

Audiência Pública
Senado Federal

Atributos do design sócio-técnico de
confiabilidade da IA: segurança, acurácia,
transparência, rastreabilidade e monitoramento

Fernanda Viégas, PhD

Google Principal Scientist

Harvard Gordon McKay Professor of Computer Science
Sally Starling Seaver Professor, Harvard Radcliffe Institute



HARVARD
School of Engineering
and Applied Sciences



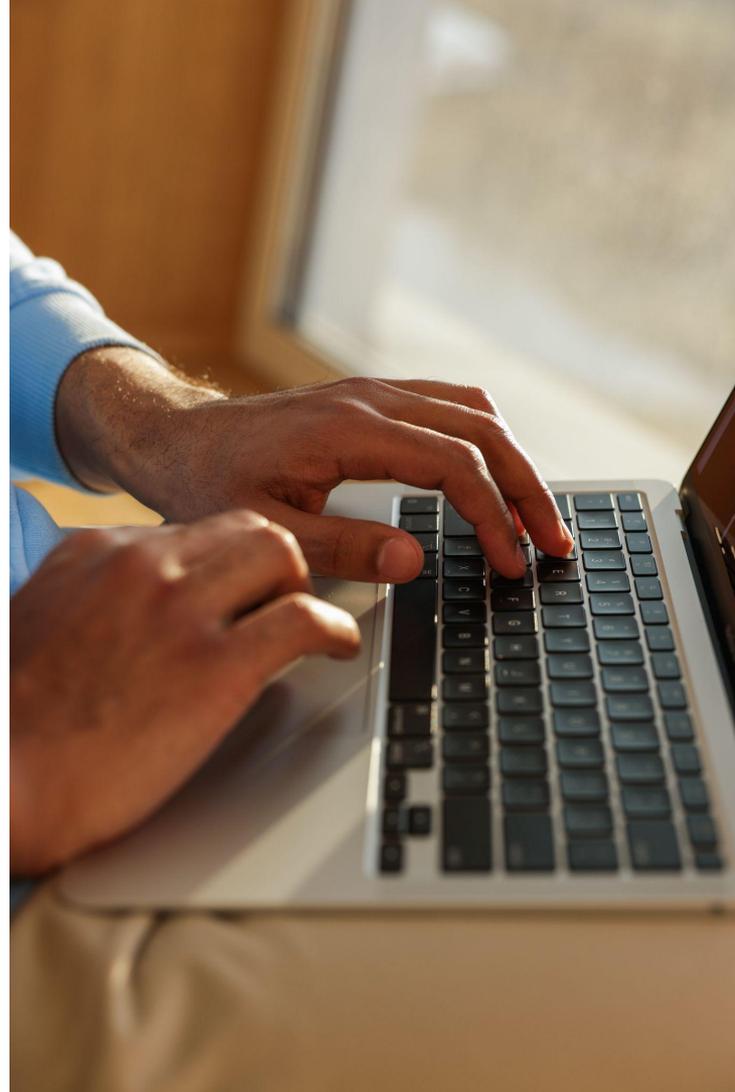
Harvard
Business
School

Minhas Áreas de Pesquisa:

Interação Humano-Computador
Interação Humano-IA

Explicabilidade em IA
Interpretabilidade em IA

Visualização de Dados



Minhas Áreas de Pesquisa:

Interação Humano-Computador

Interação Humano-IA

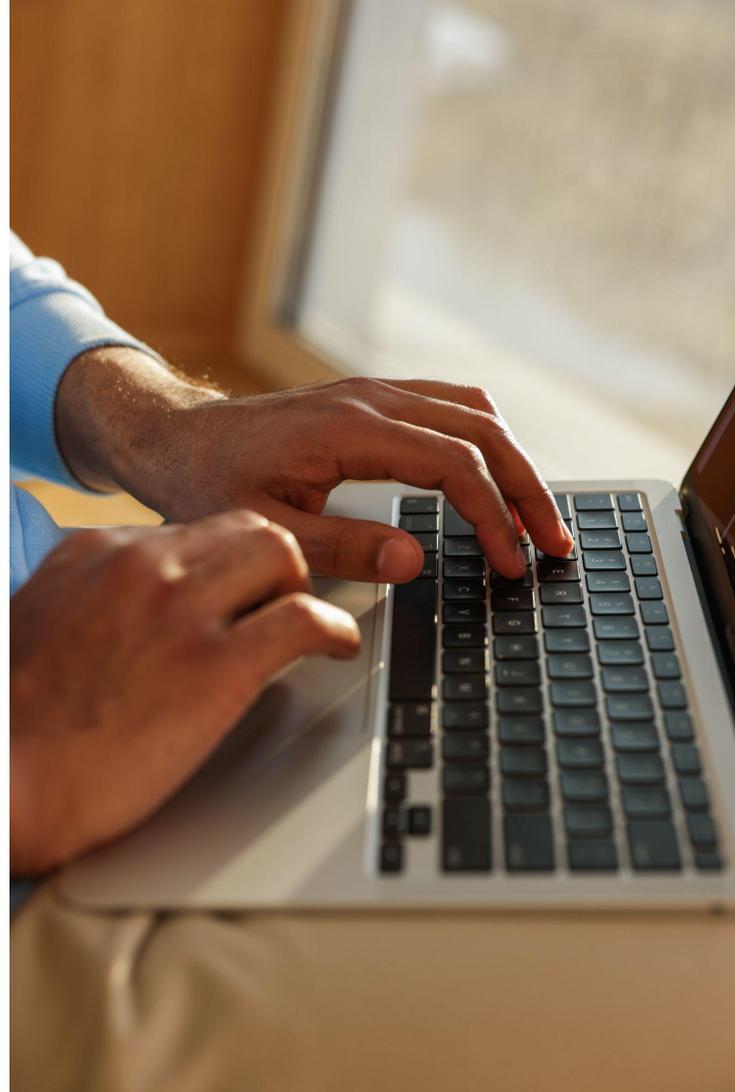
Explicabilidade em IA
Interpretabilidade em IA

Transparência
& confiabilidade

Visualização de Dados



O que tem funcionado?
O que sabemos que não funciona?



Por que razão precisamos de explicações?



As explicações são um meio para atingir um fim.

Devem ser adaptadas a diferentes usuários e diferentes utilizações.

- Ajudar programadores a identificar problemas e falhas e a melhorar o sistema
- Identificar causas de erros e parcialidade
- Capacitar os usuários com explicações relevantes e úteis
- Recurso para consumidores

Interpretabilidade / Explicabilidade para Todos



Interpretabilidade / Explicabilidade para Todos



Interpretabilidade / Explicabilidade para Todos

Expert em machine-learning

Engenheiro de software

Pesquisador de ML

Cientista de dados

Cientista da Computação

⋮

Hoje, a maioria do
trabalho acontece aqui

Profissionais

Médicos

Banqueiros

Jornalistas

Arquitetos

Advogados

Agricultores

⋮

Usuário leigo

Consumidores

Cidadãos

Crianças

Viajantes

Clientes

⋮

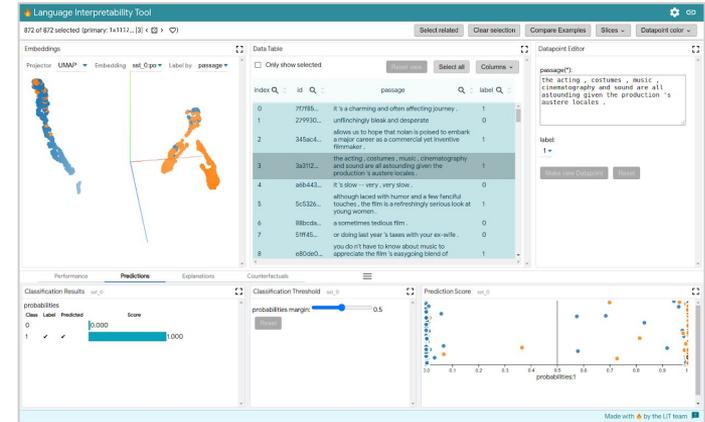
Machine learning é uma “black box”?

A ML **NÃO** é uma "black box" completa.

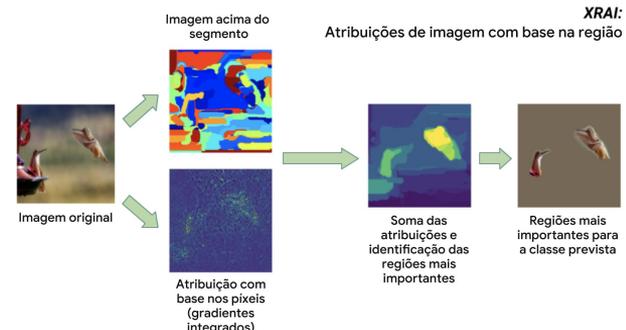
- Ferramentas e técnicas para programadores compreenderem modelos.
- Podemos sondar sistemas das mais variadas maneiras

... Mas há limitações técnicas e humanas:

- Transparência total não ajuda (nem aos experts)
- Nem sempre é possível fornecer explicações abrangentes.
- A pesquisa continua!



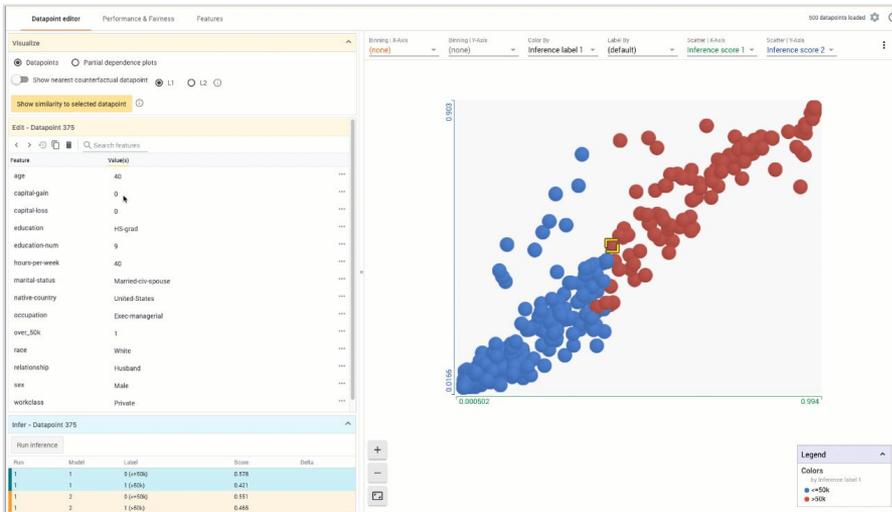
LIT (Language Interpretability Tool): ferramenta de código aberto para a sondagem de sistemas de NLP.



Podemos identificar as causas de erros e parcialidade?

Sim, em alguns casos

- Ferramentas e interpretabilidade de modelos.
- Ferramentas para compreensão e monitoramento de datasets



WIT (What-If Tool): ferramenta de código aberto para a sondagem de sistemas de machine learning, com sondagem de métricas de fairness

Interpretabilidade / Explicabilidade para Todos



Desafio do Google Flights:

Modelo altamente preciso de previsão de preços
Sem confiança do usuário



Abordagens anteriores

“Hoje é um bom dia para comprar a sua passagem”

- Foco só na venda
- Não confiável



Diretivas Simples

Abordagens anteriores

“Hoje é um bom dia para comprar a sua passagem”

- Coisa de “vendedor”
- Não confiável



Diretivas Simples

Abordagens anteriores

“Hoje é um bom dia para comprar a sua passagem”

- Coisa de “vendedor”
- Não confiável



Diretivas Simples

“É improvável que os preços caiam e há 75% de chance de que eles aumentem em US\$ 17 nos próximos 5 dias.”

- Informação demais
- Estressante



Transparência Radical

Abordagens anteriores

“Hoje é um bom dia para comprar a sua passagem”

- Coisa de “vendedor”
- Não confiável



Diretivas Simples

“É improvável que os preços caiam e há 75% de chance de que eles aumentem em US\$ 17 nos próximos 5 dias.”

- Muitas informações complexas
- Estressante



Transparência Radical

Introduction

User Needs + Defining Success

Data Collection + Evaluation

Mental Models

Explainability + Trust

Feedback + Control

Errors + Graceful Failure

Resources

Glossary

About

+ PAIR

<https://pair.withgoogle.com/guidebook/>

People + AI Guidebook

Designing human-centered AI products

User Needs + Defining Success



Identify user needs, find AI opportunities, and design your reward function.

Data Collection + Evaluation



Decide what data are required to meet your user needs, source data, and tune your AI.

Mental Models



Introduce users to the AI system and set expectations for system-change over time.

Explainability + Trust



Explain the AI system and determine if, when, and how to show model confidence.

Feedback + Control



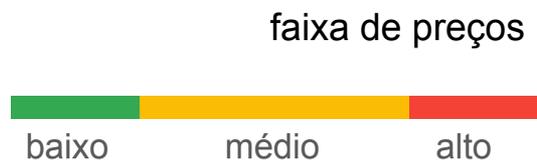
Design feedback and control mechanisms to improve your AI and the user experience.

Errors + Graceful Failure

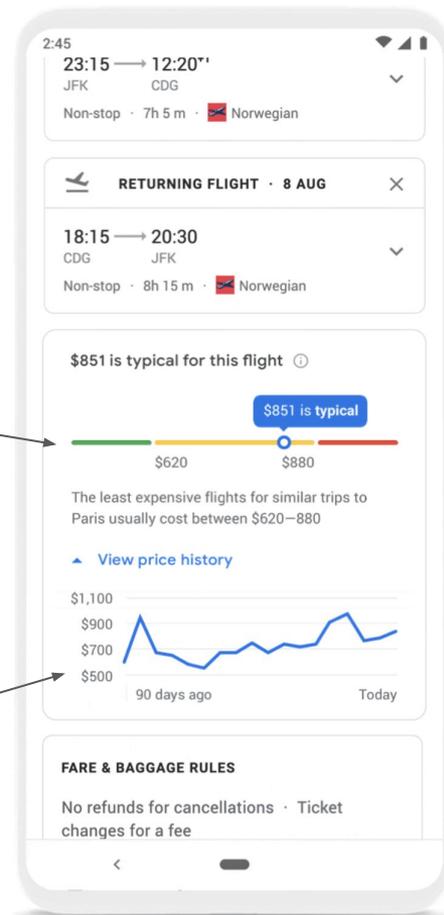


Identify and diagnose AI and context errors and communicate the way forward.

Solução: Explicabilidade + Confiança



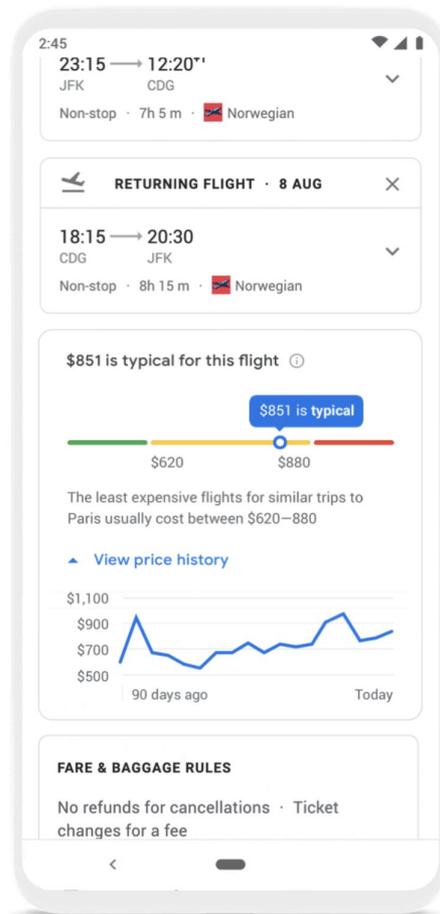
histórico de preços
probabilidade de mudança



Solução: Explicabilidade + Confiança

O que funcionou:

- User studies
- Números em contexto
 - faixa de variação de preços
 - visualização de dados
- Informações sem jargão técnico
 - **Sim:** “esse preço é típico”
 - **Não:** probabilidades
 - **Não:** confidence level



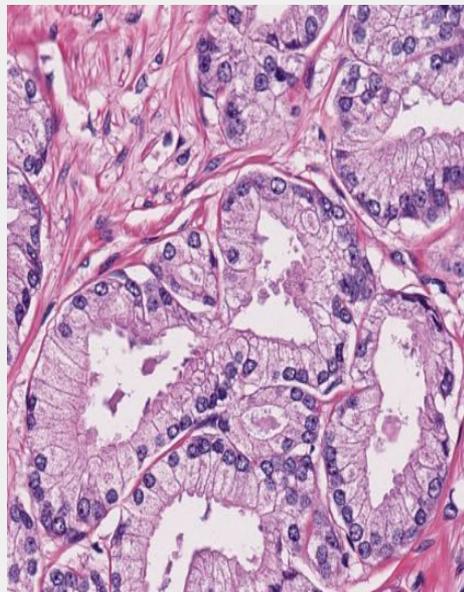
Interpretabilidade / Explicabilidade para Todos



Patologia e diagnóstico de câncer

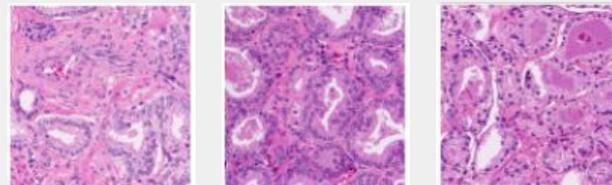
Embora haja muito trabalho na otimização do desempenho algorítmico, pouco trabalho examinou as interações reais que os médicos desejam de um sistema de busca de imagens semelhantes.

Tecido de uma biópsia

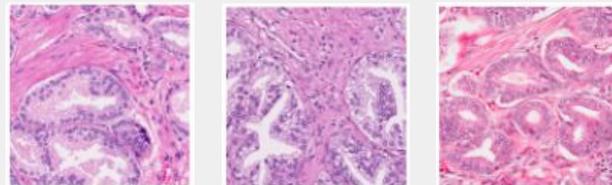


Imagens semelhantes de pacientes anteriores

Gleason 2



Gleason 3



Conceitos médicos importantes para os patologistas

Nível de estroma

Glândulas Normais

Glândulas tubulares

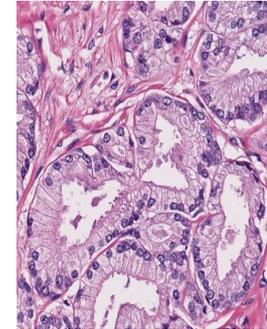
Glândulas fundidas

Borda luminal plana

Coloração de hematoxilina

Coloração de eosina

*“Gostaria de ver
imagens com mais
fused glands.”*



Conceitos médicos importantes para os patologistas

Nível de estroma

Glândulas Normais

Glândulas tubulares

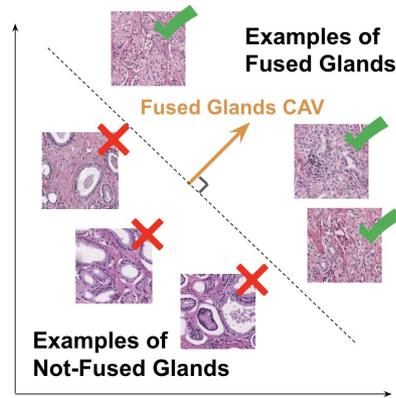
Glândulas fundidas

Borda luminal plana

Coloração de hematoxilina

Coloração de eosina

Conceito humano



Conceito de máquina
(CAV)

Conceitos médicos importantes para os patologistas

Nível de estroma

Glândulas Normais

Glândulas tubulares

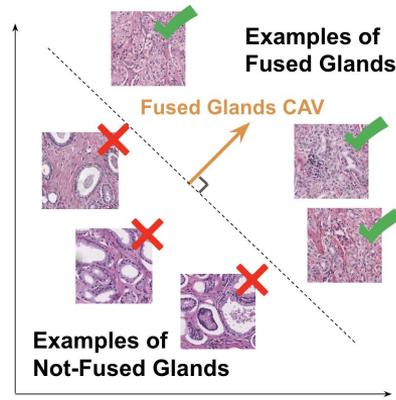
Glândulas fundidas

Borda luminal plana

Coloração de hematoxilina

Coloração de eosina

Conceito humano



**Conceito de máquina
(CAV)**

**Mais
tubular**



**Mais
Fused**

Interface acionável

Buscar resultados por conceito médico

SEARCH ON THE BOXED AREA

Magnification: 20X

Showing results 1 to 15

- Gleason: Normal
- High Grade PIN
- Gleason: 3
- Gleason: 4
- Gleason: 5

More similar Less similar

Gleason: 3 (10X) SEE MORE VARIETY

High Grade PIN (10X) SEE MORE VARIETY

Gleason: Normal (10X) SEE MORE VARIETY

ADJUST BY CONCEPT

- ignore all
- % Stroma** ignore concept
- % Normal glands** ignore concept
- % Tumor glands** ignore concept
- More Tubular** ignore concept
- More Fused** ignore concept
- Flat luminal border** ignore concept
- Eosin staining** ignore concept
- Hematoxylin staining** ignore concept

Mais transparência e poder para o usuário

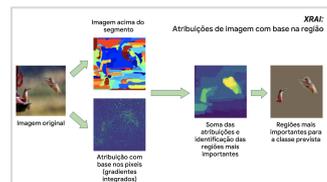
Melhor entendimento do sistema

Calibração de confiança

“Procure imagens semelhantes, mas com fused glands”

Interpretabilidade / Explicabilidade para Todos

Expert em machine-learning



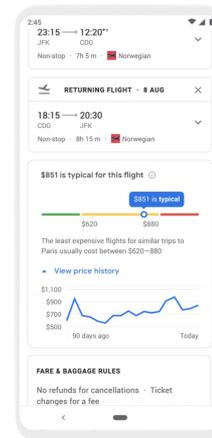
Profissionais

Showing results 1 to 15

ADJUST BY CONCEPT

- ignore all
- ignore concept

Usuário leigo



Conceitos técnicos, matemáticos

Conceitos especializados para o domínio em questão

Conceitos de entendimento geral, otimizados para compreensão

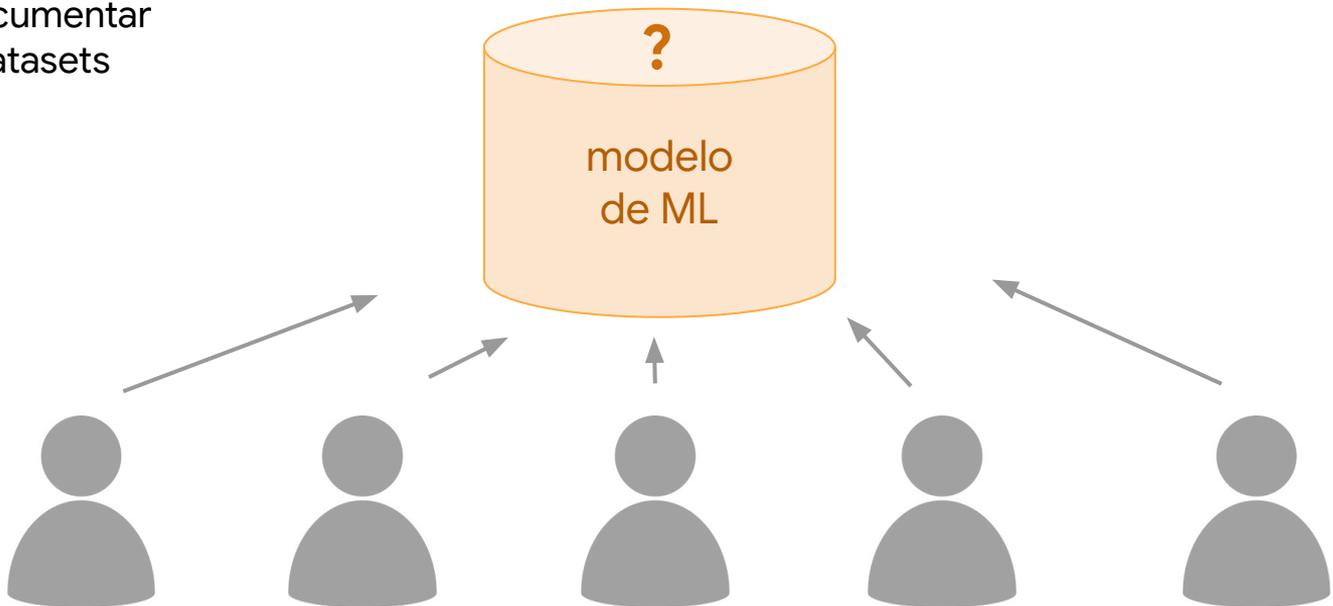
Outra forma de transparência: documentação

Hoje: dados e modelos são reutilizados em vários projetos

Há pouquíssima documentação

Erros, vieses, limitações podem ser amplificados

É preciso documentar
modelos e datasets



Explicabilidade a priori: “model cards”

Uma estrutura modular e flexível para o detalhamento do modelo:

- Proveniência
- Utilização
- Avaliação
- Limitações

Object Detection
Model Card v1 - Cloud Vision API

Overview

- Limitations
- Performance
- Test your own images
- Provide feedback

Explore

- Face Detection
- About Model Cards

Object Detection

The model analyzed in this card detects one or more physical objects within an image, from apparel and animals to tools and vehicles, and returns a box around each object, as well as a label and description for each object.

On this page, you can learn more about how the model performs on different classes of objects, and what kinds of images you should expect the model to perform well or poorly on.

MODEL DESCRIPTION

PERFORMANCE

PREDICTION 100%

RECALL 100%

0 0,02 0,04 0,06 0,08 0,10 0,12 0,14

● Open Images ● Google Internal

Input: Photo(s) or video(s)

Output: The model can detect 550+ different object classes. For each object detected in a photo or video, the model outputs:

- Object bounding box coordinates
- Knowledge graph ID ("MID")
- Label description
- Confidence score

Performance: evaluated for specific object classes recognized by the model (e.g. shirt, muffin), and for categories of objects (e.g. apparel, food).

Two performance metrics are reported:

- Average Precision (AP)

Cartão do modelo – Detecção de sorrisos em imagens

Detalhes do modelo

- Desenvolvido por investigadores da Google e da Universidade de Toronto, 2018, vl.
- Rede neural de convolução.
- Pré-treinado para o reconhecimento facial e, em seguida, aperfeiçoado com perda de centralidade cruzada para a classificação binária de sorrisos.

Utilização prevista

- Destinado a ser utilizado em aplicações divertidas, como a criação de sorrisos de desenhos animados em imagens reais; aplicações aumentativas, como o fornecimento de detalhes para pessoas cegas; ou aplicações auxiliares, como a procura automática de fotos com sorrisos.
- Especialmente destinado a públicos mais jovens.
- Não é adequado para a deteção de emoções ou para determinar o afeto; os sorrisos foram anotados com base na aparência física e não nas emoções subjacentes.

Fatores

- Baseado em problemas conhecidos com a tecnologia de visão de computador, os potenciais fatores relevantes incluem grupos de sexo, idade, raça e tipo de pele Fitzpatrick, fatores de hardware do tipo de câmara e tipo de lente e fatores ambientais de iluminação e humidade.

ário, tal como anotados no conjunto Poderá haver outros fatores não le dados de sorrisos. Sexo e idade base na apresentação visual, seguidos o/feminino e da idade jovem/idoso.

os positivos e a Taxa de falsos sproporcionados dos modelos a Taxa de falsa omissão, que rridentes) e positivas (sorridentes) as e negativas, respetivamente,

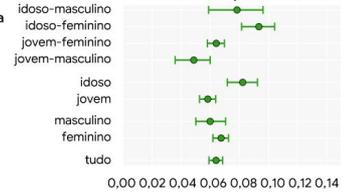
alores para diferentes erros que fusão dos sistemas de classificação

as recentes definições de (cf. [6, 26]), em que a paridade entre de a diferentes critérios de

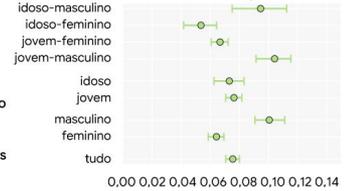
reamostragem do bootstrap. são de 0,5, em que todos os tipos de do mesmo intervalo (0,04 a 0,14).

Análises quantitativas

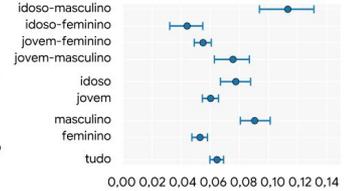
Taxa de falsos positivos a 0,5



Taxa de falsos negativos a 0,5



Taxa de falsa descoberta a 0,5



Taxa de falsa omissão a 0,5



Projetos e recursos mencionados

- Interpretabilidade para experts em ML
PAIR tools: <https://pair.withgoogle.com/tools/>
- Explicabilidade para usuários leigos
Gogle Flights: <https://medium.com/people-ai-research/pair-guidebook-google-flights-case-study-1ba8c7352141>
- Explicabilidade para usuários profissionais
Ferramentas de IA para patologistas: <https://dl.acm.org/doi/pdf/10.1145/3290605.3300234>
- Documentação
Data & Model Cards: <https://modelcards.withgoogle.com/about>
- PAIR Guidebook: <https://pair.withgoogle.com/guidebook>