Detecting and Mitigating Bias in Machine Learning Image Data through Semantic Description of the Attention Mechanism

The use-case Gender Bias in Profession Prediction from Images

George Dimitropoulos

September 16, 2019

TU Delft
Delft University of Technology
IBM

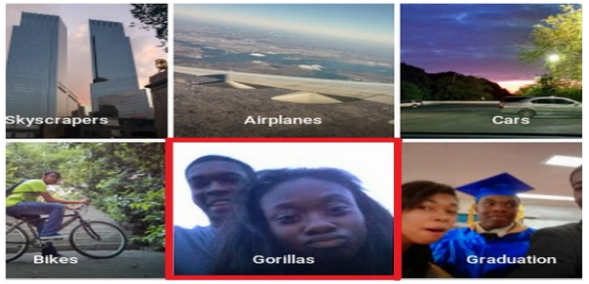
Machine Learning is now used in many high-stakes **decision making applications**

- Financial services
- Health care
- Criminal justice

Concerns have been raised about the **impact** of these **decisions** on **people's lives**

2 main **problems** have been identified in the **literature review**

Problem 1: **Discrimination** between different groups of **people** w.r.t. to their **protected attributes** (e.g. gender, age, religion, race)

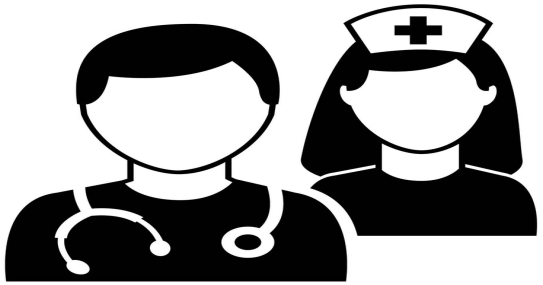


Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss



Problem 2: **Lack of methods** for the **interpretation** and **explanation** of the **predictions** made by these **ML systems**

Focus of the **Thesis Project**: **Training Data** for **Machine Learning**

- Most of the **ML models** that are used in these decision making applications are **built** from **human-generated data**

- **Human biases** produce a **skewed** and sometimes **unbalanced distribution** in the **training data**

Proposed solution in **Literature**:

- **Impose** an **equal distribution** of the **training data** with respect to the **protected attribute** of people
- **However,** as we will show later (experiments) this is **not always true**
- **We show** that**,** it is **not** the **distribution** of the **labels w.r.t.** to the **protected attribute, but** the **presence** of specific **visuals clues** that **leads** to that **bias**

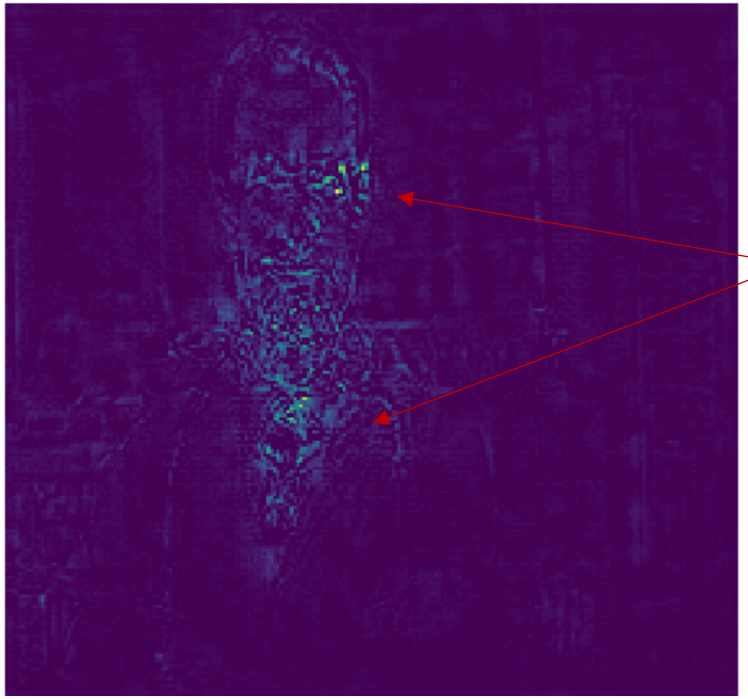
Our Goal: Develop a **methodology** that helps in **detecting**, **reasoning** upon and **compensating** for **bias** w.r.t. **protected attributes** due to **issues** in the training **data**

- **Use case**: **Profession prediction** from images
- Specific form of data -> **Images**
- Protected Attribute -> **Gender**

Hypothesis: **Presence** of **specific visual clues** in images, that **give away** the **gender, affects** the **classification** outcome and **introduces bias** on that

Original Image

SmoothGrad, Noise:10%, Samples: 500



Presence of male **face** and **tie**

MAIN RESEARCH QUESTION

How to **analyze**, **reason upon** and **fix** the **content** of Machine Learning **image training data** in order to **correct** and **reduce gender bias** in the output of the subsequent trained models?

RESEARCH SUB-QUESTIONS / CONTRIBUTIONS

RSQ1: Which are the **current methods** and their **limitations** related to **bias** in **ML data** with respect to **protected attributes** of people?

CO1: **Literature Review**

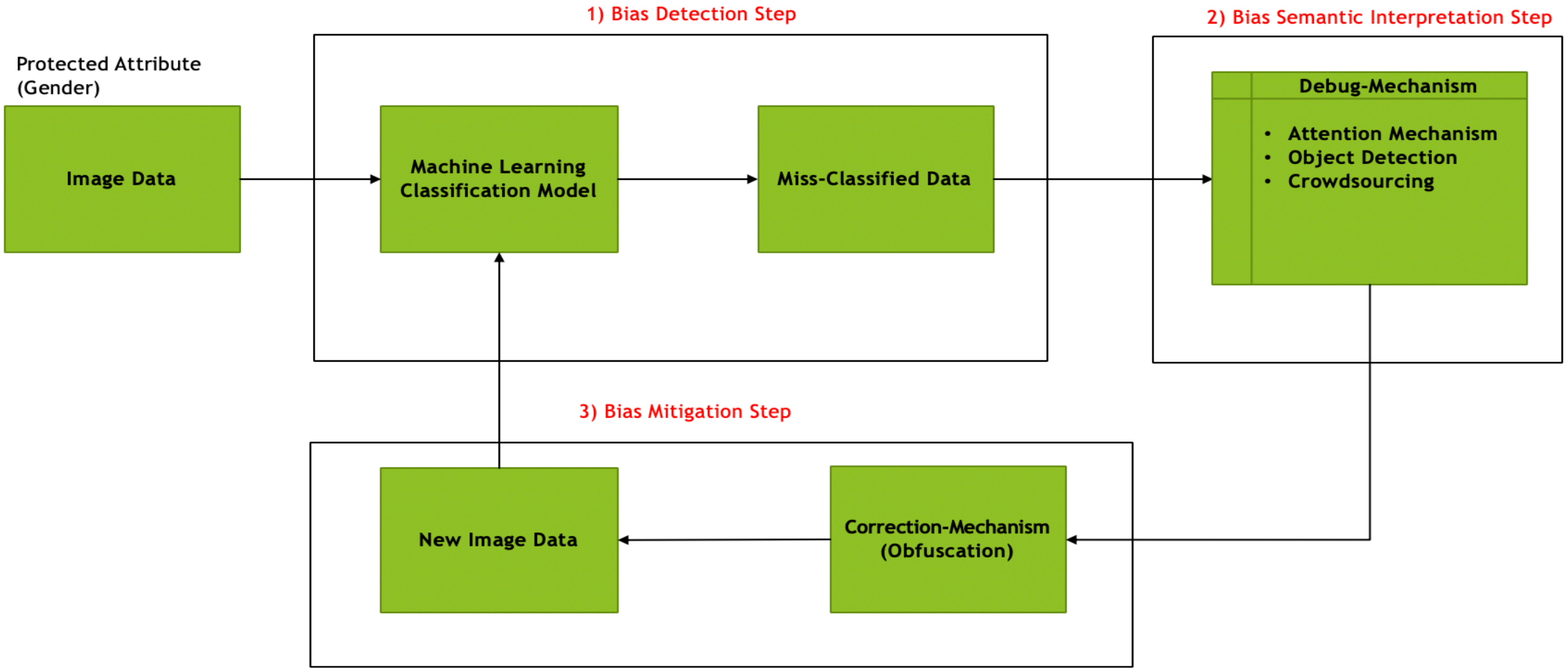
RSQ2: How can we **describe** in a **semantically rich fashion at scale** the **features** in the **data** that are likely to be **related** to a particular **biased prediction** of a **ML system**?

CO2: A **Methodology**

RSQ3: How can we **compensate** for **gender bias** that is **related** with the **content** of the **image data** in **ML systems?**

CO3: A **Methodology**

METHODOLOGY



USE CASE

- **Predicting Occupation** from Images

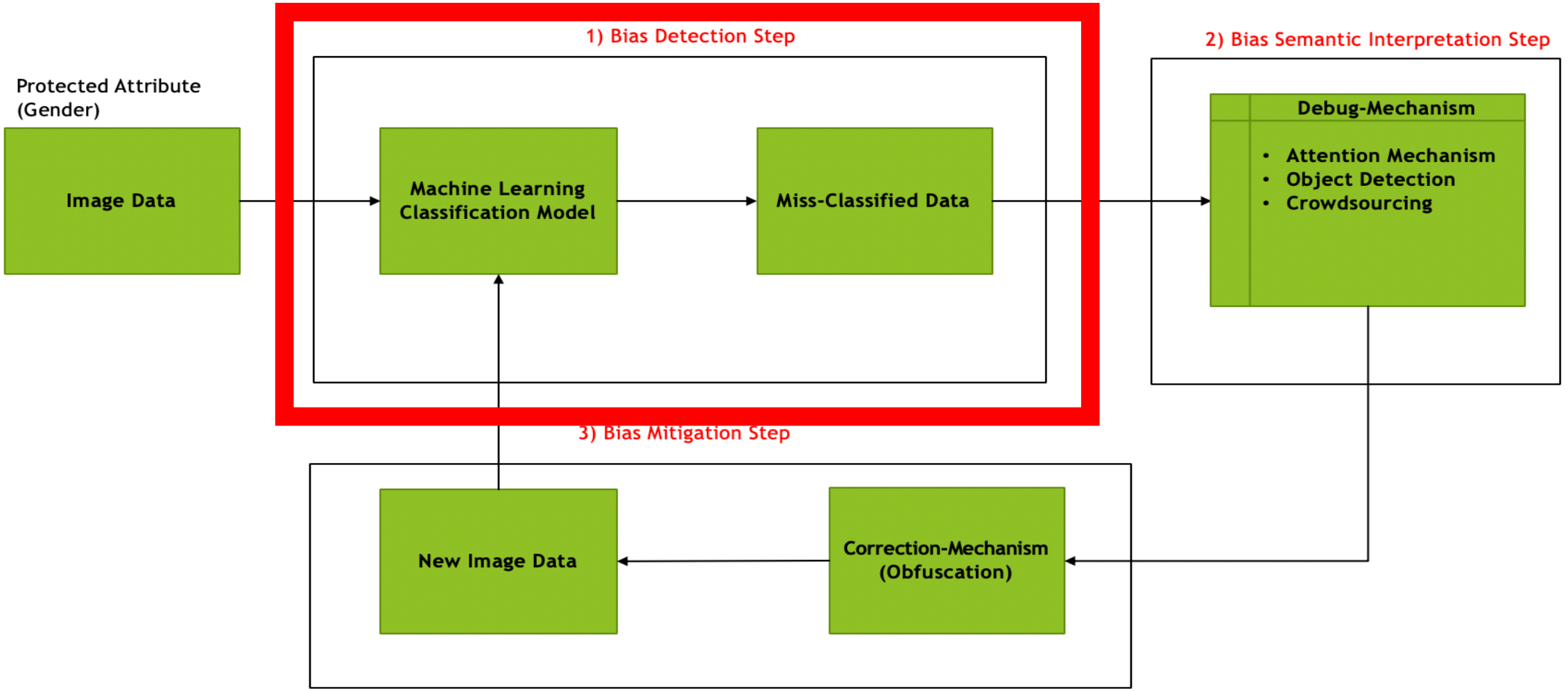
- **Equal distribution** of the **labels w.r.t.** to the **protected attribute in the training data**

3 Datasets

- Doctor/Nurse
- Chef/Waiter
- Engineer/Farmer

EXAMPLE

EXAMPLE

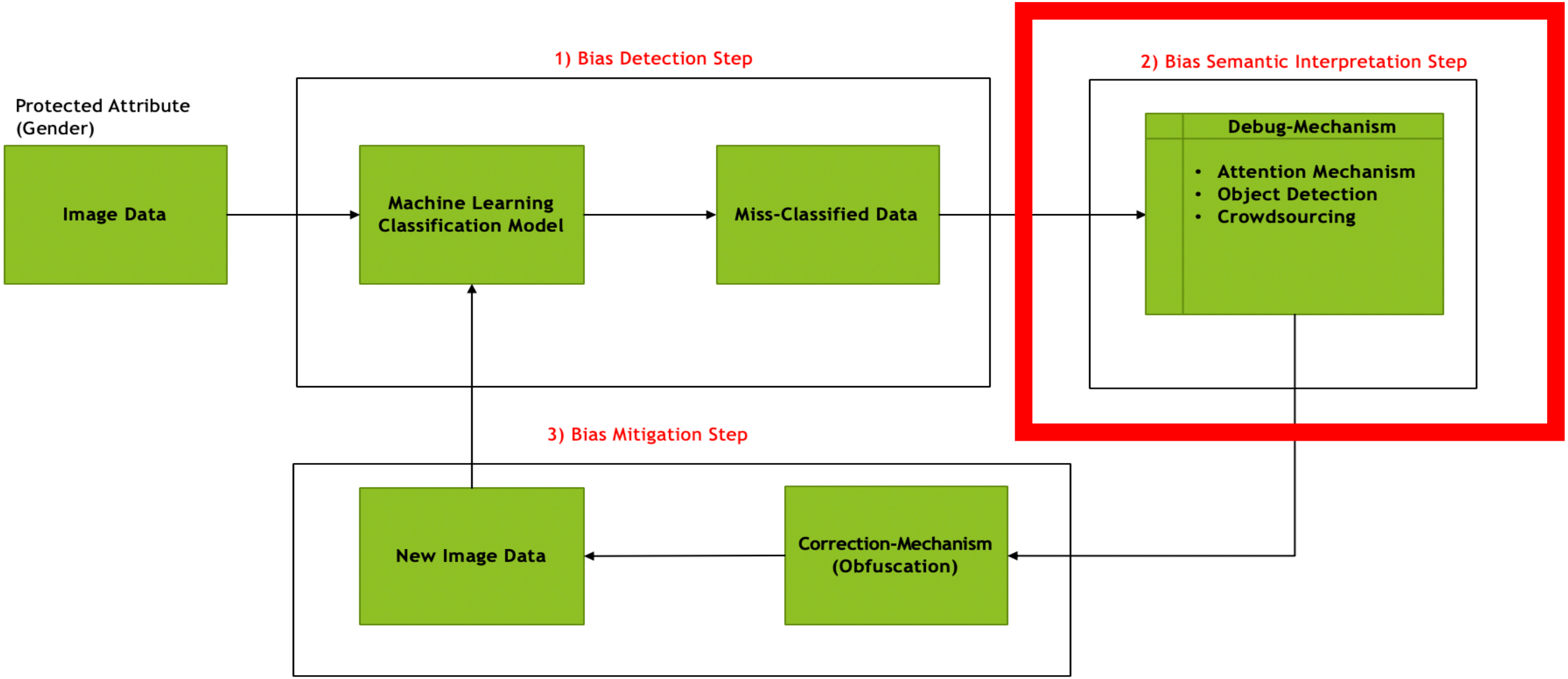

Protected Attribute
(Gender)
Image Data
1) Bias Detection Step
Machine Learning
Classification Model
Miss-Classified Data
2) Bias Semantic Interpretation Step
Debug-Mechanism
• Attention Mechanism
• Object Detection
• Crowdsourcing
3) Bias Mitigation Step
New Image Data
Correction-Mechanism
(Obfuscation)

EXAMPLE

Machine Learning Classification Model -> **Pre-Trained Residual Neural Network**

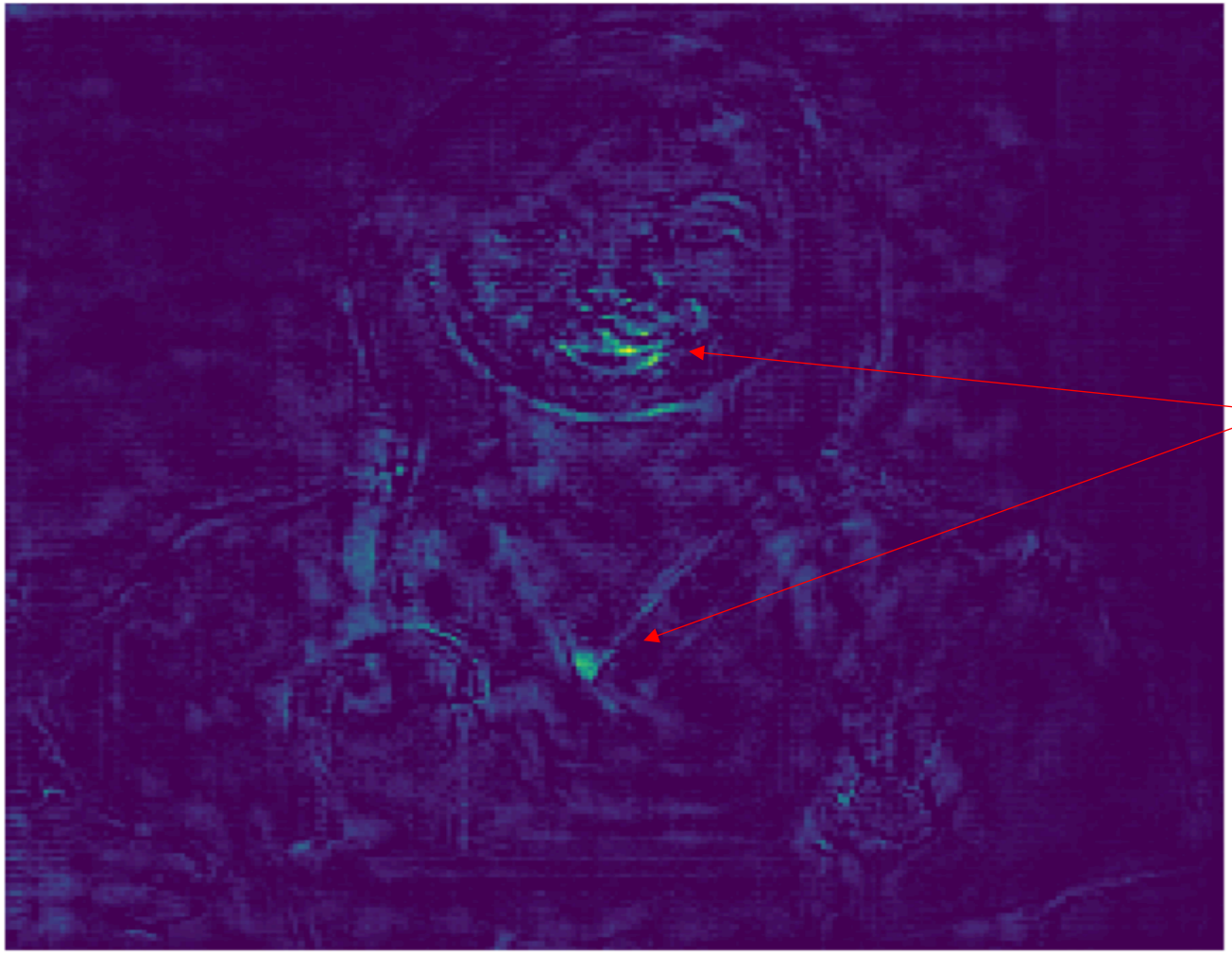
Real Label: **Doctor**
Prediction: **Nurse**
Probability: **80%**

EXAMPLE


1) Bias Detection Step
2) Bias Semantic Interpretation Step
Protected Attribute
(Gender)
Image Data
Machine Learning
Classification Model
Miss-Classified Data
Debug-Mechanism
• Attention Mechanism
• Object Detection
• Crowdsourcing
3) Bias Mitigation Step
New Image Data
Correction-Mechanism
(Obfuscation)

EXAMPLE

Attention Mechanism (AM) (**Parts** of the image that **influence** the **prediction**) -> **SmoothGrad algorithm**



Presence of face and opening in the clothing in the neck

GOAL of this **Step** -> Provide to **people insights** about the **gender bias** in the **predictions** of the ML classification model

- By **inspecting** each image we observe that the **presence** of **give way elements** of **gender affects** the classification **outcome** and **introduces bias** to that

- However, it is **difficult** to follow this procedure **at scale**

RSQ2: How can we **describe** in a **semantically rich fashion at scale** the **features** in the **data** that are likely to be **related** to a particular **biased prediction** of a **ML system**?

Goal: Describe with a **set of classes of pre-defined objects**, helping in attaching a **semantic label** on top of the attention mechanism

2 Proposed Solutions → **Object Detection**

2 Proposed Solutions → **Crowdsourcing**

Object Detection (OD) -> **Name** and **Position** of the **objects** -> **Bounding Boxes**

Goal: Observe whether the semantic description can be attained automatically

Crowdsourcing Task (CT) -> **Name** and **Position** of the **objects** (visual clues of gender bias) -> **Bounding Boxes**

Goal 1: **Understand** how the **intuition** of **people** about a potential **cause** of **gender bias** actually **compares** with the **actual reason** that **affects** the **prediction** of a ML classification model

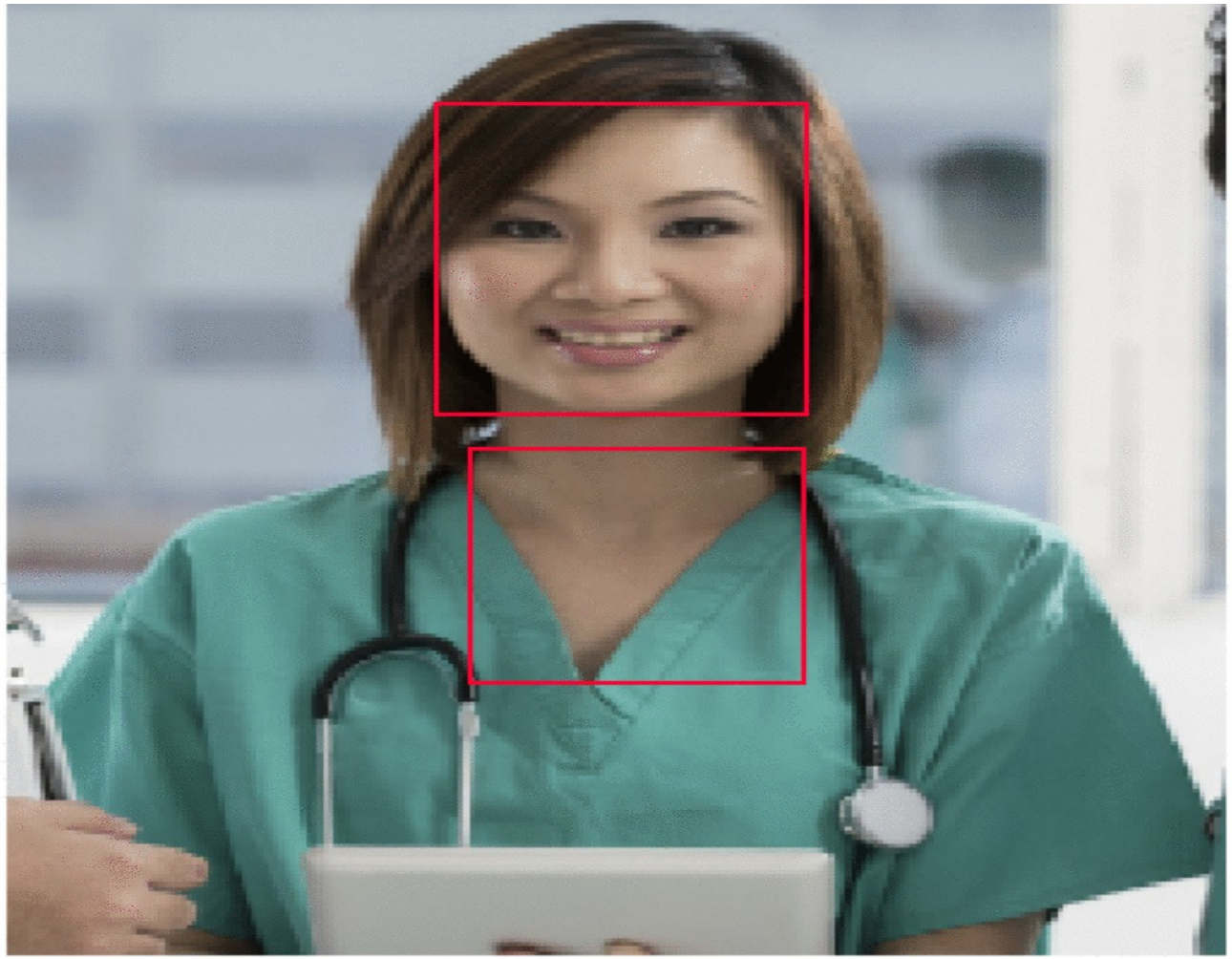
Goal 2: Gain an **insight** of how much the **intuition** of **people** about **elements** of **gender bias** actually **matches** the **semantic description** coming from the **object detection**

EXAMPLE

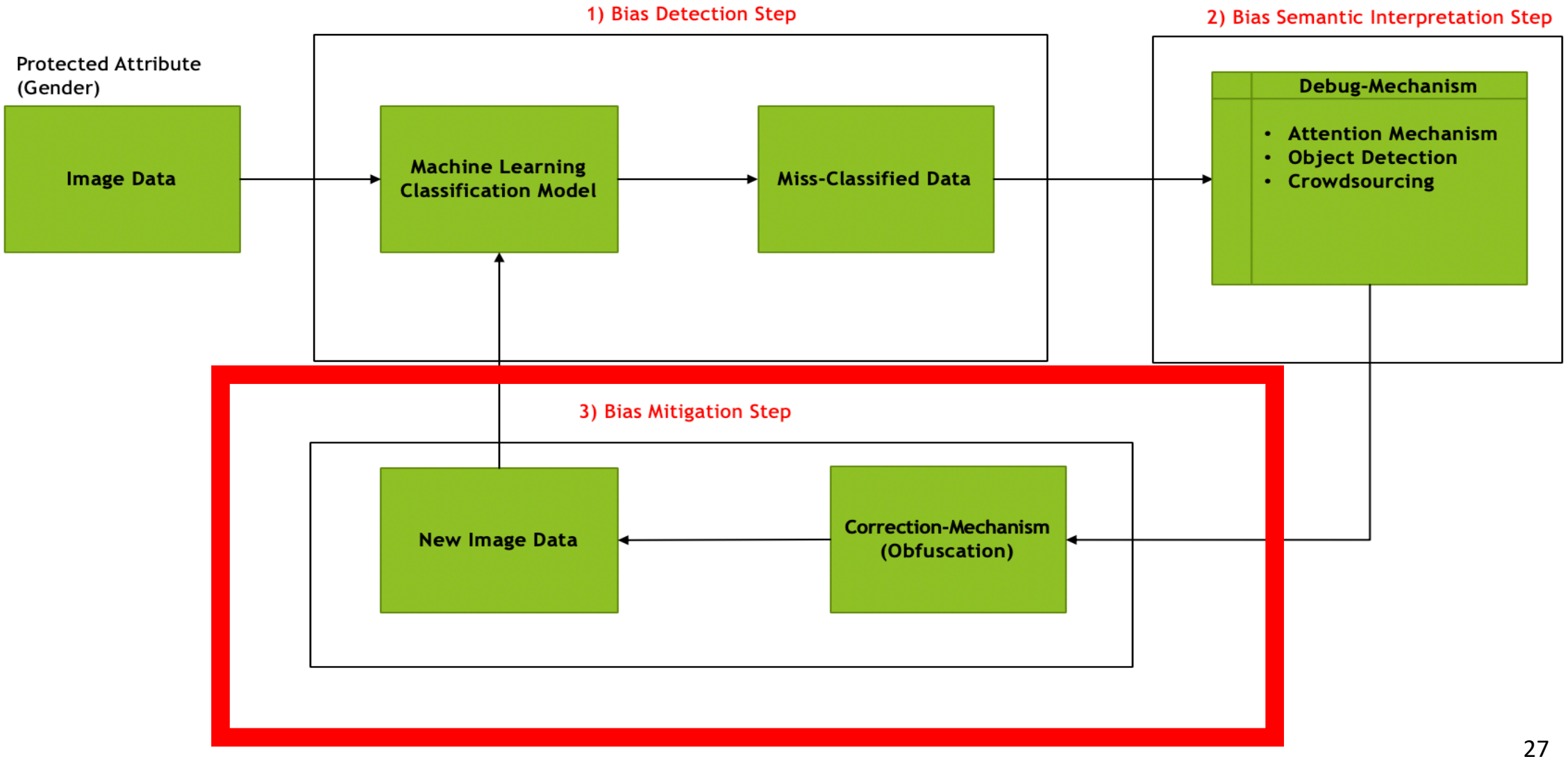
RSQ2a: Could we automatically **detect semantically meaningful visual** clues that **are related with** the prediction through **combining AM** and **OD**?

EXAMPLE

RSQ2b: Could we **detect not only** semantically meaningful visual clues **but also clues** of **gender bias** that **are related with** the prediction through **combining AM** and **CT**?



EXAMPLE


1) Bias Detection Step
2) Bias Semantic Interpretation Step
Protected Attribute
(Gender)
Image Data
Machine Learning
Classification Model
Miss-Classified Data
Debug-Mechanism
• Attention Mechanism
• Object Detection
• Crowdsourcing
3) Bias Mitigation Step
New Image Data
Correction-Mechanism
(Obfuscation)

EXAMPLE

RSQ3: Does **obfuscating** the **visual clues** coming from **overlapping** of the **AM-OD** and **AM-CT** **improve** the **predictions** ?

Before:

Real Label: **Doctor**
Prediction: **Nurse**
Probability: **80%**

After:

Real Label: **Doctor**
Prediction: **Doctor**
Probability: **88%**

EXPERIMENTAL RESULTS (BIAS DETECTION)

BEFORE

Initial Prediction performance per Gender

Doctor Class	Nurse Class
Male Accuracy: **87.4**%	Male Accuracy: 85.3%
Female Accuracy: **72.6**%	Female Accuracy: 84.5%

Gender bias in the predictions in **doctor** class (14.8% difference)

Chef Class	Waiter Class
Male Accuracy: **86.3**%	Male Accuracy: 85.4%
Female Accuracy: **58.9**%	Female Accuracy: 85%

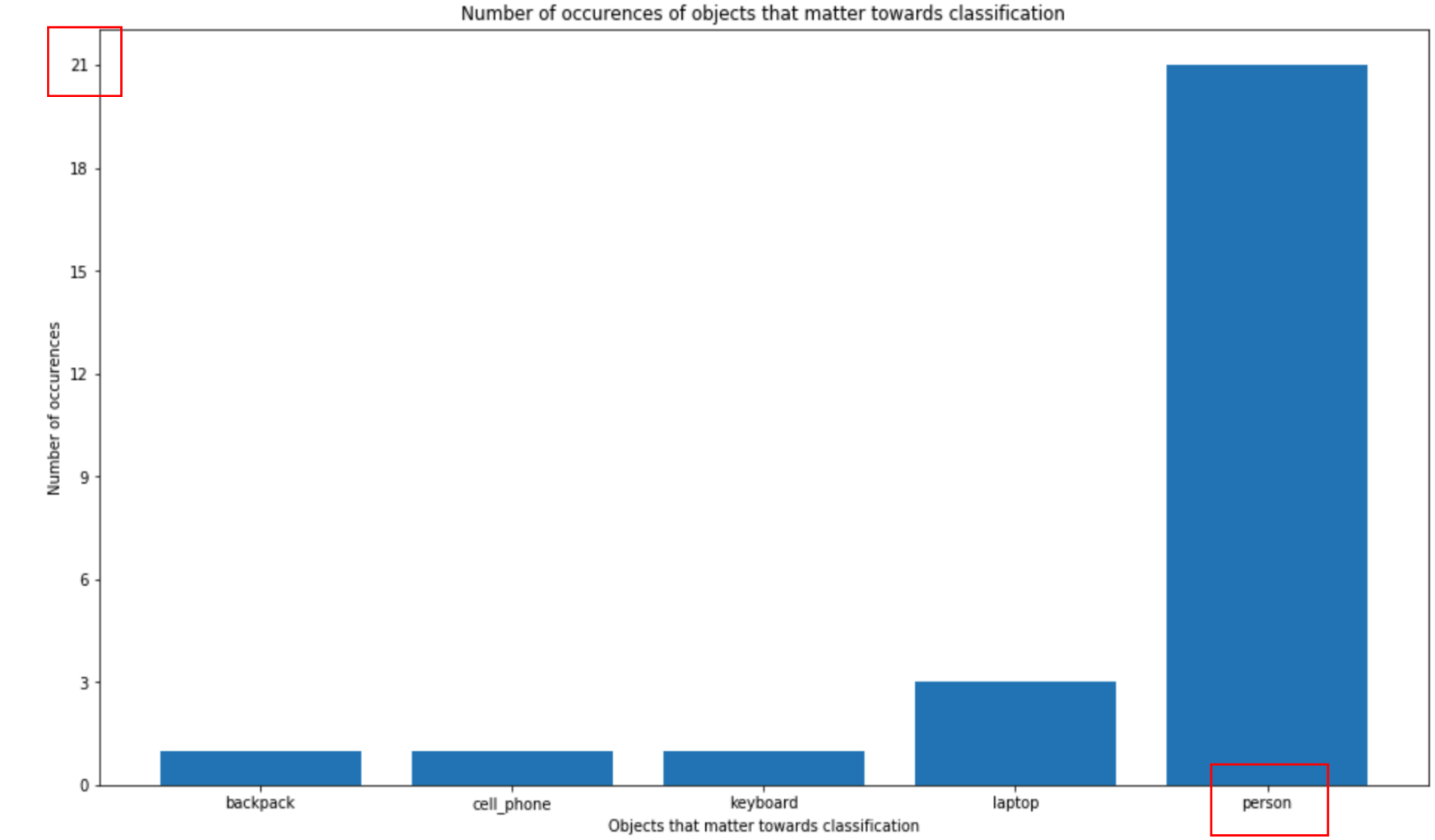
Gender bias in the predictions in **chef** class (27.4% difference)

Engineer Class	Farmer Class
Male Accuracy: **95.5**%	Male Accuracy: 94.4%
Female Accuracy: **88.8**%	Female Accuracy: 96.6%

Gender bias in the predictions in **engineer** class (6.7% difference)

EXPERIMENTAL RESULTS (BIAS SEMANTIC INTERPRETATION)

Number of **occurrences** of **objects** that **matter** towards **classification** for the **doctor class** (**A1, AM-OD**)



EXPERIMENTAL RESULTS (BIAS SEMANTIC INTERPRETATION)

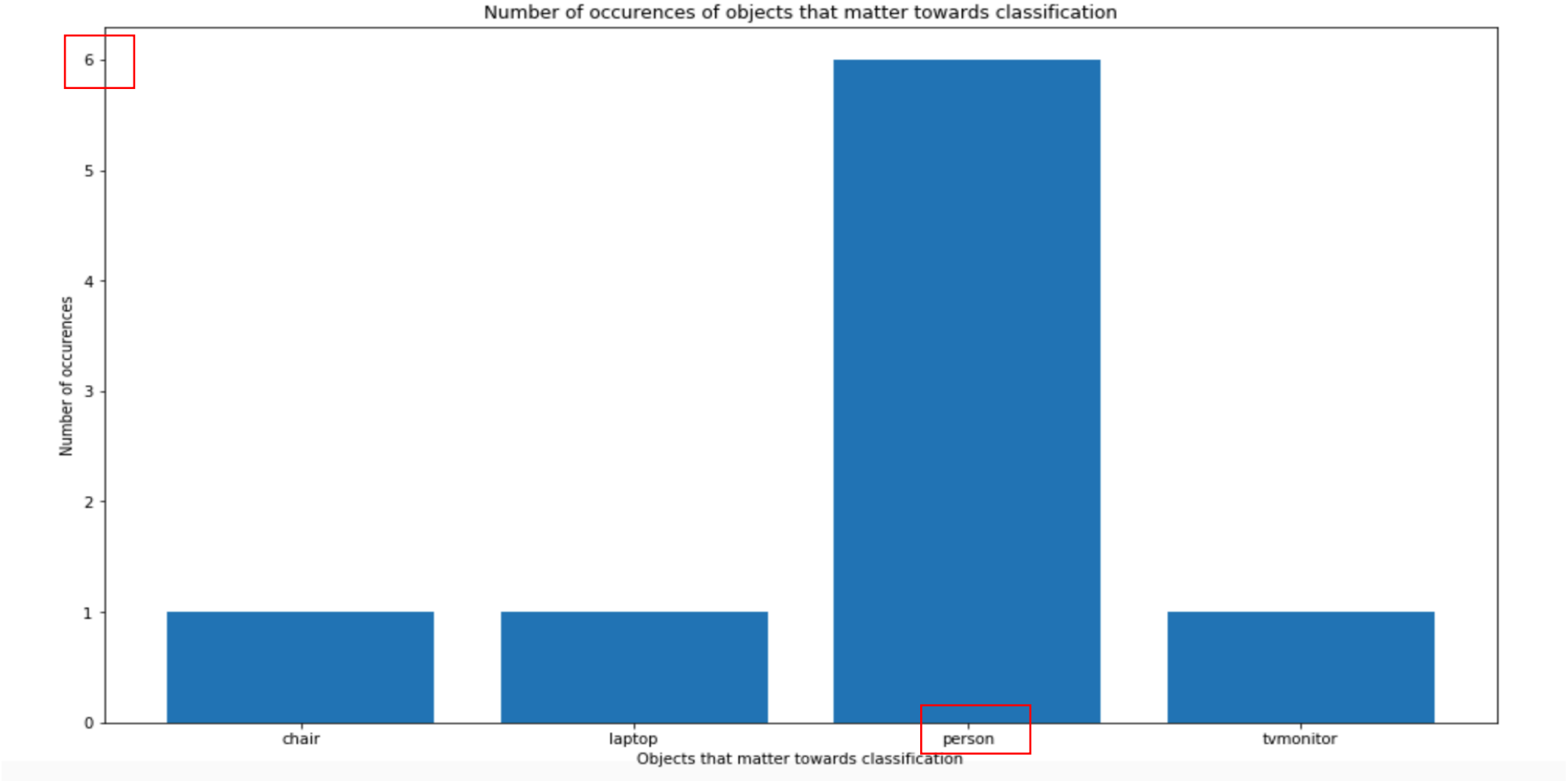
- **Primary reason** of a **classification** of a **person** as a **doctor** is the **presence** of the **face** or **body** (**class person**)

- The model does **not look** at all in the **presence** of a **stethoscope** but it looks in the **face** or the **body** (**class person**)

- The model learns to act in a **gender discriminative** way

EXPERIMENTAL RESULTS (BIAS SEMANTIC INTERPRETATION)

Number of **occurrences** of **objects** that **matter** towards **classification** for the **nurse class** (**A1, AM-OD**)



EXPERIMENTAL RESULTS (BIAS SEMANTIC INTERPRETATION)

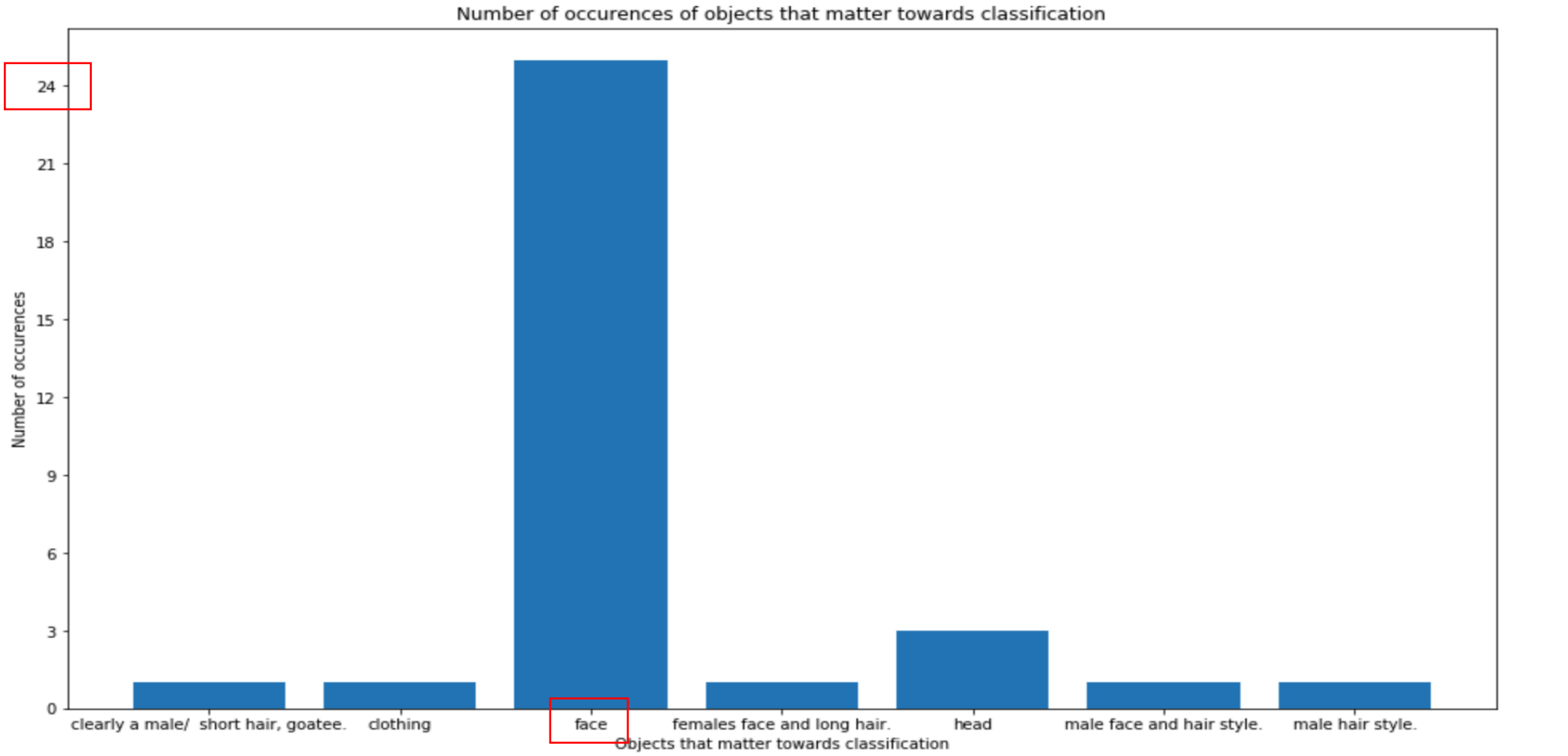
- **Primary reason** of a **classification** of a **person** as a **nurse** is the **presence** of the **face** or **body** (**class person**)

- **However**, the model now takes into account the **presence** of a **person** only in **6 images** and **not** in **21 like** in the **doctor** class

- The model does **not act** in a **gender discriminative** way

EXPERIMENTAL RESULTS (BIAS SEMANTIC INTERPRETATION)

Number of **occurrences** of **objects** that **matter** towards **classification** for the **doctor class** (**A2, AM-CT**)



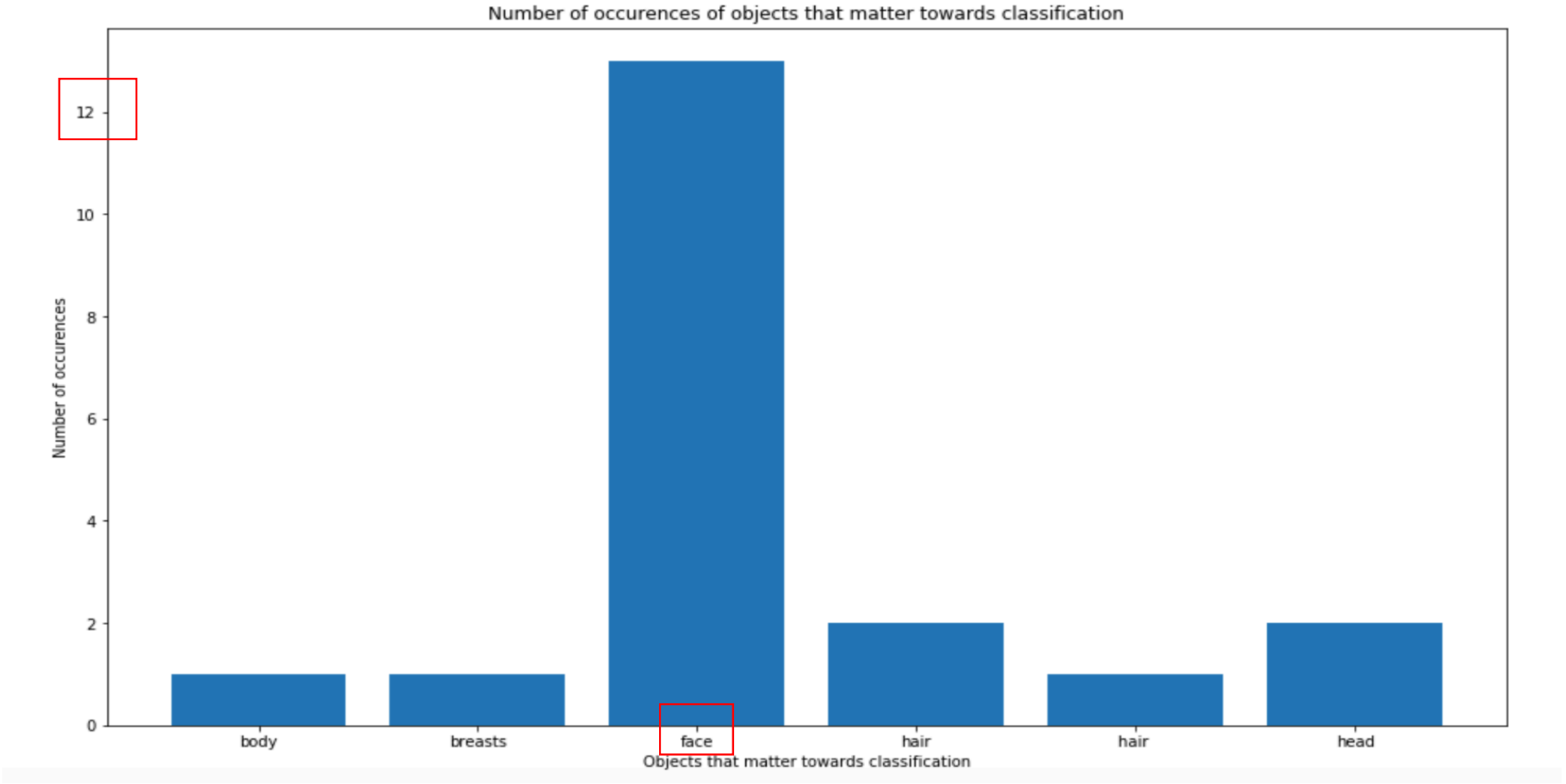
EXPERIMENTAL RESULTS (BIAS SEMANTIC INTERPRETATION)

- **Primary reason** of a **classification** of a **person** as a **doctor** is the **presence** of the **face** (and in accordance with A1)

- The model learns to act in a **gender discriminative** way

EXPERIMENTAL RESULTS (BIAS SEMANTIC INTERPRETATION)

Number of **occurrences** of **objects** that **matter** towards **classification** for the **nurse class** (**A2, AM-CT**)



EXPERIMENTAL RESULTS (BIAS SEMANTIC INTERPRETATION)

- **Primary reason** of a **classification** of a **person** as a **nurse** is the **presence** of the **face** (and in accordance with A1)

- **However**, the model now takes into account the **presence** of a **face** only in **12 images** and **not** in **24 like** in the **doctor** class

- The model does **not act** in a **gender discriminative** way

EXPERIMENTAL RESULTS (BIAS SEMANTIC INTERPRETATION)

Qualitative Analysis

- In **A1**, we end up also with **objects** that are related with the **specific profession** (e.g. **uniform**, **laptop** etc.)

- In **A2**, we have a **better understanding** of the reason of a specific prediction through having a more **detailed description** (e.g. face, hair, nails and not just person)

- Therefore, **A1** and **A2** can be used in a **complementary** way

- **Similar conclusions** are drawn for the other **datasets**

EXPERIMENTAL RESULTS (BIAS MITIGATION) AFTER

Prediction performance per Gender after applying our proposed methodology (A1, AM-OD)

Doctor Class	Nurse Class
Male Accuracy: **92.7**%	Male Accuracy: 87.3%
Female Accuracy: **84.2**%	Female Accuracy: 86.7%

Chef Class	Waiter Class
Male Accuracy: **91.6**%	Male Accuracy: 89.5%
Female Accuracy: **75.8**%	Female Accuracy: 89.3%

Engineer Class	Farmer Class
Male Accuracy: **95.5**%	Male Accuracy: 98.8%
Female Accuracy: **93.3**%	Female Accuracy: 98.8%

Huge Mitigation of gender bias in the predictions in **doctor, chef** and **engineer** class

Big increase in **accuracy** in all **classes** for both **male** and **female** gender

EXPERIMENTAL RESULTS (BIAS MITIGATION) AFTER

Prediction performance per Gender after applying our proposed methodology (A2, AM-CT)

Doctor Class	Nurse Class
Male Accuracy: **91.1**%	Male Accuracy: 88.8%
Female Accuracy: **87.7**%	Female Accuracy: 88.2%

Chef Class	Waiter Class
Male Accuracy: **92.6**%	Male Accuracy: 88.7%
Female Accuracy: **76.8**%	Female Accuracy: 88.1%

Engineer Class	Farmer Class
Male Accuracy: **96.6**%	Male Accuracy: 95%
Female Accuracy: **94.4**%	Female Accuracy: 98.8%

Huge Mitigation of gender bias in the predictions in **doctor, chef** and **engineer** class

Big increase in **accuracy** in all **classes** for both **male** and **female** gender

EXPERIMENTAL RESULTS (BIAS MITIGATION)

Statistical parity (difference in accuracy w.r.t. gender) in **doctor**, **chef** and **engineer** class, **before** and **after applying our proposed methodology (A1** and **A2)**

Statistical Parity (Before)	Statistical Parity (After Approach 1, AM-OD)	Statistical Parity (After Approach 2, AM-CT)
Doctor: **14.8%**	Doctor: **8.5%** (**6.3% improvement**)	Doctor: **3.4%** (**11.4% improvement**)
Chef: **27.4%**	Chef: **15.8%** (**11.6% improvement**)	Chef: **15.8%** (**11.6% improvement**)
Engineer: **6.7%**	Engineer: **2.2%** (**4.5% improvement**)	Engineer: **2.2%** (**4.5% improvement**)

- **Huge mitigation** of **gender bias**
- **Balance** between **A1** and **A2** with respect to the **bias mitigation**
- **A2** produces **better results** in **doctor** class (**5.1% improvement** in **statistical parity**, in comparison to **A1**)

EXPERIMENTAL RESULTS (BIAS MITIGATION)

Improvement in **accuracy** for all **datasets after applying** our **methodology** (**A1** and **A2**)

Class + Gender	Improvement in Accuracy (A1, AM-OD)	Improvement in Accuracy (A2, AM-CT)
Doctor + Male	+5.3%	+3.7%
Doctor + Female	+11.6%	+15.1%
Nurse + Male	+2%	+3.5%
Nurse + Female	+2.2%	+3.7%
Chef + Male	+5.3%	+6.3%
Chef + Female	+16.9%	+17.9%
Waiter + Male	+4.1%	+3.3%
Waiter + Female	+4.3%	+3.1%
Engineer + Male	+0%	+1.1%
Engineer + Female	+4.5%	+5.6%
Farmer + Male	+4.4%	+0.6%
Farmer + Female	+2.2%	+2.2%

- In **7/12** cases **A2** produces **better results** with respect to **accuracy**
- In **only 4/12** cases **A1** produces **better results**
- In **1** case the **result** is **identical**

CONCLUSION

We proposed a methodology that enables us to:

- **Observe** whether there is **discrimination** in the **predictions** of a ML classification model with respect to the **gender** (**bias detection**)

- **Semantically describe** at **scale** the **reason** that a particular **prediction** of a **ML system** is **made** in a **human interpretable way** (**bias semantic interpretation**)

- **Compensate** for **gender bias** that is related with the **content** of the **image data** (**bias mitigation**)

FUTURE WORK

- Application to **different use-cases** (**visual tasks** that **bias** may **appear**)

- Application to **different machine learning tasks** (**Object Detection, Text Classification, Question Answering**)

- Application to **other forms** of **data** (**text, video**)

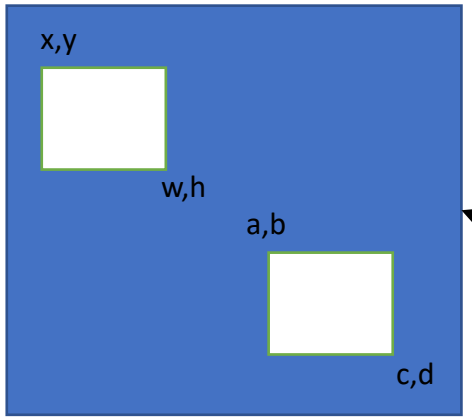
- **Creation** of a **new Image Dataset** (**images** do **not** contain specific **objects** that **give away** the **gender**)

QUESTIONS-FEEDBACK

EXTRA SLIDES

Correlation between Object Detection-Attention Mechanism and Crowd-Attention Mechanism

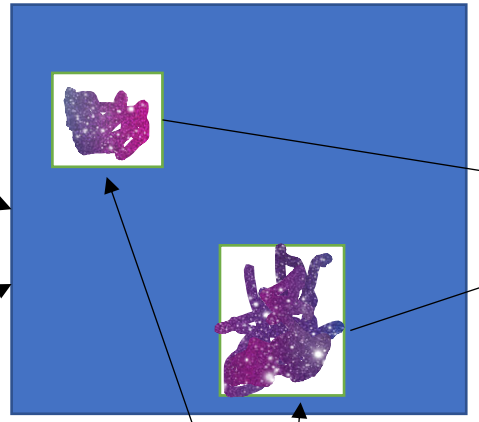
Object Detection -> Objects in the image -> **Bounding boxes** -> Coordinates



image

compare

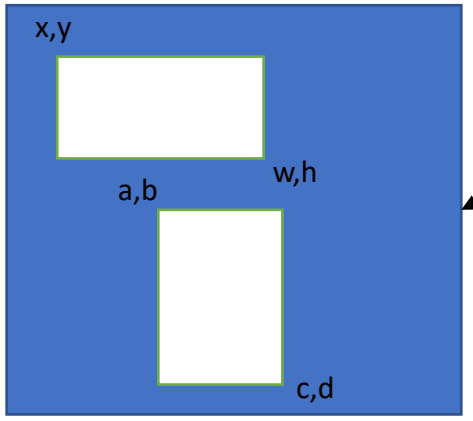
Attention Mechanism -> **Parts** of the image that the model picks up to make its prediction



Do **not** have them (bounding boxes)

image

Crowd -> Objects of gender bias -> **Bounding boxes** -> Coordinates



image

compare

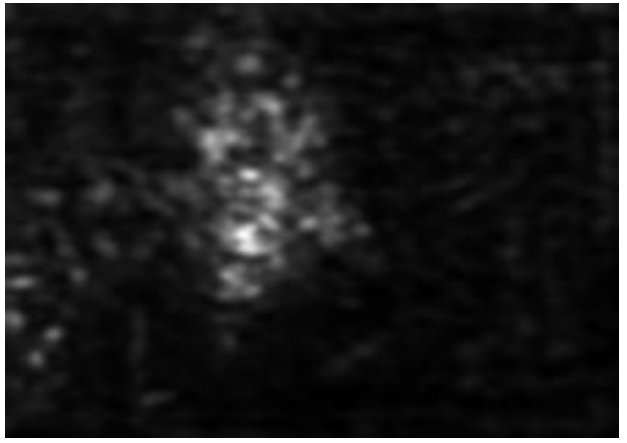
Problem:

How to find **AUTOMATICALLY** these bounding boxes (coordinates)?

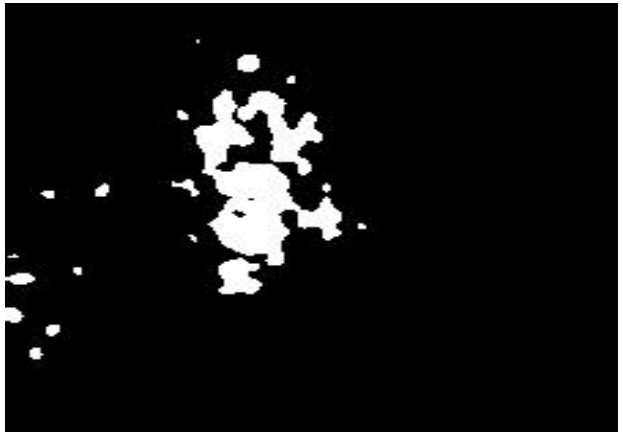
Solution to the problem

Proposed Methodology -> Detect multiple brightest spots in the images

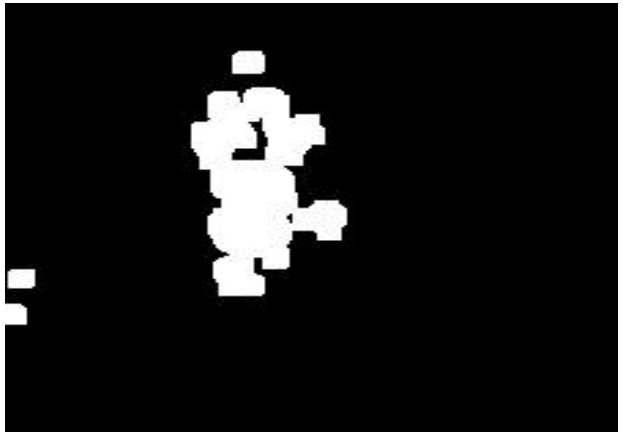
Step 1) Convert the resulted image of the Attention Mechanism to **grayscale** and smooth it (e.g. **blurring**) to reduce high frequency noise -> **Brightest** regions would be **more bright** and **less bright** regions would be more **dim**



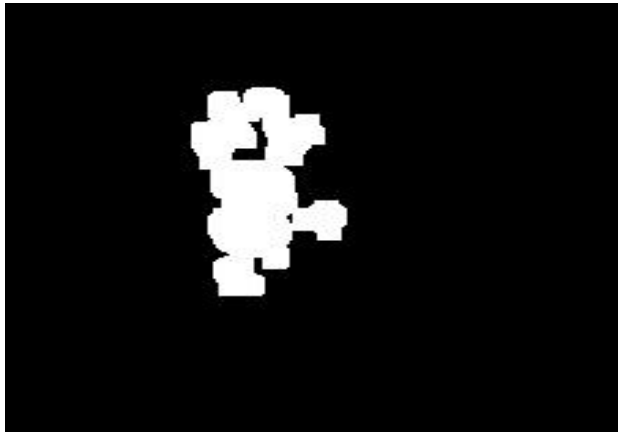
Step 2) Reveal the brightest regions in the blurred image -> Apply **thresholding**: Any pixel value *p >= pre-defined thresh* set it to white and pixel values *p < pre-defined thresh* are set to black



Step 3) Clean up the **noise** (i.e. small blobs) that is this image by performing a series of erosions and dilations



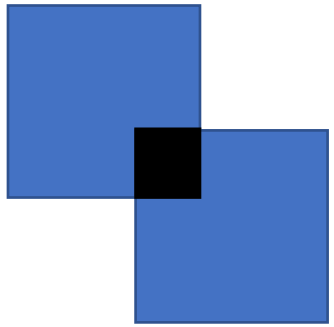
Step 4) Filter out any leftover “noisy” regions through performing a **connected-component analysis** ->
End up with a **mask** that has only the larger blobs in the image (which are also the brightest ones)



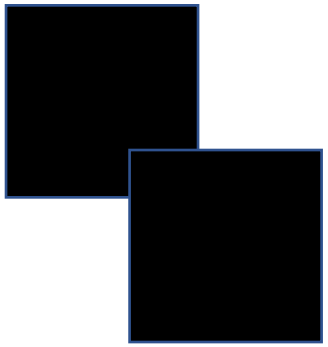
Step 5) Finally, we **detect** the **contours** in the mask and a **bounding box** is **automatically drawn** for each of them

Now that we also have a bounding box for each part in the image that the model picks up to make its prediction, we can compare Object Detection-Attention Mechanism and Crowd-Attention Mechanism

Proposed methodology: Intersection over Union (IoU)

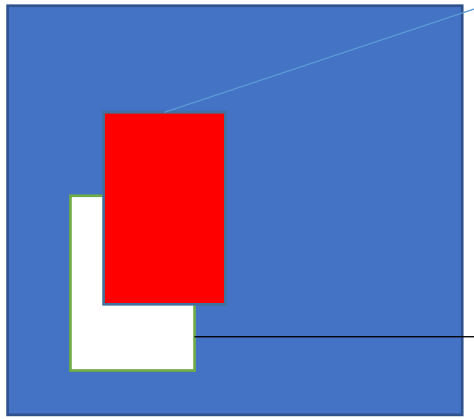


$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$



Object Detection or Crowd

image



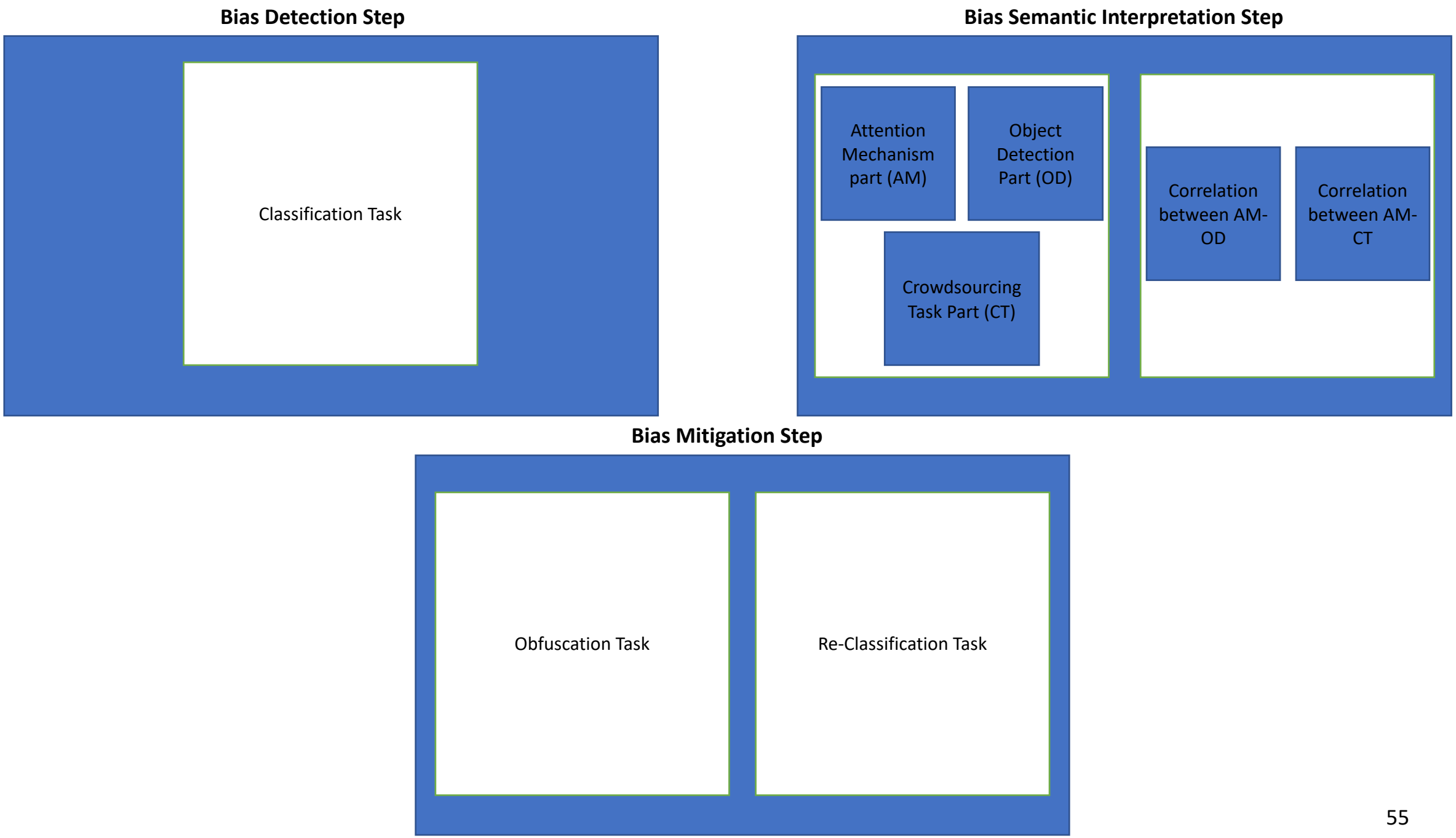
Attention Mechanism

Define threshold for IoU

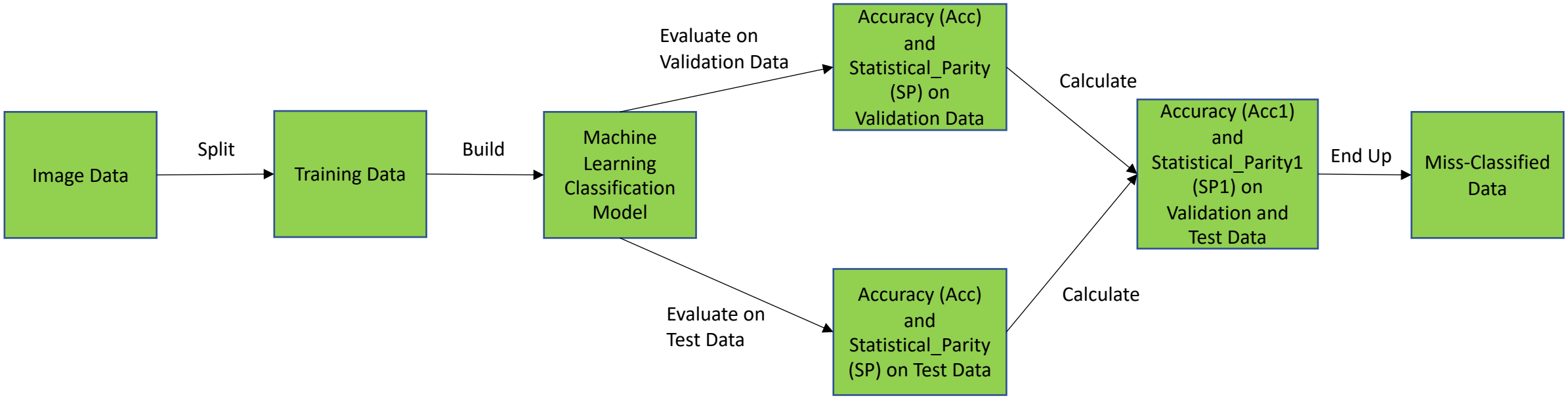
End up with objects (that come from Object Detection or Crowd) that **matter**
-> **List of objects and coordinates for each image**

Repeat this procedure for **each image** and for **each object** in the image

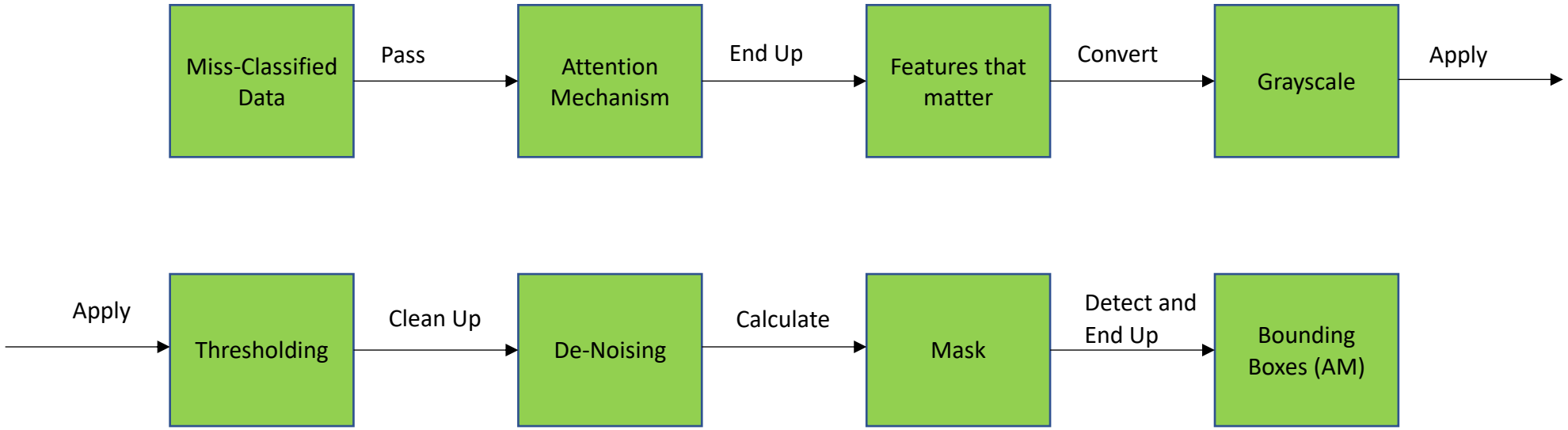
PIPELINE OF THE EXPERIMENTAL SET-UP


Bias Detection Step
Classification Task
Bias Semantic Interpretation Step
Attention Mechanism part (AM)
Object Detection Part (OD)
Crowdsourcing Task Part (CT)
Correlation between AM-OD
Correlation between AM-CT
Bias Mitigation Step
Obfuscation Task
Re-Classification Task

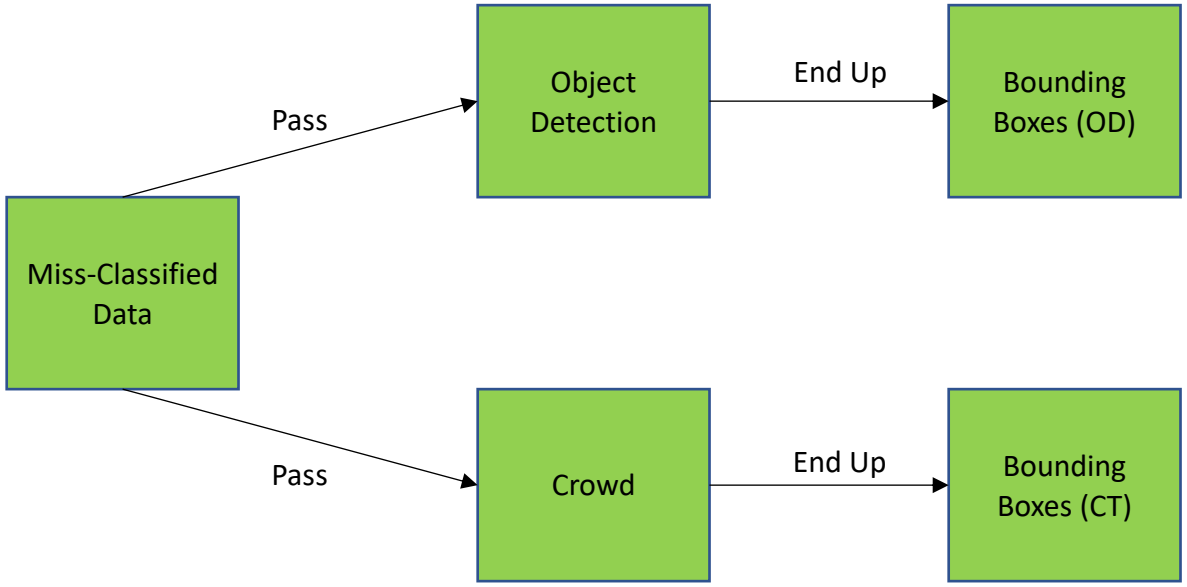
Classification Task (CLT) part



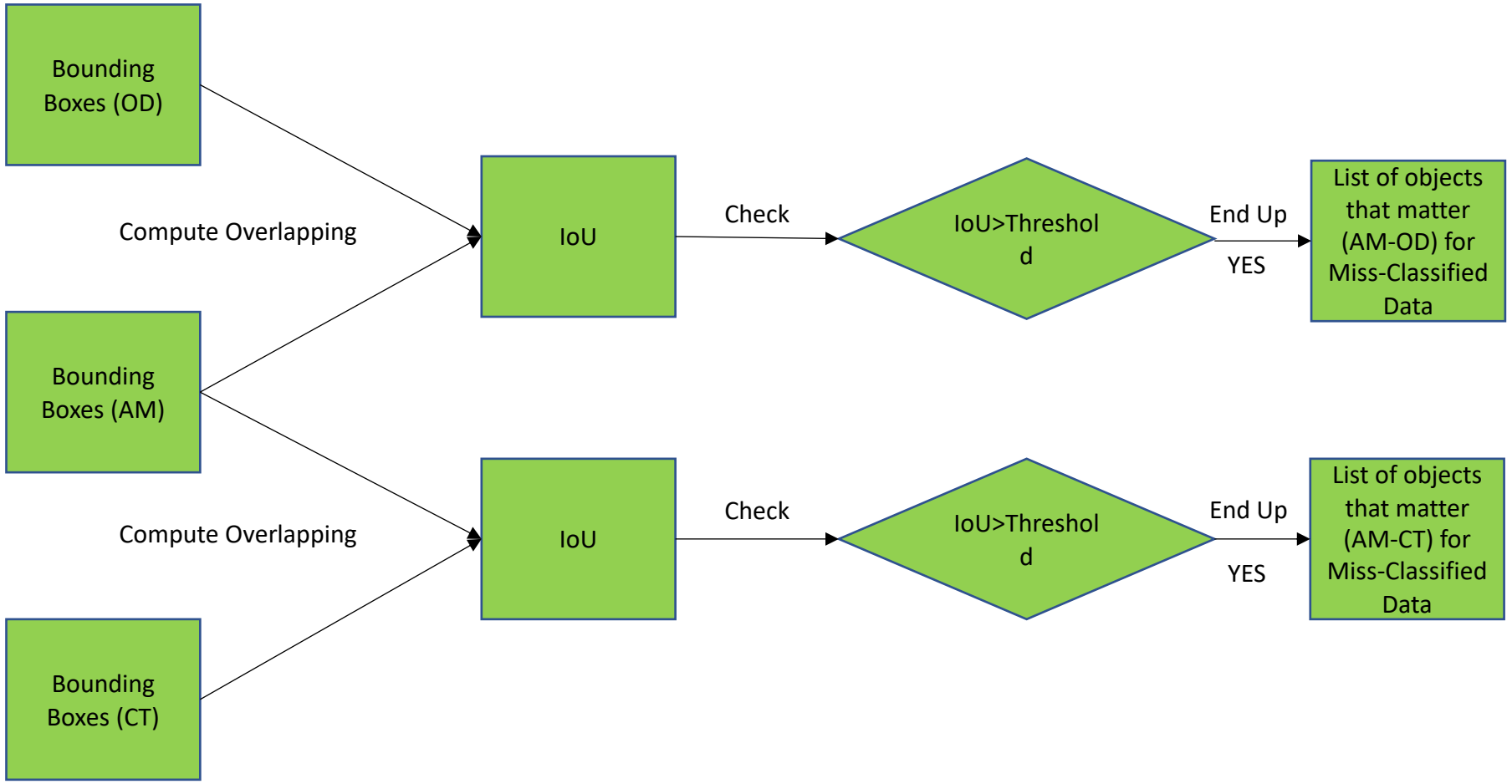
Attention Mechanism (AM) part



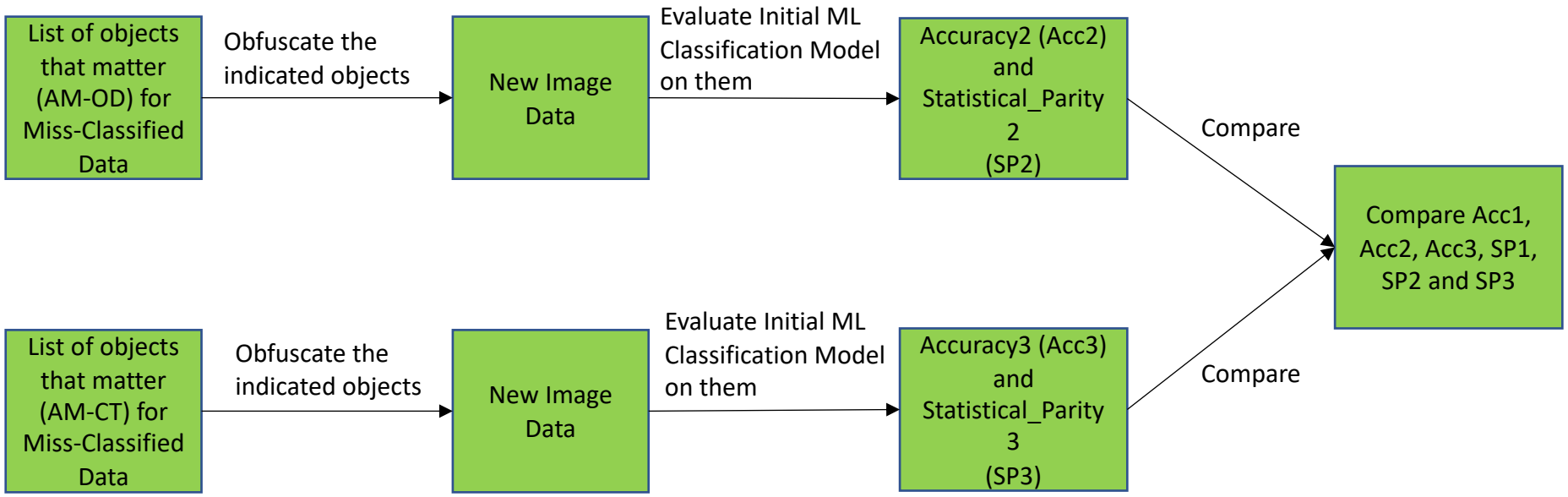
Object Detection (OD) and Crowdsourcing Task (CT) part



Correlation between AM-OD and AM-CT part



Obfuscation, Re-Classification Task and Comparison part



IMPLEMENTATION DETAILS

Information of the Datasets

Total number of data	Size of Training Set	Size of Validation Set	Size of Test Set
1000 (500 males and 500 females)	640	180	200

Doctor-Nurse

Total number of data	Size of Training Set	Size of Validation Set	Size of Test Set
1000 (500 males and 500 females)	640	180	200

Chef-Waiter

Total number of data	Size of Training Set	Size of Validation Set	Size of Test Set
1000 (500 males and 500 females)	640	180	200

Engineer-Farmer

Classification Task

- Used a **Pre-trained ResNet Model** trained on the ImageNet dataset
- **Data augmentation** configuration (rotate, rescale, flip etc. the data)
- **Transfer Learning** approach (used the convolutional layers for feature extraction)
- Add a **classifier on top** of the convolutional base (add a fully connected layer followed by a softmax layer with 2 outputs)
- Split Training data into 70% training set and 30% validation data and optimize the **hyperparameters** on that

Classification Task

Tweak of the **Hyperparameters** (optimized on the validation set)

Hyperparameters	**Value**
Learning Rate	10^{-3}
Optimizer	SGD
Multiplicative factor of learning rate decay	0.1
Period of learning rate decay	7
Dropout of the FC layers	0.5
Activation Function	Relu
Loss Function	Binary Cross Entropy
Momentum	0.9

Attention Mechanism

- **SmoothGrad** for finding features that matter
- **Connected-Component Analysis** for automatic bounding box drawing per such a feature

“Smilkov, Daniel & Thorat, Nikhil & Kim, Been & Viégas, Fernanda & Wattenberg, Martin. (2017). SmoothGrad: removing noise by adding noise”

“Hossain, Eftekhar & Anisur Rahaman, Mohammad. (2018). Detection & Classification of Tumor Cells from Bone MR Imagery Using Connected Component Analysis & Neural Network. 1-4. 10.1109/ICAEEE.2018.8642973”

Object Detection

- Used a **Pre-trained Yolo Detector**

“Redmon, Joseph et al. “You Only Look Once: Unified, Real-Time Object Detection.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 779-788.”

- Experiments performed by using weights from pre-training on ImageNet, COCO, Pascal Voc, Open Images dataset and combination of these four

- Best Results obtained from weights coming from **COCO dataset** (80 classes)

Crowdsourcing

- Figure-Eight platform
- 3 crowdworkers (coming from USA) per data point
- Each task had 5 images
- Total number of Tasks: 40