
CEARÁ  
IFCE *CAMPUS* MARACANAÚ  
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

**LEVI CORDEIRO CARVALHO**

**EXPLICAÇÕES PARA REDES NEURAIIS BASEADAS EM RACIOCÍNIO  
ABDUTIVO**

**MARACANAÚ - CE  
2021**

**LEVI CORDEIRO CARVALHO**

**EXPLICAÇÕES PARA REDES NEURAIS BASEADAS EM RACIOCÍNIO  
ABDUTIVO**

Trabalho de Conclusão de Curso apresentado ao  
Curso de Bacharelado em Ciência da Computa-  
ção do Instituto Federal de Educação, Ciência  
e Tecnologia do Ceará -IFCE - *Campus* Mara-  
canaú como requisito parcial para obtenção do  
Título de Bacharel em Ciência da Computação.  
Orientador: Prof. Dr. Thiago Alves Rocha.

**MARACANAÚ - CE**

**2021**

NOME COMPLETO

## TÍTULO DO TRABALHO

Esta Monografia foi julgada adequada para obtenção do título de Licenciado em matemática e aprovada em sua forma final pelo departamento de Matemática do Instituto Federal do Ceará-*Campus Cedro*.

Aprovado em: \_\_\_\_\_ / \_\_\_\_\_ / \_\_\_\_\_

## BANCA EXAMINADORA

---

Prof. Me. Luiz Fernando Ramos Lemos (Orientador)  
IFSULDEMINAS - *Campus Inconfidentes*.

---

Prof.(a). Ma. Mikaelle Barboza Cardoso  
IFCE - *Campus Sobral*

---

Prof. Dr. João Nunes de Araújo Neto  
IFCE - *Campus Cedro*

## **DEDICATÓRIA**

À minha mãe, ...

“Astronarta libertado  
Minha vida me urtrapassa  
Em quarqué rota que eu faça.”

(Dois mil e um - Tom Zé)

## **AGRADECIMENTOS**

Graças à vida, que me deu tanto...

Ao Instituto Federal de Educação, Ciência e Tecnologia do Ceará - *Campus* Cedro, todos os servidores, professores e alunos.

Não esqueça de agradecer às instituições que lhe forneceram algum tipo de financiamento ao longo da graduação!!!

## RESUMO

Resumo em português

**Palavras-chave:** Matemática. Educação. Função Afim. Função Definida por Partes. PDI.

## ABSTRACT

English abstract.

**Keywords:** Mathematics. Education. Affine Function. Piecewise Function. DIP.



## **LISTA DE ILUSTRAÇÕES**

Quadro 1	— LIVROS ANALISADOS . . . . .	17
----------	-------------------------------	----

## **LISTA DE FIGURAS**

Figura 1	— Representação gráfica da função afim . . . . .	18
----------	--------------------------------------------------	----

## LISTA DE CÓDIGOS

Código 1	— Método da Bissecção . . . . .	20
----------	---------------------------------	----

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>12</b>
<b>1.1</b>	<b>Resumo dos artigos selecionados</b>	<b>14</b>
<b>2</b>	<b>MOTIVAÇÃO INICIAL E JUSTIFICATIVA FUNDAMENTADA DO ESTUDO</b>	<b>17</b>
<b>3</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>18</b>
<b>3.1</b>	<b>Teoria dos Registros das Representações Semióticas</b>	<b>18</b>
<b>3.2</b>	<b>Exemplo de seção.</b>	<b>19</b>
<b>3.2.1</b>	<i>Exemplo de subseção</i>	19
<b>3.2.2</b>	<i>Imagem da função afim</i>	19
<b>3.2.3</b>	<i>Zero da função</i>	19
<b>3.2.4</b>	<i>Exemplo de subseção</i>	19
<b>3.3</b>	<b>Outra seção</b>	<b>20</b>
<b>4</b>	<b>ABORDAGEM AO PROBLEMA</b>	<b>21</b>
<b>5</b>	<b>CONSIDERAÇÕES FINAIS</b>	<b>22</b>
<b>ANEXO A</b>	<b>Apêndice</b>	<b>23</b>
<b>ANEXO A.1</b>	<b>Texto auxiliar do trabalho</b>	<b>24</b>
<b>REFERÊNCIAS</b>		<b>26</b>

## 1 INTRODUÇÃO

As redes neurais artificiais (do inglês, *Artificial Neural Network* - ANN) são utilizadas em diversas aplicações para resolução de problemas, como visão computacional, reconhecimento de fala e de padrões (LIU et al., 2017). Para que esses algoritmos alcancem um resultado satisfatório, um dos principais requisitos é realizar o treinamento da rede com o uso de um conjunto de dados que representem bem os casos do mundo real, com o objetivo de garantir a generalização do aprendizado para dados não vistos.

Apesar do grande sucesso das redes neurais, elas podem ser classificadas como um algoritmo caixa preta, sendo, basicamente, o funcionamento computado para a aquisição da resposta da ANN, dado um conjunto de entradas, não humanamente interpretável. Essa falta de explicações para o seu comportamento pode prejudicar sistemas críticos que permitem apenas uma pequena margem para falhas, como aplicações médicas e financeiras. Além disso, com o avanço de leis de proteção de dados observável em diversas regiões do mundo, como na União Europeia (COMISSÃO EUROPEIA, 2018), está ficando cada vez mais necessário disponibilizar uma explicação do que está sendo realizado com os dados dos seus respectivos donos.

A existência de exemplos adversariais é outro motivo para a necessidade do desenvolvimento de abordagens que computam explicações para algoritmos de aprendizagem de máquina, incluindo redes neurais. De acordo com o trabalho de Goodfellow, Shlens e Szegedy (2015), exemplos adversariais são instâncias classificadas erroneamente por um modelo de aprendizagem de máquina que são minimamente diferentes de uma outra instância classificada corretamente, demonstrando a fraqueza desses algoritmos para pequenas perturbações nos dados de entrada que, muitas vezes, poderiam terem sido adquiridas na própria etapa de aquisição do conjunto de dados.

Os métodos heurísticos são as principais abordagens utilizadas com o intuito de disponibilizar explicações para as saídas de qualquer modelo de aprendizagem de máquina atualmente, pelo fato de serem considerados algoritmos *model-agnostic*, como o LIME (RIBEIRO; SINGH; GUESTRIN, 2016) e ANCHOR (RIBEIRO; SINGH; GUESTRIN, 2018). Porém, de acordo com o trabalho de Ignatiev, Narodytska e Marques-Silva (2019b), como eles exploram o espaço da instância localmente, esses algoritmos não acarretam em respostas que possuam garantias formais, resultando em explicações

locais que podem ser otimistas (quando existem instâncias no espaço de instâncias que a explicação computada falha) ou pessimistas (quando alguma *feature* pertencente à explicação computada é irrelevante e pode ser descartada).

Assim, abordagens não heurísticas para a computação dessas explicações tem sido abordadas em alguns trabalhos recentes. O artigo de Ignatiev, Narodytska e Marques-Silva (2019a) apresenta um algoritmo *model-agnostic*, assim como os métodos heurísticos, porém a computação das explicações mínimas geradas carregam uma garantia formal, pois a sua proposta é baseada em lógica utilizando o raciocínio abdutivo. A abordagem basicamente codifica um modelo  $M$  de aprendizagem de máquina como um conjunto de restrições lineares  $F$  em alguma teoria, a partir de um verificador de rede neural, que no caso é usado a modelagem de um problema de Programação Linear Inteira Mista (do inglês, *Mixed Integer Linear Programming* - MILP) proposto por Fischetti e Jo (2018), então, dado uma instância  $C$  associado com a predição  $E$  tal que  $C \wedge F \models E$  (equivalente à  $C \models (F \rightarrow E)$ ), é computada a explicação mínima  $C'$  (dado  $C$ ) que é implicante principal de  $F \rightarrow E$ .

Ressalta-se, como um dos principais problemas da proposta de Ignatiev, Narodytska e Marques-Silva (2019a), a escalabilidade do algoritmo, podendo demorar um tempo de execução impraticável para ser colocado em produção quando se deseja obter explicações de redes neurais com várias camadas possuindo muitos neurônios. No artigo é apontado que os testes foram realizados apenas com redes de uma camada oculta com 10, 15 ou 20 neurônios. Além disso, os autores não utilizaram outras possíveis modelagens da rede para tentar resolver o problema da escalabilidade. Porém, é possível encontrar outras codificações de redes neurais como restrições MILP que possuem, em relação ao problema de verificar a robustez de uma ANN para exemplos adversariais, maior eficiência na literatura. Por exemplo, a abordagem apresentada no trabalho de Tjeng, Xiao e Tedrake (2019), que aponta, nos seus resultados, a sua superioridade em relação à codificação usada por Fischetti e Jo (2018) e Katz et al. (2017) (que era o antigo estado da arte).

Além disso, é possível utilizar esse algoritmo proposto por Ignatiev, Narodytska e Marques-Silva (2019a) para validar, reparar (caso sejam otimistas) e refinar (caso sejam pessimistas) as explicações de modelos de aprendizagem de máquina computadas por métodos heurísticos, como evidenciado em Ignatiev, Narodytska e Marques-Silva (2019b). Porém, os autores realizaram essas operações apenas em modelos de *boosted trees*, e também haveria o problema da escalabilidade ao utilizar redes neurais artificiais com várias camadas, pelo fato de ser o mesmo método.

Fundamentado na importância de disponibilizar explicações decorrentes de algoritmos que carregam garantias formais, este trabalho possui o objetivo de melhorar,

em termos de escalabilidade, a abordagem do artigo de Ignatiev, Narodytska e Marques-Silva (2019a) a partir do uso da codificação das ANNs proposta por Tjeng, Xiao e Tedrake (2019), utilizando o raciocínio abdutivo para a aquisição de explicações mínimas para redes neurais artificiais. Além disso, este trabalho também almeja utilizar essa abordagem implementada para a validação, reparação e refinamento de explicações heurísticas para esses modelos.

Assim, os objetivos específicos do trabalho proposto seguem a listagem abaixo:

- a) Melhorar, em termos de escalabilidade, a abordagem original de Ignatiev, Narodytska e Marques-Silva (2019a) substituindo a codificação das redes neurais como restrições MILP utilizada pela proposta em Tjeng, Xiao e Tedrake (2019);
- b) Comparar o método original com a versão implementada com o uso de redes neurais artificiais com diferentes quantidades de camadas ocultas e de neurônios;
- c) Validar, reparar e refinar explicações heurísticas computadas para redes neurais artificiais com diferentes quantidades de camadas ocultas e de neurônios com o uso da versão implementada.

Dessa forma, a hipótese deste trabalho é que ao trocar o codificador de redes neurais artificiais como restrições MILP de Fischetti e Jo (2018) pela proposta em Tjeng, Xiao e Tedrake (2019), espera-se uma melhora na escalabilidade do algoritmo de Ignatiev, Narodytska e Marques-Silva (2019a) para a computação de explicações mínimas para ANNs com o uso do raciocínio abdutivo, também possibilitando a validação, reparação e refinamento de explicações heurísticas de redes neurais mais profundas e com mais neurônios, pelo fato dessas três operações propostas por Ignatiev, Narodytska e Marques-Silva (2019b) utilizarem esse algoritmo como base.

Essa expectativa na melhora da escalabilidade do algoritmo ao trocar os codificadores é justificada nos resultados das experimentações realizadas por Tjeng, Xiao e Tedrake (2019), que apontam a sua superioridade, em termos de tempo de resolução de um problema MILP, em relação ao antigo estado da arte (KATZ et al., 2017), e por consequência também à abordagem de Fischetti e Jo (2018), para o problema de verificação de robustez de ANNs para exemplos adversariais. Logo, apesar de ser tratado outro problema neste trabalho (computação de explicações para redes neurais), provavelmente a codificação de Tjeng, Xiao e Tedrake (2019) também pode oferecer esse ganho para o algoritmo proposto, aumentando a sua escalabilidade.

## 1.1 Resumo dos artigos selecionados

O trabalho de Ignatiev, Narodytska e Marques-Silva (2019a) apresenta um algoritmo *model-agnostic*, significando que funciona em qualquer modelo desde que ele

possa ser representado por um sistema de raciocínio de restrição e que consultas de implicação possam ser decididas por um oráculo. Essa abordagem utiliza o raciocínio abdutivo para computar explicações mínimas para modelos de aprendizagem de máquina com garantias formais, fornecendo explicações de cardinalidade mínima ou sub-conjunto mínima. A abordagem basicamente codifica um modelo  $M$  de aprendizagem de máquina como um conjunto de restrições lineares  $F$  em alguma teoria, a partir de um algoritmo 0-1 MILP (FISCHETTI; JO, 2018), então, dado uma instância  $C$  associado com a predição  $E$  tal que  $C \wedge F \models E$  (equivalente à  $C \models (F \rightarrow E)$ ), é computada a explicação mínima  $C'$  (dado  $C$ ) que é implicante principal de  $F \rightarrow E$ .

O artigo de Fischetti e Jo (2018) propõe um algoritmo que realiza a codificação de uma rede neural profunda como um 0-1 MILP que usa restrições de indicação para evitar o uso da notação de *Big-M*, utilizando variáveis de ativação binárias para impor as implicações lógicas. Foi realizado a modelagem de aplicações que usavam redes com ReLUs e *max/average pooling*. Esse modelo codificado em restrições lineares não é suscetível à treinamento, pois ele se torna bilinear nessa configuração. Foi realizado experimentos do algoritmo em dois problemas de aprendizagem de máquina: visualização de características e aprendizagem de máquina adversário.

O trabalho de Tjeng, Xiao e Tedrake (2019) implementa um algoritmo MILP que codifica as partes lineares (camadas que usam transformações lineares ou funções que possuam partes lineares, como ReLU e *max pooling*) de uma rede neural como restrições lineares, minimizando o número de variáveis binárias presentes no problema MILP e melhorando o condicionamento numérico. Seja a função de ativação ReLU  $y = \max(x, 0)$  e  $l \leq x \leq u$ , para a sua formulação é considerado que há 3 fases: a unidade está estavelmente inativa (quando  $u \leq 0$ ), estavelmente ativa (quando  $l \geq 0$ ) e instável (caso contrário). Para unidades instáveis é utilizado o conjunto de restrições lineares e inteiras na Equação 1.1, com a introdução de uma variável de indicação de decisão  $a = 1_{x \geq 0}$ .

$$(y \leq x - l(1 - a)) \wedge (y \geq x) \wedge (y \leq u * a) \wedge (y \geq 0) \wedge (a \in \{0, 1\}) \quad (1.1)$$

Para indicar se uma unidade ReLU é instável ou não, é preciso definir os seus limites de entrada o mais achatado possível, porém ainda sendo válidos. Para isso, são utilizados dois procedimentos: Intervalo Aritmético e Programação Linear. Essa abordagem foi realizada na aplicação de exemplos adversariais sendo de duas a três vezes de ordem de magnitude mais rápida que o estado da arte, permitindo um aumento significativo do tamanho das redes neurais codificadas.

O artigo de Ignatiev, Narodytska e Marques-Silva (2019b) descreve os expe-



rimentos realizados que questionam as explicações de modelos de aprendizagem de máquina dadas por métodos heurísticos, que computam uma explicação explorando localmente o espaço da instância. Para realização desses experimentos é utilizado abordagens baseadas em lógica com o cálculo de implicantes principais por meio do raciocínio abdutivo, realizando os testes em *boosted trees*. Os algoritmos desenvolvidos possuem os objetivos de acessar a qualidade das explicações locais, reparar as explicações locais que são otimistas (quando existem instâncias no espaço de instâncias que a explicação computada falha) e refinar as explicações locais que são pessimistas (quando alguma literal pertencente à explicação computada é irrelevante e pode ser descartada).

O trabalho de Katz et al. (2017) implementa o algoritmo Reluplex, um SMT *solver* para a teoria da aritmética linear real que tem o objetivo de verificar redes neurais profundas utilizando uma técnica baseada no Simplex, porém adaptada para lidar com a função de ativação não-convexa ReLU sem simplificações no seu funcionamento original, apenas permitindo que suas entradas e saídas sejam temporariamente inconsistentes e corrigidas conforme a execução do algoritmo. O Reluplex foi avaliado em 45 redes neurais profundas desenvolvidas como um protótipo inicial do sistema anti-colisão de última geração para aeronaves não tripuladas ACAS Xu.

Este trabalho é fundamentado no artigo de Ignatiev, Narodytska e Marques-Silva (2019a), que utilizada um 0-1 MILP (FISCHETTI; JO, 2018) para a codificação da rede neural como um conjunto de restrições lineares em uma teoria, então são computadas explicações para os seus resultados com o uso do raciocínio abdutivo. A necessidade de tais explicações computadas carregando uma garantia formal é evidenciada em Ignatiev, Narodytska e Marques-Silva (2019b). Um dos objetivos iniciais é comparar os possíveis resultados desse algoritmo utilizando diferentes MILPs e estratégias de codificação (verificação) das redes neurais, como os implementados em Tjeng, Xiao e Tedrake (2019) e Katz et al. (2017).

## 2 MOTIVAÇÃO INICIAL E JUSTIFICATIVA FUNDAMENTADA DO ESTUDO

Apresente o que motivou o estudo e a justificativa da relevância e necessidade do estudo. O Quadro 1 é uma exemplo de quadro.

Quadro 1 – LIVROS ANALISADOS

Referência para citar no texto	Título do livro	Autor/Autores
Livro 1	Matemática Completa	Bonjorno, Giovanni Jr e Paulo Câmara
Livro 2	Matemática: Contexto e Aplicações	Luiz Roberto Dante
Livro 3	Matemática	Emanuel Paiva
Livro 4	Matemática: Ciência e Aplicações	Gelson Iezzi, Osvaldo Dulce, David Degenszajn, Roberto Périco e Nilse de Almeida
Livro 5	Matemática para compreender o mundo	Kátia Stocco Smole e Maria Ignez Diniz
Livro 6	Fundamentos de Matemática Elementar	Gelson Iezzi e Carlos Marukami

Fonte: Elaborado pela autor(a).

Um exemplo de lista de itens.

Livro 1 - Neste livro, ... Ao final das seções percebemos razoável variação de exercícios resolvidos e propostos.

Livro 2 - No segundo livro, ...

Livro 3 - Nesse exemplar ...

Livro 4 - ??), ...

Livro 5 - As autoras abordam ....

Livro 6 - Neste livro, ...

### 3 FUNDAMENTAÇÃO TEÓRICA

Apresente um resumo das teorias utilizadas de forma a facilitar o acesso ao leitor do trabalho aos pré-requisitos para o entendimento do trabalho.

#### 3.1 Teoria dos Registros das Representações Semióticas

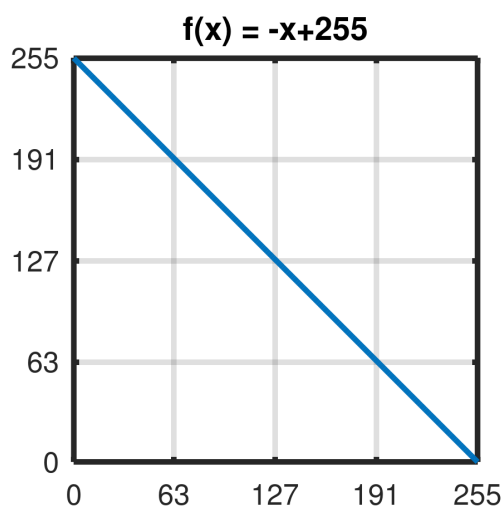
Ensinar é uma tarefa ... O psicólogo e filósofo Raymond Duval, desenvolveu a Teoria dos Registros de Representação Semiótica - (TRRS)...

A *conversão* para ??) é (exemplo de citação):

[...] a conversão de uma representação se refere às operações em que o registro inicial é transformado em outro registro; por essa razão, é considerada como uma “transformação externa”. Por exemplo, ao utilizarmos a linguagem algébrica para representar a frase “o dobro de um número resulta em oito”, estamos realizando uma conversão do registro dado na língua natural para o registro dado na linguagem algébrica (??, p. 6).

Exemplo de figura.

Figura 1 – Representação gráfica da função afim



Fonte: Elaborada pelo autor

Assim, a Figura 1..

### 3.2 Exemplo de seção.

??, p. 81), define Função como:

**Definição 3.1** *Dado dois conjuntos  $A$  e  $B$ , não vazios, uma relação  $f$  de  $A$  em  $B$  recebe o nome de aplicação de  $A$  em  $B$  ou função definida em  $A$  com imagens em  $B$  se, e somente se, para todo  $x \in A$  existe um só  $y \in B$  tal que  $(x, y) \in f$ .*

$$f \text{ aplicado de } A \text{ em } B \iff (\forall x \in A, \exists |y \in B| (x, y) \in f) \quad (3.1)$$

#### 3.2.1 Exemplo de subseção

??, p. 100), define função afim como:

**Definição 3.2** *Uma aplicação de  $\mathbb{R}$  em  $\mathbb{R}$  com  $a \neq 0$  e cada  $x \in \mathbb{R}$  associa o elemento  $(a \cdot x + b) \in \mathbb{R}$ .*

$$f(x) = a \cdot x + b \quad \text{com} \quad (a \neq 0) \quad (3.2)$$

#### 3.2.2 Imagem da função afim

??, p. 105) diz que:

reta permitindo a análise de que todos os valores de  $y$  estão relacionados com  $x$ .

#### 3.2.3 Zero da função

Vejamos que  $f(x)$  é crescente pois na medida que os valores em  $x$  vão aumentando, as suas respectivas imagens também crescem.

#### 3.2.4 Exemplo de subseção

??, p. 118), resume em:

A função afim:

$$f(x) = a \cdot x + b \text{ anula-se para } x = -\frac{b}{a}. \quad (3.3)$$

Para  $x > -\frac{b}{a}$ , temos:

$$\begin{cases} \text{se } a > 0 \text{ então } f(x) = a \cdot x + b > 0 \\ \text{se } a < 0 \text{ então } f(x) = a \cdot x + b < 0 \end{cases} \quad (3.4)$$

Isto é,  $x > -\frac{b}{a}$  a função  $f(x) = a \cdot x + b$  tem sinal de  $a$ .

Para  $x < -\frac{b}{a}$ , temos:

$$\begin{cases} \text{se } a > 0 \text{ então } f(x) = a \cdot x + b < 0 \\ \text{se } a < 0 \text{ então } f(x) = a \cdot x + b > 0 \end{cases} \quad (3.5)$$

Isto é, para  $x < -\frac{b}{a}$  a função  $f(x) = a \cdot x + b$  tem o sinal de ‘-a’ (sinal contrário ao de  $a$ ).

Exemplo de função definida por partes:

**Exemplo 3.1** Seja  $f : \mathbb{R} \rightarrow \mathbb{R}$  definida por:

$$f(x) = \begin{cases} x & \text{se } 0 \leq x \leq 128 \\ 128 & \text{se } 128 \leq x \leq 256 \\ x - 128 & \text{se } c.c \end{cases} \quad (3.6)$$

### 3.3 Outra seção

Exemplo de código.

#### Código 1 — Método da Bissecção

```
function xm=mb(f,xp,xn) % metodo da bissecao para zero de funcoes
xm=(xp+xn)/2;
y=f(xm);
while(abs(y)>0.01) % enquanto |y|>ep -> laço de repeticao
    if(y>0) % se y maior que 0
        xp=xm;
    else % senao
        xn=xm;
    end
    xm=(xp+xn)/2;
    y=f(xm);
end
end
```

Fonte: Elaborada pela autor(a) (GNU Octave)

#### **4 ABORDAGEM AO PROBLEMA**

Apresente a sua proposta de abordagem ao problema ou discussão da questão do trabalho monográfico.

## **5 CONSIDERAÇÕES FINAIS**

Apresente suas considerações finais.

## **ANEXO A    Apêndice**



**ANEXO A.1 Texto auxiliar do trabalho**

O apêndice deve ser autoral, textos externos devem ser colocados como anexo.

## REFERÊNCIAS

COMISSÃO EUROPEIA. *Reforma de 2018 das regras de proteção de dados da UE*. 2018. Disponível em: <[https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes\\_en.pdf](https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf)>.

FISCHETTI, M.; JO, J. Deep neural networks and mixed integer linear optimization. *Constraints*, v. 23, p. 296–309, Jul. 2018. Disponível em: <<https://link.springer.com/article/10.1007/s10601-018-9285-6>>.

GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. *Explaining and Harnessing Adversarial Examples*. 2015. Disponível em: <<https://arxiv.org/abs/1412.6572>>.

IGNATIEV, A.; NARODYTSKA, N.; MARQUES-SILVA, J. Abduction-based explanations for machine learning models. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 33, n. 01, p. 1511–1519, Jul. 2019a. Disponível em: <<https://ojs.aaai.org/index.php/AAAI/article/view/3964>>.

IGNATIEV, A.; NARODYTSKA, N.; MARQUES-SILVA, J. On validating, repairing and refining heuristic ML explanations. *CoRR*, abs/1907.02509, 2019b. Disponível em: <<http://arxiv.org/abs/1907.02509>>.

KATZ, G. et al. Reluplex: An efficient SMT solver for verifying deep neural networks. In: *Computer Aided Verification*. Springer International Publishing, 2017. p. 97–117. Disponível em: <[https://doi.org/10.1007/978-3-319-63387-9\\_5](https://doi.org/10.1007/978-3-319-63387-9_5)>.

LIU, W. et al. A survey of deep neural network architectures and their applications. *Neurocomputing*, v. 234, p. 11–26, 2017. ISSN 0925-2312. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0925231216315533>>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144. ISBN 9781450342322. Disponível em: <<https://doi.org/10.1145/2939672.2939778>>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Anchors: High-precision model-agnostic explanations. In: MCILRAITH, S. A.; WEINBERGER, K. Q. (Ed.). *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the*

*30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018.* AAAI Press, 2018. p. 1527–1535. Disponível em: <<https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982>>.

TJENG, V.; XIAO, K. Y.; TEDRAKE, R. Evaluating robustness of neural networks with mixed integer programming. In: *International Conference on Learning Representations*. [s.n.], 2019. Disponível em: <<https://openreview.net/forum?id=HyGIIdRqtm>>.