Biases in Neural Networks

by

Kassym Mukhanbetiyar

A Thesis Presented to the

Department of Computing Sciences

at Bocconi University

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science in Economics, Management and Computer Science.

May 2024

Table of Contents

Abstract

Facial anti-spoofing (FAS) systems are increasingly influencing critical decision-making processes. Recent studies have shown that FAS solutions exhibit significant differences in performance depending on user demographics. However, understanding the influence of a broader range of facial characteristics is critical to the development of credible and inclusive technologies. In this thesis, we examine FAS bias across a broad set of attributes. We study the impact of 43 attributes on the verification performance of a widely used FAS model, the Silent Face Anti-Spoofing model, using the publicly available CelebA dataset, which contains over 26,000 images with high-quality human annotations. Our results show that numerous non-demographic attributes, such as accessories, hairstyles and colors, face shape, and facial anomalies, significantly influence recognition performance. These observations highlight the urgent need to further develop robust systems that are more reliable and understandable. Additionally, our insights may contribute to a better understanding of FAS network functionality, improve reliability, and facilitate the development of more comprehensive solutions to mitigate FAS bias.

Chapter 1

Introduction

1.1 Background and motivation

Facial recognition technology [1] is becoming increasingly popular due to its convenience and high accuracy, especially in interactive smart applications such as mobile payments and registration. However, these systems are vulnerable to presentation attacks such as printing, rendering, makeup, and 3D masks. To address this issue, both academia and industry have focused efforts on developing face anti-spoofing (FAS) technology to protect facial recognition systems. FAS is an active area of computer vision research that has received increasing attention in recent years.

In the early stages, traditional feature-based methods [2], [3], [4] were proposed for presentation attack detection (PAD). These algorithms rely on human activity signals and hand-crafted features that require rich prior knowledge of the design task. For example, eye blinking, facial and head movements, eye tracking, and remote physiological signals have been examined for dynamic discrimination. However, these physiological activity signals are usually captured from long-term interactive videos of faces, which is inconvenient for practical applications. Moreover, these activity signals are easily imitated by video attacks, making them less reliable. In contrast, classic handcrafted descriptors (e.g., LBP [5], [6], SIFT [4], SURF [7], HOG [8] and DoG [9]) are designed to extract efficient substitution

patterns from different color spaces.

Subsequently, several hybrid (handmade + deep learning) [10], [11], [12] and end-to-end deep learning based methods [13], [14], [15] have been proposed for both static and dynamic facial PADs. Most of these methods treat FAS as a binary classification problem supervised by a simple binary cross-entropy loss. However, FAS is a self-evolving problem, which makes it even more complex. In addition, FAS relies on internal features that are irrelevant to the content and subtle, making them difficult to discern even for the human eye. Thus, convolutional neural networks (CNNs) with a single binary loss can detect clues for spoofing patterns (for example, screen bazels). The emergence of large-scale publicly available FAS datasets with a variety of attack types and recorded sensors has greatly expanded the research community. For example, CelebA-Spoof contains over 156,000 bona fide and 469,000 Presentation Attack (print, replay, 3D mask, etc.) face images for 10,177 subjects.

[16]

1.2 Problem Statement

Convolutional neural networks (CNNs) are widely used in computer vision applications, including face recognition and detection. However, recent studies [17] have shown that CNNs can exhibit biases in their decision-making processes, which can lead to unfair and discriminatory outcomes. These biases can be introduced by the training data, the network architecture, or the learning algorithms used. Therefore, there is a need to investigate the presence of biases in CNNs and develop methods to mitigate them.

In this thesis, we specifically explore solutions that do not involve retraining the model but instead identify problematic data that contribute to these biases. Our approach focuses on auditing pre-trained models to reveal their vulnerabilities rather than fine-tuning them for specific tasks. This distinction is crucial as it emphasizes our goal to assess and address biases in existing models without altering their original training.

1.3 Objective and Scope

The objective of this study is to assess the potential biases in a CNN-based face anti-spoofing model, called Silent Face Anti Spoofing , using the CelebA dataset. Specifically, we aim to investigate whether sensitive features, such as gender or race, affect the clusterization process of the embeddings extracted from the model's second to last layer. To achieve this objective, we will perform the following tasks:

- Preprocess the CelebA dataset to extract relevant facial features and annotations.
- Run the Silent Face Antispoofing model on the preprocessed CelebA dataset.
- Extract the embeddings from the model's second to last layer for each image in the CelebA dataset.
- Perform clusterization on the embeddings using various clustering algorithms.
- Evaluate the statistical significance of sensitive features, such as race and gender, in the clusterization process.

The scope of this study is limited to assessing the potential biases in the Silent Face Antispoofing model using the CelebA dataset. We will not investigate other sources of biases, such as the training data or network architecture. Additionally, we will not propose any bias-mitigating solutions in this study, but our findings can inform the development of such solutions in future research.

[18]

Chapter 2

Literature Review

2.1 Generalized Adversarial Adaptation

2.1.1 Generalized Adversarial Adaptation

Generalized Adversarial Adaptation provides a comprehensive framework that unifies various adversarial domain adaptation methods. This approach, as described by [19], offers a flexible structure to analyze and compare different domain adaptation techniques based on their design choices, such as weight sharing, base models, and adversarial losses.

2.1.1.1 Framework Overview

The Generalized Adversarial Adaptation framework identifies key components and design choices that influence the performance of domain adaptation methods:

- **Base Model**: The underlying neural network architecture used for feature extraction, which can be either generative or discriminative.
- **Weight Sharing**: The degree to which weights are shared between the source and target encoders. Methods can use shared weights (symmetric mapping) or unshared weights (asymmetric mapping).

- **Adversarial Loss**: The loss function used to train the domain discriminator and the target encoder. Common choices include minimax loss and domain confusion loss.

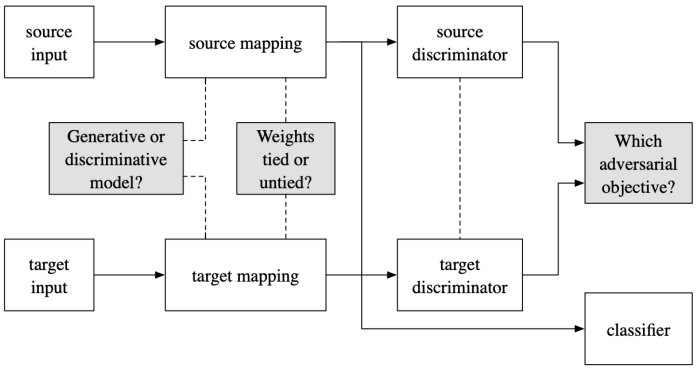


Figure 2.1: Overview of the Generalized Adversarial Adaptation framework. The framework allows for various design choices, including base models, weight sharing, and adversarial losses.. Source: [19]

2.1.1.2 Notable Instantiations

Several notable methods can be viewed as specific instances of the Generalized Adversarial Adaptation framework:

- **Gradient Reversal (GRL)** [20]: Uses a discriminative base model with shared weights and a gradient reversal layer to align source and target distributions by maximizing the domain classifier's loss.

$$\mathcal{L}_{adv_M} = -\mathcal{L}_{adv_D} = -\mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t}[\log D(M_t(\mathbf{x}_t))]$$

- **Domain Confusion** [21]: Employs a discriminative model with shared weights and a domain confusion loss to achieve domain invariance.

- **CoGAN** [22]: Utilizes two generative adversarial networks with unshared weights to generate source and target samples, achieving domain adaptation through tied high-level layers.

By understanding and leveraging these design choices, researchers can develop more effective domain adaptation methods tailored to specific tasks and challenges.

Generalized Adversarial Adaptation provides a valuable lens through which to view and develop domain adaptation methods, offering insights into the trade-offs and benefits of different design choices.

2.2 Adversarial Discriminative Domain Adaptation

Adversarial Discriminative Domain Adaptation (ADDA) [19] is a state-of-the-art unsupervised domain adaptation method that combines discriminative modeling with adversarial learning. ADDA aims to reduce the domain shift between a labeled source domain and an unlabeled target domain, thus improving the generalization performance of models on the target domain. In this section we discuss the model architecture, training process, and loss functions used in ADDA.

2.2.1 Model Architecture

The ADDA model contains four main components:

- **Source Encoder (M_s)**: A convolutional neural network (CNN) pre-trained on the source domain data to extract features from source images.
- **Target Encoder (M_t)**: A CNN initialized with the source encoder's weights, adapted to the target domain through adversarial training.
- **Domain Discriminator (D)**: A binary classifier that distinguishes between features extracted from the source and target domains.
- **Classifier (C)**: A classifier that takes the output of the encoder and produces the final classification of the model.

The source encoder remains fixed after pre-training, while the target encoder is trained to generate features that confuse the domain discriminator, making it unable to distinguish between source and target features.

2.2.2 Training Process

The training process of ADDA involves three stages:

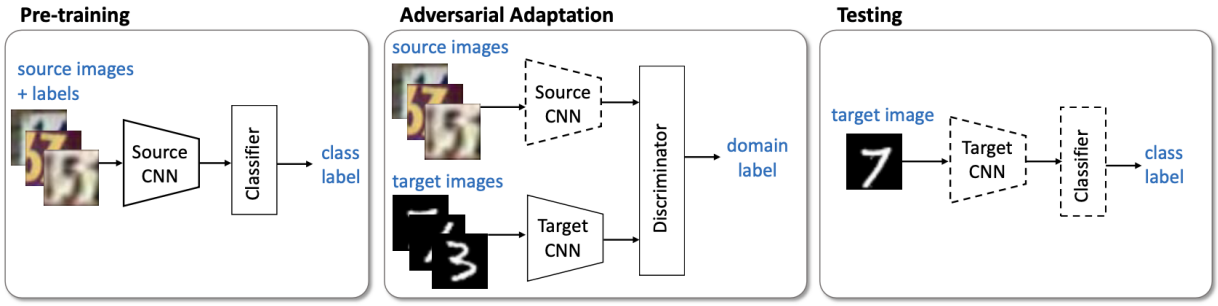


Figure 2.2: Overview of the ADDA architecture. The source encoder is pre-trained on the source domain. Next,the target encoder is trained to fool the domain discriminator during adversarial adaptation. The final model consists of target encoder and classifier. The dashed lines indicate fixed network parameters. Source: [19]

2.2.2.1 Pre-training the Source Encoder

The source encoder is trained on labeled source domain data using a standard supervised loss (e.g., cross-entropy loss). The objective is to learn a discriminative representation for the source domain.

$$\min_{M_s,C} \mathcal{L}_{cls}(\mathbf{X}_s, Y_s) = -\mathbb{E}_{(\mathbf{x}_s,y_s)\sim(\mathbf{X}_s,Y_s)} \sum_{k=1}^{K} \mathbb{1}_{[k=y_s]} \log C(M_s(\mathbf{x}_s)) \tag{2.1}$$

2.2.2.2 Adversarial Adaptation

The target encoder is trained using an adversarial loss. The domain discriminator is trained to distinguish between source and target features, while the target encoder is trained to fool the discriminator.

$$\min_{D} \mathcal{L}_{adv_D}(\mathbf{X}_s, \mathbf{X}_t, M_s, M_t) = -\mathbb{E}_{\mathbf{x}_s \sim \mathbf{X}_s}[\log D(M_s(\mathbf{x}_s))] - \mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t}[\log(1 - D(M_t(\mathbf{x}_t)))] \quad (2.2)$$

$$\min_{M_t} \mathcal{L}_{adv_M}(\mathbf{X}_s, \mathbf{X}_t, D) = -\mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t}[\log D(M_t(\mathbf{x}_t))] \quad (2.3)$$

2.2.2.3 Testing

During testing, target images are passed through the target encoder, and the resulting features are classified using the source classifier.

2.2.3 Loss Functions

The training of ADDA involves optimizing the following loss functions:

- **Supervised Loss for Source Classification (L_{cls})**: Ensures the source encoder learns discriminative features for the source domain.
- **Adversarial Loss for Domain Discrimination (L_{advD})**: Ensures that the domain discriminator can distinguish between source and target domain features.
- **Adversarial Loss for Target Encoder (L_{advM})**: Ensures that the target encoder learns to produce features indistinguishable from the source domain.

By minimizing these loss functions, ADDA effectively reduces the domain shift and enhances the model's performance on the target domain.

ADDA's ability to learn domain-invariant features makes it a powerful approach for unsupervised domain adaptation, providing significant improvements in performance on various domain adaptation tasks.

Chapter 3

Methodology

3.1 Face Anti-spoofing Models

In this thesis we will focus on deep learning-based methods that use convolutional neural networks (CNNs) to learn discriminative features from the input data. CNNs have shown superior performance in various computer vision tasks, including FAS. Most deep learning-based FAS models treat FAS as a binary classification problem, where the goal is to classify an input image as either real or fake. These models are trained on large-scale datasets with various PAs, such as CASIA-SURF, Replay-Attack, and MSU-MFSD. The most common loss function used in FAS models is binary cross-entropy.

The Silent Face Antispoofing model used in this study is a deep learning-based FAS model that uses CNNs to learn features from the input face images. It was proposed by Liu et al. (2018) and achieved state-of-the-art performance on the OULU-NPU dataset. The model consists of three main components: a feature extraction module, a fusion module, and a classification module. The feature extraction module uses a ResNet-18 network to extract features from the input face images. The fusion module combines the extracted features from multiple frames to enhance the discriminative power of the model. The classification module uses a softmax function to classify the input images as either real or fake.

In this study, we aim to assess the potential biases in the Silent Face Antispoofing model

by analyzing the embeddings extracted from the second to last layer of the model. We will use these embeddings to perform clusterization and evaluate the statistical significance of sensitive features, such as race and gender, in the clusterization process.

3.2 Dataset description (CelebA)

[23]

3.3 Pretrained Model ("Silent Face Anti-Spoofing")

The Silent Face Anti-spoofing model [24] mentioned in the previous chapter uses a silent living detection method based on the auxiliary supervision of the Fourier spectrum. This means that the model not only learns features from the input face images but also uses the Fourier spectrum as an auxiliary input to enhance its discriminative power.

The model architecture consists of two main branches: the main classification branch and the auxiliary supervision branch of the Fourier spectrum. The main classification branch is a ResNet-18 network that extracts features from the input face images. The auxiliary supervision branch takes the Fourier spectrum of the input face images as input and extracts features using a separate ResNet-18 network. The output features from both branches are then concatenated and fed into a fusion module that combines them to enhance the discriminative power of the model. Finally, the fused features are classified as either real or fake using a softmax function.

By incorporating the Fourier spectrum as an auxiliary input, the Silent Face Antispoofing model can better distinguish between real and fake faces, even in challenging scenarios where traditional handcrafted features may not be effective. This approach has shown promising results in several benchmark datasets [13], [18].

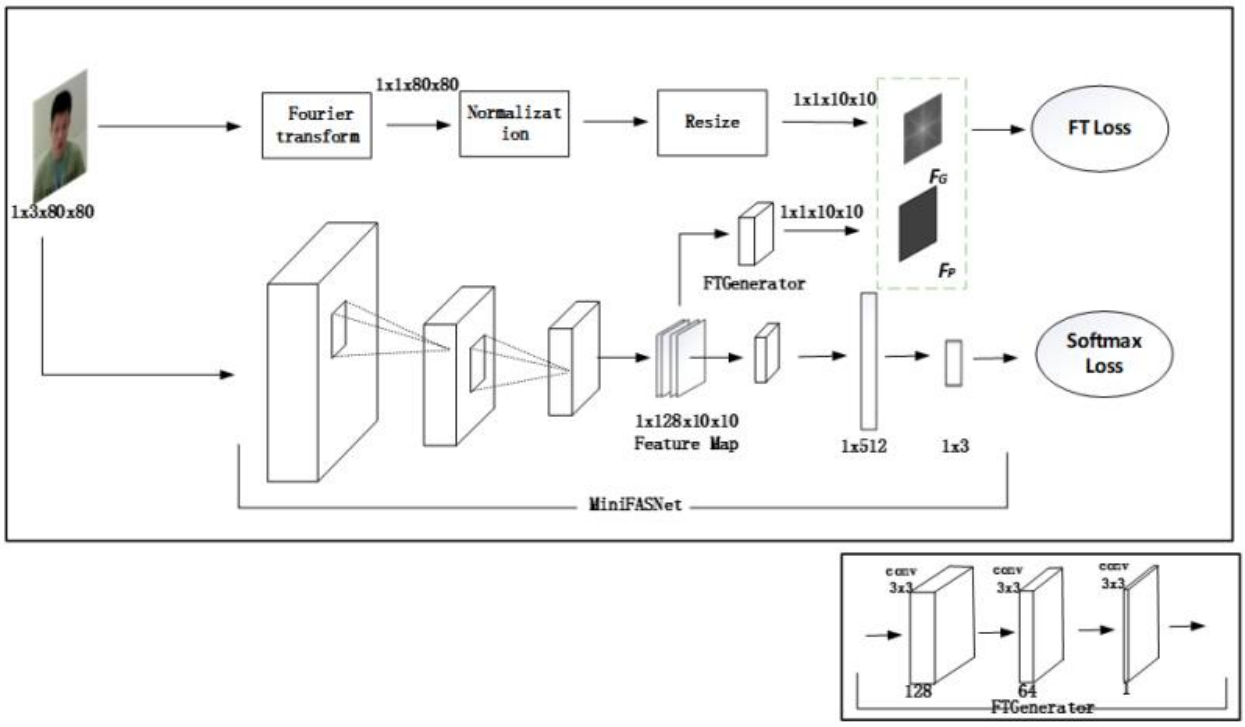


Figure 3.1: The architecture of the Silent Face Anti-spoofing Model. Source: [24]

3.4 Sensitive features in face recognition

Sensitive features are attributes of an individual that may be used to discriminate against them, such as race, gender, age, and ethnicity. In the context of face anti-spoofing, absence of sensitive features in the training process can potentially introduce biases in the decision-making process of the model, leading to unfair and discriminatory outcomes. Therefore, it is important to investigate the presence and impact of sensitive features in face anti-spoofing models.

In this study, we extracted 43 sensitive features from the CelebA dataset, which for over 25000 face images with annotations for various attributes. The sensitive features we considered include gender, race, facial hair, makeup, accessories, etc. We examined these features on their potential to introduce biases in the face anti-spoofing model.

Our findings suggest that it is important to consider the presence and impact of sensitive

features in face anti-spoofing models and develop methods to mitigate their effects. Future research can explore various methods, such as adversarial training, data augmentation, and fairness constraints, to address the issue of biases in face anti-spoofing models.

3.5 Extraction of Embeddings

To evaluate the impact of sensitive features on the Silent Face Antispoofing model, we performed clusterization on the embeddings extracted from the model's second to last layer (Dropout-202 Layer). Each embedding is a 128-dimensional tensor.

We then evaluated the statistical significance of each sensitive feature in the clusterization process using a chi-squared test.

Chapter 4

Results and Discussion

4.1 Visualizations of Embeddings

The following figure 4.1 is the visualization of sensitive features in the t-SNE representation of embeddings. [25].

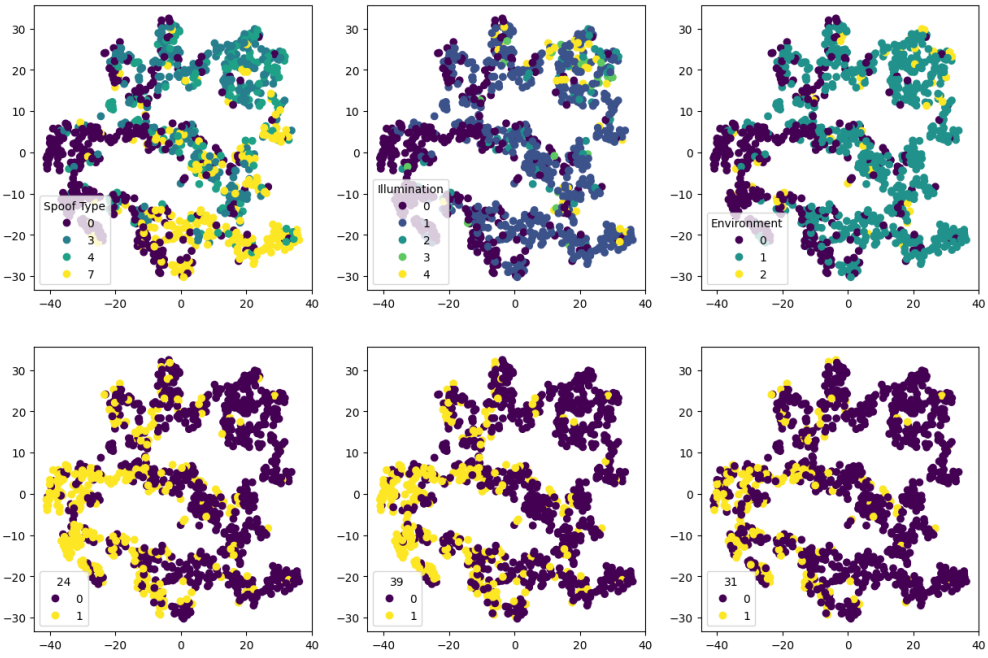


Figure 4.1: Color graded t-SNE representation of sensitive features

The Spoof Type 0, Illumination type 0 and Environment type 0 represent the live photos (purple dots on the top row of the Figure 4.1). On the bottom row of the Figure 4.1 we

can see that the model was able to successfully differentiate live photos from the spoofs with respect to these sensitive features.

4.2 Cluster analysis

The following figure represents the t-SNE representation of clusters in the embeddings 4.2.

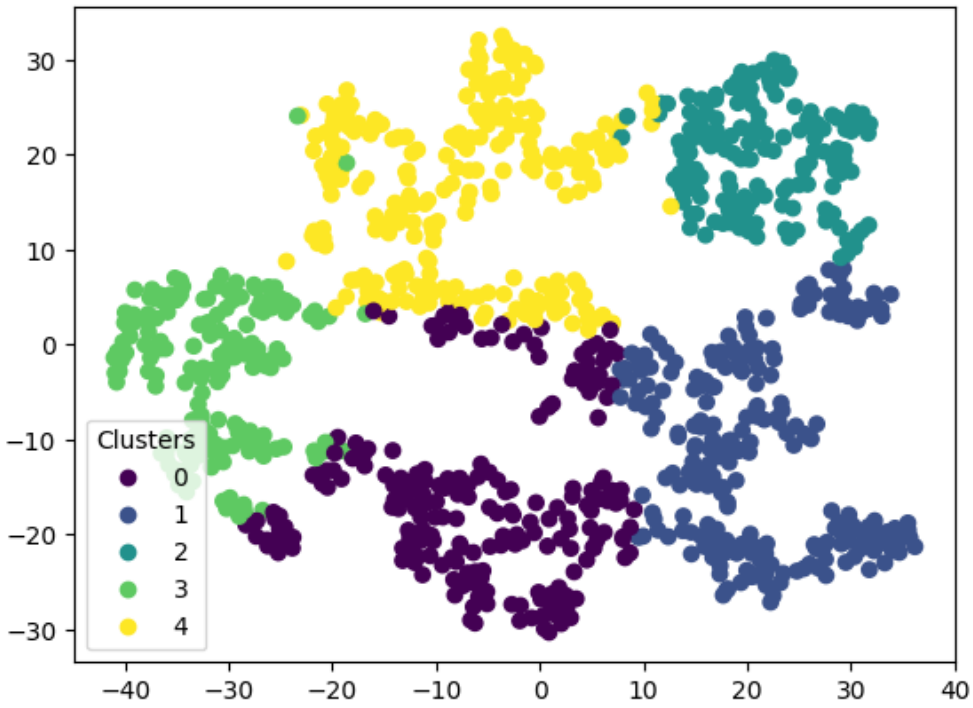
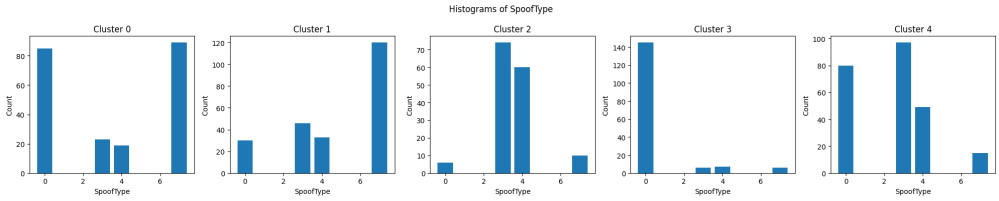
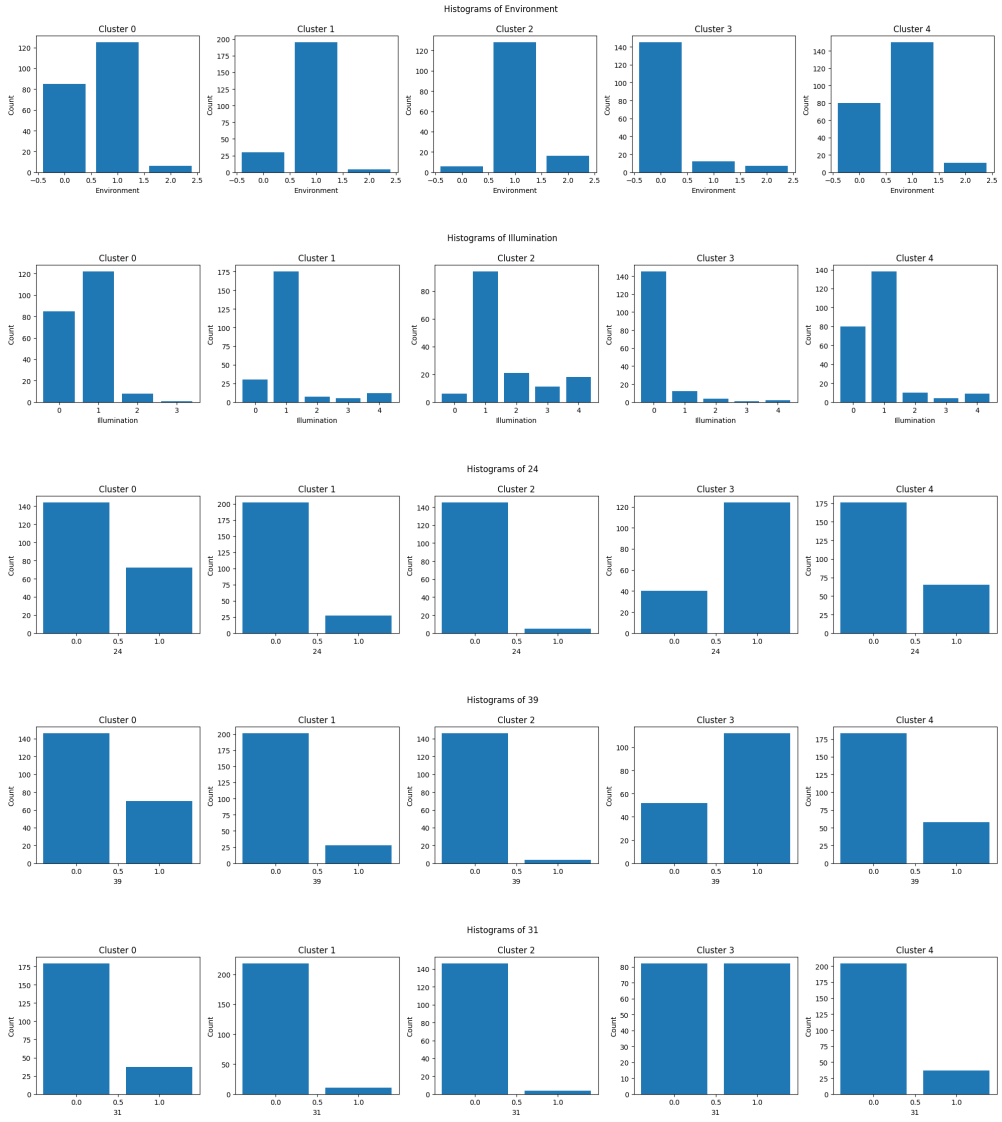


Figure 4.2: Embeddings Clusters

We can observe that the distribution of the sensitive features in the third cluster is drastically different from the rest indicating that the model is able to identify these features on images.





4.3 Chi-squared test for feature significance

Chi-squred test showed that several sensitive features, such as illumination, environment and spoof type were the most statistically significant in the clusterization process, which can be explained as the 0 value for these three parameters indicates that the photo is live and not a spoof. But, we also can see that the sensitive features 24, 39, 31, 19 and 2 had the most significant chi-sqaured value, as indicated in the figure 4.3.

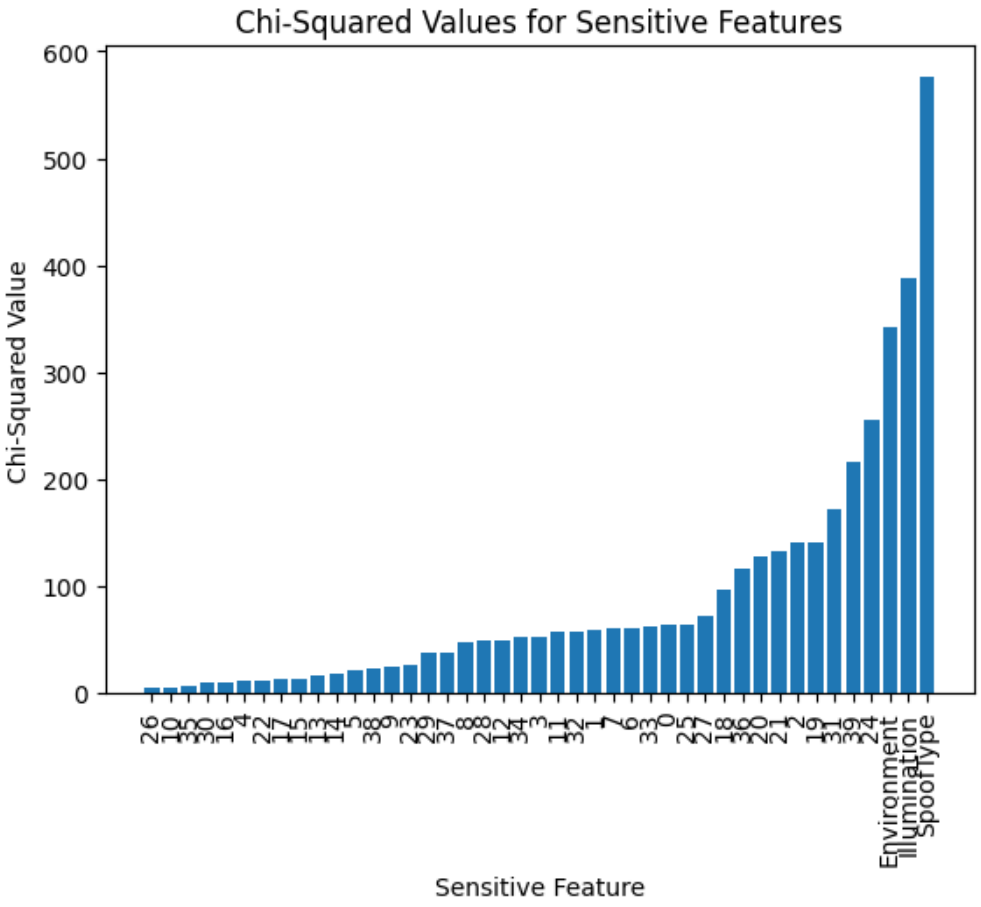


Figure 4.3: Chi-Sqaured value of the sensitive feature in the cluster formation

4.4 Top features and their impact on cluster formations

Chapter 5

Conclusion

5.1 Summary of findings

5.2 Limitations and future work

Chapter 6

Additional Information

6.1 Dimensionality Reduction and Clustering Techniques (t-SNE and GMM)

The Gaussian Mixture Model was used for the clusterization of embeddings.

Bibliography

[1] Jianzhu Guo et al. "Learning Meta Face Recognition in Unseen Domains". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[2] Gang Pan et al. "Eyeblink-based anti-spoofing in face recognition from a generic webcamera". In: *2007 IEEE 11th international conference on computer vision*. IEEE. 2007, pp. 1–8.

[3] Xiaobai Li et al. "Generalized face anti-spoofing by detecting pulse from face videos". In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, pp. 4244–4249.

[4] Keyurkumar Patel, Hu Han, and Anil K Jain. "Secure face unlock: Spoof detection on smartphones". In: *IEEE transactions on information forensics and security* 11.10 (2016), pp. 2268–2283.

[5] Tiago de Freitas Pereira et al. "LBP- TOP based countermeasure against face spoofing attacks". In: *Computer Vision-ACCV 2012 Workshops: ACCV 2012 International Workshops, Daejeon, Korea, November 5-6, 2012, Revised Selected Papers, Part I 11*. Springer. 2013, pp. 121–132.

[6] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. "Face anti-spoofing based on color texture analysis". In: *2015 IEEE international conference on image processing (ICIP)*. IEEE. 2015, pp. 2636–2640.

[7] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. "Face antispoofing using speeded-up robust features and fisher vector encoding". In: *IEEE Signal Processing Letters* 24.2 (2016), pp. 141–145.

[8] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. "Context based face antispoofing". In: *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE. 2013, pp. 1–8.

[9] Xiaoyang Tan et al. "Face liveness detection from a single image with sparse low rank bilinear discriminative model". In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11*. Springer. 2010, pp. 504–517.

[10] Xiao Song et al. "Discriminative representation combinations for accurate face spoofing detection". In: *Pattern Recognition* 85 (2019), pp. 220–231.

[11] Muhammad Asim, Zhu Ming, and Muhammad Yaqoob Javed. "CNN based spatio-temporal feature extraction for face anti-spoofing". In: *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. IEEE. 2017, pp. 234–238.

[12] Yasar Abbas Ur Rehman, Lai-Man Po, and Jukka Komulainen. "Enhancing deep discriminative feature maps via perturbation for face presentation attack detection". In: *Image and Vision Computing* 94 (2020), p. 103858.

[13] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. "Learning deep models for face anti-spoofing: Binary or auxiliary supervision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 389–398.

[14] Zitong Yu et al. "Searching central difference convolutional networks for face anti-spoofing". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5295–5305.

[15] Zitong Yu et al. "Face anti-spoofing with human material perception". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer. 2020, pp. 557–575.

[16] Zitong Yu et al. "Deep Learning for Face Anti-Spoofing: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.5 (2023), pp. 5609–5631. DOI: 10.1109/TPAMI.2022.3215850.

[17] Philipp Terhörst et al. "A comprehensive study on face recognition biases beyond demographics". In: *IEEE Transactions on Technology and Society* 3.1 (2021), pp. 16–30.

[18] Jiangwei Li et al. "Live face detection based on the analysis of fourier spectra". In: *Biometric technology for human identification*. Vol. 5404. SPIE. 2004, pp. 296–303.

[19] Eric Tzeng et al. "Adversarial discriminative domain adaptation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7167–7176.

[20] Edward Raff and Jared Sylvester. "Gradient reversal against discrimination". In: *arXiv preprint arXiv:1807.00392* (2018).

[21] Eric Tzeng et al. "Simultaneous deep transfer across domains and tasks". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4068–4076.

[22] Ming-Yu Liu and Oncel Tuzel. "Coupled generative adversarial networks". In: *Advances in neural information processing systems* 29 (2016).

[23] Yuanhan Zhang et al. "CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset with Rich Annotations". In: *European Conference on Computer Vision (ECCV)*. 2020.

[24] MinivisionAI. *Silent Face Anti-Spoofing*. https://github.com/minivision-ai/Silent-Face-Anti-Spoofing. 2020.

[25] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).