Correlation Analysis of Facial Attributes

with Respect to Face Identity

Bachelor Thesis

Raffael Mogicato

18-742-767

Submitted on

July 22 2021

Thesis Supervisor

Prof. Dr. Manuel Günther

University of Zurich^UZH

Department of Informatics

Bachelor Thesis

Author: Raffael Mogicato, raffael.mogicato@uzh.ch

Project period: 22.01.2021 - 22.07.2021

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

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Manuel Günther for guiding me through this project. His valuable advice and knowledge were instrumental when writing this thesis. I am also thankful for the help of my family, particularly that of my sister, Alina, who supported me each step of the way and helped me accomplish this task. I would also like to express my deep appreciation to my friends Colin and Gianluca for their insightful advice and support.

Abstract

In recent years, research has made tremendous progress in the classification of facial attributes through convolutional neural networks (CNNs). Such neural networks specialize in extracting various facial attributes from images. In this thesis, we investigate how these attributes correlate with identity and whether such correlations can be used to improve the classification accuracy of extractions from neural networks. We introduce a method for calculating correction terms based on statistical metrics of each attribute of ran identity. These correction terms are then applied to the extractions of neural networks, resulting in reweighted values which possess a lower average error rate than the original extractions. We offer approaches for both unbalanced and balanced CNNs. For the balanced networks, we adapt our correction terms to the domain distribution in order to consider the imbalance of classes. We show that this approach achieves similar results for manually annotated identity labels and identity labels that are inferred from the clustering of features that are extracted with neural networks. Our results show that it is indeed possible to lower the average attribute classification error rates for both neural networks based on the correlation of these attributes with identity.

Zusammenfassung

In den letzten Jahren gab es in der Forschung enorme Fortschritte bei der Klassifizierung von Gesichtsattributen durch Convolutional Neural Networks (CNN). Diese künstlichen neuronalen Netze sind darauf spezialisiert, verschiedene Gesichtsattribute aus Bildern zu extrahieren. In der vorliegenden Arbeit wird untersucht, wie solche Attribute mit der Identität korrelieren und wie diese Korrelation benutzt werden kann, um die Klassifizierungsgenauigkeit von Extraktionen neuronaler Netze zu verbessern. In dieser Arbeit wird eine Methode zur Berechnung von Korrekturtermen für jedes Attribut einer Identität vorgestellt. Diese Korrekturterme basieren auf statistischen Kennzahlen der Attribute einer Identität. Die Extraktionen von neuronalen Netzen werden mit diesen Korrekturtermen neu berechnet, was zu neu gewichteten Werten führt, die eine niedrigere durchschnittliche Fehlerrate besitzen als die Extraktionen selbst. In dieser Arbeit werden Ansätze für unausgewogene und ausgewogene neuronale Netze präsentiert. Beim ausgewogen Ansatz werden die Korrekturterme an die Klassenverteilung einzelner Attribute angepasst. Schliesslich wird gezeigt, dass der gewählte Ansatz ähnliche Resultate mit Identitätsinformationen erzielt, die von menschlicher Beschriftung stammen, wie auch mit solchen Identitätsinformationen, die durch eine Clusteranalyse von Extraktionen eines weiteren CNN gewonnen werden.

Contents

Chapter 1

Introduction

In recent years, the prediction and classification of facial attributes have been the subject of increasing attention in the literature, due to the various and widespread uses of facial attribute information. These attributes are not only machine detectable, but also have a semantic meaning that is understandable by humans (Kalayeh et al., 2017). Examples of such attributes are "Big Nose", "Bushy Eyebrows" or "Rosy Cheeks". An important driver of this research is large-scale data sets such as the CelebA data set (Liu et al., 2015). This data set contains 202,599 images of 10,177 identities. Each image was manually annotated with 40 binary facial attributes. The current state-of-the-art approach consists of the usage of convolutional neural networks (CNNs) that employ a multi-label approach to classify attributes, trained with and tested on the aforementioned manual labels. Two examples of this approach are the Mixed Objective Optimization Network (MOON) for the Recognition of Facial Attributes (Rudd et al., 2016) and the Alignment-Free Facial Attribute Classification Technique (AFFACT) (Günther et al., 2017).

Some of the 40 binary facial attributes are independent of the person's identity, (e.g. "Smiling", "Mouth Slightly Open"), though most attributes are bound to the subject's identity. In this thesis, we explore the stability of these facial attributes across subjects' identity and assess the predictive quality of this stability. A high stability of an attribute indicates that the attribute is related to the subject's identity. For example, if a subject possesses the attribute "Male" in 9 out of 10 pictures, it is highly likely that they will also possess the attribute in the 10th picture. While some approaches have already successfully incorporated identity information into their CNNs (Cao et al., 2018), the relationship between attributes and face identity remains mostly unexploited.

In the first step in this research, we identify how stable attributes are based on the ground truth annotations provided by the CelebA data set. In a second step, we apply the findings from the manually annotated attributes to automatically extracted annotations which are extracted using AFFACT (Günther et al., 2017) to improve the classification accuracy. To achieve this, we calculate correction terms for each attribute of each identity and then apply these terms to the extractions from two CNNs. This process of reweighting has a larger effect on the stable attributes of an identity and a smaller (or no effect at all) on the unstable ones. We conduct a small number of experiments to identify the most fitting formula for our reweighting process.

A major challenge of machine learning is biases in data sets. In this bachelor thesis, we address the issue of the class imbalance of attributes using two different CNNs, one that considers class imbalance, and one that does not. The core of this issue is that there is an unbalanced distribution of the existence and absence of certain attributes. Most attributes have a majority class; for example, the absence of a certain attribute, such as "Bushy Eyebrows", may be more common in a data set than its existence. This unequal distribution can lead to biased classifiers that are better at classifying attributes of the majority class, as there is a larger number of samples to learn from (Rudd et al., 2016). Both networks are based on the aforementioned AFFACT approach; while one CNN is optimized for maximal attribute classification accuracy, the other, balanced approach also considers the distribution of these attributes. This balanced approach uses a domain-adapted loss function that combats effects related to the class imbalance of the data set. The goal of this thesis is to reweight the extractions of these two CNNs using identity-related statistical measurements, namely the mean and standard deviation of each attribute of an identity. The two CNNs require different approaches, as we the issue of class imbalance for the balanced network must be considered. This means that the calculated terms we apply to the extraction should also consider class imbalance. At first, this identity-based reweighting approach uses the manually annotated identity labels provided by the data set.

Lastly, we use another CNN that specializes in facial recognition to extract embedded features, a representation of facial information through a vector using ArcFace (Deng et al., 2019). By clustering these features, we gain identity labels that are automatically extracted, rather than manually labeled to minimize the reliance on ground truth labels. We thus propose an approach that improves the classification accuracy by exploiting attribute correlation with respect to face identities. This approach is shown to successfully lower the average error rate for extractions from two AFFACT CNNs. One of these neural networks employs a balanced approach and one employs an unbalanced one. Further, we show that our approach can be applied to both manually labeled identities and identities that are clustered based on automatically extracted facial information.

Chapter 2

Related Work

2.1 Facial Attribute Classification

Facial attribute classification has received considerable attention in the field of computer vision, due to the various real-world applications of these attributes. Facial attributes are applied in a variety of tasks, such as face verification (Song et al., 2014), face recognition (Manyam et al., 2011), face image retrieval (Fang and Yuan, 2018), and image searches using descriptive attributes (Kumar et al., 2011). Facial attributes are semantically meaningful, meaning that not only machines are able to, and classify attributes, but that this classification is human understandable. This property of human understandability allows humans to understand the results of attribute classification easily and naturally. A key driver of progress in this field of research is large-scale image data sets that are labeled with facial attributes, such as CelebA (Liu et al., 2015) or Labeled Faces in the Wild (Huang et al., 2012). The CelebA data set is of particular interest in attribute prediction, as faces are hand-labeled with numerous facial attributes, a process that is made possible by the easy comprehensibility of these attributes by humans. Manually labeled data sets are a great asset for supervised machine learning techniques, as it provides the ground truth on which to train and test neural networks.

The use of CNNs to extract attributes is the current state-of-the-art approach to facial attribute classification. Examples of such work are papers by Zhang et al. (2014), Liu et al. (2015), and Günther et al. (2017). CNNs are neural networks that use a mathematical operation called convolution in at least one of the networks layers (Goodfellow et al., 2016) and are well-suited for pattern recognition tasks (Valueva et al., 2020), such as the recognition and classification of facial attributes. These networks employ key concepts that are drawn from neuroscience, which means that their structure is somewhat similar to the human visualization process. The way CNNs classify attributes can be divided into two general approaches: Single-label and multi-label.

The single-label approach, which was employed by Kang et al. (2015) and Zhang et al. (2014), considers the classification of each attribute as a single, independent problem and, therefore, disregards the correlations between attributes. In the multi-label approach, multiple attributes are predicted simultaneously in an end-to-end trained network. This method is well-suited for facial attribute classification tasks, as each image of a face is naturally related to multiple attributes (Mao et al., 2020). This method has been applied by Rudd et al. (2016), Kang et al. (2015), and Zhuang et al. (2018) and is currently the most successful approach to facial attribute classification. However, this multi-label approach also faces challenges, such as an increased difficulty in dealing with class imbalance. Class imbalance is a general problem faced by CNNs in relation facial attribute data: Some attribute classes have many more samples than others. These classes can be divided into majority and minority classes. Minority classes are classes with very few samples while majority classes have many samples. Zheng et al. (2020) determined that the largest ratio between an attribute's majority and minority class in the CelebA data set was 43:1. Such large imbalance ratios lead to biased classifiers, which fail to properly evaluate features learned from the minority class.

Rudd et al. (2016)proposed the MOON as a solution to the issues that stem from class imbalance. This network seeks to simultaneously maximize the prediction accuracy for both classes of each attribute through multi-tasked training on multi-labeled data sets, such as the CelebA data set. This method allows for a better fitting approach for determining minority classes, as the distribution of classes is considered, which allows for more accurate extractions for attributes of inferior classes. Further, Rudd et al. (2016) proposed a way to measure the accuracy with inclusion of the domain distribution, allowing for a balanced accuracy measurement that considers the issue of class imbalance.

2.2 Attribute Grouping and Correlation

In recent years, several novel approaches based on attribute correlation have been introduced to further improve facial attribute classification. These approaches involve grouping several attributes under the assumption that there exist strong correlations in and between these groups. Hand and Chellappa (2017) used a multi-task deep CNN (MCNN) that considered attribute correlation on the mid-level of the neural network. This correlation is based on attribute groups, namely nine manually determined attribute groups that categorized 40 attributes. These groupings are based on attribute location and semantics, which led to categories such as *Nose*, with the attributes "Big Nose", "Pointy Nose" or *Mouth* with "Big Lips", "Smiling", "Lipstick", and "Mouth Slightly Open". At the final layer, an auxiliary network receives the scores from the MCNN and finalized the prediction by allowing for interaction between attributes on the score level. Cao et al. (2018) adopted a similar approach. Like in the MCNN approach, the authors split the attributes into groups based on location. Rather than including semantically meaningful groups they solely focused on the attributes' location in the image: The four groups described by them are an upper, middle, lower and whole-image group. The authors extend the MCNN by introducing a partially shared structure to allow for more interaction on the higher levels of the neural network. A further unique element of this approach is the inclusion of identity information to the partially shared MCNN, resulting in the incorporation of this information with local constraints, achieving a state-of-the-art error rate of just 7%. Further approaches to attribute grouping exist, such as grouping according to holistic vs. local and nominal vs. ordinal attributes, as proposed by Han et al. (2018) or grouping by objective and subjective attributes, as outlined Mao et al. (2020).

Chapter 3

CelebA

3.1 Data Set

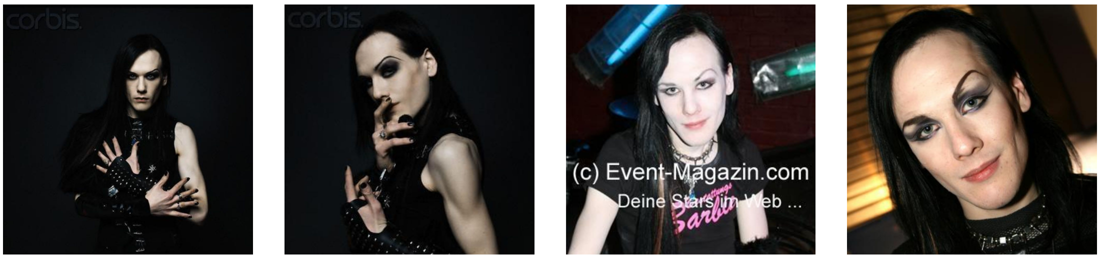
This data set was created by Liu et al. (2015) with the aim of producing a large-scale data set to train and test CNNs on. The data set consists of 10,177 identities and a total of 202,599 images. Each of the images was manually annotated with 40 binary facial attributes and five key facial landmarks. The five key facial landmarks are the left eye center, right eye center, tip of the nose, left mouth corner and right mouth corner. More importantly, each image is annotated with one identity. This identity annotation is referred to as the ground truth identity label. All 40 attributes are provided in Table 3.1.

We refer to these manual annotations as the ground truth. The ground truth is essential for training neural networks and subsequently for measuring their accuracy, as the ground truth describes direct observations. However, these manual annotations are not without flaws: Some of these attribute are highly subjective, as their labeling may heavily depend on the person labeling them, e.g., attributes such as "Attractive" or "Young". Further, some images may be labeled inconsistently, illustrated in Figure 3.1. This is partially caused by the decision to only use binary labels for all attributes, even though most attributes exist on a continuous range in the real world. While this leads to a disparity between the real world and the ground truth used to train neural networks, it increases the feasibility of creating such a large-scale data set. The images in this data set are of celebrities. Most images are taken in a relatively controlled environment, compared with a real-world scenario. Still, the images exhibit a large variety of poses and background clutter.

The CelebA data set is partitioned into three sets. The first and largest partition is the training set, which contains 8,192 identities with 162,770 images. The remaining two partitions are the

Table 3.1: ATTRIBUTES OF CELEBA.

"5 o’ Clock Shadow"	"Arched Eyebrows"	"Attractive"	"Bags Under Eyes"
"Bald"	"Bangs"	"Big Lips"	"Big Nose"
"Black Hair"	"Blond Hair"	"Blurry"	"Brown Hair"
"Bushy Eyebrows"	"Chubby"	"Double Chin"	"Eyeglasses"
"Goatee"	"Gray Hair"	"Heavy Makeup"	"High Cheekbones"
"Male"	"Mouth Slightly Open"	"Mustache"	"Narrow Eyes"
"No Beard"	"Oval Face"	"Pale Skin"	"Pointy Nose"
"Receding Hairline"	"Rosy Cheeks"	"Sideburns"	"Smiling"
"Straight Hair"	"Wavy Hair"	"Wearing Earrings"	"Wearing Hat"
"Wearing Lipstick"	"Wearing Necklace"	"Wearing Necktie"	"Young"



(a) Image 014378.jpg (b) Image 025802.jpg (c) Image 027499.jpg (d) Image 106131.jpg

Figure 3.1: INCONSISTENT GENDER. An example of inconsistent labeling: This person (Id 6101 in CelebA) was labeled as "Male" in only 6 out of 10 images, which means that some images are labeled incorrectly. In subfigures (a) and (b) the values of the attribute "Male" are 1, in (c) and (d) -1.

validation and test sets, which are both similar in size. The validation set contains 985 identities with 19,867 images, the test set 1,000 identities with 19,962 images. Neural networks are trained on the training set, which is also the reason why it is by far the largest partition. The validation set is then used to optimize parts of the architecture, e.g., for selecting a loss function. Finally, the results are then tested on the test set. This set is independent of the others, and its main purpose is to measure performance. The reason for these partitions is to avoid overfitting and to reduce the introduction of bias through the data set as much as possible. In this thesis, we focus on the validation and test sets. The validation set is used to decide on specific reweighting functions and the test set to measure how accurate the results are. Each image in the data set has a corresponding identity. The number of images for each identity varies between 1 and 35, with the median being 19.9 images per identity. In total, 1,055 of the 10,177 identities have five or fewer images attached to them.

3.2 Class Imbalance

A further notable property of the data set is the distribution of binary attributes. Attributes have a value of either 1 or -1, denoting the presence or absence of said attribute. The distribution of presence and absence of attributes in the CelebA data set is not always equal, this means that there is a minority and a majority class for all attributes in the data set, as no attribute has the same number of samples for both classes. Most of the time, the absence of an attribute is more common than its presence. This imbalance is most likely intrinsic, thus a result of the nature of the data space, as the 40 - somewhat arbitrarily chosen - attributes are most likely not distributed evenly in the human population. However, some of the imbalance may stem from the sampling of the data set, also known as extrinsic imbalance (He and Garcia, 2009). A possible example is the attribute "Attractive". There may be a sample bias for this attribute, as the subjects in the CelebA data set are celebrities, and, therefore, represent a sample of individuals that are commonly more attractive than the general population. An issue that arises from this unequal distribution is that of class imbalance when training CNNs on an imbalanced data set. Due to the fact that the minority class contains significantly fewer samples, neural networks tend to over-classify the majority class, which means that members of the minority class are misclassified to be a part of the majority class (Johnson and Khoshgoftaar, 2019). For this reason, it is important to consider class imbalance during the whole process; otherwise, the resulting approach to attribute classification may be unable to discriminate features of the minority class.

The ratio of positive and negative labels, denoting the existence and absence of an attribute

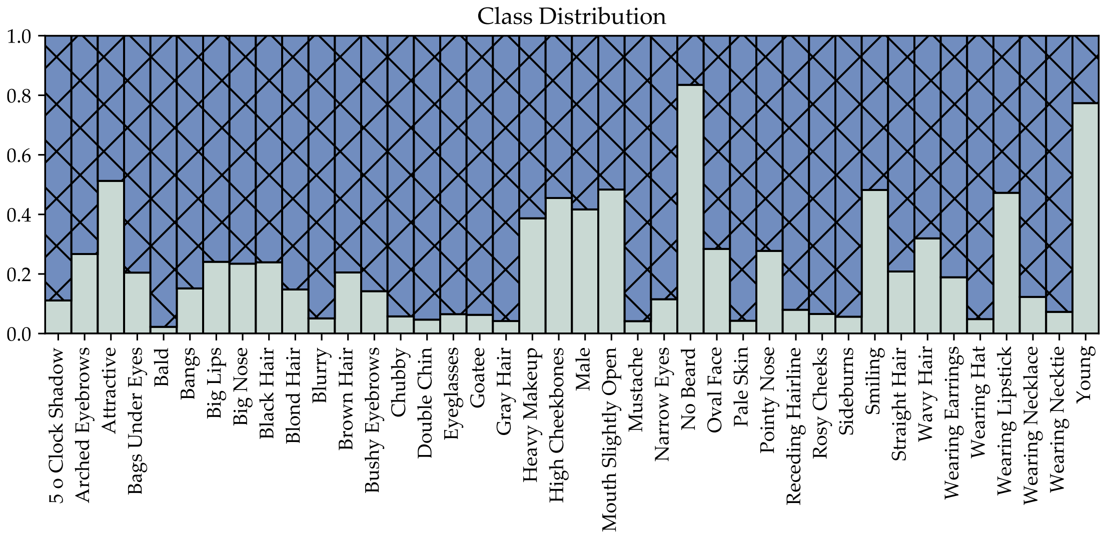


Figure 3.2: CLASS IMBALANCE. The ratio of positive (light) and negative (dark, hatched) classes. The former denotes the presence of an attribute, while the latter the absence of one.

respectively, shown in Figure 3.2. The most extreme case in the CelebA data set is the attribute "Bald", where only 2.25% of all samples are positive: This means that we can blindly classify each instance as negative and still achieve a remarkably high accuracy of 97.75%. Such a severe imbalance illustrates the need for an approach that not only seeks to optimize accuracy, but rather also consider the underlying imbalance of the data set. As mentioned before, the negative class, i.e., the absence of an attribute is more common than the positive class. Only the attributes "No Beard" and "Young" have positive majority classes.

Class imbalance even has an impact on how attributes are learned by CNNs if the class distribution is rather equal, as shown by Rudd et al. (2016).

Chapter 4

CNNs

4.1 AFFACT

Different approaches exist for the task of facial attribute classification through neural networks. In this thesis, we use the AFFACT introduced by Günther et al. (2017). This technique employs a residual learning framework (He et al., 2016) for training, allowing for effective training of deep convolutional neural networks. This training results in deep Residual Networks (ResNets), which are first trained for generic image recognition and then fine-tuned on the training partition of the CelebA data set. Günther et al. (2017) showed that using an ensemble of three such networks outperform a single ResNet. A distinctive feature of AFFACT is that it is less reliant on the alignment of images. To lower the reliance on alignment, random perturbations regarding scale, angle, shift, and blur are added to the faces in the training set, increasing the robustness to facial misalignment. This technique allows for similar performances for detected bounding boxes and detected facial landmarks. However, in our experiments, we use bounding boxes that are determined by manual annotations, to keep errors due to misalignment to a minimum.

4.2 Preprocessing

In our experiments we align and crop the images from the data set in the same manner as Günther et al. (2017) using the manually annotated facial landmarks to determine a bounding box. We use four of the five provided landmark labels: The left and right eye $\mathbf{t}_{e_r}, \mathbf{t}_{e_l}$ and the left and right mouth corner $\mathbf{t}_{m_r}, \mathbf{t}_{m_l}$ where $\mathbf{t} = (x, y)^{\mathrm{T}}$. Using these landmarks, we calculate the eye center $\mathbf{t}_e$, mouth center $\mathbf{t}_e$ and consequently the eye-mouth distance d:

$$\mathbf{t}_e = \frac{\mathbf{t}_{e_r} + \mathbf{t}_{e_l}}{2}, \mathbf{t}_m = \frac{\mathbf{t}_{m_r} + \mathbf{t}_{m_l}}{2}, d = \|\mathbf{t}_e - \mathbf{t}_m\| \tag{4.1}$$

The eyes are then aligned on a horizontal line before the bounding box is added. These three measures determine the size and location of the square bounding box, denoted with the top left corner at x_l, y_t with the side length $s = d \cdot 5.5$:

$$x_l = x_e - 0.5 \cdot s, y_t = y_e - 0.45 \cdot s \tag{4.2}$$

The images are then cropped to their bounding box and resized to 224 x 224 pixels. We save the images as PNGs, rather than in their original JPG format, to avoid lossy compression.

4.3 Balanced vs. Unbalanced Networks

In our experiments we use two CNNs that use AFFACT to extract the attributes. The difference between these two networks is that one considers the issue of class imbalance, while the other one does not. If an attribute has a large majority class, the predictions of an unbalanced network are better for this majority class and comparatively worse for the minority class. This means that the unbalanced approach can result in a high overall accuracy, but a low accuracy score for classifying attributes of a minority class. To address this imbalance, which is the consequence of biases in the data set, Rudd et al. (2016) suggest a solution that utilizes a loss function that considers the distribution of classified attributes: Through a mixed-objective function, domain-adapted weights are calculated that consider the source and target distribution of each class for each attribute. This domain adaption is then incorporated into the multi-task loss layer of the CNN, resulting in a loss layer that can adapt the biased distribution in the training partition to a target distribution. Such a domain-adapted multi-task loss layer can be combined with the above described AFFACT, resulting in a balanced CNN that is well suited to dealing with class imbalance. Such balanced approaches generally achieve lower accuracy values than their unbalanced counterparts. However, the measurement of overall accuracy fails to consider the distribution of classes. This leads to networks that excel at classifying the majority class, but often fail to identify minority samples.

In this thesis, we consider both a balanced and an unbalanced AFFACT CNN. The balanced network is denoted as AFFACT-B, and the unbalanced network is denoted as AFFACT-U. AFFACT-U was published by Günther et al. (2017), while AFFACT-B was presented by same authors but remains unpublished.

4.4 Extraction and Accuracy Evaluation

The difference between balanced and unbalanced networks means that the evaluation of the results should be approached differently. The accuracy measures the number of correct classifications in relation to the number of classified objects. However, this measurement does not consider the issue of class imbalance, as members of a small minority class can be classified incorrectly most of the time while still achieving a high accuracy. While we can simply evaluate the unbalanced network AFFACT-U with the accuracy and consequently the unbalanced error rate ER_u, we need to consider the balanced error rate ER_b for the balanced network AFFACT-B. These two different approaches to accuracy have also been used by Rudd et al. (2016) and are calculated as follows using true positives T_P, false positives F_P, true negatives T_N and false negatives F_N:

$$A_u = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \qquad ER_u = 1 - A_u \tag{4.3}$$

For ER_b we calculate the false negative rate FNR and the false positive rate FPR. FNR describes the relation of incorrectly classified positive values (F_N) and the total number of positives values. The FPR the number of incorrectly classified negative values (F_P) in relation to all negative values. We take the mean of these two rates to get ER_b:

$$\begin{aligned} FNR &= \frac{F_N}{T_P + F_N} \\ FPR &= \frac{F_P}{F_P + T_N} \\ ER_b &= \frac{FNR + FPR}{2} \end{aligned} \tag{4.4}$$

The CNNs then extract values on a continuous range for each of the 40 attributes from an image. The value of each attribute indicates how the neural network evaluates the presence or absence of an attribute in an image. The exact range of this value varies between attributes. Figure 4.1 provides an example of how these values are distributed by visualizing the extracted values from AFFACT-B in a box plot. The extractions of AFFACT-B and AFFACT-U are very similar in their distribution: The vast majority of values are approximately between -2 and 2 with very few outliers beyond this. Nearly all attributes have at least half their values between -1 and 1. However, the ground truth labels of the data are annotated binary, this means that we must map the continuous values of the extraction to a binary space: In this thesis we simply classify all values x as either 1 or -1, as follows:

$$\mathrm{bin}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{4.5}$$

A downside of this mapping is that a binary classification of facial attributes does not necessarily represent reality, as most facial attributes lie on a continuous range, just like the extractions of our CNNs. This means that forcing all values to be binary leads to issues, especially in cases in which the existence or absence of an attribute is ambiguous, and the output of the CNNs is close to 0. This impacts the accuracy evaluation, as ambiguous attributes are still assigned binary values in the manual annotations.

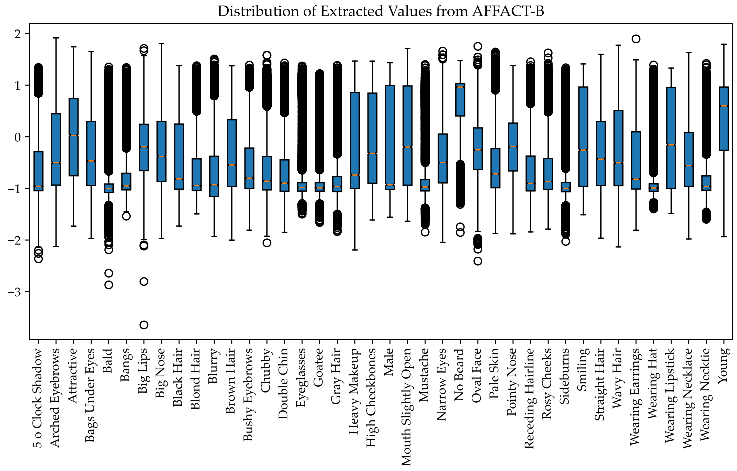


Figure 4.1: Distribution of Extractions. This plot depicts the distribution of values extracted with AFFACT-B from the validation set. The blue box represents the first to the third quartile, while the orange line indicates the median. The whiskers extend to the last data point contained within the interquartile distance from the upper and lower quartile. Data points beyond that are considered outliers and are depicted as circles.

Chapter 5

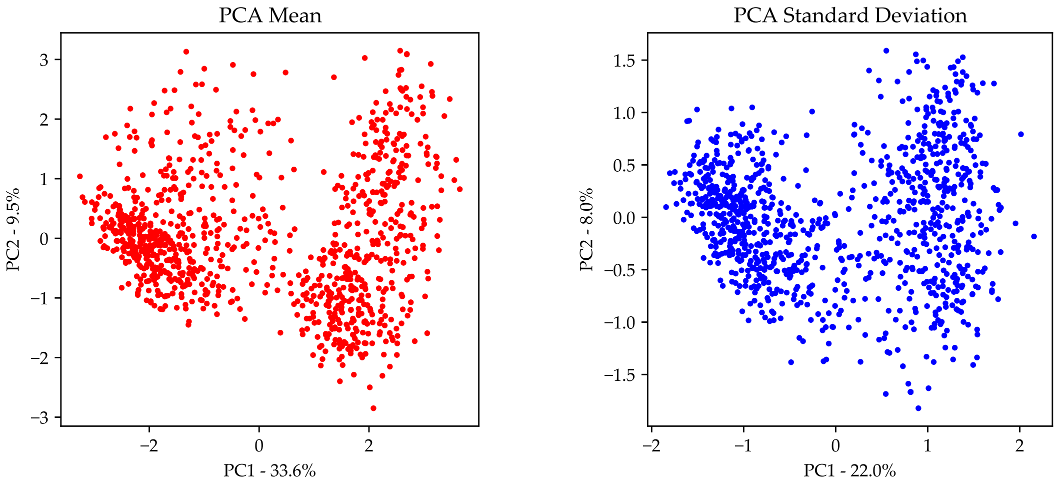
Approach

5.1 Stability of Ground Truth Attributes

A prerequisite to our approach is to explore the relationship between identity and the stability of its attributes. In previsions research manual grouping has been used to divide attributes into groups, such as the location-based groups presented by Cao et al. (2018) and Hand and Chellappa (2017). The approach of this thesis is somewhat related to such grouping, though rather than manually determining which attributes we consider stable for an identity, we base our approach on statistical metrics. While this approach is more complex, it is also more versatile, as the stability of an attribute may depend more on the identity rather than on the attribute itself. The two most important statistical measurements used to model stability are the mean μ and standard deviation σ. To explore the relationship between identities and attribute stability, we take the annotations for each of the n pictures and calculate the two measurements for each attribute a from the value of that attribute x in a single image j.

$$\begin{aligned} \mu_{ia} &= \frac{1}{n}\sum_{j}^{n} x_j \\ \sigma_{ia} &= \sqrt{\frac{\sum_j^n (x_j - \mu_{ia})^2}{n}} \end{aligned} \tag{5.1}$$

A key challenge involves the multidimensional nature of our data: The large number of identities combined with the fact that each identity has multiple attributes, makes visualizing and consequently analyzing how our two statistical measurements correlate with their identity challenging. To explore this correlation, we employ principal component analysis (PCA). PCA is a dimension reduction technique that utilizes the eigenvectors of the correlation matrix to calculate the principal components (Jolliffe, 2011). In our scenario, we have m identities, we represent each identity with a vector $\mathbf{v}$ containing a values, one for each of the 40 attributes. We aim to reduce these 40 dimensions to just 2. PCA achieves this by finding linear combinations $\mathbf{c}_1^{\mathrm{T}}\mathbf{v}, \mathbf{c}_2^{\mathrm{T}}\mathbf{v}, ..., \mathbf{c}_{40}^{\mathrm{T}}\mathbf{v}$ called principal components. These 40 principal components successively have maximum variance for the data and are uncorrelated with previous $\mathbf{c}_k{}^{\mathrm{T}}\mathbf{v}$. When we solve this maximization problem, we find that $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_{40}$ are the eigenvectors of the covariance matrix, $\mathbf{S}$, of the data (Jolliffe, 2011) and correspond to the 40 largest eigenvalues. To gain the desired two-dimensional model of this data, we can simply take the first two principal components $\mathbf{c}_1^{\mathrm{T}}\mathbf{v}, \mathbf{c}_2^{\mathrm{T}}\mathbf{v}$ to plot our data points, i.e. identities, as an ellipsoid with $\mathbf{c}_1^{\mathrm{T}}\mathbf{v}, \mathbf{c}_2^{\mathrm{T}}\mathbf{v}$ as the axes. These first two principal components explain the highest amount of variance from all 40 principal components.



(a) PCA of the mean

(b) PCA of the standard deviation

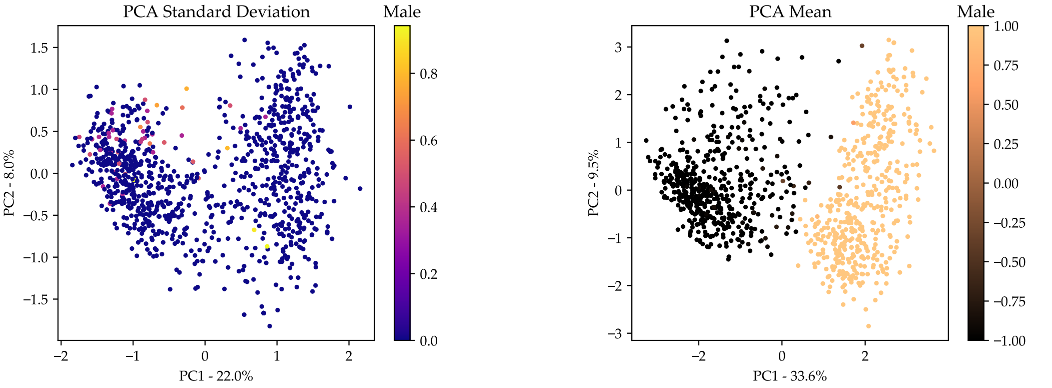
Figure 5.1: BASIC PCA PLOTS. These two plots show the first two principal components for the mean and standard deviation of the manually labeled attributes in CelebA. Each dot represents one identity. The axes correspond to the two principal components that explain the most variance. This allows for a two dimensional plot that preserves a large portion of the variation of our data. Subfigure (b) shows the distribution according to standard deviation, that means dots, i.e. identities, closer to each other have similar standard deviations for their attributes and dots further away less similar ones. Subfigure (a) depicts the same effect, but for the mean of identities rather than the standard deviation.

For the concrete implementation of this PCA we use the scikit-learn module for PCA[1]. To maintain consistency in the separation between the training, validation and test sets, we first fit the PCA model on the training set and then apply the dimensionality reduction to the data of the validation set. We generate two models, one for each of our statistical metrics. In the first model we use $\mathbf{v}^{\mathrm{T}} = (\mu_{i1},\ \mu_{i2},\ ...,\ \mu_{i40})$ for each of our identities. This means our first PCA model uses the mean of each attribute of each identity as a measurement. Analog to this, our second model uses $\mathbf{v}^{\mathrm{T}} = (\sigma_{i1},\ \sigma_{i2},\ ...,\ \sigma_{i40})$. This means that both models are first fitted with $m = 8192$, the number of identities in the training model and then apply the dimensionality reduction to the data of the validation set where $m = 985$.

The plots we generate can be seen in Figure 5.1. In these two plots we visualize how the two measurements vary between identities. A larger distance between two dots indicates a larger variance of the measurements, i.e. the mean and standard deviation, of attributes from an identity. The first model explains 43.1% of the variance of the mean and the second 30.0% of the variance of the standard deviation. These values are rather low, as PCA usually aims to model a larger amount of variance. However, the two models still allow us to analyze how the two measurements relate to identity. Subfigure 5.1a depicts the first model of the mean. We can see that there are two clusters. This indicates that both these clusters consist of attributes that have similar values for the means of their attributes. The same goes for the our second model in Subfigre 5.1b, however, here, the two clusters are not as well separated.

To explore how these clusters relate to specific attributes - and their stability - we generated plots that colorize the values of both the mean and standard deviation for each attribute, resulting in 80 plots: Two plots depicting the statistical measurements for each of the 40 attributes. This

[1] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html



(a) PCA of the standard deviation colored by attribute "Male"

(b) PCA of the mean colored by attribute "Male"

Figure 5.2: STABILITY OF ATTRIBUTE "MALE". In Subplot (a) we can see that most identities are colored blue, indicating a low standard deviation: This means that this attribute is very stable. Subplot (b) shows how the two clusters of the PCA of the mean heavily correlate with the mean of the attribute "Male". Copper indicates the presence of this attribute, black the absence of it.

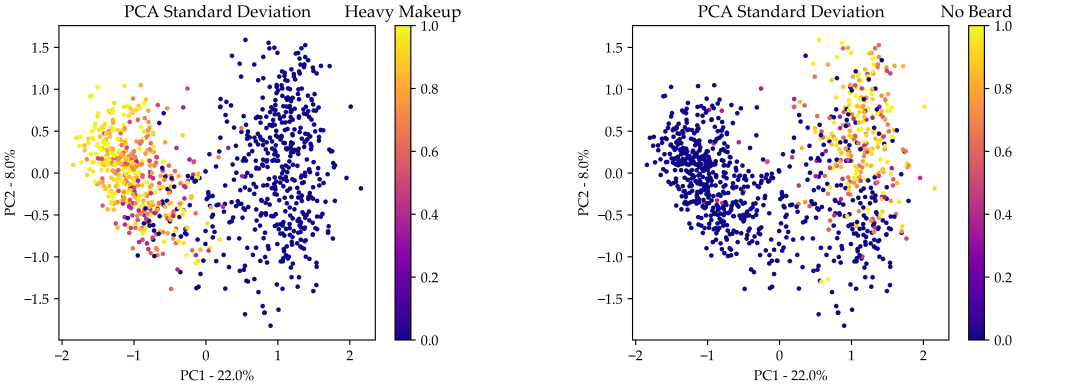
analysis is based on the ground truth labels provided by Liu et al. (2015) in their CelebA data set. This allows us to analyze the human labeled attributes first, before applying the gained insights to the data that is extracted by CNNs.

A prime example of how this visual analysis can help determine how stable certain attributes are, is illustrated in Figures 5.2 and 5.3: The standard deviation in Subfigure 5.2(a) is consistently low for the vast majority of identities, while the mean depicted in subfigure 5.2(b) is close to -1.00 for the vast identities in the left cluster, indicating that female subjects are in the left cluster, while male ones are in the right cluster.

This gender specific cluster interpretation can also be applied to the PCA of the standard deviation, the clusters for which are less visually distinct compared with those of the mean. In this case, colorizing by the attribute "Male" does not give us any indication of whether these two clusters are gender related, as this attribute is very stable, a fact that is expressed through a low standard deviation for most identities. When we look at other gender related attributes such as the attribute "No Beard" in figure 5.3(b) or "Heavy Makeup" in figure 5.3(a) there is a clearly visible difference in the standard deviation according to cluster. In the left cluster, "Heavy Makeup" has a high standard deviation, while "No Beard" has a consistently low one. This leads to the interpretation that the left cluster consists of identities, which vary in their usage of makeup but not in the absence of facial hair. This visualization shows that both measurements used in the PCA offer insights into correlation of attributes and their identity. Our approach for reweighting is based on a combination of both measurements: We consider the stability using standard deviation as a measurement, which is then combined with the mean to gain an idea to which class, positive or negative, the resulting correction term should belong to.

5.2 Reweighting of Attributes

The most challenging aspect of this thesis involved correcting the automatically extracted attributes. This reweighing of attributes has two main components: The subject's identity and



(a) PCA of the standard deviation colored by attribute "Heavy Makeup"

(b) PCA of the standard deviation colored by attribute "No beard"

Figure 5.3: PCA FOR GENDER RELATED ATTRIBUTES. Subplot (a) shows that in one cluster the attribute "Heavy Makeup" varies to a much larger degree than in the other one. In Subplot (b) the same effect can be observed for the attribute "No Beard", indicating that these clusters also correlate with the gender of the identities.

the statistical measurements relating to said identity. The approach in this thesis to attribute correction is to use the subject's identity to determine how consistent an attribute is across the same identity. If an attribute is stable across most images of an identity, we want to correct values that deviate from this consistency. To achieve this correction, we calculate a correction term that is added to or subtracted from the automatically extracted values. The absolute value of this term should be high for consistent attributes and low for inconsistent ones, as the consistency of attributes indicate a strong relationship between the person's identity and the existence or absence of that attribute. This correction term can be positive or negative, and the sign of the term for an attribute depends on the class of the mean. If an attribute is inconsistent across an identity, we expect a weak or no relationship between that attribute and the person's identity. This approach results in 40 correction terms, one per attribute for each identity. These terms are then added to the extracted values. Due to the fact that we measure accuracy through binary classification, only the sign of the resulting value must be correct. This means that "over-correcting" values to be too high or too low is an effect we can ignore. Determining the exact terms is a challenging task: On one hand it is difficult to determine a formula for the correction terms, on the other hand the issue of class imbalance is still something we must consider, particularly for the balanced network.

Consequently, we should consider the issue of class imbalance for our statistical measures μ_{ia} and σ_{ia}. For our unbalanced CNN AFFACT-U, we calculate the mean μ_{ia} and standard deviation σ_{ia} with the same formula shown in Equation 5.1, but instead of using the ground truth information, we use the automatically extracted attribute information. Again, here we calculate the mean and standard deviation for each attribute a for all identities i. Each i has n pictures, with an extracted values x for each attribute in a picture j.

$$\begin{aligned} \mu_{ia}^U &= \frac{1}{n}\sum_{j}^{n}(x_j) \\ \sigma_{ia}^U &= \sqrt{\frac{\sum_j^n (x_j - \mu_{ia}^U)^2}{n}} \end{aligned} \tag{5.2}$$

For the balanced CNN, we want to use balanced statistical measurements to calculate our correction terms. We achieve this, by adapting the extracted values to the class distribution. To do so, we employ the domain adaption function that was introduced by Rudd et al. (2016). The same function was used in the architecture of AFFACT-B for the domain-adapted multi-task loss layer. This function assigns a probability to the two classes of each attribute a. Rudd et al. (2016) first calculate the source distribution S_a of the training set for each attribute a by counting the relative occurrences of positive samples S_a^+ and negative samples S_a^- and then assign the probability by considering the binary target distribution T_a^+ and T_a^-. As shown by Rudd et al. (2016), using $T_a^+ = T_a^- = 0.5$ as a target distribution is appropriate.

$$p(a|+1) = \begin{cases} 1 & \text{if } T_a^+ > S_a^+ \\ \frac{S_a^- T_a^+}{S_a^+ T_a^-} & \text{otherwise} \end{cases} \text{ and } p(a|-1) = \begin{cases} 1 & \text{if } T_a^- > S_a^- \\ \frac{S_a^+ T_a^-}{S_a^- T_a^+} & \text{otherwise} \end{cases} \tag{5.3}$$

All values for this calculated probability are given in Table A.3. We integrate this probability by multiplying the extracted value x of an attribute a of an identity i in each image j out of the n images of that identity, giving us a mean μ_{ia}^B and consequently a standard deviation μ_{ia}^B, which is based on balanced values of x:

$$\begin{aligned} \mu_{ia}^B &= \frac{1}{n}\sum_j^n \mathrm{b}(x_j) \\ \sigma_{ia}^B &= \sqrt{\frac{\sum_j^n (\mathrm{b}(x_j) - \mu_{ia}^B)^2}{n}} \end{aligned} \tag{5.4}$$

with

$$\mathrm{b}(x) = \begin{cases} x \cdot p(+1) & \text{if } x \geq 0 \\ x \cdot p(-1) & \text{otherwise} \end{cases} \tag{5.5}$$

This domain adaption aimed to counteract the over-representation of the majority classes by reducing their values with the calculated probability. In essence, this means that we lower the values of the majority class while leaving those of the minority class unchanged. Thies yielded a more balanced consideration of the extracted attributes, as values of the minority class appear less frequent due to the class imbalance of the data set. However, this adaption of our statistical measurements for reweighting only makes sense if we consequently evaluate our results after reweighting based on the balanced error rate, since ER_b considers the effect of class imbalance. This means that we only use domain-adapted, or balanced, correction terms for extractions of AFFACT-B, since we hypothesize that lowering ER_b results in a higher ER_u.

This consideration of class imbalance requires us to use two different statistical measurements to compute our correction terms. To keep this thesis legible, we will simply refer to the mean of an attribute for an identity as μ_{ia} and the standard deviation as σ_{ia}, while their calculation depends from which CNN, AFFACT-B or AFFACT-U, they are calculated:

$$\mu_{ia} = \begin{cases} \mu_{ia}^B & \text{if calculated from extractions from AFFACT-B} \\ \mu_{ia}^U & \text{if calculated from extractions from AFFACT-U} \end{cases} \tag{5.6}$$

$$\sigma_{ia} = \begin{cases} \sigma_{ia}^B & \text{if calculated from extractions from AFFACT-B} \\ \sigma_{ia}^U & \text{if calculated from extractions from AFFACT-U} \end{cases} \tag{5.7}$$

Since we calculate σ_{ia} from extractions, it can occur that the standard deviation is larger than 1, as the range of the extracted values goes beyond the maximum values of ground truth labels which are either -1 or 1, as can be seen in Figure 4.1. However, effectively the values of σ_{ia} are rarely larger than 1 and when they are, they are only slightly larger: We found the highest σ_{ia} of all values from the validation set extracted with AFFACT-U, was σ_{ia} = 1.21. Still, the vast majority of σ_{ia} are below 1, only a few outliers actually have values larger than 1. As described above, the correction term for a stable positive / negative attribute should be higher / lower than for an unstable one. This means that the correction term should be high for lower σ_{ia}. We model this with a polynomial term for 1-σ_{ia}. For the previously mentioned outliers that possess a σ_{ia} larger than 1, we simply set the correction term to 0, as we consider their stability too low to compute a proper correction term. We found the most success using either 2 or 3 as an exponent for 1-σ_{ia}, resulting in the square and cubic approaches. For the correction towards the mean, we differentiate between directly correcting towards the mean and correcting towards the sign of the mean. The former is an approach that uses the continuous values that we receive from the CNN extractions. This results in an approach that corrects values of attributes with a mean closer to zero to a lesser degree and values with more extreme means to a larger degree. In contrast, the sign approach is a binary approach, that simply multiplies the sign of the mean with the result of the polynomial function.

$$
\begin{aligned}
w^{ia}_{\text{square mean}} &= \begin{cases} 0 & \text{if } \sigma_{ia} > 1 \\ (1-\sigma_{ia})^2 \cdot \mu_{ia} & \text{otherwise} \end{cases} \\
w^{ia}_{\text{square sign}} &= \begin{cases} 0 & \text{if } \sigma_{ia} > 1 \\ (1-\sigma_{ia})^2 \cdot \operatorname{sgn}(\mu_{\text{ia}}) & \text{otherwise} \end{cases} \\
w^{ia}_{\text{cube mean}} &= \begin{cases} 0 & \text{if } \sigma_{ia} > 1 \\ (1-\sigma_{ia})^3 \cdot \mu_{ia} & \text{otherwise} \end{cases} \\
w^{ia}_{\text{cube sign}} &= \begin{cases} 0 & \text{if } \sigma_{ia} > 1 \\ (1-\sigma_{ia})^3 \cdot \operatorname{sgn}(\mu_{\text{ia}}) & \text{otherwise} \end{cases}
\end{aligned}
\tag{5.8}
$$

with

$$
\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}
\tag{5.9}
$$

This results in 40 correction terms for each identity. The correction terms for AFFACT-B consider class imbalance through the modified mean μ^B_{ia} and standard deviation σ^B_{ia}, while the correction terms for AFFACT-U simply use the mean μ^U_{ia} and standard deviation σ^U_{ia}.

5.3 Identity Clustering

A central idea of this thesis is to exploit the identity of subjects to improve the prediction of attributes. In early experiments images are grouped according to the ground truth identity. This information was provided by the authors Liu et al. (2015) for their CelebA data set. However, using this ground truth information is problematic, as in most real-life scenarios we do not know which identity a face belongs to. Our approach to the issue of unknown identities is to acquire identity information through a face recognition CNN. A suitable candidate for this task is ArcFace by Deng et al. (2019), a state-of-the-art neural network that uses Additative Angular Margin

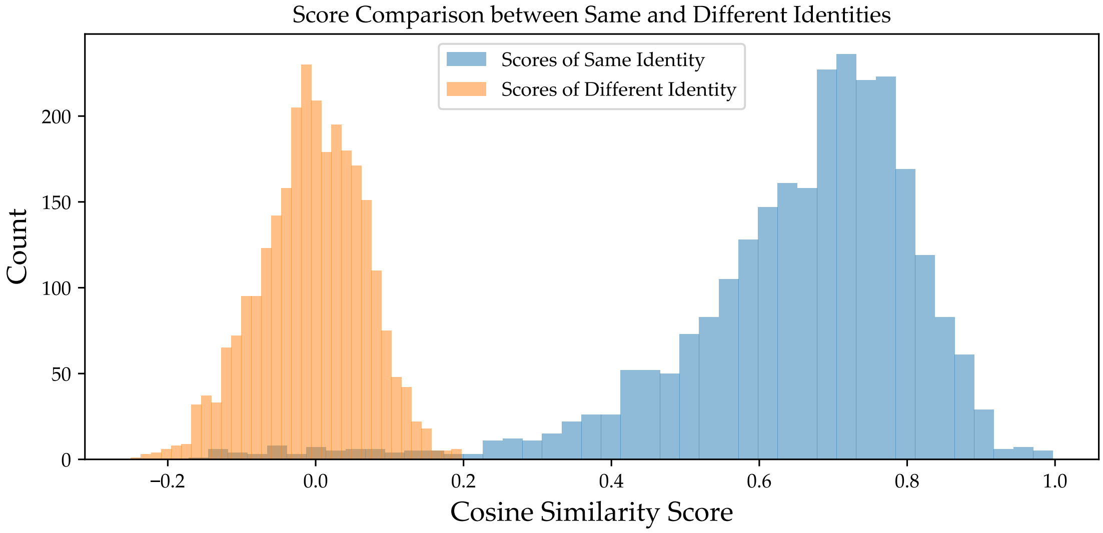


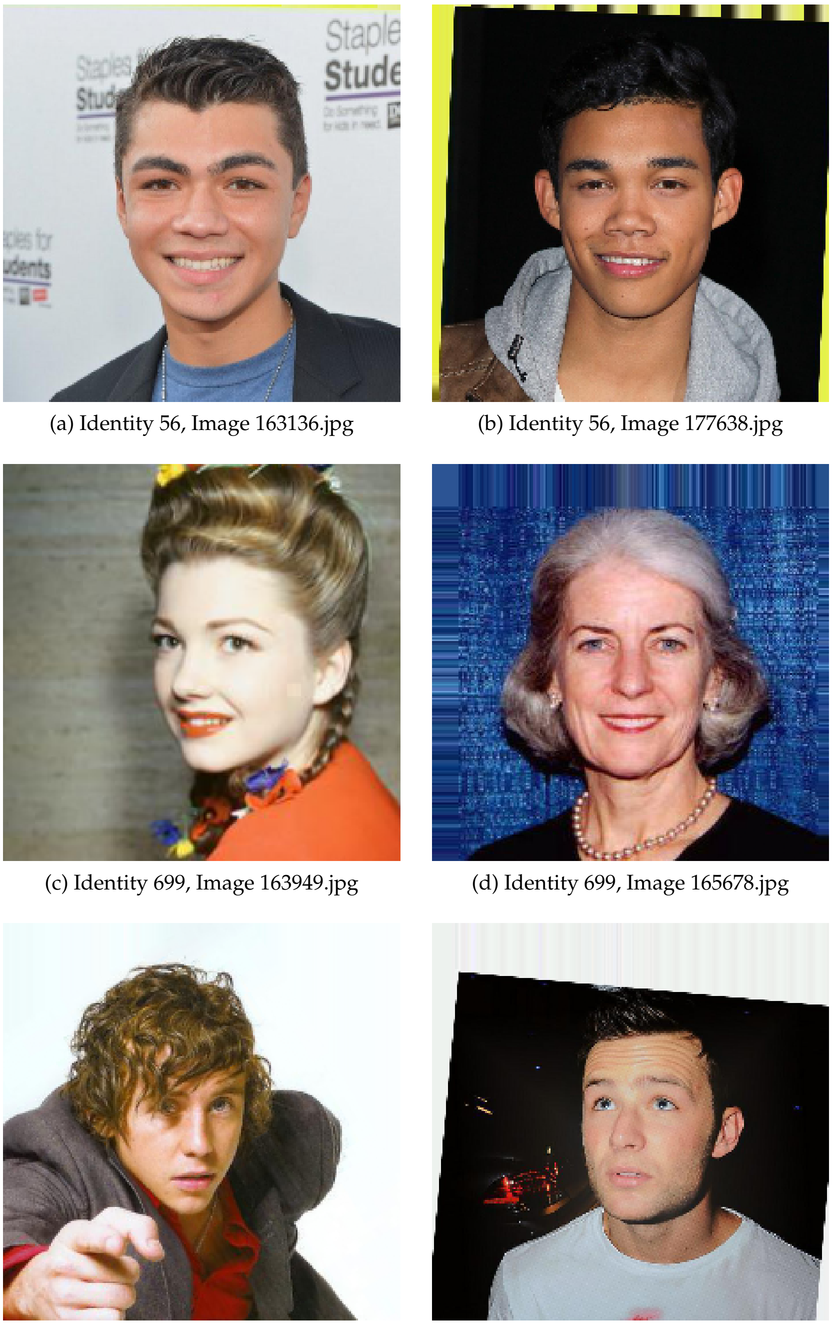
Figure 5.4: HISTOGRAM OF COSINE SIMILARITY. This plot shows how similar images of a sample size of 300 identities are using cosine similarity as a score. The blue bars show the count of scores reached when pictures of the same identity are compared to one another, orange bars show the scores of pictures from different identities.

Loss for face recognition. ArcFace is easy to implement and only requires a low amount of computational resources. Several pretrained models exist, this means that no time must be spent on training a model. The model used in this thesis is called LResNet100E-IR,ArcFace@ms1m-refine-v2[2]. This CNN requires different image preprocessing than the two neural networks used for attribute classification. The image resolution is 112 x 112 and unfortunately no instructions for facial landmark alignment are provided. However, since samples of input images exist, we are able to estimate the location of the eye and mouth sample, allowing for a close approximation to the actual alignment: Both eye and mouth center are aligned to the horizontal center. The eye center is on the 52nd pixel from the top and the mouth center on the 90th. Analog to our cropping procedure for AFFACT described in Section 4.2, we now crop the images to be of size 112 x 112 and set the new side length as $s = d \cdot \frac{90-52}{112}$, i.e. multiplying the eye mouth distance with the ratio between eye mouth distance and image height.

The outputs for each image are embedded features as 512 dimensional vectors. These embedded features are representations of the face of each image, this means that vectors of the same identity have a high similarity, while vectors of different identities have a low one. This can be seen in figure 5.4 where a sample size (n=300) of identities are compared with each other using cosine similarity. In a first step we choose one identity and select one image. This image is then compared to images of its own identity and images of a different control identity. This process is then repeated for half of the identities in the sample size, the other half are used as control identities. This histogram shows that images of the same identity consistently score a higher similarity. Interestingly, there is still an overlap between scores of same and different identities. Namely,image pairs of the same identity have low scores. These false negatives have several reasons, the most prominent being mislabeled identities in the data set, a large amount of time having passed between both pictures, and the face being generally obstructed. Examples of such false negatives are given in Figure 5.5.

[2] https://github.com/deepinsight/insightface/wiki/Model-Zoo

Clustering the vectors into identities is a challenging task, as there are a large number of clustering algorithms, each with their own advantages and disadvantages. Most clustering algorithms require a similarity or distance measurement and the number of final clusters. The cosine similarity is an appropriate choice as a similarity measurement, as we anticipated similar scores for the same identities. As the number of resulting clusters is unknown, at least without using ground truth information, hierarchical clustering is a well-suited approach to the clustering problem: The most appropriate hierarchical clustering method is agglomerative clustering. In this method, each object starts as its own cluster. These initial clusters are then successively merged, resulting in a dendogram consisting of leaves and nodes that can be cut at any desired level. The objects of this cut represent the clusters themselves (Rokach and Maimon, 2005), allowing us to label each cluster as an identity. One key advantage of this method is that we can first obtain a complete tree based on the cosine similarity and maximum linkage as a linkage criterion before finally determining the number of clusters. Evaluating the resulting clusters is also a challenging task: In this thesis we consider both the purity and the normalized mutual information score to gain an idea how accurate the resulting face clusters are.



(a) Identity 56, Image 163136.jpg

(b) Identity 56, Image 177638.jpg

(c) Identity 699, Image 163949.jpg

(d) Identity 699, Image 165678.jpg

(e) Identity 2277, Image 163095.jpg

(f) Identity 2277, Image 171085.jpg

Figure 5.5: SAME IDENTITY, LOW SIMILARITY. Some handpicked examples where images of the same ground truth identity have considerably low cosine similarity. The two most common reasons for this disparity are either a large amount of change of the face from a person or mistakes in the labeling of the identity of a person. We can see that in the image pair Subfigures (a) and (b) and in the pair Subfigures (e) and (f) that despite having the same ground truth identity labels, the two images clearly belong to different identities. In Subfigures (c) and (d) the same person is identified as two different identities, most likely due to the age difference of the individuals in the images.

Chapter 6

Experiments

6.1 Data Set and Neural Networks

In a first step, we extract the attributes of the validation and training partition with both AFFACT-B and AFFACT-U. This allows us to calculate the error rates of each attribute. As already mentioned, we focus on the balanced error rate ER_b for AFFACT-B and the unbalanced error rate ER_u for AFFACT-U, as described in Equation 4.4 and Equation 4.3. It is apparent that certain attributes are being classified substantially more accurate than others: For the attribute "Male", AFFACT-B achieves an ER_b of only 1.62%, AFFACT-U an ER_u of 1.45%. This is a significant difference compared with other attributes such as "Big Lips" (AFFACT-B ER_b = 31.35%, AFFACT-U ER_u = 27.17%) or "Pointy Nose" (AFFACT-B ER_b = 27.62%, AFFACT-U ER_u = 22.29%).

Furthermore, we can see the importance of evaluating both the AFFACT CNNs with their corresponding error rate: While AFFACT-B achieves similar scores with both metrics, AFFACT-U has a considerably worse balanced error rate. This is an effect that we anticipated, as this CNN completely neglects to consider class imbalance and as a result often fails to correctly assign the minority class. This increases the number of false negatives for positive minority classes and the number of false positives for negative minority classes considerably, leading to a high balanced error rate even though the unbalanced error rate is low.

6.2 ArcFace Identities

In order to automatically cluster faces into identities, we first use ArcFace to extract a feature vector of facial information from the images and then use this facial information to cluster the images. Our goal is to get clusters of faces that correspond to the identity. The method we use is agglomerative hierarchical clustering with the cosine similarity as a distance metric. As a linkage criterion we use complete linkage, which is also known as maximum linkage. This means that the distance between two clusters is "considered equal to the longest distance from any member of one cluster to any member of the other cluster" (Rokach and Maimon, 2005). This approach typically results in more compact clusters and more useful hierarchies compared with other linkage criteria, such as single-linkage (Rokach and Maimon, 2005). However, agglomerative clustering requires more similar pairs to have a lower distance, but the cosine similarity results in a higher score for more similar pairs. Consequently, this also has an impact on maximum-linkage as a linkage criterion. It is difficult to evaluate the impact of this factor on our clustering, but since the implementation[1] we use only has a limited number of linkage criteria, maximum-linkage still is

[1] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

an appropriate choice for our approach.

In a first step we calculate the entire clustering tree, but a key challenge remains: Cutting our tree at the right level, i.e. determining the number of clusters. Two commently used approaches to this issue are the elbow method and the silhouette score (Rousseeuw, 1987). We were unable to determine an exact number of clusters using the elbow method[2], leaving us with the silhouette score. This method indicates how separated and cohesive clusters are. Cohesiveness describes how similar objects are to other objects in the same cluster, while separation describes an object's dissimilarity to other objects in different clusters. This results in a score between -1 and 1 for each cluster, a higher value indicates more cohesiveness and better separation of that cluster (Rousseeuw, 1987). As the number of expected clusters is rather large, manual optimizations regarding the number of clusters are infeasible. For this reason, we simply compute the average silhouette score for each number of clusters and then choose the number of clusters where this score is at its maximum. With this method we calculate the number of clusters to be 1,240 for the validation set and 1,142 clusters for the test set. The number of hand labeled ground truth identities for both sets are 987 and 1,000 respectively. In both cases we overestimate the number of identities, however, this helps us to achieve a high purity. Purity is a measure of how uniform a cluster is regarding to its classes: In our case, we calculate the purity of a single cluster by assigning the cluster to the most frequent ground truth identity in that cluster before counting how many objects in that cluster are of that same identity. We repeat this process for each cluster and sum up the number of correctly assigned objects and divide it by the total number of objects (Schütze et al., 2008). For our purpose we want to achieve high purity, since we use the clustered identities to perform identity-related reweighting of attributes. This is why we try to avoid scenarios where there are multiple different ground truth identities in the same clusters, as this would mean that we consider attributes from different identities too. However, it should be noted that a high purity can be achieved by vastly overestimating the number of clusters, e.g. if the number of clusters is equal to the number of objects, we achieve a purity of 100%. With our silhouette score based agglomerative clustering approach we achieve a purity of 97.57% for the test set and 98.81% for the validation partition. Purity only measures how uniform the clusters are, but disregards clustering errors where one identity has been split into multiple clusters. A second measurement is the normalized mutual information (NMI) score. Mutual information is a quantification of how much information is shared between two variables: In our case the clustered labels and the ground truth labels. This score shows how much information we gain of one label by knowing about the other (Schütze et al., 2008). To calculate this score, we use a scikit-learn module that provides us with the NMI score based on the arithemtic average.[3] The NMI scores for the training partition and validation partition are 98.88% and 99.02% respectively.

6.3 Reweighting

6.3.1 Approaches

In a first step, we apply different reweighting formulas to the validation set in order to narrow the selection down to a number of candidates without introducing possible biases by optimizing them on the test partition. The four most successful approaches on the validation partition are described in Equation 5.8: $w_{square\ mean}$, $w_{square\ sign}$, $w_{cube\ sign}$, $w_{cube\ mean}$. These functions have two components: An exponent that lowers the term for unstable attributes, i.e. attributes with a low standard deviation and a factor that corrects the term towards the mean.

[2] https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html

When we compare these two methods for reweighting attributes with the same stability factor, the mean method results in a higher term if the mean has a higher absolute value and a lower term if the absolute value of the mean is lower. Alternative approaches involve sigmoid and cosine functions to model the desired effect for unstable attributes. Two examples can be found in Appendix A.1. Still, the most successful results on the validation sets stem from one of the four correction term formulas.

6.3.2 Reweighting with Ground Truth Identities

In a first step, we simply use the hand-labeled identities to calculate the correction terms for each attribute of an identity. This means that we calculate the σ and μ of each attribute a for an identity i, giving us σ_{ia} and μ_{ia} to which we apply all four reweighting approaches w from Equation 5.8. For the balanced network AFFACT-B, we adapt the correction terms to our source distribution as described in Equation 5.5, resulting in balanced correction terms. To gain meaningful insights, we reweight the extractions from the test partition. The goal is to lower the balanced error rate ER_b for AFFACT-B and the unbalanced error rate ER_u for AFFACT-U compared with the error rates of the unperturbed, original extraction. The results in Table 6.1 show that all approaches result in lower error rates than the uncorrected extractions. Interestingly, the best performing correction term formulas change depending on the CNN we use: For the balanced network AFFACT-B $w_{\textit{square sign}}$ with domain adaption lowers the average balanced error rate by 0.23%, a relative improvement of 2%. The average unbalanced error rate of the unbalanced network AFFACT-U is lowered by the correction term function $w_{\textit{square mean}}$ by 0.1%, a relative improvement of 1.2%.

Table 6.1: Average Error Rates Using Ground Truth. Average error rates after and before reweighting the extractions of both CNNs using the ground truth identity information.

	AFFACT-B		AFFACT-U	
	Average ER_b	Average ER_u	Average ER_b	$Average ER_u$
uncorrected	11.46	12.16	17.16	8.26
square sign	**11.23**	12.90	18.94	8.24
square mean	11.28	12.33	18.15	**8.16**
cube sign	11.24	12.51	18.24	8.18
cube mean	11.33	12.25	17.79	8.17

6.3.3 Reweighting with Clustered Identities

Finally, we use the information from the facial clustering instead of the manually annotated identities. The rest of our approach remains the same, but the reweighting function now uses σ and μ that we calculated for each attribute of each cluster, and then applies the resulting correction term to the other object of that cluster. The results in Table 6.2 are nearly identical with those from the reweighting with ground truth information. For the AFFACT-B, the correction term function $w_{\textit{square sign}}$ performs slightly worse, lowering the average balanced error rate by 0.21%, a relative improvement of 1.8% on par with $w_{\textit{cube sign}}$. The results for AFFACT-U remain essentially the same, $w_{\textit{square mean}}$ still performs the best by lowering the average unbalanced error by 0.1%, a relative improvement of 1.2%.

Table 6.2: AVERAGE ERROR RATES USING CLUSTERED LABELS. Average error rates after and before reweighting the extractions of both CNNs using the clustering of ArcFace extractions. The bold text denote the best performance of a formula for the appropriate error rate.

	AFFACT-B		AFFACT-U	
	Average ER_b	Average ER_u	Average ER_b	Average ER_u
unweighted	11.46	12.16	17.16	8.26
$w_{\text{square sign}}$	**11.25**	12.90	18.94	8.24
$w_{\text{square mean}}$	11.29	12.33	18.15	**8.16**
$w_{\text{cube sign}}$	**11.25**	12.51	18.24	8.18
$w_{\text{cube mean}}$	12.51	12.25	17.78	8.17

6.3.4 Balancing of Correction Terms

For the calculation of the average error rates, we only use balanced correction terms, i.e. terms that consider how classes are distributed in the domain. We achieved this by adapting the statistical measurements to the class distribution with σ_{ia}^{B} and μ_{ia}^{B}. In the experiments above, we used these two domain-adapted measurements for AFFACT-B and the unadapted measurements σ_{ia}^{U} and μ_{ia}^{U} for AFFACT-U. The reason why we only use the balanced correction terms that are calculated with σ_{ia}^{B} and μ_{ia}^{B} for the extractions of AFFACT-B can be seen in Table 6.3:

When we compare the performance of balanced correction terms and unbalanced ones, we can see that using σ_{ia}^{B} and μ_{ia}^{B} nearly always results in a lower ER_b. Interestingly, this works for extractions of both AFFACT-B and AFFACT-U, showing that our balanced correction terms are effective at reducing the balanced error rate. The same observation can be made for unbalanced correction terms. Using unbalanced correction terms results in a lower ER_u for extractions from both CNNs.

Table 6.3: BALANCED VS. UNBALANCED CORRECTION TERMS. Comparison of average error rates between balanced correction terms that use σ_{ia}^{B} and μ_{ia}^{B} and the unbalanced correction terms that use σ_{ia}^{U} and μ_{ia}^{U} with reweighting based on ground truth identities. The lower error rate is marked in bold.

	Average ER_b		Average ER_u	
	balanced correction term	unbalanced correction term	balanced correction term	unbalanced correction term
AFFACT-B				
$w_{\text{square sign}}$	**11.23**	11.38	12.90	**11.36**
$w_{\text{square mean}}$	11.28	11.28	12.33	**11.63**
$w_{\text{cube sign}}$	**11.24**	11.28	12.51	**11.59**
$w_{\text{cube mean}}$	11.33	**11.31**	12.25	**11.80**
AFFACT-U				
$w_{\text{square sign}}$	**16.53**	18.94	8.24	8.24
$w_{\text{square mean}}$	**16.89**	18.15	8.20	**8.16**
$w_{\text{cube sign}}$	**16.79**	18.24	8.22	**8.18**
$w_{\text{cube mean}}$	**17.00**	17.79	8.21	**8.17**

Chapter 7

Discussion

A central finding of these experiments is that we were not only able to lower the error rates with ground truth identity labels, but also by using labels from clustered face identities. In our experiments, the number of clustered identities is larger than the number of ground truth identities. This over estimation is most likely a solid approach to automatically clustered identity labels, as it is quite challenging to accurately determine the ground truth information from visual information only, since facial appearance can vary greatly over time. This visual based approach to identity may be better for the purpose of reweighting identity-related attributes, especially in cases where images of the same identity exhibit different features over time. There are various reasons for large changes of facial features such as aging, facial surgery, or gender-reassignment surgery. A comparison of the relative improvements of AFFACT-U can be seen in Table ??. The difference in the performance of ground truth and clustered identity labels is insignificant. To compare the performance between clustered and ground truth labels for AFFACT-B, we must consider both the false negative and false positive rate due to the balanced approach of this network. As can be seen in Figure 7.1, the false positive and false negative rates based on which we calculate the balanced error rate are also similar after reweighting with ground truth labels and reweighting with clustered labels.

While our clustering method is by not perfect, as indicated by the differences in error rates between reweighting using ground truth and clustered identities, we receive high values for both purity and NMI. The clusters of the test set possess a purity of 97.57%. Ideally, we want this value to be as high as possible, as mixing identities together defeats the purpose of reweighting attributes based on their correlation with an identity. However, simply focusing on maximizing purity is not ideal either: The purity of a cluster tends to increase the smaller the cluster becomes. However, the smaller these clusters become, the fewer images are available to calculate the correction terms.

This is why we also introduced NMI as a measurement to evaluate our clustering. This measurement considers the mutual information shared between ground truth labels and clustered labels, which means it also considers the effect of one identity being split into multiple clusters. The NMI score of the test set is 98.88%, a relatively high score, which indicates that the clustering in our experiments is fairly accurate. The task of accurate identity clustering is also hindered by incorrect labels of the ground truth identities. It is difficult to estimate how large the impact of such incorrect labels is, but they certainly influence the evaluation of our clustering method. Generally, the performance of reweighting with ground truth identity labels and with clustered identity labels is very similar. Especially the resulting error rates after reweighting extractions of AFFACT-U are basically identical for both labels. In AFFACT-B we can observe a slight decrease in performance when reweighting with the clustered labels.

Not only is determining the number of identities difficult, but also modelling the effect of identity-related stability: One of the most challenging aspects of the approach introduced in this

	uncorrected	ground truth labels		clustered labels	
	ER_u	ER_u	change	ER_u	change
5 o Clock Shadow	5.23	5.11	2.29%	5.11	2.29%
Arched Eyebrows	15.73	15.39	2.16%	15.39	2.16%
Attractive	17.02	16.76	1.53%	16.76	1.53%
Bags Under Eyes	14.76	**14.34**	**2.85%**	14.36	2.71%
Bald	0.95	0.93	2.11%	0.93	2.11%
Bangs	3.83	**3.83**	**0.00%**	3.86	-0.78%
Big Lips	27.17	**27.28**	**-0.40%**	27.30	-0.48%
Big Nose	15.67	14.90	4.91%	14.90	4.91%
Black Hair	9.57	9.44	1.36%	**9.43**	**1.46%**
Blond Hair	3.93	3.91	0.51%	3.91	0.51%
Blurry	3.61	**3.77**	**-4.43%**	3.78	-4.71%
Brown Hair	10.59	10.25	3.21%	**10.24**	**3.31%**
Bushy Eyebrows	7.14	**6.94**	**2.80%**	6.96	2.52%
Chubby	4.35	4.22	2.99%	4.22	2.99%
Double Chin	3.45	3.38	2.03%	3.38	2.03%
Eyeglasses	0.36	0.34	5.56%	**0.32**	**11.11%**
Goatee	2.41	2.32	3.73%	2.32	3.73%
Gray Hair	1.74	1.71	1.72%	1.71	1.72%
Heavy Makeup	8.16	8.06	1.23%	**8.03**	**1.59%**
High Cheekbones	11.86	11.79	0.59%	**11.78**	**0.67%**
Male	1.45	**1.29**	**11.03%**	1.31	9.66%
Mouth Slightly Open	5.90	**5.86**	**0.68%**	5.89	0.17%
Mustache	2.87	**2.71**	**5.57%**	2.72	5.23%
Narrow Eyes	12.21	12.62	-3.36%	**12.59**	**-3.11%**
No Beard	3.54	3.57	-0.85%	**3.56**	**-0.56%**
Oval Face	23.11	**22.92**	**0.82%**	22.97	0.61%
Pale Skin	2.82	**2.91**	**-3.19%**	2.92	-3.55%
Pointy Nose	22.29	22.32	-0.13%	**22.27**	**0.09%**
Receding Hairline	5.96	5.95	0.17%	5.95	0.17%
Rosy Cheeks	4.77	4.71	1.26%	4.71	1.26%
Sideburns	2.24	2.18	2.68%	**2.17**	**3.13%**
Smiling	6.83	6.72	1.61%	**6.71**	**1.76%**
Straight Hair	14.55	**14.59**	**-0.27%**	14.60	-0.34%
Wavy Hair	13.57	13.55	0.15%	**13.54**	**0.22%**
Wearing Earrings	9.17	**9.05**	**1.31%**	9.06	1.20%
Wearing Hat	0.81	**0.81**	**0.00%**	0.82	-1.23%
Wearing Lipstick	6.12	5.81	5.07%	**5.78**	**5.56%**
Wearing Necklace	10.71	10.69	0.19%	**10.67**	**0.37%**
Wearing Necktie	2.70	2.68	0.74%	2.68	0.74%
Young	11.06	10.63	3.89%	10.63	3.89%
Average	8.26	8.16	1.21%	8.16	1.21%

s

thesis is the calculation of the correction terms that are added to the extractions. The four methods of calculation that were introduced are all similar to each other: In all methods a low standard deviation results in a significantly lower absolute value of the correction term, due to the polynomial nature of our formulas. This part of the formula represents the stability of an attribute from an identity. Attributes with a low stability are corrected to a smaller degree; attributes with a standard deviation higher than 1 are not corrected at all. A reason for this lowered consideration of unstable attributes is to avoid changing correctly identified attributes as much as possible. However, a considerable weakness of our approach is that the standard deviation does not give us any information about the binary classes it stems from: The standard deviation from which we infer stability might be the same for an attribute that has values that are all below 0 as one that has half its values below 0 and the other half above, even though the former one is stable from a binary attribute classification view, while the latter is not. This effect is somewhat counteracted by the second part of our correction term calculation formula which also considers the mean. Still, most likely there is room improvement for representing stability in such formulas. Interestingly, how we consider the mean results in different best performing reweighting formulas: For the balanced extractions we receive better results if the correction terms only take the sign of the mean, while the unbalanced one corrects it directly towards the mean. Furthermore, there most likely exist formulas that model the desired effect more accurately, however, the four formulas shown in this thesis show a first, simple and straightforward approach for calculating such terms.

All in all, there are considerable differences between the results after reweighting the extractions from our two CNNs. An example of this is the attribute "Big Lips", which has a better performance after reweighting extractions from AFFACT-B but a worse performance after reweighting extractions from AFFACT-U. On one hand, this is of course due to the different extraction approach of these two neural networks, on the other hand it is also an artifact from our two different approaches to reweighting: The correction terms for the extractions from AFFACT-B are domain-adapted, furthermore, the best-performing reweighting formulas are two different ones for the balanced and unbalanced approaches. We also show that by adapting our correction terms to the domain we receive a lower average ER_b for both AFFACT-B and AFFACT-U. If we do not adapt these correction terms, we receive a lower average ER_u. As we focus on lowering ER_b for AFFACT-B and ER_u for AFFACT-U, we only use domain-adapted correction terms for AFFACT-B. This difference in approaches makes comparisons between the reweighting of both neural networks difficult, even when comparing the same apparent reweighting formulas, as we use μ_{ia}^{B} σ_{ia}^{B} for extractions of AFFACT-B and μ_{ia}^{U} σ_{ia}^{U} for extractions of AFFACT-U.

The approach shown in this thesis uses a very pragmatic understanding of identity and the attributes that relate to it. This is one of its biggest strengths: Rather than solely basing the reweighting of attributes on the human understanding of attributes and their consistency, we rely on statistical measurements of each identity to perform corrections. This allows for a more individualistic approach rather than a global one, as different attributes may vary greatly in their stability for different individuals. However, this approach also leads to certain issues, as the stability of attributes is always interpreted as a correlation with identity. This interpretation is not true for all attributes, a great example of this effect leading to a worse performance can be seen with the attributes "Blurry". We can seen this by looking at the increased false positive rate in Figure 7.1 of the balanced network AFFACT-B and the increased ER_u of AFFACT-U in Table ??. From a human understanding it is apparent that "Blurry" is a property of the image, rather than an identity-related facial attribute. However, since our approach only considers in how many pictures of the same identity an attribute is present or absent, we incorrectly assume that this attribute is stable. The attribute "Pale Skin" also increases for reweighting of both CNNs. Examples of other increases in error rates can be observed for the attributes like "Big Lips", "Blurry", and "Narrow Eyes" for AFFACT-U and attributes like "Arched Eyebrows", "Chubby", and "Double Chin" for AFFACT-B. Many of these attributes for which the reweighting results in a higher error rate than before, are attributes that are heavily imbalanced. This indicates that, despite our efforts to consider class imbalance, the class distribution of the samples still is an obstacle for our reweighting process.

From a purely statistical point of view, it is difficult to determine why these attributes show a low standard deviation, but should not be reweighted. This task is even more difficult due to labeling issues, as some attributes are of a very subjective nature. Examples for this are attributes such as "Big Lips" or "Narrow Eyes". We expect such attributes to be highly related to the identity, but labeling such attributes is a challenging task. This means that such attributes introduce an inherent bias through whoever is labeling them and in which cases they consider a person to possess these attributes.

A further issue is the available sample size for each identity. It is increasingly difficult to determine whether the stability of an attribute is truly related to the identity if the number of images of that identity is small. Our formulas are well defined for identities that occur in one image, as the standard deviation is 0 for all attributes. In such cases, we simply correct the extracted values toward the mean of the attributes of that identity. If that identity only has one image, then the reweighting has no effect on the error rate, due to the binary nature of our evaluation. However, if the number of images of an identity is small but greater than 1, then our statistical measures are substantially less meaningful due to the small sample size.

A global consideration of the correction terms might help to solve this problem. We could differentiate between attributes that globally correlate more often and such that correlate less often, allowing us to decrease the correction terms of attributes that show a globally low correlation with their identity. Such an approach should also differentiate between the correlation of the absence of an attribute and its presence, as these two correlations may be substantially different for some attributes. This would allow us exploit information of cross-correlating attributes and subsequently result in a more robust reweighting approach. Another possibility would be to combine our statistical approach with attribute grouping, a method that is widespread in related works for to improve the performance of CNNs. However, manual attribute grouping is most likely not an appropriate choice to calculate correction terms for attributes, as there are disparities between the human understanding of how certain facial attributes correlate with identity and how they actually correlate in such a data set.

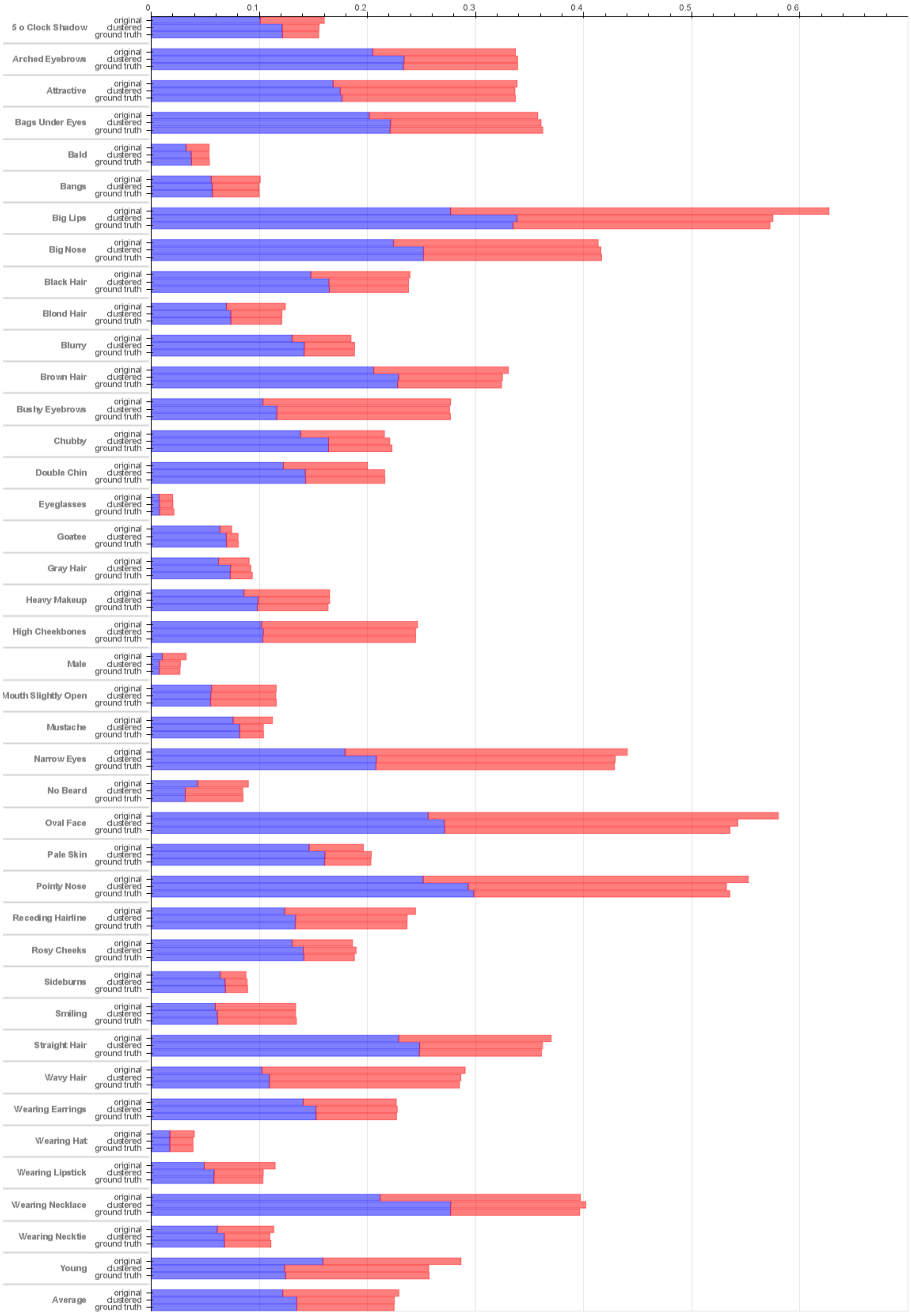


Figure 7.1: Balanced and Unbalanced Error Rates AFFACT-B. Each attribute has three bars. The stacked bar *original* shows the number of mistakes of the unweighted extractions. The stacked bars *clustered* and *ground truth* show the error rates after being reweighted using the respective identity labels. Both were reweighted with the correction terms calculated from $w_{square\ sign}$. The blue bars show the false positive rate, while the red bars indicate the false negative rate. The balanced error rate is the average of these two rates.

Chapter 8

Conclusion

The results of our experiments show that it is indeed possible to improve the accuracy of CNN extractions using identity-related attribute correlation. This thesis introduced a method for calculating correction terms using simple statistical measurements such as the standard deviation and mean from each identity. While there is room for improvement considering the exact formulas for these correction terms, the terms we calculate are able to effectively reweight extractions from CNNs. A major issue for CNNs is the class imbalance in the data set. We address this issue by using extractions from both an unbalanced and a balanced neural neural network: Both these networks are based on the approach introduced by Günther et al. (2017). Our approach to this issue is to modify correction terms by adapting them to the distribution of classes, resulting in balanced correction terms, which improve the balanced accuracy rate of the balanced CNN. However, while we are able to lower the overall error rates for both networks, we fail to improve the classification accuracy for every attribute. The main reason for this is that attributes may appear stable, which we then interpret as correlating with identity, though this interpretation is incorrect for some cases, leading to a lowered performance. While we are able to successfully consider the issue of class imbalance in our reweighting approaches to a certain degree by utilizing statistical measures that are adapted to a target domain, class imbalance may very well be the reason why some attributes have higher error rates after reweighting.

Furthermore, we also show that this improvement can be achieved to a similar degree when using automatically detected identities, rather than just ground truth ones. In this thesis the identities are determined by agglomerative clustering of facial information that was extracted using ArcFace (Deng et al., 2019). We hypothesize that it is even possible to achieve a higher performance with automatically extracted identity labels than with ground truth ones, due to a higher correlation between certain attributes and visual identity compared with the ground truth one.

There is potential to further improve our approach. A major possibility for improvement is the specific calculation of our correction terms. While we show that we can already achieve improvements with basic formulas, there most likely is room for improvement by utilizing formulas that model our desired effect more accurately. Currently the reweighting approach only considers attributes correlation with identity. It could be enhanced by not only considering how a single attribute correlates with identity but also how attributes correlate with each other with respect to an identity. Such an approach, though more complex, could significantly enhance performance. Furthermore, our current approach could benefit by some sort of global consideration to reduce the effect of reweighting apparently stable but identity unrelated attributes. Such incorrect reweighting is a major obstacle for our stability-based approach, as we cannot infer whether an attribute is identity-related or not from the standard deviation and mean especially if there are only few sample images for our calculations.

Our results show that reweighting attributes based on their correlation to identity is possible.

While the results of this thesis may seem minor, since we are only able to improve the average error rates by a few tenths of a percent, it offers a basic approach that was tested successfully on extractions of two neural networks, lowering the error rates for both. This basic approach has shown significant of potential, and we hope to refine and improve it in future research.

Appendix A

Attachments

A.1 Further Approaches to Reweighting

In addition to the calculation formulas for the correction terms described in Equation 5.8, we tested several different formulas. As most of these formulas were rather unsuccessful, we did not describe them in detail. Here we outline two examples of formulas that were able to lower the average balanced error rate of extractions from AFFACT-B:

$$\begin{aligned} w^{ia}_{\text{sigmoid}} &= \begin{cases} 0 & \text{if } \sigma_{ia} > 1 \\ \frac{\operatorname{sgn}(\mu_{\text{ia}})}{1-e^{-10(0.5-\sigma_{ia})}} & \text{otherwise} \end{cases} \\ w^{ia}_{\text{cosine}} &= \begin{cases} 0 & \text{if } \sigma_{ia} > 1 \\ \frac{\cos((1-\sigma_{ia})\cdot\Pi)+1}{2} \cdot \operatorname{sgn}(\mu_{\text{ia}}) & \text{otherwise} \end{cases} \end{aligned} \tag{A.1}$$

These were not included in the main text, as their performance was worse than that of the four correction term formulas using polynomials. However, this shows that there are a large number of approaches for calculating correction terms. Some of these formulas most likely lead to a better performance than the formulas used in this thesis. The complete results of these two formulas can be found in Table A.2.

A.2 Additional Tables

Table A.1: SIGMOID AND COSINE. This table shows the performance off AFFACT-B before and after reweighting with $\mathrm{w}_{\mathrm{sigmoid}} and w_{\mathrm{cosine}}.AFFACT - BisthebalancedCNN, making\mathrm{ER}_b$ the relevant error rate. The averages of the relevant error rate are marked in bold.

	no reweight		w_{sigmoid}		w_{cosine}	
Error rates	ER_b	ER_u	ER_b	ER_u	ER_b	ER_u
5 o Clock Shadow	8.00	9.67	7.98	12.48	8.07	12.63
Arched Eyebrows	16.84	18.41	17.49	20.89	17.52	21.04
Attractive	16.92	16.92	17.43	17.44	17.41	17.43
Bags Under Eyes	17.88	19.25	18.41	21.37	18.71	21.58
Bald	2.68	3.21	2.85	3.99	2.92	4.12
Bangs	5.04	5.40	5.03	5.50	5.00	5.48
Big Lips	31.35	30.08	28.33	30.87	28.30	30.74
Big Nose	20.67	21.68	21.10	24.68	21.06	24.54
Black Hair	11.97	13.25	12.17	14.72	12.14	14.83
Blond Hair	6.20	6.76	6.03	7.22	6.04	7.29
Blurry	9.24	12.65	9.71	14.25	9.78	14.75
Brown Hair	16.52	19.12	16.56	21.54	16.72	21.82
Bushy Eyebrows	13.85	11.25	14.26	13.26	14.17	13.45
Chubby	10.79	13.50	11.03	17.18	11.08	17.26
Double Chin	10.00	12.02	11.12	15.14	11.27	15.44
Eyeglasses	0.99	0.77	1.07	0.86	1.08	0.87
Goatee	3.72	6.11	4.07	6.97	4.03	7.00
Gray Hair	4.53	6.12	4.99	7.89	5.06	8.03
Heavy Makeup	8.25	8.32	8.37	8.79	8.36	8.85
High Cheekbones	12.32	12.24	12.39	12.32	12.48	12.41
Male	1.62	1.48	1.21	1.08	1.18	1.05
Mouth Slightly Open	5.78	5.78	5.85	5.85	5.88	5.88
Mustache	5.61	7.43	5.20	8.32	5.25	8.55
Narrow Eyes	22.02	19.14	21.67	22.57	21.71	23.40
No Beard	4.49	4.64	4.29	5.29	4.33	5.46
Oval Face	29.00	27.62	26.98	28.15	26.91	28.10
Pale Skin	9.81	14.21	10.28	16.42	10.48	17.57
Pointy Nose	27.62	26.57	27.43	29.73	27.29	29.49
Receding Hairline	12.23	12.33	11.81	13.72	11.69	13.89
Rosy Cheeks	9.31	12.50	9.70	13.95	9.90	14.50
Sideburns	4.38	6.20	4.52	6.95	4.59	7.08
Smiling	6.68	6.68	6.79	6.79	6.77	6.77
Straight Hair	18.50	21.05	18.49	23.04	18.73	23.33
Wavy Hair	14.52	13.36	14.20	13.54	14.26	13.67
Wearing Earrings	11.33	12.93	11.54	14.19	11.70	14.53
Wearing Hat	2.00	1.75	1.94	1.74	1.93	1.74
Wearing Lipstick	5.73	5.77	5.22	5.15	5.15	5.08
Wearing Necklace	19.85	20.81	20.70	29.25	20.90	30.07
Wearing Necktie	5.67	6.06	5.61	6.98	5.50	7.21
Young	14.33	13.53	12.73	13.50	12.67	13.37
Average	**11.46**	12.16	**11.41**	13.59	**11.45**	13.76

Table A.2: Error Rates of CNN Extractions. This table lists both error rates for extractions of both CNNs. AFFACT-B is the balanced CNN, making ER_b the relevant error rate while ER_u is relevant for the extractions of AFFACT-U.

	AFFACT-B		AFFACT-U	
Error rates	ER_b	ER_u	ER_b	ER_u
5 o Clock Shadow	8.00	9.67	13.59	5.23
Arched Eyebrows	16.84	18.41	19.91	15.73
Attractive	16.92	16.92	17.03	17.02
Bags Under Eyes	17.88	19.25	23.07	14.76
Bald	2.68	3.21	10.55	0.95
Bangs	5.04	5.40	7.59	3.83
Big Lips	31.35	30.08	37.58	27.17
Big Nose	20.67	21.68	24.29	15.67
Black Hair	11.97	13.25	13.15	9.57
Blond Hair	6.20	6.76	9.63	3.93
Blurry	9.24	12.65	26.41	3.61
Brown Hair	16.52	19.12	17.05	10.59
Bushy Eyebrows	13.85	11.25	19.72	7.14
Chubby	10.79	13.50	23.62	4.35
Double Chin	10.00	12.02	25.53	3.45
Eyeglasses	0.99	0.77	1.38	0.36
Goatee	3.72	6.11	10.11	2.41
Gray Hair	4.53	6.12	12.68	1.74
Heavy Makeup	8.25	8.32	8.99	8.16
High Cheekbones	12.32	12.24	11.93	11.86
Male	1.62	1.48	1.59	1.45
Mouth Slightly Open	5.78	5.78	5.90	5.90
Mustache	5.61	7.43	25.48	2.87
Narrow Eyes	22.02	19.14	36.05	12.21
No Beard	4.49	4.64	6.22	3.54
Oval Face	29.00	27.62	35.16	23.11
Pale Skin	9.81	14.21	25.09	2.82
Pointy Nose	27.62	26.57	34.01	22.29
Receding Hairline	12.23	12.33	23.85	5.96
Rosy Cheeks	9.31	12.50	20.74	4.77
Sideburns	4.38	6.20	8.27	2.24
Smiling	6.68	6.68	6.83	6.83
Straight Hair	18.50	21.05	25.33	14.55
Wavy Hair	14.52	13.36	16.31	13.57
Wearing Earrings	11.33	12.93	13.58	9.17
Wearing Hat	2.00	1.75	4.70	0.81
Wearing Lipstick	5.73	5.77	6.03	6.12
Wearing Necklace	19.85	20.81	29.71	10.71
Wearing Necktie	5.67	6.06	10.14	2.70
Young	14.33	13.53	17.64	11.06
Average	11.46	12.16	17.16	8.26

Table A.3: CALCULATED PROBABILITY. This probability was calculated using the formula shown in 5.3 and outlined by Rudd et al. (2016).

Attributes	*positive*	*negative*
5 o Clock Shadow	1.0	0.13
Arched Eyebrows	1.0	0.36
Attractive	0.95	1.0
Bags Under Eyes	1.0	0.26
Bald	1.0	0.02
Bangs	1.0	0.18
Big Lips	1.0	0.32
Big Nose	1.0	0.31
Black Hair	1.0	0.31
Blond Hair	1.0	0.18
Blurry	1.0	0.05
Brown Hair	1.0	0.26
Bushy Eyebrows	1.0	0.17
Chubby	1.0	0.06
Double Chin	1.0	0.05
Eyeglasses	1.0	0.07
Goatee	1.0	0.07
Gray Hair	1.0	0.04
Heavy Makeup	1.0	0.62
High Cheekbones	1.0	0.83
Male	1.0	0.72
Mouth Slightly Open	1.0	0.93
Mustache	1.0	0.04
Narrow Eyes	1.0	0.13
No Beard	0.2	1.0
Oval Face	1.0	0.4
Pale Skin	1.0	0.04
Pointy Nose	1.0	0.38
Receding Hairline	1.0	0.09
Rosy Cheeks	1.0	0.07
Sideburns	1.0	0.06
Smiling	1.0	0.92
Straight Hair	1.0	0.26
Wavy Hair	1.0	0.47
Wearing Earrings	1.0	0.23
Wearing Hat	1.0	0.05
Wearing Lipstick	1.0	0.89
Wearing Necklace	1.0	0.14
Wearing Necktie	1.0	0.08
Young	0.28	1.0

Table A.4: ERROR RATES AFFACT-B WITH $w_{\text{SQUARE SIGN}}$. This table compares the error rates of the uncorrected extractions to those of extractions that are reweighted with $w_{\text{square sign}}$. The column *ground truth labels* shows the results when ground truth identity labels were used for reweighting and *clustered labels* when clustered labels were used. The lowest ER_bs are marked in bold.

	uncorrected		ground truth lables		clustered labels	
Error rates	ER_b	ER_u	ER_b	ER_u	ER_b	ER_u
5 o Clock Shadow	8.00	9.67	**7.75**	11.26	7.76	11.25
Arched Eyebrows	**16.84**	18.41	16.95	19.70	16.94	19.72
Attractive	16.92	16.92	16.84	16.85	**16.82**	16.83
Bags Under Eyes	**17.88**	19.25	18.11	20.48	18.03	20.47
Bald	**2.68**	3.21	2.69	3.69	**2.68**	3.66
Bangs	5.04	5.40	**5.00**	5.45	**5.00**	5.45
Big Lips	31.35	30.08	**28.62**	30.30	28.75	30.51
Big Nose	**20.67**	21.68	20.83	23.34	20.80	23.33
Black Hair	11.97	13.25	**11.90**	13.99	**11.90**	13.97
Blond Hair	6.20	6.76	**6.04**	7.02	6.05	7.01
Blurry	**9.24**	12.65	9.40	13.67	9.41	13.68
Brown Hair	16.52	19.12	**16.20**	20.42	16.25	20.49
Bushy Eyebrows	13.85	11.25	13.84	12.20	**13.80**	12.19
Chubby	**10.79**	13.50	11.13	15.85	11.03	15.83
Double Chin	**10.00**	12.02	10.81	13.96	10.80	13.94
Eyeglasses	**0.99**	0.77	1.05	0.82	1.01	0.80
Goatee	**3.72**	6.11	4.03	6.70	4.02	6.67
Gray Hair	**4.53**	6.12	4.68	7.14	4.61	7.16
Heavy Makeup	8.25	8.32	**8.18**	8.49	8.25	8.56
High Cheekbones	12.32	12.24	**12.23**	12.16	**12.23**	12.16
Male	1.62	1.48	**1.33**	1.20	1.35	1.22
Mouth Slightly Open	5.78	5.78	5.79	5.79	**5.76**	5.76
Mustache	5.61	7.43	5.20	7.96	**5.19**	7.94
Narrow Eyes	22.02	19.14	**21.43**	20.97	21.46	21.01
No Beard	4.49	4.64	4.25	5.02	**4.24**	5.01
Oval Face	29.00	27.62	**26.77**	26.93	27.13	27.12
Pale Skin	**9.81**	14.21	10.16	15.54	10.18	15.57
Pointy Nose	27.62	26.57	26.76	28.08	**26.60**	27.77
Receding Hairline	12.23	12.33	**11.83**	13.08	11.84	13.09
Rosy Cheeks	**9.31**	12.50	9.40	13.44	9.48	13.41
Sideburns	**4.38**	6.20	4.46	6.64	4.44	6.61
Smiling	**6.68**	6.68	6.72	6.72	**6.68**	6.68
Straight Hair	18.50	21.05	**18.05**	21.97	18.10	22.01
Wavy Hair	14.52	13.36	**14.26**	13.36	14.32	13.40
Wearing Earrings	**11.33**	12.93	11.35	13.64	11.38	13.64
Wearing Hat	2.00	1.75	1.94	1.74	**1.93**	1.73
Wearing Lipstick	5.73	5.77	**5.17**	5.14	5.18	5.15
Wearing Necklace	19.85	20.81	**19.82**	25.52	20.10	25.60
Wearing Necktie	5.67	6.06	5.54	6.61	**5.49**	6.58
Young	14.33	13.53	12.85	13.06	**12.84**	13.09
Average	11.46	12.16	**11.23**	12.90	11.25	12.90

Table A.5: ERROR RATES AFFACT-U WITH $w_{\text{SQUARE MEAN}}$. This table compares the error rates of the uncorrected extractions to those of extractions that are reweighted with $w_{\text{square mean}}$. The column *ground truth labels* shows the results when ground truth identity labels were used for reweighting and *clustered labels* when clustered labels were used. The lowest ER_us are marked in bold.

	uncorrected		ground truth lables		clustered labels	
Error rates	ER_b	ER_u	ER_b	ER_u	ER_b	ER_u
5 o Clock Shadow	13.59	5.23	13.92	**5.11**	13.94	**5.11**
Arched Eyebrows	19.91	15.73	20.09	**15.39**	20.09	**15.39**
Attractive	17.03	17.02	16.76	**16.76**	16.76	**16.76**
Bags Under Eyes	23.07	14.76	24.45	**14.34**	24.50	14.36
Bald	10.55	0.95	11.11	0.93	11.00	0.93
Bangs	7.59	**3.83**	8.24	**3.83**	8.29	3.86
Big Lips	37.58	**27.17**	38.61	27.28	38.63	27.30
Big Nose	24.29	15.67	24.32	**14.90**	24.27	**14.90**
Black Hair	13.15	9.57	13.31	9.44	13.31	**9.43**
Blond Hair	9.63	3.93	9.89	**3.91**	9.87	**3.91**
Blurry	26.41	**3.61**	31.89	3.77	31.89	3.78
Brown Hair	17.05	10.59	18.08	10.25	18.03	10.24
Bushy Eyebrows	19.72	7.14	20.63	**6.94**	20.65	6.96
Chubby	23.62	4.35	24.80	**4.22**	24.76	**4.22**
Double Chin	25.53	3.45	27.22	**3.38**	27.37	**3.38**
Eyeglasses	1.38	0.36	1.52	0.34	1.40	**0.32**
Goatee	10.11	2.41	10.89	**2.32**	10.94	**2.32**
Gray Hair	12.68	1.74	13.66	**1.71**	13.58	**1.71**
Heavy Makeup	8.99	8.16	8.82	8.06	8.79	**8.03**
High Cheekbones	11.93	11.86	11.86	11.79	11.85	**11.78**
Male	1.59	1.45	1.43	**1.29**	1.45	1.31
Mouth Slightly Open	5.90	5.90	5.86	**5.86**	5.89	5.89
Mustache	25.48	2.87	27.21	**2.71**	27.34	2.72
Narrow Eyes	36.05	12.21	39.71	12.62	39.60	**12.59**
No Beard	6.22	**3.54**	6.30	3.57	6.30	3.56
Oval Face	35.16	23.11	36.20	**22.92**	36.25	22.97
Pale Skin	25.09	**2.82**	29.52	2.91	29.52	2.92
Pointy Nose	34.01	22.29	35.16	22.32	35.10	**22.27**
Receding Hairline	23.85	5.96	25.74	**5.95**	25.69	**5.95**
Rosy Cheeks	20.74	4.77	23.74	**4.71**	23.74	**4.71**
Sideburns	8.27	2.24	9.00	2.18	9.00	**2.17**
Smiling	6.83	6.83	6.72	6.72	6.71	**6.71**
Straight Hair	25.33	**14.55**	27.07	14.59	27.10	14.60
Wavy Hair	16.31	13.57	16.52	13.55	16.52	**13.54**
Wearing Earrings	13.58	9.17	13.95	**9.05**	13.97	9.06
Wearing Hat	4.70	**0.81**	5.55	**0.81**	5.56	0.82
Wearing Lipstick	6.03	6.12	5.76	5.81	5.73	**5.78**
Wearing Necklace	29.71	10.71	32.04	10.69	31.98	**10.67**
Wearing Necktie	10.14	2.70	11.19	**2.68**	11.19	**2.68**
Young	17.64	11.06	17.38	**10.63**	17.36	**10.63**
Average	17.16	8.26	18.15	**8.16**	18.15	**8.16**

List of Figures

List of Tables

Bibliography

Cao, J., Li, Y., and Zhang, Z. (2018). Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4290–4299.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Fang, Y. and Yuan, Q. (2018). Attribute-enhanced metric learning for face retrieval. *EURASIP Journal on Image and Video Processing*, 2018(1):44.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Günther, M., Rozsa, A., and Boult, T. E. (2017). Affact: Alignment-free facial attribute classification technique. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 90–99.

Han, H., Jain, A. K., Wang, F., Shan, S., and Chen, X. (2018). Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2597–2609.

Hand, E. and Chellappa, R. (2017). Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Huang, G., Mattar, M., Lee, H., and Learned-miller, E. (2012). Learning to align from scratch. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27.

Jolliffe, I. (2011). *Principal Component Analysis*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kalayeh, M. M., Gong, B., and Shah, M. (2017). Improving facial attribute prediction using semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4227–4235, Los Alamitos, CA, USA. IEEE Computer Society.

Kang, S., Lee, D., and Yoo, C. D. (2015). Face attribute classification using attribute-aware correlation map and gated convolutional neural networks. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4922–4926.

Kumar, N., Berg, A., Belhumeur, P., and Nayar, S. (2011). Describable visual attributes for face verification and image search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33:1962 – 1977.

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)*, pages 3730–3738.

Manyam, O. K., Kumar, N., Belhumeur, P., and Kriegman, D. (2011). Two faces are better than one: Face recognition in group photographs. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–8.

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*, pages 1–1.

Rokach, L. and Maimon, O. (2005). *Clustering Methods*, pages 321–352. Springer US, Boston, MA.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

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.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Song, F., Tan, X., and Chen, S. (2014). Exploiting relationship between attributes for improved face verification. *Computer Vision and Image Understanding*, 122:143–154.

Valueva, M., Nagornov, N., Lyakhov, P., Valuev, G., and Chervyakov, N. (2020). Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation*, 177:232–243.

Zhang, N., Paluri, M., Ranzato, M., Darrell, T., and Bourdev, L. (2014). Panda: Pose aligned networks for deep attribute modeling. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644.

Zheng, X., Guo, Y., Huang, H., Li, Y., and He, R. (2020). A survey of deep facial attribute analysis. *International Journal of Computer Vision*, 128.

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.