

GAN, SEGURANÇA, VIÉS E ÉTICA

Lucas Bastos

Gabriel Tormin

Marco Aurélio

Filipe Ribeiro

20 de maio de 2021

UnB - Departamento de Estatística

G - Generative
A - Adversarial
N - Network

Redes Adversárias Generativas

Duas redes competindo:

Generativa vs Discriminativa

Introduzido por Ian Goodfellow et al. (2014)

Podemos pensar em algoritmos generativos como o oposto dos algoritmos discriminativos.

Podemos pensar em algoritmos generativos como o oposto dos algoritmos discriminativos.

Discriminativos reduzem a informação para obter rótulos.

Podemos pensar em algoritmos generativos como o oposto dos algoritmos discriminativos.

Discriminativos reduzem a informação para obter rótulos.

Generativos criam informação para um dado rótulo.

Exemplo: classificação de e-mails em *spam* ou não *spam*.

Exemplo: classificação de e-mails em *spam* ou não *spam*.

Discriminativas: $p(x|y)$

Exemplo: classificação de e-mails em *spam* ou não *spam*.

Discriminativas: $p(x|y)$

Generativo: Como obter x ? Otimizando $p(x|y)$

Loop de feedback duplo:

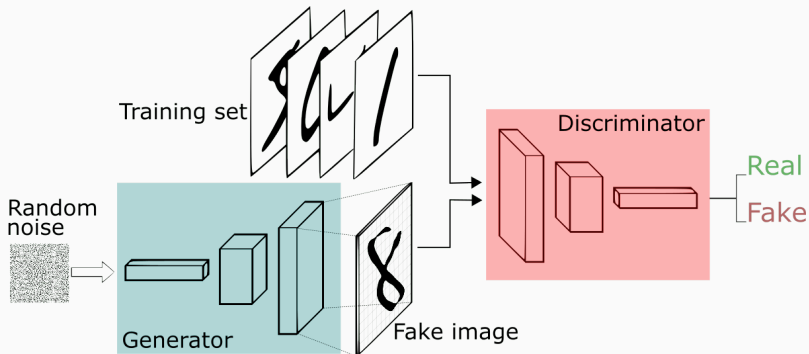


Figura: deeplearningbook.com.br - Capítulo 54

GANs podem aprender qualquer distribuição.

Imagens, música, fala, prosa, etc.

Combinando esses, pode gerar Deep Fakes.

Contudo, em 2019 DeepMind mostrou que os Autoencoders Variacionais (VAEs) poderiam superar as GANs na geração de faces.



Figura: Two of Barrat's GAN-generated nudes. Images by Robbie Barrat

Mais exemplos:

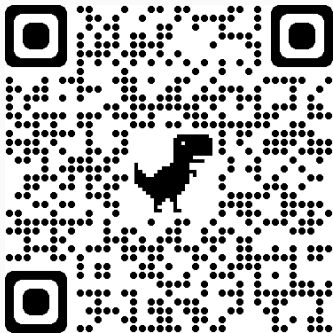
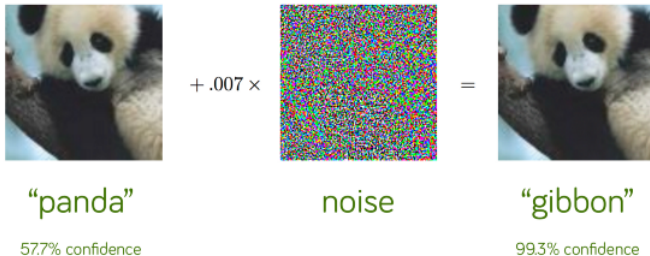


Figura: Fake Obama - BBC



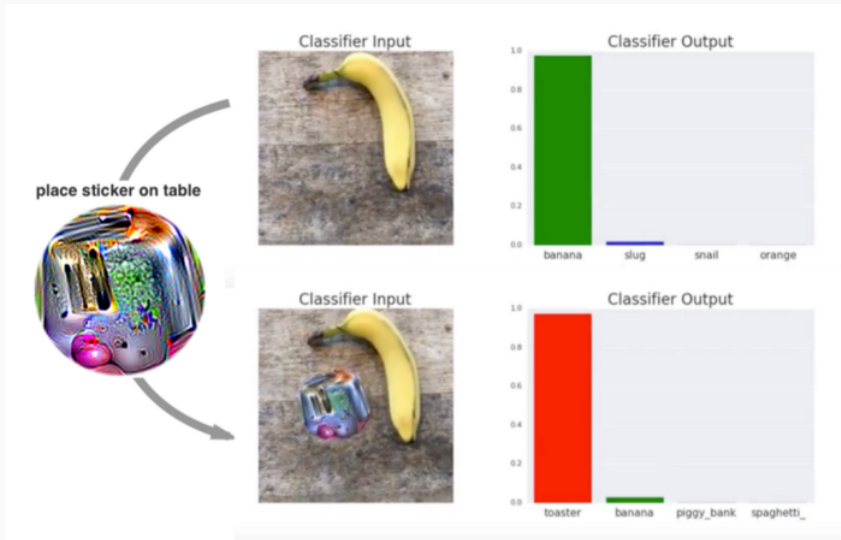
Figura: GANs interativas Nvidia

ATAQUES ADVERSARIAIS



Fonte: <https://towardsdatascience.com/adversarial-attacks-in-machine-learning-and-how-to-defend-against-them>.

ATAQUES ADVERSARIAIS



Como lidar com esses ataques?

- Denoising;
- Inserção de ataques para que a rede seja treinada a identificá-los;
- BaRT (Barrage of Random Transforms).

Perigos:

- roubo de identidade em sistemas de reconhecimento facial;
- sinais de trânsito usados por veículos autônomos;
- detectores de notícias falsas;

...

SEGURANÇA, VIÉS E ÉTICA

- Dados históricos e seu viés intrínseco

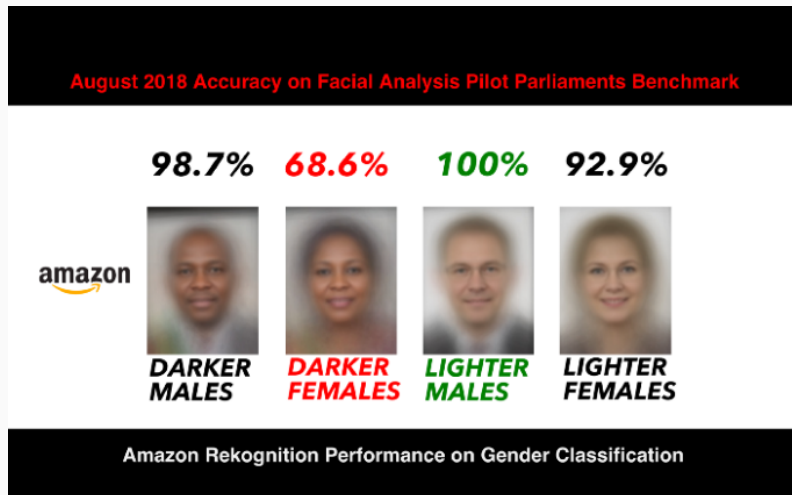


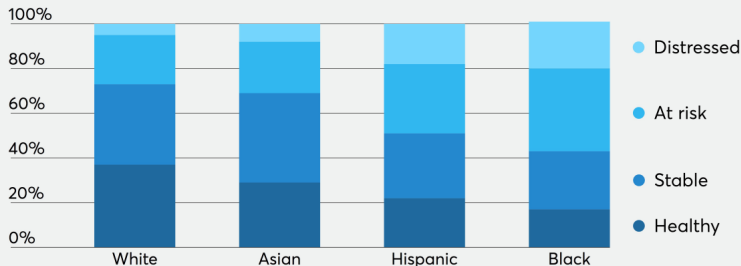
Figura: Fonte: Joy Buolamwini - Racial and Gender bias in Amazon Rekognition — Commercial AI System for Analyzing Faces.

SEGURANÇA, VIÉS E ÉTICA

- Viés na avaliação de crédito pelos bancos: taxa de crédito negado para negros nos EUA de 25%. Para brancos, 10%.

Unequal footing

Before the pandemic hit, Black- and Hispanic-owned businesses were more likely to be deemed "at risk" or "distressed." Profits, credit scores and ability to lean on retained earnings were the measures of health.



Source: New York Fed (data from late 2019)

Figura: Fonte: A House Divided - How Race Colors the Path to Homeownership.

SEGURANÇA, VIÉS E ÉTICA

Avaliação de professores do sistema público de ensino estadunidense a fim de obterem estabilidade.

- Caso: Daniel Santos, professor premiado e muito bem avaliado em diversas escolas em que lecionou, não era considerado apto pelo algoritmo.



Avaliação de detentos em liberdade provisória nos EUA.

- O algoritmo utilizado pelo Departamento de Justiça tem a última palavra acima do juiz e do oficial de condicional.
- "Profecia autorrealizada"

Tech policy / AI Ethics

AI is sending people to jail —and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao**

January 21, 2019

SEGURANÇA, VIÉS E ÉTICA

- Uso de dados pessoais e a violação da privacidade.

Conheça os 12 principais pontos sobre a LGPD

ESCOPO DE APLICAÇÃO
Afeta qualquer atividade que envolva utilização de dados pessoais, incluindo o tratamento pela internet, de consumidores, empregados, entre outros.

AUTORIZAÇÃO PARA O TRATAMENTO DE DADOS
O consentimento será umas das 10 possibilidades que legitimarão o tratamento de dados pessoais

PRINCÍPIOS DE PROTEÇÃO DE DADOS
Introduzidos 10 princípios da proteção de dados, incluindo-se o de demonstrar medidas adotadas para cumprir a lei (prestação de contas)

DIREITOS DOS TITULARES DE DADOS
Titulares dos dados terão amplos direitos: informação, acesso, retificação, cancelamento, oposição, portabilidade, entre outros.

AUTORIDADE
Previsão de Autoridade Nacional de Proteção de Dados, responsável por garantir cumprimento da Lei

NOTIFICAÇÕES OBRIGATÓRIAS
em caso de incidentes de segurança envolvendo os dados, nas situações aplicáveis

APLICAÇÃO EXTRATERRITORIAL
Aplica-se também a empresas que não possuem estabelecimento no Brasil

DADOS: SENSÍVEIS, DE MENORES E TRANSF. INTERNACIONAL
Regras específicas para tratar dados sensíveis, transferência internacional de dados e utilizar dados de crianças e adolescentes

ASSESSMENT SOBRE O TRATAMENTO DE DADOS
Necessidade de realizar assessment de impacto à proteção de dados (semelhante ao DPIA)

MAPEAMENTO DO TRATAMENTO DE DADOS
Atividades de tratamento de dados devem ser registradas em relatório

SANÇÕES
Multas de até 50 milhões de reais por infrações, entre outras sanções

DATA PROTECTION OFFICER (DPO)
Toda empresa responsável por tratamento de dados deverá nomear Encarregado da Proteção de Dados Pessoais

OPICE BLUM
OPICE BLUM • BRASÍLIA • SÃO PAULO • RIO DE JANEIRO
www.opiceblum.com.br

SEGURANÇA, VIÉS E ÉTICA

No Reino Unido, o reconhecimento facial de pedestres identificou milhares de pessoas erroneamente. Em Wales, o índice de falha chegou a 92%.



SEGURANÇA, VIÉS E ÉTICA

- Devido Processo Legal;
- Contraditório e Ampla Defesa;
- Inviolabilidade da pessoa humana;
- Princípio da não-autoincriminação;
- ...



SEGURANÇA, VIÉS E ÉTICA

- Desinformação e manipulação de massa.



SEGURANÇA, VIÉS E ÉTICA

- Remodelagem comportamental



Fonte: <https://thehackernews.com/2016/03/tay-artificial-intelligence.html>

SEGURANÇA, VIÉS E ÉTICA

- Escore de crédito social.



SEGURANÇA, VIÉS E ÉTICA

- Automação de armas e eficiência bélica.



<https://digitalbusiness.law/2017/10/the-use-of-ai-in-weapons-systems-the-uk-and-us-legal-and-regulatory-framework/>