Deep Fair Models for Complex Data: Graphs Labeling and Explainable Face Recognition

Danilo Franco[a], Nicolò Navarin[b], Michele Donini[c], Davide Anguita[a], Luca Oneto[a,*]

[a] *University of Genoa - Via Opera Pia 11a, 16145, Genova, Italy*
[b] *University of Padua - Via Trieste 63, 35121, Padova, Italy*
[c] *Amazon - Rocket Tower, Charlottenstrasse 4, 10969, Berlin, Germany*

Abstract

The central goal of algorithmic fairness is to develop AI-based systems which do not discriminate subgroups in the population with respect to one or multiple notions of inequity, knowing that data is often humanly biased. Researchers are racing to develop AI-based systems able to reach superior performance in terms of accuracy, increasing the risk of inheriting the human biases hidden in the data. An obvious tension exists between these two lines of research that are currently colliding due to increasing concerns regarding the widespread adoption of these systems and their ethical impact. The problem is even more challenging when the input data is complex (e.g. graphs, trees, or images) and deep uninterpretable models need to be employed to achieve satisfactory performance. In fact, it is required to develop a deep architecture to learn a data representation able, from one side, to be expressive enough to describe the data and lead to highly accurate models and, from the other side, to discard all the information which may lead to unfair behavior. In this work we measure fairness according to Demographic Parity, requiring the probability of the model decisions to be independent of the sensitive information. We investigate how to impose this constraint in the different layers of deep neural networks for complex data, with particular reference to

*Corresponding author
Email addresses: danilo.franco@edu.unige.it (Danilo Franco), nnavarin@math.unipd.it (Nicolò Navarin), donini@amazon.com (Michele Donini), davide.anguita@unige.it (
Davide Anguita), luca.oneto@unige.it (Luca Oneto)

deep networks for graph and face recognition. We present experiments on different real-world datasets, showing the effectiveness of our proposal both quantitatively by means of accuracy and fairness metrics and qualitatively by means of visual explanation.

Keywords: Algorithmic Fairness, Learning Fair Representation, Demographic Parity, Deep Learning, Structured Data, Graphs, Face Recognition, Visual Explanation

---

1. Introduction

It has been argued that Artificial Intelligence (AI), and Machine Learning (ML) especially [1–3], is experiencing a fast process of commodification. This phenomenon means that ML-based systems are reaching the society at large and, therefore, the societal and ethical issues related to their use need to be directly addressed. Designing ML models from this human-centered perspective, thus building a more responsible AI [4], means incorporating human-relevant requirements such as safety, fairness, privacy, and interpretability, but also considering broad societal issues such as ethics and legislation. While these are essential aspects to foster the acceptance of AI and ML-based technologies, the research community has identified two main research directions [5]. The first one studies how AI-based systems can learn moral notions or ethical behaviors and then autonomously behave ethically. In this framework, Comparative Moral Turing Test [6] or Ethical Turing Test [7] have been proposed to assess the morality of the choices of automated systems. Because of the strong connection between philosophical, ethical, and technical problems, this branch of research is currently quite unexplored. The second one focuses on how humans should design and develop AI-based systems minimizing the possible harms derived from poor design, inappropriate application, or misuse. Algorithmic Fairness [8], Privacy Preserving Data Mining [9], Explainable AI [10], Adversarial Learning [11] are examples of efforts made in this direction. Our work falls within this category and, in particular, in the context of Algorithmic Fairness.

Generally speaking, Algorithmic Fairness deals with the problem of developing AI-based systems able to treat subgroups in the population equally. These subgroups are often determined by means of sensitive attributes, which should not be taken into account for decision purposes. Examples of these attributes are gender, ethnicity, and sexual or political orientation, etc. For

example, let us consider the emerging problem of political opinion polarization, where individuals' opinions are in opposition [12–15]. AI-based systems may exacerbate this disparity by creating echo chambers, consequently pushing towards an increased polarization. Social network algorithms try to maximize users engagement by providing limitative feeds on a particular subject, failing to furnish broader points of views. These echo chambers are quite dangerous since they reinforce and radicalize existing opinions [16–18]. Algorithmic Fairness, in this context, should help in designing algorithms able to avoid these filtered bubbles by showing to users interesting, plural, and informative feeds independently, for example, from their specific political orientations. Another representative case is the use of Face Recognition software by government agencies [19]. Many recent evidences [20–22] show how these algorithms can be biased against black people and females. In reaction to these issues, according to CNN [23], some governments banned the usage of face recognition systems in law enforcement agencies and public-facing businesses. Again, algorithmic fairness should help in designing algorithms able to mitigate and rectify these kinds of discrimination.

More specifically, one of the main questions in Algorithmic Fairness is how to enhance ML models, with fairness requirements. In fact, when these models take a decision in an human-oriented environment (e.g. decide whether to hire, to grant a loan, or to approve an insurance), it is ethically as well as legally discriminating to ground the choice on one or more sensitive attributes [24]. More formally, two types of discrimination are mainly considered in the literature: disparate treatment and disparate impact [25]. Disparate treatment highlights the case where the outputs depends directly on the knowledge of the sensitive attributes, while disparate impact describes decisions that end up being biased due to the correlation between non-sensitive and sensitive attributes, even if the latter are unknown. These observations give immediately hints on the complexity of formally defining fairness. Nevertheless, several notions of fairness already exist in the literature [26]. The most common notions are surely Demographic Parity (DP) [27] or Equal Odds and Equal Opportunities [28]. The idea behind the general notions of fairness is that the learned ML model should behave equally, or at least similarly, no matter whether it is applied to one subgroup in the population or to another one (e.g. females respect to males or black people respect to white people). For example, DP implies that the probability of a certain ML model output should not depend on the value of one or more specific sensitive attributes. Nevertheless, these definitions, also called Group (or Statistical)

Fairness [29], are far from being perfect. Specifically, many Group Fairness definitions proved to be mathematically incompatible [30, 31]. Moreover, Group Fairness definitions, despite removing averaged discrimination from the models, might still allow unfair treatments between specific individuals. In this regard, Individual Fairness [29] definitions try to fill these gaps. Once one or more formal definitions of fairness are chosen depending on the specific problem under exam [32], it is possible to plug them into the process of building models through ML training algorithms. For this purpose, three main mitigation approaches exists [8]. The first approach consists in pre-processing the data to remove historical biases and then feeding this refined data to classical ML models. This approach is quite convenient when one wants to make out-of-the-box ML tools fairer, slightly tweaking the data, without actually changing the tool. The second approach consists in post-processing the output of an already learned ML model. This approach is particularly useful to avoid the retraining or fine-tuning of already trained complex models for fairness reasons. The third approach, called in-processing, consists in imposing fairness constraints directly into the learning phase, enforcing fairness in the models inner structures.

A specific approach, which lies in the middle between pre- and in-processing, is to extract a fair representations of the data that can be transferred and used for other tasks while ensuring that every model trained over this representation will be again fair [33–35]. This particular approach, referred to as fair representation learning [8, 26, 36], is becoming increasingly important nowadays due to the intensive use of Deep Learning (DL) architectures. In fact, nowadays, DL [37] represents the state-of-the-art alternative for a wide variety of real world applications which require to automatically learn a compact yet rich representation of complex data. Visual Understanding [38], Natural Language Processing [39], Drug Discovery [40], Medicine [41], and Graph Analysis [42] are just few examples of domains where DL outperformed classical ML methods. For example in Face Recognition, traditional methods attempted to extract handcrafted shallow features (e.g. Viola-Jones [43], Gabor [44], LBP [45], LGBHPS [46]) and, before the advent of DL, they represented the state of the art for classical benchmark datasets [47]. DL approaches have recently shown to outperform these methods being more robust to changes in illumination, face pose, aging, expressions, and occlusions [48].

Besides face recognition applications, in some other cases DL is even able to surpass human performances (e.g. melanoma classification [49] and logic

based game [50]) or is expected to do it in the near future [51]. Classical ML methods exploit features engineered from the raw data based on the domain knowledge [52]. DL, instead, is able to actually learn a compact and rich data representation by means of a multilayered deep network designed to comply with the particular raw data format under exam and a huge amount of (un-)labeled samples [37]. As a positive side effect, these representations can be reused, incrementally enriched, and fine-tuned for many different tasks [53–56]. This preamble clearly points to the fact that DL is much more effective than classical ML when the input data are complex, or more precisely, structured (e.g., in the form of graphs [42], trees [57], general images [58], or faces [59]). The complex geometrical and topological relations present in these kind of data, usually require ad-hoc pre-processing or kernels [60] methods, based on the experience of domain specialists. DL revolutionized this field by actually being able to learn those representations directly from the data. Nevertheless, this ability of relying ultimately on data transformations for shaping highly performant models also increases the risk of carelessly inheriting the human historical biases hidden in the data itself.

For this reason, in this paper, extending our previous work [61] and leveraging on the theoretical results of [33], we investigate how to impose the fairness constraint in the representation layers of Deep Neural Networks (DNNs) for complex input data, with particular reference to DNNs for Graphs and Face Recognition. We decided to measure fairness according to DP, requiring the probability of a data representation, and consequently of possible model decisions, to be independent of the sensitive information.

The contribution of our work can be summarized as follows:

- We analyze how the layers of a DNN for complex data have to be constrained in order to obtain fairer and accurate results, measuring the effects by means of different fairness and accuracy metrics. Then we specialize our analysis to two tasks with particularly complex input data: Graph labelling and Face Recognition;
- We impose the fairness constraint by means of different Tikhonov [62] regularizers. All the proposed constraints are differentiable, and in some cases also convex approximations of the DP, which is computationally hard to handle in its naive formulation. Specifically, we define the fairness regularizers by means of (i) a simple first order convex approximation [63] of the DP, (ii) the Maximum Mean Discrepancy [64–66], and (iii) the Sinkhorn Divergence [67, 68];
- Apart from using classical metrics like the Difference of Demographic

Parity (DDP) [33] for characterizing the models fairness and misclassification error or the area under the receiver operating characteristic curve for characterizing the models accuracy, we try to reach a deeper understanding of the effects of the proposed fairness constraints on the learning process of the DNNs. For this reason, in order to see the modifications of the DNNs perception for image recognition, we exploit a state-of-the-art Explainable-AI [10, 69] tool: visual explanation through the use of attention maps [70] employing the Grad-CAM technique [71]. This step allows us to show the effectiveness of our proposal not just from a quantitative point of view but also qualitatively via visual explanation;

- We consider also real-world state-of-the-art datasets in our study, namely FairFace [72] for Face Recognition and Pokec[73] for Graph labeling. Finally, in case of Graph labeling, we also generate a new dataset, which we named Marvel, combining two publicly available datasets on the Marvel universe [74, 75].

The rest of the paper is organized as follows. Section 2 discusses the related works on the topic of Learning Fair Representation. Section 3 introduces some preliminary definitions. Section 4 presents the specific DL architectures for Graphs and Face Recognition. Section 5 presents our proposal to make the architectures described in Section 4 fairer. Results on real world datasets are presented in Section 6. Section 7 concludes the paper.

2. Related Works

Recently, Fair Representations Learning has attracted the attention of the scientific community due to the inherent fairness guarantee characterizing any predictor trained on top of an unbiased representation space. Similarly to what is proposed in this work, the vast majority of the literature on this topic advances to learn fair representations using regularizers to balance utility and fairness.

The idea of mapping data points from the original input space to a new, so called, representation space where any implicit or explicit information regarding the sensitive attribute is removed was firstly introduced by Zemel et al. [36]. In this work, fairness is ensured through a probabilistic mapping from the original input space to a set of prototypes satisfying DP. Specifically, each data point in the input space is assigned to a particular prototype with a probabilistic rule constructed by the Euclidean distance between the origi-

nal and the representation spaces. In order to satisfy DP, the probability of mapping two random individuals belonging to two different subgroups to the same prototype should be the same no matter the value of the sensitive attribute. Moreover, other constraints for an optimal mapping are introduced, such as the preservation of both the non-sensitive information of the original input space and the original prediction accuracy.

Lahoti et al. [76] extend the work of Zemel et al. [36] by considering a fairness definition close to individual fairness [77]. Specifically, given a set of input data and their respective non-sensitive counterparts (where the sensitive attributes are removed), the task is to find an optimal representation that minimizes both the reconstruction loss and the individual fairness loss for a chosen binary distance function. Again, the intuition behind individual fairness, which can be easily observed in the previous formulation, is that any difference in the new representation space needs to be justified by a non-sensitive difference in the original space. Note how this formulation automatically enables the support for multiple sensitive attributes with unknown values (as long as the distance can be defined both on the original and representation spaces).

Another line of works [78–81] on the regularization framework exploits auto-encoders [82] for ensuring fairness. These works formulate a generic Bayesian model which admits two distinct independent sources: one which determines the sensitive information, and one which models all the remaining "legal" information. The input datapoint is then generated according to a conditional probabilistic rule that takes into account both the prior sources. Consequently, the problem results in finding a representation that is invariant to the values of the sensitive prior through modeling the parameters of the posterior distribution. Moreover, Luoizos et al [78] also penalize any leakage of the sensitive attribute into the posterior of the representation by comparing the two posterior distributions for a binary sensitive attribute through the Maximum Mean Discrepancy [64].

Oneto et al. [33] tackle the problem of learning transferable fair representations through the regularization over compositional models with a shared representation. In this setting, fairness is enforced at the representation level by imposing the DP via Maximum Mean Discrepancy [64] and Sinkhorn Divergence [67]. They complement their work also proving the existence of learning bounds on the accuracy and fairness of the learned model in the lifelong setting.

Many works [34, 83–99] exploits the Generative Adversarial Networks [100]

to achieve fairness. These works find an optimal fair representation through an optimization process where two entities, an encoder and a decoder (a.k.a. the adversary) are opposed in a minimax game. Specifically, the adversary tries to maximize an unfairness measure, while the encoder tries to fool the adversary finding a representation that is able to minimize the dependency with the protected attributes and to minimize a reconstruction error from the original space (thus preserving the non-sensitive information). In a supervised setting, the encoder will also consider a prediction error for finding a suitable representation vector.

Recently, Song et al. [101] unified some of the previous works and develops a general framework relying on the concept of mutual information between the legitimate and protected attributes.

Other works in the literature enforce Fair Representation Learning by following intuitions different from imposing fairness through a regularization framework. Since these works are quite different from the one proposed in this work, we just briefly analyze their contributions (for more details please refer to recent reviews on algorithmic fairness [26, 29]):

- *Rank of Conditional Distributions* [102–104], the learned representation aims at removing discrimination while preserving the rank of the legitimate variables distributions conditioned on the protected attributes;
- *Fairness Graph* [105], the extracted representation is constrained by a graph structure and, specifically, it preserves the graph local neighborhoods. Rather than learning fair embeddings for a specific graph (for example, through adversarial regularization on protected attributes decoding [106]), this approach associates similar points in the learned representation to connected individuals in the fairness graph, enforcing individual fairness.
- *Fair Dimensionality Reduction* [107], the unbiased representation is extracted from the orthogonal complement of the feature projection that captures the information related to the protected attribute to obtain a fair subspace for high predictive kernel models.
- *Fair Disentangled Representation* [108, 109], the aim is to find a generative model composed by latent independent ground factors. A fair predictor could be then just trained on the latent independent factors (known as disentangled representations) that are not related to the protected attributes.
- *Semantical Meaning Preservation* [90, 110, 111], in addition to the task of learning fair representation, these approaches aim at preserving se-

mantical meaning. For example, in the context of face recognition, they aim at conditioning images input data by removing features correlated with a protected attribute (the presence of a beard in the example of gender discrimination), while retaining the structure of a face in the extracted representation.

- *Theoretical Guarantees*, the trade-off between utility and different notions of fairness when learning invariant representation is theoretically characterized. Zhao et al. [112] proved the existence of a lower bound on the joint error across groups when the base prediction rates differ. Another line of works quantified the reduction in discrimination capabilities achieved by a certain representation mapping [113, 114].

3. Preliminaries

Let us consider the problem of assigning a binary label to structured inputs[1] with fairness requirements.

The problem is identified by a probability distribution μ on $\mathcal{I} \times \mathcal{S} \times \mathcal{Y}$, where $\mathcal{I}$ is a (semi-)structured input space, $\mathcal{S} = \{1, 2\}$ is the set of values of a binary sensitive variable[2] and $\mathcal{Y} = \{-1, 1\}$. We let $\mathcal{D} = (I_i, s_i, y_i)_{i=1}^n \in (\mathcal{I} \times \mathcal{S} \times \mathcal{Y})^n$ be a set of data, which is sampled independently from μ. For each sensitive $s \in \{1, 2\}$ we also let $\mathcal{D}^1 = \{(I, s, y) \in \mathcal{D} : s = 1\}$ and $\mathcal{D}^2 = \{(I, s, y) \in \mathcal{D} : s = 2\}$ be the set of inputs in the first and second group, respectively. The goal is to learn a model $\mathsf{h} : \mathcal{I} \times \mathcal{S} \to \mathcal{Y}$ able to well approximate $\mathbb{P}\{y \,|\, I, s\}$. Note that using, implicitly of explicitly, the sensitive attribute in the functional form of the model may be illegal with respect to some jurisdictions [115, 116]; in these cases $\mathsf{h} : \mathcal{I} \to \mathcal{Y}$ approximates $\mathbb{P}\{y \,|\, I\}$. Formally, we will indicate $\mathsf{h} : \mathcal{Z} \to \mathcal{Y}$ such that it well approximates $\mathbb{P}\{y \,|\, Z\}$ where $Z \in \mathcal{Z}$ may contain ($\mathcal{Z} = \mathcal{I} \times \mathcal{S}$) or not ($\mathcal{Z} = \mathcal{I}$) the sensitive attribute, depending on the actual legislation. The ability of h in approximating $\mathbb{P}\{y \,|\, Z\}$ is usually measured with different indices of performance $\mathsf{P}(\mathsf{h})$ based on the different task under exam [37]. For example, when tackling a binary classification problem, typical formulation

[1] Our method will be tested on binary classification of graph nodes and images but it naturally extends to multiclass and regression and other kinds of structured data such as natural language and trees.

[2] Our method naturally extends to multiple sensitive variables but to ease the presentation we consider only the binary case in the paper.

of $\mathsf{P}(\mathsf{h})$ are the Accuracy measure, the Binary Cross-Entropy, or the Area Under the Receiver Operating Characteristic curve [117].

With the increased use of deep learning models, the model h is actually a composition of models $\mathsf{m}(\mathsf{r}(Z))$, where $\mathsf{m} : \mathbb{R}^d \to \mathcal{Y}$ is usually a (non-)linear function approximating the desired prediction and $\mathsf{r} : \mathcal{Z} \to \mathbb{R}^d$ is a function mapping the input into a vector, which is usually referred to as the representation. Note that r can be a composition of functions too $\mathsf{r} : \mathsf{r}_l \circ \cdots \circ \mathsf{r}_2 \circ \mathsf{r}_1$, for example, in a deep architectures of l layers [37]. In other words, the function r synthesizes the information needed to well describe the structured input and to learn an accurate model m.

Moreover the model h should be fair with respect to one or more notions of fairness [8]. As recently theoretically studied in [33] and practically demonstrated in many works [34–36, 78, 85, 88, 113, 114, 118], when deep learning models are developed, learning fair representation is a much more effective and cognitively grounded way of learning fair models. In fact, learning a fair representation implies (i) to learn fair models no matter the task that will leverage on this representation, and (ii) being able to remove the historical biases not just in the last layer of the network but also from the deeper layers making the entire network fairness-aware. Specifically, we require that the representation vector satisfies the Demographic Parity[3] (DP) constraint [29, 33]. Namely, we require that

$$\mathbb{P}_Z\{\mathsf{r}(Z) \in \mathcal{C} \,|\, s = 1\} = \mathbb{P}_Z\{\mathsf{r}(Z) \in \mathcal{C} \,|\, s = 2\},\ \forall \mathcal{C} \subseteq \mathbb{R}^d, \tag{1}$$

that is, the two conditional distributions of the representation vector, the one for nodes with $s = 1$ and the one with $s = 2$, should be the same. Note that our method naturally extends to constraining all ($\forall\, i \in \{1, \cdots, l\}$) or some ($\forall\, i \in \mathcal{L} \subseteq \{1, \cdots, l\}$) of the layers composing the representations as follows

$$\mathbb{P}_Z\{\mathsf{r}_i(Z) \in \mathcal{C} \,|\, s = 1\} = \mathbb{P}_Z\{\mathsf{r}_i(Z) \in \mathcal{C} \,|\, s = 2\},\ \forall \mathcal{C} \subseteq \mathcal{R}(\mathsf{r}_i), \tag{2}$$

where $\mathcal{R}(\mathsf{r}_i)$ is the domain of the r_i-th layer. The constraint of Eq. (1) (or Eq. (2)) implies also that any model learned on top of a fair representation

[3]Other notion of fairness could be exploited in this paper like Equal Opportunity and Equal Odds [28], but the extension of the proposal is quite simple and out of the scope of this paper.

will be again fair [33]

$$\mathbb{P}_Z\{\mathsf{m}(\mathsf{r}(Z)) = y \,|\, s = 1\} = \mathbb{P}_Z\{\mathsf{m}(\mathsf{r}(Z)) = y \,|\, s = 2\}, y \in \mathcal{Y}. \tag{3}$$

with respect to the notion of DP for binary classification, namely the probability of assigning a particular label should not depend on the value of the sensitive attribute.

The final model performances $\mathsf{P}(\mathsf{h})$ will be evaluated through the Accuracy measure ($\mathsf{ACC}_y(\mathsf{h})$) or the empirical Area Under the Receiver Operating Characteristic curve ($\mathsf{AUROC}_y(\mathsf{h})$), which is more informative in the case of unbalanced datasets. Such measures will be computed on a test test not exploited during the model training phase in order to avoid biased results [119].

The fairness of the final models h, instead, will be measured with the Difference of Demographic Parity [33] ($\mathsf{DDP}(\mathsf{h})$)

$$\mathsf{DDP}(\mathsf{h}) = \left| \frac{1}{|\mathcal{D}^1|} \sum_{(Z,y)\in\mathcal{D}^1} [\mathsf{h}(Z) > 0] - \frac{1}{|\mathcal{D}^2|} \sum_{(Z,s,y)\in\mathcal{D}^2} [\mathsf{h}(Z) > 0] \right|, \tag{4}$$

where the Iverson bracket notation is exploited, together with the accuracy of h in learning s (e.g. by means of $\mathsf{ACC}_s(\mathsf{h})$ or $\mathsf{AUROC}_s(\mathsf{h})$). The latter is a sanity check on the fairness of $\mathsf{r}(Z)$ since we are measuring the ability of the same model m to learn s instead of y from $\mathsf{r}(Z)$ itself. Also the fairness will be computed on a test set not exploited during the model learning phase in order to avoid biased results [119].

3.1. Graph Binary Classification

Let us introduce the graph nodes binary classification problem faced in this paper. A training graph $\mathcal{G}^{\text{Tr}} = (\mathcal{V}, \mathcal{E}, X, \boldsymbol{s}, \boldsymbol{y})$ is given, where $\mathcal{V} = \{v_1, \ldots, v_{d_d}\}$ is the set of d_d nodes (or vertices), $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, $X \in \mathbb{R}^{d_d \times d_x}$ is the matrix of non-sensitive attributes ($\boldsymbol{x}_i$, the i-th row of X, is the vector of attributes associated to the vertex v_i), $s_i \in \{1, 2\}$ is the sensitive attribute associated to node v_i, and $y_i \in \{\pm 1\}$ is label associated to node v_i. Let us define the neighborhood of a vertex v_i as $\mathcal{N}(v_i) = \{v_j | (v_i, v_j) \in E\}$. The training set is composed by all nodes in the training graph. The goal is to learn a model $\mathsf{h}(Z)$, where we indicate Z its input composed by v, $\mathcal{E}$, $\boldsymbol{x}$, and possibly s if legally possible. We consider the challenging inductive setting, where two different graphs are taken, one for training and one for testing. In other words, the test is the set of nodes from a second graph $\mathcal{G}_{\text{Te}}$. A dataset, $\mathcal{D}_{\text{Tr}} = \{(Z_i, s_i, y_i) | i \in \{1, \cdots, d_d\}\}$, is generated from $\mathcal{G}_{\text{Tr}}$ and, analogously, $\mathcal{D}_{\text{Te}}$ is generated from $\mathcal{G}_{\text{Te}}$.

3.2. Faces Binary Classification

Let us introduce the faces binary classification problem. In this case $\mathcal{I}$ is the space of all RGB images of human faces. Then $\mathcal{I} \subseteq \mathbb{R}^{h \times w \times 3}$ where h and w are the height and width of the image, and then we have the three color channels (red, green, and blue). Images have two type of information: one is the actual color of each pixel and the other one is the relative position of the pixels. For this reason, the model needs to take into account both information in order to achieve acceptable performances. In this case $\mathcal{D}_{\text{Tr}}$ and $\mathcal{D}_{\text{Te}}$ are set of different and mutually exclusive human faces.

4. Deep Neural Networks

In this section, we will recall the DNN exploited in this paper for binary labeling of graph nodes and faces.

4.1. Deep Neural Networks for Graphs

In this paper, we consider the GraphSAGE DGNN model [120], since, contrarily to other architectures in literature (e.g. [121]), it is designed to deal with large graphs (such as social network graphs) sampling a fixed-size set of neighbors, while achieving competitive predictive performance, in particular on the considered inductive setting. The representation of a node v at layer k is defined as:

$$\mathsf{r}_{k,v} = \mathsf{ReLU}\left(W_k \cdot \mathsf{mean}\left(\{\mathsf{r}_{k-1,v}\} \cup \{\mathsf{r}_{k-1,u}, \forall u \in \mathsf{sample}(\mathcal{N}(v), n_s)\}\right)\right), \quad (5)$$

where W_k is the matrix of parameters for the k-th layer, ReLU [122] is the rectified linear activation function, mean is the function returning the mean vector over a set of vectors[4], and sample is a function randomly sampling a subset of n_s elements in the set of neighbors $\mathcal{N}(v)$. We then stack multiple (d_l) layers like the one of Eq. (5), and a fully connected output layer. For more details about the network, we refer the reader to the original paper [120]. The DGNN has been trained using the Adam optimizer, minimizing the empirical Binary Cross-Entropy ($\mathsf{BCE}(\mathsf{h})$).

[4]In the original work [120] mean can be substituted with any aggregation operator.

4.2. Deep Neural Networks for Face Recognition

For the facial recognition task, we rely on the VGGNet-16 convolutional neural pre-trained network [123, Configuration D]. The peculiarity of the VGG-based nets is that they exploit deeper architectures, leading to more accurate results for a variety of different tasks, while maintaining low computational requirements thanks to the use of small filters. In fact, stacking convolution layers with small kernels is equivalent, yet less computationally demanding, to use a single layer with larger kernels [123]. Moreover, the use of multiple stacked layers allows to easily increase the non linearity of the network by adding an activation function at each intermediate step (e.g. the ReLU [122]) if compared to a single layer with larger kernel. The VGGNet-16 embeds the faces in a 25 088-dimensional vector space r by means of 14 million parameters. The VGGNet-16 deployed in this work has been pre-trained on the VGG-Face face recognition dataset [124]. Exploiting the VGGNet-16 embeddings it is possible to easily achieve almost state-of-the-art results, in terms of accuracy, in multiple face recognition related tasks [125–128].

Using VGGNet-16, allows us to test the effectiveness of our methods for learning fair representations also in complex tasks which would require huge amount of data and computational power for simply training the network. Instead, using a pre-trained network allows us to start already from "good" (in terms of accuracy) embeddings and then fine-tune them, exploiting different alternatives and regularizers, toward the definition of "good" (accurate) and "fair" (in terms of DDP) embeddings. In order to use the VGGNet-16 embeddings for our final face recognition task, we stacked on top of the embeddings a single hidden layer neural network (a universal approximator). The hidden layer has a sigmoid activation and the output layer has a softmax activation. The weights of these last layers are initialized randomly according to a 0-mean and .01-variance Gaussian distribution. The VGGNet-16 has been fine-tuned using the ADADELTA [129] minimizing the empirical Binary Cross-Entropy ($\mathsf{BCE}(\mathsf{h})$). Based on the final scope of the analysis, we kept fixed or fine-tuned its weights (e.g. just the ones of the embedding or also the deeper ones).

4.2.1. Visual Explanation

In order to visualize how the different DNNs for Face Recognition reacts to the input images, we will exploit a state of the art Visual Explanation tool: Grad-CAM [71]. Visual Explanation produces visual attention maps (namely, heatmaps images) that highlight the most predictive image regions

for a particular task. In our context, attention maps can graphically represent any divergence between the heatmaps of images belonging to different populations. More specifically, we extract the Grad-CAMs relative to the last convolution layer (since it represents the face embedding/representations) and analyze any difference based on the sensitive attribute.

Grad-CAM works as follows: fixed a prediction target Y, a non-normalised network score Y_s for the target Y (namely prior to the softmax activation at the end of the DNN), a convolutional layer output $A \in \mathbb{R}^{K \times U \times V}$ where we extract the matrix $A_k \in \mathbb{R}^{U \times V}$ relative to the channel $k \in \{1, \cdots, K\}$ (U, V are the output matrices dimensions for any of the k channels), the gradient $G_{Y,k} \in \mathbb{R}^{U \times V}$ of Y_s with respect to A_k is then defined as

$$G_{Y,k} = \frac{\partial Y_s}{\partial A_k} \tag{6}$$

We can then obtain the importance weight of the dimension k with respect to the class Y as the average $\alpha_{Y,k}$ across the convolutional layer matrix entries (or pixels)

$$\alpha_{Y,k} = \frac{1}{UV} \sum_{i=1}^{U} \sum_{j=1}^{V} G_{Y,k,i,j} \tag{7}$$

The latter quantity captures the importance of the channel k (across the whole layer feature mapping) when trying to predict Y. Finally, the Grad-CAM map with respect to a target Y is defined as L_Y, namely the weighted sum across all the dimensions k

$$L_Y = \mathsf{ReLU}\left(\sum_{k=1}^{K} \alpha_{Y,k} A_k \right) \tag{8}$$

where the ReLU [122] simply suppresses the negative values highlighting the interest for the features that have only a positive influence towards a certain target.

Although gradient-based methods might not be the optimal solution for visual explanation (e.g. saturation, zero-gradient image regions, and false confidence in the output score phenomena [70]) the computational cost Grad-CAMs is negligible compared to other methods that require multiple network forward-passes per image [70, 130]. Moreover, in most recent works, Grad-CAM is used as the baseline methods from improvements margins [130–134].

5. Deep Fair Neural Networks

In this section we will propose different approaches for imposing the fairness constraint of Eq. (1) into the DNNs described in Section 4.

In particular, we propose to add the fairness constraint as regularizer $\mathsf{F}(\mathsf{h})$, through the Tikhonov philosophy [37, 62], in the cost function to be minimized for training the DNNs, together with the $\mathsf{P}(\mathsf{h})$, as follows

$$\mathsf{h}^* = \arg\min_{\mathsf{h}} \ (1-\lambda)\mathsf{P}(\mathsf{h}) + \lambda\mathsf{F}(\mathsf{h}), \tag{9}$$

where $\lambda \in [0,1]$ trades off accuracy and fairness as we will also see in Section 6. Note that the constraint could have been imposed using the Ivanov philosophy [135], and the results would be the the following optimization problem

$$\mathsf{h}^* = \arg\min_{\mathsf{h}} \ \mathsf{P}(\mathsf{h}), \quad \text{s.t. } \mathsf{F}(\mathsf{h}) \leq \eta, \tag{10}$$

where $\eta \in [0,1]$ regulates the level of accepted fairness, which is cognitively more close to the problem of imposing a certain level of fairness to the final model. Nevertheless note that, for some values of η and λ the two problems are equivalent and that Problem (9) is much less computationally demanding with respect to Problem (10) [136]. Note also that setting $\eta = 0$ in Problem (10) (or $\lambda \to 1$ in Problem (9)) to impose the DP, does not guarantees fairness in terms of generalization since Problem (10) (or Problem (9)) exploit empirical quantities. $\eta, \lambda \in [0,1]$ allow to avoid to overfit the particular sample.

As previously described, the constraint, and then the regularizers, can act on the representation layers in different ways. On way is to impose the constraint just on the last layer of the representation, namely $\mathsf{F}(\mathsf{h}) \to \mathsf{F}(\mathsf{r})$. The other way is to impose the constraint on all or some of the layers of the representation, namely $\mathsf{F}(\mathsf{h}) \to \mathsf{F}(\mathsf{r}_i | \forall i \in \mathcal{L})$ where $\mathcal{L} \subseteq \{1, \cdots, l\}$.

Unfortunately, the constraint of Eq. (4) is hard to handle and transform in an effective yet computationally efficient regularizer. In this work, we propose three different alternatives to reach this goal.

The first one is based on the work of [33], where a convex approximation and relaxation of the constraint of Eq. (1) is proposed. In particular, the regularizer assumes the following form

$$\mathsf{AVG}(\mathsf{r}) = \frac{1}{d} \left\| \frac{1}{|\mathcal{D}^1|} \sum_{(Z,y)\in\mathcal{D}^1} \mathsf{r}(Z) - \frac{1}{|\mathcal{D}^2|} \sum_{(Z,y)\in\mathcal{D}^2} \mathsf{r}(Z) \right\|_1, \tag{11}$$

where $\|\cdot\|_1$ is the Manhattan norm, which means that the average representation output, conditioned on the sensitive features, should be the same independently from the sensitive features. Note that Eq. (11) is the first order approximation of Eq. (1).

This second and third regularizers are theoretically studied in [33] and need some preliminary definitions to be defined. Let $\mathcal{P}(\mathcal{Z})$ be the set of probability measures on $\mathcal{Z}$. Let us define a metric as a function mapping $\mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to \mathbb{R}$. Let $\mathsf{K} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a positive definite kernel and let $\Psi : \mathcal{Z} \to \mathbb{H}$ be the corresponding feature map [137] where $\mathbb{H}$ is the corresponding Hilber space. If $P, Q \in \mathcal{P}(\mathcal{Z})$ we can define a metric, called squared Maximum Mean Discrepancy (MMD) [65, 66], relative to the kernel K as

$$\mathsf{MMD}(P, Q) = \|\mathbb{E}_{Z\sim P}\Psi(Z) - \mathbb{E}_{Z\sim Q}\Psi(Z)\|^2_{\mathbb{H}}. \tag{12}$$

Moreover, the Optimal Transport (OT) problem with entropic regularization (ϵ) is defined as [138]

$$\mathsf{OT}_\epsilon(P, Q) = \min_{\pi\in\Pi(P,Q)} \int_{\mathcal{Z}^2} \|z_1 - z_2\|^2 d\pi(z_1, z_2) + \epsilon\mathsf{KL}(\pi|P \otimes Q), \epsilon \geq 0 \tag{13}$$

where $\mathsf{KL}(\pi|P\otimes Q)$ is the Kullback-Leibler divergence [139] between the candidate transport plan π and the product distribution $P \otimes Q$, and $\Pi(P, Q) = \{\pi \in \mathbb{P}(\mathcal{Z} \times \mathcal{Z})|\pi_1 = P, \pi_2 = Q\}$, with π_1 and π_2 the marginals of π. The case $\epsilon = 0$ corresponds to the classic Optimal Transport problem introduced by Kantorovich [140]. Sinkhorn divergence SNK is defined as

$$\mathsf{SNK}(P, Q) = \mathsf{OT}_\epsilon(P, Q) - \frac{1}{2}\mathsf{OT}_\epsilon(P, P) - \frac{1}{2}\mathsf{OT}_\epsilon(Q, Q) \tag{14}$$

and was shown in [141] to be non-negative, biconvex and to metrize the convergence in law under mild assumptions. The Sinkhorn divergence is a fast approximation of the Wasserstein distance which is a quite well suited way of e to quantify how much two probability densities differ. The Wasserstein distance has appealing geometrical properties but it also raises important statistical and computational challenges [138, 142, 143]. Sinkhorn divergence is one of the state-of-the-art approaches that allows to overcame such challenges [68]. Note that when $\epsilon \to 0$ $\mathsf{SNK}(P, Q)$ converges to the Wasserstein distance [68]. Note also that when $\epsilon \to \infty$, $\mathsf{SNK}(P, Q)$ converges to $\mathsf{MMD}(P, Q)$-like distance [141, 144].

Surely other metrics could be exploited, but the best known metrics cannot be effectively adopted in our work for reasons due to numerical or geometrical problems. For example the Rényi divergence and Kullback-Leibler divergence which show good statistical properties, including convexity and continuity [139], can assume a value up to infinite, which make them hard to exploit in practice, and are not able to well describe some properties, namely they are not symmetric and do not satisfy the triangle inequality [139].

Now that we have defined these two metrics, we can show how to exploit them to define a regularizer F able to impose the fairness constraint of Eq. (1). First, note that the fairness constraint of Eq. (1) forces the distribution of the representation vector to be the same no matter the value of the sensitive feature. This means that the distribution of the representation vector when $s = 1$, namely P, should be equal to the one when $s = 2$, namely Q. In other words $\mathsf{MMD}(P, Q) = 0$ or alternatively $\mathsf{SNK}(P, Q) = 0$. Unfortunately, P and Q are unknown but, thanks to $\mathcal{D}^1$ and $\mathcal{D}^2$, we have the corresponding empirical distributions $\hat{P}$ and $\hat{Q}$

$$\hat{P}(\cdot) = \frac{1}{|\mathcal{D}^1|} \sum_{Z \in \mathcal{D}^1} \delta[\mathsf{r}(Z) - \ \cdot\ ], \quad \hat{Q}(\cdot) = \frac{1}{|\mathcal{D}^2|} \sum_{Z \in \mathcal{D}^2} \delta[\mathsf{r}(Z) - \ \cdot\ ],$$

where δ is the Dirac delta function. Then we can impose $\mathsf{MMD}(\hat{P}, \hat{Q}) = 0$ or $\mathsf{SNK}(\hat{P}, \hat{Q}) = 0$ which, as described above, is not the most effective way of imposing the constraint since we risk to overfit our data. So we impose to $\mathsf{MMD}(\hat{P}, \hat{Q})$ or $\mathsf{SNK}(\hat{P}, \hat{Q})$ to be small, or, in other words, we can set $\mathsf{F}(\mathsf{r}) = \mathsf{MMD}(\hat{P}, \hat{Q})$ or $\mathsf{F}(\mathsf{r}) = \mathsf{SNK}(\hat{P}, \hat{Q})$ in Problem (9).

Note that $\mathsf{MMD}(\hat{P}, \hat{Q})$ can be easily computed noting that

$$\begin{aligned}\mathsf{MMD}(\hat{P}, \hat{Q}) =& \frac{1}{|\mathcal{D}^1|^2} \sum_{Z_a \in \mathcal{D}^1} \sum_{Z_b \in \mathcal{D}^1} \mathsf{K}(\mathsf{r}(Z_a), \mathsf{r}(Z_b)) \\ &+ \frac{1}{|\mathcal{D}^2|^2} \sum_{Z_a \in \mathcal{D}^2} \sum_{Z_b \in \mathcal{D}^2} \mathsf{K}(\mathsf{r}(Z_a), \mathsf{r}(Z_b)) \\ &- 2\frac{1}{|\mathcal{D}^1||\mathcal{D}^2|} \sum_{Z_a \in \mathcal{D}^1} \sum_{Z_b \in \mathcal{D}^2} \mathsf{K}(\mathsf{r}(Z_a), \mathsf{r}(Z_b)) \end{aligned} \tag{15}$$

while the computation of $\mathsf{SNK}(\hat{P}, \hat{Q})$ is a bit more complex since we have to solve the optimization problem reported in Eq. (13) for three cases $\mathsf{OT}_\epsilon(\hat{P}, \hat{Q})$,

$\mathsf{OT}_\epsilon(\hat{P}, \hat{P})$, and $\mathsf{OT}_\epsilon(\hat{Q}, \hat{Q})$. The solution of the optimization problem, because of its convexity, can be solved iteratively using the Sinkhorn iterations [67].

6. Experimental Results

In this section we will presents the results of applying the methodology presented in Section 5 on the DNNs presented in Section 4 on real world datasets.

6.1. Results on Graphs Datasets

In this section we will presents the results of applying the methodology presented in Section 5 on the DNN for graphs presented in Section 4.1 on two real-world social network datasets: Pokec [73] and Marvel. The experiments reported in this subsection were deployed on machines running Ubuntu 18.04.5 (LTS) OS equipped with 2 Intel Xeon E5-2650 @2.30 GHz CPUs and 160 GB of RAM. Our experiments have been coded in Python 3.8.5, and are based on Deep Graph Library[5] 0.5.3 and the PyTorch [145] 1.7 framework. Note that we did not use GPUs for these experiments.

6.1.1. The Datasets

While during the last few years the number of popular on-line social networks has been steadily increasing, it is hard to access real-world data from such social networks for research purposes. Pokec is the most popular on-line social network in Slovakia. Its popularity has not changed even after the rise of Facebook. Pokec released an anonymized version of the data of the whole network, including user profiles and connections. We consider gender as the sensitive attribute and marital status as the target (in this work, simplified in the binary attribute single/in-a-serious-relationship). Our dataset, after removing users with missing data, comprehends a total of $361,450$ users: $184,862$ males and $176,588$ females. Table 1 reports the statistics of this dataset.

Given the lack of other similar resources in the literature, we decided to study another kind of social network that closely resemble real-world ones without posing any concern on user's privacy: the Marvel universe [74]. To mimic a real-world social network, we need both the graph structure and

[5] *https://www.dgl.ai*

	Single	**In a relationship**	*sensitive marginals*
Females	35.44 % 128117	13.41 % 48471	48.86 % 176588
Males	41.58 % 150293	9.56 % 34569	51.14 % 184862
class marginals	77.02 % 278410	22.98 % 83040	361450

Table 1: Pokec dataset labels distribution when gender is chosen as sensitive feature and marital status as target.

some features associated to the users. To this end, we built a novel dataset merging two existing ones:

- *The Marvel Universe* [74] that encodes the Marvel characters as nodes and connects them if they appear in at least one comic together. This dataset, however, does not provide any information about the characters beside their name;
- Data behind the story *Comic Books Are Still Made By Men, For Men, and About Men*[75]. The dataset contains information about Marvel and DC characters, including their gender, alignment (good, bad, neutral) and other information.

We matched the characters in the two datasets using fuzzy string matching, obtaining a social network where users have associated gender and alignment information. While there may be a small amount of errors in the automatic character matching between the two datasets, this is not relevant for our purposes since it may be resembled to the noise intrinsically present in social networks. We consider gender as the sensitive attribute, while our task will be to predict the character's alignment (for simplicity, we considered only good and bad characters, removing neutral ones). The dataset is composed by $2,612$ characters: 745 females and $1,867$ males. The dataset statistics are reported in Table 2.

We split both datasets considering 50% of the nodes as the training set and the other half as the test set.

6.1.2. How fair is the learned representation using different constraints?

In this section we evaluate the effectiveness of the different regularizers in terms of effects on both the final Area Under ROC curve AUROC_y and fairness DDP on the test set. The reference case is always when no fairness regularizer (NOR) is introduced (namely $\lambda = 0$). Each constraint is applied at either one of three different network representation layers: the first graph

	Bad	**Good**	*sensitive marginals*
Females	10.00 % 261	18.52 % 484	28.52 % 745
Males	44.64 % 1166	26.84% 701	71.48 % 1867
class marginals	54.64 % 1427	45.36 % 1185	2612

Table 2: Marvel dataset labels distribution when gender is chosen as sensitive feature and alignment as target.

convolutional layer (FCL), the second (and last) graph convolutional layer (LCL), and the output layer (OUT). As previously described in Sections 1 and 3, in order to create a fair deep model we can impose the constraint just on the output or in one or more of the deeper layers. Specifically, the latter allows one to extract a data representation able to automatically deliver fair models. For our experiments, we applied the constraints in a subset of all the many possibilities since all the possibilities cannot be explored due to space constraints. This subset is still general enough to be applied also in other applications, as well as architectures. We train the graph neural network for 10 epochs, and we report the mean of 10 repetitions of the experiment. We sample 25 neighbors for each GraphSAGE layer. We use the Adam optimizer, 64 neurons for each graph convolutional layer and a batch size of 512.

Let us now analyze our results starting from the Pokec dataset. Figure 1a reports the AUROC_y against the DDP for the different constraints (AVG, SNK, and MMD) applied on the different layers of the Graph DNN (NOR, FCL, LCL, and OUT) when different values of λ are exploited. Figure 1a clearly shows the effectiveness of the proposed approaches in learning fair models. Each constraint forces the network to discard an increasing amount of sensitive information as the regularization parameter λ (Eq. (9)) strengthen, resulting in fairer but less accurate predictions. Note that all constraints works quite well but, in our experimental setting AVG and MMD resulted the most effective one, significantly improving the DDP (obtaining a value < 0.02 compared to 0.11 of NOR) without compromising the AUROC_y (losing just around 2% performance w.r.t. NOR). The SNK method, while slightly less effective compared to the others, still performs quite well when the fairness constraint is applied to the first graph convolution layer (FCL). Let us now consider the Marvel dataset. Figure 1b reports, similarly as before, the AUROC_y against the DDP for the different constraints and applying the fair-

ness constraint on different layers of the Graph DNN. Note that in this case, the model without fairness constraints (NOR) is very unfair, with a DDP of 0.64. While also in this case we can obtain substantially fairer models with all the considered fairness constraints, AVG provides consistently good results no matter where we insert the fairness constraint in the network. SNK constraint shows good performance as well, while being slightly less effective than AVG. MMD shows good performance with FCL and LCL, while OUT seems to be less effective in reducing the DDP.

6.1.3. Is the fair representation able to "forget" the sensitive attribute?

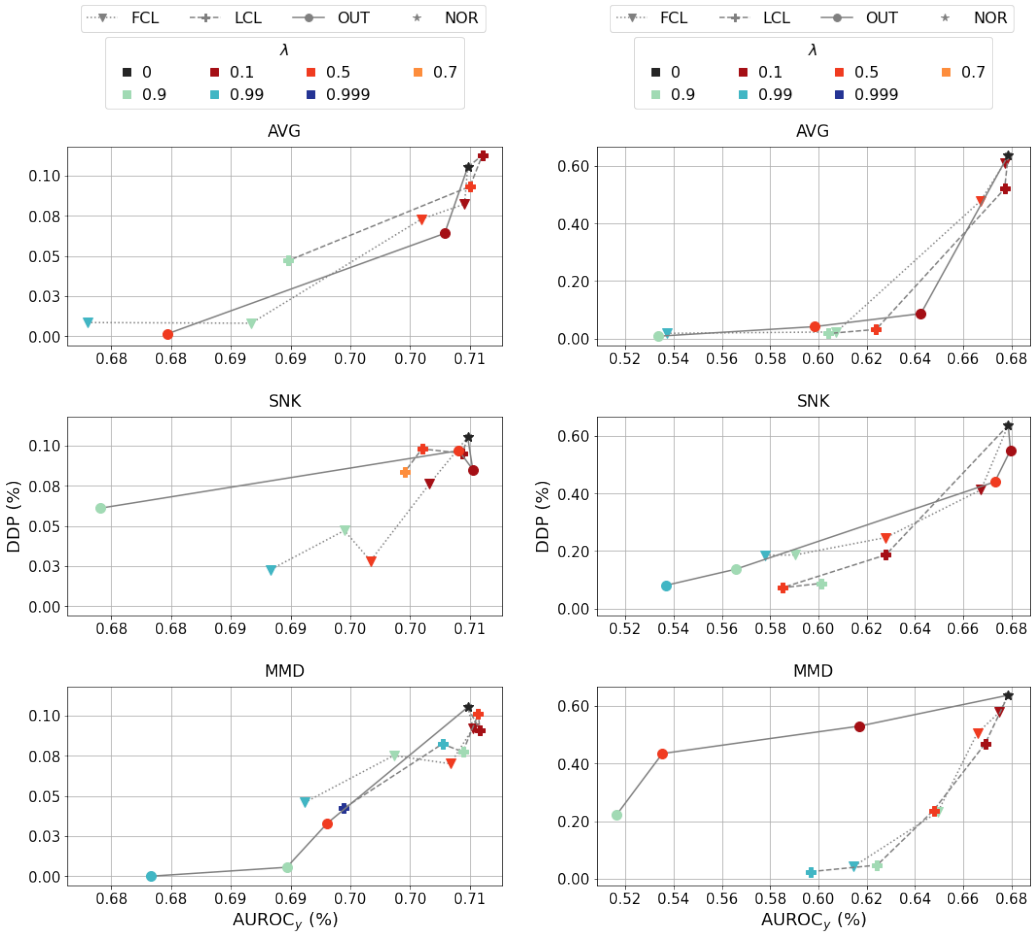
In this section, we focus on a setting similar to the one in previous section but instead of comparing the AUROC_y against the fairness DDP, we will compare AUROC_y against the performance in reconstructing the sensitive feature AUROC_s. In other words, we will test how much the fair representation is able to "forget" the sensitive attribute. To compute AUROC_s, we train an SVM on the hidden representation and we optimize the SVM regularization hyperparameter via a 5-fold cross validation on the training set.

Figure 2 reports the AUROC_y against the AUROC_s for the different constraints (AVG, SNK, and MMD) applied on the different layers of the Graph DNN (NOR, FCL, LCL, and OUT) when different values of λ are exploited for both the analyzed graphs datasets.

Figure 2 shows that all the fairness constraints work quite well in reducing the AUROC_s. On both datasets, AVG and MMD achieve an AUROC_s close to 0.5 without much reducing the AUROC_y. SNK, while performing well, cannot achieve AUROC_s as low as the other two methods. Moreover, we can observe that increasing the λ parameter (thus the weight assigned to the fairness constraint in the loss) always results in a decrease on the reconstruction of the sensitive feature AUROC_s. Note that by increasing the λ parameter, all methods are able to achieve an AUROC_s of 0.5. However, if the resulting AUROC_y was too low, we decided not to report them to not impact the readability of the plots, since those models are not interesting due to the low predictive performance. In general, we can conclude that the models that exhibited good fairness values in the previous section also encode few information about the sensitive attribute in the representation they learn.

6.2. Results on Face Recognition Datasets

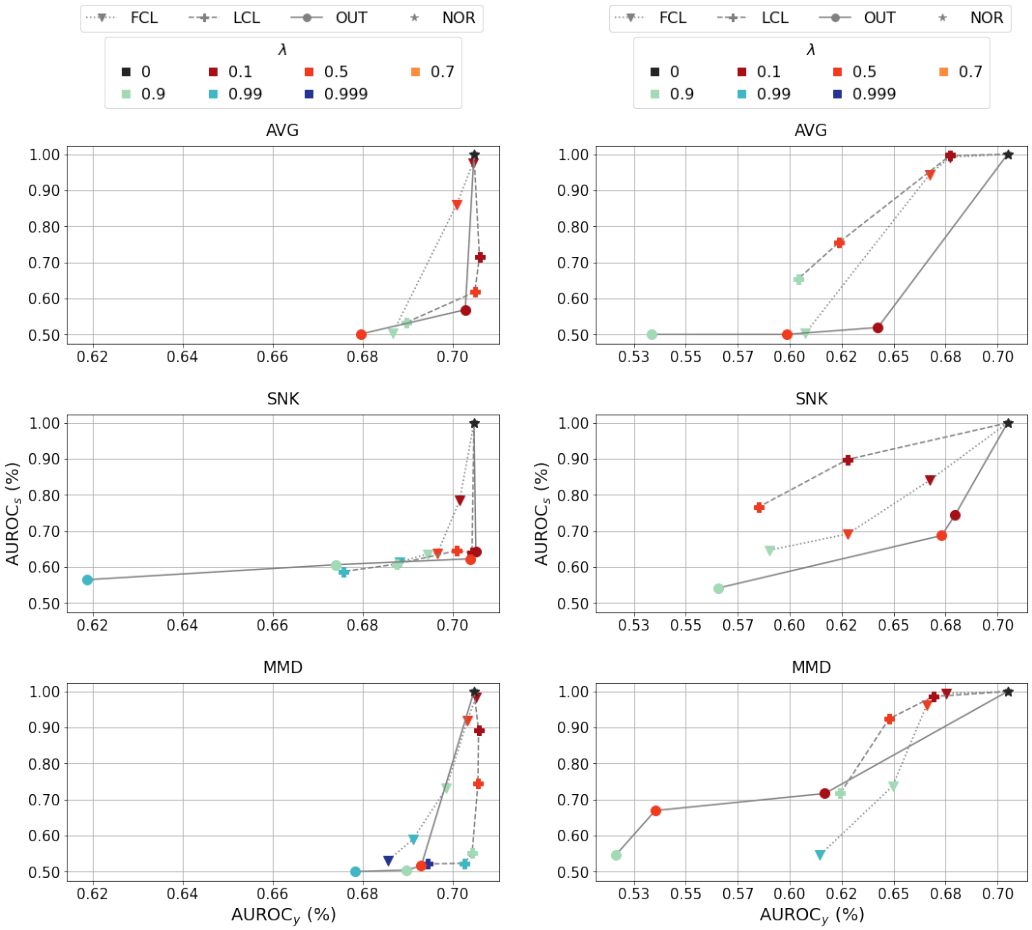
In this section we present the results of applying the methodology presented in Section 5 to the DNN for Face Recognition presented in Section 4.2



(a) Pokec Dataset

(b) Marvel Dataset

Figure 1: Graph DNNs on the Pokec and Marvel Datasets: AUROC$_y$ against the DDP for the different constraints (AVG, SNK, and MMD) applied on the different layers of the DNN (NOR, OUT, FCL, and LCL) when different values of λ are exploited.



(a) Pokec Dataset

(b) Marvel Dataset

Figure 2: Graph DNNs on the Pokec and Marvel Datasets: AUROC$_y$ against the AUROC$_s$ for the different constraints (AVG, SNK, and MMD) applied on the different layers of the DNN (NOR, OUT, FCL, and LCL) when different values of λ are exploited.

on the real-world FairFace dataset [72]. These experiments are deployed on an Ubuntu 18.04.5 (LTS) OS-based server equipped with an Intel Xeon @2.30GHz dual core CPU, 12GB of RAM and one NVidia Tesla T4 GPU. Experiments have been coded in Python 3.7 leveraging on the PyTorch 1.7 framework.

6.2.1. The Dataset

FairFace dataset [72] is a collection of ≈100 thousand facial images extracted from the YFCC-100M Flickr dataset [146]. It also provides age group[6], gender[7], and ethnicity[8]. Gender and Ethnicity can be used as sensitive attributes. Our task consists in predicting whether a face belongs to a person with more (+1) or less (-1) than 30 years old adopting binary sex (Females and Males) or selecting two categories from ethnicity (in our experiments, Blacks and Western Whites). Table 3 reports some statistics about the FairFace dataset when gender or ethnicity are chosen as sensitive features. The training and test sets are composed of 86.7 thousand and 10.9 thousand images respectively (same split as in the original paper [72]) when gender is exploited as sensitive feature, while they will be composed of 28.7 thousand and 3.6 thousand respectively when considering the binary ethnicity (Blacks and Whites) sensitive feature.

6.2.2. How fair is the learned representation using different constraints?

In this section, analogously to what have been done for the DNNs for Graphs, we evaluate the effectiveness of the different regularizers in terms of effects on both the final accuracy ACC_y and fairness DDP on the test set. The reference case is always when no fairness regularizer (NOR) is introduced (namely $\lambda = 0$). Each constraint is applied at either one of three different network layers: the output layer (OUT), the first dense layer (FDL), and the last convolution layer (LCL). The same considerations that we did for the graph labelling experiments also apply here: also in this case we chose to apply the fairness regularizers in a subset of all the many possibilities, which cannot be explored due to space constraints, that is general enough to

[6]The groups are [0-2], [3-9], [10-19], [20-29], [30-39], [40-49], [50-59], [60-69], and [70+].

[7]For this dataset, the attribute gender refers to the perceived binary phisical sex (Male and Female) of an individual.

[8] The different ethnicities are Western White, Middle Eastern White, East Asian, Southeast Asian, Black, Indian, and Latinx.

	Age ≥ 30	**Age** < 30	*sensitive marginals*
Females	18.60 % 18174	28.40 % 27746	47.00 % 45920
Males	27.21 % 26587	25.79 % 25191	53.00 % 51778
class marginals	45.82 % 44761	54.18 % 52937	97698

(a) Sensitive features: Gender

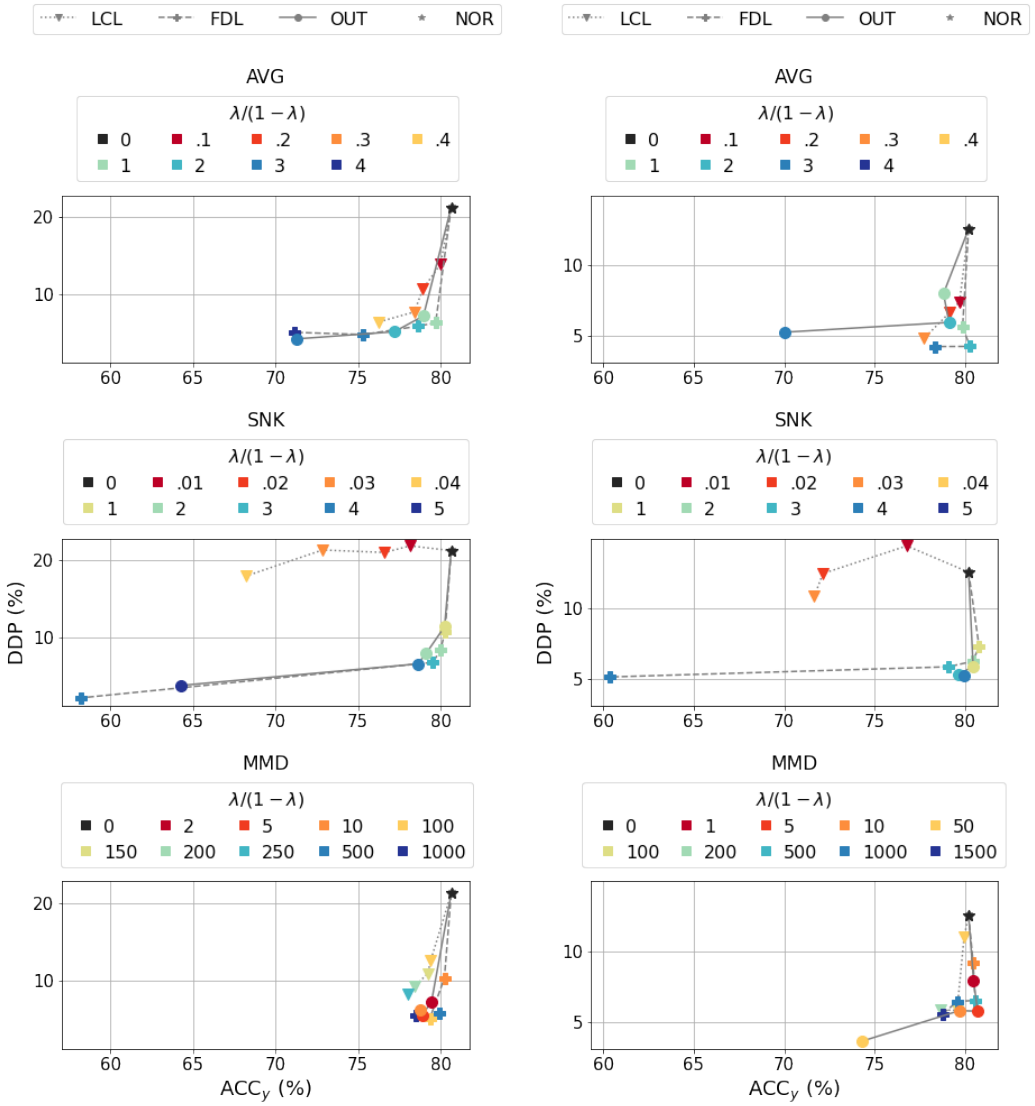
	Age ≥ 30	**Age** < 30	*sensitive marginals*
Blacks	18.1 % 5862	24.5 % 7927	42.6 % 13789
Whites	29.5 % 9559	27.9 % 9053	57.4 % 18612
class marginals	47.6 % 15421	52.4 % 16980	32401

(b) Sensitive features: Ethnicity

Table 3: Fairface dataset labels distribution when gender or ethnicity are chosen as sensitive features.

be exploited also in other architectures and applications. Each experimental run exploits a random selection of 20 thousand training and 10 thousand test images from the training and test sets respectively. We train every model for a total of 10 epochs using the ADADELTA [129] with mini batches of 200 images. The layers until the last convolution one excluded have not been fine-tuned.

Figure 3 reports the ACC_y against the DDP for the different constraints (AVG, SNK, and MMD) applied on the different layers of the DNN for for Face Recognition (NOR, OUT, FDL, and LCL) when different values of λ are exploited. Both the cases when gender and ethnicity are considered as sensitive feature are reported. Figure 3 clearly shows the effectiveness of the proposed approaches in learning fair models. Each constraint forces the network to discard an increasing amount of sensitive information as the regularization parameter λ (Eq. (9)) strengthen, resulting in fairer but less accurate predictions. Note that all constraints work quite well but, in our experimental setting MMD resulted to be the most effective one, namely the one which improves more the DDP without compromising the ACC_y. SNK resulted to be the worst performing method (especially when applied to LCL) while AVG performs averagely well.



(a) Gender as the sensitive attribute

(b) Ethnicity as the sensitive attribute

Figure 3: DNNs for Face Recognition on the FairFace dataset: ACC_y against the DDP for the different constraints (AVG, SNK, and MMD) applied on the different layers of the DNN (NOR, OUT, FDL, and LCL) when different values of λ are exploited.

6.2.3. Is the fair representation able to "forget" the sensitive attribute?

In this section, we focus on the very same setting of the previous section but, instead of comparing the final accuracy ACC_y against the fairness DDP, we will compare ACC_y against the sensitive accuracy ACC_s.

Figure 4 reports the ACC_y against the ACC_s for the different constraints (AVG, SNK, and MMD) applied on the different layers of the DNN for for Face Recognition (NOR, OUT, FDL, and LCL) when different values of λ are exploited. Both the cases when gender and ethnicity are considered as sensitive features are reported. Figure 4 clearly shows the effectiveness of the proposed approaches in forgetting the sensitive feature from the representation. Each constraint forces the network to forget an increasing amount of sensitive information as the regularization parameter λ strengthen, resulting in fairer but less accurate predictions. Also in this case, all constraints work quite well, analogously to what have been seen for ACC_y against the DDP except for the case when the constraint is applied on FDL and OUT. In these cases, the effect on the representation is less evident due to the fact that we are not imposing the constraint directly on the LCL (the representation layer). This was expected from theory and in fact the effect here is quite evident.

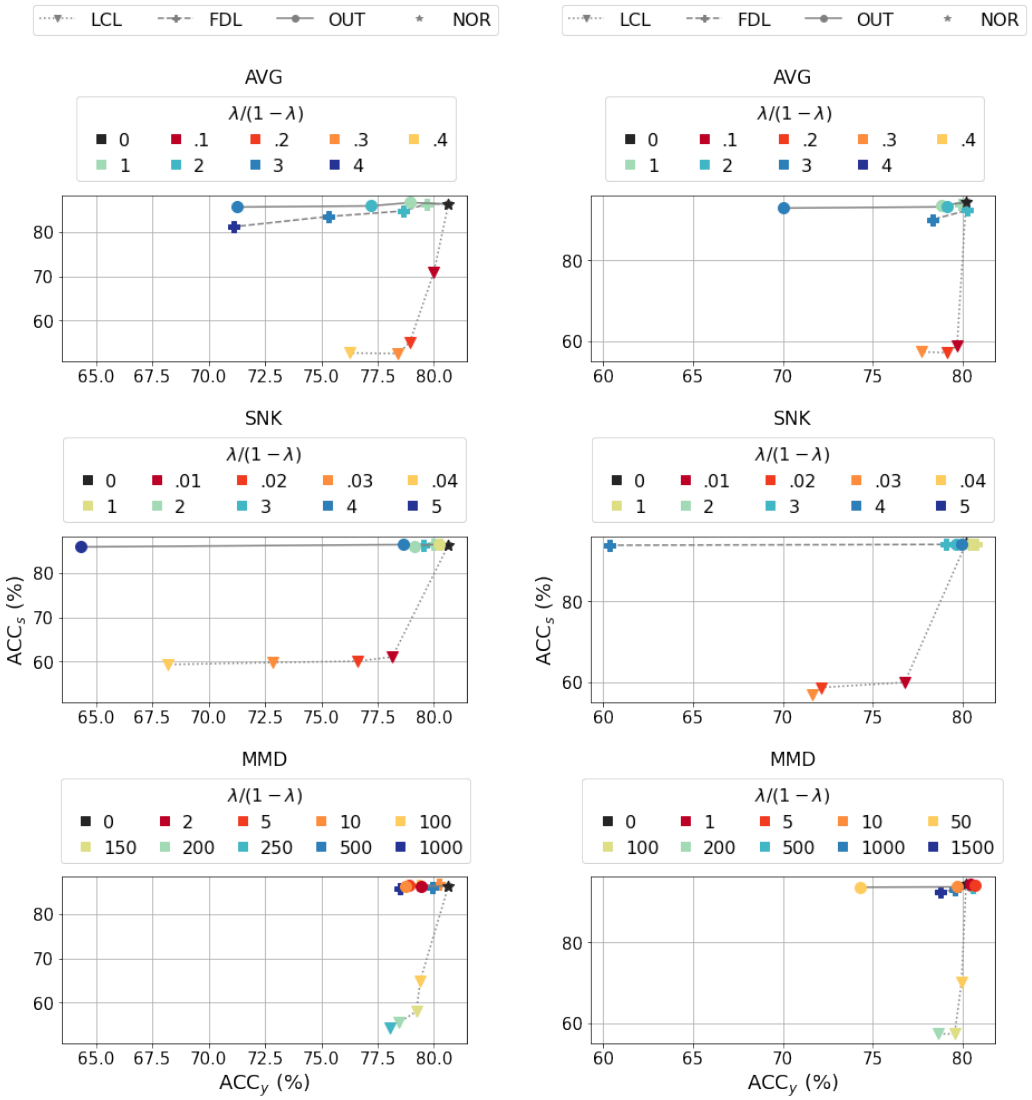
6.2.4. Visualizing the Effects of the Fairness Constraints by means of Visual Explanation

In this section, we aim at assessing a possible discriminatory attention behavior carried out by the DNNs for face recognition (see Sections 4.2) and observe whether the application of different fairness regularizers produces less discriminatory attention mechanisms.

In our visualization experiments, in order to standardize the image face regions, we exploit a set of 50 thousand images of frontal faces extracted from the Diversity in Faces dataset as proposed in [147].

Firstly, we want to analyze the dataset average attention map to assess whether the trained DNNs show any discriminatory attention behaviors. Hence, for each face we extract the attention map corresponding to the LCL using Grad-CAM. Then, we take the average attention map of both males and females. Finally we compute the difference between these two average attention maps through the Frobenius distance [148].

More formally, for each image in the dataset we compute L_Y (see Section 4.2.1). Then, we define $M_s \in \mathbb{R}^{U \times V}$, with $s \in \{\text{males}, \text{females}\}$, as the averaged L_Y for each subgroup (males, females) in the populations and



(a) Gender as the sensitive attribute

(b) Ethnicity as the sensitive attribute

Figure 4: DNNs for Face Recognition on the FairFace dataset: ACC_y against the ACC_s for the different constraints (AVG, SNK, and MMD) applied on the different layers of the DNN (NOR, OUT, FDL, and LCL) when different values of λ are exploited.

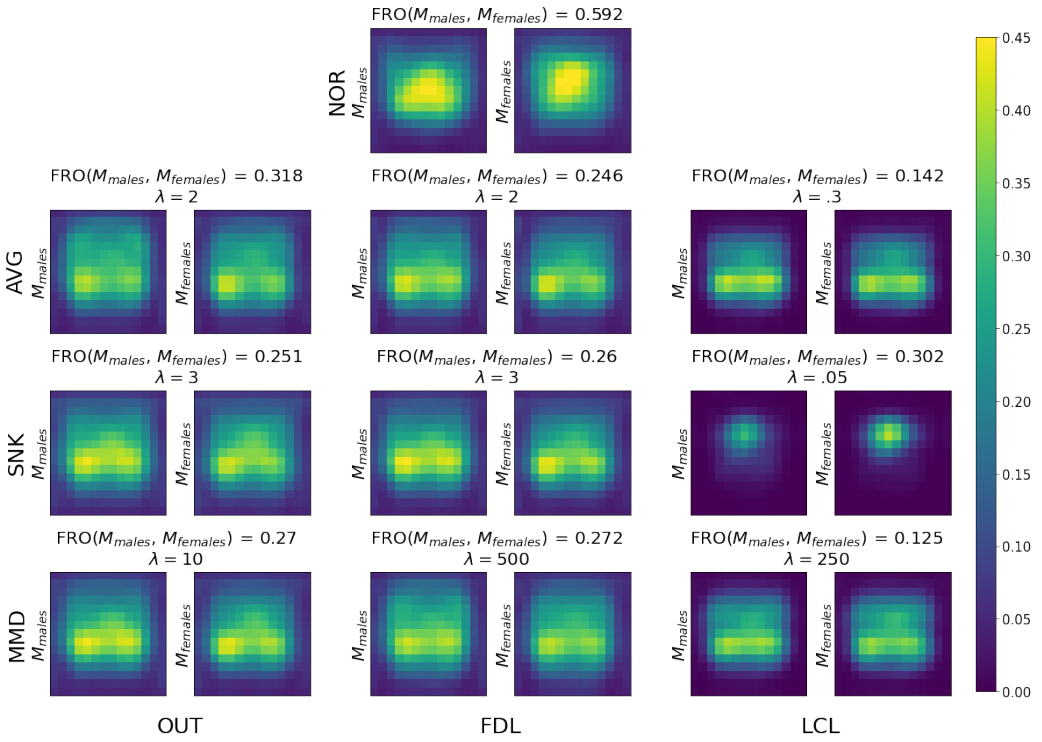


Figure 5: M_{males}, M_{females}, and $\mathsf{FRO}(M_{\text{males}}, M_{\text{females}})$ for the different constraints (AVG, SNK, and MMD) applied on the different layers of the DNN for for Face Recognition (NOR, OUT, FDL, and LCL) for λ which showed the best accuracy/fairness trade-off (i.e. the best DDP allowing a maximum of 5% of loss in ACC_y).

compute thee Frobenius distance of M_{males} and M_{females}

$$\mathsf{FRO}(M_{\text{males}}, M_{\text{females}}) = \sqrt{\sum_{i=1}^{U}\sum_{j=1}^{V}(M_{males,i,j} - M_{females,i,j})^2} \qquad (16)$$

Figure 5 reports M_{males}, M_{females}, and $\mathsf{FRO}(M_{\text{males}}, M_{\text{females}})$ for the different constraints (AVG, SNK, and MMD) applied on the different layers of the DNN for for Face Recognition (NOR, OUT, FDL, and LCL) for λ which showed the best accuracy/fairness trade-off (i.e. the best DDP allowing a maximum of 5% of loss in ACC_y).

Figure 5 clearly shows the positive effect of each one of the different regularizers in reducing the networks' discriminatory attention mechanism, which is quite evident when no fairness regularizer is imposed.

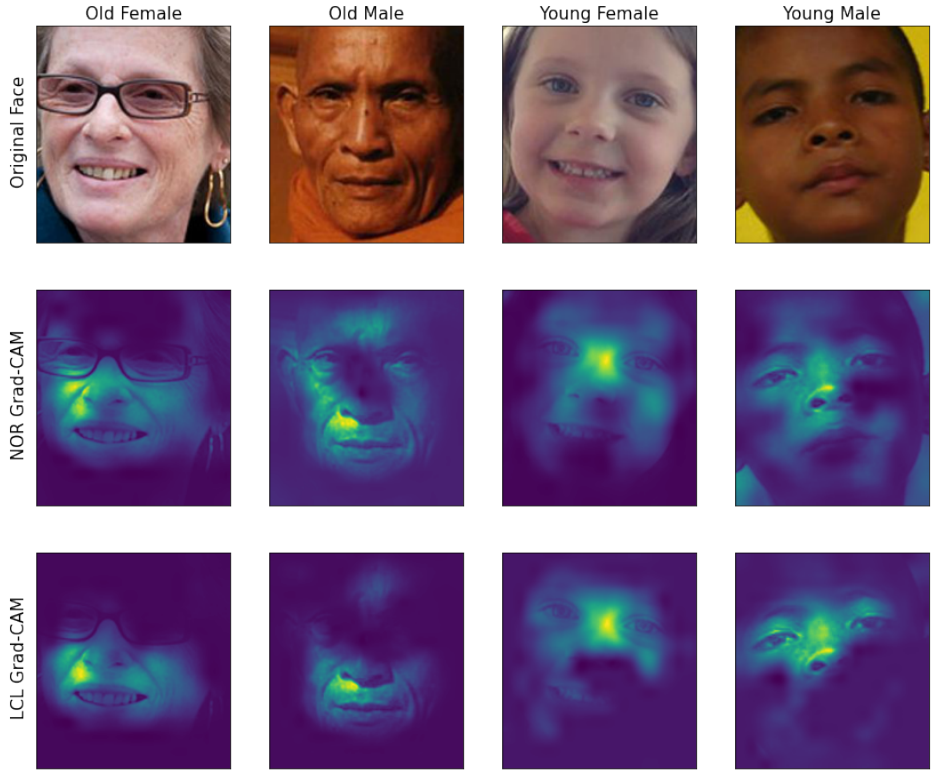


Figure 6: Examples of attention maps for a young ($Y = -1$) male, for a young ($Y = -1$) female, an old ($Y = 1$) male, and for an old ($Y = 1$) female before (NOR) and after (LCL) the application of the fairness constraint. Results show the use of the MMD constraint with $\lambda = 250$ which showed the best accuracy/fairness trade-off.

For sake of completeness, we also report in Figure 6 for a young ($Y = -1$) male, for a young ($Y = -1$) female, an old ($Y = 1$) male, and for an old ($Y = 1$) female their attention map before (NOR) and after (LCL) the application of the fairness constraint. Due to space limitations, we report just the results with the MMD and $\lambda = 250$ which show the best accuracy/fairness trade-off.

Figure 6 shows how the fairness regularizer is able to restrict the DNN receptive field to class-specific face regions. These face areas can present distinctive traits for age related tasks: a network activation in the eyes region is observed as an indicator for the negative class (age less than 30), while a strong activation on the skin portion below the nose, on the cheeks, and around the mouth represents a clear trait of the positive class (arguably, these

area are the most affected from seniority-related markers, as the presence of wrinkles or beard).

7. Conclusions

In this work, we focused our attention on the problem of algorithmic fairness, namely developing AI-based systems which do not discriminate with respect to one or multiple notions of inequity. A challenge exists between the race for superior performance in AI-based systems and the effort to not inherit also the human biases hidden in the data.

We considered a particularly challenging task: how to make models for structured input data (graphs and images) fairer. We addressed this issue by means of learning fairer representations that are on the one hand expressive enough to well describe the data and lead to highly accurate models, while on the other hand are simultaneously able to discard the information which may lead to unfair behaviors. Exploiting the fairness notion of Demographic Parity, we investigate how to impose these fairness constraints in the different layers of deep neural networks for complex data through the use of different regularizers.

We present experiments on different real-world datasets, showing the effectiveness of our proposal both quantitatively by means of accuracy and fairness metrics and qualitatively by means of visual explanation.

In the future, we plan to extend our work to a larger number of architectures and datasets, providing more insights and guidelines on the best practice of building fairer models for complex input data. Moreover, we will investigate the possibility of including human oriented requirements, such as robustness and privacy.

Acknowledgments

This work was partially supported by Amazon Web Services.

References

[1] X. He, K. Zhao, X. Chu, Automl: A survey of the state-of-the-art, arXiv preprint arXiv:1908.00709.

[2] L. Tuggener, M. Amirian, K. Rombach, S. Lörwald, A. Varlet, C. Westermann, T. Stadelmann, Automated machine learning in practice: state of the art and recent results, in: Swiss Conference on Data Science, 2019.

[3] P. Das, N. Ivkin, T. Bansal, L. Rouesnel, P. Gautier, Z. Karnin, L. Dirac, L. Ramakrishnan, A. Perunicic, I. Shcherbatyi, W. Wu, A. Zolic, H. Shen, A. Ahmed, F. Winkelmolen, M. Miladinovic, C. Archembeau, A. Tang, B. Dutt, P. Grao, K. Venkateswar, Amazon sagemaker autopilot: a white box automl solution at scale, in: International Workshop on Data Management for End-to-End Machine Learning, 2020.

[4] Gartner, Two megatrends dominate the gartner hype cycle for artificial intelligence, 2020, `https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020/`, accessed: 2020-11-2.

[5] A. F. Winfield, K. Michael, J. Pitt, V. Evers, Machine ethics: the design and governance of ethical ai and autonomous systems, Proceedings of the IEEE 107 (3) (2019) 509–517.

[6] C. Allen, G. Varner, J. Zinser, Prolegomena to any future artificial moral agent, Journal of Experimental & Theoretical Artificial Intelligence.

[7] M. Anderson, S. L. Anderson, Geneth: A general ethical dilemma analyzer, Paladyn, Journal of Behavioral Robotics 12 (3) (2018) 251–261.

[8] L. Oneto, S. Chiappa, Fairness in machine learning, in: Recent Trends in Learning From Data, 2020.

[9] R. Agrawal, R. Srikant, Privacy-preserving data mining, in: ACM SIGMOD International Conference on Management of Data, 2000.

[10] D. Gunning, Explainable artificial intelligence (xai), Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2).

[11] B. Biggio, F. Roli, Wild patterns: Ten years after the rise of adversarial machine learning, Pattern Recognition 84 (2018) 317–331.

[12] P. DiMaggio, J. Evans, B. Bryson, Have american's social attitudes become more polarized?, American journal of Sociology 102 (3) (1996) 690–755.

[13] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, B. Nyhan, Social media, political polarization, and political disinformation: A review of the scientific literature, in: SSRN, 2018.

[14] Muskaan, M. P. Dhaliwal, A. Seth, Fairness and diversity in the recommendation and ranking of participatory media content, arXiv preprint arXiv:1907.07253.

[15] Scientific American, Why social media makes us more polarized and how to fix it, https://www.scientificamerican.com/article/why-social-media-makes-us-more-polarized-and-how-to-fix-it/, accessed: 2020-11-2.

[16] M. D. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, A. Flammini, Political polarization on twitter, International AAAI Conference on Weblogs and Social Media 133 (26) (2011) 89–96.

[17] H. A. Prasetya, T. Murata, A model of opinion and propagation structure polarization in social media, Computational Social Networks 7 (1) (2020) 1–35.

[18] A. Bessi, F. Zollo, M. Del Vicario, M. Puliga, A. Scala, G. Caldarelli, B. Uzzi, W. Quattrociocchi, Users polarization on facebook and youtube, PloS one 11 (8) (2016) e0159641.

[19] New York Times, A case for banning facial recognition, https://www.nytimes.com/2020/06/09/technology/facial-recognition-software.html, accessed: 2020-11-2.

[20] J. Buolamwini, T. Gebru, Gender shades: intersectional accuracy disparities in commercial gender classification, in: Conference on Fairness, Accountability and Transparency, 2018.

[21] I. D. Raji, J. Buolamwini, Actionable auditing: investigating the impact of publicly naming biased performance results of commercial ai products, in: AAAI/ACM Conference on AI Ethics and Society, 2019.

[22] The Verge, A black man was wrongfully arrested because of facial recognition, https://www.theverge.com/2020/6/24/21301759/facial-recognition-detroit-police-wrongful-arrest-robert-williams-artificial-intelligence, accessed: 2020-11-2.

[23] CNN, Portland passes broadest facial recognition ban in the us, https://edition.cnn.com/2020/09/09/tech/portland-facial-recognition-ban/index.html, accessed: 2020-12-29.

[24] A. Romei, S. Ruggieri, A multidisciplinary survey on discrimination analysis, The Knowledge Engineering Review 29 (5) (2014) 582–638.

[25] S. Barocas, A. D. Selbst, Big data's disparate impact, California Law Review 104 (2016) 671.

[26] L. Oneto, Learning fair models and representations, Intelligenza Artificiale 14 (1) (2020) 151–178.

[27] T. Calders, F. Kamiran, M. Pechenizkiy, Building classifiers with independency constraints, in: IEEE International Conference on Data mining, 2009.

[28] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Advances in Neural Information Processing Systems, 2016.

[29] S. Verma, J. Rubin, Fairness definitions explained, in: IEEE/ACM International Workshop on Software Fairness, 2018, pp. 1–7.

[30] J. M. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, in: Innovations in Theoretical Computer Science Conference, 2017.

[31] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big data 5 (2) (2017) 153–163.

[32] S. Chiappa, W. S. Isaac, A causal bayesian networks viewpoint on fairness, in: Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data, 2018.

[33] L. Oneto, M. Donini, G. Luise, C. Ciliberto, A. Maurer, M. Pontil, Exploiting mmd and sinkhorn divergences for fair and transferable representation learning, in: Advances in Neural Information Processing Systems, 2020.

[34] D. Madras, E. Creager, T. Pitassi, R. Zemel, Learning adversarially fair and transferable representations, in: International Conference on Machine Learning, 2018.

[35] H. Edwards, A. Storkey, Censoring representations with an adversary, in: International Conference on Learning Representations, 2016.

[36] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: International Conference on Machine Learning, 2013.

[37] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[38] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M. S. Lew, Deep learning for visual understanding: A review, Neurocomputing 187 (2016) 27–48.

[39] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, IEEE Computational intelligenCe magazine 13 (3) (2018) 55–75.

[40] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, The rise of deep learning in drug discovery, Drug discovery today 23 (6) (2018) 1241–1250.

[41] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. Do, G. P. Way, E. Ferrero, P. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter, C. S. Greene, Opportunities and obstacles for deep learning in biology and medicine, Journal of The Royal Society Interface 15 (141) (2018) 20170387.

[42] D. Bacciu, F. Errica, A. Micheli, M. Podda, A gentle introduction to deep learning for graphs, Neural Networks.

[43] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[44] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Transactions on Image processing 11 (4) (2002) 467–476.

[45] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: Application to face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12) (2006) 2037–2041.

[46] W. Zhang, S. Shan, W. Gao, X. Chen, H. Zhang, Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition, in: IEEE International Conference on Computer Vision, 2005.

[47] M. Wang, W. Deng, Deep face recognition: A survey, Neurocomputing 429 (2021) 215–244.

[48] U. Jayaraman, P. Gupta, S. Gupta, G. Arora, K. Tiwari, Recent development in face recognition, Neurocomputing 408 (2020) 231–245.

[49] A. Hekler, J. S. Utikal, A. H. Enk, W. Solass, M. Schmitt, J. Klode, D. Schadendorf, W. Sondermann, C. Franklin, F. Bestvater, et al., Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images, European Journal of Cancer 118 (2019) 91–96.

[50] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. P. Lillicrap, F. Hui, L. Sifre, G. Van Den Driessche, T. Graepel, D. Hassabis, Mastering the game of go without human knowledge, Nature 550 (7676) (2017) 354–359.

[51] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, O. Evans, Viewpoint: When will AI exceed human performance? evidence from AI experts, J. Artif. Intell. Res. 62 (2018) 729–754.

[52] A. Zheng, A. Casari, Feature engineering for machine learning: principles and techniques for data scientists, O'Reilly Media, Inc., 2018.

[53] R. Ribani, M. Marengoni, A survey of transfer learning for convolutional neural networks, in: Conference on Graphics, Patterns and Images Tutorials, 2019.

[54] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: International Conference on Artificial Neural Networks, 2018.

[55] Y. Bengio, Deep learning of representations for unsupervised and transfer learning, in: International Conference on Unsupervised and Transfer Learning, 2011.

[56] H. W. Ng, V. D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, in: ACM International Conference on Multimodal Interaction, 2015.

[57] D. Castellana, D. Bacciu, Tensor decompositions in recursive neural-networks for tree-structured data, arXiv preprint arXiv:2006.10619.

[58] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural computation 29 (9) (2017) 2352–2449.

[59] M. O. Oloyede, G. P. Hancke, H. C. Myburgh, A review on face recognition systems: recent approaches and challenges, Multimedia Tools and Applications 79 (37) (2020) 27891–27922.

[60] T. Gärtner, A survey of kernels for structured data, ACM SIGKDD Explorations Newsletter 5 (1) (2003) 49–58.

[61] N. Navarin, L. Oneto, M. Donini, Learning deep fair graph neural networks, in: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2020.

[62] A. N. Tikhonov, V. I. A. Arsenin, F. John, Solutions of ill-posed problems, Winston Washington, DC, 1977.

[63] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, M. Pontil, Empirical risk minimization under fairness constraints, in: Advances in Neural Information Processing Systems, 2018.

[64] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A. Smola, A kernel method for the two-sample-problem, in: Advances in neural information processing systems, 2006.

[65] K. M. Borgwardt, A. Gretton, M. J. Rasch, H. P. Kriegel, B. Schölkopf, A. J. Smola, Integrating structured biological data by kernel maximum mean discrepancy, Bioinformatics 22 (14) (2006) e49–e57.

[66] L. Song, Learning via hilbert space embedding of distributions, in: PhD Thesis, 2008.

[67] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in: Advances in Neural Information Processing Systems, 2013.

[68] L. Chizat, P. Roussillon, F. Léger, F. X. Vialard, G. Peyré, Faster wasserstein distance estimation with the sinkhorn divergence, arXiv preprint arXiv:2006.08172.

[69] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.

[70] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-cam: Score-weighted visual explanations for convolutional neural networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.

[71] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, 2017.

[72] K. Kärkkäinen, J. Joo, Fairface: Face attribute dataset for balanced race, gender, and age, arXiv preprint arXiv:1908.04913.

[73] L. Takac, M. Zabovsky, Data analysis in public social networks, in: International Scientific Conference and International Workshop Present Day Trends of Innovations, 2012.

[74] R. Alberich, J. Miro-Julia, F. Rosselló, Marvel universe looks almost like a real social network, arXiv preprint cond-mat/0202174.

[75] FiveThirtyEight, Comic books are still made by men, for men and about men, 2014, https://fivethirtyeight.com/features/women-in-comic-books/, accessed: 2021-03-23.

[76] P. Lahoti, K. P. Gummadi, G. Weikum, ifair: Learning individually fair data representations for algorithmic decision making, in: IEEE International Conference on Data Engineering, 2019.

[77] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Innovations in Theoretical Computer Science Conference, 2012.

[78] C. Louizos, K. Swersky, Y. Li, M. Welling, R. S. Zemel, The variational fair autoencoder, in: International Conference on Learning Representations, 2016.

[79] F. P. Calmon, D. Wei, K. N. Ramamurthy, K. R. Varshney, Optimized data pre-processing for discrimination prevention, arXiv preprint arXiv:1704.03354.

[80] D. Moyer, S. Gao, R. Brekelmans, A. Galstyan, G. Ver Steeg, Invariant representations without adversarial training, in: Advances in Neural Information Processing Systems, 2018.

[81] P. Botros, J. M. Tomczak, Hierarchical vampprior variational fair auto-encoder, arXiv preprint arXiv:1806.09918.

[82] D. P. Kingma, M. Welling, Auto-encoding variational bayes, in: International Conference on Learning Representations, 2014.

[83] H. Edwards, A. J. Storkey, Censoring representations with an adversary, in: International Conference on Learning Representations, 2016.

[84] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, Y. LeCun, Disentangling factors of variation in deep representation using adversarial training, in: Advances in Neural Information Processing Systems, 2016.

[85] A. Beutel, J. Chen, Z. Zhao, E. H. Chi, Data decisions and theoretical implications when adversarially learning fair representations, arXiv preprint arXiv:1707.00075.

[86] Q. Xie, Z. Dai, Y. Du, E. Hovy, G. Neubig, Controllable invariance through adversarial feature learning, in: Advances in Neural Information Processing Systems, 2017.

[87] D. Xu, S. Yuan, L. Zhang, X. Wu, Fairgan: Fairness-aware generative adversarial networks, in: IEEE International Conference on Big Data, 2018.

[88] Y. Wang, T. Koike-Akino, D. Erdogmus, Invariant representations from adversarially censored autoencoders, arXiv preprint arXiv:1805.08097.

[89] B. Kim, H. Kim, K. Kim, S. Kim, J. Kim, Learning not to learn: Training deep neural networks with biased data, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[90] T. Wang, J. Zhao, M. Yatskar, K. Chang, V. Ordonez, Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations, in: IEEE International Conference on Computer Vision, 2019.

[91] T. Adel, I. Valera, Z. Ghahramani, A. Weller, One-network adversarial fairness, in: AAAI Conference on Artificial Intelligence, 2019.

[92] M. Bertran, N. Martinez, A. Papadaki, Q. Qiu, M. Rodrigues, G. Reeves, G. Sapiro, Adversarially learned representations for information obfuscation and inference, in: International Conference on Machine Learning, 2019.

[93] R. Feng, Y. Yang, Y. Lyu, C. Tan, Y. Sun, C. Wang, Learning fair representations via an adversarial framework, arXiv preprint arXiv:1904.13341.

[94] P. C. Roy, V. N. Boddeti, Mitigating information leakage in image representations: A maximum entropy approach, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[95] B. Sadeghi, R. Yu, V. Boddeti, On the global optima of kernelized adversarial representation learning, in: IEEE International Conference on Computer Vision, 2019.

[96] J. Kim, S. Cho, Fair representation for safe artificial intelligence via adversarial learning of unbiased information bottleneck., in: Workshop on Artificial Intelligence Safety, 2020.

[97] H. Zhao, A. Coston, T. Adel, G. J. Gordon, Conditional learning of fair representations, in: International Conference on Learning Representations, 2020.

[98] X. Gitiaux, H. Rangwala, Learning smooth and fair representations, arXiv preprint arXiv:2006.08788.

[99] A. Morales, J. Fierrez, R. Vera-Rodriguez, R. Tolosana, Sensitivenets: Learning agnostic representations with application to face images, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[100] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014.

[101] J. Song, P. Kalluri, A. Grover, S. Zhao, S. Ermon, Learning controllable fair representations, in: International Conference on Artificial Intelligence and Statistics, 2019.

[102] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

[103] P. Adler, C. Falk, S. A. Friedler, G. Rybeck, C. Scheidegger, B. Smith, S. Venkatasubramanian, Auditing black-box models for indirect influence, in: IEEE International Conference on Data Mining, 2016.

[104] J. E. Johndrow, K. Lum, An algorithm for removing sensitive information: application to race-independent recidivism prediction, The Annals of Applied Statistics 13 (1).

[105] P. Lahoti, K. P. Gummadi, G. Weikum, Operationalizing individual fairness with pairwise fair representations, VLDB Endowment 13 (4) (2019) 506–518.

[106] A. J. Bose, W. L. Hamilton, Compositional fairness constraints for graph embeddings, in: International Conference on Machine Learning, 2019.

[107] Z. Tan, S. Yeom, M. Fredrikson, A. Talwalkar, Learning fair representations for kernel models, in: International Conference on Artificial Intelligence and Statistics, 2020.

[108] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, O. Bachem, On the fairness of disentangled representations, in: Advances in Neural Information Processing Systems, 2019.

[109] E. Creager, D. Madras, J. Jacobsen, M. A. Weis, K. Swersky, T. Pitassi, R. S. Zemel, Flexibly fair representation learning by disentanglement, in: International Conference on Machine Learning, 2019.

[110] V. Mirjalili, S. Raschka, A. Namboodiri, A. Ross, Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images, in: International Conference on Biometrics, 2018.

[111] N. Quadrianto, V. Sharmanska, O. Thomas, Discovering fair representations in the data domain, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[112] H. Zhao, G. Gordon, Inherent tradeoffs in learning fair representations, in: Advances in Neural Information Processing Systems, 2019.

[113] D. McNamara, C. S. Ong, R. C. Williamson, Provably fair representations, arXiv preprint arXiv:1710.04394.

[114] D. McNamara, C. S. Ong, B. Williamson, Costs and benefits of fair representation learning, in: AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, 2019.

[115] C. Dwork, N. Immorlica, A. T. Kalai, M. D. M. Leiserson, Decoupled classifiers for group-fair and efficient machine learning, in: Conference on Fairness, Accountability and Transparency, 2018.

[116] L. Oneto, M. Donini, A. Elders, M. Pontil, Taking advantage of multitask learning for fair classification, in: AAAI/ACM Conference on AI, Ethics, and Society, 2019.

[117] C. D. Brown, H. T. Davis, Receiver operating characteristics curves and related decision measures: A tutorial, Chemometrics and Intelligent Laboratory Systems 80 (1) (2006) 24–38.

[118] F. Johansson, U. Shalit, D. Sontag, Learning representations for counterfactual inference, in: International Conference on Machine Learning, 2016.

[119] L. Oneto, Model Selection and Error Estimation in a Nutshell, Springer, 2020.

[120] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning in large attributed graphs, in: Advances in Neural Information Processing Systems, 2017.

[121] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907.

[122] K. A., I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012.

[123] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

[124] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: British Machine Vision Conference, 2015.

[125] M. Wang, W. Deng, Deep face recognition: A survey, arXiv preprint arXiv:1804.06655.

[126] I. Masi, Y. Wu, T. Hassner, P. Natarajan, Deep face recognition: A survey, in: Conference on Graphics, Patterns and Images, 2018.

[127] G. Guo, N. Zhang, A survey on deep learning based face recognition, Computer Vision and Image Understanding 189.

[128] H. Du, H. Shi, D. Zeng, T. Mei, The elements of end-to-end deep face recognition: A survey of recent advances, arXiv preprint arXiv:2009.13290.

[129] M. D. Zeiler, Adadelta: an adaptive learning rate method, arXiv preprint arXiv:1212.5701.

[130] S. Desai, H. G. Ramaswamy, Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization, in: IEEE Winter Conference on Applications of Computer Vision, 2020.

[131] S. Rebuffi, R. Fong, X. Ji, A. Vedaldi, There and back again: Revisiting backpropagation saliency methods, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[132] A. Taha, X. Yang, A. Shrivastava, L. Davis, A generic visualization approach for convolutional neural networks, in: IEEE European Conference on Computer Vision, 2020.

[133] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, K. Müller, Toward interpretable machine learning: Transparent deep neural networks and beyond, arXiv preprint arXiv:2003.07631.

[134] S. Sattarzadeh, M. Sudhakar, A. Lem, S. Mehryar, K. Plataniotis, J. Jang, H. Kim, Y. Jeong, S. Lee, K. Bae, Explaining convolutional neural networks through attribution-based input sampling and block-wise feature aggregation, arXiv preprint arXiv:2010.00672.

[135] V. V. Ivanov, The theory of approximate methods and their application to the numerical solution of singular integral equations, Springer, 1976.

[136] L. Oneto, S. Ridella, D. Anguita, Tikhonov, ivanov and morozov regularization for support vector machine learning, Machine Learning 103 (1) (2015) 103–136.

[137] A. Berlinet, C. Thomas-Agnan, Reproducing kernel Hilbert spaces in probability and statistics, Springer Science & Business Media, 2011.

[138] G. Peyré, M. Cuturi, Computational optimal transport: With applications to data science, Foundations and Trends® in Machine Learning 11 (5-6) (2019) 355–607.

[139] T. Van Erven, P. Harremos, Rényi divergence and kullback-leibler divergence, IEEE Transactions on Information Theory 60 (7) (2014) 3797–3820.

[140] L. Kantorovich, On the transfer of masses (in russian), in: Doklady Akademii Nauk USSR, 1942.

[141] J. Feydy, T. Séjourné, F. X. Vialard, S. Amari, A. Trouvé, G. Peyré, Interpolating between optimal transport and mmd using sinkhorn divergences, in: International Conference on Artificial Intelligence and Statistics, 2019.

[142] F. Santambrogio, Optimal transport for applied mathematicians, Springer, 2015.

[143] C. Villani, Optimal transport: old and new, Springer Science & Business Media, 2008.

[144] A. Ramdas, N. G. Trillos, M. Cuturi, On wasserstein two-sample testing and related families of nonparametric tests, Entropy 19 (2) (2017) 47.

[145] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Neural Information Processing Systems 32, 2019.

[146] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L. Li, YFCC100M: the new data in multimedia research, Communications of the ACM 59 (2) (2016) 64–73.

[147] M. Merler, N. Ratha, R. S. Feris, J. R. Smith, Diversity in faces, arXiv preprint arXiv:1901.10436.

[148] C. F. Van Loan, G. H. Golub, Matrix computations, Johns Hopkins University Press Baltimore, 1983.