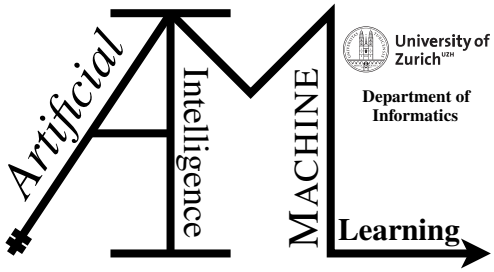
Visualization of Facial Attribute Classifiers via Class Activation Mapping

Bachelor Thesis

Johanna Bieri
18-731-661

Submitted on
August 9 2023

Thesis Supervisor
Prof. Dr. Manuel Günther



Bachelor Thesis

Author: Johanna Bieri, johanna.bieri@uzh.ch

Project period: February 9 2023 - August 9 2023

Artificial Intelligence and Machine Learning Group
Department of Informatics, University of Zurich

Acknowledgements

First and foremost I would like to thank my supervisor Prof. Dr. Manuel Günther for his guidance and support throughout the process. I always appreciated his valuable advice and patience. I would also like to thank him for giving me the opportunity to dive into the topic of machine learning even without much prior knowledge. Not only did I learn about machine learning but Prof. Günther also taught me how to approach such a big scientific project.

Furthermore, I would like to express my gratitude to my friends, family, and flatmates who supported me in various ways. A special thank you goes to Alex and Manuel.

Abstract

The use of convolutional neural networks (CNNs) in image classification tasks is a rapidly progressing field of research, including the classification of facial attributes. However, it is not yet completely understood how CNNs make decisions. To improve the transparency of the decision-making process and thus enhance interpretability and trustworthiness of CNNs, methods have been developed to visualize this process. In this thesis, we use the Gradient-weighted Class Activation Mapping (Grad-CAM) technique proposed by Selvaraju et al. (2017) to identify the regions of an image that the CNN uses for classification. This technique produces class-specific heatmaps that are intuitively interpretable. In order to evaluate the class activation maps, we define a set of masks, one for each of the 40 facial attributes that we examine. By using an approach called Acceptable Mask Ratio (AMR) we quantify how much of the activated area lies within the masked area. The higher the value of the AMR the more active is the CNN within the area that we expect, which usually corresponds to the location of the attribute being classified. We compare two different CNNs, one considers the class imbalance inherent to the data set (balanced CNN), and the other does not (unbalanced CNN). Our results show that overall the balanced CNN more often uses image regions that lie within the masked area. Furthermore, the results show an unexpected pattern for the unbalanced CNN namely for highly biased attributes the Grad-CAMs for the majority class show no activity at all.

Contents

Chapter 1

Introduction

Convolutional neural networks (CNNs) have emerged as powerful tools for image classification tasks because of their ability to recognize patterns. Various techniques have been explored to improve CNN performance, including deeper architectures, residual connections (He et al., 2016), and making use of large-scale data collections like CelebA (Liu et al., 2015).

Despite the tremendous progress that has been made in the past years, it is still not fully understood what neural networks learn exactly, i.e. which part of an image they use to make predictions. One approach to better understand which regions of an image a CNN uses to classify it is by making use of a technique called Class Activation Mapping (CAM) introduced by Zhou et al. (2016). It is able to identify the class-specific regions of an image, which the CNN uses for classification. The method used in this thesis is called Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017), which is similar to the CAM method, but it makes use of the gradients. By combining the gradient information with the feature map activations, this technique generates a heatmap that highlights the regions in the input image that are crucial for the network's decision-making process.

In this thesis, we perform facial attribute classification. This differs from default categorical image classification in a few aspects. For categorical image classification, the output is a single value representing the probability with which the image belongs to the predicted class or category. For example, classifying images of cats and dogs into the categories "cat" and "dog". In facial attribute classification, on the other hand, the output is a set of binary scores indicating the presence or absence of specific attributes. For example, predicting attributes like "Smiling", "Male" or "Rosy Cheeks", each with its own binary prediction.

We use a binary classifier to perform facial image classification on the CelebA dataset predicting 40 binary facial attributes in each image. As not all of the attributes are equally present in the CelebA dataset we face the problem of class imbalance. Few attributes are overrepresented and many are mostly absent throughout the dataset. This leads to a biased classifier because it sees one of the two classes much more often during the training process. Rudd et al. (2016) showed that such a biased classifier performs well on the majority class but rather poorly on the minority class.

The experiments are conducted using two CNNs. They differ in that one considers the class imbalance (balanced network) and the other one does not (unbalanced network). Both of them employ the Alignment-Free Facial Attribute Classification Technique (Günther et al., 2017) to extract facial attributes. By utilizing the Grad-CAM method to visualize the discriminative image regions we examine the following questions:

1. Is the classification accuracy correlated to a Grad-CAM that highlights the expected region[1]

[1] We expect the network to use the part of the image where the attribute being classified is located. E.g. when classifying the attribute 'Bushy Eyebrows' we expect the network to focus on the eye region rather than for example the

for the balanced and the unbalanced network?

2. Are the balanced and the unbalanced network more likely to rely on the expected location when classifying a sample correctly compared to classifying it incorrectly?
3. For the balanced and the unbalanced network: Do the Grad-CAMs for images where the attribute is present highlight other regions than the Grad-CAMs for images where the attribute is absent?
4. Does the balanced network more often rely on the expected location than the unbalanced one?
5. Do the Grad-CAMs for the balanced and the unbalanced network differ?

mouth region.

Chapter 2

Related Work

2.1 Facial Attribute Classification

Over the past few years, there has been a growing focus on the prediction and classification of facial attributes, primarily driven by the wide range of applications and extensive utilization of facial attribute information for example for face verification (Kumar et al., 2009), face recognition (Chan et al., 2017), face image retrieval (Nguyen et al., 2018), or face re-identification (Su et al., 2018).

The progress in this research field is strongly influenced by the availability of large-scale data collections of face images, such as Labeled Faces in the Wild (Huang et al., 2007) containing roughly 13k images or the even larger dataset CelebA (Liu et al., 2015) which consists of over 200k images. The latter provides an extensive collection of images that exhibit a natural variability in aspects like lighting, pose, expression, background, and other parameters. Furthermore, the CelebA dataset is of great value for facial attribute classification because every image contains manually annotated labels for all the attributes. For supervised machine learning[1] in particular hand-labeled datasets represent a highly valuable resource because they offer a ground truth reference for training neural networks and evaluating their performance.

As Günther et al. (2017) point out, contrary to many other facial features used in the field of face recognition, facial attributes have a unique characteristic, they have a semantic meaning. This property makes them interpretable for humans while also being detectable for neural networks.

Currently, the state-of-the-art method for classifying facial attributes is the utilization of CNNs. They are a type of artificial neural network that consists of at least one convolutional layer, pooling layers, and fully-connected layers (O'Shea and Nash, 2015). They find their primary application in the domain of image pattern recognition, which is essential in facial attribute recognition and classification.

According to Mao et al. (2020) there are two different approaches for facial attribute classification, namely single-label learning based methods which make predictions for each attribute separately, thereby disregarding the relationships or correlations among them, and multi-label learning based methods where multiple facial attributes are predicted concurrently. An example of the use of the first approach is Zhong et al. (2016). But in the context of this thesis, we are more interested in the second approach, since it is also applied in our work. Examples of the multi-label approach are Hand et al. (2018) or Zhuang et al. (2018). Rudd et al. (2016) as well applied the multi-label approach and they proposed a solution for the problem of imbalanced labels[2] for

[1] In the case of supervised machine learning the neural network is trained with pre-labeled data (O'Shea and Nash, 2015).

[2] Class imbalance occurs when the number of samples for one class is significantly larger than that for the other class for one attribute (Mao et al., 2020).

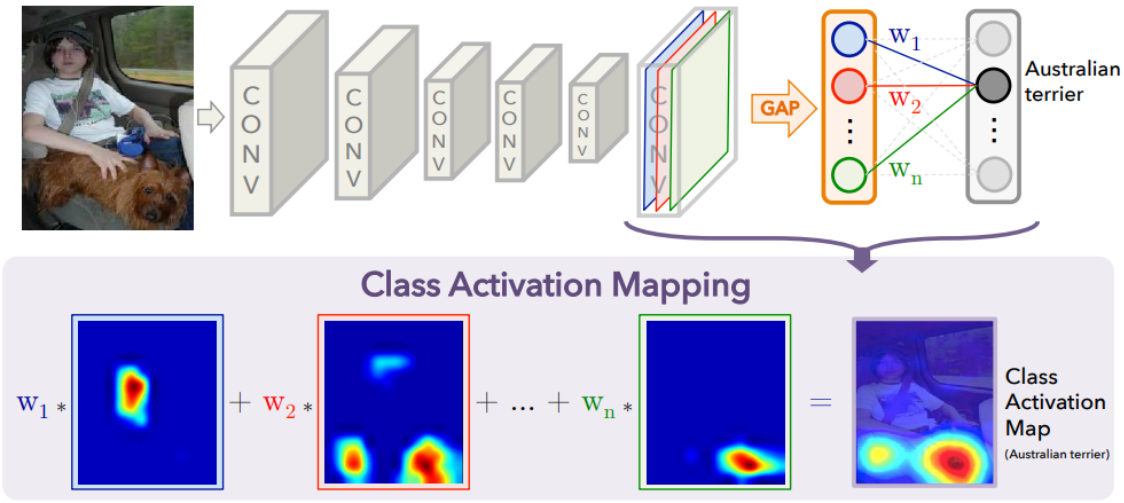


Figure 2.1: CLASS ACTIVATION MAPPING. *This figure by Zhou et al. (2016) illustrates the process of Class Activation Mapping. The class score predicted by the model is mapped back to the last convolutional layer. The resulting Class Activation Map highlights the important image region for the predicted class.*

some attributes, a difficulty inherent to this approach. They dealt with the class imbalance problem by introducing a novel multi-objective neural network architecture, which mixes the tasks of multi-label classification and domain adaptation under one unified objective function. Another example of the multi-label approach is Günther et al. (2017) who introduced the Alignment-Free Facial Attribute Classification Technique (AFFACT). The two CNNs used in this thesis both use this technique. This will be discussed in more detail in section 3.2.

2.2 Visualizing CNNs

When it comes to making the decision-making process of CNNs more transparent several approaches exist. Gradient-based methods (Simonyan and Zisserman, 2014), perturbations-based methods (Ribeiro et al., 2016), and CAM-based methods are among the most popular ones. For the sake of brevity, we will focus on the third kind. CAM stands for *Class Activation Mapping*, a technique that computes a weighted linear sum of the feature maps of the last convolutional layer which results in a heatmap highlighting the regions most relevant for the predicted class (Figure 2.1). After the introduction of this visualization technique by Zhou et al. (2016) researchers have been developing other CAM-based methods attempting to further improve it. The primary distinguishing element among various CAM techniques lies in the approach employed to calculate the mentioned weights, which as said before are utilized in combination with the feature maps obtained from the convolutional layer being targeted (not always the last layer) (Dugăeșescu and Florea, 2022). Selvaraju et al. (2017) introduced the *Gradient-weighted CAM* (Grad-CAM) method that uses the gradients to compute the weights. Compared to the original CAM method, Grad-CAM eliminates the need for a penultimate[3] global average pooling[4] layer directly following the last convolutional layer, thus making it available to a much broader

[3] As the penultimate layer the global average pooling layer passes its output to the fully-connected layer.

[4] Global average pooling is a dimensionality reduction technique that reduces each feature map to one single value. It is explained in more detail in section 3.3.

range of CNN architectures. However, Grad-CAM fails when there is more than one occurrence of a specific class within one image. Furthermore, due to working with an unweighted average of the gradients Grad-CAM often lacks to localize every part of an object. Building on the Grad-CAM method Chattopadhay et al. (2018) presented *Grad-CAM++*, a slightly different approach that takes a weighted combination of the gradients instead of a global average (like Grad-CAM) which addresses the problem of partial object localization. Two years later Draelos and Carin (2020) proposed *HiRes-CAM*, a method similar to Grad-CAM but instead of averaging over the gradients they use them directly as weights. Another Grad-CAM-based approach is the so-called Axiom-based Grad-CAM or *XGrad-CAM* introduced by Fu et al. (2020), further improving Grad-CAM and providing clear theoretical support for it. One year later Jiang et al. (2021) proposed a method called *Layer-CAM* which uses the class-specific gradients as weights in case of positive gradients, for negative gradients the weight is zero.

In view of the fact that gradient-based CAM methods too have their drawbacks, that is gradient saturation[5] or false confidence[6], gradient-independent approaches such as *Score-CAM* (Wang et al., 2020), *Eigen-CAM* (Muhammad and Yeasin, 2020) or *Ablation-CAM*(Desai and Ramaswamy, 2020) have been developed.

However, all of these techniques typically visualize the class with the highest predicted probability for categorical classification tasks. So far there has been little research on visualizing binary classifiers such as the ones used in this thesis.

[5]The gradient saturation problem causes the backpropagating gradients to diminish and therefore, adversely affect the quality of visualizations (Desai and Ramaswamy, 2020).

[6]In cases of false confidence activation maps with higher weights show lower contribution to the network's output compared to a zero baseline (Wang et al., 2020).

Chapter 3

Background

This chapter talks about the data set, the CNNs, and the key methods used in this thesis.

3.1 Dataset

The CelebA dataset is the most widely used image dataset in the field of facial attribute classification. It originates from the CelebFaces dataset by Sun et al. (2014) and has been constructed by Liu et al. (2015) as a large-scale data collection for the training of CNNs. It contains over 200k images of approximately 10k different identities. In the field of machine learning it is common to split the dataset into three distinct sets: the training set, the validation set, and the test set. Each of these sets serves a specific purpose in the process. The CelebA dataset as well is divided into three partitions. With 162'770 images (roughly 80%) the training set is by far the largest of the three and is used to train the model. It is the set on which the model learns patterns, relationships, and rules to make predictions. The validation set consists of 19'867 images (roughly 10%) and is typically used to fine-tune the hyperparameters[1], e.g. adjust the loss function. By evaluating the model's performance on the validation set, one can make adjustments to achieve better generalization and avoid overfitting. The validation set helps in selecting the best-performing model architecture and configuration. Eventually, the model is then tested on the test set. In the case of the CelebA dataset, it is about the same size as the validation set, containing 19'962 images (roughly 10%). It represents new, unseen images of identities that the model has not encountered during training or validation. The test set's results give an indication of the model's true performance and help estimate its effectiveness in real-world applications. Since the experiments in this thesis were done using a pre-trained network and there was no need for retraining, only validation and test partition were used.

Each of the images of this dataset was hand-labeled with a binary label for each of the 40 facial attributes provided in Figure 3.1. These labels represent the ground truth. They serve as a reference to measure the accuracy of the model. Furthermore, each image was annotated with 5 facial landmarks. They mark the location of the eyes, the tip of the nose, and the corners of the mouth.

When working with the CelebA dataset it is important to consider its class imbalance. As shown in Figure 3.1 many of the attributes such as "Bald", "Double Chin" or "Chubby" can hardly be found in the dataset, while others such as "No Beard" or "Young" are overrepresented.

[1] Hyperparameters such as the depth or the stride are used to optimize the model output.

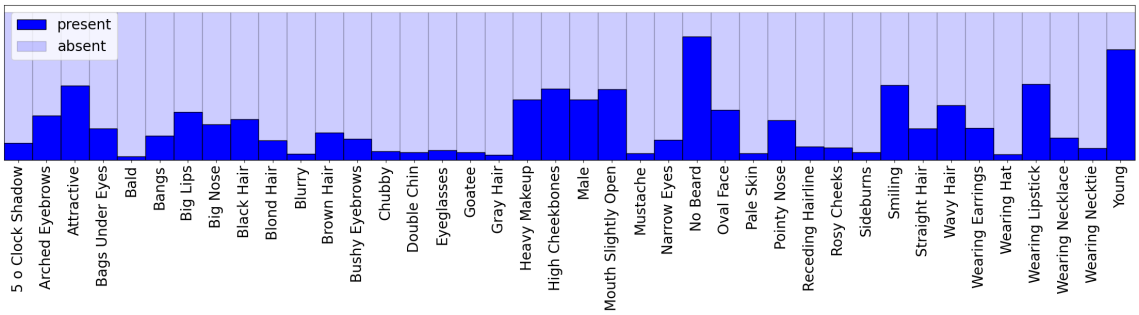


Figure 3.1: DISTRIBUTION OF ATTRIBUTES. *This figure shows the distribution of the binary facial attributes throughout the CelebA dataset.*

3.2 CNNs

Residual networks (ResNets) are deep CNNs that use a residual learning framework introduced by He et al. (2016). This learning framework enables the training of significantly deeper networks and facilitates their optimization, considering that deeper neural networks are more challenging to train. Furthermore, ResNets overcome the problem of vanishing or exploding gradients, by inserting shortcut connections that skip a few layers. The skipped layers form a residual block that allows the network to learn residual functions which represent the difference between the input and output of a residual block. The ResNet architecture is typically composed of multiple stacked residual blocks.

For this thesis, a CNN employing the Alignment-Free Facial Attribute Classification Technique (AFFACT) (Günther et al., 2017) to extract attributes was used. AFFACT is a data augmentation technique that applies random perturbations such as scaling, rotating, shifting, and blurring of images during training. This approach makes the network less dependent on the alignment of images. Predictions are made solely based on detected bounding boxes. Introduced by Günther et al. (2017), the AFFACT network is based on a ResNet-50, which is a ResNet containing 50 layers. After pre-training the ResNet-50 on the ILSVRC2012 subset of Deng et al. (2009)'s ImageNet dataset it was modified by adding an extra fully connected convolutional layer with 40 output units (matching the 40 attributes), resulting in a ResNet-51 (Günther et al., 2017). In order to fine-tune the network for facial attribute classification it subsequently was trained on the CelebA dataset.

In this thesis, two CNNs using AFFACT were used. One, taking the class imbalance (Figure 3.1) present in the CelebA dataset into consideration, and another one, which does not. The former is referred to as the balanced network or AFFACT-b and the latter as the unbalanced network or AFFACT-ub. The reason for using two CNNs is that the AFFACT-ub while performing well on majority class samples that dominate the training data, it often fails to correctly classify samples of the minority class. Rudd et al. (2016) dealt with this by presenting a Mixed Objective Optimization Network (MOON) that uses a domain-adapted multitask loss function (3.1). This network architecture incorporates attribute correlations and is able to adapt the bias of the training dataset (comparable to the bias in Figure 3.1) to a desired target distribution.

$$L(\mathbf{X}, \mathbf{Y}) = \sum_{j=1}^{N} \sum_{i=1}^{M} p(i|Y_{ji}) \, ||f_i(X_j) - Y_{ji}||^2 \tag{3.1}$$

M represents the number of attributes, $\mathbf{X}$ the data tensor containing N input images, and $\mathbf{Y}$ a $N \times M$ matrix with the corresponding labels. For each attribute i with target value $Y_{ji} \in \{-1, +1\}$

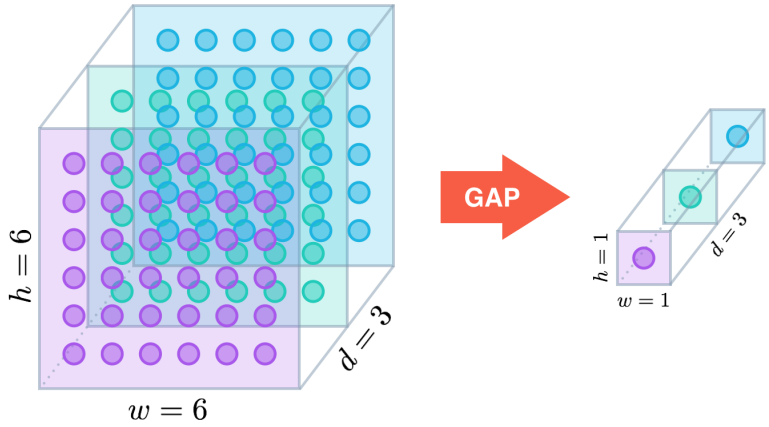


Figure 3.2: GAP. *This figure*[2] *shows how Global Average Pooling reduces dimensions. Each* $h \times w$ *feature map is reduced to a single value and this is done for all channels* d.

the error is only backpropagated with the probability $p(i|Y_{ji})$ (derived via sampling), otherwise the gradient for attribute i is set to 0. The more the distribution of the attributes in the training set differs from the desired target distribution, the more elements in the gradient are reset (Rudd et al. (2016)). The AFFACT-b was designed based on a similar approach but remains yet unpublished.

3.3 Global Average Pooling

Global Average Pooling (GAP) is a pooling operation commonly used in CNNs performing image classification tasks. It was first introduced by Lin et al. (2014) and has since become a popular alternative to traditional pooling techniques like Global Max Pooling (GMP). The main idea behind GAP is to summarize the entire feature map generated by the last convolutional layer into a single value for each channel (Figure 3.2).

$$v_k = \frac{1}{w \cdot h} \sum_{x,y} f_k(x, y) \tag{3.2}$$

Here v_k represents the average of the feature map $f_k(x, y)$ at spatial location (x, y) for channel k (equal to d in Figure 3.2). This is done for all channels which results in a fixed-length vector representation (Figure 3.2 right side) that can be fed into a fully-connected layer for classification.

Zhou et al. (2016) showed that GAP outperforms GMP when it comes to object localization. GMP aims to identify only one discriminative region. Whereas in GAP the value can be maximized by finding all discriminative regions because all low-activation regions reduce the output of the map. This is the reason why Zhou et al. (2016)'s Class Activation Mapping technique is designed for a network architecture that contains a GAP layer. They use this characteristic of GAP to enable their technique to localize objects.

As shown in Figure 3.3 the network architecture used in this thesis also contains a GAP layer right before the fully-connected layer.

[2]Figure from: https://alexisbcook.github.io/assets/global_average_pooling.png

3.4 Class Activation Mapping

Class Activation Mapping (CAM) is a technique introduced by Zhou et al. (2016). It reveals the specific region in an image that a CNN utilizes to discern and identify a particular category. Given a network architecture that performs a GAP on the feature maps of the last convolutional layer and then passes them to a fully-connected layer to produce the final output, this technique determines the significance of image regions by projecting the weights of the output layer back onto the convolutional feature maps. The GAP layer calculates the spatial average of the feature map for each channel in the last convolutional layer. A weighted sum of these values is then employed to generate the final output. Likewise, the class activation map is obtained by computing a weighted sum of the feature maps in the last convolutional layer (Figure 2.1).

As stated in (3.3), every spatial element M^c_{CAM} of the class activation map for class c at spatial location (x, y), is calculated as the sum of the weighted activations f_k, where w^c_k represents the weight factor, summed up over all channels k. Thus M^c_{CAM} indicates the importance of the activation for the classification of an image to class c. The weights w^c_k equal the predicted class score that the fully-connected layer outputs for each class.

$$M^c_{CAM}(x, y) = \sum_k w^c_k f_k(x, y) \tag{3.3}$$

Gradient-weighted Class Activation Mapping is a generalization to CAM since it does not require a specific network architecture[3] or retraining[4] (Selvaraju et al., 2017). As described in (3.4) Grad-CAM computes the gradients by taking the partial derivative of the logit y^c for class c (before the softmax function[5]), with respect to the partial derivative of the feature map of the last convolutional layer $f_k(x, y)$ for channel k at spatial location (x, y), i.e. $\frac{\partial y^c}{\partial f_k(x,y)}$. Similarly to the feature maps passing the GAP layer in the classification process, these gradients are global average pooled to derive the neuron importance weight α^c_k corresponding to class c. Analogous to w^c_k in (3.3) α^c_k represents the importance of feature map f_k for target class c.

To obtain the final Grad-CAM $M^c_{Grad-CAM}$ a Rectified Linear Unit Activation Function[6] (ReLU) is applied to the weighted linear sum of the feature maps (3.5), because Grad-CAM is only interested in pixels that increase the score y^c when their intensity is increased. Negative pixels are likely to belong to another category (Selvaraju et al., 2017).

$$\alpha^c_k = \overbrace{\frac{1}{w \cdot h} \sum_{x,y}}^{\text{GAP}} \underbrace{\frac{\partial y^c}{\partial f_k(x, y)}}_{\text{gradient}} \tag{3.4}$$

$$M^c_{Grad-CAM}(x, y) = ReLU\left(\sum_k \alpha^c_k f_k(x, y)\right) \tag{3.5}$$

[3] CAM is limited to CNNs with a Global Average Pooling penultimate layer.

[4] CAM requires retraining of one linear classifier for each class (Chattopadhay et al., 2018).

[5] The softmax function converts the logits to probabilities, i.e. normalizes them. The AFFACT network architecture does not contain a softmax function, it directly outputs the logit for each attribute.

[6] $ReLU(x) = max(0, x)$ (from https://pytorch.org/docs/stable/generated/torch.nn.ReLU.html)

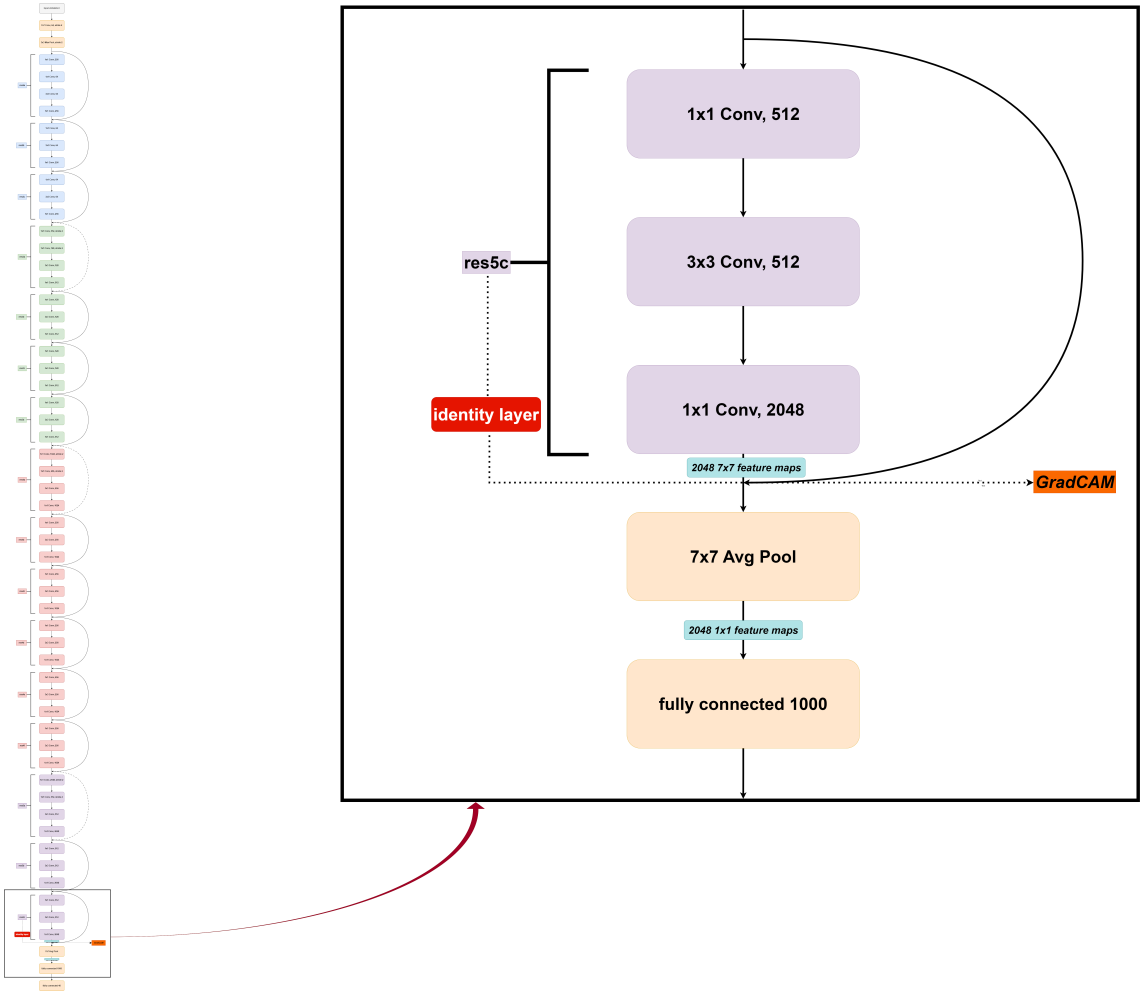


Figure 3.3: AFFACT ARCHITECTURE WITH GRAD-CAM. *This figure shows the AFFACT network architecture with 16 residual blocks and shows where in the network architecture the Grad-CAM method acts.*

Chapter 4

Approach

4.1 Image Preprocessing

For the experiments in this thesis, the images of the CelebA dataset were preprocessed in the same way as in Günther et al. (2017). First, the center of the eyes t_e and the center of the mouth t_m along with their respective distance d are computed using the landmarks of the eyes t_{el}, t_{er} and the mouth corners t_{ml}, t_{mr}, with $t = (x, y)^T$:

$$\vec{t_e} = \frac{\vec{t_{er}} + \vec{t_{el}}}{2},\ \vec{t_m} = \frac{\vec{t_{mr}} + \vec{t_{ml}}}{2},\ d = ||\vec{t_e} - \vec{t_m}|| \tag{4.1}$$

After aligning the eyes horizontally, a square bounding box with top left corner x_l, y_t and bottom right corner x_r, y_b and edge length l is added:

$$x_l = x_e - 0.5 \cdot l,\ \ y_t = y_e - 0.45 \cdot l$$
$$x_r = x_e + 0.5 \cdot l,\ \ y_b = y_e + 0.55 \cdot l$$

Finally, the images are cropped to the size of the bounding box which results in 224 x 224 pixels, and are saved as PNGs.

4.2 Landmarks

As described in section 3.1 each image was annotated with five landmarks: eyes, tip of the nose, and mouth corners. Because these landmarks were marked before cropping the images to their bounding box, they had to be shifted in order to match the cropped images that were used in our experiments. Since the image size was known, a first estimation of how much the landmarks had to be shifted could be done. After some fine-tuning, the results were evaluated by visually checking a set of images that was randomly chosen from the data set. Eventually, they were shifted as follows:

$$x_{shifted} = x_{orignial} + 24$$
$$y_{shifted} = y_{original} - 10$$

The result can be found in Figure 4.1.

Figure 4.1: LANDMARKS. *This table shows the original landmarks as yellow dots (top) and the shifted landmarks as green dots (bottom).*

4.3 Grad-CAM

The implementation[1] of the Grad-CAM technique that was used in this thesis needed some adjustments in order to fit our experiments.

4.3.1 Layer

As described in section 3.4 the feature maps of the last convolutional layer are used to derive the class activation map. In the AFFACT network used in this thesis, this would correspond to the output of the last residual block `res5c` (Figure 3.3). For implementational reasons, these feature maps were initially not accessible. In order to access them we inserted an identity layer using PyTorch's identity operator[2]. The following line was inserted after loading the model:

```
model.identity = torch.nn.Identity()
```

Listing 4.1: Inserting an identity layer

This line was inserted in the forward pass function to retrieve the data stored in the variable `res5c`:

```
res5c = self.identity(res5c)
```

Listing 4.2: Retrieve data from res5c via identity operator

4.3.2 Target Function

The Grad-CAM method contains a target function that is used to calculate the loss. At first, we worked with the already implemented `ClassifierOutputTarget` target function. As shown in Figure 4.2, it seems to not work properly whenever the prediction is negative (referring to the activation in the lower left corner). If the prediction is negative, the output of the target function is negative as well, since it simply returns the output of the model. We did not only observe this behavior with the `ClassifierOutputTarget` but also with other target functions that we implemented in attempting to solve this problem (Figure A.1, Figure A.2, Figure A.3). The Grad-CAM method seemed to fail whenever the output of the target function was negative. Since all the implementations we tested showed the same behavior for negative values, we decided to work with the implementation shown in Listing 4.3 which always returns a positive value (Figure 4.3).

[1] The source code is available online: https://github.com/jacobgil/pytorch-grad-cam

[2] More information can be found at the following link: https://pytorch.org/docs/stable/generated/torch.nn.Identity.html

Figure 4.2: AVERAGE GRAD-CAMS EXAMPLES. *Average Grad-CAMs from AFFACT-ub for different attributes when prediction > 0 (top) and when prediction < 0 (bottom). Attributes from left to right: "Bags Under Eyes", "Brown Hair", "Oval Face", "Wavy Hair", "Wearing Necklace".*

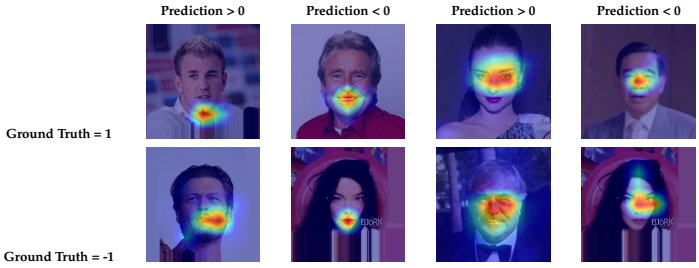


Figure 4.3: TARGET FUNCTION 2. *Grad-CAMs when the target function returns the absolute model output for "5 o Clock Shadow" (left square) and "Bushy Eyebrows" (right square).*

This approach works for the AFFACT-b but for classification with the AFFACT-ub the Grad-CAM method still returns Grad-CAMs with a sole activation in the corners (Figure A.5).

```
1 class BinaryCategoricalClassifierOutputTarget:
2     def __init__(self, category):
3         self.category = category
4 
5     def __call__(self, model_output):
6         if len(model_output.shape) == 1:
7             return abs(model_output[self.category])
8         return abs(model_output[:, self.category])
```

Listing 4.3: Implementation of target function 2

4.4 Evaluation Metrics

The classification accuracy is evaluated by calculating the classification error rate.

$$ErrorRate(a) = \frac{1}{N}\sum_{n=1}^{N} e(pr_{a,n}, l_{a,n}) \tag{4.2}$$

$$e(pr, l) = \begin{cases} 1, & \text{if } (pr \cdot l) < 0 \\ 0, & \text{if } (pr \cdot l) > 0 \end{cases} \tag{4.3}$$

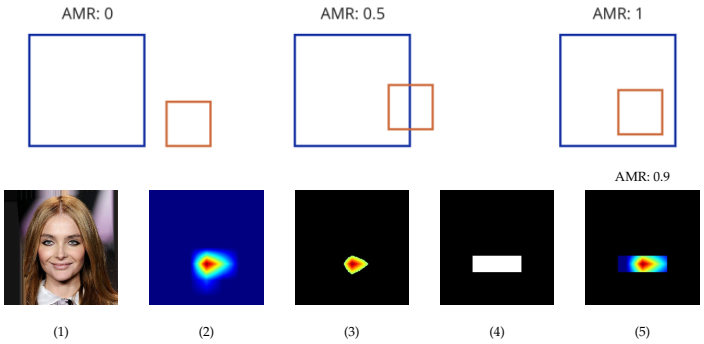


Figure 4.4: AMR. ***Top:*** *this figure by Chen (2022) illustrates the concept of AMR. The blue box represents the masked area, the red box represents the discriminative area.* ***Bottom:*** *these are the different components that are needed to calculate the AMR. From left to right: (1) image, (2) Grad-CAM, (3) discriminative area, (4) mask with masked area in white, (5) overlay of mask and Grad-CAM.*

The error rate is calculated by summing up the errors over all images N in the filtered test set, and then dividing by N to get the ratio. An error occurs when for attribute a and image n the value pr that the model predicted does not have the same sign as the corresponding label $l \in \{-1, 1\}$ in the ground truth.

As for the Grad-CAMs, they are being evaluated using the Acceptable Mask Ratio which is explained in the following subsection.

4.4.1 Acceptable Mask Ratio

In order to evaluate the accuracy of Grad-CAMs, an evaluation metric named Acceptable Mask Ratio (AMR) proposed by Chen (2022) was used. It measures how much of the activation in an image is located within a predefined masked area. The activated area is referred to as the discriminative region and the mask simply as the masked area (Figure 4.4). The masks that were used to calculate the AMR can be found in Figure A.4. There are two versions for each attribute. Version 1 was calculated using the shifted landmarks; therefore the location differs slightly for each attribute (Table A.1). The second version of masks consists of blocks of size 32 x 32 px and is independent of the landmarks, hence exactly the same for all images (Table A.2). The latter version takes into account that the feature maps from which the Grad-CAM is derived have a size of 7 x 7 px (Figure 3.3). To derive the final CAM they are upsampled to the image size of 224 x 224 px (Zhou et al., 2016). It can therefore be concluded that activations in the 7 x 7 Grad-CAM afterward correspond to an area of size 32 x 32 px. So activations probably tend to occur within 32 x 32 px areas and if the mask is smaller than that the AMR will likely be lower. As shown in the attachments in Figure A.7 the AMR with the second version of masks improved for roughly 80% of the attributes.

As stated in (4.4), the AMR calculates the ratio of the intersection between the discriminative area and the masked area to the discriminative area itself. Our approach is slightly different from the one by Chen (2022), as we additionally consider the masked area in regard to the image size. This adjustment makes the AMR from different attributes more comparable since the mask size varies from attribute to attribute. The effect of this adjustment is shown in the attachments in Figure A.6.

$$AMR = \frac{\#(Discriminative\ Area \cap Masked\ Area)}{\#Discriminative\ Area} \cdot \frac{\#Image\ Size - \#Masked\ Area}{\#Image\ Size} \tag{4.4}$$

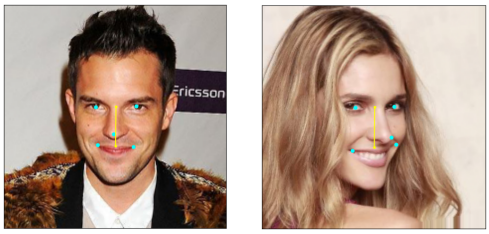


Figure 4.5: EXAMPLE FRONTAL VS. NON-FRONTAL IMAGE. *Examples for frontal image (left) and non-frontal image (right). The yellow line represents the centerline of the face. In the frontal image, the ratio of the centerline and the distance of the centerline to the tip of the nose is smaller than in the non-frontal image.*

For a given image the AMR is calculated as follows:

$$
\begin{aligned}
0 \leq AMR &= \frac{\sum_{x,y}^{I} f(g_{xy}, m_{xy})}{D} \cdot \frac{I - M}{I} \leq 1 \\
f(g_{xy}, m_{xy}) &= \begin{cases} 1, & \text{if } g_{xy} > 0 \text{ and } m_{xy} = 255 \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
\tag{4.5}
$$

D is the number of pixels of the discriminative area, M represents the number of pixels of the masked area, and I equals the number of pixels in the image (224·224). g_{xy} represents a pixel of the Grad-CAM at spatial location (x, y) and m_{xy} represents a pixel of the mask at the same spatial location. The Grad-CAM as well as the mask have dimension 224 x 224 px. Every pixel within the masked area has the value 255 (white), and every pixel outside the masked area has the value 0 (black). Thus to get the number of pixels of the intersection between mask and discriminative area it is checked for every pixel g_{xy} at spatial location (x, y) in the grayscale version of the Grad-CAM whether its value is greater than 0 and whether the pixel m_{xy} at the same spatial location in the mask is 255. If this is both the case, the pixel is part of the discriminative area and lies within the masked area.

The CelebA dataset does mostly but not only contain frontal-pose images, some half-profile and a few profile faces are also included. This posed a problem when defining the masks for calculating the AMR. The non-frontal images were problematic as the mask would not fit properly. Thus the validation and test partition were both filtered in order to only contain frontal-pose images. A frontal image was defined as follows. With the centerline[3] of the face c and the distance from the centerline to the tip of the nose s, the ratio $\frac{c}{s}$ had to be smaller than 0.1 in order for the image to be considered frontal (Figure 4.5). By using the vertical centerline of the face as a reference, we can expect the tip of the nose to be relatively centered and balanced. A threshold of 0.1 allows for minor natural variations while still ensuring a predominantly frontal view.

Through the filtering, the validation set was reduced to 10'539 which corresponds to approximately 50% of the original amount. The test set was also reduced to roughly 50% of its original size with 10'458 remaining images.

4.4.2 Statistical Methods

Research question 1 aims to determine whether there is a correlation between the two variables classification error and AMR. In order to evaluate this correlation, Numpy's `cov` function was used which returns the covariance matrix and from that Pearson's correlation coefficient $\rho \in$

[3] A straight line from the center of the eyes to the center of the mouth (yellow line in Figure 4.5).

$[-1, 1]$ can be calculated. Pearson's correlation coefficient measures the linear correlation between two variables. In case of a positive slope, the closer the data points are to a straight line, the closer ρ is to 1. In case of a negative slope, the closer the data points are to a straight line, the closer ρ is to -1 (Fahrmeir, 2016). It is formally defined as:

$$\rho_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{4.6}$$

Research question 5 aims to determine whether the Grad-CAMs from the AFFACT-b differ from the ones from AFFACT-ub. In order to quantify this difference we made use of the Kullback-Leibler (KL) information or distance. The KL distance quantifies the distance from one probability distribution to another (Burnham, 2002). More formally speaking:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x)\, log \frac{P(x)}{Q(x)} \tag{4.7}$$

where P and Q are discrete probability distributions and X is the space of the input data. The KL distance is not symmetric, meaning $KL(P||Q) \neq KL(Q||P)$. Since the AFFACT-b considers the class imbalance, we consider its Grad-CAMs as the base. We take the attribute-wise average of the balanced Grad-CAMs and the attribute-wise average of the unbalanced Grad-CAMs and calculate the distance from the balanced to the unbalanced ones leading to $KL(Grad{-}CAM_{balanced}||Grad{-}CAM_{unbalanced})$.

In order to calculate the KL distance, the `rel_entr`[4] function from the subpackage 'special' of the python library SciPy was used. It calculates the element-wise KL distance for two input arrays.

[4]More information can be found at the following link: https://docs.scipy.org/doc/scipy/reference/generated/scipy.special.rel_entr.html#scipy.special.rel_entr

Chapter 5

Experiments

In the beginning, some adjustments had to be made and tested in order to obtain optimal or at least nearly optimal Grad-CAMs. Said adjustments were done using the validation partition of the CelebA dataset, which is commonly used for fine-tuning and finding the best-performing configurations. We filtered out non-frontal images, made the output of the last residual block of the AFFACT networks accessible for the Grad-CAM method, and adapted the target function in the Grad-CAM algorithm. The following experiments were all performed on the test partition of the CelebA dataset. The filtered image set of the test partition was given to the AFFACT-b and the AFFACT-ub and the Grad-CAM method was used to generate Grad-CAMs. For every image, the following data was recorded for each attribute respectively: error (1)/no error (0), ground truth value $\in \{-1, 1\}$, prediction value, and AMR $\in [0, 1]$. The diagrams, tables, and average Grad-CAMs/AMR were all generated using the same set of results, but they were aggregated in different ways, according to the research questions, which will be described for each research question separately.

The masks used for calculating the AMR are shown in Figure A.4. For the sake of brevity, we focus on the results obtained using the second version (consisting of 32 x 32 blocks) as they achieve better AMRs and - unlike the first version - were derived in a logical and scientific manner. The results for the experiments with the first version of masks can be found in the attachments in Table A.3.

In all the diagrams (except for Figure 5.1 and Figure 5.2), figures, and tables of this chapter, the attributes are ordered by increasing class imbalance to facilitate spotting patterns which are connected to the class imbalance.

5.1 Experiment 1

Research question 1 aims to determine whether there is a correlation between the error rate of the AFFACT-b and the AFFACT-ub and the AMR. In order to answer this research question the error rate and the mean of the AMR were calculated for each attribute. Both can be found in Table 5.1.

Based on the research of Wu et al. (2023) we expected a negative correlation between the AMR and the error rate. They showed that their modified version of the MOON network (Rudd et al., 2016) with improved classification accuracy focused more on the mouth region when classifying the attribute "Mouth Slightly Open". Therefore when the error rate is low meaning the attribute is classified well, we expected the CNN to focus on the region where the attribute is located thus leading to a high AMR. And when the error rate is high meaning the attribute is often classified incorrectly, we expected the network to focus less on the location of the attribute and maybe focus more on other locations thus resulting in a low AMR.

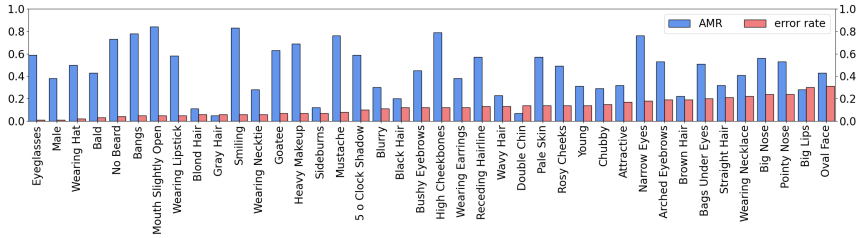


Figure 5.1: AMR VS. ERROR RATE 1. *AMR (blue) and error rate (red) for **AFFACT-b** ordered by ascending error rate.*

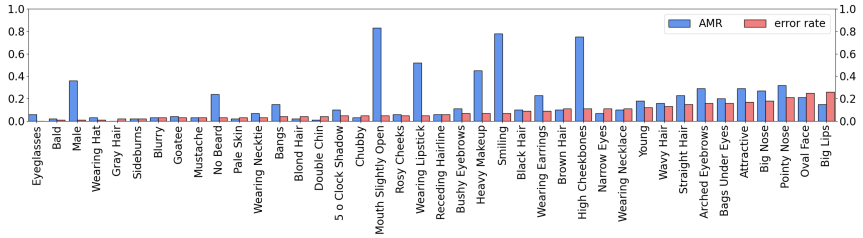


Figure 5.2: AMR VS. ERROR RATE 2. *AMR (blue) and error rate (red) for **AFFACT-ub** ordered by ascending error rate.*

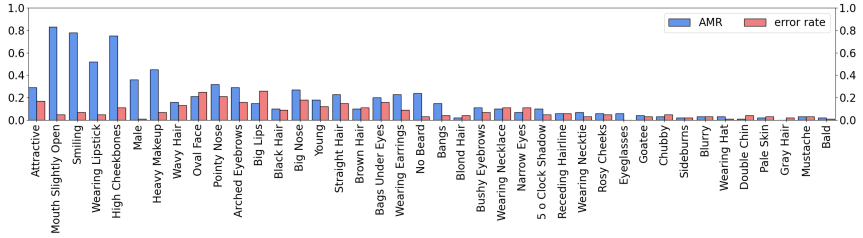


Figure 5.3: AMR VS. ERROR RATE 3. *AMR (blue) and error rate (red) for **AFFACT-ub** ordered by increasing class imbalance.*

In a first step, we plotted the results in a grouped bar chart to check whether the data fitted a straight line with a negative slope (which is the definition of the correlation coefficient). But neither for the results of AFFACT-b nor for those of AFFACT-ub a negative linear correlation became apparent. So in a next step we tried to order the bar chart by increasing error rate to maybe make a hidden correlation more apparent. When the error rate was ordered in an ascending fashion, the AMR was expected to be found in a descending fashion. But as shown in Figure 5.1 for the AFFACT-b and Figure 5.2 for the AFFACT-ub no negative correlation became apparent. For the AFFACT-ub it actually looked more like a positive correlation. Besides a few exceptions, the AMR seemed to increase together with the error rate. In Figure 5.3 a trend associated with class imbalance can be observed. With increasing class imbalance the error rate and the AMR both decrease.

Since it was difficult to make a statement solely based on the bar chart, the relationship between the two variables had to be quantified. In order to do so we calculated the covariance matrix and the correlation coefficient. The covariance matrix for the AFFACT-b cov_b (5.1) indicates that there is a negative correlation between error rate and AMR. But as the corresponding correlation coefficient ρ_b shows, this correlation is very weak. The covariance matrix for the AFFACT-ub cov_{ub} (5.2) indicates that there is in fact a positive correlation between the AMR and the error rate.

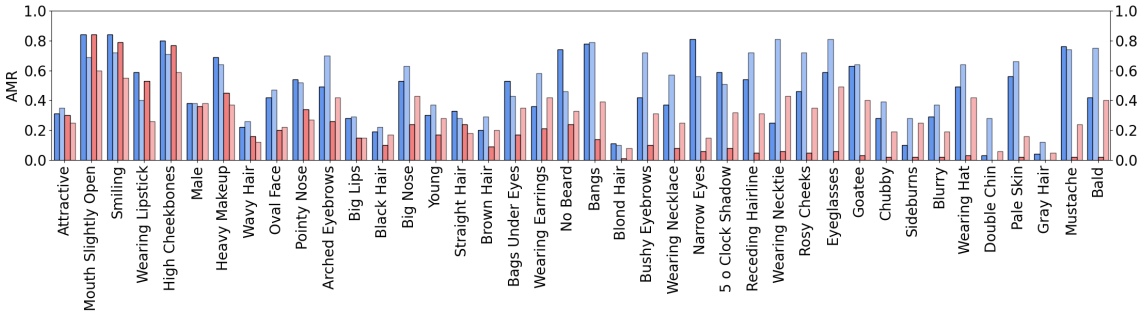


Figure 5.4: AMR PREDICTION CORRECT VS. INCORRECT. *The AMR for AFFACT-b in blue and for AFFACT-ub in red. For each pair of bars of the same color the left bar (darker) shows the AMR for correctly classified samples and the right bar (more transparent) shows the AMR for incorrectly classified samples. The attributes are ordered by increasing class imbalance.*

But according to the correlation coefficient ρ_{ub} the correlation is rather weak.

$$cov_b = \begin{bmatrix} 0.0477 & -0.00287 \\ -0.00287 & 0.0058 \end{bmatrix} \qquad \rho_b = \textbf{-0.1706} \tag{5.1}$$

$$cov_{ub} = \begin{bmatrix} 0.0451 & 0.0032 \\ 0.0032 & 0.0046 \end{bmatrix} \qquad \rho_{ub} = \textbf{0.2211} \tag{5.2}$$

5.2 Experiment 2

Research question 2 compares the Grad-CAMs for correctly and incorrectly classified samples for both the AFFACT-b and the AFFACT-ub. In order to do so the attribute-wise average Grad-CAM for correctly classified samples and for incorrectly classified samples was calculated respectively once for the AFFACT-b and once for the AFFACT-ub (Figure 5.8). Similarly, the mean of the AMR was calculated for correctly and incorrectly classified images. The numbers can be found in Table 5.1.

The results show that generally, the AMR is higher for the incorrectly classified samples with an overall AMR value of 0.51 for AFFACT-b and 0.3 for AFFACT-ub. For the AFFACT-b 26 of 40 attributes have a higher AMR for the incorrectly classified samples, for the AFFACT-ub 30 of 40 have a higher AMR for the incorrectly classified samples. There are 7 attributes for which the AMR is higher for the correctly classified samples for AFFACT-b as well as for AFFACT-ub, namely "Heavy Makeup", "High Cheekbones", "Mouth Slightly Open", "Pointy Nose", "Smiling", "Straight Hair", and "Wearing Lipstick" (Figure 5.4). They all happen to be rather balanced attributes.

5.3 Experiment 3

Research question 3 attempts to determine whether the AFFACT-b and the AFFACT-ub more often use the expected area when the attribute is present compared to when it is not. To answer this question the attribute-wise average Grad-CAM was calculated once for samples where the attribute is present (ground truth = 1) and once for samples where the attribute is absent (ground truth = -1). The Grad-CAMs for AFFACT-b and for AFFACT-ub can be found in Figure 5.9. Likewise, the mean of the AMR was calculated for both groups (Table 5.1).

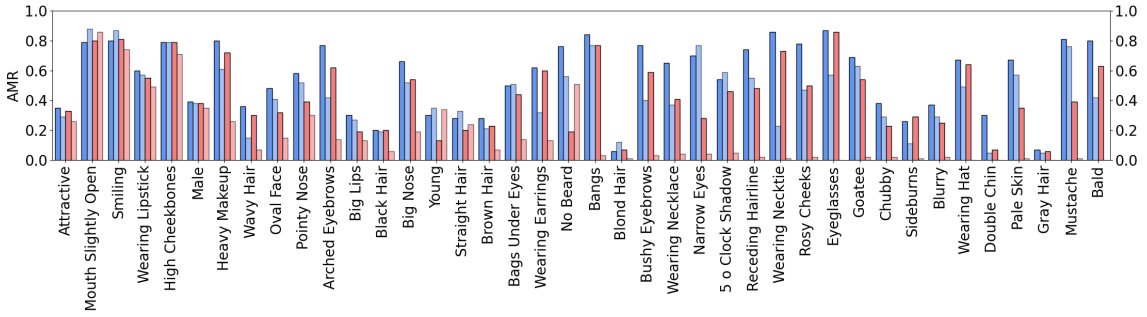


Figure 5.5: AMR GROUND TRUTH POSITIVE VS. NEGATIVE. *The AMR for AFFACT-b in blue and for AFFACT-ub in red. For each pair of bars of the same color, the left bar (darker) shows the AMR for samples where the attribute is present and the right bar (more transparent) shows the AMR for samples where the attribute is absent. The attributes are ordered by increasing class imbalance.*

The results show that the overall AMR is higher when the ground truth is positive, so when the attribute is present. For the AFFACT-b 31 of 40 attributes have a higher AMR when the attribute is present, resulting in an overall AMR value of 0.56. For the AFFACT-ub 36 of 40 attributes have a higher AMR when the attribute is present, leading to an overall AMR value for ground truth positive of 0.43. There are 3 attributes for which the AMR is higher for the ground truth negative samples for AFFACT-b as well as for AFFACT-ub, namely "Mouth Slightly Open", "Young", and "Straight Hair" (Figure 5.5). They all happen to be rather balanced attributes. "Mouth Slightly Open" is the second most balanced attribute and "Young" and "Straight Hair" follow on rank 15 and 16.

5.4 Experiment 4

Research question 4 aims to explore whether the Grad-CAMs of AFFACT-b focus more on the expected location than those from AFFACT-ub. The focus here lies on the AMR, which can be found in the columns "o" of Table 5.1.

Wu et al. (2023) showed that models with higher accuracy also produce CAMs (they used Score-CAM) where the discriminative area corresponds to the location where the attribute is located. Therefore we expected the balanced Grad-CAMs all to highlight more or less the expected area. And for the unbalanced Grad-CAMs, we expected those for rather balanced attributes to highlight the expected area and those for more unbalanced attributes to maybe highlight different areas. The results show that the more unbalanced the attribute, the lower the AMR for the unbalanced Grad-CAMs (Figure 5.6) and the less intense the heat maps. For the balanced Grad-CAMs, the AMR varies from attribute to attribute and there is no trend apparent. For the unbalanced Grad-CAMs a trend is apparent, the AMR decreases with increasing class imbalance. With an overall value of 0.46, the AMR for the balanced Grad-CAMs is higher than for the unbalanced Grad-CAMs with an overall value of 0.19.

5.5 Experiment 5

Research question 5 also compares the results of AFFACT-ub and AFFACT-b, but here the focus lies on the Grad-CAMs themselves. The average Grad-CAM for each attribute for AFFACT-b and AFFACT-ub are shown in Figure 5.10. As described in section 4.4.2 we used the Kullback-Leibler

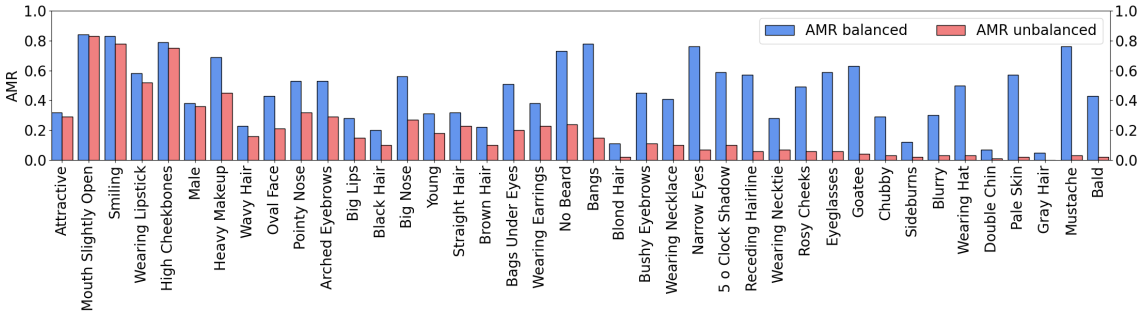


Figure 5.6: AMR AFFACT-B VS. AMR AFFACT-UB. *The overall AMR for AFFACT-b in blue and for AFFACT-ub in red. The attributes are ordered by increasing class imbalance.*

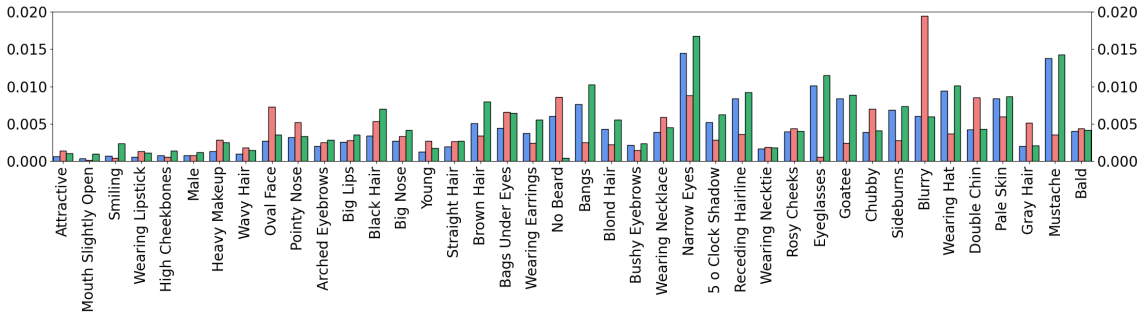


Figure 5.7: KL DISTANCE. *This diagram shows the attribute-wise Kullback-Leibler distance. It denotes the distance from the probability distribution of the balanced Grad-CAM to the probability distribution of the unbalanced Grad-CAM. Blue represents the value for the overall average Grad-CAMs, red represents the values for the average Grad-CAMs of correctly classified samples and green represents the values for the average Grad-CAMs of negative samples. The attributes are ordered by increasing class imbalance.*

distance to quantify how much the unbalanced Grad-CAMs (from AFFACT-b) differ from the balanced Grad-CAMs (from AFFACT-ub). As shown in Figure 5.7 (blue bars) the KL distance for the rather balanced attributes is fairly small so the two Grad-CAMs do not differ much. Towards the unbalanced side (on the right) the KL distance and therefore the difference increases. The attributes for which the Grad-CAMs differ the most are "Brown Hair", "No Bread", "Bangs", "Narrow Eyes", "5 o Clock Shadow", "Receding Hairline", "Eyeglasses", "Goatee", "Sideburns", "Blurry", "Pale Skin", and "Mustache". This is also apparent when looking at the Grad-CAMs (Figure 5.10). In these cases, there is a clearly visible activity in the balanced Grad-CAM and no visible activity in the unbalanced Grad-CAM.

Attribute	balanced						unbalanced					
	ER	AMR					ER	AMR				
		o	gt p	gt n	pr c	pr nc		o	gt p	gt n	pr c	pr nc
Attractive	0.17	**0.32**	**0.35**	0.29	0.31	**0.35**	0.17	0.29	**0.33**	0.26	**0.3**	0.25
Mouth Slightly Open	0.05	**0.84**	0.79	**0.88**	**0.84**	0.69	0.05	0.83	0.8	**0.86**	**0.84**	0.6
Smiling	0.06	**0.83**	0.8	**0.87**	**0.84**	0.72	0.07	0.78	**0.81**	0.74	**0.79**	0.55
Wearing Lipstick	0.05	**0.58**	**0.6**	0.57	**0.59**	0.4	0.05	0.52	**0.55**	0.49	**0.53**	0.26
High Cheekbones	0.12	**0.79**	**0.79**	**0.79**	**0.8**	0.71	0.11	0.75	**0.79**	0.71	**0.77**	0.59
Male	0.01	0.38	**0.39**	0.38	**0.38**	**0.38**	0.01	**0.36**	**0.38**	0.35	0.36	**0.38**
Heavy Makeup	0.07	**0.69**	**0.8**	0.61	**0.69**	0.64	0.07	0.45	**0.72**	0.26	**0.45**	0.37
Wavy Hair	0.13	**0.23**	**0.36**	0.15	0.22	**0.26**	0.13	0.16	**0.3**	0.07	**0.16**	0.12
Oval Face	0.31	**0.43**	**0.48**	0.41	0.42	**0.47**	0.25	0.21	**0.32**	0.15	0.2	**0.22**
Pointy Nose	0.24	**0.53**	**0.58**	0.52	**0.54**	0.52	0.21	0.32	**0.39**	0.3	**0.34**	0.27
Arched Eyebrows	0.19	**0.53**	**0.77**	0.42	0.49	**0.7**	0.16	0.29	**0.62**	0.14	0.26	**0.42**
Big Lips	0.3	**0.28**	**0.3**	0.27	0.28	**0.29**	0.26	0.15	**0.19**	0.13	**0.15**	**0.15**
Black Hair	0.12	**0.2**	**0.2**	0.19	0.19	**0.22**	0.09	0.1	**0.2**	0.06	0.1	**0.17**
Big Nose	0.24	**0.56**	**0.66**	0.52	0.53	**0.63**	0.18	0.27	**0.54**	0.19	0.24	**0.43**
Young	0.14	**0.31**	0.3	**0.35**	0.3	**0.37**	0.12	0.18	0.13	**0.34**	0.17	**0.28**
Straight Hair	0.21	**0.32**	0.28	**0.33**	**0.33**	0.28	0.15	0.23	0.2	**0.24**	**0.24**	0.18
Brown Hair	0.19	**0.22**	**0.28**	0.21	0.2	**0.29**	0.11	0.1	**0.23**	0.07	0.09	**0.2**
Bags Under Eyes	0.2	**0.51**	0.5	**0.51**	**0.53**	0.43	0.16	0.2	**0.44**	0.14	0.17	**0.35**
Wearing Earrings	0.12	**0.38**	**0.62**	0.32	0.36	**0.58**	0.09	0.23	**0.6**	0.13	0.21	**0.42**
No Beard	0.04	**0.73**	**0.76**	0.56	**0.74**	0.46	0.03	0.24	0.19	**0.51**	0.24	**0.33**
Bangs	0.05	**0.78**	**0.84**	0.77	0.78	**0.79**	0.04	0.15	**0.77**	0.03	0.14	**0.39**
Blond Hair	0.06	**0.11**	0.06	**0.12**	**0.11**	0.1	0.04	0.02	**0.07**	0.01	0.01	**0.08**
Bushy Eyebrows	0.12	**0.45**	**0.77**	0.4	0.42	**0.72**	0.07	0.11	**0.59**	0.03	0.1	**0.31**
Wearing Necklace	0.22	**0.41**	**0.65**	0.37	0.37	**0.57**	0.11	0.1	**0.41**	0.04	0.08	**0.25**
Narrow Eyes	0.18	**0.76**	0.7	**0.77**	**0.81**	0.56	0.11	0.07	**0.28**	0.04	0.06	**0.15**
5 o Clock Shadow	0.1	**0.59**	0.54	**0.59**	**0.59**	0.51	0.05	0.1	**0.46**	0.05	0.08	**0.32**
Receding Hairline	0.13	**0.57**	**0.74**	0.55	0.54	**0.72**	0.06	0.06	**0.48**	0.02	0.05	**0.31**
Wearing Necktie	0.06	**0.28**	**0.86**	0.23	0.25	**0.81**	0.03	0.07	**0.73**	0.01	0.06	**0.43**
Rosy Cheeks	0.14	**0.49**	**0.78**	0.47	0.46	**0.72**	0.05	0.06	**0.5**	0.02	0.05	**0.35**
Eyeglasses	0.01	**0.59**	**0.87**	0.57	0.59	**0.81**	0.0	0.06	**0.86**	0.0	0.06	**0.49**
Goatee	0.07	**0.63**	**0.69**	0.63	0.63	**0.64**	0.03	0.04	**0.54**	0.02	0.03	**0.4**
Chubby	0.15	**0.29**	**0.38**	0.29	0.28	**0.39**	0.05	0.03	**0.23**	0.02	0.02	**0.19**
Sideburns	0.07	**0.12**	**0.26**	0.11	0.1	**0.28**	0.02	0.02	**0.29**	0.01	0.02	**0.25**
Blurry	0.11	**0.3**	**0.37**	0.29	0.29	**0.37**	0.03	0.03	**0.25**	0.02	0.02	**0.19**
Wearing Hat	0.02	**0.5**	**0.67**	0.49	0.49	**0.64**	0.01	0.03	**0.64**	0.0	0.03	**0.42**
Double Chin	0.14	**0.07**	**0.3**	0.05	0.03	**0.28**	0.04	0.01	**0.07**	0.0	0.0	**0.06**
Pale Skin	0.14	**0.57**	**0.67**	0.57	0.56	**0.66**	0.03	0.02	**0.35**	0.01	0.02	**0.16**
Gray Hair	0.06	**0.05**	**0.07**	0.05	0.04	**0.12**	0.02	0.0	**0.06**	0.0	0.0	**0.05**
Mustache	0.08	**0.76**	**0.81**	0.76	**0.76**	0.74	0.03	0.03	**0.39**	0.01	0.02	**0.24**
Bald	0.03	**0.43**	**0.8**	0.42	0.42	**0.75**	0.01	0.02	**0.63**	0.0	0.02	**0.4**
OVERALL	0.12	**0.46**	**0.56**	0.44	0.45	**0.51**	0.08	0.19	**0.43**	0.16	0.19	**0.3**

Table 5.1: AMR AND ERROR RATE MASK VERSION 2. *This table shows the error rate and the AMR for the AFFACT-b and the AFFACT-ub (with masks version 2). The attributes are ordered by increasing class imbalance. The greater value of the column pairs o/o, gt p/gt n, and pr c/pr nc is highlighted with bold font and exceptions from the norm are in color. o = overall, gt p = ground truth positive, gt n = ground truth negative, pr c = prediction correct, pr nc = prediction not correct*

Figure 5.8: GRADCAMS PREDICTION CORRECT VS. INCORRECT. *Each group of four shows the attribute-wise average GradCAM for correctly classified samples (left) and incorrectly classified samples (right). The upper pair is from AFFACT-b, the lower pair from AFFACT-ub. The attributes are ordered by increasing class imbalance.*

Figure 5.9: GradCAMs attribute present vs. absent. *Each group of four shows the attribute-wise average GradCAM for presence (left) and absence of the attribute (right). The upper pair is from AFFACT-b, the lower pair from AFFACT-ub. For the pair from AFFACT-ub, the majority class is indicated with a red frame. The attributes are ordered by increasing class imbalance.*

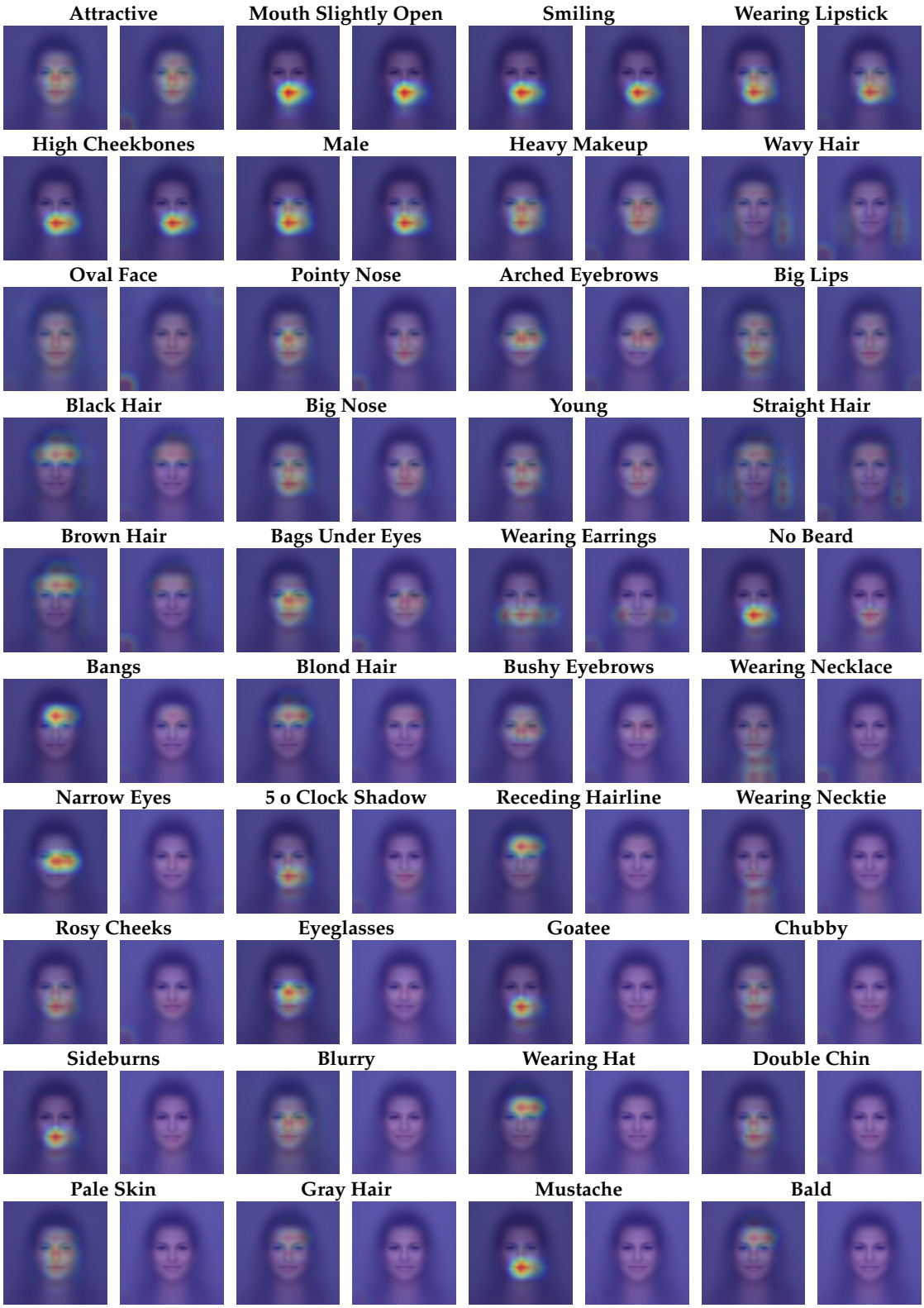


Figure 5.10: CAMs BALANCED VS. UNBALANCED. *This figure shows the attribute-wise average Grad-CAMs for AFFACT-b (left) and AFFACT-ub (right). The attributes are ordered by increasing class imbalance.*

Chapter 6

Discussion

The main finding of the experiments that were conducted as part of this thesis is that for the AFFACT-ub the Grad-CAMs for the majority class show less activity the more the class imbalance increases while for the minority class, it is rather stable. As already mentioned the AFFACT-ub performs well on the majority class but rather poorly on the minority class. Considering this together with the findings of Wu et al. (2023) which showed that the Class Activation Map improves[1] when the classification accuracy improves, we expected the Grad-CAMs of the majority class to highlight the expected region while for the minority class, we expected the discriminative area to may be somewhere else.
This is primarily important for experiment 3, but indirectly it also affects the other experiments. In the following, we will discuss the results of each experiment and try to establish the link between the respective results and this main finding.

In the *first experiment* the results are not very clear, the negative correlation between the AMR and the error rate for the AFFACT-b is not significant and for the AFFACT-ub the correlation coefficient even indicated a positive relationship. A positive correlation means, that the error rate and the AMR increase/decrease simultaneously and it can be observed in Figure 5.3 that they both decrease with increasing class imbalance. With increasing imbalance, the overall error rate decreases because the majority class which has a low error rate becomes bigger and therefore has a bigger impact. For example for the attribute "Bald" the dominant class makes up about 98% of the samples so the network performs well on almost all of the samples and achieves a very low error rate. The AMR also decreases with increasing class imbalance because of the phenomenon that we observed for the Grad-CAMs of the AFFACT-ub namely that the Grad-CAMs for the majority class do hardly show any activity. E.g. for the attribute "Bald" 9'931 of 10'548 Grad-CAMs show no activity at all which results in an AMR of 0 and of the remaining 527 Grad-CAMs many have an activity in the corner of the image which also results in an AMR of 0. This explains its extremely low AMR of 0.02.

The results for the *second experiment* show that the AMR is generally higher for the incorrectly classified samples. While for the results of the AFFACT-b, it is hard to spot a pattern that would somehow explain the results for the AFFACT-ub it seems to be a similar pattern that we also observed in other experiments. That is with increasing class imbalance the heatmaps of the average Grad-CAMs for correctly classified samples tend to become weaker and weaker and with them decreases the AMR. The decreasing AMR is nicely shown in Figure 5.4 as the darker red bars. Now how could this be explained. As already discussed, the Grad-CAMs for the dominating class of highly imbalanced attributes seem to often show no activity at all leading to an AMR of 0 which then results in a low average AMR. And we know that the network performs well on the dominating class. Therefore the average Grad-CAM of correctly classified samples mainly

[1] e.g. is located at the mouth region when classifying "Mouth Slightly Open" rather than at the eye region

contains samples of the dominating class which leads to a mostly blue average Grad-CAM and a low average AMR. The increasing difference between the correctly and incorrectly classified average Grad-CAM of the AFFACT-ub can be observed in Figure 5.7 (red bars). Even though it was initially not planned to calculate the KL distance for this experiment, it helps to emphasize the observed pattern.

In the *third experiment*, the results indicate that the average AMR for positive samples (attribute present) is higher than for negative samples (attribute absent). In case of the AFFACT-ub, this contradicts what we expected because we know that the network predicts the majority class well and with increasing accuracy the Class Activation Map should be more likely to actually highlight the expected area (Wu et al., 2023). Interestingly the opposite can be observed in the results in Figure 5.9. In order to quantify the difference between the positive and negative average Grad-CAM of AFFACT-ub we calculated the KL distance (Figure 5.7, green bars). Even though it was initially not planned to calculate the KL distance for this experiment, it helps to visualize the observed pattern. It can be observed in Figure 5.7 that for the rather balanced attributes, the difference between the positive and negative AFFACT-ub-Grad-CAMs is fairly small. Towards the unbalanced side, the difference increases. For most of the highly imbalanced attributes, the average Grad-CAM for the positive samples looks close to what we expected while the average Grad-CAM for the negative samples is almost completely blue (Figure 5.9. At first, this looks like the Grad-CAM for the negative class is always the better one. But there are three exceptions namely "Attractive", "No Beard", and "Young". In exactly these three cases the positive class is the dominating class while for all other attributes, it is the negative class. It seems like for classifying the majority class the network does not use much of the image to classify. The question is why this happens. A first theory could be that the network does not do much when the feature maps show patterns that are characteristic of the majority class because this is the default. But the network looks closer when the feature maps show patterns that are different from the default. We leave it for future research to further examine this phenomenon.

The results of the *fourth experiment* show that the overall AMR is higher for AFFACT-b than for AFFACT-ub. This is because especially for the highly unbalanced attributes the average AMR for AFFACT-ub is fairly small. This again can be explained by the phenomenon we observed in experiment 3. As already discussed, the average Grad-CAM for the dominating class shows little to no activity the more unbalanced the attribute. So when the AMR of the dominating class - which mainly contributes to the average - decreases then consequently the average AMR decreases as well.

The *fifth experiment's* results confirm what we already saw in the other experiments, namely that the heatmaps of the average Grad-CAMs generated by AFFACT-b do not vary much throughout the attributes (except for the location of course) while those of AFFACT-ub get weaker with increasing class imbalance until they as of the attribute "Eyeglasses" visually disappear.

Besides the results of the single experiments, for some attributes, the discriminative area does not correspond exactly to the location of the attribute. There are a few attributes for which the focus lies mainly on the mouth rather than the respective attribute location. E.g."High Cheekbones", "Rosy Cheeks", "Sideburns" and "Big Nose". Also for "Male" the focus is on the mouth instead of - as one may expect - the whole face. Despite "Wavy Hair" and "Straight Hair" for all the other hair-associated attributes the discriminative area is located at the forehead, apparently ignoring hair on the side of the head. Other than that for most attributes the AFFACT-b as well as the AFFACT-ub indeed use the part of the image where the attribute to be classified is located.

Chapter 7

Conclusion

In this thesis, the goal was to gain a deeper understanding of the decision-making process of two binary classifiers when classifying 40 binary facial attributes. We used two pre-trained CNNs which are both based on a ResNet-50 structure and apply the AFFACT technique to classify attributes (Günther et al., 2017). One of them, referred to as the AFFACT-b, considers class imbalance inherent to the dataset while the other one, the AFFACT-ub, does not. We then conducted several experiments on the CelebA dataset, used the Grad-CAM method to identify the discriminative regions by generating a heatmap for every image, and created masks built of 32 x 32 pixel blocks to evaluate the heatmaps. With the Acceptable Mask Ratio (AMR), we measured how much of the discriminative area lies within the masked area. Through these experiments, we sought to answer five research questions for the AFFACT-b and the AFFACT-ub. *First*, we examined whether there is a correlation between the classifier's error rate and the location of the discriminative area. The results indicated an insignificantly weak negative correlation for the AFFACT-b and a fairly weak positive correlation for the AFFACT-ub. *Second*, the location of the discriminative area between correctly classified and incorrectly classified samples was inspected and we observed that for both the AFFACT-b and the AFFACT-ub the incorrectly classified samples are more likely to rely on the area of the image where the attribute is located. *Third*, we compared the Grad-CAMs for samples containing the attribute to samples where the attribute is absent and discovered that when the attribute is present the classifier is more likely to rely on the expected area when the attribute is present. *Fourth*, we contrasted the Grad-CAMs of the AFFACT-b and the AFFACT-ub and found that the discriminative area of the balanced Grad-CAMs more often lies within the masked area. *Fifth*, we measured the difference between the Grad-CAMs from AFFACT-b and those from AFFACT-ub. The results showed that the more unbalanced the attribute, the more different the Grad-CAMs of the two CNNs. Throughout all five experiments, we observed the phenomenon that for highly biased attributes the Grad-CAMs for the dominant class do not show any activity. It remains to be explained why this happens.

While the experiments conducted in this thesis have provided valuable insights, there remain potential areas for improvement that could enhance the robustness and comprehensiveness of the results. For example, the threshold for defining the discriminative area could be raised. For our experiments, any value greater than zero was counted as part of the discriminative area. Also, one could consider the shape of an attribute and design the masks respectively instead of creating them all in a rectangular shape. Moreover, our findings leave us with some questions that have yet to be explored. For instance, it is to be examined whether the Grad-CAM method is suited for binary classifiers. Because we did not manage to completely solve the problem of activations in the corners of images. Some adjustments in the code of the Grad-CAM method might be able to solve this issue. In addition, it would be of utmost interest to further examine the observation that the AFFACT-ub's average Grad-CAMs for the majority class do not highlight any area when the attribute is highly biased.

Appendix A

Attachments

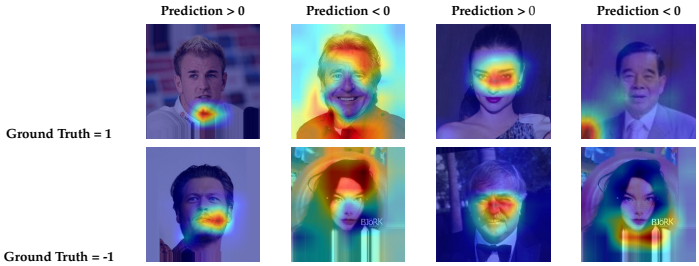


Figure A.1: TARGET FUNCTION 1. *CAMs when the target function returns the raw model output (as in 'ClassifierOutputTarget') for "5 o Clock Shadow" (left square) and "Bushy Eyebrows" (right square).*

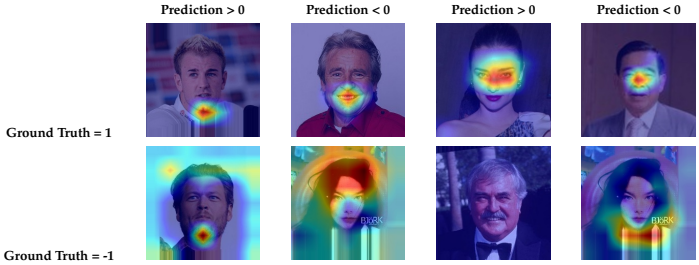


Figure A.2: TARGET FUNCTION 3. *CAMs when the target function returns the absolute model output multiplied by the ground truth value for "5 o Clock Shadow" (left square) and "Bushy Eyebrows" (right square).*

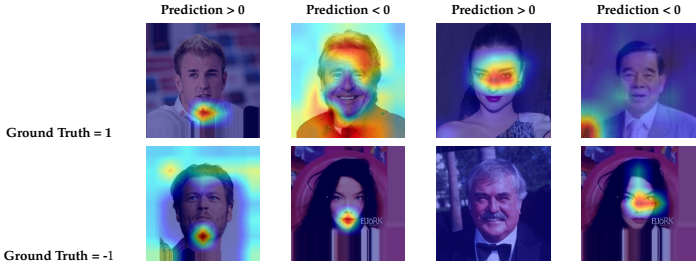


Figure A.3: TARGET FUNCTION 4. *CAMs when the target function returns the model output multiplied by the ground truth value for "5 o Clock Shadow" (left square) and "Bushy Eyebrows" (right square).*

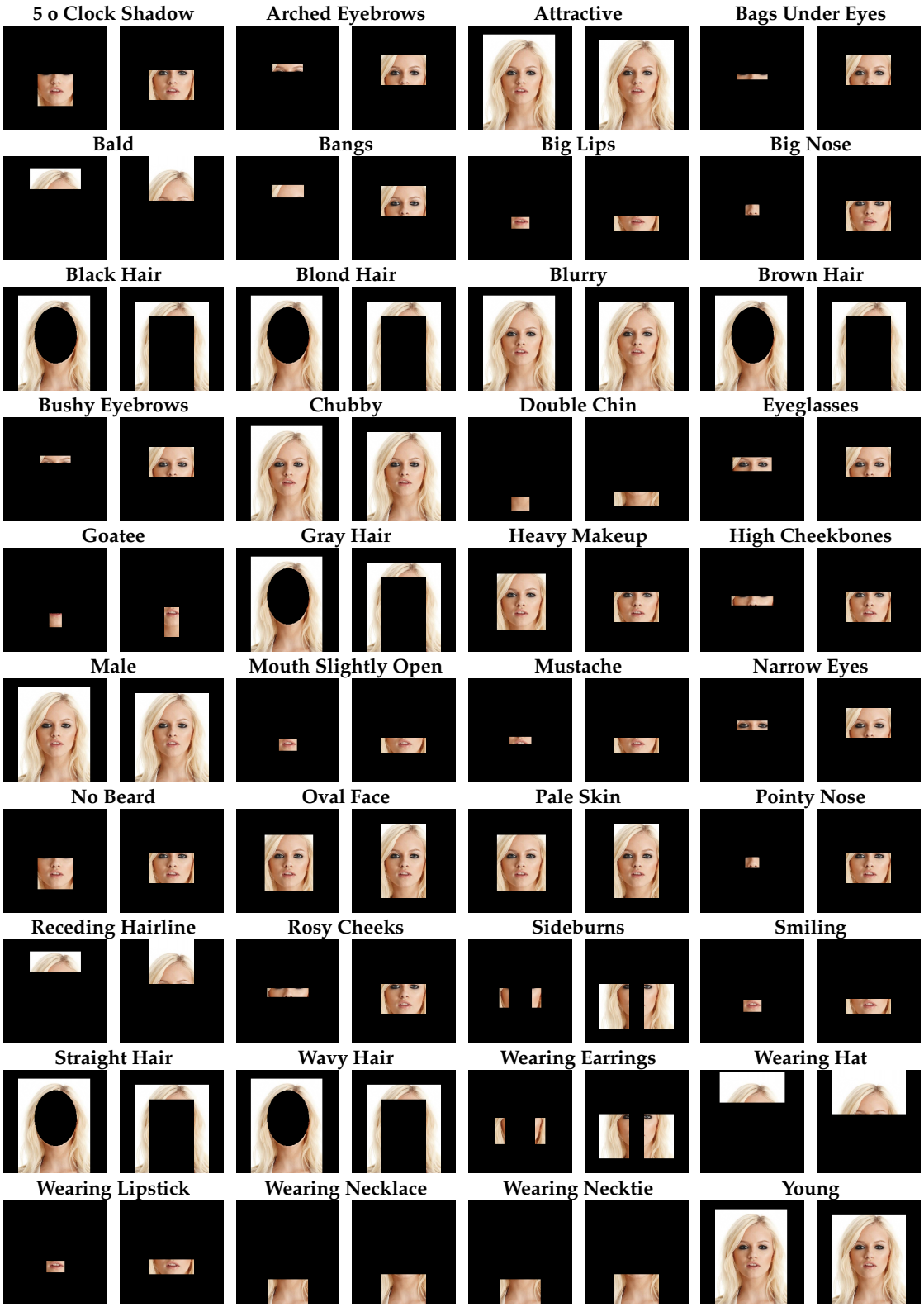


Figure A.4: MASKS. *These are the masks used to calculate the AMR, version 1 (left) and version 2 (right).*

Attributes	x_l	x_r	y_t	y_b
5 o Clock Shadow	$x_{m,l} - 25$	$x_{m,r} + 25$	$y_{e,c} + 4$	$y_{m,c} + 32$
Arched Eyebrows	$x_{e,l} - 13$	$x_{e,r} + 13$	$y_{e,c} - 18$	$y_{e,c} - 2$
Attractive	$x_n - 80$	$x_n + 75$	$y_n - 100$	224
Bags Under Eyes	$x_{e,l} - 13$	$x_{e,r} + 13$	$y_{e,c} + 4$	$y_{e,c} + 15$
Bald	$x_{head,c} - 55$	$x_{head,c} + 55$	$y_{hairline} - 30$	$y_{hairline} + 15$
Bangs	$x_{e,l} - 15$	$x_{e,r} + 15$	$y_{f,c} - 12$	$y_{f,c} + 15$
Big Lips	$x_{m,l} - 6$	$x_{m,r} + 6$	$y_{m,c} - 10$	$y_{m,c} + 15$
Big Nose	$x_n - 15$	$x_n + 15$	$y_n - 15$	$y_n + 8$
Black Hair	$mask_{global} - ellipse(center = (x_n, y_n - 15), r_{long} = 60, r_{short} = 45)$			
Blond Hair	$mask_{global} - ellipse(center = (x_n, y_n - 15), r_{long} = 60, r_{short} = 45)$			
Blurry	$x_n - 80$	$x_n + 75$	$y_n - 100$	224
Brown Hair	$mask_{global} - ellipse(center = (x_n, y_n - 15), r_{long} = 60, r_{short} = 45)$			
Bushy Eyebrows	$x_{e,l} - 13$	$x_{e,r} + 13$	$y_{e,c} - 18$	$y_{e,c} - 2$
Chubby	$x_n - 80$	$x_n + 75$	$y_n - 100$	224
Double Chin	$x_{m,l} - 6$	$x_{m,r} + 6$	$y_{m,c} + 30$	$y_{m,c} + 60$
Eyeglasses	$x_{e,l} - 22$	$x_{e,r} + 22$	$y_{e,c} - 15$	$y_{e,c} + 15$
Goatee	$x_{m,l}$	$x_{m,r}$	$y_{m,c}$	$y_{m,c} + 30$
Gray Hair	$mask_{global} - ellipse(center = (x_n, y_n - 15), r_{long} = 60, r_{short} = 45)$			
Heavy Makeup	$x_n - 50$	$x_n + 55$	$y_{e,c} - 45$	$y_{m,c} + 35$
High Cheekbones (left)	$x_{e,l} - 25$	x_n	$y_{e,c} + 4$	$y_n + 5$
High Cheekbones (right)	x_n	$x_{e,r} + 25$	$y_{e,c} + 4$	$y_n + 5$
Male	$x_n - 80$	$x_n + 75$	$y_n - 100$	224
Mouth Slightly Open	$x_{m,l} - 6$	$x_{m,r} + 6$	$y_{m,c} - 10$	$y_{m,c} + 15$
Mustache	$x_{m,l} - 10$	$x_{m,r} + 10$	$y_{m,c} - 15$	$y_{m,c}$
Narrow Eyes	$x_{e,l} - 13$	$x_{e,r} - 13$	$y_{e,c} - 10$	$y_{e,c} + 10$
No Beard	$x_{m,l} - 25$	$x_{m,r} + 25$	$y_{e,c} + 4$	$y_{m,c} + 32$
Oval Face	$x_n - 50$	$x_n + 55$	$y_{e,c} - 45$	$y_{m,c} + 35$
Pale Skin	$x_n - 50$	$x_n + 55$	$y_{e,c} - 45$	$y_{m,c} + 35$
Pointy Nose	$x_n - 15$	$x_n + 15$	$y_n - 15$	$y_n + 8$
Receding Hairline	$x_{head,c} - 55$	$x_{head,c} + 55$	$y_{hairline} - 30$	$y_{hairline} + 15$
Rosy Cheeks (left)	$x_{e,l} - 25$	x_n	$y_{e,c} + 4$	$y_n + 5$
Rosy Cheeks (right)	x_n	$x_{e,r} + 25$	$y_{e,c} + 4$	$y_n + 5$
Sideburns (left)	$x_{e,l} - 25$	$x_{e,l} - 5$	$y_{e,c} + 4$	$y_n + 28$
Sideburns (right)	$x_{e,r} + 5$	$x_{e,r} + 25$	$y_{e,c} + 4$	$y_n + 28$
Smiling	$x_{m,l} - 6$	$x_{m,r} + 6$	$y_{m,c} - 10$	$y_{m,c} + 15$
Straight Hair	$mask_{global} - ellipse(center = (x_n, y_n - 15), r_{long} = 60, r_{short} = 45)$			
Wavy Hair	$mask_{global} - ellipse(center = (x_n, y_n - 15), r_{long} = 60, r_{short} = 45)$			
Wearing Earrings (left)	$x_{e,l} - 34$	$x_{e,l} - 12$	$y_n - 15$	$y_n + 42$
Wearing Earrings (right)	$x_{e,r} + 12$	$x_{e,r} + 34$	$y_n - 15$	$y_n + 42$
Wearing Hat	$x_{head,c} - 70$	$x_{head,c} + 70$	5	$y_{hairline} + 15$
Wearing Lipstick	$x_{m,l} - 6$	$x_{m,r} + 6$	$y_{m,c} - 10$	$y_{m,c} + 15$
Wearing Necklace	$x_{m,l} - 30$	$x_{m,r} + 30$	$y_{m,c} + 30$	224
Wearing Necktie	$x_{m,l} - 30$	$x_{m,r} + 30$	$y_{m,c} + 30$	224
Young	$x_n - 80$	$x_n + 75$	$y_n - 100$	224
Additional Variables	$y_{e,c} = \frac{y_{e,l}+y_{e,r}}{2}$ $x_{head,c} = \frac{x_{e,l}+x_{e,r}}{2}$ $y_{hairline} = \frac{y_{e,l}+y_{e,r}}{2}$ $y_{f,c} = y_{e,c} - 27$ $y_{m,c} = \frac{y_{m,l}+y_{m,r}}{2}$ $mask_{global} : x_l = x_n - 80, x_r = x_n + 75, y_t = y_n - 100, y_b = 224$			

Table A.1: MASKS VERSION 1 COORDINATES. *This table shows the corner coordinates of the masks (version 1). They are computed using the shifted landmarks. Every variable that is not mentioned in the last row is a landmark. e = eyes, n = nose, m = mouth, l = left, r = right, c = center*

Attributes	x_l	x_r	y_t	y_b
5 o Clock Shadow	$2 \cdot 32$	$5 \cdot 32$	$3 \cdot 32$	$5 \cdot 32$
Arched Eyebrows	$2 \cdot 32$	$5 \cdot 32$	$2 \cdot 32$	$4 \cdot 32$
Attractive	$1 \cdot 32$	$6 \cdot 32$	$1 \cdot 32$	$7 \cdot 32$
Bags Under Eyes	$2 \cdot 32$	$5 \cdot 32$	$2 \cdot 32$	$4 \cdot 32$
Bald	$2 \cdot 32$	$5 \cdot 32$	$0 \cdot 32$	$3 \cdot 32$
Bangs	$2 \cdot 32$	$5 \cdot 32$	$2 \cdot 32$	$4 \cdot 32$
Big Lips	$2 \cdot 32$	$5 \cdot 32$	$4 \cdot 32$	$5 \cdot 32$
Big Nose	$2 \cdot 32$	$5 \cdot 32$	$3 \cdot 32$	$5 \cdot 32$
Black Hair	$mask_{global} - mask_{face}$			
Blond Hair	$mask_{global} - mask_{face}$			
Blurry	$1 \cdot 32$	$6 \cdot 32$	$1 \cdot 32$	$7 \cdot 32$
Brown Hair	$mask_{global} - mask_{face}$			
Bushy Eyebrows	$2 \cdot 32$	$5 \cdot 32$	$2 \cdot 32$	$4 \cdot 32$
Chubby	$1 \cdot 32$	$6 \cdot 32$	$1 \cdot 32$	$7 \cdot 32$
Double Chin	$2 \cdot 32$	$5 \cdot 32$	$5 \cdot 32$	$6 \cdot 32$
Eyeglasses	$2 \cdot 32$	$5 \cdot 32$	$2 \cdot 32$	$4 \cdot 32$
Goatee	$3 \cdot 32$	$4 \cdot 32$	$4 \cdot 32$	$6 \cdot 32$
Gray Hair	$mask_{global} - mask_{face}$			
Heavy Makeup	$2 \cdot 32$	$5 \cdot 32$	$3 \cdot 32$	$5 \cdot 32$
High Cheekbones	$2 \cdot 32$	$5 \cdot 32$	$3 \cdot 32$	$5 \cdot 32$
Male	$1 \cdot 32$	$6 \cdot 32$	$1 \cdot 32$	$7 \cdot 32$
Mouth Slightly Open	$2 \cdot 32$	$5 \cdot 32$	$4 \cdot 32$	$5 \cdot 32$
Mustache	$2 \cdot 32$	$5 \cdot 32$	$4 \cdot 32$	$5 \cdot 32$
Narrow Eyes	$2 \cdot 32$	$5 \cdot 32$	$2 \cdot 32$	$4 \cdot 32$
No Beard	$2 \cdot 32$	$5 \cdot 32$	$3 \cdot 32$	$5 \cdot 32$
Oval Face	$2 \cdot 32$	$5 \cdot 32$	$1 \cdot 32$	$6 \cdot 32$
Pale Skin	$2 \cdot 32$	$5 \cdot 32$	$1 \cdot 32$	$6 \cdot 32$
Pointy Nose	$2 \cdot 32$	$5 \cdot 32$	$3 \cdot 32$	$5 \cdot 32$
Receding Hairline	$2 \cdot 32$	$5 \cdot 32$	$0 \cdot 32$	$3 \cdot 32$
Rosy Cheeks	$2 \cdot 32$	$5 \cdot 32$	$3 \cdot 32$	$5 \cdot 32$
Sideburns (left)	$1 \cdot 32$	$3 \cdot 32$	$3 \cdot 32$	$6 \cdot 32$
Sideburns (right)	$4 \cdot 32$	$6 \cdot 32$	$3 \cdot 32$	$6 \cdot 32$
Smiling	$2 \cdot 32$	$5 \cdot 32$	$4 \cdot 32$	$5 \cdot 32$
Straight Hair	$mask_{global} - mask_{face}$			
Wavy Hair	$mask_{global} - mask_{face}$			
Wearing Earrings (left)	$1 \cdot 32$	$3 \cdot 32$	$3 \cdot 32$	$6 \cdot 32$
Wearing Earrings (right)	$4 \cdot 32$	$6 \cdot 32$	$3 \cdot 32$	$6 \cdot 32$
Wearing Hat	$1 \cdot 32$	$6 \cdot 32$	$0 \cdot 32$	$3 \cdot 32$
Wearing Lipstick	$2 \cdot 32$	$5 \cdot 32$	$4 \cdot 32$	$5 \cdot 32$
Wearing Necklace	$2 \cdot 32$	$5 \cdot 32$	$5 \cdot 32$	$7 \cdot 32$
Wearing Necktie	$2 \cdot 32$	$5 \cdot 32$	$5 \cdot 32$	$7 \cdot 32$
Young	$1 \cdot 32$	$6 \cdot 32$	$1 \cdot 32$	$7 \cdot 32$
Additional Variables	$mask_{face} = x_l = 2 \cdot 32, x_r = 5 \cdot 32, y_t = 2 \cdot 32, y_b = 7 \cdot 32$			
	$mask_{global} = x_l = 1 \cdot 32, x_r = 6 \cdot 32, y_t = 1 \cdot 32, y_b = 7 \cdot 32$			

Table A.2: MASKS VERSION 2 COORDINATES. *This table shows the corner coordinates of the masks (version 2). They are built of multiples of 32 because initially, the Grad-CAMs have a size of 7 x 7 px, which after upsampling them to image size corresponds to an area of 32 x 32. e = eyes, n = nose, m = mouth, l = left, r = right, c = center*

Figure A.5: GRAD-CAM ERROR EXAMPLES. *Grad-CAMs for the attribute "Bald" from AFFACT-b (left of pair) and AFFACT-ub (right of pair) generated with target function 2. For the AFFACT-ub the Grad-CAM method still fails sometimes (activation in image corner).*

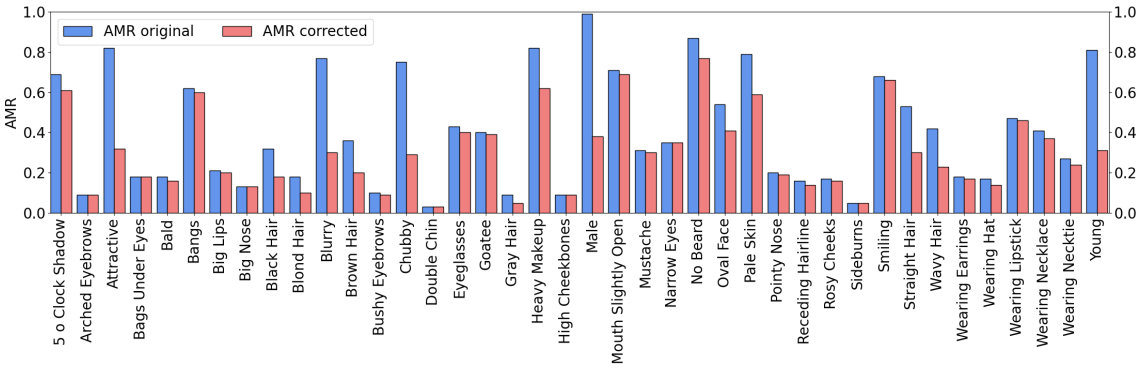


Figure A.6: AMR CORRECTED (AFFACT-B). *This diagram compares the attribute-wise average AMR when using the formula from Chen (2022) (blue) to the attribute-wise average AMR obtained using our adjusted formula (red) which considers the various sizes of the masks w.r.t. the image size. Both AMRs were calculated from CAMs generated using the Grad-CAM method and the AFFACT-b. It shows that when using the formula from Chen (2022) attributes like "Attractive", "Blurry", "Chubby" whose masked areas contain almost the whole image have a very high AMR compared to other attributes. This makes it hard to compare them to other attributes whose masks are considerably smaller. As shown in this diagram our formula relativizes the AMR of attributes with a big mask, i.e. "Attractive", "Blurry", "Chubby", "Male", and "Young".*

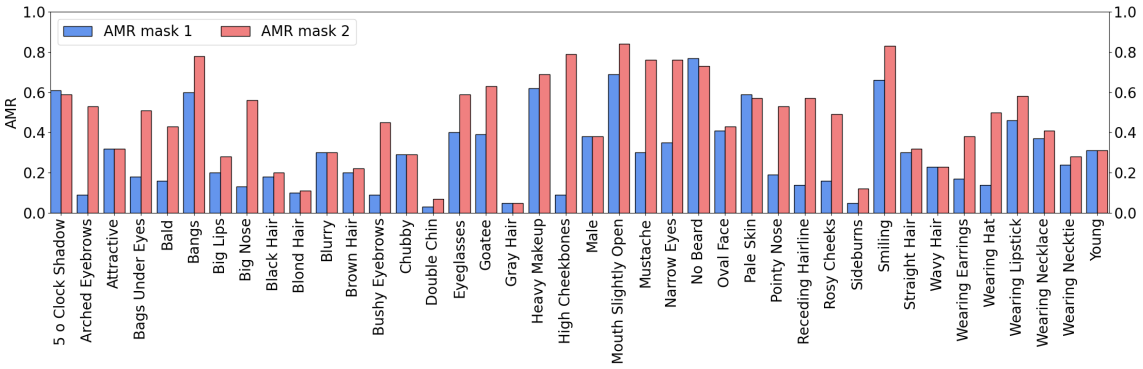


Figure A.7: AMR MASK 1 VS. MASK 2. *This diagram compares the attribute-wise average AMR when using the first version of masks (blue) to the attribute-wise average AMR when using the second version of masks consisting of multiples of 32 (red). Both are obtained from Grad-CAMs of AFFACT-b. The AMR with the second version of masks improves for roughly 80% of the attributes.*

Attribute	balanced						unbalanced					
	ER	AMR					ER	AMR				
		o	gt p	gt n	pr c	pr nc		o	gt p	gt n	pr c	pr nc
5 o Clock Shadow	0.1	0.61	0.71	0.6	0.61	0.65	0.05	0.12	0.61	0.06	0.1	0.42
Arched Eyebrows	0.19	0.09	0.14	0.06	0.08	0.13	0.16	0.06	0.15	0.02	0.06	0.1
Attractive	0.17	0.32	0.35	0.29	0.31	0.35	0.17	0.29	0.33	0.25	0.3	0.25
Bags Under Eyes	0.2	0.18	0.18	0.18	0.19	0.14	0.16	0.07	0.15	0.05	0.06	0.12
Bald	0.03	0.16	0.19	0.16	0.16	0.17	0.01	0.01	0.23	0.0	0.01	0.14
Bangs	0.05	0.6	0.7	0.58	0.6	0.65	0.04	0.11	0.53	0.02	0.1	0.3
Big Lips	0.3	0.2	0.22	0.19	0.2	0.2	0.26	0.11	0.13	0.1	0.11	0.1
Big Nose	0.24	0.13	0.18	0.11	0.12	0.16	0.18	0.07	0.16	0.04	0.06	0.11
Black Hair	0.12	0.18	0.19	0.18	0.18	0.21	0.09	0.08	0.15	0.05	0.08	0.14
Blond Hair	0.06	0.1	0.06	0.11	0.1	0.09	0.04	0.02	0.08	0.01	0.02	0.08
Blurry	0.11	0.3	0.36	0.29	0.29	0.36	0.03	0.03	0.24	0.02	0.02	0.19
Brown Hair	0.19	0.2	0.27	0.18	0.19	0.26	0.11	0.09	0.21	0.07	0.08	0.17
Bushy Eyebrows	0.12	0.09	0.18	0.08	0.09	0.17	0.07	0.03	0.15	0.01	0.02	0.08
Chubby	0.15	0.29	0.37	0.28	0.27	0.38	0.05	0.03	0.23	0.02	0.02	0.19
Double Chin	0.14	0.03	0.14	0.03	0.02	0.12	0.04	0.0	0.02	0.0	0.0	0.02
Eyeglasses	0.01	0.4	0.63	0.39	0.4	0.57	0.0	0.04	0.63	0.0	0.04	0.36
Goatee	0.07	0.39	0.41	0.39	0.39	0.38	0.03	0.02	0.3	0.01	0.02	0.22
Gray Hair	0.06	0.05	0.08	0.05	0.05	0.11	0.02	0.0	0.07	0.0	0.0	0.06
Heavy Makeup	0.07	0.62	0.73	0.54	0.61	0.68	0.07	0.42	0.68	0.23	0.42	0.4
High Cheekbones	0.12	0.09	0.11	0.07	0.08	0.12	0.11	0.07	0.1	0.05	0.07	0.09
Male	0.01	0.38	0.38	0.38	0.38	0.37	0.01	0.36	0.38	0.34	0.36	0.37
Mouth Slightly Open	0.05	0.69	0.64	0.74	0.7	0.52	0.05	0.72	0.65	0.77	0.73	0.49
Mustache	0.08	0.3	0.31	0.3	0.3	0.28	0.03	0.01	0.15	0.0	0.01	0.09
Narrow Eyes	0.18	0.35	0.3	0.35	0.37	0.22	0.11	0.03	0.13	0.02	0.03	0.07
No Beard	0.04	0.77	0.78	0.75	0.78	0.65	0.03	0.28	0.2	0.69	0.27	0.46
Oval Face	0.31	0.41	0.46	0.38	0.39	0.45	0.25	0.2	0.33	0.14	0.2	0.21
Pale Skin	0.14	0.59	0.7	0.59	0.57	0.7	0.03	0.02	0.38	0.01	0.02	0.17
Pointy Nose	0.24	0.19	0.17	0.2	0.21	0.15	0.21	0.08	0.1	0.06	0.08	0.07
Receding Hairline	0.13	0.14	0.18	0.14	0.14	0.17	0.06	0.02	0.17	0.01	0.02	0.11
Rosy Cheeks	0.14	0.16	0.28	0.15	0.15	0.26	0.05	0.02	0.16	0.01	0.01	0.12
Sideburns	0.07	0.05	0.07	0.05	0.05	0.08	0.02	0.01	0.1	0.0	0.01	0.08
Smiling	0.06	0.66	0.62	0.71	0.67	0.52	0.07	0.64	0.63	0.65	0.66	0.41
Straight Hair	0.21	0.3	0.27	0.31	0.31	0.27	0.15	0.23	0.2	0.24	0.24	0.17
Wavy Hair	0.13	0.23	0.36	0.16	0.23	0.27	0.13	0.16	0.3	0.07	0.16	0.13
Wearing Earrings	0.12	0.17	0.28	0.14	0.16	0.25	0.09	0.11	0.29	0.06	0.1	0.2
Wearing Hat	0.02	0.14	0.19	0.14	0.14	0.36	0.01	0.01	0.27	0.0	0.01	0.26
Wearing Lipstick	0.05	0.46	0.43	0.48	0.47	0.27	0.05	0.41	0.38	0.43	0.42	0.18
Wearing Necklace	0.22	0.37	0.62	0.32	0.32	0.52	0.11	0.09	0.39	0.04	0.08	0.24
Wearing Necktie	0.06	0.24	0.73	0.2	0.21	0.69	0.03	0.06	0.67	0.01	0.06	0.39
Young	0.14	0.31	0.3	0.34	0.3	0.37	0.12	0.18	0.13	0.34	0.17	0.28
OVERALL	0.12	**0.3**	**0.36**	0.29	0.3	**0.33**	0.08	0.13	**0.28**	0.12	0.13	**0.2**

Table A.3: ERROR RATE AND AMR 1. *This table shows the error rate and the amr for the AFFACT-b and the AFFACT-ub (with masks version 1). In the "OVERALL" row the greater value of the column pairs o/o, gt p/gt n, and pr c/pr nc is highlighted with bold font and color. o = overall, gt p = ground truth positive, gt n = ground truth negative, pr c = prediction correct, pr nc = prediction not correct*

List of Figures

List of Tables

List of Listings

Bibliography

Burnham, K. P. (2002). *Model selection and multimodel inference : : a practical information-theoretic approach*. Springer, New York, New York, second edition.

Chan, J.-S., Hsu, G.-S. J., Shie, H.-C., and Chen, Y.-X. (2017). Face Recognition by Facial Attribute Assisted Network. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3825–3829.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847.

Chen, Y. (2022). Explainable Classification of COVID-19 in Chest X-ray Images. Master's thesis, University of Zurich.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Desai, S. and Ramaswamy, H. G. (2020). Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Draelos, R. L. and Carin, L. (2020). Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*.

Dugăeșescu, A. and Florea, A. M. (2022). Evaluation of Class Activation Methods for Understanding Image Classification Tasks. In *2022 24th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 165–172. IEEE.

Fahrmeir, L. (2016). *Statistik : Der Weg zur Datenanalyse*. Springer-Lehrbuch. Springer Berlin Heidelberg, Berlin, Heidelberg, 8th ed. 2016. edition.

Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., and Li, B. (2020). Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. *The 31st British Machine Vision Conference*.

Günther, M., Rozsa, A., and Boult, T. E. (2017). AFFACT: Alignment-Free Facial Attribute Classification Technique. In *International Joint Conference on Biometrics (IJCB)*.

Hand, E., Castillo, C., and Chellappa, R. (2018). Doing the Best We Can with What We Have: Multi-Label Balancing with Selective Learning for Attribute Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. Technical Report 07-49.

Jiang, Pengand Zhang, C., Hou, Q., Cheng, M., and Wei, Y. (2021). LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. *IEEE Transactions on Image Processing*, 30:5875–5888.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and Simile Classifiers for Face Verification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 365–372.

Lin, M., Chen, Q., and Yan, S. (2014). Network In Network. In *International Conference on Learning Representations*.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*.

Mao, L., Yan, Y., Xue, J.-H., and Wang, H. (2020). Deep Multi-task Multi-label CNN for Effective Facial Attribute Classification. *IEEE Transactions on Affective Computing*, 13(2):818–828.

Muhammad, M. B. and Yeasin, M. (2020). Eigen-CAM: Class Activation Map using Principal Components. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE.

Nguyen, H. M., Ly, N. Q., and Phung, T. T. T. (2018). Large-Scale Face Image Retrieval System at Attribute Level Based on Facial Attribute Ontology and Deep Neuron Network. In Nguyen, N. T., Hoang, D. H., Hong, T.-P., Pham, H., and Trawiński, B., editors, *Intelligent Information and Database Systems*, pages 539–549, Cham. Springer International Publishing.

O'Shea, K. and Nash, R. (2015). An Introduction to Convolutional Neural Networks. *arXiv preprint arXiv:1511.08458*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rudd, E. M., Günther, M., and Boult, T. E. (2016). MOON: A Mixed Objective Optimization Network for the Recognition of Facial Attributes. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 19–35, Cham. Springer International Publishing.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L. S., and Gao, W. (2018). Multi-Task Learning with Low Rank Attribute Embedding for Multi-Camera Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1167–1181.

Sun, Y., Chen, Y., Wang, X., and Tang, X. (2014). Deep Learning Face Representation by Joint Identification-Verification. *Advances in Neural Information Processing Systems*, 27.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. (2020). Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25.

Wu, H., Bezold, G., Günther, M., Boult, T., King, M. C., and Bowyer, K. W. (2023). Consistency and Accuracy of CelebA Attribute Values. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3257–3265.

Zhong, Y., Sullivan, J., and Li, H. (2016). Leveraging Mid-Level Deep Representations for Predicting Face Attributes in the Wild. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3239–3243. IEEE.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhuang, N., Yan, Y., Chen, S., and Wang, H. (2018). Multi-task Learning of Cascaded CNN for Facial Attribute Classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2069–2074.