Ethics, Bias, Privacy

Computer Vision – Lecture 20

Further Reading

- Timnit Gebru and Emily Denton, CVPR 2020 Tutorial on FATE/CV
- Kate Crawford, "The Trouble with Bias", NeurIPS 2017 Keynote
- Barocas, Hardt, Narayanan, "*Fairness and machine learning*"
- ACM Conference on Fairness, Accountability, and Transparency
- Law and Computer Science Course
- Oxford Internet Institute, Sandra Wachter

Why do we build ML systems?

Automate decision making, so machines can make decision instead of people.

Ideal: Automated decisions can be cheaper, more accurate, more impartial, improve our lives

Reality: automated decisions can encode bias, harm people, make lives worse

Case Study: COMPAS

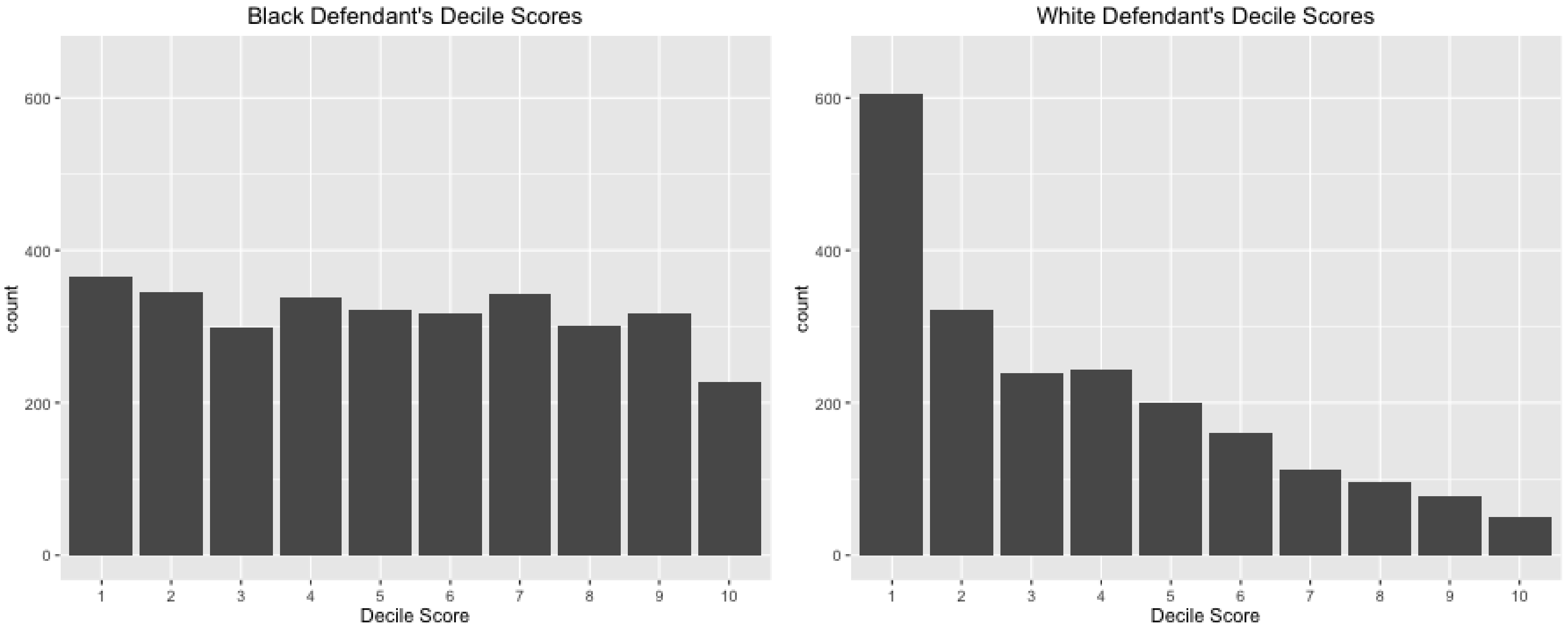
1. Person commits a crime, is arrested
2. COMPAS software predicts the chance that the person will commit another crime in the future (*recidivism)*
3. Recidivism scores impact criminal sentences: if a person is likely to commit another crime, shouldn't they get a longer sentence?

Real system that has been used in New York, Wisconsin, California, Florida, etc.

Slide credit: J Johnson (this and following)

Case Study: COMPAS

2016 ProPublica article analyzed COMPAS scores for >7000 people arrested in Broward county, Florida



Question: How many of these people ended up committing new crimes within 2 years?

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Recap: Error Metrics

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

Error Metrics: Error Rate

	Prediction: Low Risk	**Prediction: High Risk**
Outcome: No Recidivism	**True Negative (TN)**	**False Positive (FP)**
Outcome: Recidivated	**False Negative (FN)**	**True Positive (TP)**

$$\text{Error Rate} = \frac{\boldsymbol{FP}+\boldsymbol{FN}}{\boldsymbol{TN}+\boldsymbol{FP}+\boldsymbol{FN}+\boldsymbol{TP}}$$

How often is the prediction wrong?

Error Metrics: False Positive Rate

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	**True Negative (TN)**	**False Positive (FP)**
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

Error Rate = $\frac{FP+FN}{TN+FP+FN+TP}$

How often is the prediction wrong?

False Positive Rate = $\frac{\boldsymbol{FP}}{\boldsymbol{FP}+\boldsymbol{TN}}$

How often were non-offenders predicted to reoffend?

Error Metrics: False Negative Rate

	Prediction: Low Risk	**Prediction: High Risk**
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	**False Negative (FN)**	**True Positive (TP)**

Error Rate = $\frac{FP+FN}{TN+FP+FN+TP}$

How often is the prediction wrong?

False Positive Rate = $\frac{FP}{FP+TN}$

How often were non-offenders predicted to reoffend?

False Negative Rate = $\frac{\boldsymbol{FN}}{\boldsymbol{FN}+\boldsymbol{TP}}$

How often were offenders predicted not to reoffend?

Error Metrics: Different Stakeholders

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

Error Rate = $\frac{FP+FN}{TN+FP+FN+TP}$ *How often is the prediction wrong?*

Defendants care about this False Positive Rate = $\frac{FP}{FP+TN}$ *How often were non-offenders predicted to reoffend?*

False Negative Rate = $\frac{FN}{FN+TP}$ *How often were offenders predicted not to reoffend?*

Error Metrics: Different Stakeholders

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

Error Rate = $\frac{FP+FN}{TN+FP+FN+TP}$ *How often is the prediction wrong?*

Defendants care about this — False Positive Rate = $\frac{FP}{FP+TN}$ *How often were non-offenders predicted to reoffend?*

Judges care about this — False Negative Rate = $\frac{FN}{FN+TP}$ *How often were offenders predicted not to reoffend?*

Case Study: COMPAS

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	2681 (TN)	1282 (FP)
Outcome: Recidivated	1216 (FN)	2035 (TP)

$$\text{Error Rate} = \frac{FP+FN}{TN+FP+FN+TP} \approx 34.6\%$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} \approx 32.4\%$$

$$\text{False Negative Rate} = \frac{FN}{FN+TP} \approx 37.4\%$$

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Case Study: COMPAS

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Case Study: COMPAS

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

Error Rate $\approx 36.2\%$

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Error Rate $\approx 33.0\%$

Similar error rates between white and black defendants

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Case Study: COMPAS

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Error Rate $\approx 36.2\%$

Error Rate $\approx 33.0\%$

False Positive Rate $\approx 44.9\%$

False Positive Rate $\approx 23.5\%$

Black defendants have 1.9x higher False Positive Rate!

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Case Study: COMPAS

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Error Rate $\approx 36.2\%$

False Positive Rate $\approx 44.9\%$

False Negative Rate $\approx 28.0\%$

Error Rate $\approx 33.0\%$

False Positive Rate $\approx 23.5\%$

False Negative Rate $\approx 47.7\%$

White defendants have 1.7x higher False Negative Rate

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Case Study: COMPAS

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

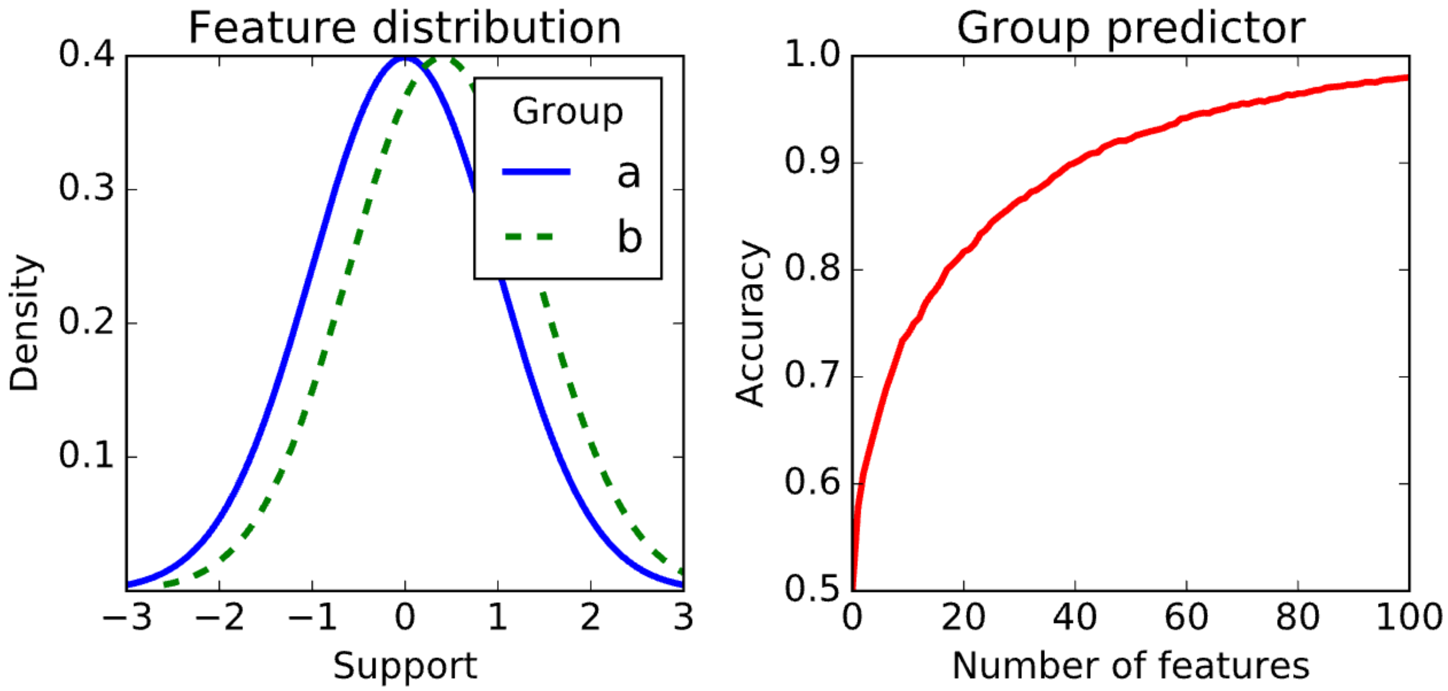
White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Surprising fact: COMPAS gives very different outcomes for white vs black defendants, but it does not use race as an input to the algorithm!

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

No Fairness Through Unawareness

Even if a sensitive feature (e.g. race) is not an input to the algorithm, other features (e.g. zip code) may correlate with the sensitive feature



Source: https://fairmlbook.org/classification.html

Formalizing Fairness

Y: Target variable (e.g. recidivism)

R: Classifier response (e.g. predicted recidivism)

A: Sensitive attribute (e.g. race)

Fairness Definition 1: Independence

The classifier response is *independent* (as a random variable) from the sensitive attribute

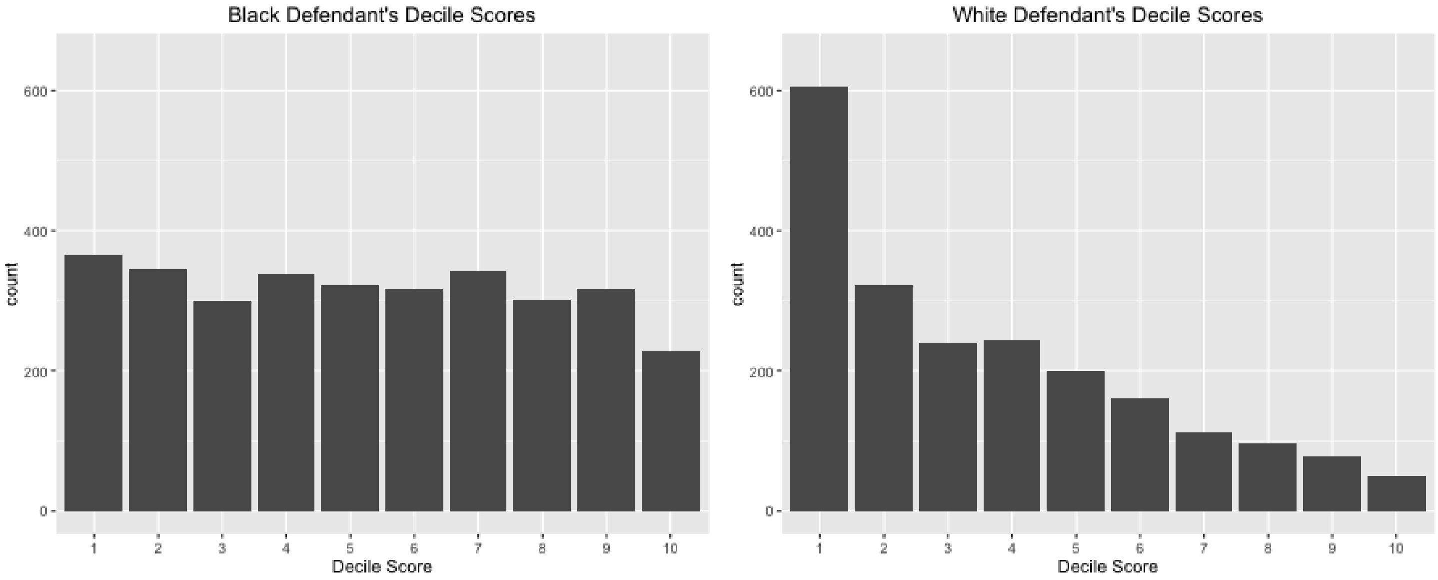
$$P(R, A) = P(R)P(A)$$

$$= P(R \mid A)P(A) \text{ (Chain Rule)}$$

$$\implies P(R \mid A) = P(R)$$

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness



COMPAS predictions are not independent – different distributions for black vs white

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Y: Target variable (e.g. recidivism)

R: Classifier response (e.g. predicted recidivism)

A: Sensitive attribute (e.g. race)

Fairness Definition #2: Separation

The classifier response is *conditionally independent* from the sensitive attribute given the target

$$P(R, A \mid Y) = P(R \mid Y) P(A \mid Y)$$

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Fairness Definition #2: Separation

The classifier response is *conditionally independent* from the sensitive attribute given the target

$$P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$$

Error rate parity:

Requires that all groups experience

- the same false negative rate.
- the same false positive rate.

COMPAS scores do not satisfy separation

Barocas, Hardt, and Narayanan. “Fairness and Machine Learning”, https://fairmlbook.org/index.html

Formalizing Fairness

Y: Target variable
R: Classifier response
A: Sensitive attribute

Independence: $P(R, A) = P(R)P(A)$
Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

Assume Y is binary, A is not independent of Y, and R is not independent of Y. Then, independence and separation cannot both hold.

(Proof in “Fairness and Machine Learning”)

Barocas, Hardt, and Narayanan. “Fairness and Machine Learning”, https://fairmlbook.org/index.html

Formalizing Fairness: Takeaways

There are **multiple ways** to formalize notions of fairness mathematically.

It is often impossible to achieve all notions of fairness at the same time

Fairness in ML **is not only a technical problem! We need to think about context, stakeholders, etc.**

There are many notions of fairness:
e.g. Arvind Narayanan, "21 fairness definitions and their politics"

Allocative Harms – Immediate Effect

- A system decides how to *allocate resources*
- If the system is biased, it may allocate resources unfairly or perpetuate inequality
- Examples:
  - Sentencing criminals
  - Loan applications
  - Mortgage applications
  - Insurance rates
  - College admissions
  - Job applications

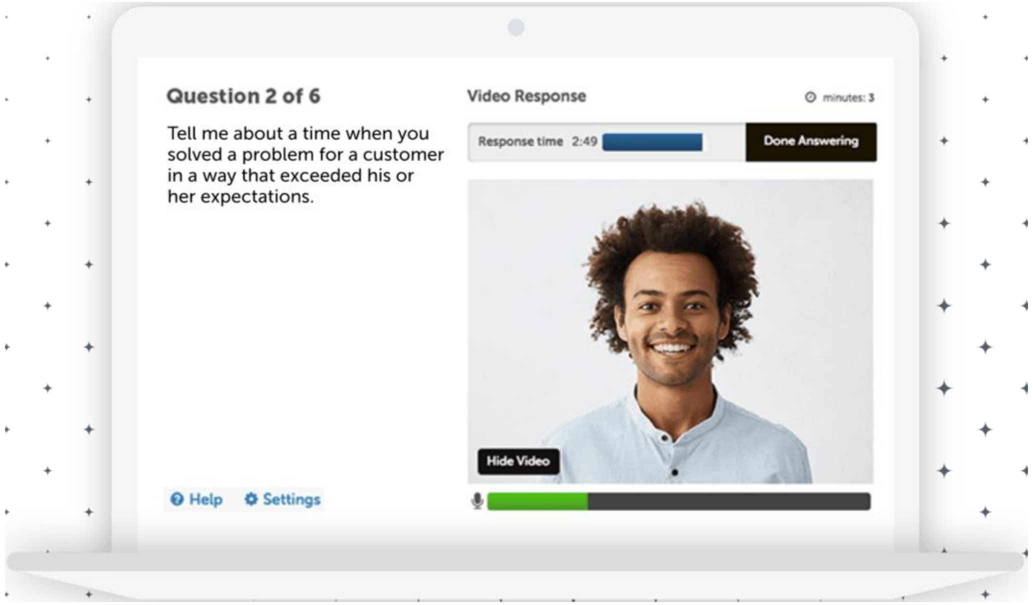
Barocas et al, "The Problem With Bias: Allocative Versus Representational Harms in Machine Learning", SIGCIS 2017
Kate Crawford, "The Trouble with Bias", NeurIPS 2017 Keynote

Example: Video Interviewing

Technology

A face-scanning algorithm increasingly decides whether you deserve the job

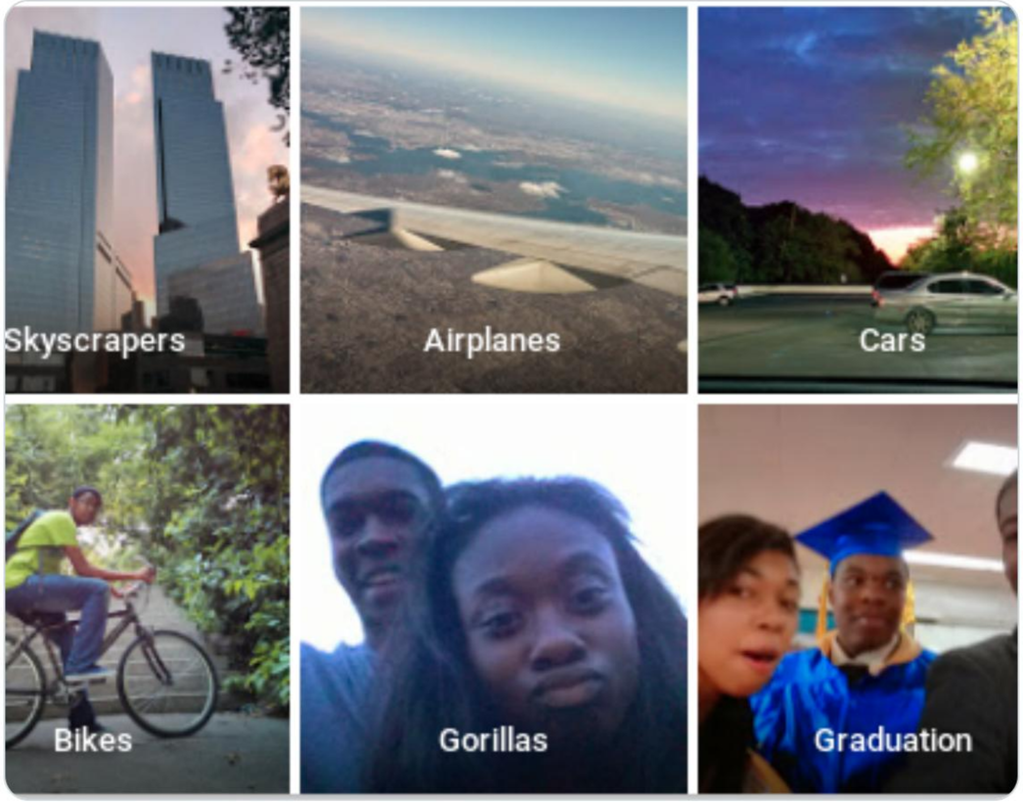
HireVue claims it uses artificial intelligence to decide who's best for a job. Outside experts call it 'profoundly disturbing.'



Source: https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/
https://www.hirevue.com/platform/online-video-interviewing-software
Example Credit: Timnit Gebru

Representational Harms

A system reinforces harmful stereotypes: denigration.



Barocas et al, “The Problem With Bias: Allocative Versus Representational Harms in Machine Learning”, SIGCIS 2017
Kate Crawford, “The Trouble with Bias”, NeurIPS 2017 Keynote
Source: https://twitter.com/jackyalcine/status/615329515909156865 (2015, tweet no longer avaliable)

Representational Harms – Long Term

- Harder to quantify
- Cultural

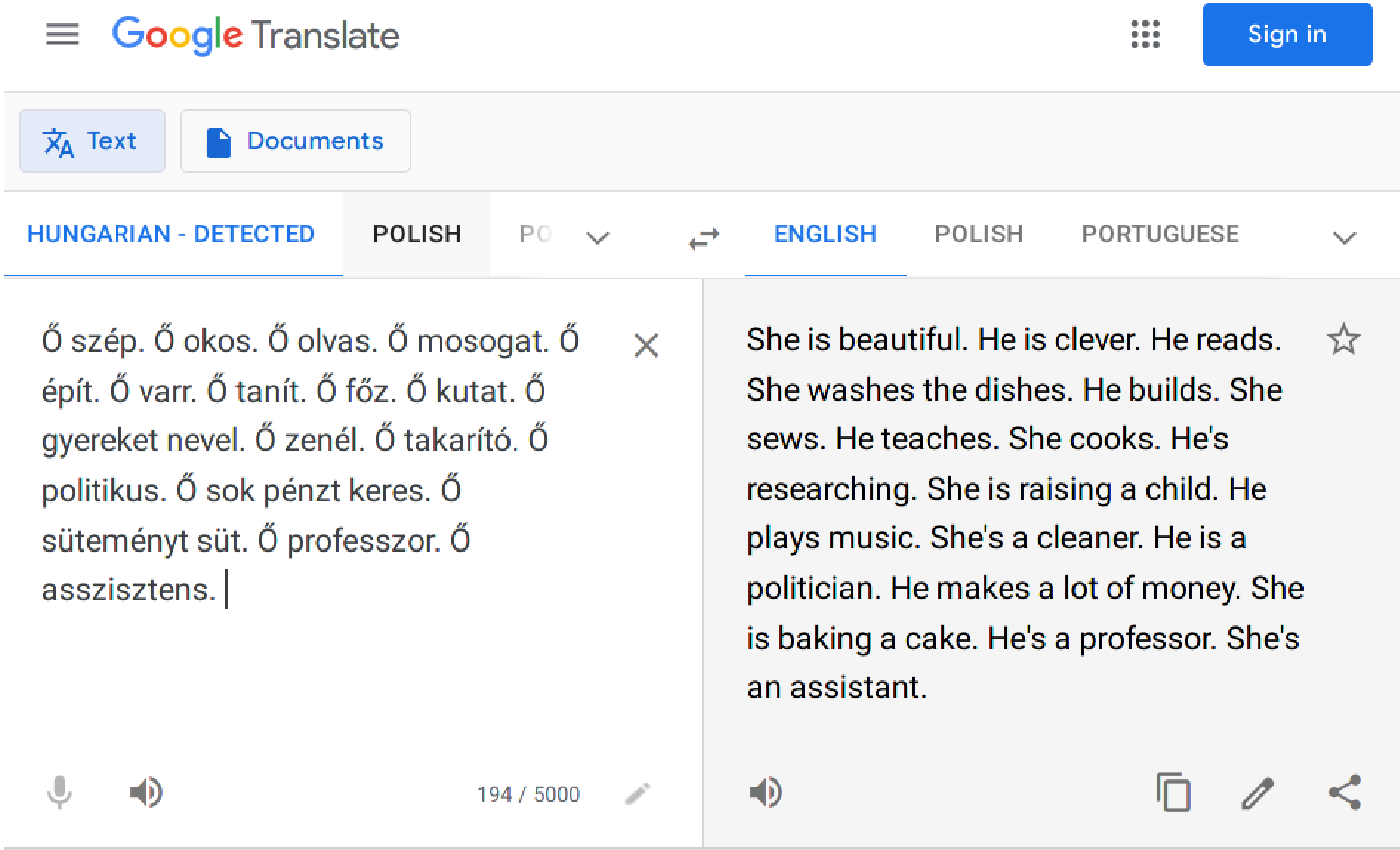
Types

• Denigration:	use of culturally disparaging terms
• Stereotype:	reinforces stereotypes
• Recognition:	a group is erased or made invisible
• Under-Representation:	a group is under-represented
• Ex-Nomination:	represent ideology as common sense

Solon Barocas, Kate Crawford, Aaron Shapiro, Hanna Wallach, 2017 'The Problem With Bias: Allocative Versus Representational Harms in Machine Learning', *SIGCIS Conference*

Hungarian -> English Translation

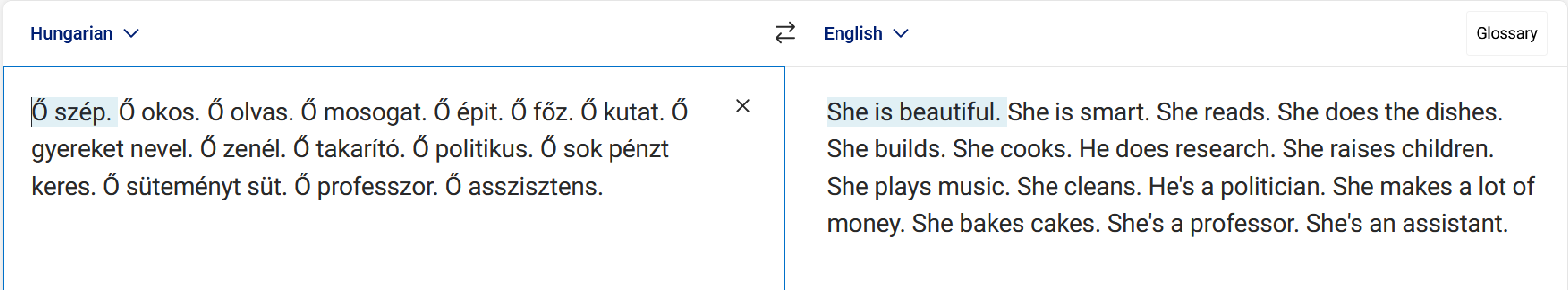
Hungarian does not use gendered pronouns



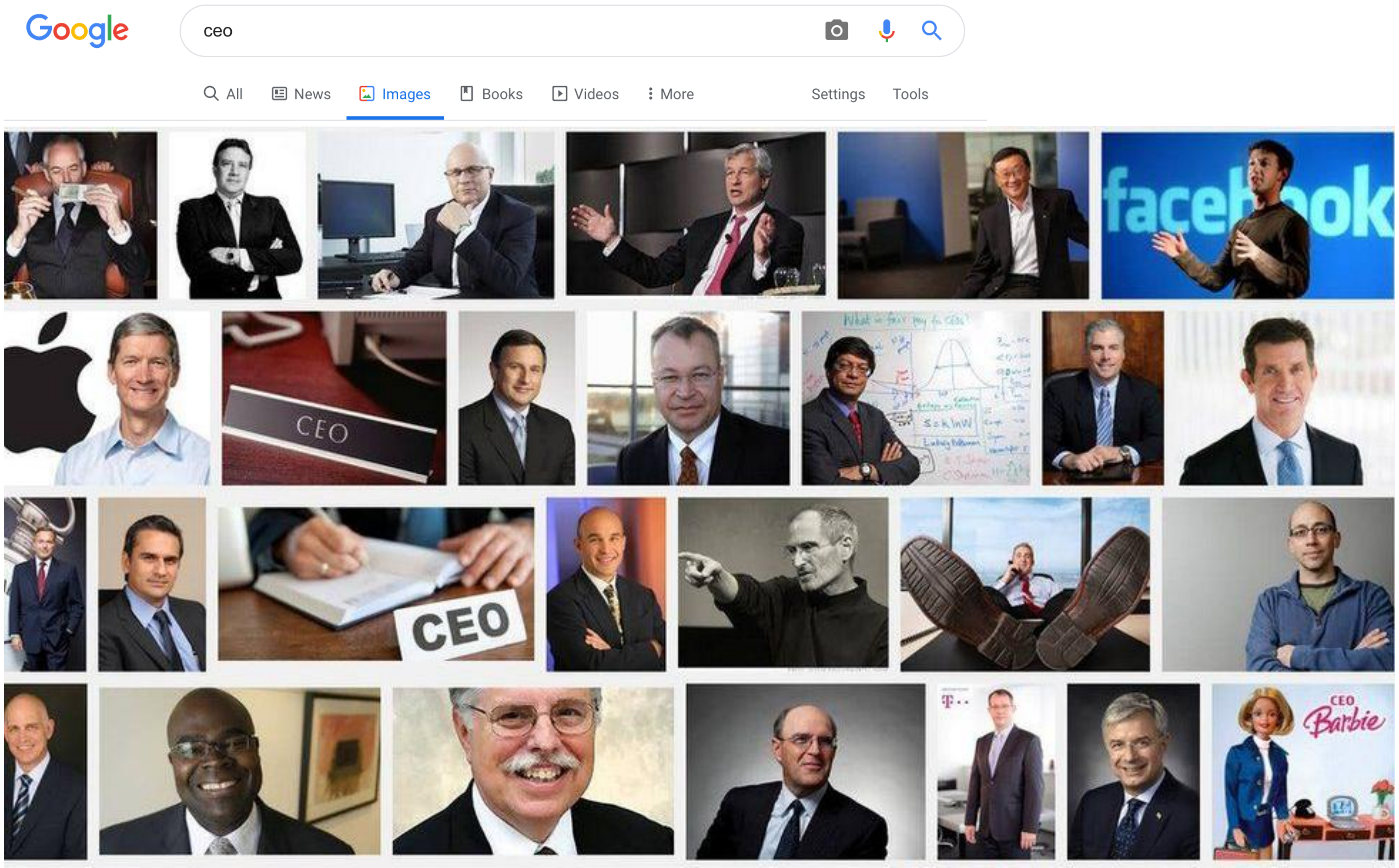
English translation makes assumptions

Source:
https://www.reddit.com/r/europe/comments/m9uphb/hungarian_has_no_gendered_pronouns_so_google

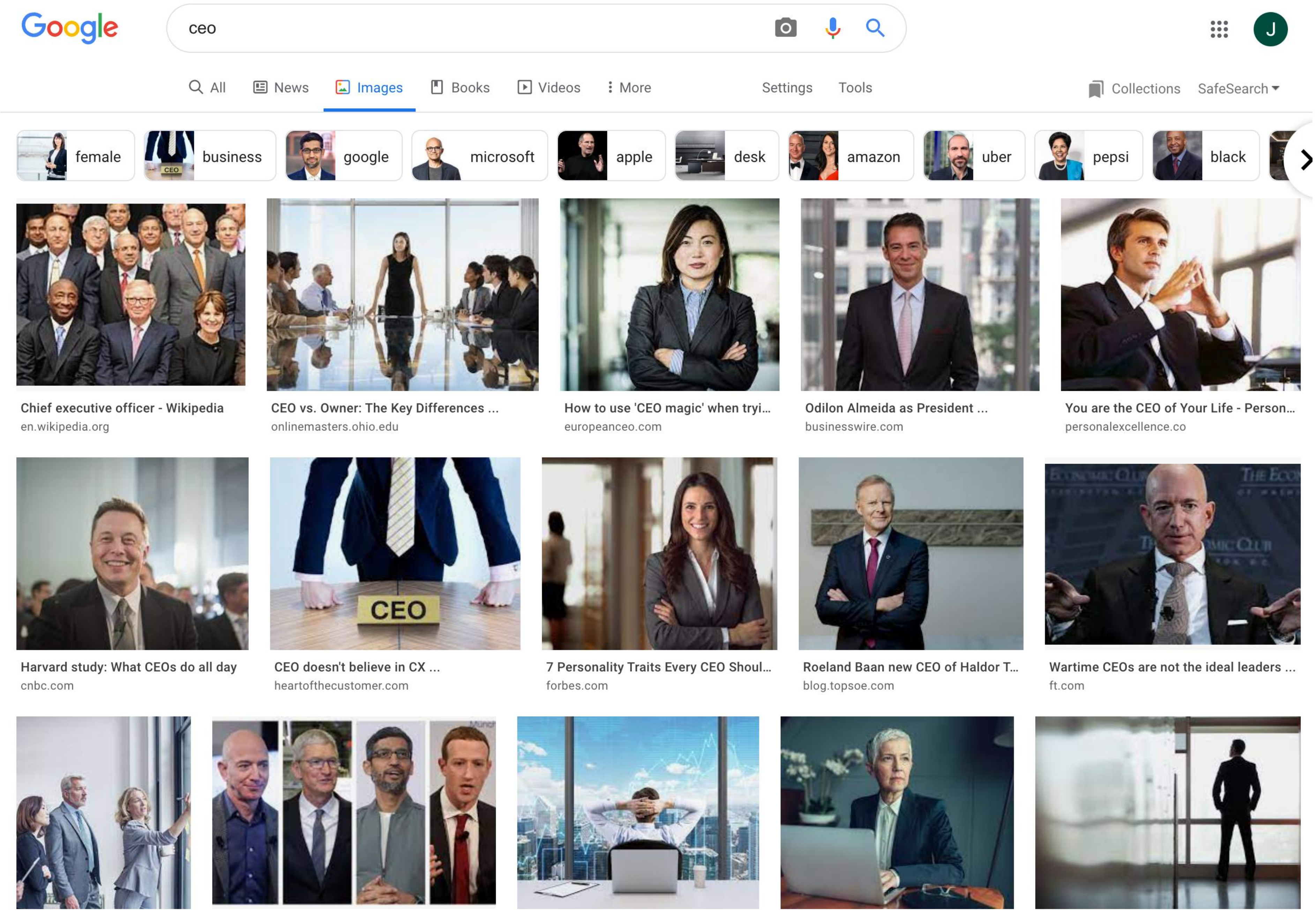
DeepL



Hungarian
English
Glossary
Ő szép. Ő okos. Ő olvas. Ő mosogat. Ő épit. Ő főz. Ő kutat. Ő gyereket nevel. Ő zenél. Ő takarító. Ő politikus. Ő sok pénzt keres. Ő süteményt süt. Ő professzor. Ő asszisztens.
She is beautiful. She is smart. She reads. She does the dishes. She builds. She cooks. He does research. She raises children. She plays music. She cleans. He's a politician. She makes a lot of money. She bakes cakes. She's a professor. She's an assistant.

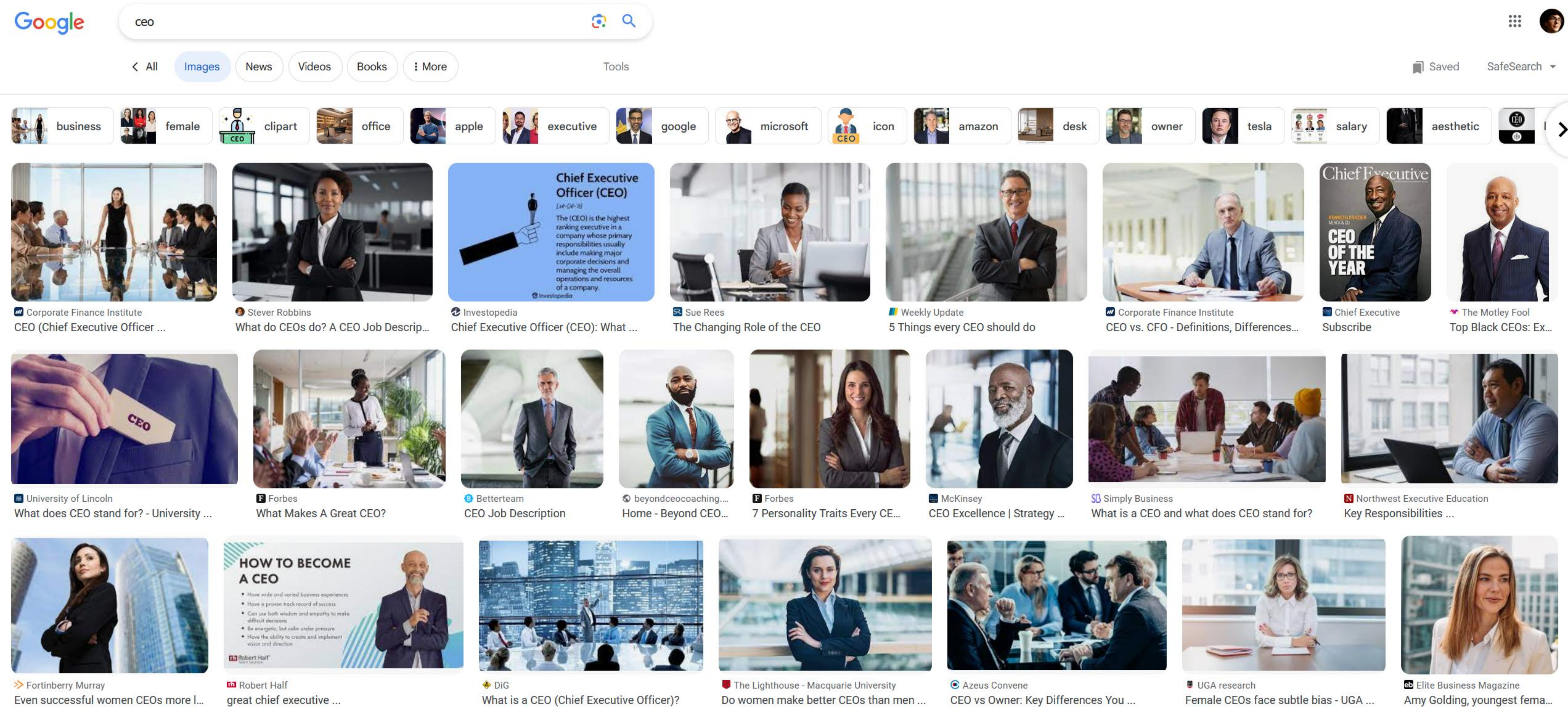


Source: https://www.bbc.com/news/newsbeat-32332603 , 2015

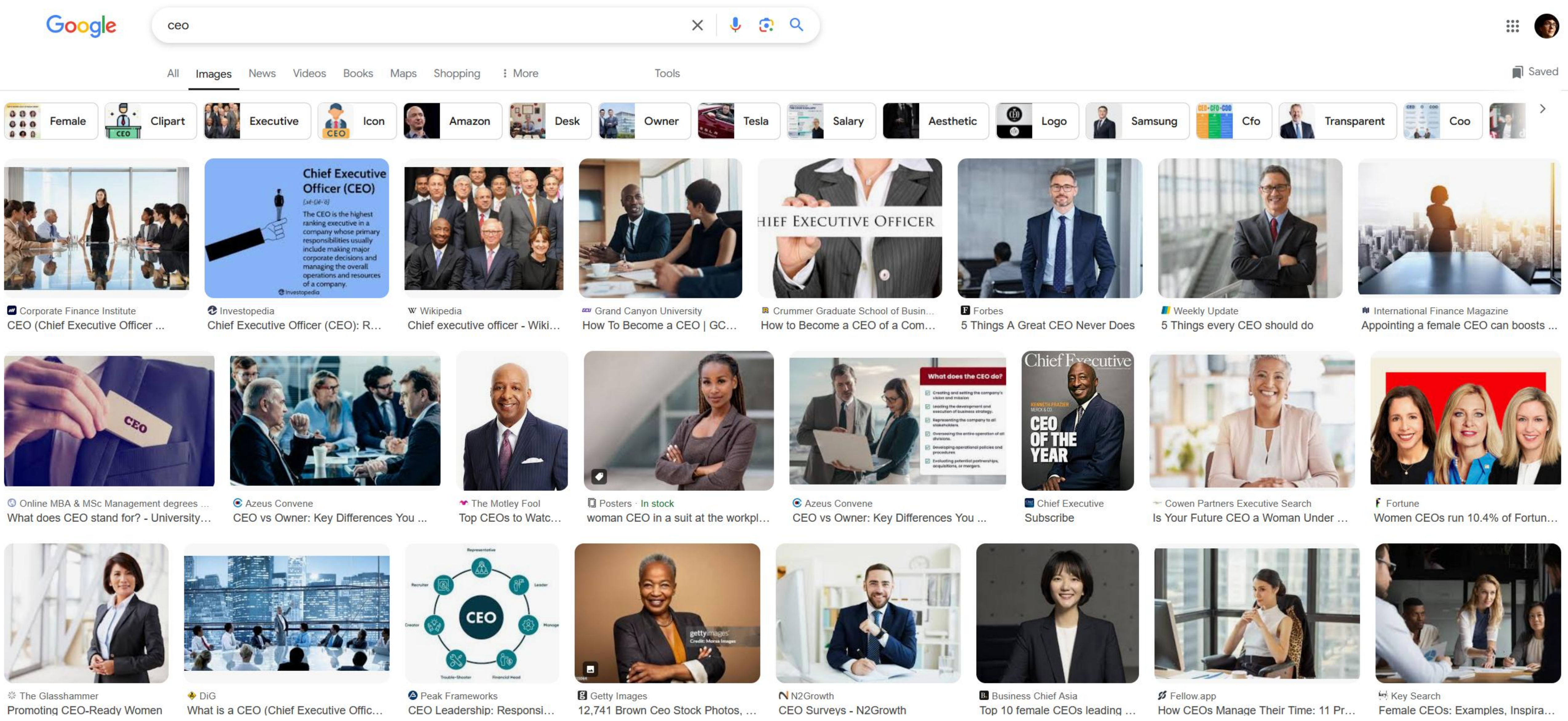


2021 results
more diverse

2024


Google
ceo
All
Images
News
Videos
Books
More
Tools
Saved
SafeSearch
business
female
clipart
office
apple
executive
google
microsoft
icon
amazon
desk
owner
tesla
salary
aesthetic
Corporate Finance Institute
CEO (Chief Executive Officer ...
Stever Robbins
What do CEOs do? A CEO Job Descrip...
Investopedia
Chief Executive Officer (CEO): What ...
Sue Rees
The Changing Role of the CEO
Weekly Update
5 Things every CEO should do
Corporate Finance Institute
CEO vs. CFO - Definitions, Differences...
Chief Executive
Subscribe
The Motley Fool
Top Black CEOs: Ex...
University of Lincoln
What does CEO stand for? - University ...
Forbes
What Makes A Great CEO?
Betterteam
CEO Job Description
beyondceocoaching....
Home - Beyond CEO...
Forbes
7 Personality Traits Every CE...
McKinsey
CEO Excellence | Strategy ...
Simply Business
What is a CEO and what does CEO stand for?
Northwest Executive Education
Key Responsibilities ...
Fortinberry Murray
Even successful women CEOs more l...
Robert Half
great chief executive ...
DiG
What is a CEO (Chief Executive Officer)?
The Lighthouse - Macquarie University
Do women make better CEOs than men ...
Azeus Convene
CEO vs Owner: Key Differences You ...
UGA research
Female CEOs face subtle bias - UGA ...
Elite Business Magazine
Amy Golding, youngest fema...

2025



Google
ceo
All
Images
News
Videos
Books
Maps
Shopping
More
Tools
Saved
Female
Clipart
Executive
Icon
Amazon
Desk
Owner
Tesla
Salary
Aesthetic
Logo
Samsung
Cfo
Transparent
Coo
Corporate Finance Institute
CEO (Chief Executive Officer ...
Investopedia
Chief Executive Officer (CEO): R...
Wikipedia
Chief executive officer - Wiki...
Grand Canyon University
How To Become a CEO | GC...
Crummer Graduate School of Busin...
How to Become a CEO of a Com...
Forbes
5 Things A Great CEO Never Does
Weekly Update
5 Things every CEO should do
International Finance Magazine
Appointing a female CEO can boosts ...
Online MBA & MSc Management degrees ...
What does CEO stand for? - University...
Azeus Convene
CEO vs Owner: Key Differences You ...
The Motley Fool
Top CEOs to Watc...
Posters · In stock
woman CEO in a suit at the workpl...
Azeus Convene
CEO vs Owner: Key Differences You ...
Chief Executive
Subscribe
Cowen Partners Executive Search
Is Your Future CEO a Woman Under ...
Fortune
Women CEOs run 10.4% of Fortun...
The Glasshammer
Promoting CEO-Ready Women
DiG
What is a CEO (Chief Executive Offic...
Peak Frameworks
CEO Leadership: Responsi...
Getty Images
12,741 Brown Ceo Stock Photos, ...
N2Growth
CEO Surveys - N2Growth
Business Chief Asia
Top 10 female CEOs leading ...
Fellow.app
How CEOs Manage Their Time: 11 Pr...
Key Search
Female CEOs: Examples, Inspira...

Image Super-Resolution

Input: Low-Resolution Face

Output: High-Resolution Face

Menon et al, "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models", CVPR 2020
Example source: https://twitter.com/Chicken3gg/status/1274314622447820801

Representational Harms

	denigration	Stereotype	Recognition	Under-representation	Ex-nomination
Image Search for “CEO” yields all white men on the first page			x		x
Google Photo mislabels black people as “gorillas”	x				
YouTube speech-to-text does not recognize women’s voices			x		x
HP Cameras’ facial recognition does not recognize Asian peoples’ faces			x	x	x
Amazon labels LGBTQ literature as ‘adult content’ and removes sales ranking		x	x		x
Word embeddings contain implicit biases [Bolukbasi et al.]	x	x	x	x	x
Searches for African-American-sounding names yields ads for criminal background checks [Sweeney, 2013]	x	x		x	

Band-Aid Solutions

- Fairness & bias are often only an afterthought.
- Leads to band-aid solutions.
- E.g.: *“let’s make everything as diverse as possible!”*

ARTIFICIAL INTELLIGENCE / TECH / WEB

Google apologizes for ‘missing the mark’ after Gemini generated racially diverse Nazis

Gemini's results for the prompt “generate a picture of a US senator from the 1800s.”

Source: https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

Representational Harms

- Representational harms often transcend the scope of technical interventions.
- Technical approaches are necessary but not sufficient.
- Complicated political and cultural factors.

Economic Bias in Visual Classifiers

Ground-Truth: Soap
Source: UK, $1890/month

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave

Ground-Truth: Soap
Source: Nepal, $288/month

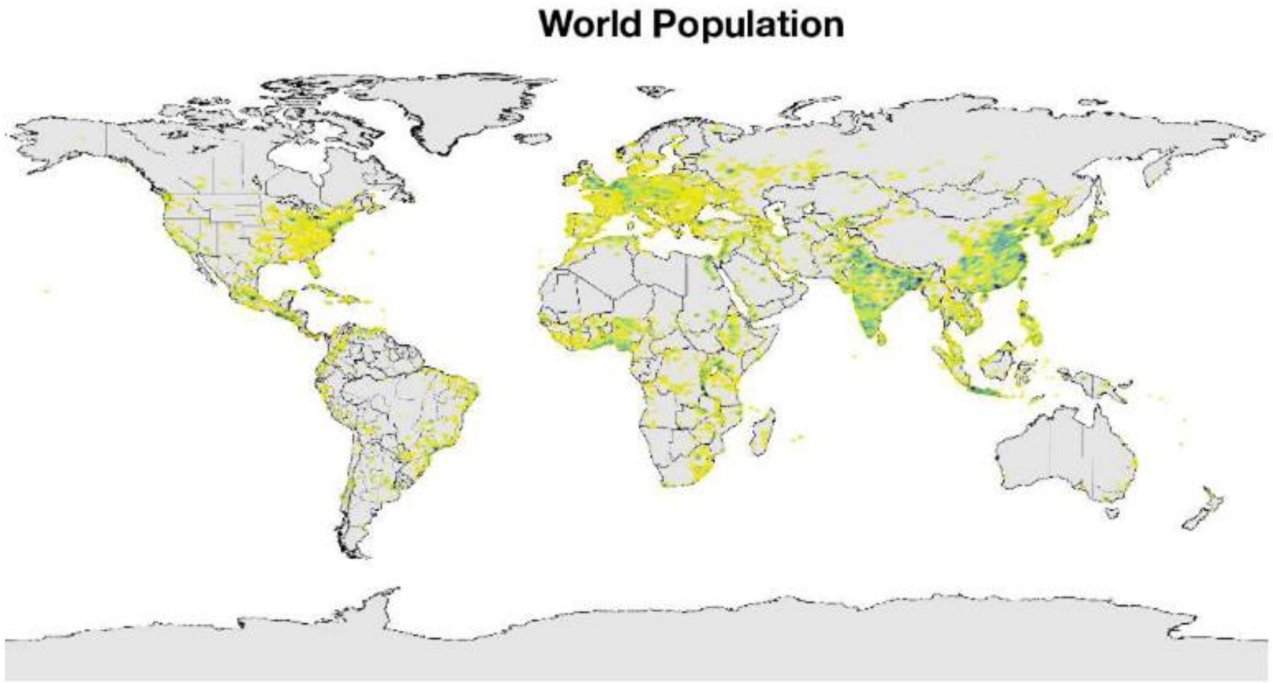
Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment

DeVries et al, "Does Object Recognition Work for Everyone?", CVPR Workshops, 2019

Data Geolocation

- More data: increased diversity?

- Strong socio-economic bias on who has the means to access and upload data to the internet.



form DeVries et al., CVPRW '19

The Data Excuse

- It is tempting to dismiss these issues "*of course the system is biased - this is just a training data issue*".

- As soon as our research affects people, this is not an excuse anymore.

- Affects people: publishing papers, open-source, used in applications, etc.!

- We cannot only benefit from the hype, we need to also deal with the consequences.

Gender Bias

- Studying gender biases is complicated post-hoc.
- Ideal: ask subjects to specify their gender.
- Many datasets have been scraped from the internet.
- Many studies currently (knowingly) conflate: binary sex, gender, perceived gender.
- Known, obvious limitations, yet can still be useful in absence of annotations.

COCO Dataset: Multi-label Classification

<u>Multilabel Classification</u>

Person
Umbrella
Cat

Define "gender bias" of object category C as:

$$\frac{\#(C, Man)}{\#(C, Man) + \#(C, Woman)}$$

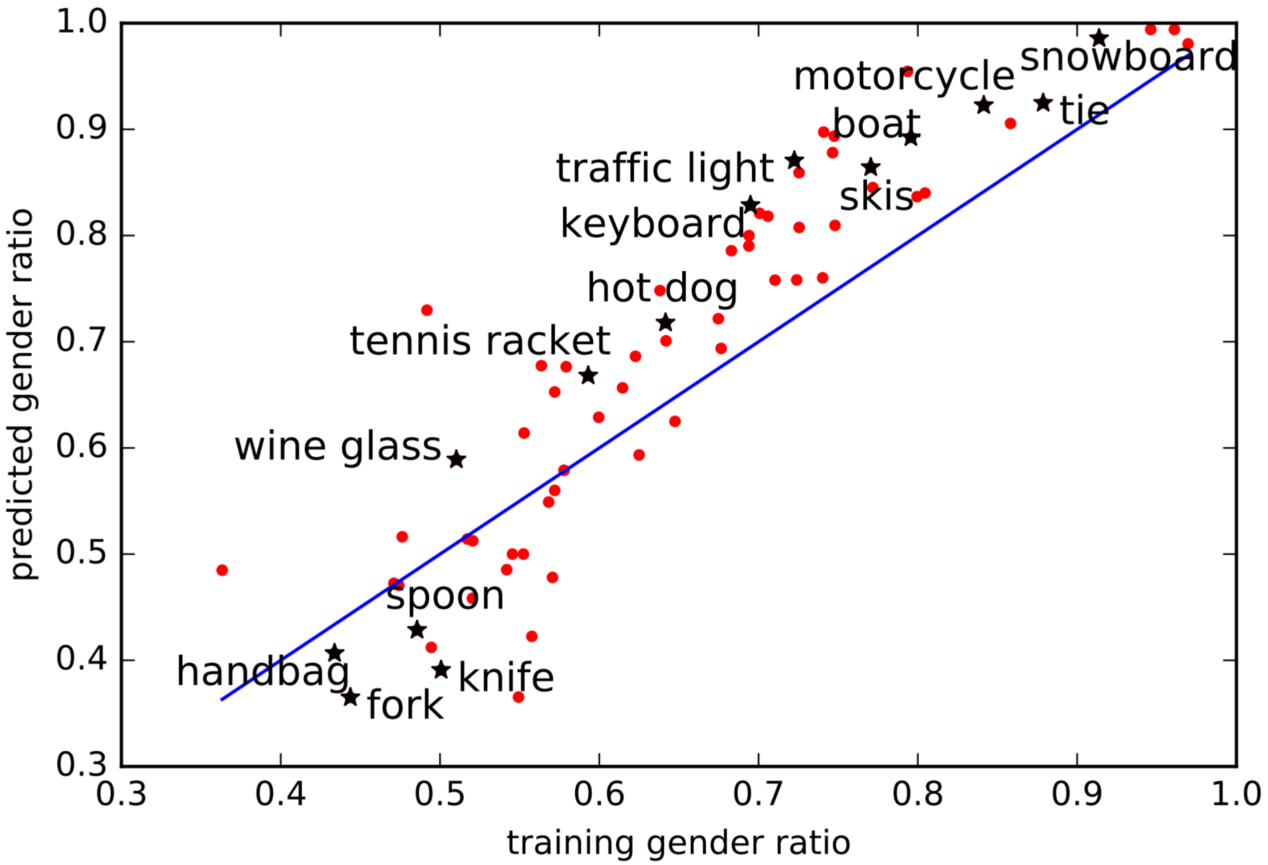
Example: "Snowboards" are 90% biased towards men

Zhao et al, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", EMNLP 2017

Problem: Bias Amplification

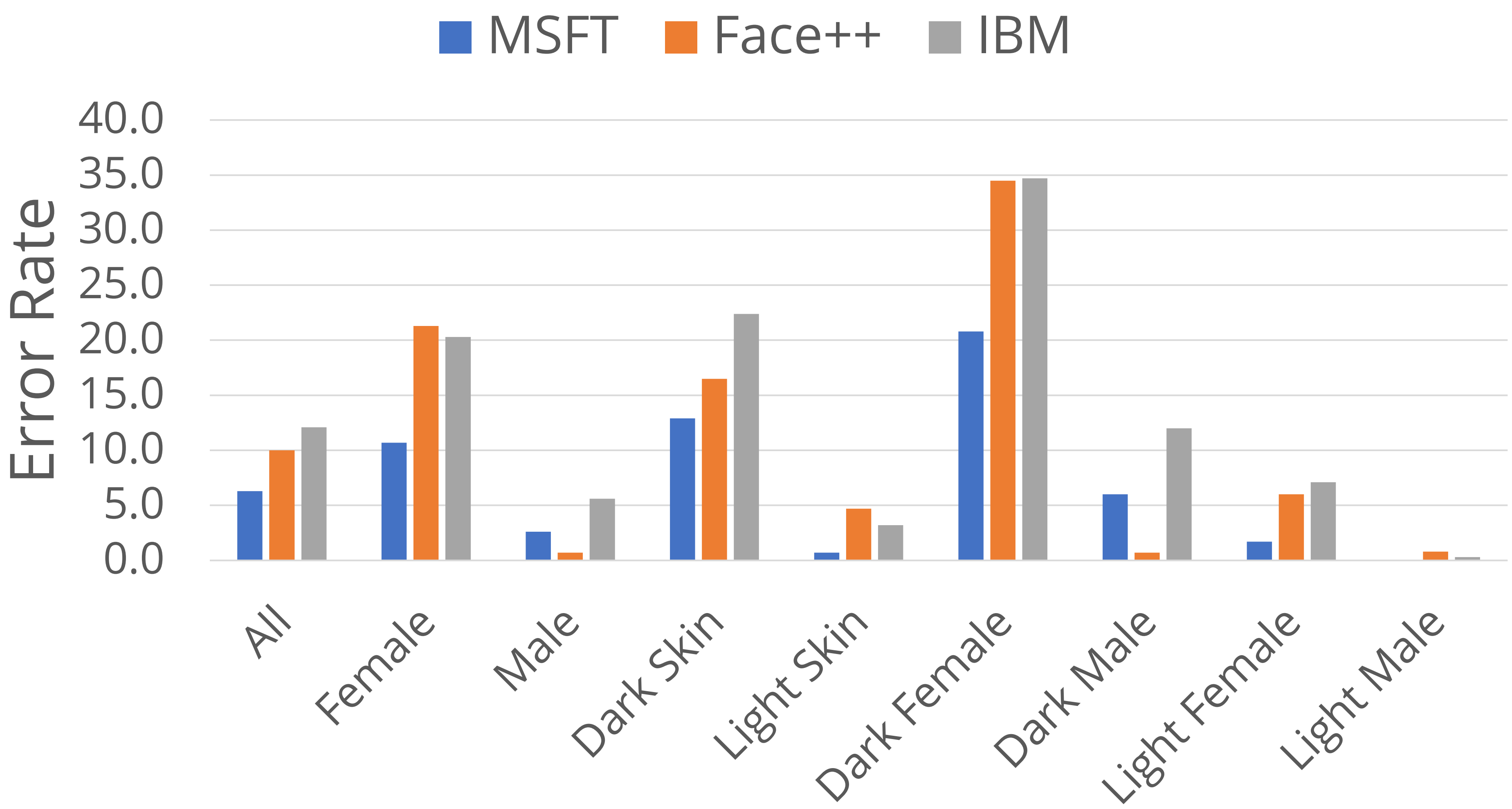
CNN predictions are **more biased** than their training data!

Reducing bias in datasets is **not enough**



Zhao et al, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", EMNLP 2017

Gender Shades: Intersectionality



Buolamwini and Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, FAT* 2018

Think Critically about Datasets

CelebA Dataset: 202k images labeled with 40 binary attributes

Liu et al, “Deep Learning Face Attributes in the Wild”, ICCV 2015

Think Critically about Datasets

5_o_Clock_Shadow
Arched_Eyebrows
Attractive
Bags_Under_Eyes
Bald
Bangs
Big_Lips
Big_Nose
Black_Hair
Blond_Hair
Blurry
Brown_Hair
Bushy_Eyebrows
Chubby
Double_Chin
Eyeglasses
Goatee
Gray_Hair
Heavy_Makeup
High_Cheekbones
Male
Mouth_Slightly_Open
Mustache
Narrow_Eyes
No_Beard
Oval_Face
Pale_Skin
Pointy_Nose
Receding_Hairline
Rosy_Cheeks
Sideburns
Smiling
Straight_Hair
Wavy_Hair
Wearing_Earrings
Wearing_Hat
Wearing_Lipstick
Wearing_Necklace
Wearing_Necktie
Young

Liu et al, "Deep Learning Face Attributes in the Wild", ICCV 2015

Think Critically about Datasets

Many attributes seem subjective. Who chose the attributes? Why? How are they defined? Who labeled the images?

5_o_Clock_Shadow
Arched_Eyebrows
Attractive
Bags_Under_Eyes
Bald
Bangs
Big_Lips
Big_Nose
Black_Hair
Blond_Hair
Blurry
Brown_Hair
Bushy_Eyebrows
Chubby

Double_Chin
Eyeglasses
Goatee
Gray_Hair
Heavy_Makeup
High_Cheekbones
Male
Mouth_Slightly_Open
Mustache
Narrow_Eyes
No_Beard
Oval_Face
Pale_Skin

Pointy_Nose
Receding_Hairline
Rosy_Cheeks
Sideburns
Smiling
Straight_Hair
Wavy_Hair
Wearing_Earrings
Wearing_Hat
Wearing_Lipstick
Wearing_Necklace
Wearing_Necktie
Young

Liu et al, “Deep Learning Face Attributes in the Wild”, ICCV 2015

Think Critically about Datasets

Almost no detail in the paper

images of $5,749$ identities. Each image in CelebA and LFWA is annotated with forty face attributes and five key points by a professional labeling company. CelebA and LFWA have over eight million and five hundred thousand attribute labels, respectively.

Datasheets for Datasets

Idea: A standard list of questions to answer when releasing a dataset. Who created it? Why? What is in it? How was it labeled?

A Database for Studying Face Recognition in Unconstrained Environments — **Labeled Faces in the Wild**

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

The dataset does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is accompanied by a label indicating the name of the person in the image.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included in the dataset.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources

Gebru et al, "Datasheets for Datasets", FAccT 2018

Model Cards

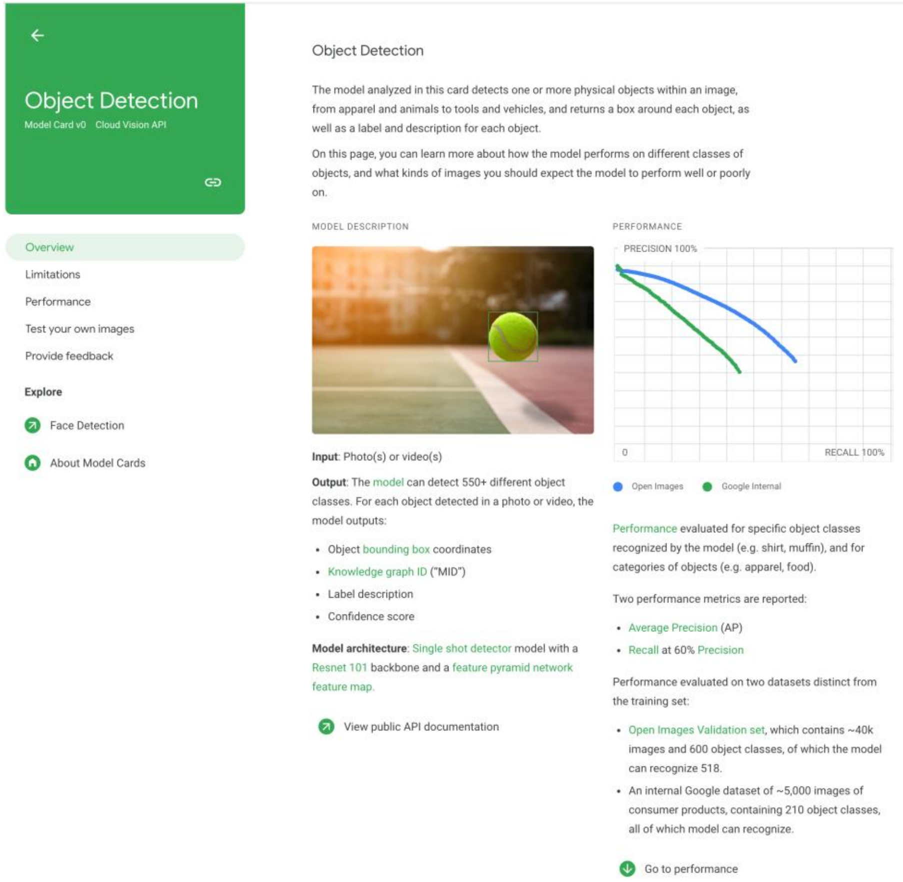
Idea: A standard list of questions to answer when releasing a trained model. Who created it? What data was it trained on? What should it be used for? What should it **not** be used for?

Model Card

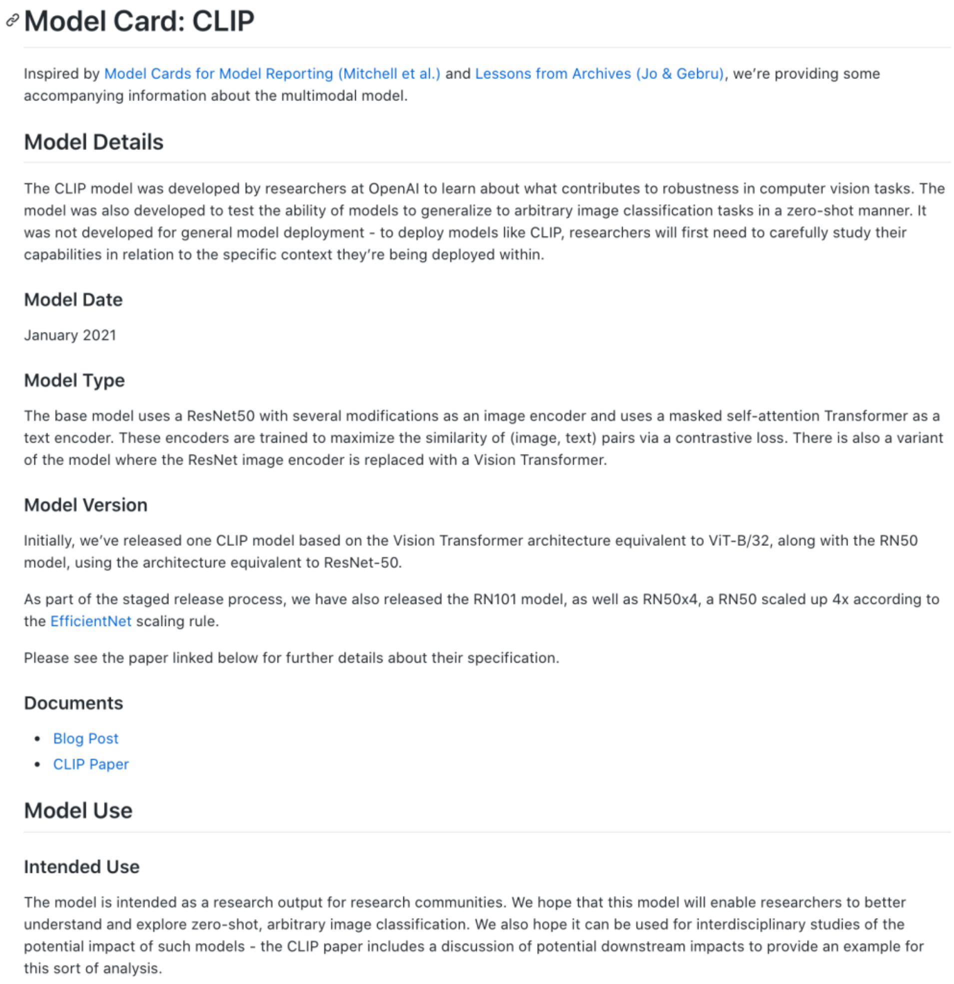
- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Mitchell et al, “Model Cards for Model Reporting”, FAccT 2019

Model Cards



https://modelcards.withgoogle.com/object-detection



Model Card: CLIP

Inspired by Model Cards for Model Reporting (Mitchell et al.) and Lessons from Archives (Jo & Gebru), we're providing some accompanying information about the multimodal model.

Model Details

The CLIP model was developed by researchers at OpenAI to learn about what contributes to robustness in computer vision tasks. The model was also developed to test the ability of models to generalize to arbitrary image classification tasks in a zero-shot manner. It was not developed for general model deployment - to deploy models like CLIP, researchers will first need to carefully study their capabilities in relation to the specific context they're being deployed within.

Model Date

January 2021

Model Type

The base model uses a ResNet50 with several modifications as an image encoder and uses a masked self-attention Transformer as a text encoder. These encoders are trained to maximize the similarity of (image, text) pairs via a contrastive loss. There is also a variant of the model where the ResNet image encoder is replaced with a Vision Transformer.

Model Version

Initially, we've released one CLIP model based on the Vision Transformer architecture equivalent to ViT-B/32, along with the RN50 model, using the architecture equivalent to ResNet-50.

As part of the staged release process, we have also released the RN101 model, as well as RN50x4, a RN50 scaled up 4x according to the EfficientNet scaling rule.

Please see the paper linked below for further details about their specification.

Documents

- Blog Post
- CLIP Paper

Model Use

Intended Use

The model is intended as a research output for research communities. We hope that this model will enable researchers to better understand and explore zero-shot, arbitrary image classification. We also hope it can be used for interdisciplinary studies of the potential impact of such models - the CLIP paper includes a discussion of potential downstream impacts to provide an example for this sort of analysis.

https://github.com/openai/CLIP/blob/main/model-card.md

Adopted by Google, OpenAI (sometimes)

Model Cards

Some models are just for research and not to be deployed. Make it clear!

Out-of-Scope Use Cases

Any deployed use case of the model - whether commercial or not - is currently out of scope. Non-deployed use cases such as image search in a constrained environment, are also not recommended unless there is thorough in-domain testing of the model with a specific, fixed class taxonomy. This is because our safety assessment demonstrated a high need for task specific testing especially given the variability of CLIP's performance with different class taxonomies. This makes untested and unconstrained deployment of the model in any use case currently potentially harmful.

Certain use cases which would fall under the domain of surveillance and facial recognition are always out-of-scope regardless of performance of the model. This is because the use of artificial intelligence for tasks such as these can be premature currently given the lack of testing norms and checks to ensure its fair use.

Model Cards

- CLIP Model Card: do not use in a deployed system.

- LAION-5B dataset: filtered with CLIP to remove “bad” images.

SAFETY REVIEW FOR LAION 5B

by: LAION.ai, 19 Dec, 2023

There have been reports in the press about the results of a research project at Stanford University, according to which the LAION training set 5B contains potentially illegal content in the form of CSAM. We would like to comment on this as follows:

LAION is a non-profit organization that provides datasets, tools and models for the advancement of machine learning research. We are committed to open public education and the environmentally safe use of resources through the reuse of existing datasets and models.

LAION datasets (more than 5.85 billion entries) are sourced from the freely available Common Crawl web index and offer only links to content on the public web, with no images. We developed and published our own rigorous filters to detect and remove illegal content from LAION datasets before releasing them.

LAION collaborates with universities, researchers and NGOs to improve these filters and are currently working with the Internet Watch Foundation (IWF) to identify and remove content suspected of violating laws. LAION invites the Stanford researchers to join its Community to improve our datasets and to develop efficient filters for detecting harmful content.

LAION has a zero tolerance policy for illegal content and in an abundance of caution, we are temporarily taking down the LAION datasets to ensure they are safe before republishing them.

Following a discussion with the Hamburg State Data Protection Commissioner, we would also like to point out that the CSAM data is data that must be deleted immediately for data protection reasons in accordance with Art. 17 GDPR.

All Cyber News / Blogs / December 20, 2023

Investigation Finds AI Image Generation Models Trained on Child Abuse

A new report identifies hundreds of instances of exploitative images of children in a public dataset used for AI text-to-image generation models.

Consent vs Copyright

- Datasets often scraped from the internet without regard for copyright or consent.
- Even if the image has a permissive copyright license, consent of the subjects is still missing!
- Many datasets are being withdrawn, taken offline.

Birhane and Prabhu, “Large Image Datasets: A Pyrrhic Win for Computer Vision?”, WACV 2021

The Dataset Crisis

BREAKING • BUSINESS

Clearview AI Fined $9.4 Million In U.K. For Illegal Facial Recognition Database

Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content

/ Getty Images claims Stability AI 'unlawfully' scraped millions of images from its site. It's a

Were your Flickr photos used in biometric surveillance research?

Enter your Flickr username, photo URL, or #tag to find out

Flickr username, #tag, or photo URL | Search

Places365 CNNs

Convolutional neural networks (CNNs) trained on the Places2 Database can be used fo generic deep scene features for visual recognition. We share the following pre-trained

- Github page for Places365-CNNs.
- List of the categories
- Scene hierarchy

Dataset is under maintenance. If you have urgent use of the dataset, please contact

What's Really Behind Those AI Art Images?

What feels like magic is actually incredibly complicated and ethically fraught.

We therefore have decided to formally withdraw the dataset. It has been taken offline and it will not be put back online. We ask the community to refrain from using it in future and also delete any existing copies of the dataset that may have been downloaded.

Not Found

The requested URL was not found on this server.

The future of datasets and models

- Datasheets for Datasets [Gebru et al, FAccT ‘18]
- Ethics checks/boards
- Ethics, limitations and social impact statements

- Synthetic Datasets [Carla, Dosovitsky, CoRL’17]
- Remove, replace and open [Asano et al., NeurIPS D&B’21]
- Obfuscate humans/faces [Yang et al., ICML’22]
- Consent [Ego4D, Graumann et al., CVPR’22]

PASS Dataset

0

Humans

1.4M

images

1.4M

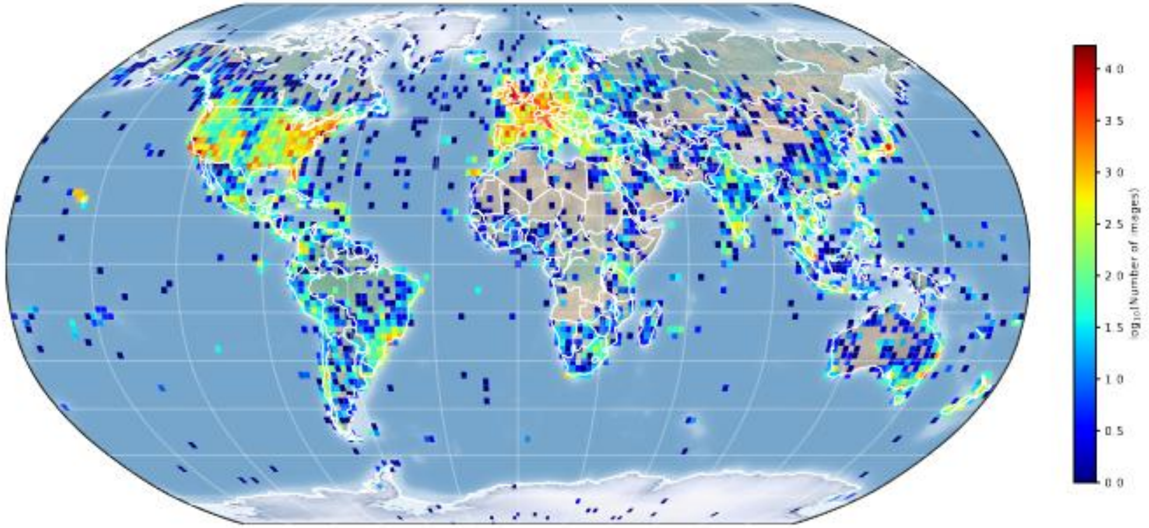
license files

Init.	Bounding-box AP	AP_{50}	AP_{75}	🧍-AP	Segmentation AP	AP_{50}	AP_{75}	🧍-AP
Random	26.4	44.0	27.8		29.3	46.9	30.8	
MoCo-v2								
on IN-1k	38.7	59.2	42.3	55.5	35.2	56.2	37.9	47.6
on IN-1k*	38.4	58.7	41.8	55.4	35.0	55.8	37.4	47.2
on Places	38.3	58.4	41.7	55.9	34.8	55.4	37.4	47.8
on PASS	38.0	58.5	41.5	55.5	34.7	55.4	37.1	47.5

Init.	Dense-pose AP^{dp}_{GPS}	AP^{dp}_{GPSm}	B-box AP^{bb}	Seg. AP^{sg}
Random	*—does not train—*			
MoCo-v2				
on IN-1k	65.0	66.3	61.7	67.6
on IN-1k*	64.8	66.1	61.4	67.0
on Places	64.9	66.0	61.8	67.1
on PASS	64.9	65.7	61.5	66.8

[Asano et al.; NeurIPS Datasets&Benchmarks '21]

PASS Dataset



Appendix

Table of Contents

Prompt Engineering

Specially designed input that elicits a desirable response

Research Assistant/AI Prompt Engineer
ENVIROGEN WATER TECHNOLOGIES LIMIT…
Alfreton
via Jobrapido.com
19 hours ago Full–time

Research Assistant/AI Prompt Engineer
ENVIROGEN WATER TECHNOLOGIES LIMIT…
Alfreton
via Totaljobs
8 days ago £18K–£22K a year

Prompt Engineer for AI Singapore (Makerspace)
National University of Singapore
Kent
via Jooble
10 days ago Full–time

Prompt Engineer
Team Rehab
London
via Glassdoor
5 days ago Full–time

Prompt Engineer - Manchester
ChatGPT Consultancy
Manchester
via Prompt Engineering Jobs
22 days ago Contractor

ChatGPT & Bard Prompt Engineer
Datasumi
London
via Prompt Engineering Jobs
13 days ago Full–time

GPT Legal Prompt Engineer
Mishcon de Reya Group
London
via Prompt Engineering Jobs
28 days ago Full–time

Prompt Engineer
vidIQ
Anywhere
via App.otta.com
11 days ago Work from home Full–time

Prompt Engineering

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 **X**

T. Kojima et. al 2022 Large Language Models are Zero-Shot Reasoners

Prompt Engineering

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: ***Let's think step by step.***

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.* ✓

1. Kojima et. al 2022 Large Language Models are Zero-Shot Reasoners

Rob Borovsky, Cater news

Hosted inference API

Zero-Shot Image Classification

Examples

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.716 s

tree 0.443

https://huggingface.co/openai/clip-vit-large-patch14

Rob Borovsky, Cater news

Hosted inference API

Zero-Shot Image Classification

Examples

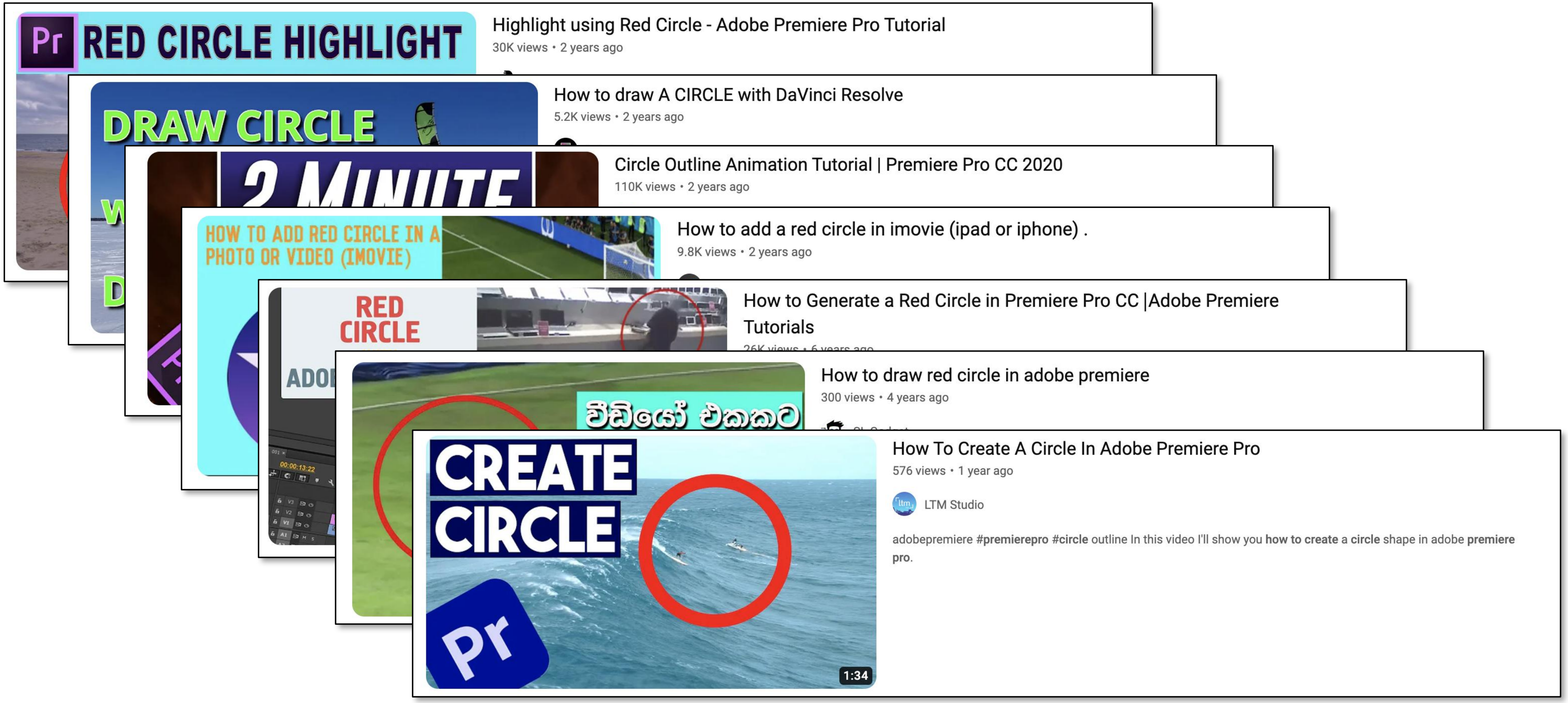
Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.871 s

owl 0.925

https://huggingface.co/openai/clip-vit-large-patch14

Annotating with a red circle



r/usefulredcircle

35,303 members • 11 online

Join

Videos, images, and GIFs of useful instances of red circles.

u/MesopotamiaSong • 4y • i.redd.it

Wow

1822 22 Share

r/uselessredcircle

179,716 members • 24 online

Join

For images or videos where something obvious got highlighted with a red circle or outline.

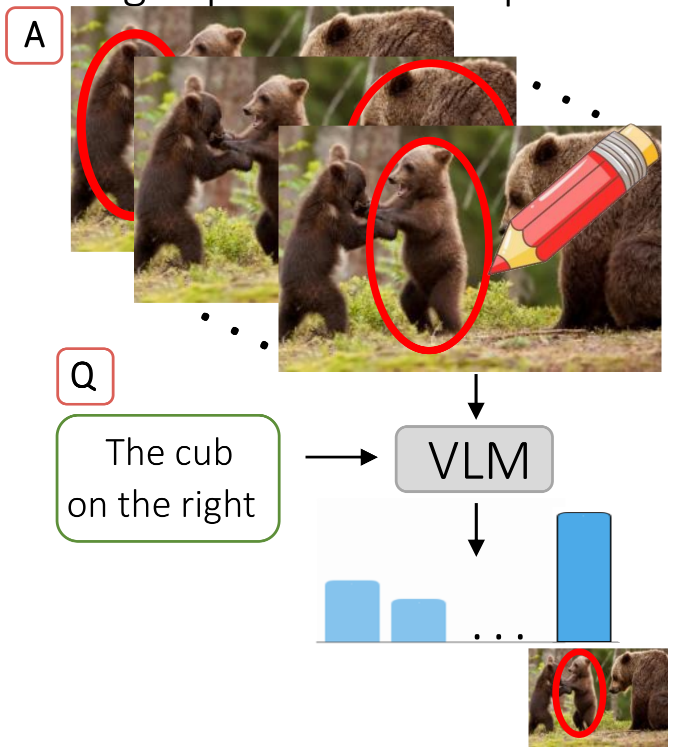
u/dark_night01 • 215d • i.redd.it

No fu[redacted] way

1566 18 Share

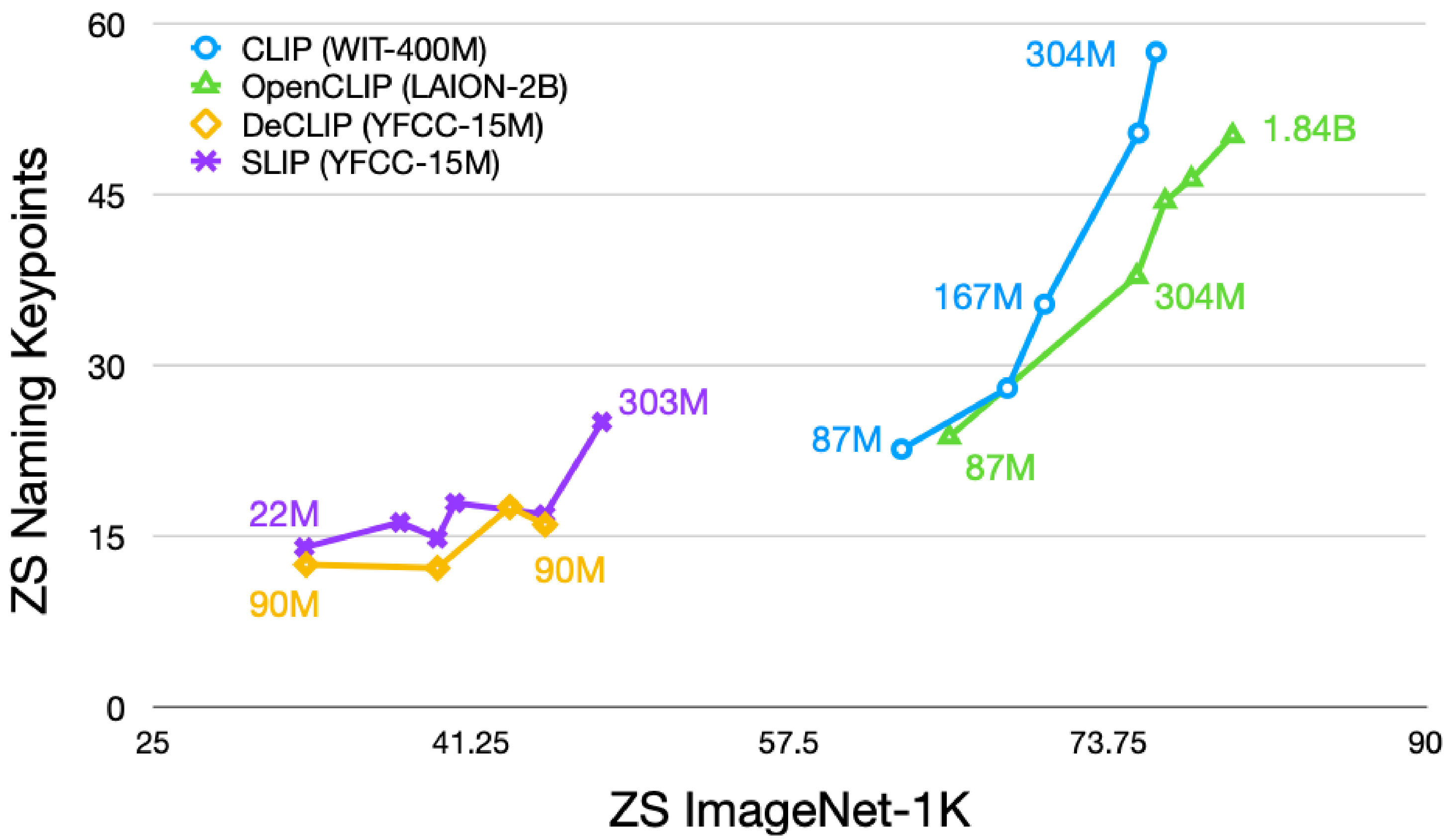
Using VLMs for zero-shot inference

Referring expressions comprehension



- Generate images with a circle in different locations
- Observation: adding a red circles steers the global descriptor to the annotated region
- Choose the image with the highest correlation to the text

Model Size



Model size only matters when trained on very large datasets

Bias Considerations

Ranking 4 classes – man, woman, missing person, murderer

1. woman
2. man
3. missing person
4. murderer

1. murderer
2. missing person
3. man
4. woman

1. missing person
2. woman
3. murderer
4. man

Bias from (unknown!) training data reflected in the model

Keeping Bias in Mind

When building a system, ask yourself

- who will benefit and
- who will be harmed.

Act accordingly, be transparent, be clear with limitations.

Thanks!