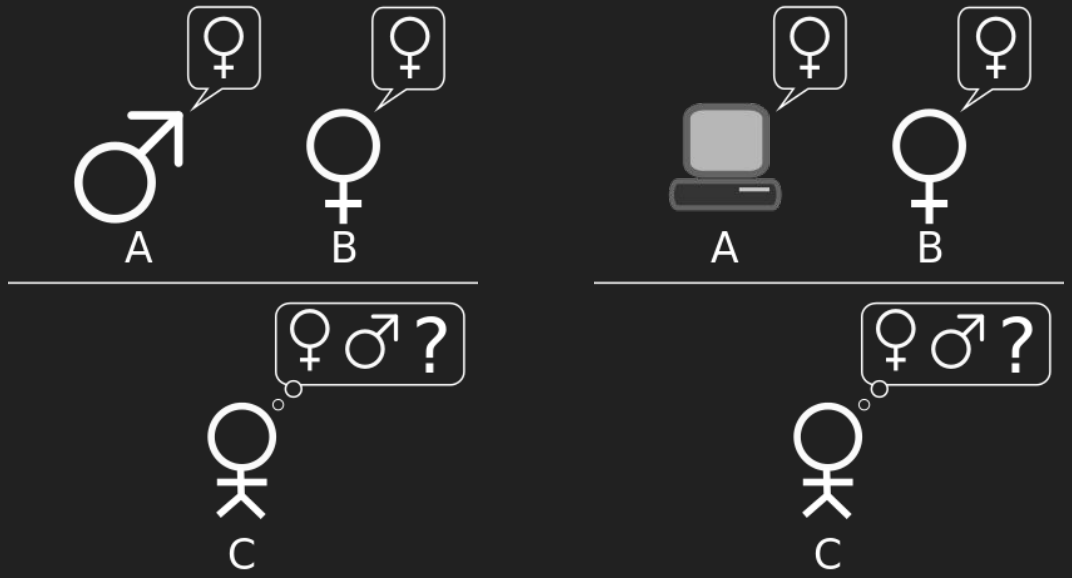
Segurança de *HIP*s além da obscuridade:

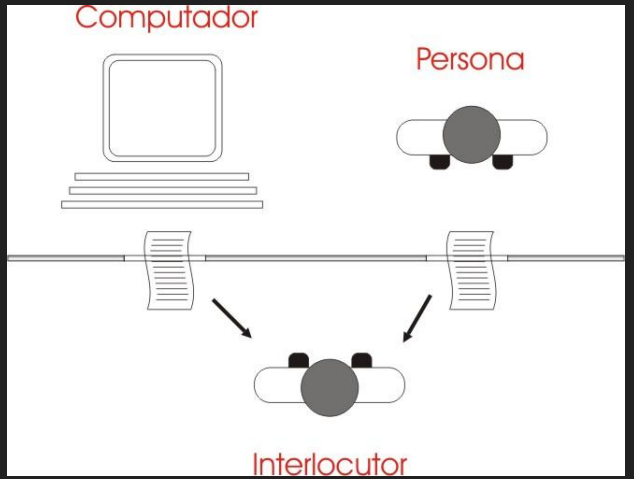
*CAPTCHA*s, de volta ao básico

Por **Carybé Gonçalves Silva** sob supervisão de **Ronaldo Fumio Hashimoto**

Can machines think?

O Teste de Turing e o Jogo da Imitação





Fonte: Wikimedia Commons

O Teste de Turing e o Jogo da Imitação

Fonte: HBO's "WestWorld" (2016)

Histórico da Inteligência Artificial

de “*estupidez artificial*” a *sistemas especialistas*

1966 — Atualidade

Fonte: Wikimedia Commons

Fonte: Reddit

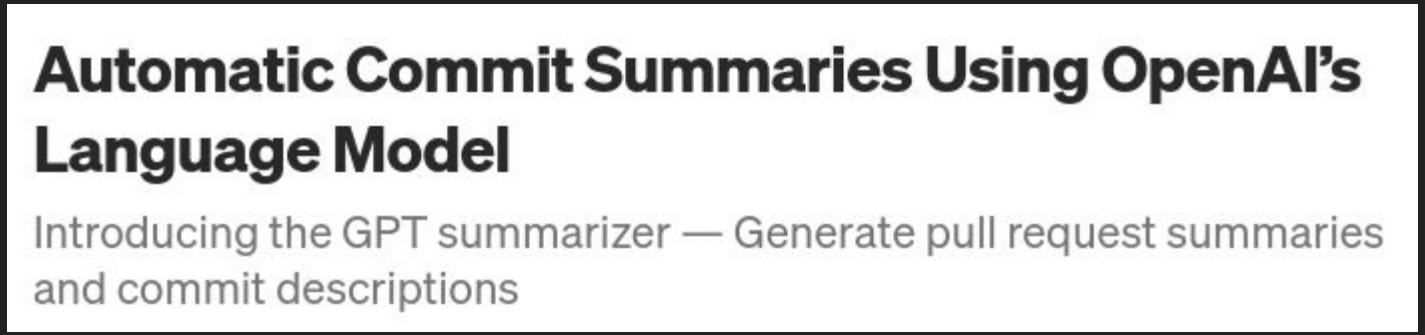
Using GPT-3 to explain how code works

One of my favourite uses for the GPT-3 AI language model is generating explanations of how code works. It's shockingly effective at this: its training set clearly include a vast amount of source code.

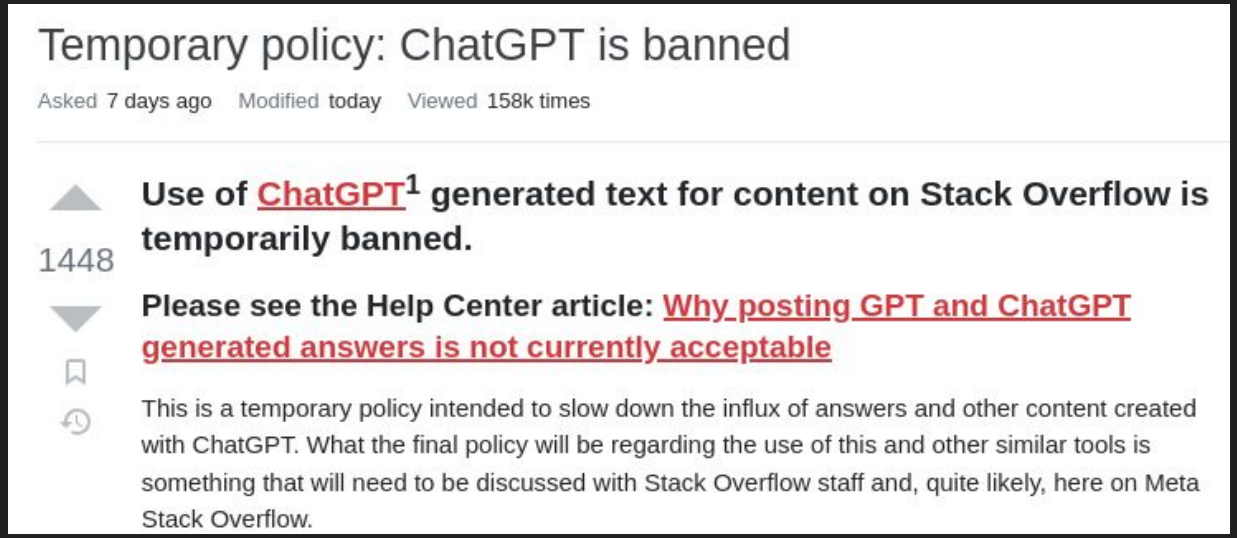
(I initially thought this was related to GitHub Copilot being built on GPT-3, but actually that's built on a GPT-3 descendent called OpenAI Codex.)

Here are a few recent examples.

Fonte: Simon Willison's Weblog

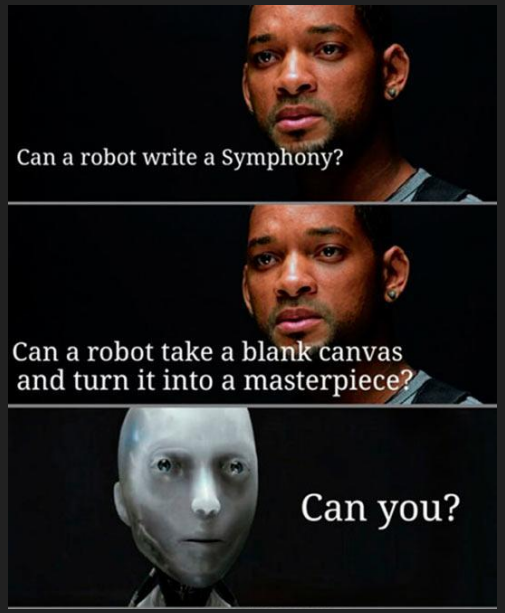


Fonte: Better Progamming



Fonte: Stack Overflow

Quais são as capacidades *estritamente* humanas?



Fonte: “I, Robot” (2004)

StableDiffusion prompt: *Film still of blade runner interrogation scene*

O Teste de Turing Reverso

e o desafio de se automatizar um comprovante de interação humana

HIP - *Human Interaction Proof*

Fonte: *AltaVista (1997)*

Fonte: *Yahoo EZ-Gimpy (2000)*

O Teste de Turing Reverso

e o desafio de se automatizar um comprovante de interação humana

CAPTCHA - *Completely Automated Public Turing test to tell Computers and Humans Apart*
Luis von Ahn (2003)

Fonte: *reCAPTCHA (2008)*

“A captcha is a cryptographic protocol whose underlying hardness assumption is based on an AI problem.”

von Ahn (2003)

*HIP*s como sistemas de segurança

- Protegem dados e recursos sensíveis a acessos automatizados

- Diferentemente de sistemas criptográficos, há uma **tolerância de acerto do *atacante***, assim como uma **tolerância de erro do *usuário*** (desafio da Acessibilidade)

*“We do not allow captchas to base their security in the secrecy of a **database or a piece of code**”*

von Ahn (2003)

Segurança por obscuridade de Kerckhoffs a Shannon

1. The system must be practically, if not mathematically, indecipherable;
2. **It should not require secrecy, and it should not be a problem if it falls into enemy hands;**
3. It must be possible to communicate and remember the key without using written notes, and correspondents must be able to change or modify it at will;
4. It must be applicable to telegraph communications;
5. It must be portable, and should not require several persons to handle or operate;
6. Lastly, given the circumstances in which it is to be used, the system must be easy to use and should not be stressful to use or require its users to know and comply with a long list of rules.

Kerckhoffs (1883)

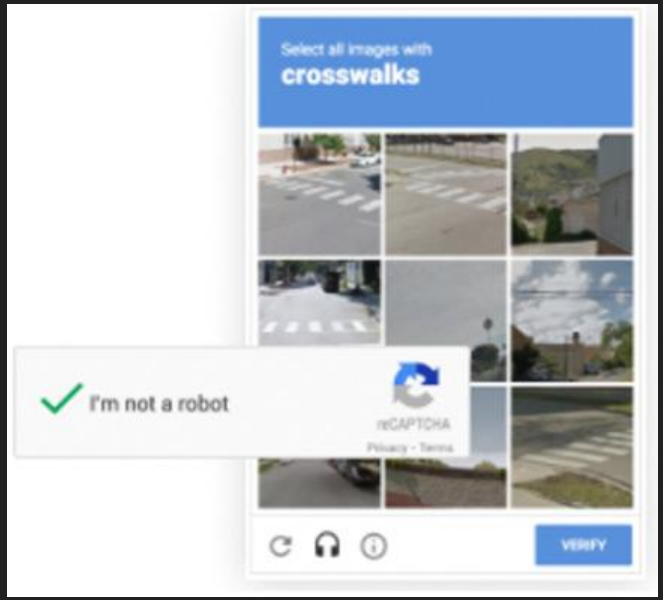
"The enemy knows the system"
Shannon (1949)

Evolução dos *HIP*s

Fonte: *AltaVista (1997)*

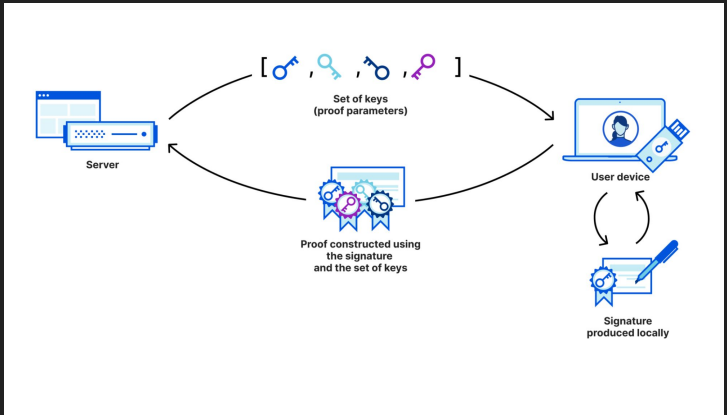
Fonte: *reCAPTCHA (2008)*

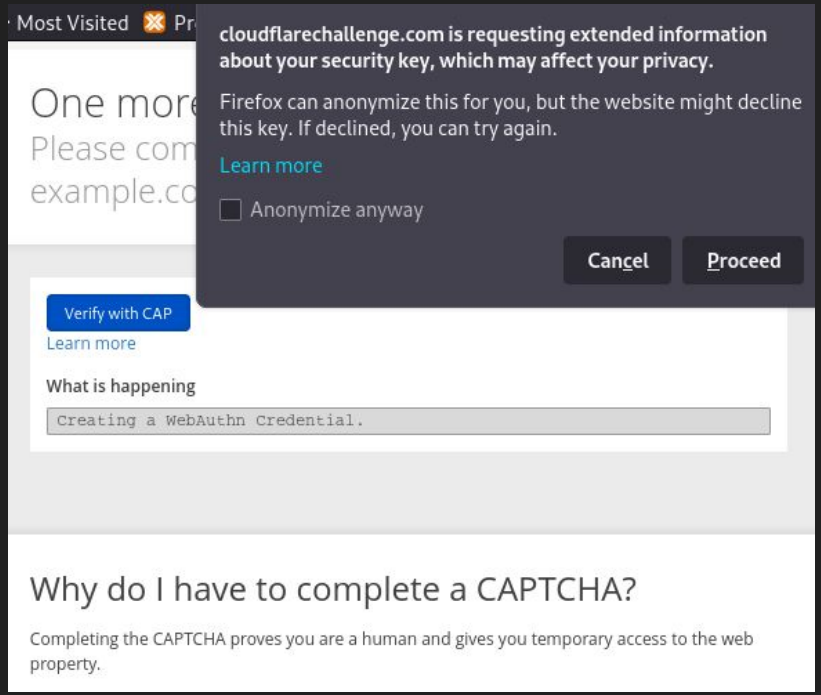
Fonte: Google's *reCAPTCHA (2012)*



Fonte: Google's *reCAPTCHAv2 (2014)*

Evolução dos *HIP*s





Fonte: Cloudflare *(2021)*

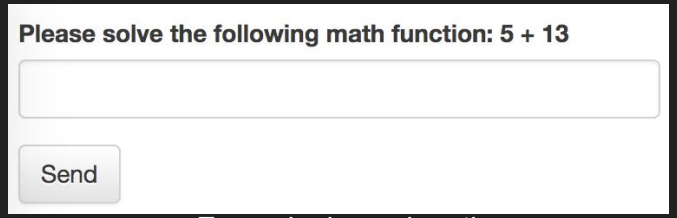
Taxonomia dos HIPs

Textual

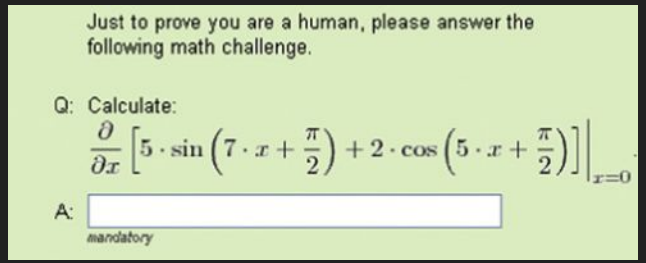
Exemplo: captcha.py

Visual
Generativo
Independente de dados*
Média acessibilidade

Lógico



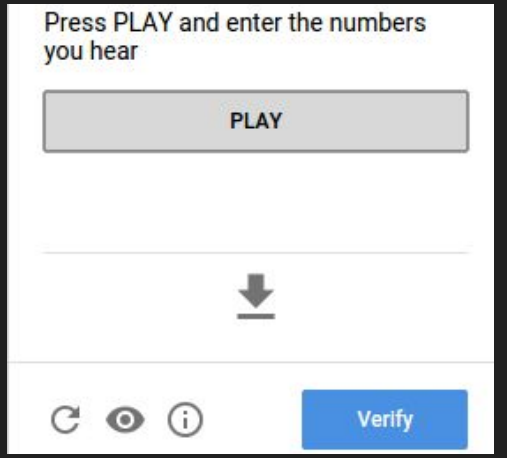
Exemplo: laravel math



Exemplo: QRBGS

Visual
Generativo
Independente de dados
Média acessibilidade*

Auditivo



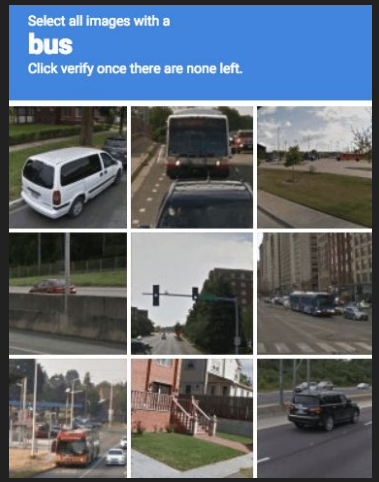
Exemplo: reCAPTCHAv2 áudio

Sonoro
Generativo
Dependente de dados*
Média acessibilidade*

***Com ressalvas**

Taxonomia dos HIPs

Imagético de Seleção



Exemplo: reCAPTCHAv2

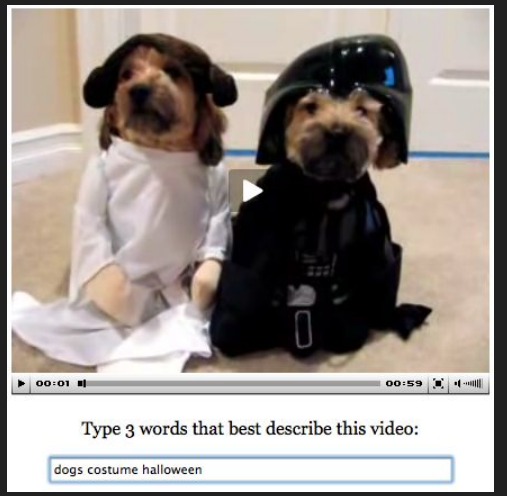
Visual
Não-Generativo
Dependente de dados
Média acessibilidade

Imagético de Transformação

Exemplo: FunCaptcha

Visual
Generativo
Dependente de dados
Média acessibilidade*

Vídeo



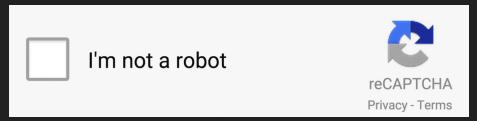
Exemplo: Video Captcha

Visual
Não-Generativo
Dependente de dados
Baixa acessibilidade

***Com ressalvas**

Taxonomia dos HIPs

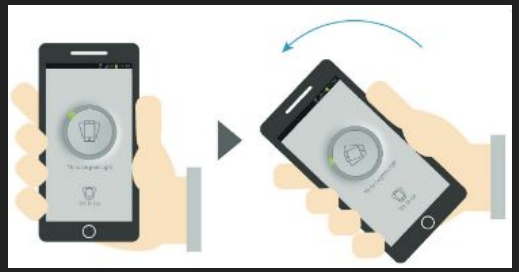
Comportamental



Exemplo: reCAPTCHAv2 NoCAPTCHA

Telemétrico
Não-Generativo
Dependente de dados
Alta acessibilidade

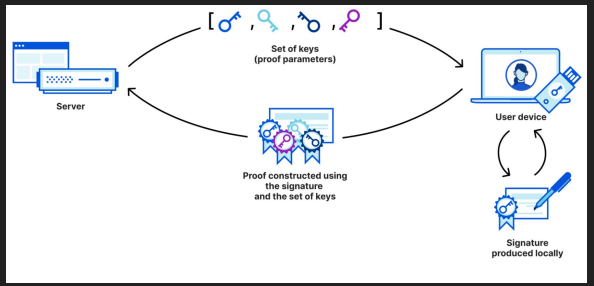
Sensorial



Exemplo:CAPPCHA

Telemétrico
Não-Generativo
Dependente de dados
Média acessibilidade*

Token



Exemplo:Turnstile

Criptográfico
Não-Generativo
Dependente de entidade certificadora
Alta acessibilidade*

***Com ressalvas**

Estudo de caso: *CAPTCHA* textual

Modelo mais ubíquo entre as diferentes categorias:

muitos modelos abertos permitindo uma fácil geração de conjuntos de dados

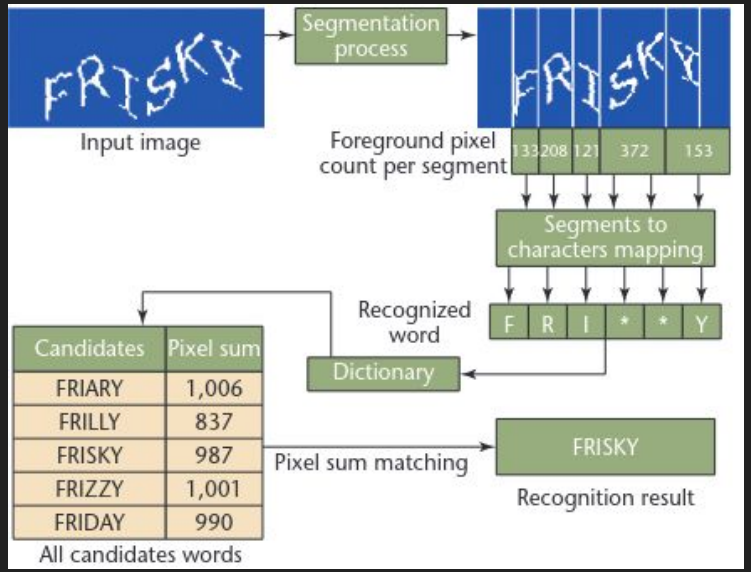
Não-dependente de dados:

segurança não-baseada em obscuridade

Biblioteca: ***captcha.py*** implementa: aleatorização de fonte, distorção e rotação independete de caracteres, coloração aleatória do fundo e do texto, ruído, linhas anti-segmentação, sobreposição de caracteres e suavização de bordas

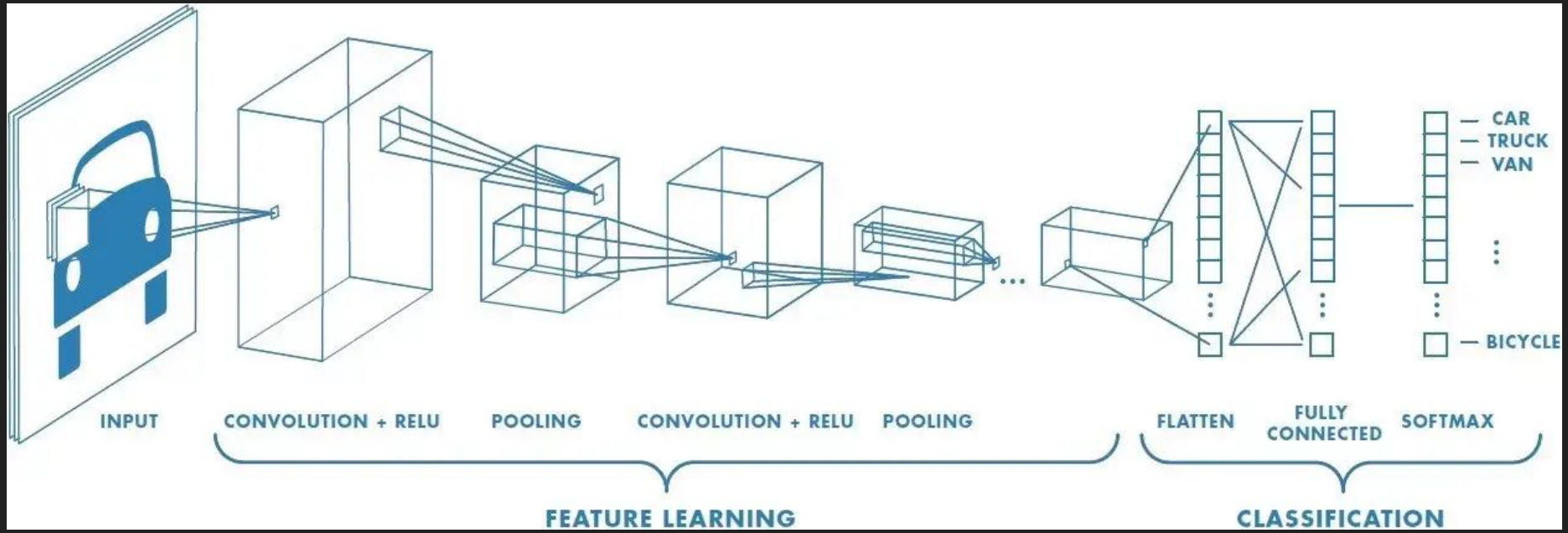
CAPTCHA textual: ataque clássico

Baseado em OCR (*Optical character recognition)*, segmentação e processamento de imagens em uma estrutura de *pipeline*



Fonte: *CAPTCHA Security: a case study*

CAPTCHA textual: ataque ingênuo usando CNNs



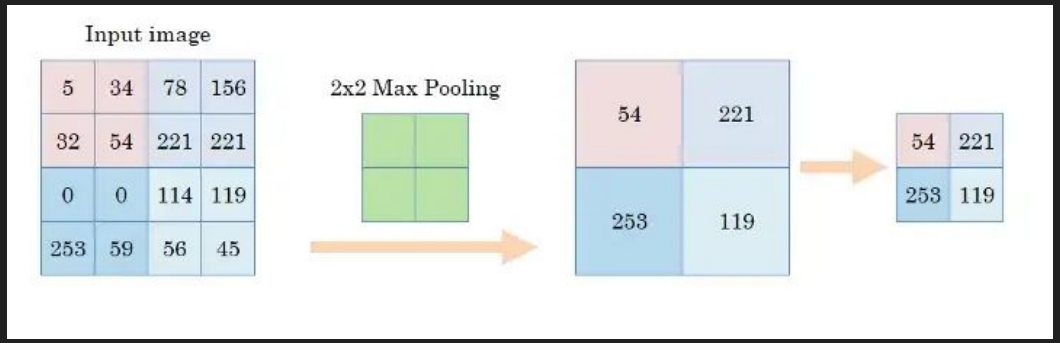
Fonte: *towards data science*

CAPTCHA textual: ataque ingênuo usando CNNs

Convolução

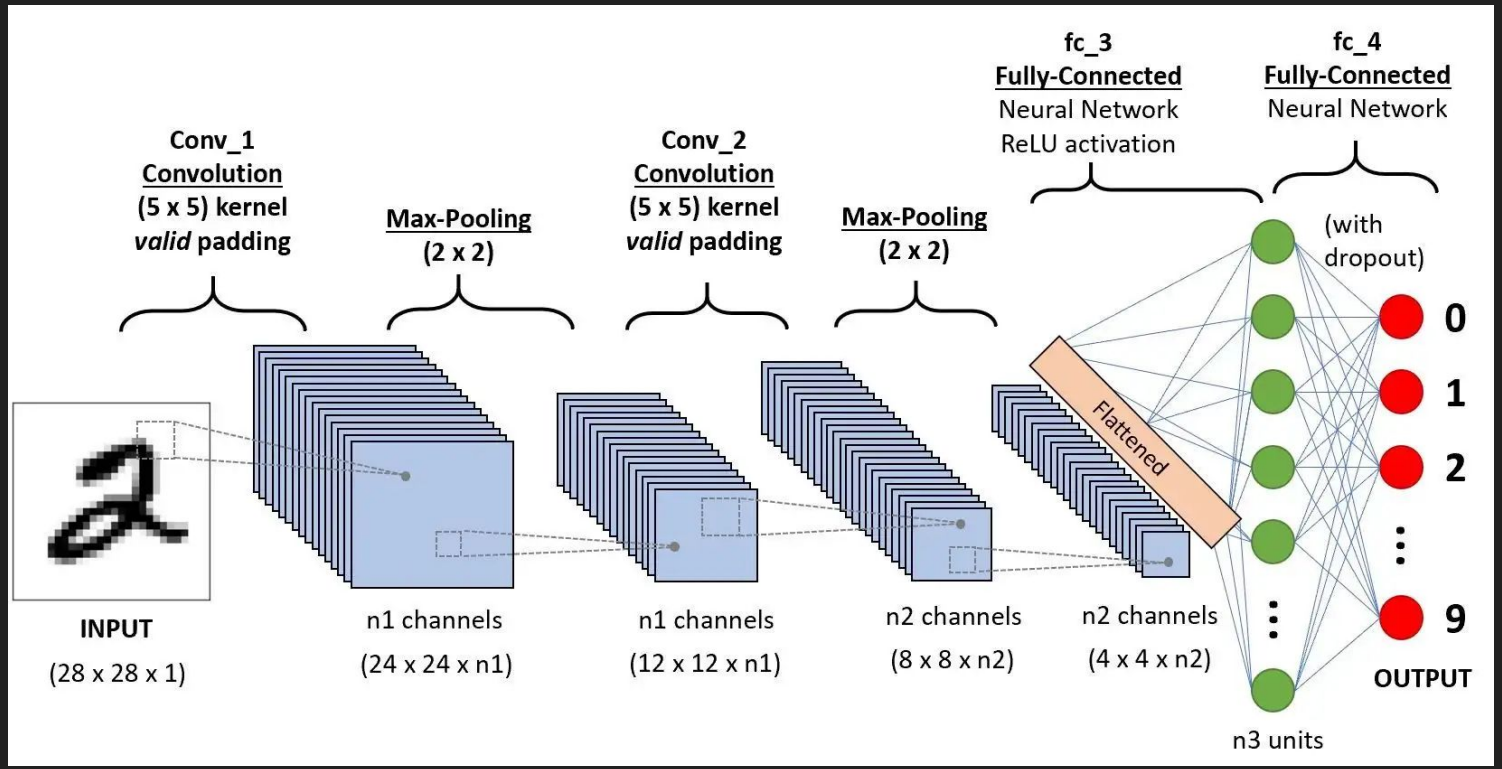
Fonte: *Convolution Arithmetic*

Pooling



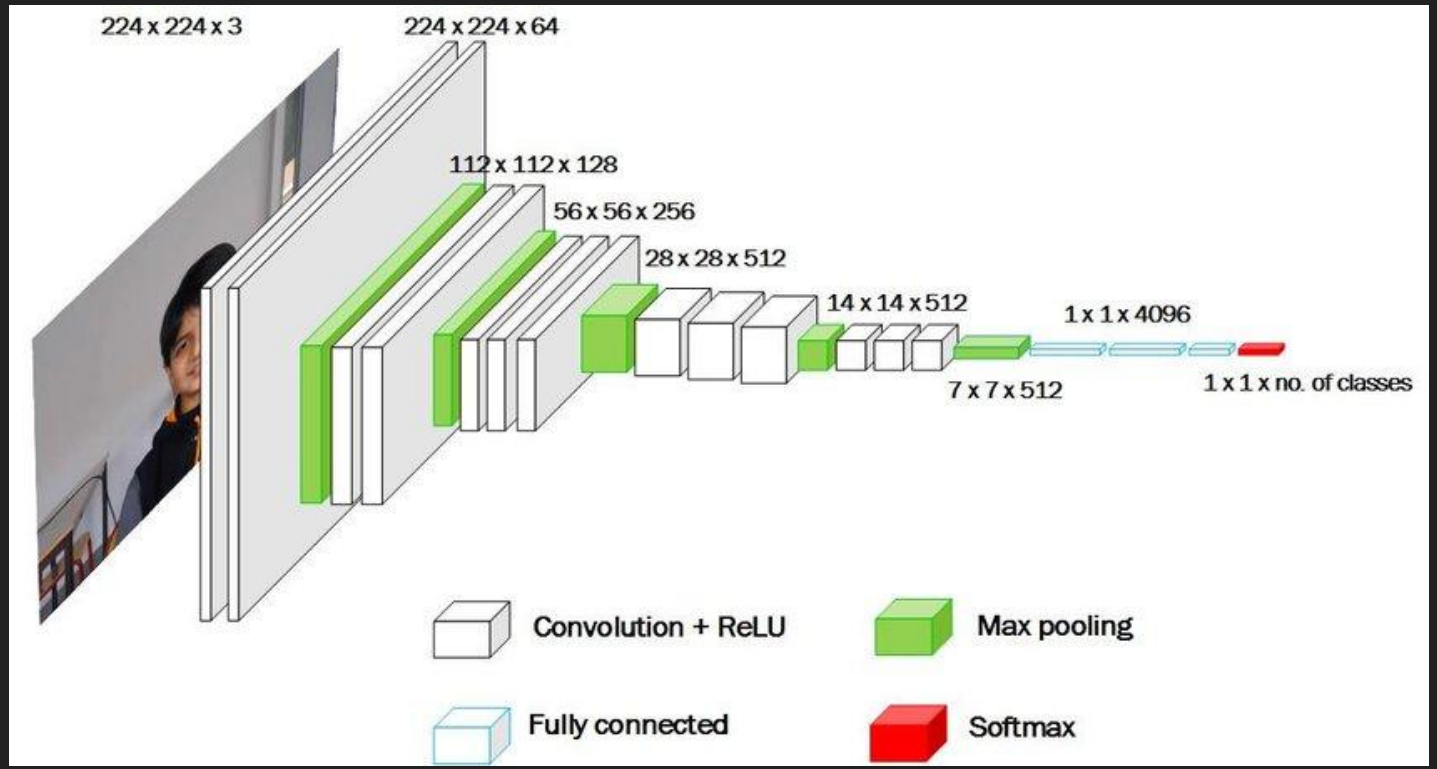
Fonte: *towards data science*

CAPTCHA textual: ataque ingênuo usando CNNs



Fonte: *towards data science*

CAPTCHA textual: ataque ingênuo usando CNNs



Estrutura VGG16 utilizada
Fonte: *LIRIS-CSE*

Resultados e generalização dos modelos

Um modelo mais genérico é capaz de atacar de forma eficiente um maior número de tipos de captchas textuais, por isso variou-se em 3 modelos disponíveis pela biblioteca:

1. Wheezy (depreciado sem aleatorização de algumas distorções)
2. Default (padrão da biblioteca)
3. Fonts (com aleatorização de um conjunto maior de fontes)

Além disso, variou-se os alfabetos e a extensão dos textos em (letras minúsculas (**a**), letras maiúsculas (**A**) e números (**n**)) e (2, 4, 8 e 16 caracteres), respectivamente

Resultados e generalização dos modelos

MODEL\ DATASET	w_A	w_n	w_a	w_An	w_Aa	w_na	w_Ana	d_A	d_n	d_a	d_An	d_Aa	d_na	d_Ana	f_A	f_n	f_a	f_An	f_Aa	f_na	f_Ana
w_A	99.9%			57.3%	27.1%		20.6%														
w_n		100.0%																			
w_a			99.9%		26.9%	56.9%	20.9%														
w_An	99.6%	99.3%		99.5%	26.3%		34.7%														
w_Aa	99.1%		99.5%	52.4%	99.3%	52.3%	70.2%														
w_na		99.9%	99.9%		25.2%	99.9%	33.9%														
w_Ana	98.7%	99.1%	99.5%	98.8%	99.1%	99.4%	99.1%														
d_A	34.1%			20.7%	9.8%			99.8%			56.9%	27.5%		21.2%	69.9%			40.4%	19.5%		
d_n		28.8%							100.0%							91.1%					
d_a										99.9%		26.8%	58.8%	21.1%			73.9%		20.2%	43.4%	16.2%
d_An								98.3%	93.7%		96.9%	24.9%		33.2%	62.1%	67.0%		62.9%			
d_Aa								96.3%		96.7%	54.9%	96.5%	55.0%	70.9%	56.6%		51.1%	31.8%	52.8%	28.5%	38.8%
d_na									99.2%	99.7%		24.7%	99.6%	33.3%		78.9%	66.4%			70.0%	
d_Ana								94.7%	95.2%	96.4%	94.8%	95.6%	95.9%	95.5%	52.4%	65.0%	49.4%	55.7%	50.3%	53.0%	52.5%
f_A	43.5%			24.6%				99.2%			56.9%	27.4%			99.0%			56.7%			
f_n									99.8%							99.9%					
f_a			23.1%			14.2%				99.7%		26.8%	59.8%	21.4%			98.9%		26.8%	57.9%	21.1%
f_An		33.0%		22.9%				94.3%	95.6%		94.4%			32.8%	96.8%	95.4%		96.6%			34.4%
f_Aa								81.8%		96.7%	46.3%	88.9%	54.4%	65.1%	88.6%		92.6%	48.7%	90.5%	50.3%	65.5%
f_na									95.0%	98.1%		24.3%	97.1%	32.6%		96.2%	96.2%		24.2%	96.3%	32.3%
f_Ana								84.2%	89.6%	93.9%	85.5%	88.6%	92.5%	88.9%	89.8%	91.5%	89.4%	90.7%	89.5%	90.0%	89.9%

Modelos treinados com 1 milhão de captchas de 2 caracteres de extensão

Resultados e generalização dos modelos

MODEL\ DATASET	w_A	w_n	w_a	w_An	w_Aa	w_na	w_Ana	d_A	d_n	d_a	d_An	d_Aa	d_na	d_Ana	f_A	f_n	f_a	f_An	f_Aa	f_na	f_Ana
w_A	99.3%			32.6%																	
w_n		99.9%																			
w_a			99.8%			32.6%															
w_An	98.8%	99.3%		98.9%																	
w_Aa	97.6%		99.2%		98.6%																
w_na		99.7%	99.7%			99.7%															
w_Ana	97.1%	99.0%	99.2%	97.6%	98.2%	99.1%	98.5%														
d_A								99.2%			32.2%				42.9%						
d_n									99.9%							84.1%					
d_a										99.6%			33.9%				49.3%				
d_An								95.8%	93.3%		95.2%				33.7%	54.5%		38.5%			
d_Aa								93.2%		93.3%		93.4%	28.5%	49.7%	28.9%		22.2%		24.3%		
d_na									98.8%	99.1%			99.1%			63.1%	38.8%			44.8%	
d_Ana								89.4%	90.7%	93.1%	90.2%	91.2%	92.3%	91.3%	25.9%	46.6%	21.8%	30.1%	22.8%	27.3%	25.9%
f_A								98.0%			32.2%				98.0%						
f_n									99.6%							99.8%					
f_a										98.7%			34.0%				96.9%				
f_An								89.0%	90.7%		88.9%				93.9%	90.7%		93.2%			
f_Aa								71.7%		89.8%		80.2%	27.9%	42.6%	83.1%		82.6%		83.2%		
f_na									85.6%	96.5%			93.9%			91.6%	93.1%			92.8%	
f_Ana								71.9%	82.7%	85.4%	74.6%	78.2%	84.0%	78.7%	81.3%	86.5%	77.4%	83.2%	79.8%	80.0%	80.9%

Modelos treinados com 1 milhão de captchas de 4 caracteres de extensão

Resultados e generalização dos modelos

MODEL\ DATASET	w_A	w_n	w_a	w_An	w_Aa	w_na	w_Ana	d_A	d_n	d_a	d_An	d_Aa	d_na	d_Ana	f_A	f_n	f_a	f_An	f_Aa	f_na	f_Ana
w_A	96.0%																				
w_n		99.8%																			
w_a			99.2%																		
w_An	94.5%	97.1%		96.2%																	
w_Aa	91.9%		98.2%		96.3%																
w_na		98.2%	99.1%			98.9%															
w_Ana	90.7%	96.0%	98.2%	93.5%	95.9%	97.6%	96.0%														
d_A								97.6%													
d_n									99.7%												
d_a										98.0%											
d_An								89.5%	85.4%		88.8%										
d_Aa								82.9%		84.6%		83.9%									
d_na									95.9%	96.5%			96.4%								
d_Ana								79.1%	79.6%	83.1%	79.4%	81.1%	81.9%	80.8%							
f_A								93.6%							95.2%						
f_n									98.9%							99.6%					
f_a										95.9%							94.5%				
f_An								76.5%	82.0%		76.5%				87.8%	82.2%		86.9%			
f_Aa								53.9%		75.7%		63.7%			72.7%		67.5%		70.5%		
f_na									77.0%	90.1%			86.8%			85.8%	86.5%			86.9%	
f_Ana								45.0%	65.1%	70.6%	48.8%	55.8%	68.9%	56.9%	65.1%	72.0%	61.4%	67.4%	64.0%	64.6%	65.3%

Modelos treinados com 1 milhão de captchas de 8 caracteres de extensão

Resultados e generalização dos modelos

MODEL\ DATASET	w_A	w_n	w_a	w_An	w_Aa	w_na	w_Ana	d_A	d_n	d_a	d_An	d_Aa	d_na	d_Ana	f_A	f_n	f_a	f_An	f_Aa	f_na	f_Ana
w_A	50.1%																				
w_n		19.3%																			
w_a			85.2%																		
w_An	62.7%	56.2%		66.6%																	
w_Aa	57.9%		69.0%		69.2%																
w_na		76.3%	84.3%			82.2%															
w_Ana	53.5%	57.8%	69.0%	60.7%	67.1%	66.3%	67.2%														
d_A								22.8%													
d_n									38.2%							11.6%					
d_a										23.8%											
d_An								17.2%	15.4%		17.3%										
d_Aa								10.4%		10.5%		10.9%									
d_na									15.8%	17.6%			18.3%								
d_Ana																					
f_A								16.4%							21.6%						
f_n									32.9%							38.7%					
f_a										13.6%							12.4%				
f_An								8.6%	9.9%		8.0%				15.4%	12.6%		15.3%			
f_Aa										5.1%							10.1%		7.8%		
f_na									7.2%	7.9%			8.0%			8.5%	9.3%			8.4%	
f_Ana																		7.6%			

Modelos treinados com 1 milhão de captchas de 16 caracteres de extensão