THE DEVELOPER'S CONFERENCE

Trilha – Computação Cognitiva

DEEP FAKE E #TRAKINAGENS PARA FUGIR DAS IA(S)

Pedro Bezerra

DEEP FAKE E #TRAKINAGENS

PARA FUGIR DAS IA(S)

THE DEVELOPER'S CONFERENCE

TRILHA – COMPUTAÇÃO COGNITIVA

Por que falar sobre isso?

A computação cognitiva está sendo implementada, em larga escala, **sem a diminuição da probabilidade do sucesso de ataques de cibernéticos.**

Sejamos realistas ...

- Segurança necessita de um esforço político e financeiro;
- Risco é probabilístico, assim não é garantido que todos os sistemas serão atacados;
- 99% das pessoas não querem ser vítimas de um ataque, seja por "ego", multas ou perdas de contratos.

https://www.linkedin.com/in/pedrobezerragrc

LinkedIn **pedrobezerragrc**

Pedro Bezerra

Consultor de Segurança da Informação com foco em **Governança, Risco e Conformidade (GRC)** desde 2008. Atuou em projetos nacionais e internacionais em varejistas, indústria médica e instituições financeiras.

Desde 2013 desenvolve estudos e projetos independentes sobre a adoção de tecnologias disruptivas como **IoT (Internet das Coisas) e Inteligência Artificial**.

THE DEVELOPER'S CONFERENCE

Trilha – Computação Cognitiva

security H1V3

THE DEVELOPER'S CONFERENCE

Comunidades

AI Brasil
an artificial intelligence community

https://www.meetup.com/pt-BR/Security-Hive/

Procurar por Security H1V3

https://web.facebook.com/H1V3Sec

Grupos em Apps: Entre na nossa página do Meetup.com e fale com Pedro Bezerra.

https://web.telegram.org/#/im?p=@H1V3SecResurrection

https://www.meetup.com/pt-BR/ai-brasil

https://www.youtube.com/c/AIBrasilCommunity

https://web.facebook.com/BrasilAI

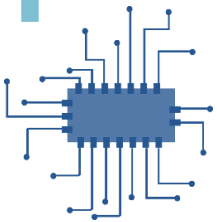
Grupos em Apps: Entre na nossa página do Meetup.com e fale com Pedro Bezerra.

https://web.telegram.org/#/im?p=g310549344

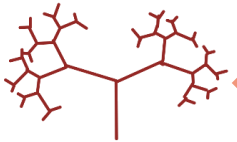
AI, Machine Learning, Deep Learning e NLP

THE DEVELOPER'S CONFERENCE

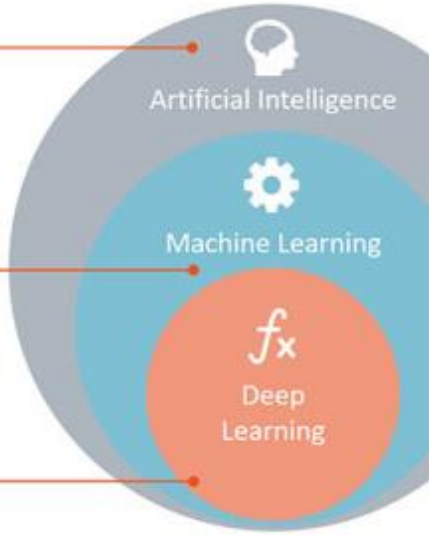
Artificial Intelligence - AI
Qualquer técnica que permita aos computadores **imitar** o comportamento humano



Machine Learning - ML
Subconjunto de técnicas de AI que utilizam **métodos estatísticos** para permitir que as máquinas melhorem com experiências.

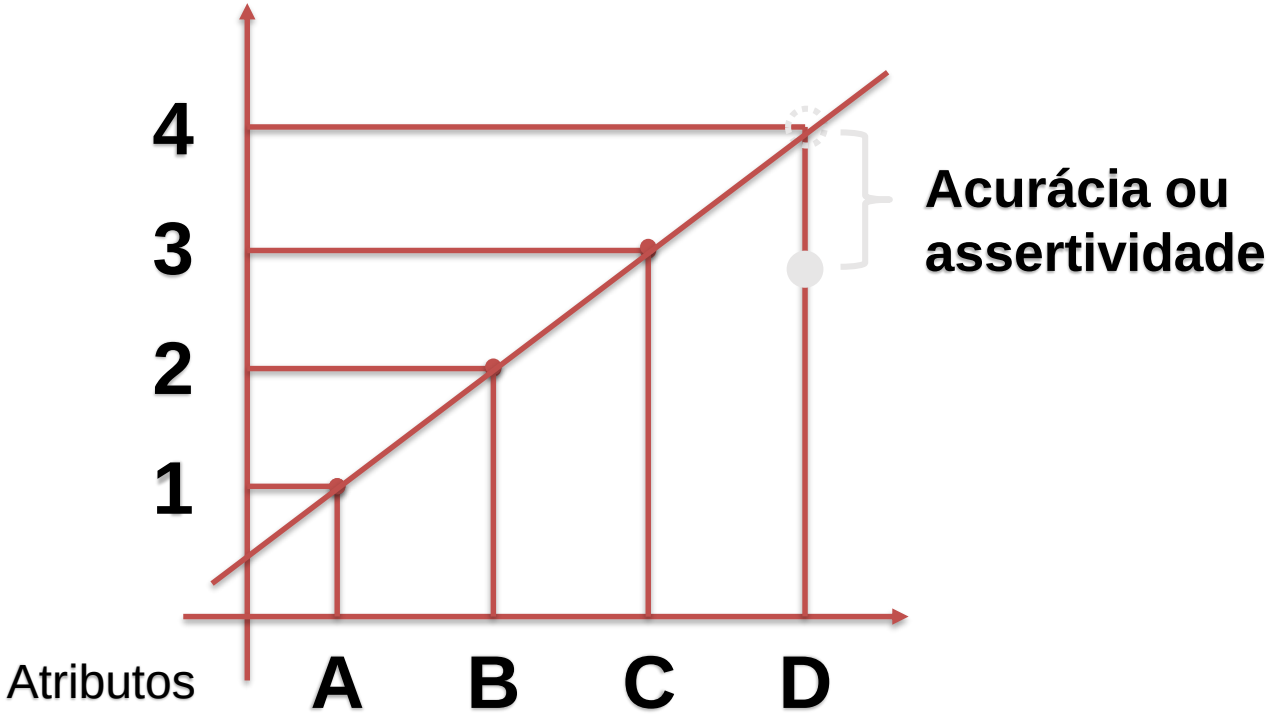


Deep Learning - DL
Subconjunto de ML que possibilita a computação de **redes neurais** de várias camadas.



Fontes:
https://content-static.upwork.com/blog/uploads/sites/3/2017/06/27091427/image-43.png
http://biogeocarlos.blogspot.com.br/2009/04/arte-zoologia-iv-camaleon.html

Tentando simplificar



Tentando simplificar

THE DEVELOPER'S CONFERENCE

Somos enganados por IA.

THE DEVELOPER'S CONFERENCE

Deep Fake

Vídeo

Fonte: BBC NEWS BRASIL: https://www.youtube.com/watch?v=OrGLT6-pKqs

Somos enganados por IA.

THE DEVELOPER'S CONFERENCE

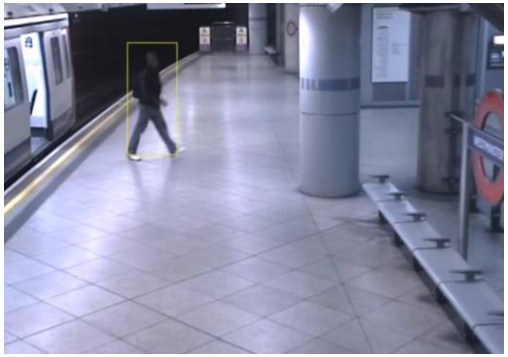
Deep Fake

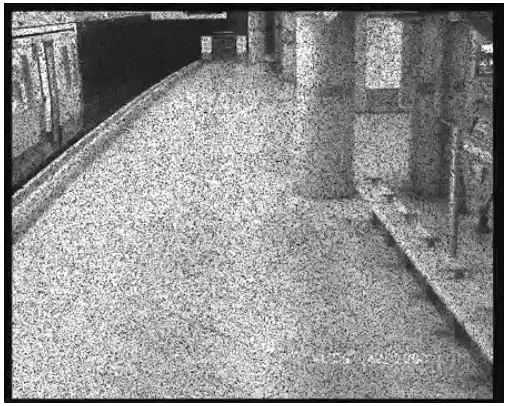
Texto

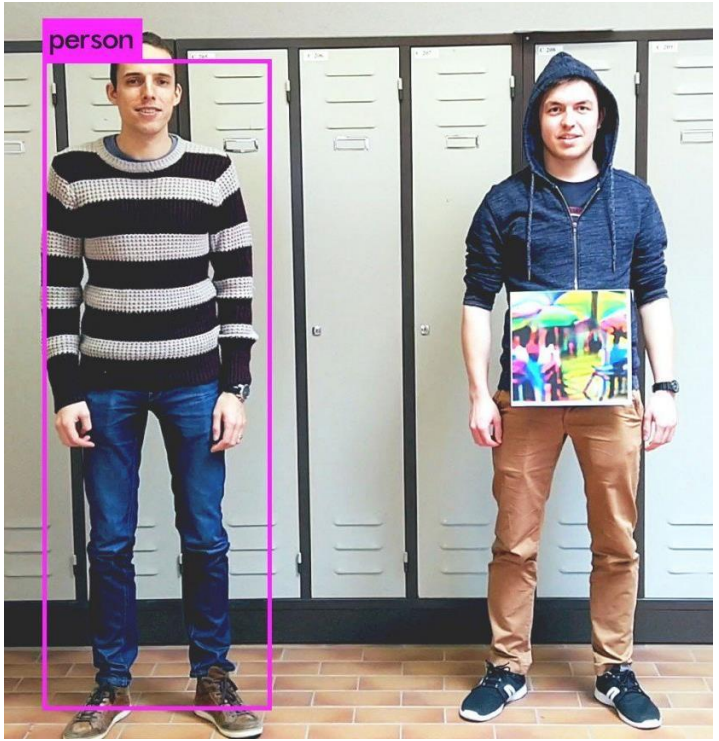
Fonte: THE GUARDIAN: https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

Podemos enganar a IA !
#Trakinagem

THE DEVELOPER'S CONFERENCE







Podemos enganar a IA !
#Trakinagem

THE DEVELOPER'S CONFERENCE

NEXTCon
Online AI Tech Talk Series
Friday, Jan 18

Adversarial Attacks on A.I. Systems

Anant Jain
Co-founder, commonlounge.com (Compose Labs)

https://commonlounge.com
https://index.anantja.in

toaster

imagens identificadas

DLP AI - Imagens

imagem

DLP AI - Imagens

imagem

Conclusões

Adversarial Patch
Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer
https://arxiv.org/abs/1712.09665

THE DEVELOPER'S CONFERENCE

1. Os patches são universais porque **podem ser usados para atacar qualquer cena**, robustos porque funcionam sob uma ampla variedade de transformações e direcionados porque podem fazer com que um classificador produza qualquer classe-alvo.
2. Os sistemas de aprendizagem profunda são amplamente vulneráveis a exemplos contraditórios, **inputs cuidadosamente escolhidos** que fazer com que a rede mude a saída sem uma mudança visível para um ser humano [15, 5].
3. Porque esse patch é independente de cena, permite que **atacantes criem um ataque no mundo físico sem conhecimento prévio de as condições de iluminação, ângulo da câmera, tipo de classificador sendo atacado**, ou até mesmo os outros itens dentro a cena.

Conclusões

Adversarial Patch
Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer
https://arxiv.org/abs/1712.09665

THE DEVELOPER'S CONFERENCE

4. Este ataque é significativo porque o atacante não precisa saber que imagem ele está atacando ao construir o ataque. Depois de gerar um patch adversário, o patch **poderia ser amplamente distribuídos pela Internet para outros invasores imprimirem e usarem**.
5. As técnicas de **defesa existentes que se concentram na defesa contra pequenas perturbações** podem não ser robustas a perturbações maiores como estas. Na verdade, o trabalho recente demonstrou que os modelos treinados por adversários de **última geração sobre o MNIST ainda são vulneráveis** a perturbações
6. Muitos modelos ML **operam sem validação humana** de cada entrada e, assim, atacantes mal-intencionados não se preocupam com a imperceptibilidade de seus ataques.
7. Mesmo que os seres humanos sejam capazes de perceber patches, eles podem não entender a intenção do patch e **vê-lo como uma forma de arte**.

Trilha – Computação Cognitiva

DEEP FAKE E #TRAKINAGENS PARA FUGIR DAS IA(S)

Pedro Bezerra

#Perguntas

THE DEVELOPER'S CONFERENCE