Trustworthy Computer Vision

Dr. Julio C. S. Jacques Junior
julio.silveira@ub.edu

Summary

- Trustworthy computer vision
  - Interpretability and Explainability
  - Fairness in Computer Vision
  - Privacy and Ethics

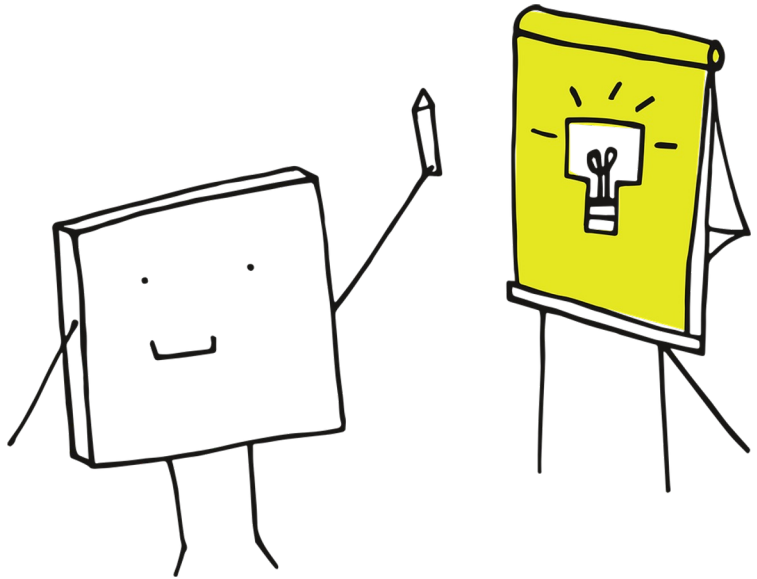


Image source: https://pixabay.com

Trustworthy AI

- Trustworthy AI refers to the use of Artificial Intelligence in a way that it is **reliable**, **transparent**, and **respects ethical principles** such as privacy and non-discrimination.
- Trustworthy AI seeks to ensure that AI systems are designed and **used for good**, in a way that they are **fair**, **accountable**, and **transparent**, also taking into account their potential impacts on society.

    - Additional Information:
        - Ethics Guidelines for Trustworthy AI:

          https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Trustworthy AI or Responsible AI?

- Trustworthy AI
  - Often used to reference the technical implementation of AI.
  - E.g., ensuring **fairness** through the detection and mitigation of bias
  - E.g., ensuring AI models are **transparent** and **explainable**.

- Responsible AI
  - Actions taken to mitigate harm to people and the planet.

Both relate to the principles behind the **design**, **development**, and **implementation** of AI systems, in a manner that benefits individuals, society, and businesses while reinforcing **human centricity and societal value**.

image source: https://www.vecteezy.com

Introduction

- Current machine learning models are obtaining **impressive results** due to the recent advances in artificial intelligence (AI), having a huge impact on real world applications;
- Some systems are obtaining **better than human performance** in a wide number of tasks;
  - E.g., Image and object recognition, playing video-games, voice generation / recognition.

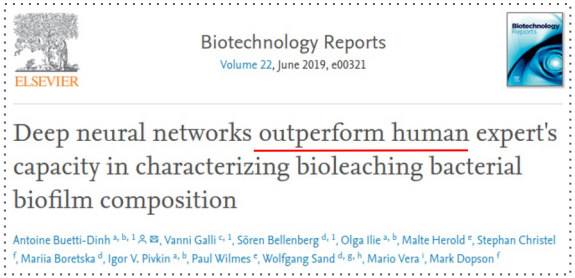

nature > news > article

NEWS | 30 October 2019

Google AI beats top human players at strategy game *StarCraft II*

DeepMind's AlphaStar beat all but the very best humans at the fast-paced sci-fi video game.

Dan Garisto


ELSEVIER

Biotechnology Reports

Volume 22, June 2019, e00321

Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition

Antoine Buetti-Dinh [a, b, 1], Vanni Galli [c, 1], Sören Bellenberg [d, 1], Olga Ilie [a, b], Malte Herold [e], Stephan Christel [f], Mariia Boretska [d], Igor V. Pivkin [a, b], Paul Wilmes [e], Wolfgang Sand [d, g, h], Mario Vera [i], Mark Dopson [f]


NEWS AND VIEWS | 09 June 2021

AI system outperforms humans in designing floorplans for microchips

A machine-learning system has been trained to place memory blocks in microchip designs. The system beats human experts at the task, and offers the promise of better, more-rapidly produced chip designs than are currently possible.

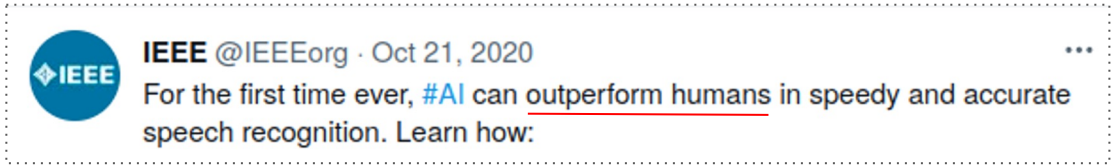
Andrew B. Kahng


nature > nature communications > articles > article

Article | Open Access | Published: 24 February 2021

Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images

Wenying Zhou, Yang Yang, Cheng Yu, Juxian Liu, Xingxing Duan, Zongjie Weng, Dan Chen, Qianhong Liang, Qin Fang, Jiaojiao Zhou, Hao Ju, Zhenhua Luo, Weihao Guo, Xiaoyan Ma, Xiaoyan Xie, Ruixuan Wang & Luyao Zhou


IEEE @IEEEorg · Oct 21, 2020

For the first time ever, #AI can outperform humans in speedy and accurate speech recognition. Learn how:

The need of Explainable AI

- Such shift in performance, has been achieved through increased model complexity, turning such systems into **“black box”** approaches and causing uncertainty regarding the way they operate and come to decisions;

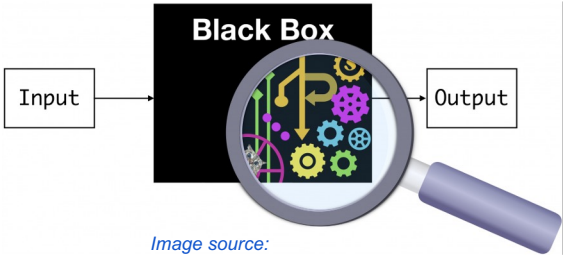


Image source:
https://towardsdatascience.com/black-boxes-and-their-intrusion-620aa3c4c56b

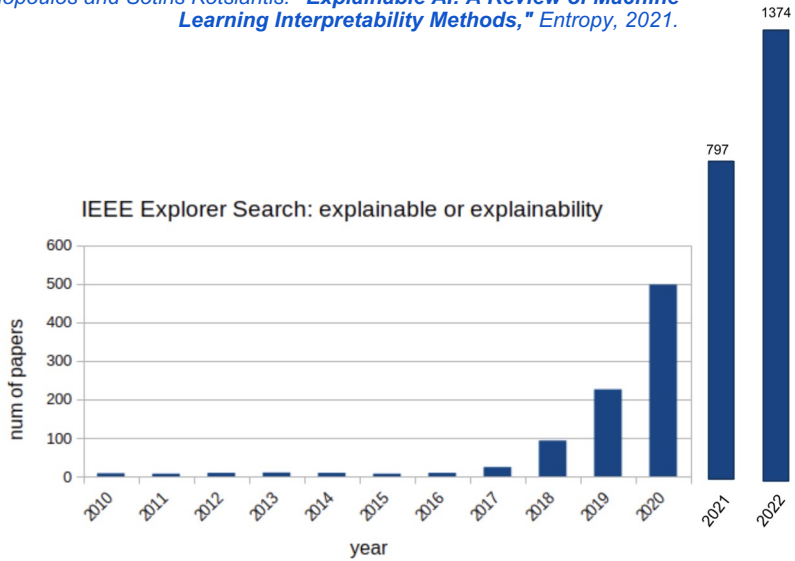
- This imposes a **strong limitation** for machine learning systems to be adopted in sensitive yet critical domains, such as healthcare, autonomous driving or education.
- **Models must be trusted**, and one way to achieve trustworthy is through interpretability and explainability.

Pantelis Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsiantis. ***"Explainable AI: A Review of Machine Learning Interpretability Methods,"*** *Entropy, 2021.*

The need of Explainable AI

- As a result, the interest on **Explainable AI**, a field that is concerned with the development of new methods that can explain or support the interpretability of such “black boxes”, has increased significantly over the past few years.

Pantelis Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsiantis. ***"Explainable AI: A Review of Machine Learning Interpretability Methods,"*** *Entropy, 2021.*



What contributed to such tremendous shift in performance?

- Revolution of **Deep Learning** methods
- Easy access to vast amounts of **data**

- Increased computational power (GPUs) and “user friendly” frameworks

- Popularity of Benchmarks and Competitions

The Deep Learning Revolution (ImageNet error history)

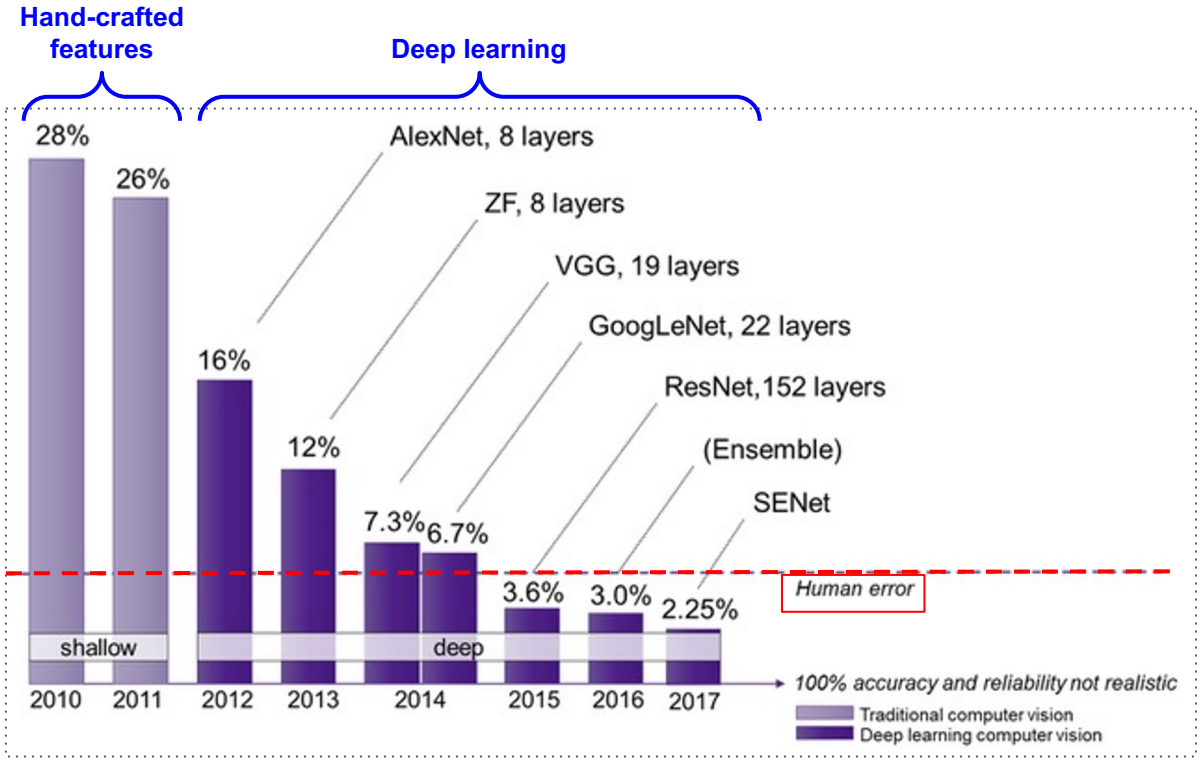


Image source: https://machinelearningknowledge.ai/keras-implementation-of-resnet-50-architecture-from-scratch/

The Deep Learning Paradigm

- Beyond the outstanding results deep learning models may have in a wide variety of tasks, **the big advantage** is that they can automatically discover, learn, and extract the *hierarchical data representations* that are needed for modeling different kind of problems.
- This hierarchy of **increasing complexity** combined with the fact that **vast amounts of data are used** to train and develop such complex systems can inherently reduce their ability to explain their inner workings and mechanisms.
- **As a consequence**, the reason behind their decisions becomes quite **hard to understand** and, therefore, their predictions **hard to interpret**.

There is a clear trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions.

Pantelis Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsiantis. ***"Explainable AI: A Review of Machine Learning Interpretability Methods,"*** *Entropy, 2021.*

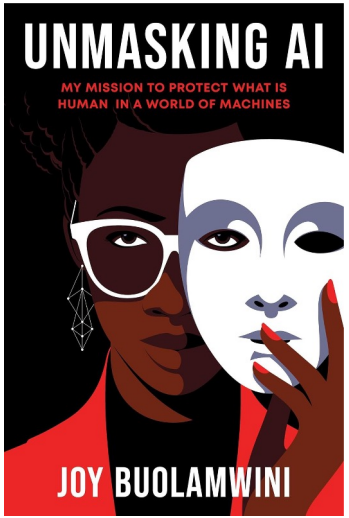
Thrusted AI

- Systems whose decisions cannot be well-interpreted are **difficult to be trusted**.
- The **need of trustworthy**, fairness, robustness and high performing models for real-world applications resulted on the revival of **Explainable AI**;
- **Before** this emergent interest on explainable models, most research have been focusing on the predictive power of algorithms rather than the understanding behind these predictions.

Thrusted AI

- Just being very accurate may not be enough, as **you may want to understand how the outcomes of your model are being generated**.

- For instance, imagine you have a model that obtained 95% of accuracy on a particular task (e.g. face detection). The question is, **can you explain what happened on the cases it didn't work?**



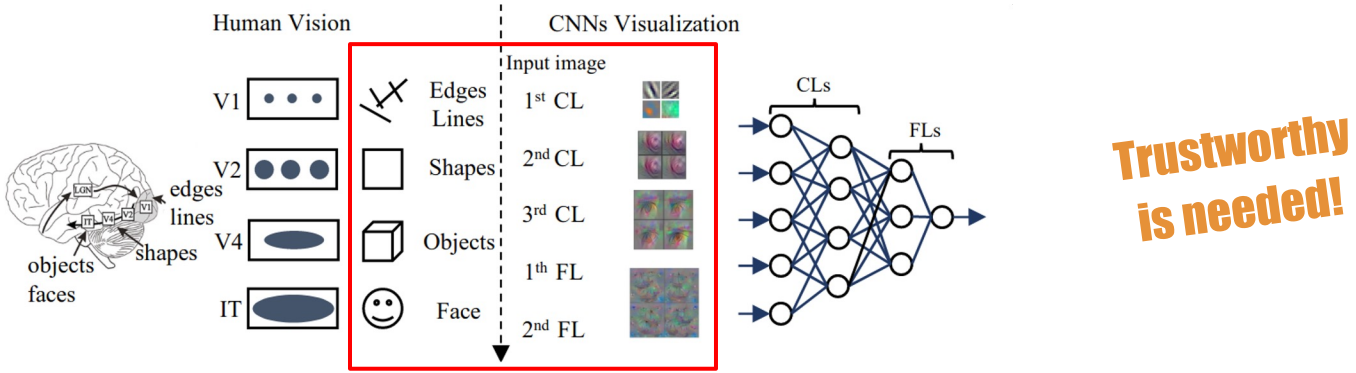
https://www.unmasking.ai/

First attempts to explain how CNNs works

- Past works proposed to **visualize filter activations** to understand how CNN works. They revealed that:

*Zeiler and Fergus, **"Visualizing and Understanding Convolutional Networks,"** ECCV, 2014*

  - Basic visual features, such as **edges and lines** are learned in the first layers
  - Whereas more complex structures, such as **shapes and objects**, are learned in deeper layers



Trustworthy is needed!

*Qin et al. **"How convolutional neural networks see the world: A survey of convolutional neural network visualization methods,"** Mathematical Foundations of Computing, 2018.*

Explainability and Interpretability

- Two critical aspects of decision support systems
  - While these terms are very closely related, some works identify their differences
- **Explainability**
  - *“focuses on mechanisms that can tell what is the explanation behind the decision or recommendation made by the model”* (Escalante et al. TAC’2020)
- **Interpretability**
  - *“focuses on revealing which part(s) of the model structure influences its recommendations”* (*Escalante et al. TAC’2020*)
  - *“the degree to which a human can understand the cause of a decision”* (Miller, AI’2019)

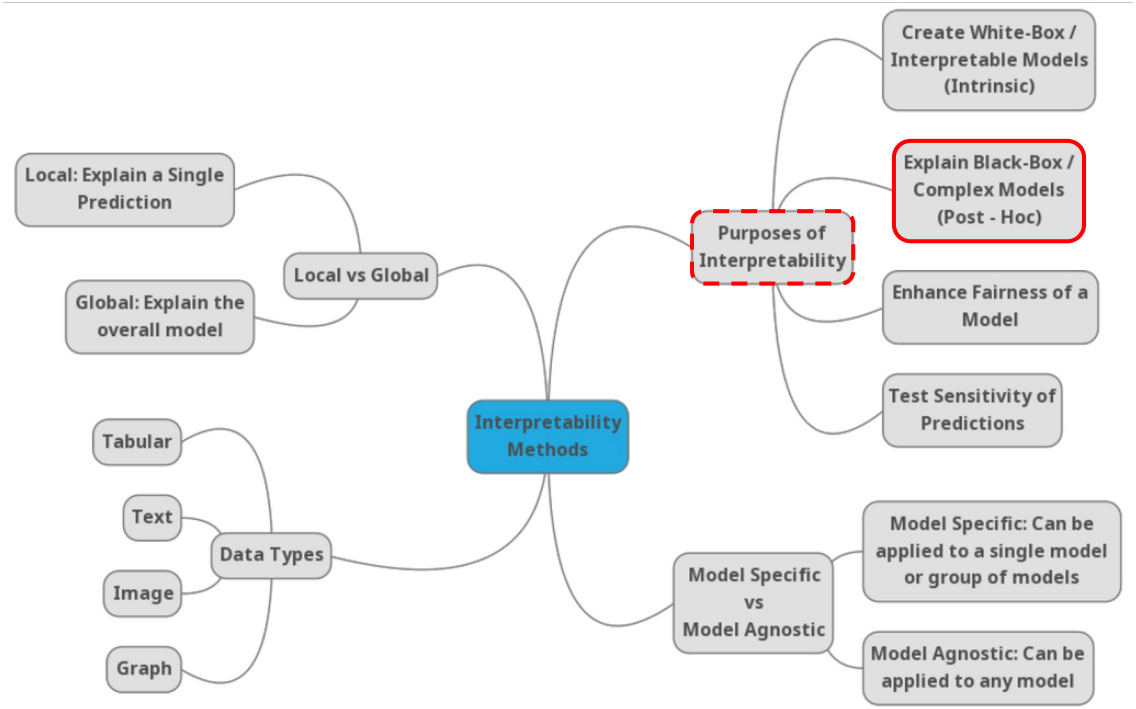
Escalante et al., ***"Modeling, Recognizing, and Explaining Apparent Personality from Videos,"*** *IEEE Transactions on Affective Computing, 2020.*
Miller, T. ***“Explanation in artificial intelligence: Insights from the social sciences,”*** *in Artif. Intell. 2019.*

Explainability and Interpretability

- Both aspects are decisive when applications can have serious implications, such as in health care, security and education scenarios.
- The more explainable a model, **the deeper the understanding** one can achieve in terms of the internal procedures that take place while the model is training or making decision
- The more interpretable a machine learning system is, **the easier it is to identify cause-and-effect relationships** within the system's inputs and outputs.

Interpretability does not naturally lead to explainability, or vice versa.

Machine Learning Interpretability Techniques



Pantelis Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsiantis. ***"Explainable AI: A Review of Machine Learning Interpretability Methods,"*** *Entropy, 2021.*

Interpretability Methods to Explain Deep Learning Models

“such methods do not try to create interpretable models, but, instead, try to interpret already trained, often complex models, such as deep neural networks”

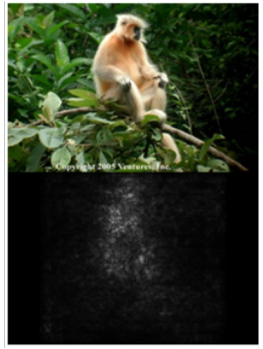
(Linardatos et al. Entropy’2021)

Interpretability Methods to Explain Deep Learning Models

- *A substantial portion of **"attention"** tools is focused on deep learning for images (e.g., classification tasks), more specifically on the concept of **saliency maps***
  - *Saliency refers to unique features, such as pixels or regions in the context of visual processing*
  - *They can give us **insights** into failure modes of analysed models*
    - *Salience maps (e.g., gradient-based attribution methods)*
    - *Class Activation Maps (CAM)*
    - *Gradient weighted Class Activation Maps (Grad-CAM)*

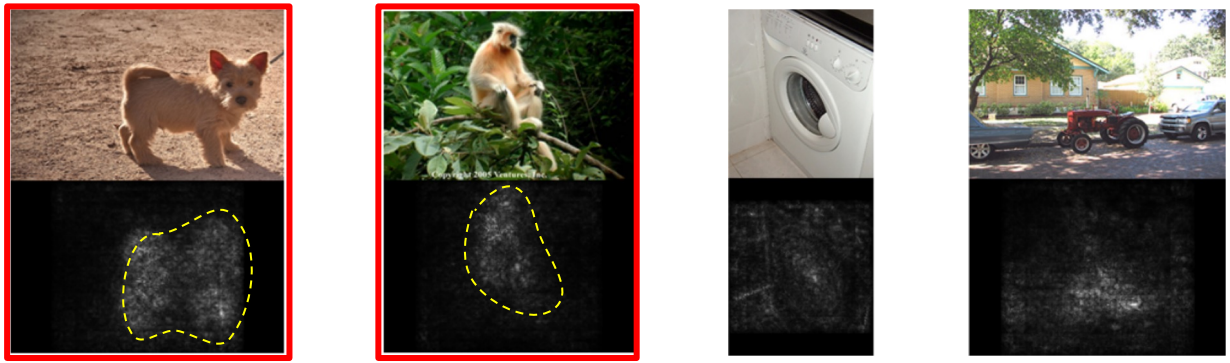
Salience Maps

- First proposed in *Simonyan et al.* ***"Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,"*** *arXiv, 2014*
- The method computes an image-specific **class saliency map** corresponding to the gradient of an output neuron with respect to the input, highlighting the areas of the given image, discriminative with respect to the given class.
    - highlight fine-grained details in the image, but are **in general not class-discriminative**

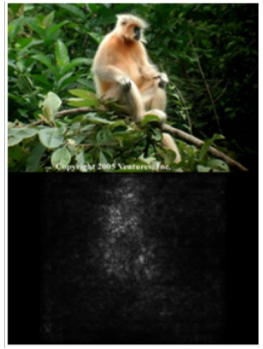
Salience Maps

- First proposed in *Simonyan et al.* ***"Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,"*** *arXiv, 2014*
- The method computes an image-specific **class saliency map** corresponding to the gradient of an output neuron with respect to the input, highlighting the areas of the given image, discriminative with respect to the given class.
    - highlight fine-grained details in the image, but are in general not class-discriminative



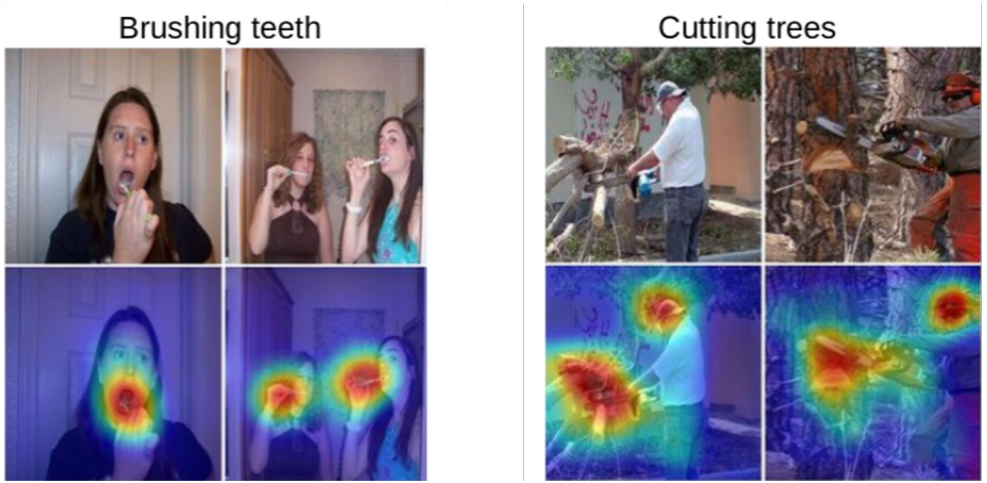
Salience Maps

- First proposed in *Simonyan et al.* ***"Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,"*** *arXiv, 2014*
- The method computes an image-specific **class saliency map** corresponding to the gradient of an output neuron with respect to the input, highlighting the areas of the given image, discriminative with respect to the given class.
    - highlight fine-grained details in the image, but are in general not class-discriminative

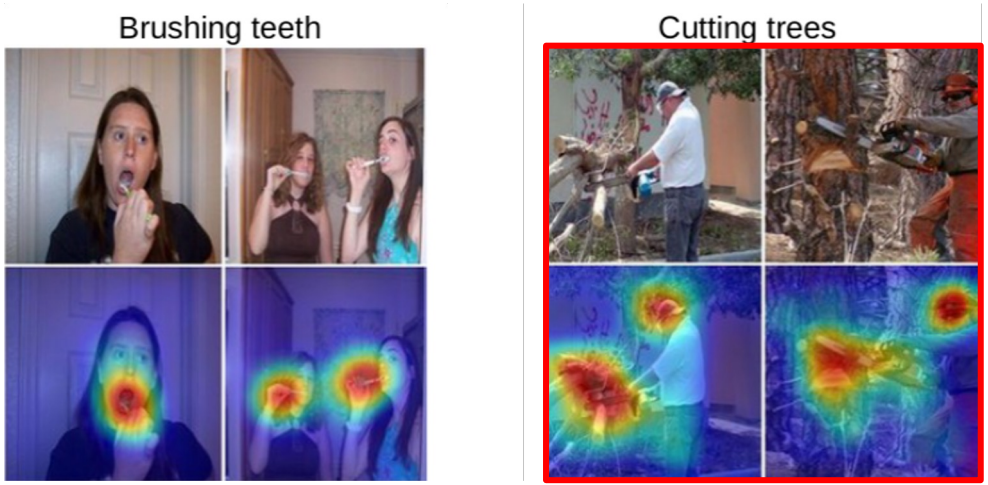
Class Activation Maps (CAMs)

- First introduced in *B. Zhou et al.* ***"Learning Deep Features for Discriminative Localization,"*** *in CVPR, 2016*
- Able to localize the discriminative image regions on a variety of tasks while not being trained for them.



Class Activation Maps (CAMs)

- First introduced in *B. Zhou et al.* ***"Learning Deep Features for Discriminative Localization,"*** *in CVPR, 2016*
- Able to localize the discriminative image regions on a variety of tasks while not being trained for them.



Class Activation Maps (CAMs)

- A feature vector is created from the **Global Average Pooling (GAP)**, just before the output layer. Then, a weighted sum of this vector is fed to the final softmax loss layer.
- Discriminative image regions can be identified by **projecting back the weights** of the output layer **on the convolutional feature maps**

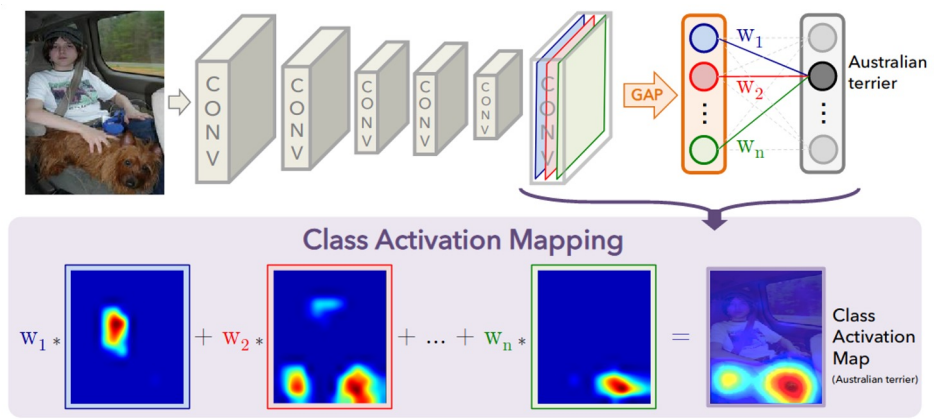
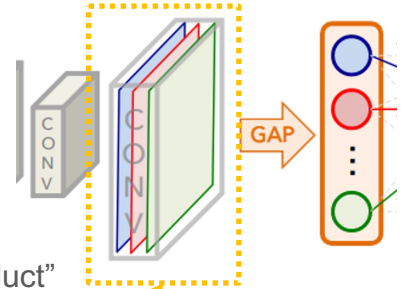


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Class Activation Maps (CAMs)

- Suppose we are looking at ResNet
- Last few layers of ResNet
  - w = W[:,some_cat_index] → “size 2048”
  - F = 2048 x 7 x 7
  - 2048 weights for 1000 classes
  - C = F[0]*w[0] + F[1]*w[1] + … + F[2047]*w[2047] → “dot product”
  - Result is a 7x7 heat map



activation_49 (Activation)	(None, 7, 7, 2048)	0	add_16[0][0]
avg_pool (AveragePooling2D)	(None, 1, 1, 2048)	0	activation_49[0][0]
flatten_1 (Flatten)	(None, 2048)	0	avg_pool[0][0]
fc1000 (Dense)	(None, 1000)	2049000	flatten_1[0][0]

Class Activation Maps (CAMs)

- Final step: rescale the 7x7 image to the original image size (e.g., 224x224, of ResNet), and generate a visualization where both the class activation map and the input image are combined.

Class Activation Maps (CAMs) - **Drawbacks** & Results

- It requires that neural networks have a very specific structure in their final layers;
  - If not, the structure needs to be **changed** and the network needs to be **re-trained**.
- CAMs are useful for interpreting the very last stages of the network's image classification, but are **unable to provide any insight into the previous stages**.

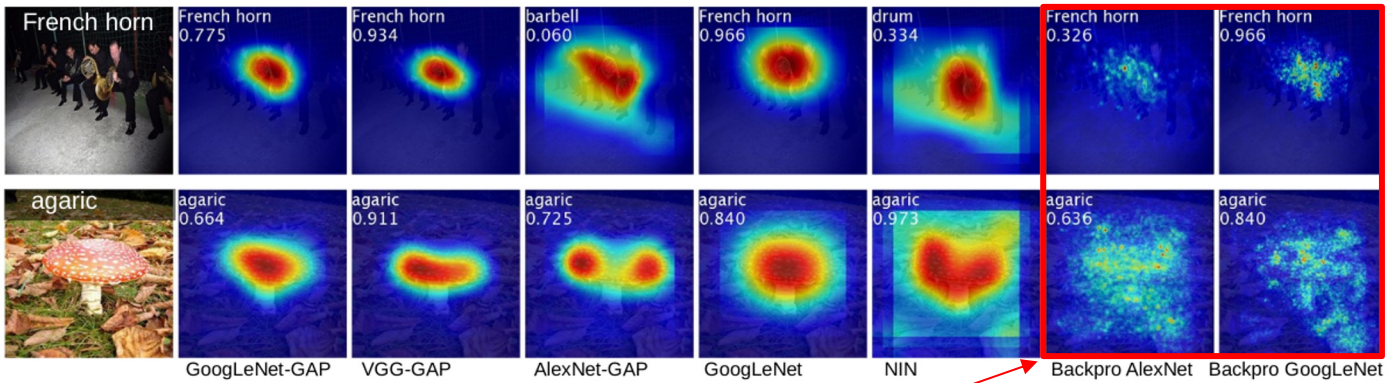


Figure 5. Class activation maps from CNN-GAPs and the class-specific saliency map from the backpropagation methods.

Class Activation Maps (CAMs): **Application Case**

- Explaining First Impressions with CAMs in *C. Ventura et al.* ***"Interpreting CNN Models for Apparent Personality Trait Regression,"*** *CVPRW, 2017.*

Figure 5. Discriminative localization (Class Activation Maps) obtained for the 20 images that give the highest predicted value for the agreeableness personality trait in the test subset.

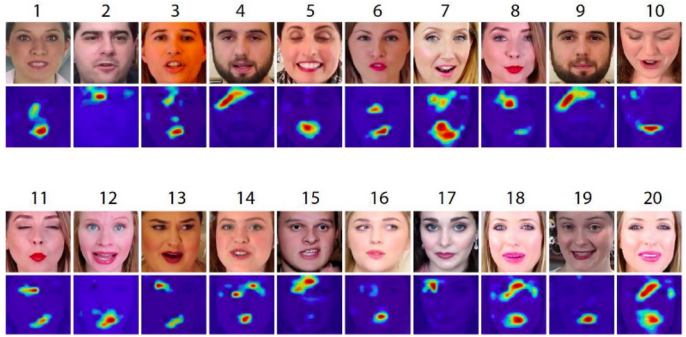


Figure 6. Discriminative localization (class activation maps) obtained for the 20 images that give the highest predicted value for the agreeableness personality trait in the test subset.

Class Activation Maps (CAMs): **Application Case**

- The network **focused on the face regions** to discriminate among the different traits.

Input: the whole image

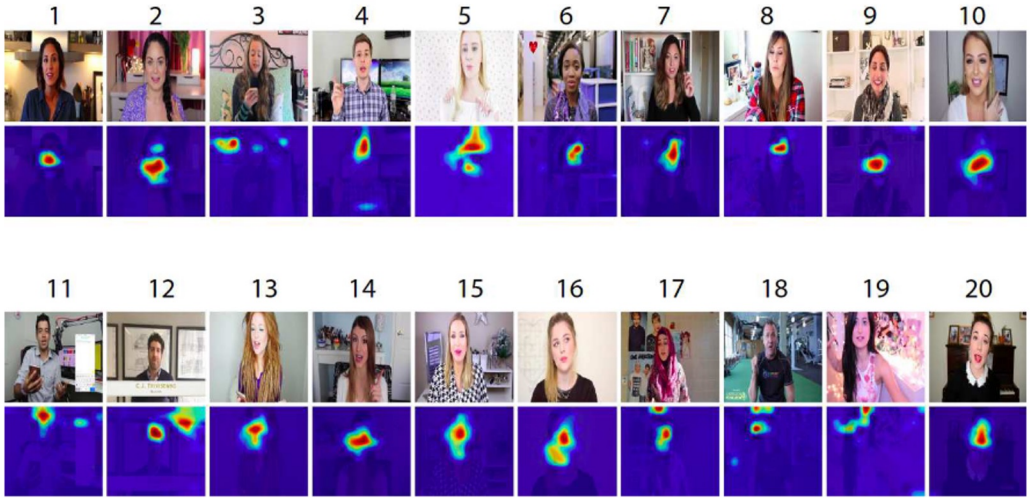


Figure 5. Discriminative localization (Class Activation Maps) obtained for the 20 images that give the highest predicted value for the agreeableness personality trait in the test subset.

Class Activation Maps (CAMs): **Application Case**

- When focusing on the face region, **facial part detectors** automatically emerged from last layers with no supervision provided on this task.

Input: the face region

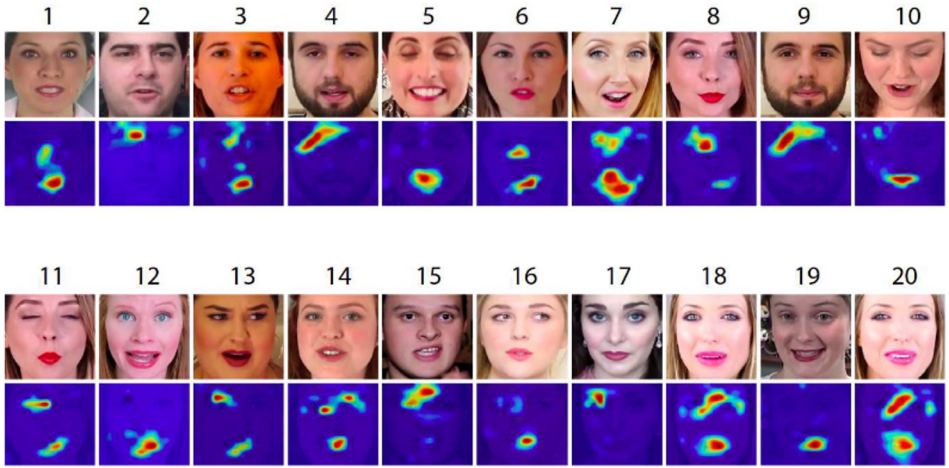


Figure 6. Discriminative localization (class activation maps) obtained for the 20 images that give the highest predicted value for the agreeableness personality trait in the test subset.

Class Activation Maps (CAMs): **Application Case**

- These are very interesting findings and can **support a better understanding of the problem being modelled** and the adopted solution.

Figure 5. Discriminative localization (Class Activation Maps) obtained for the 20 images that give the highest predicted value for the agreeableness personality trait in the test subset.

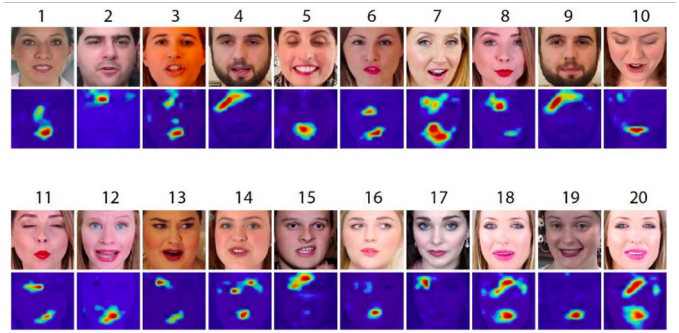
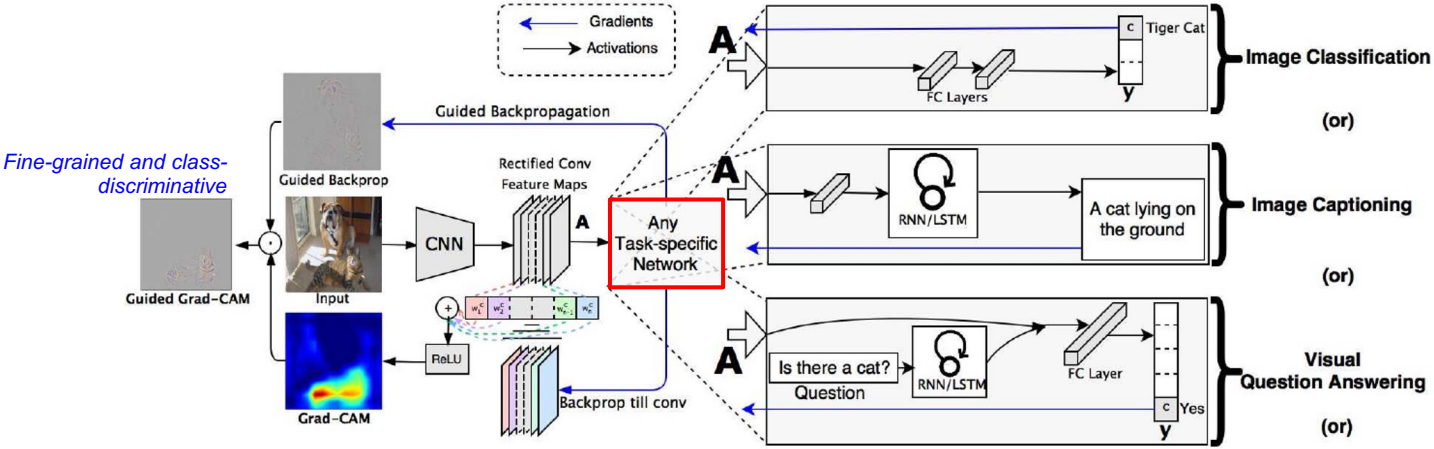


Figure 6. Discriminative localization (class activation maps) obtained for the 20 images that give the highest predicted value for the agreeableness personality trait in the test subset.

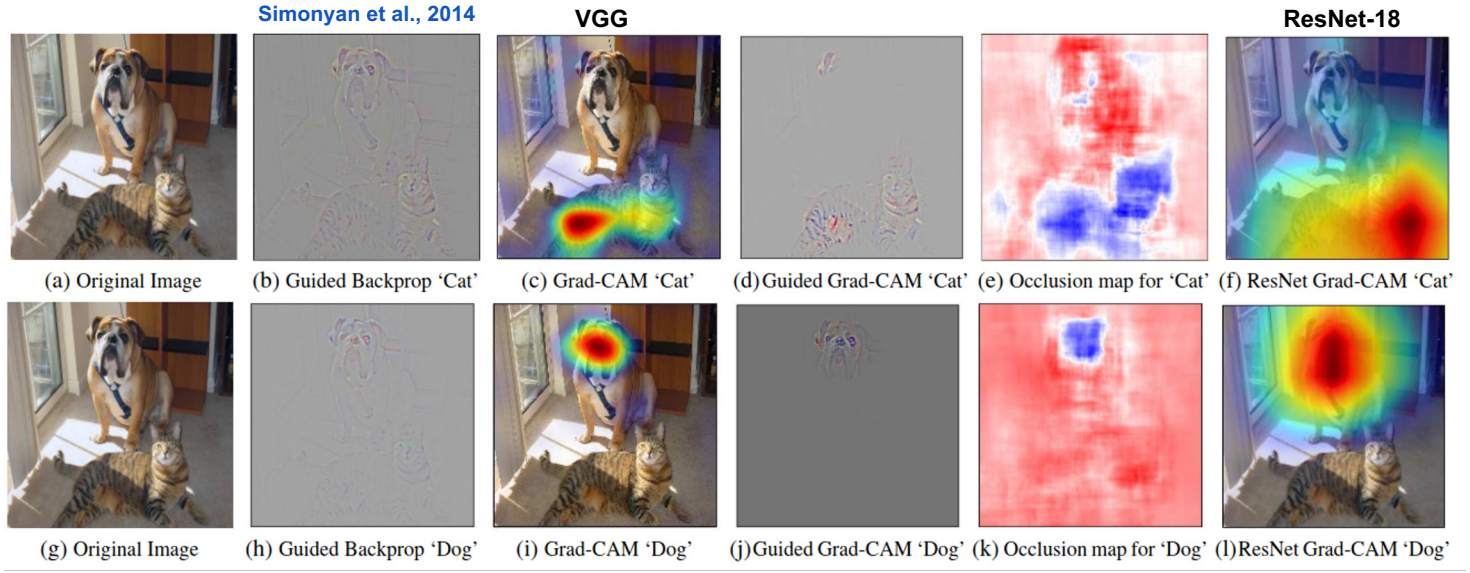
Gradient weighted Class Activation Maps (Grad-CAM)

- Generalization of CAM that can produce visual explanations for any CNN without requiring architectural changes or re-training, thus **overcoming one of the limitations of CAM** - *(Selvaraju et al. **"Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,"** ICCV, 2017)*
- It uses the gradient information flowing into the last convolutional layer to understand the importance of each neuron for a decision of interest.



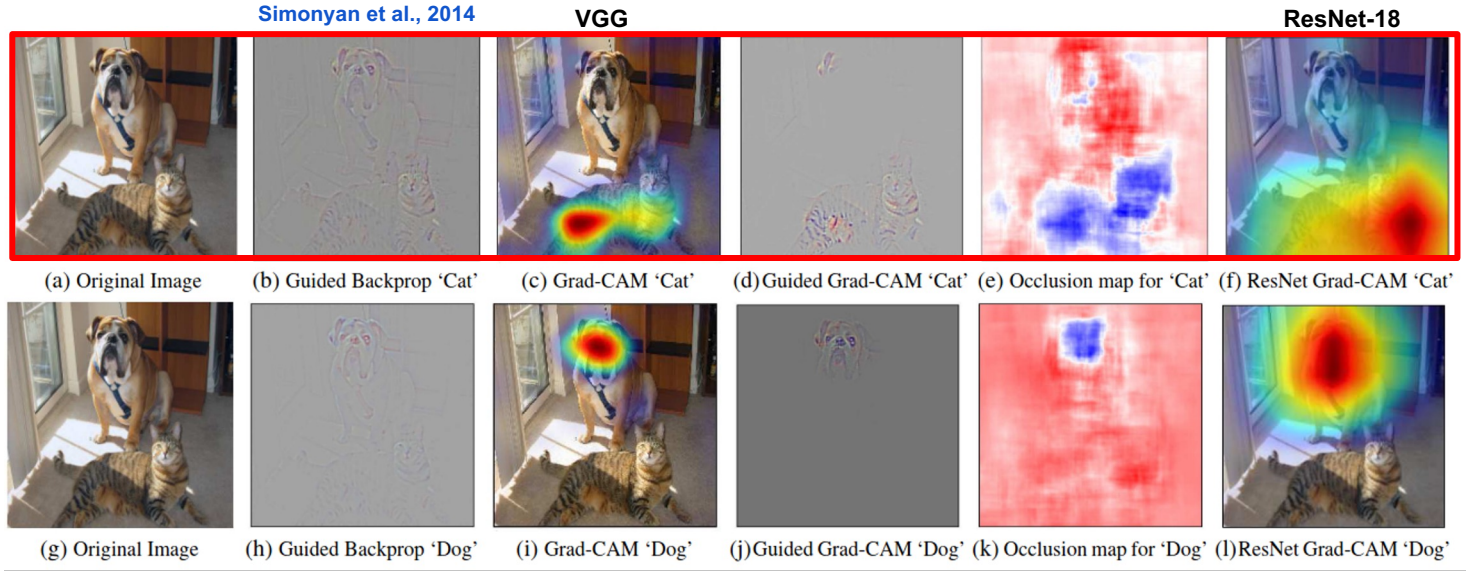
Gradient weighted Class Activation Maps (Grad-CAM)

- Results obtained by Grad-CAM method given different architectures



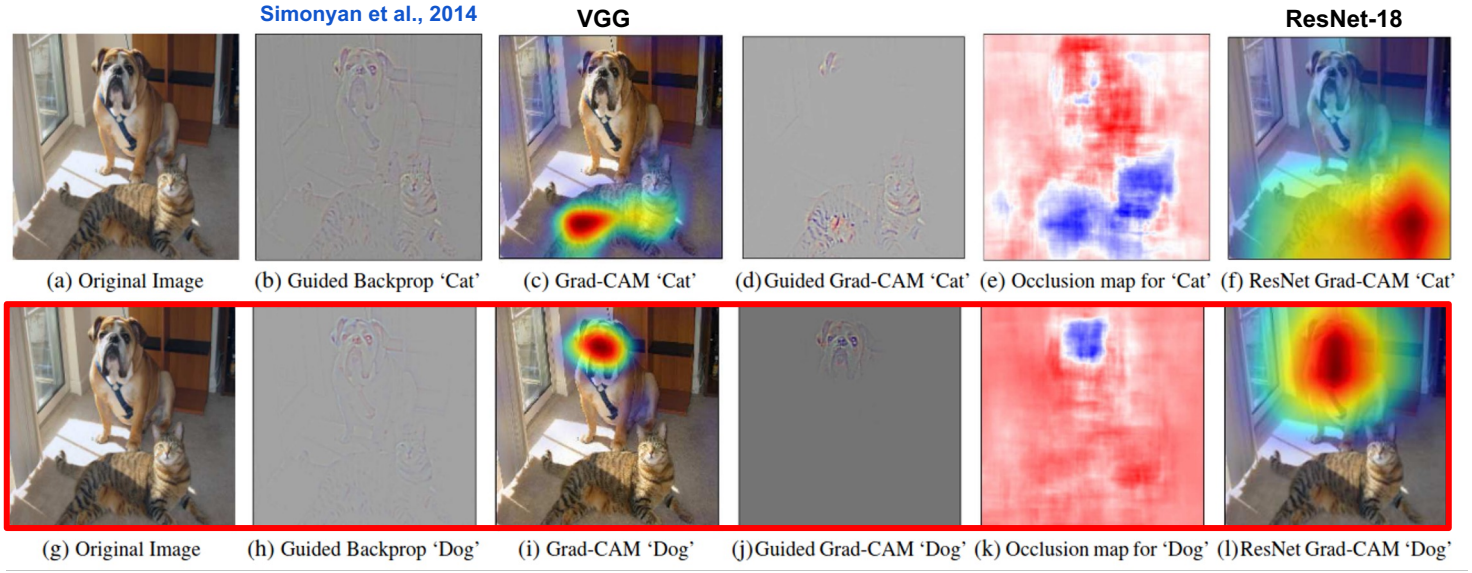
Gradient weighted Class Activation Maps (Grad-CAM)

- It is interesting to see where the networks were focusing when trying to classify a cat…



Gradient weighted Class Activation Maps (Grad-CAM)

- … or a dog.



Grad-CAM: Identifying dataset biases

- Models trained on biased datasets may not generalize to real world scenarios, or worse, **may perpetuate biases and stereotypes** (w.r.t. gender, race, age, etc.).
  - Finetune an ImageNet trained VGG-16 model for the task of classifying “doctor” vs. “nurse”
  - Although the trained model achieved a good validation accuracy, it did not generalize as well.

- Grad-CAM visualizations of the model predictions **revealed** that the model had **learned to look at the person’s face / hairstyle** to distinguish nurses from doctors, thus learning a gender stereotype.
  - The model misclassified several female doctors to be a nurse and male nurses to be a doctor.
  - Thus, they turned out that the database used for training were **gender-biased** (78% of images for doctors were men, and 93% images for nurses were women).

Grad-CAM: Identifying dataset biases

- Models trained on biased datasets may not generalize to real world scenarios, or worse, **may perpetuate biases and stereotypes** (w.r.t. gender, race, age, etc.).
  - Finetune an ImageNet trained VGG-16 model for the task of classifying “doctor” vs. “nurse”
  - Although the trained model achieved a good validation accuracy, it did not generalize as well.

- Grad-CAM visualizations of the model predictions **revealed** that the model had **learned to look at the person’s face / hairstyle** to distinguish nurses from doctors, thus learning a gender stereotype.
  - The model misclassified several female doctors to be a nurse and male nurses to be a doctor.
  - Thus, they turned out that the database used for training were **gender-biased** (78% of images for doctors were men, and 93% images for nurses were women).

Salience-based Maps Limitations

- Aggregate the results (Maps) of large amounts of data (e.g., a dataset)
- Regression problems
- Multimodal information
- Temporal Information

Fairness in Computer Vision

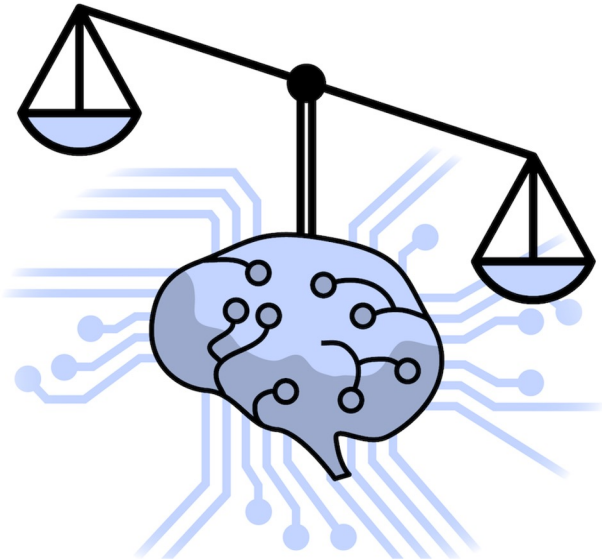


Image source: https://sitn.hms.harvard.edu/uncategorized/2020/fairness-machine-learning/

Introduction

- Past works in Machine Learning and Computer Vision were mainly focused on **improving accuracy (only)** on a varied set of subjects
  - Let's take *face detection* as example:
    - "This paper describes a machine learning approach for visual object detection which is capable of processing images extremely **rapidly** and **achieving high detection rates**."
    - On that time, fairness in computer vision was not a problem. Nowadays, it cannot be ignored.

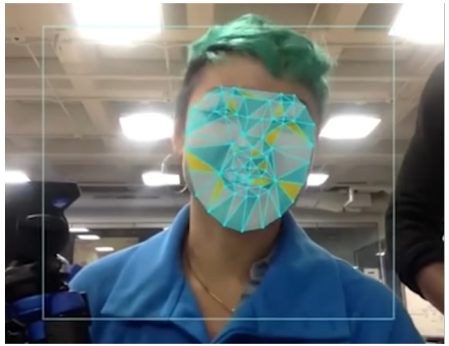
Hand-crafted features + classical machine learning

Figure 3: The first and second features selected by AdaBoost. The two features are shown in the top row and then overlayed on a typical training face in the bottom row. The first feature measures the difference in intensity between the region of the eyes and a region across the upper cheeks. The feature capitalizes on the observation that the eye region is often darker than the cheeks. The second feature compares the intensities in the eye regions to the intensity across the bridge of the nose.

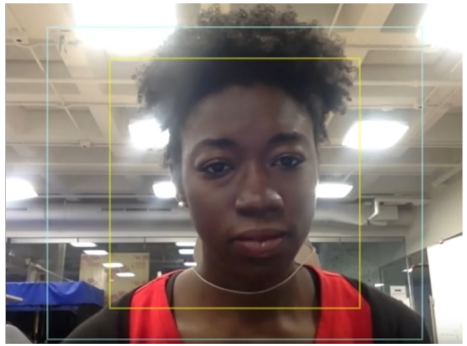
Viola and Jones ***"Rapid Object Detection using a Boosted Cascade of Simple Features,"*** *CVPR 2001.*

Introduction

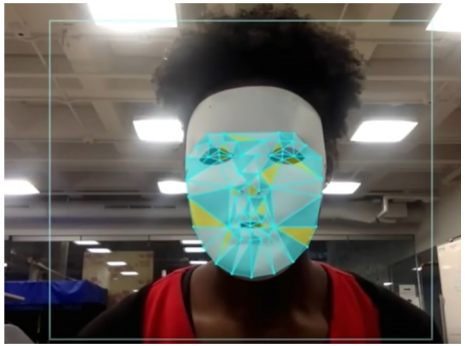
- Beyond the emergent and increasing interest on **explainable models**, current works on ML/CV are interested on the research and development of **fair models** and **bias mitigation methods**, particularly for sensitive applications where reducing bias is crucial.



Face Detected
(lighter-skin male)



Face Not Detected
(darker-skin female)



Face Detected
(wearing a white mask)

Joy Buolamwini, Timnit Gebru ***"Gender shades: Intersectional accuracy disparities in commercial gender classification,"***
ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), 2018.

How did fairness become a hot topic in ML/CV ?

- Mainly: **Biased predictions!**
  - Face detection,
  - Image captioning,
  - Recommendation systems,
  - Natural Language Processing (NLP), among others.
- Possible sources of bias:
  - **General** bias problems
  - **Person perception** bias (subjective tasks)

General Bias Problems

- Extensively studied by the machine learning / computer vision communities
- Found in almost any machine learning-based task
  - **Imbalance bias**
  - **Sample selection bias**
  - **Covariate shift**

General Bias Problems

- **Imbalance bias**: denotes the situation in which there are considerably fewer examples for one **specific decision** (or label) than for the other(s).

General Bias Problems

- **Sample selection bias**: occurs when a subset of the data is systematically excluded due to a particular attribute, **failing to ensure that the sample obtained is representative** of the population intended to be analyzed.

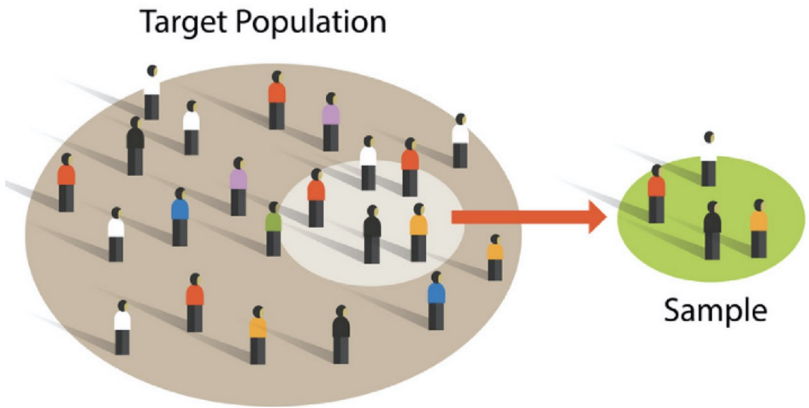
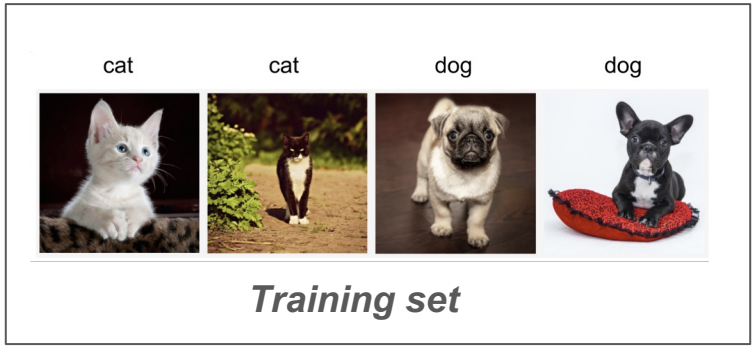


Image source: https://www.vectorstock.com/royalty-free-vector/sample-from-population-statistics-research-survey-vector-16452707

General Bias Problems

- **Covariate shift**: occurs when the distribution of input data shifts between the training environment and the test environment.



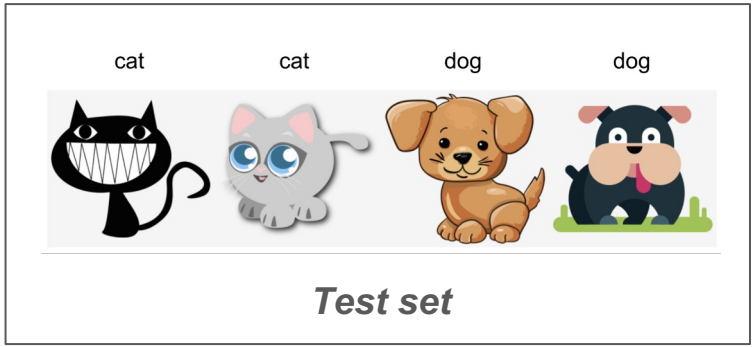


Image source: https://www.doc.ic.ac.uk/~bkainz/teaching/DL/T04_covariateShift.pdf

This is an extreme case of covariate shift, used for illustration purpose. While the labels are the same, our samples have very different features.

General Bias Problems

- Extensively studied by the machine learning / computer vision communities
- Found in almost any machine learning-based task
  - **Imbalance bias**
  - **Sample selection bias**
  - **Covariate shift**

In reality these types of bias do not necessarily occur separately, but often a biased dataset contains a mixture of these.

Jindong Gu and Daniela Oelke. ***"Understanding bias in machine learning,"*** *arXiv, 2019.*

Person perception

- Is all about the **information we gather when we meet another person**. This is part of social cognition, which basically explores how people think and act and **how we process information from our social world**.

https://study.com/academy/lesson/person-perception-definition-theory.html

Perception bias

- **The tendency to be subjective about people and events**, causing biased information to be collected in a study or biased interpretation of a study's results.

https://catalogofbias.org/biases/perception-bias/

- The biases produced by human perception, which have been widely studied in sociology and psychology (*w.r.t* gender, age, cultural, social, etc.), **have a strong influence in subjective tasks** such as automatic personality perception, emotion recognition or image captioning.

*Oh et al. "Revealing hidden **gender biases** in competence impressions of faces," Psychological Science, 2019*

*Kaufmann et al. "**Age Bias** in Selection Decisions: The Role of Facial Appearance and Fitness Impressions," Front Psychol. 2017*

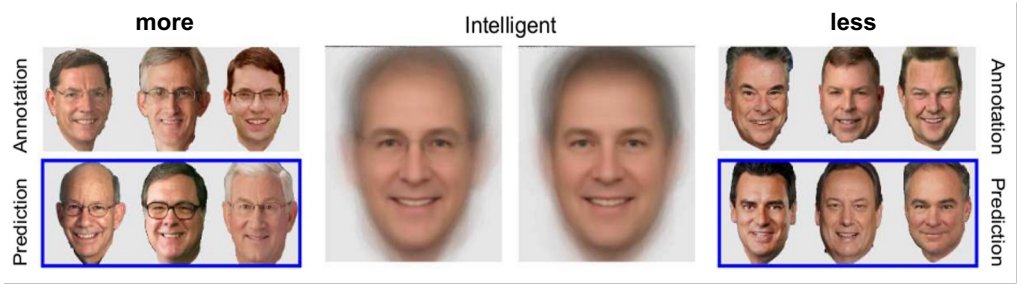
Sofer et al. "For your local eyes only: **Culture-specific** face typicality influences perceptions of trustworthiness," Perception, 2017

Sutherland, et al. "Facial first impressions from another angle: How **social judgements** are influenced by changeable and invariant facial properties," British Journal of Psychology, 2017

*Palmer and Peterson. "**Halo effects** and the attractiveness premium in perceptions of political expertise," American Politics Research, 2016*

Person perception bias

- Perceive someone as **more intelligent** just because he/she is **using glasses**.
- *Attractiveness halo-effect:* More **positive impressions** are given to more **attractive people**.



Joo, Jungseock et al. "Automated Facial Trait Judgment and Election Outcome Prediction: Social Dimensions of Face." ICCV, *2015*

- May have a **strong and negative impact on data labeling**, and consequently on the outcomes of machine learning methods trained from such data.
- Mitigating any type of bias is crucial, particularly for sensitive applications under certain controlled dimensions.

Inspiring works *that helped to push the research on explainability, transparency, fairness and bias mitigation methods*

Face detection case: *Gender Shades paper*

https://youtu.be/TWWsW1w-BVo

Joy Buolamwini, Timnit Gebru ***"Gender shades: Intersectional accuracy disparities in commercial gender classification,"***
ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), 2018.

Face detection case: *Gender Shades paper*

- The study revealed that ML algorithms can **discriminate** based on classes like race and gender.
- Two common datasets used for evaluating face classification methods are mainly composed of **lighter-skinned subjects**
  - **79.6%** for ***IJB-A*** *(US government benchmark, release by NIST in 2015 - 500 unique individuals)*
  - **86.2%** for ***Audience*** *(2284 unique individuals)*
- Introduced a new facial analysis dataset (1270 individuals) which is balanced by gender and skin type
  - Evaluated 3 commercial gender classification systems using their dataset and showed that **darker-skinned females** are the most **misclassified group**
    - With error rates of up to 34.7%
    - The maximum error rate for lighter-skinned males is 0.8%
- Highlighted that algorithms trained with **biased data** result in algorithmic discrimination

Joy Buolamwini, Timnit Gebru ***"Gender shades: Intersectional accuracy disparities in commercial gender classification,"***
ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), 2018.

Face detection case: *Gender Shades paper*

- The study revealed that ML algorithms can **discriminate** based on classes like race and gender.
- Two common datasets used for evaluating face classification methods are mainly composed of **lighter-skinned subjects**
  - **79.6%** for ***IJB-A*** *(US government benchmark, release by NIST in 2015 - 500 unique individuals)*
  - **86.2%** for ***Audience*** *(2284 unique individuals)*
- Introduced a new facial analysis dataset (1270 individuals) which is balanced by gender and skin type
  - Evaluated 3 commercial gender classification systems using their dataset and showed that **darker-skinned females** are the most **misclassified group**
    - With error rates of up to **34.7%**
    - The maximum error rate for lighter-skinned males is **0.8%**
- Highlighted that algorithms trained with **biased data** result in algorithmic discrimination

Joy Buolamwini, Timnit Gebru ***"Gender shades: Intersectional accuracy disparities in commercial gender classification,"***
ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), 2018.

Face detection case: *Gender Shades paper*

- Pilot Parliaments Benchmark (PPB)

Much more balanced representation of all subgroups

Set	n	F	M	Darker	Lighter	DF	DM	LF	LM
All Subjects	1270	44.6%	55.4%	46.4%	53.6%	21.3%	25.0%	23.3%	30.3%
Africa	661	43.9%	56.1%	86.2%	13.8%	39.8%	46.4%	4.1%	9.7%
South Africa	437	41.4%	58.6%	79.2%	20.8%	35.2%	43.9%	6.2%	14.6%
Senegal	149	43.0%	57.0%	100.0%	0.0%	43.0%	57.0%	0.0%	0.0%
Rwanda	75	60.0%	40.0%	100.0%	0.0%	60.0%	40.0%	0.0%	0.0%
Europe	609	45.5%	54.5%	3.1%	96.9%	1.3%	1.8%	44.2%	52.7%
Sweden	349	46.7%	53.3%	4.9%	95.1%	2.0%	2.9%	44.7%	50.4%
Finland	197	42.6%	57.4%	1.0%	99.0%	0.5%	0.5%	42.1%	56.9%
Iceland	63	47.6%	52.4%	0.0%	100.0%	0.0%	0.0%	47.6%	52.4%

Table 2: Pilot Parliaments Benchmark decomposition by the total number of female subjects denoted as F, total number of male subjects (M), total number of darker and lighter subjects, as well as female darker/lighter (DF/LF) and male darker/lighter subjects (DM/LM). The group compositions are shown for all unique subjects, Africa, Europe and the countries in our dataset located in each of these continents.

➔ Darker females (DF) are the least represented in ***IJB-A*** (**4.4%**)

➔ Darker males (DM) are the least represented in ***Adience*** (**6.4%**)

Face detection case: *Gender Shades paper*

- Comercial **Gender** Classification

Highlighted boxes: higher scores = better results

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

"Not that bad", but given the gender, and skin-tone, and sub-groups

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Face detection case: *Gender Shades paper*

- Comercial **Gender** Classification

Highlighted boxes: higher scores = better results

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

"Not that bad", but given the gender, and skin-tone, and sub-groups

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Face detection case: *Gender Shades paper*

- Comercial **Gender** Classification

Highlighted boxes: higher scores = better results

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

"Not that bad", but given the gender, and skin-tone, and sub-groups

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Face detection case: *Gender Shades paper*

- Comercial **Gender** Classification

Highlighted boxes: higher scores = better results

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

"Not that bad", but given the gender, and skin-tone, and sub-groups

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Face detection case: *Gender Shades paper*

- Comercial **Gender** Classification

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7

One important lesson we can take from this, is that **if you have a subgroup in your test set having a very small representation**, this group may have a low impact on the evaluated score. Thus, one algorithm may have high and global accuracy but it **does not guarantee it will be able to generalize well in a real world scenario**.

rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Face detection case: *Gender Shades paper*

To conclude, the paper emphasized:

- **The need of *Transparency***
  - For human-centered computer vision, they define transparency as *"**providing information** on the demographic and phenotypic composition of training and testing datasets"*
    - Ex.: data distribution, taking into account the multiple attributes and subgroups
- **The need of *Accountability***
  - *"**Reporting algorithmic performance** on demographic and phenotypic subgroups and actively working to close performance gaps where they arise"*

AI in dermatology

Does your dermatology AI app work for your skin colour?

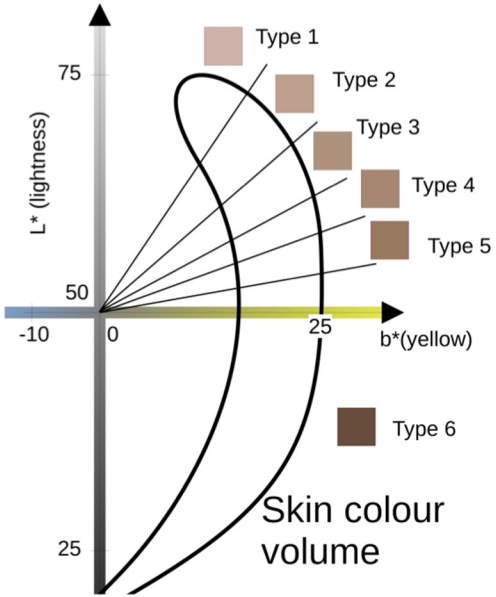
- Let’s imagine you wake up one day and find a new dark spot on your skin.
- Before having a doctor appointment, you decide to check it with a **skin lesion classifier**, which was trained by looking at many medical images of different skin lesions.

*Dangerous **melanoma** or harmless beauty spot?*

https://www.linkedin.com/pulse/does-your-dermatology-ai-app-work-skin-colour-unsolved-thorsten-kalb-ijlwf?trk=public_profile_article_view

AI in dermatology

- The usability of ***such AI tools should not depend on the skin colour of the patient***.
- But their performance depends on the training data and **often there are more images of light skin available than dark skin images**, which can cause any dermatological AI tool to likely perform worse on darker skin subjects.



https://www.linkedin.com/pulse/does-your-dermatology-ai-app-work-skin-colour-unsolved-thorsten-kalb-ijlwf?trk=public_profile_article_view

AI in dermatology

- In this work, they reviewed and compared four approaches for **skin tone classification on** a common benchmark for assessing skin cancer classification fairness in the literature, and revealed that:
  - Common algorithms for skin tone estimation are **unreliable and contradictory**.
  - A benchmark dataset for skin tone fairness **contains likely no image of brown or dark skin.**

Revisiting Skin Tone Fairness in Dermatological Lesion Classification

Thorsten Kalb, Kaisar Kushibar, Celia Cintas, Karim Lekadir, Oliver Diaz & Richard Osuala

In CLIP 2023, EPIMI 2023, FAIMI 2023: Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging

https://www.linkedin.com/pulse/does-your-dermatology-ai-app-work-skin-colour-unsolved-thorsten-kalb-ijlwf?trk=public_profile_article_view

Image Captioning Case

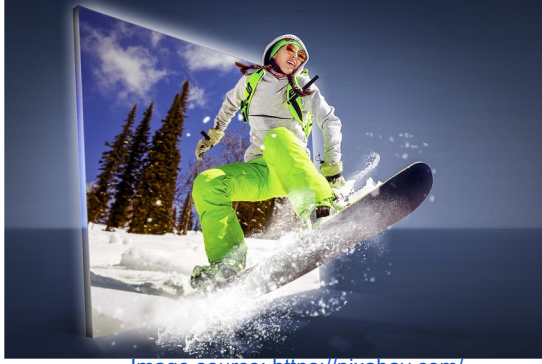


Image source: https://pixabay.com/

- *Hendricks et al.* investigated the generation of **gender-specific caption words** (e.g. man, woman) based on the person's appearance or the image context.
  - Motivated by the fact that **exploiting contextual cues** can frequently lead to better performance in computer vision tasks
  - However, in some cases making decisions based on context can lead to incorrect, and perhaps even offensive predictions → "***Women Also Snowboard***"
- They Introduced a model that **encourages equal gender probability** in the absence of gender information and **discriminative** when gender evidence is present.

Hendricks et al. ***"Women Also Snowboard: Overcoming Bias in Captioning Models"****, ECCV, 2018*

Image Captioning Case

- To do so, two complementary loss terms are considered:
  - Appearance Confusion Loss: is based on the intuition that, **given an image in which evidence of gender is absent**, description models should be *unable to accurately predict a gendered word*.
  - Confident Loss: helps to increase the model's confidence **when gender is in the image**.

$$\mathcal{L} = \alpha\mathcal{L}^{CE} + \beta\mathcal{L}^{AC} + \mu\mathcal{L}^{Con},$$

where α=μ=1 and β=10

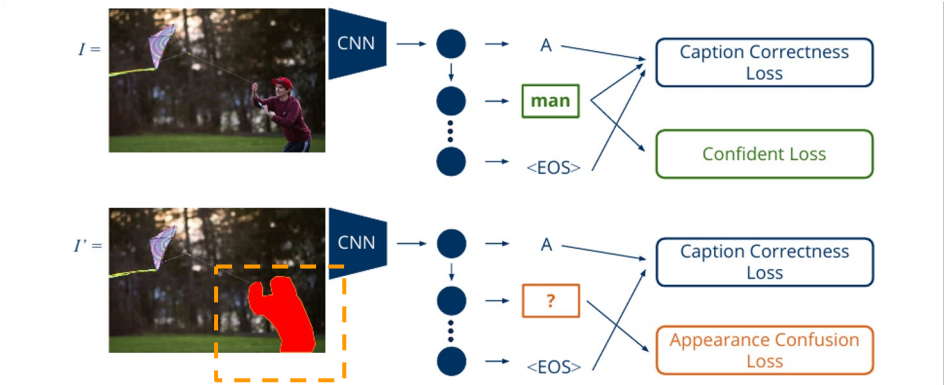


Fig. 2: Equalizer includes two novel loss terms: the Confident Loss on images with men or women (top) and the Appearance Confusion Loss on images where men and women are occluded (bottom). Together these losses encourage our model to make correct predictions when evidence of gender is present, and be cautious in its absence. We also include the Caption Correctness Loss (cross entropy loss) for both image types.

Hendricks et al. ***"Women Also Snowboard: Overcoming Bias in Captioning Models"****, ECCV, 2018*

Image Captioning Case

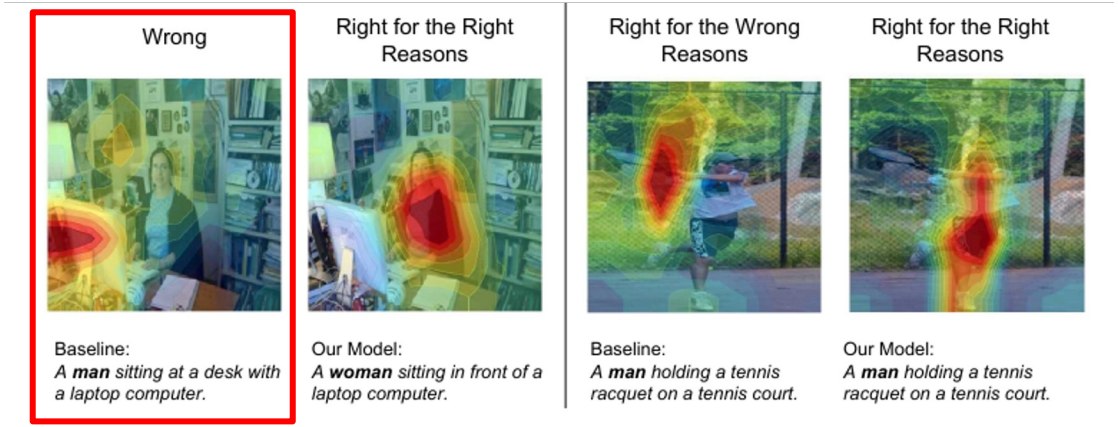


Fig. 1: Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.

Hendricks et al. ***"Women Also Snowboard:*** ***Overcoming Bias in Captioning Models"****, ECCV, 2018*

Image Captioning Case

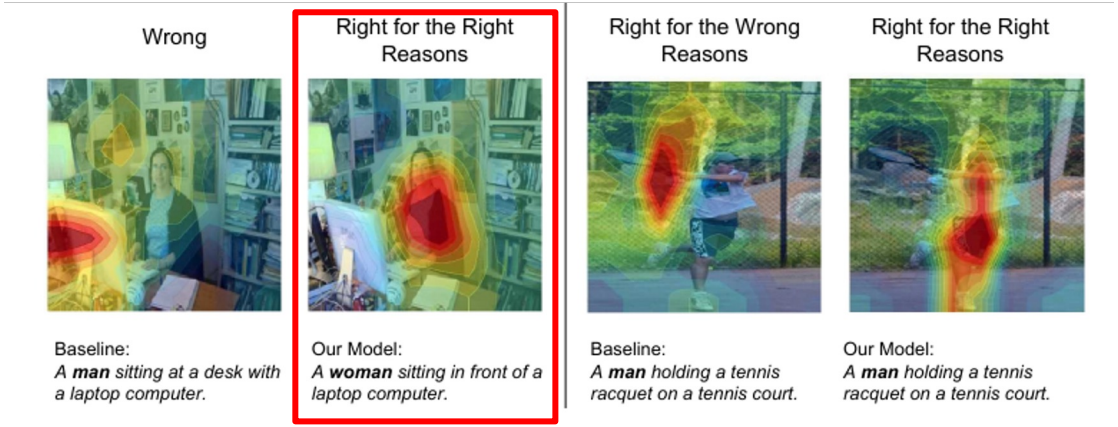


Fig. 1: Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.

Hendricks et al. ***"Women Also Snowboard:*** ***Overcoming Bias in Captioning Models"****, ECCV, 2018*

Image Captioning Case

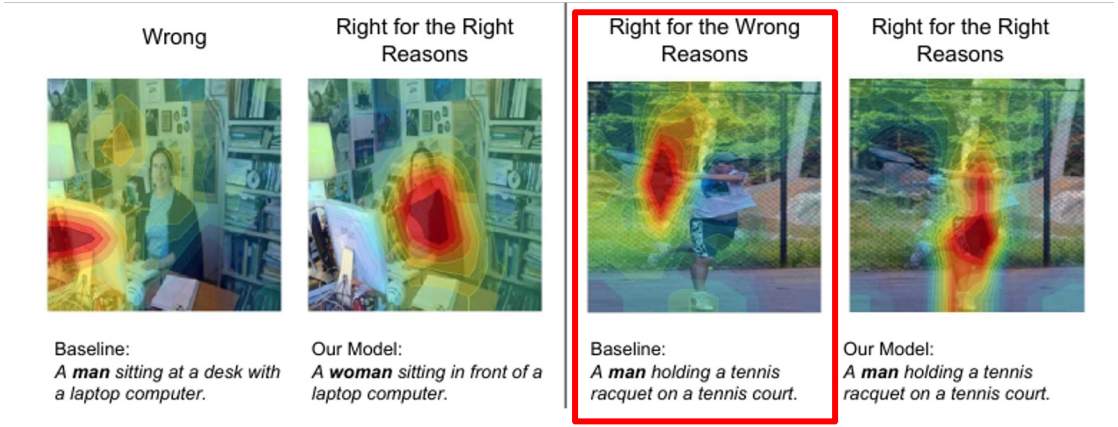


Fig. 1: Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.

Hendricks et al. ***"Women Also Snowboard: Overcoming Bias in Captioning Models"****, ECCV, 2018*

Image Captioning Case

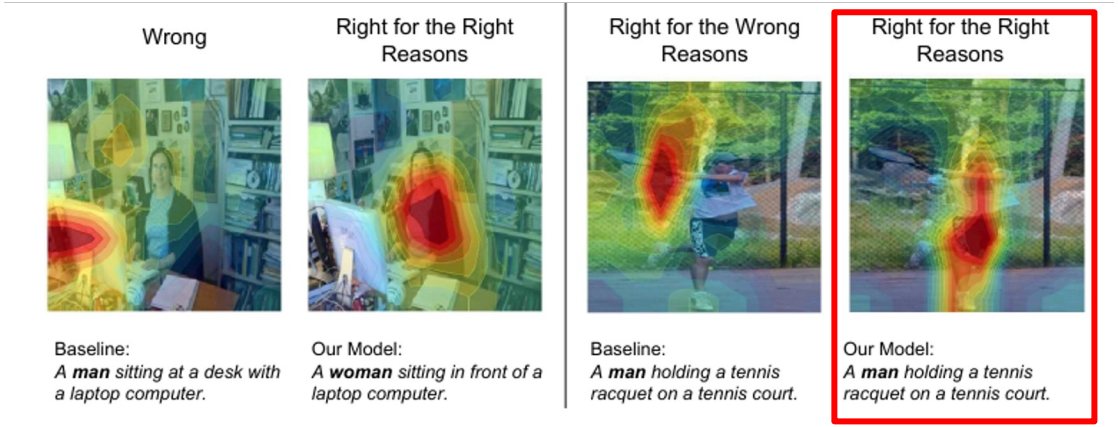


Fig. 1: Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.

Hendricks et al. ***"Women Also Snowboard: Overcoming Bias in Captioning Models"****, ECCV, 2018*

But how to define fairness?

How to define Fairness?

- For simplicity, let's assume A is encoded as a binary attribute (e.g., sex), but this can be generalized

- Definition 1: **Fairness Through Unawareness (FTU)**
  - An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process.
    - Any mapping $\hat{\mathrm{Y}} : \mathrm{X} \rightarrow \mathrm{Y}$ that excludes A satisfies this.
    - However, elements of X can contain discriminatory information analogous to A that may not be obvious at first (i.e., “features can be biased”)

Kusner et al. ***"Counterfactual Fairness,"*** *NeurIPS, 2017.*

How to define Fairness?

- Definition 2: **Individual Fairness (IF)**
  - An algorithm is fair if it gives similar predictions to similar individuals.
    - Formally, if individuals i and j are similar under a given a metric d (i.e., $d(i,j)$ is small) then their predictions should be similar:

$$\hat{Y}\left(X(i),A(i)\right) \approx \hat{Y}\left(X(j),A(j)\right)$$

    - The challenge here is that the metric $d(\cdot,\cdot)$ must be carefully chosen, requiring a clear understanding of the domain at hand.

Kusner et al. ***"Counterfactual Fairness,"*** *NeurIPS, 2017.*

How to define Fairness?

- Definition 3: **Demographic Parity (DP)**
  - States that the proportion of each segment of a protected class (e.g., sex) should receive the positive outcome at equal rates.
  - A predictor $\hat{Y}$ satisfies demographic parity if

$$P(\hat{Y} | A = 0) = P(\hat{Y} | A = 1)$$

➔ **Other definitions of fairness can be found in the literature** (e.g., Equality of Opportunity, Counterfactual Fairness, etc).

Even if we can't state exactly **what fair should look like**, we often have a **good idea of** what unfair is.

Ways to mitigate bias

- According to *Friedler et al.*, fairness-aware machine learning approaches can be categorised as:
  a. **Preprocessing** techniques which aim to modify the input data;
  b. **Algorithm modification** techniques, which modify existing algorithms by adding constraints or regularisation; and
  c. **Postprocessing** techniques which modify the output of an existing method to be fair.

*These categories consider the **data is already available** and ready to use*

- One step back → Intelligent data **collection** and **preparation + transparency**

Friedler et al. ***"A comparative study of fairness-enhancing interventions in machine learning,"***
Conference on Fairness, Accountability, and Transparency, 2019

Ways to mitigate bias

- According to *Friedler et al.*, fairness-aware machine learning approaches can be categorised as:
  a. **Preprocessing** techniques which aim to modify the input data;
  b. **Algorithm modification** techniques, which modify existing algorithms by adding constraints or regularisation; and
  c. **Postprocessing** techniques which modify the output of an existing method to be fair.

*These categories consider the **data is already available** and ready to use*

- One step back → Intelligent data **collection** and **preparation + transparency**

Friedler et al. ***"A comparative study of fairness-enhancing interventions in machine learning,"***
Conference on Fairness, Accountability, and Transparency, 2019

Ways to mitigate bias

- According to *Friedler et al.*, fairness-aware machine learning approaches can be categorised as:
  a. **Preprocessing** techniques which aim to modify the input data;
  b. **Algorithm modification** techniques, which modify existing algorithms by adding constraints or regularisation; and
  c. **Postprocessing** techniques which modify the output of an existing method to be fair.

*These categories consider the **data is already available** and ready to use*

- One step back → Intelligent data **collection** and **preparation + transparency**

Friedler et al. ***"A comparative study of fairness-enhancing interventions in machine learning,"***
Conference on Fairness, Accountability, and Transparency, 2019

Ways to mitigate bias

- According to *Friedler et al.*, fairness-aware machine learning approaches can be categorised as:
  a. **Preprocessing** techniques which aim to modify the input data;
  b. **Algorithm modification** techniques, which modify existing algorithms by adding constraints or regularisation; and
  c. **Postprocessing** techniques which modify the output of an existing method to be fair.

*These categories consider the **data is already available** and ready to use*

- One step back → Intelligent data **collection** and **preparation + transparency**

Friedler et al. ***"A comparative study of fairness-enhancing interventions in machine learning,"***
Conference on Fairness, Accountability, and Transparency, 2019

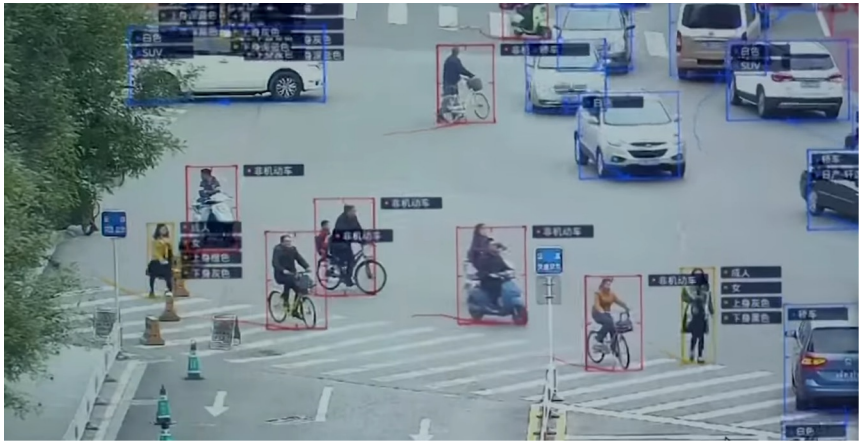
Challenges & possible future research directions

- Human Analysis:
  - **Multiple attributes**
    - Age, gender, race, culture, facial expression, hair-style, skin color, image background, image quality, etc
  - **Multimodality and temporal information**
    - Audio, Video, Text
- **Identifying** and **discovering** the unknown biases
  - Most works focus on bias mitigation given a particular set of known attributes (e.g., gender, race, etc).

Challenges & possible future research directions

- Human Analysis:
  - **Multiple attributes**
    - Age, gender, race, culture, facial expression, hair-style, skin color, image background, image quality, etc
  - **Multimodality and temporal information**
    - Audio, Video, Text

- **Identifying** and **discovering** the unknown biases
  - Most works focus on bias mitigation given a particular set of known attributes (e.g., gender, race, etc).



Privacy & Ethics

Images source: https://www.youtube.com/watch?v=p8WAIFatAqw

Privacy & Ethics

- The **revolution** created with the age of big-data and deep learning has changed the world we live in different ways
- Some applications have been practically reinvented, e.g.:
  - Smart healthcare,
  - Financial technology,
  - Visual surveillance systems.
- We use smartphones and smart-watches with different types of sensors;
- We have cameras everywhere
- Some people agree with terms and conditions without reading one single sentence
- In this world, **privacy and ethics have emerged as a big concern**.

Liu et al. ***"When Machine Learning Meets Privacy: A Survey and Outlook,"***
ACM Comput. Surv. 2021.

Promises and Concerns

Visual Human Behavior analysis based applications

- Increased safety at airports, public and private spaces
- Healthcare applications
- Counter terrorism
- Non-touch access and login
- Locating missing people
- Games
- Etc, etc

- Violation of personal data and privacy
- Lack of transparency
- Biases and unfairness
- Unauthorized or inappropriate surveillance
- Improper uses, e.g., predict criminality from face images
- Psychological impact of being watched
- Etc, etc

A dedicated **discussion** and the production of guidelines around these themes are crucial to advance in the direction of **AI for good**.

Regulation

- Different initiatives have been proposed to advance in the direction of IA for good.
- EU: General Data Protection Regulation (GDPR, 2016)
  - The aim of the GDPR is to **protect fundamental rights and freedoms** and, in particular, the right to personal data protection.
  - The regulation defines personal data as **“all the information on an identified or identifiable natural person”**, that is, any data that allows the identity of a natural person to be determined, directly or indirectly, without a disproportionate effort, such as:
    - Name and surnames
    - Address, telephone number, email address
    - Date and place of birth
    - Passport numbers, other documents, vehicle registration number, etc
    - Employment and educational information
    - Biometric data (fingerprint, facial image)
    - Genetic data (DNA)
    - Health data
    - IP address

European Commission

Additional information: *https://ec.europa.eu/info/law/law-topic/data-protection_en*

Regulation

- GDPR defines entities like
  - Data controller
  - Data processor
- Having distinct responsibilities about the data
  - Ensuring the data processing is performed in accordance with the regulation.

European Commission

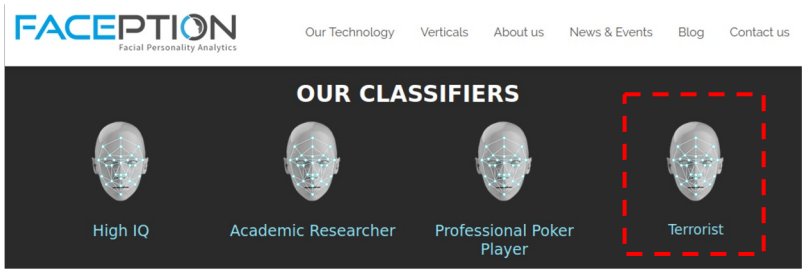
Additional information: *https://ec.europa.eu/info/law/law-topic/data-protection_en*

Questionable AI examples

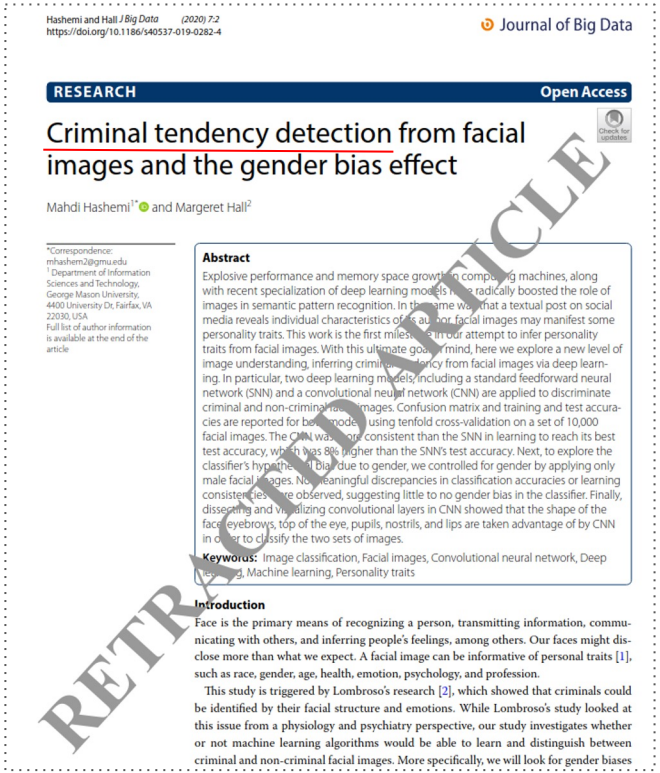
This article is more than 3 years old

Cambridge Analytica scandal: the biggest revelations so far

Source:https://www.theguardian.com/uk-news/2018/mar/22/cambridge-analytica-scandal-the-biggest-revelations-so-far



Source: https://www.faception.com/



Hashemi and Hall *J Big Data* (2020) 7:2
https://doi.org/10.1186/s40537-019-0282-4

Journal of Big Data

RESEARCH Open Access

Criminal tendency detection from facial images and the gender bias effect

Mahdi Hashemi[1*] and Margeret Hall[2]

*Correspondence: mhashem2@gmu.edu
[1] Department of Information Sciences and Technology, George Mason University, 4400 University Dr, Fairfax, VA 22030, USA
Full list of author information is available at the end of the article

Abstract

Explosive performance and memory space growth in computing machines, along with recent specialization of deep learning models have radically boosted the role of images in semantic pattern recognition. In the same way that a textual post on social media reveals individual characteristics of its author, facial images may manifest some personality traits. This work is the first milestone in our attempt to infer personality traits from facial images. With this ultimate goal in mind, here we explore a new level of image understanding, inferring criminal tendency from facial images via deep learning. In particular, two deep learning models, including a standard feedforward neural network (SNN) and a convolutional neural network (CNN) are applied to discriminate criminal and non-criminal facial images. Confusion matrix and training and test accuracies are reported for both models using tenfold cross-validation on a set of 10,000 facial images. The CNN was more consistent than the SNN in learning to reach its best test accuracy, which was 8% higher than the SNN's test accuracy. Next, to explore the classifier's hypothetical bias due to gender, we controlled for gender by applying only male facial images. No meaningful discrepancies in classification accuracies or learning consistencies were observed, suggesting little to no gender bias in the classifier. Finally, dissecting and visualizing convolutional layers in CNN showed that the shape of the face eyebrows, top of the eye, pupils, nostrils, and lips are taken advantage of by CNN in order to classify the two sets of images.

Keywords: Image classification, Facial images, Convolutional neural network, Deep learning, Machine learning, Personality traits

Introduction

Face is the primary means of recognizing a person, transmitting information, communicating with others, and inferring people's feelings, among others. Our faces might disclose more than what we expect. A facial image can be informative of personal traits [1], such as race, gender, age, health, emotion, psychology, and profession.

This study is triggered by Lombroso's research [2], which showed that criminals could be identified by their facial structure and emotions. While Lombroso's study looked at this issue from a physiology and psychiatry perspective, our study investigates whether or not machine learning algorithms would be able to learn and distinguish between criminal and non-criminal facial images. More specifically, we will look for gender biases

RETRACTED ARTICLE

The power and the danger of deepfakes

- We should not forget to mention the power and danger of deep fakes
- Have a potential and dangerous power in the creation of **fake news** and **mass manipulation**

Source: https://www.bbc.com/news/business-51204954

The power and the danger of deepfakes

https://youtu.be/1OqFY_2JE1c

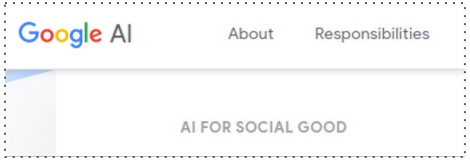
Virtual interviews

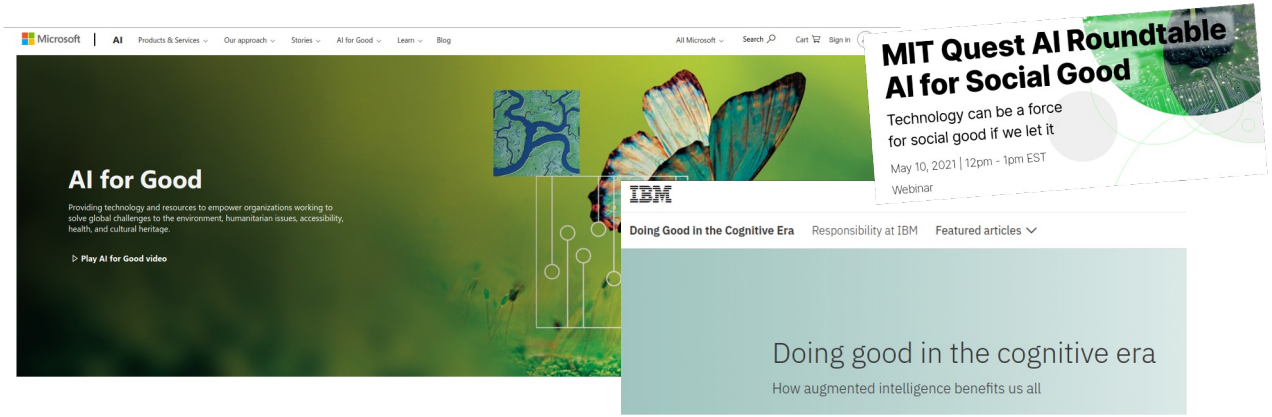
- Nowadays, some companies are using computer vision tools for screening job candidates, which is another very sensitive topic.
- **Can we remove the humans from the loop** and let the algorithm to take the final decision?

Image source: https://www.theguardian.com/technology/2018/mar/04/robots-screen-candidates-for-jobs-artificial-intelligence

AI for Good

- Emerging expressions you may find in different discussions around AI
- The discussions around this topic has been attracted a huge attention of the whole society, from public/private sectors, academia to industry.
- The discussion, however, is in its infancy.
- **Involve:** Society + Research + Industry + Academy + ...





Challenges: *https://www.xprize.org/aidataforgood*

The take-home message

We are responsible for

promoting and building

the next generation of AI for good!

Questions?