How Does Loss Function Affect Generalization Performance of Deep Learning? Application to Human Age Estimation

Ali Akbari [1] **Muhammad Awais** [1] **Manijeh Bashar** [2] **Josef Kittler** [1]

Abstract

Good generalization performance across a wide variety of domains caused by many external and internal factors is the fundamental goal of any machine learning algorithm. This paper theoretically proves that the choice of loss function matters for improving the generalization performance of deep learning-based systems. By deriving the generalization error bound for deep neural models trained by stochastic gradient descent, we pinpoint the characteristics of the loss function that is linked to the generalization error, and can therefore be used for guiding the loss function selection process. In summary, our main statement in this paper is: *choose a stable loss function, generalize better.* Focusing on human age estimation from the face which is a challenging topic in computer vision, we then propose a novel loss function for this learning problem. We theoretically prove that the proposed loss function achieves stronger stability, and consequently a tighter generalization error bound, compared to the other common loss functions for this problem. We have supported our findings theoretically, and demonstrated the merits of the guidance process experimentally, achieving significant improvements.

1. Introduction

The human age estimation from a face image has received increasing attention in a wide variety of applications, including advanced video surveillance, age-specific advertising, demographic statistics collection, customer profiling, or search optimization in large datasets. Nevertheless, it is one of the most challenging topics in computer vision due to the large variations of factors, such as lightening conditions, camera quality, head pose, makeup application and face expression, which have a negative effect on the estimation accuracy. These factors, mixed with other personal attributes such as gene, gender, race and personal life style, make the problem even more challenging. Collecting large-scale datasets may offer opportunities to advance the state-of-the-art in the field, however, training datasets can never represent all the aforementioned factors fully. Consequently, methods with substantial robustness need to be developed in order to get better accuracy in unseen scenarios.

The ultimate objective of an ideal machine learning algorithm is to build an optimal model being able to generalize to any test data that might be distributed differently than the data used for training. However, deep neural networks (DNN) have the tendency to over-fit the training data. In view of this over-fitting issue, it is very important to understand the factors, that impact on the generalization performance of DNNs. In order to link the generalization performance of a deep neural model to its causal factors, one approach is to derive an upper bound for the generalization error, which defines the potential gap between the error on the training and test data.

One established way to upper-bound the generalization error is achieved by exploiting form the notion of uniform stability (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010; Hardt et al., 2016; Jakubovitz et al., 2018; Wu et al., 2020). Roughly speaking, the stability measures the sensitivity of a learning algorithm to perturbations in the training set. The pioneering work of Bousquet and Elisseeff (Bousquet & Elisseeff, 2002; Elisseeff et al., 2005) show that, with high probability, a tighter generalization bound is achieved by the help of using a more stable learning algorithm. However, this result is valid for deterministic learning algorithms. In a following work, Hardt *et al.* (Hardt et al., 2016) extend the notion of uniform stability to randomized learning algorithms in order to derive upper-bounds for the generalization error of a neural model trained by stochastic gradient descent (SGD). Conceptually, they demonstrate that SGD is more stable, provided that the number of iterations taken by SGD is sufficiently small.

It is common in practice to choose an appropriate loss func-

[1] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK. [2] Institute for Communication Systems (ICS), University of Surrey, Guildford, UK. Correspondence to: Ali Akbari <ali.akbari@surrey.ac.uk>.

Proceedings of the 38^{th} *International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

tion to assess each data point during training of DNNs. This task can be addressed in a heuristic way by trying every possible loss function, or in a principled way, advocated in our paper, which guides the selection process. In this paper, we build on the notion of the algorithmic stability defined in (Hardt et al., 2016) to develop new insights into designing loss functions for training deep neural models. To do so, we link the generalization capability of the trained model to the loss function. In a nutshell, our result establishes that the generalization error of any model trained with SGD is dependent on the properties of the loss adopted function. More concretely, we identify the properties that define a stable and effective loss function, by deriving an upper-bound for the generalization error as a vanishing function of the loss function attributes. We demonstrate that a tighter bound on the generalization error of a DNN model could be expected when using a loss function that follows the guidelines emerging from the above analysis.

As our second contribution in this paper, we propose a novel loss function which provides a better generalization performance, compared with the commonly used loss functions for training DNN based age estimation systems. We theoretically analyze the generalization performance of a DNN model trained via SGD using the proposed loss function. To this end, we first prove that the magnitude value and Lipschitz constant of the proposed loss function are upper bounded by the loss function itself. These results help us to prove that the stability of SGD is directly related to the rate of change of the adopted loss function. Consequently, we show that SGD is generally more stable when the speed of change of the loss function is lower. These results then build a fundamental connection between the generalization error and the stability of a DNN model trained by the loss function. Finally, we experimentally validate our theoretical findings on real data and show that a DNN based age estimation system trained via SGD using the proposed loss function provides a model with a higher generalization capability across a variety of unseen scenarios. The results hold true in a broad range of settings, including small and large-scale datasets and different model architectures. The results suggest that it might be highly beneficial for practitioners in other application ?? fields to focus on designing an effective loss function for which stochastic gradient method converges to a better model.

2. Related Work

Learning with powerful models such as deep neural networks (DNN) has achieved a step change in performance over recent years across a wide variety of tasks (Khalid et al., 2020; Akbari et al., 2021b; Bashar et al., 2020a;b; Akbari et al., 2020b; Khalid et al., 2021). Different learning algorithms for different applications have been proposed in literature (Khalid et al., 2021; Akbari et al., 2020c; 2017b;a; 2016). In order to design an effective learning method, a deep understanding of the impact of various design choices on the generalization performance is particularly critical.

There is a variety of successful methods in the machine learning literature, which try to gain theoretical insight into the factors affecting the generalization performance of learning algorithms (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010; Devroye & Wagner, 1979; Lin, 2019). One classical approach to assess the generalization performance is to derive upper bounds for the generalization error. There are various approaches to set an upper bound on the generalization error, including the use of algorithmic stability (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010; Hardt et al., 2016; Jakubovitz et al., 2018), Vapnik-Chervonenkis (VC) dimension (Jakubovitz et al., 2018), robustness (Xu & Mannor, 2012), the PAC-Bayesian theory (Neyshabur et al., 2018), etc. Each of these approaches reveals some factors that are critical for analyzing the generalization capability of the model obtained by a learning algorithm.

The concept of uniform stability which was firstly introduced by Bousquet and Elisseeff (Bousquet & Elisseeff, 2002), has been widely used for analyzing the generalization error of deterministic or randomized learning algorithms (Elisseeff et al., 2005; Hardt et al., 2016). The theory defines the assumptions on regularity or convexity of the loss function, under which the output of a learning algorithm is uniformly stable. The key consequence of the above-mentioned theoretical work is that the generalization error is upper-bounded by a vanishing function of the number of training samples. This implies the bound becomes tighter as the size of the training data increases. However, there is still a lack of knowledge on how other factors impact on the generalization error of deep neural networks.

Recently, Hardt *et al.* (Hardt et al., 2016) derived an upper-bound on the generalization error as a function of the number of iterations of SGD. In contrast, our focus in this paper is on the link between the properties of the loss function used by SGD for training DNNs and their generalization performance. There are two key differences with the existing stability analysis derived by Hardt *et al.* (Hardt et al., 2016). i) We derive a "high-probability" generalization bound instead of "expected" generalization bound derived by Hardt *et al.* (Hardt et al., 2016). High-probability bounds are stronger than expected ones[1]. ii) More importantly, Hardt *et al.* (Hardt et al., 2016) connect the generalization to number of SGD iterations, while our work connects the generalization to properties of loss *i.e.* its Lipschitzness. Our derivation of the relationship between loss functions and the error bound provides a theoretical insight and guidelines on

[1] http://www.cs.cmu.edu/afs/cs/academic/class/15210-s15/www/lectures/random-notes.pdf

how the loss function should be constructed. To the best of our knowledge, we are the first to adopt this concept for designing an efficient loss function.

In the rest of the paper, we first present a framework for the age estimation problem and then introduce the proposed loss function and its properties. Next, we theoretically analyze the proposed loss function and validate the theoretical findings experimentally.

3. Problem Formulation

Let $(\mathbf{x}, y)$ represents a training instance, where $\mathbf{x}$ represent the input image of an individual's face and y is the corresponding age label being a scalar number from the set of possible age labels $\mathcal{L} = \{1, \cdots, K\}$. The objective in the facial age estimation is finding a function which maps the input face $\mathbf{x}$ to its corresponding age label y. Naturally, there is a semantic similarity between the facial features of an individual at a certain age and those at the immediately preceding and following ages. Typical age estimation algorithms adopt regression or classification approaches (Rothe et al., 2018; Carletti et al., 2019) which might not be efficient. In fact, it is well-known that regression based methods show the instability during training phase and classification based methods ignore the correlation among the neighboring ages at the training stage.

Recently, an efficient learning framework, namely label distribution learning (LDL) (Geng, 2016; Gao et al., 2017; Akbari et al., 2021a), was developed by which the cross-age correlation is exploited in the training phase. In this approach, each scalar age label is encoded as a vector $\mathbf{y} = [y_1, y_2, \cdots, y_K] \in \mathbb{R}^K$, where $\forall y_k, 0 \leq y_k \leq 1$ and $\sum_{k=1}^{K} y_k = 1$ and the expected value of $\mathbf{y}$ is set to equal the true scalar age label. This vector, called label distribution, shares the same properties with probability distribution. That means each y_k expresses the probability of the face sample $\mathbf{x}$ belonging to the k-th age label in $\mathbb{L}$. As is standard (Geng, 2016; Gao et al., 2017), the label vector $\mathbf{y}$ is usually assumed to be a normal distribution function, centered at the true age y with a standard deviation σ), controlling the shape (width) of the label distribution at each age. With this kind of label modeling, the objective of the age estimation problem is to find a mapping function between $\mathbf{x}$ and $\mathbf{y}$.

4. Preliminaries

Our objective is to learn the deep model $f^\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by $\theta \in \mathcal{H}$, that maps the input space $\mathcal{X}$ to the corresponding output space $\mathcal{Y}$. Given input pair $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, drawn according to an unknown distribution $\mathbb{P}$, a typical setting for such learning problem is described as

$$\underset{f^\theta \in \mathcal{F}}{\operatorname{argmin}} \, \mathbb{E}_{\mathbf{z} \sim \mathbb{P}}[\ell(f^\theta; \mathbf{z})], \tag{1}$$

wherein a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^N \to \mathcal{Y}^{\mathcal{X}}$ looks for a solution $f^\theta \in \mathcal{F}$ by minimizing the expected (true) risk $R_{\text{true}}(f^\theta) \triangleq \mathbb{E}_{\mathbf{z} \sim \mathbb{P}}[\ell(f^\theta; \mathbf{z})]$. $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ denotes the loss function which evaluates the precision of the hypothesis f^θ on the basis of difference between the expected and true outputs. Throughout the paper, we sometime use $\ell(f^\theta(\mathbf{x}); \mathbf{y})$ as $\ell(f^\theta; \mathbf{z})$.

As $\mathbb{P}$ is unknown, the optimization problem (1) cannot be solved directly. So, the true risk $R_{\text{true}}(f^\theta)$ is alternatively estimated by the empirical risk defined as $R_{\text{emp}}(f^\theta) \triangleq \frac{1}{N} \sum_{i=1}^{N} \ell(f^\theta; \mathbf{z}_i)$, where $\mathcal{S} = \{\mathbf{z}_i, i = 1, 2, \cdots, N\}$ denotes a finite set of N input pairs $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$, i.i.d. drawn according to $\mathbb{P}$.

In the context of deep neural models, SGD is the widely used learning algorithm $\mathcal{A}$ for dealing with the minimization problem (1). It is a randomized algorithm due to either the random initialization of the model's weights or the random order of passing training samples in $\mathcal{S}$ through SGD. For simplicity in notation, throughout this paper, we assume the only nature of randomness of SGD appears only by the random choice of training samples. Let $\mathcal{R} = \{r_1 \cdots r_T\}$ represent the set of random indices of instances in $\mathcal{S}$. Let $f^\theta_{\mathcal{S},\mathcal{R}}$ be the output of SGD applied to a training dataset $\mathcal{S}$ and the set $\mathcal{R}$.

The final aim of SGD is to provide an ideal solution f^θ by adopting a suitably chosen loss function ℓ. The output model should be able to provide a small gap of performance over the training set $\mathcal{S}$ and any other test set drawn with an unknown distribution $\mathbb{P}$. One approach to assess the efficiency of SGD is to derive an upper bound for the *generalization error*:

Definition 1 (Generalization Error)**.** *Given a training set $\mathcal{S}$ and staring with a set of random indices $\mathcal{R}$ of samples in $\mathcal{S}$, the generalization error of the output model $f^\theta_{\mathcal{S},\mathcal{R}}$, trained by SGD, is defined as the difference between the empirical risk and true risk, i.e. $E(\mathcal{S}, \mathcal{R}) = R_{\text{true}}(f^\theta_{\mathcal{S},\mathcal{R}}) - R_{\text{emp}}(f^\theta_{\mathcal{S},\mathcal{R}})$. It should be noted that due to the randomness of $\mathcal{S}$ and $\mathcal{R}$, $f^\theta_{\mathcal{S},\mathcal{R}}$ and consequently $E(\mathcal{S}, \mathcal{R})$ are random variable.*

Roughly speaking, if SGD provides a tighter bound on the generalization error, the generalization performance of the output model would be better. In this paper, our aim is to express the generalization error bound of the output model $f^\theta_{\mathcal{S},\mathcal{R}}$ achieved by SGD as a function of the properties of the loss function adopted for training. In this study, we will use the notion of uniform stability (Hardt et al., 2016) to uncover the connection between the properties of loss functions and the generalization error. For brevity, in the following, $f_{\mathcal{S},\mathcal{R}}$, $R_{\text{true}}(f_{\mathcal{S},\mathcal{R}})$ and $R_{\text{emp}}(f_{\mathcal{S},\mathcal{R}})$ are sometimes used as a shorthand for $f^{\theta}_{\mathcal{S},\mathcal{R}}$, $R_{\text{true}}(f^{\theta}_{\mathcal{S},\mathcal{R}})$ and $R_{\text{emp}}(f^{\theta}_{\mathcal{S},\mathcal{R}})$ if their meaning is clear from the context.

The bounded difference inequality (BDI), proved by McDiarmid (McDiarmid, 1989), is central to our analysis:

Definition 2 (BDI)**.** *Let $\mathcal{Z}$ be some set and $G : \mathcal{Z}^n \to \mathbb{R}$ be any measurable function. Consider two sets $\mathcal{Q}$, $\mathcal{Q}' \in \mathcal{Z}^n$, such that $\mathcal{Q}$ and $\mathcal{Q}'$ differ in at most one element. If there exists constant ρ such that the following condition, namely bounded difference condition (BDC),*

$$\sup_{\mathcal{Q},\mathcal{Q}' \in \mathcal{Z}^n} |G(\mathcal{Q}) - G(\mathcal{Q}')| \leq \rho, \tag{2}$$

holds, then $\forall \epsilon > 0$

$$\mathbb{P}_{\mathcal{Q}}\big[G(\mathcal{Q}) - \mathbb{E}_{\mathcal{Q}}[G(\mathcal{Q})] \geq \epsilon\big] \leq \exp(-2\epsilon^2/n\rho^2). \tag{3}$$

In other words, BDC (2) holds provided that $G(\cdot)$ does not change much by changing only one element of $\mathcal{Q}$,. Intuitively, these types of functions are slightly clustered around their average, and this intuition is made precise by Eq. (3).

5. Loss Functions

Given a typical face sample $\mathbf{x}$, let $\mathbf{y}$ and $\hat{\mathbf{y}} = f^{\theta}(\mathbf{x})$ represent the corresponding ground-truth label distribution and the label distribution estimated by the the deep model f^{θ}, respectively. Further, consider y_k and $\hat{y}_k$ as the k-element of $\mathbf{y}$ and $\hat{\mathbf{y}}$, respectively. To optimize the model's parameters within the deep LDL framework, we need to choose an appropriate loss function to accurately compute the meaningful distance between the predicted and ground-truth label distributions. The well-known Kullback-Leibler (KL) divergence is widely employed as the loss function to measure the similarity between the predicted and the ground-truth label distributions. The KL loss function is defined as

$$\ell_{KL}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{k=1}^{K} y_k \log(\frac{y_k}{\hat{y}_k}). \tag{4}$$

In this section, we propose a novel parametric loss function, namely Generalized Jeffries-Matusita (GJM) distance, for use in a deep LDL framework by generalizing the Jeffries-Matusita distance (Cha, 2007) as

$$\ell_{GJM}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{k=1}^{K} |y_k^{\alpha} - \hat{y}_k^{\alpha}|^{\frac{1}{\alpha}} = \sum_{k=1}^{K} y_k \left|1 - \left(\frac{\hat{y}_k}{y_k}\right)^{\alpha}\right|^{\frac{1}{\alpha}}, \tag{5}$$

where α is in the range $(0, 1]$. As will be discussed in Section 7, the best performance is achieved when the parameter α ranges between 0.3 and 0.6. In the rest of this paper, we consider α as 0.5, unless otherwise stated.

Our generalization error analysis with respect to the KL and GJM loss functions is explained in the next sections. Before starting, we first explain some properties of the GJM loss function in comparison with the other measure, *i.e.* KL divergence. Throughout this paper, the model's architecture is assumed to be the same. The proofs of the following statements and theorems are provided in the supplementary material.

Loss Function Properties

The following definitions and theorems provide the foundation of our generalization error analysis.

Definition 3 (Lipschitzness)**.** *A loss function $\ell(\hat{\mathbf{y}}, \mathbf{y})$ is γ-Lipschitz with regard to the estimated output vector $\hat{\mathbf{y}}$, if for $\gamma \geq 0$ and $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^K$ we have*

$$|\ell(\mathbf{u}, \mathbf{y}) - \ell(\mathbf{v}, \mathbf{y})| \leq \gamma \|\mathbf{u} - \mathbf{v}\|, \tag{6}$$

where $\|\cdot\|$ denotes the ℓ_2-norm of vectors. Intuitively, a Lipschitz loss function is upper-bounded in terms of its rate of change.

Definition 4 (Smoothness)**.** *A loss function $\ell(\hat{\mathbf{y}}, \mathbf{y})$ is η-smooth with regard to the estimated output vector $\hat{\mathbf{y}}$, if its gradient $\nabla\ell(\hat{\mathbf{y}}, \mathbf{y})$ is η-Lipschitz, that is for $\eta \geq 0$ and $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^K$ we have*

$$\|\nabla\ell(\mathbf{u}, \mathbf{y}) - \nabla\ell(\mathbf{v}, \mathbf{y})\| \leq \eta \|\mathbf{u} - \mathbf{v}\|. \tag{7}$$

Intuitively, the curvature of the loss function is upper-bounded by the η-smoothness property.

Theorem 1. *Let function $h : (0, \infty) \to \mathbb{R}$ be convex, such that $h(1) = 0$. Let's define the following function:*

$$I(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{k=1}^{K} y_k h\left(\frac{\hat{y}_k}{y_k}\right). \tag{8}$$

If $h(\cdot)$ is γ-Lipschitz, i.e.

$$|h(x) - h(z)| \leq \gamma |x - z| \quad \forall x, z, \tag{9}$$

then $I(\hat{\mathbf{y}}, \mathbf{y})$ is also γ-Lipschitz. Furthermore, since $h(\cdot)$ is convex, $I(\hat{\mathbf{y}}, \mathbf{y})$ is also convex with regard to its first argument.

Remark. With $h_{KL}(x) = -\log(x), x > 0$, then it can easily be inferred $\ell_{KL}(\hat{\mathbf{y}}, \mathbf{y}) = I(\hat{\mathbf{y}}, \mathbf{y})$. It is also straightforward to show that $\ell_{GJM}(\hat{\mathbf{y}}, \mathbf{y}) = I(\hat{\mathbf{y}}, \mathbf{y})$, if $h_{GJM}(x) = |1 - x^{\alpha}|^{\frac{1}{\alpha}}, x > 0$. Consequently, since $h_{KL}(x)$ and $h_{GJM}(x)$ are convex functions, then $\ell_{KL}(\hat{\mathbf{y}}, \mathbf{y})$ and $\ell_{GJM}(\hat{\mathbf{y}}, \mathbf{y})$ are also convex with regard to the first argument.

Lemma 1. *A function $h : (0, \infty) \to \mathbb{R}$ is γ-Lipschitz, if γ satisfies the following condition:*

$$\gamma = \sup_{x} |h'(x)|. \tag{10}$$

This implies the value of γ must be equal to the maximum value of $|h'(x)|$.

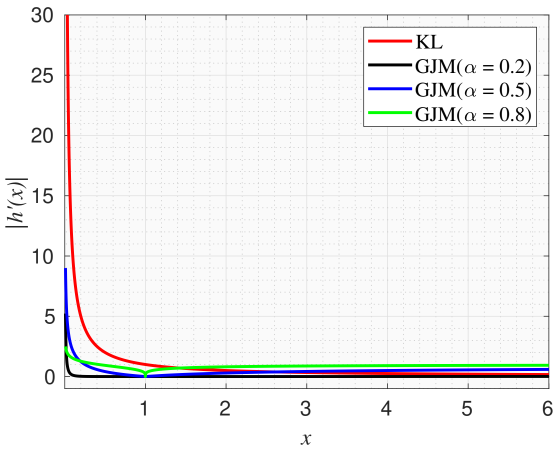

Figure 1. Absolute value of the derivative of the loss functions at different points x.

Proof. This lemma can be easily proved from the definition of the Lipschitz property. ■

The practical result is expressed in the following statement which is the basis for our theoretical analysis in the next section.

Corollary 1. *Let the GJM and KL loss functions are γ_{GJM}-Lipschitz and γ_{KL}-Lipschitz, respectively. Then, the following inequality holds:*

$$\gamma_{GJM} \leq \gamma_{KL}. \tag{11}$$

Fig. 1 shows the absolute value of the derivative for the KL and GJM loss functions as a function of x. As can be seen $|h'_{GJM}(x)|$ is smaller than $|h'_{KL}(x)|$. From Lemma 1, this implies the inequality in (11) holds. We also theoretically prove that $|h'_{GJM}(x)| \leq |h'_{KL}(x)|$ for $\alpha = 0.5$, *i.e.*

$$\left|1 - \frac{1}{\sqrt{x}}\right| \leq \left|\frac{1}{x}\right|. \tag{12}$$

Eq. (12) is equivalent to $|x - \sqrt{x}| \leq 1$, which results in the condition $x \leq 2.6$ after some mathematical simplification. We experimentally found that the variable x always satisfies this condition when the model starts to converge. Note that $|h'_{GJM}(x)|$ and $|h'_{KL}(x)|$ meet each other at some point. For instance, for $\alpha = 0.2, 0.4, 0.5, 0.8$, the intersection point is $x_p = 22.06, 3.75, 2.61, 1.42$ respectively. After this point, $|h'_{GJM}(x)|$ starts to be slightly larger than $|h'_{KL}(x)|$, but, the difference in this area is very small and negligible compared with the points smaller than the intersection point.

As the last fundamental statement in this section, we now provide a connection between the two above-mentioned loss functions.

Theorem 2. *For two distribution $\mathbf{y}, \hat{\mathbf{y}} \in \mathbb{R}^K$, the GJM loss function with $\alpha = 0.5$ is upper-bounded by the KL divergence, i.e. we have the following inequality:*

$$\ell_{GJM}(\hat{\mathbf{y}}, \mathbf{y}) \leq \ell_{KL}(\hat{\mathbf{y}}, \mathbf{y}). \tag{13}$$

6. Stability and Generalization Error Bound

In this section, we follow the notion of stability, introduced by Hardt*et al.* (Hardt et al., 2016), to analyze the generalization error of the DNN model trained by SGD. Roughly speaking, stability refers to the robustness of the output model achieved by a learning algorithm with respect to small changes in its input. The classical result, derived by Bousquet and Elisseeff (Bousquet & Elisseeff, 2002) shows that the generalization error of the output model, obtained by a deterministic learning algorithm $\mathcal{A}$, is upper bounded by a factor of the stability measure, provided that $\mathcal{A}$ satisfies the uniform stability condition. This implies the following statement: a tighter bound on the generalization error can be expected for the output model, if $\mathcal{A}$ satisfies the stability condition with a stricter stability measure. However, these results are valid for deterministic learning algorithms and may not be accurate for the learning algorithms, such as SGD, which have a random element. Unlike the concept of stability used by Bousquet and Elisseeff, we rely on the notion of uniform stability presented in (Shalev-Shwartz et al., 2010; Hardt et al., 2016) to take into account the concerns regarding the randomness of SGD.

Definition 5 (Uniform Stability)**.** *Let $\mathcal{S}'$ and $\mathcal{S}$ denote two training sets of equal size, following an unknown distribution $\mathbb{P}$, such that $\mathcal{S}$ and $\mathcal{S}'$ vary in one entity. Let $f_{\mathcal{S},\mathcal{R}}$ and $f_{\mathcal{S}',\mathcal{R}}$ be the optimal models obtained by SGD, with the set of random indices $\mathcal{R}$ of the training samples in $\mathcal{S}$ and $\mathcal{S}'$, respectively. SGD is then β-uniformly stable with regard to a certain loss function ℓ, if the following inequality holds:*

$$\forall\, \mathcal{S},\, \mathcal{S}' \quad \sup_{\mathbf{z}} \mathbb{E}_{\mathcal{R}}\left[\,|\ell(f_{\mathcal{S},\mathcal{R}}; \mathbf{z}) - \ell(f_{\mathcal{S}',\mathcal{R}}; \mathbf{z})|\,\right] \leq \beta, \tag{14}$$

where the expectation is taken over the randomness of SGD which is a function of the random choice of data $\mathcal{S}$ for training.

Intuitively, if SGD is β-uniformly stable, then it has this property that altering one pair in the training set $\mathcal{S}$ and holding others fixed makes at most β-change in the error of the output model by SGD with any random permutation of the training samples in $\mathcal{S}$.

Now, we link the concept of stability with the loss function and then derive an upper bound for the generalization error which depends on some attributes of the employed loss function. This reveals the relation between the loss function and the generalization error. This renders it amenable for analyzing the generalization performance of DNNs, trained by SGD, with regard to the loss function employed for training. These assertions are stated in Theorems 3 – 5.

Theorem 3. *Consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, L]$. Let $f_{\mathcal{S},\mathcal{R}}$ denote the optimal model obtained by SGD with the set of random indices $\mathcal{R}$ of the training samples in $\mathcal{S}$. Let SGD be β-uniformly stable with regard to the employed loss function ℓ. Furthermore, assume there is a constant ρ for which $\ell(f_{\mathcal{S},\mathcal{R}}, \mathbf{z})$ satisfies the bounded difference condition* (2) *with respect to $\mathcal{R}$. Then, with probability at least $1-\delta$, the following bounds hold $\forall \mathcal{S}, \mathcal{R}$:*

$$R_{\text{true}}(f_{\mathcal{S},\mathcal{R}}) - R_{\text{emp}}(f_{\mathcal{S},\mathcal{R}}) \leq \rho\sqrt{T\log(2/\delta)} + \beta\left(1 + \sqrt{2N\log(2/\delta)}\right) + L\sqrt{\frac{\log(2/\delta)}{2N}}. \tag{15}$$

The stability parameter β and the BDI constant ρ depend on the properties of the loss function used by SGD. We now state the following theorem which derives the upper bounds for β and ρ.

Theorem 4. *Assume that SGD is run for T iterations with an annealing learning rate λ_t to find the optimal solution of the minimization problem* (1). *Let $\ell(f^{\theta}(\mathbf{x}), \mathbf{y})$ be convex, γ-Lipschitz and η-smooth with regard to its first argument for each $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. Then SGD is β-uniformly stable and holds the ρ-BDC* (2) *with regard to $\ell(f_{\mathcal{S},\mathcal{R}}, \mathbf{z})$ and $\mathcal{R}$. Consequently, we have*

$$\beta \leq \frac{2\gamma^2}{N}\sum_{t=1}^{T}\lambda_t \qquad \textit{and} \qquad \rho \leq \frac{4\gamma^2}{T}\sum_{t=1}^{T}\lambda_t. \tag{16}$$

Combining Theorem 3 and Theorem 4 gets the following result.

Theorem 5. *Consider a loss function ℓ such that $0 \leq \ell(f(\cdot;\mathbf{z}) \leq L$ for any point $\mathbf{z}$. Suppose that the SGD update rule is executed for T iterations with an annealing learning rate λ_t to solve the optimization problem* (1). *Then, we have the following generalization error bound with probability at least $1-\delta$:*

$$R_{\text{true}}(f_{\mathcal{S},\mathcal{R}}) - R_{\text{emp}}(f_{\mathcal{S},\mathcal{R}}) \leq L\sqrt{\frac{\log(2/\delta)}{2N}} + 2\gamma^2\sum_{t=1}^{T}\lambda_t\left(2\sqrt{\frac{\log(2/\delta)}{T}} + \sqrt{\frac{2\log(2/\delta)}{N}} + \frac{1}{N}\right). \tag{17}$$

Remark. Theorem 5 implies that the generalization error diminishes when the number of training samples increases. On the other hand, both the first and second term in (17), depend on some attributes of the loss function, including its Lipschitz constant γ and the maximum value ℓ can assume. As a result, using a loss function with a smaller value of γ give us the ability to control the uniform stability and the generalization error bound of the trained model.

As proved in Section 5, the GJM and KL loss functions satisfy the Lipschitzness and smoothness properties. Thus, Theorem 4 and Theorem 5 are valid when the GJM or KL are used as the loss function for training[2]. Following Corollary 1 and Theorem 2, in the following corollary, we link the generalization error bound of a model trained by the GJM loss function, to that trained by the KL divergence.

Corollary 2. *Consider two models $f^{GJM}_{\mathcal{S},\mathcal{R}}$ and $f^{KL}_{\mathcal{S},\mathcal{R}}$ trained under the same settings using the GJM and KL loss functions, respectively using the training set, $\mathcal{S}$. We have the following inequality:*

$$E(f^{GJM}_{\mathcal{S},\mathcal{R}}) \leq E_{CE}(f^{KL}_{\mathcal{S},\mathcal{R}}), \tag{18}$$

where $E(f_{\mathcal{S},\mathcal{R}}) = R_{\text{true}}(f_{\mathcal{S},\mathcal{R}}) - R_{\text{emp}}(f_{\mathcal{S},\mathcal{R}})$.

This implies that a tighter bound on the generalization error is achieved using the GJM loss function. In other words, DNNs which are trained with the proposed loss function exhibit a better generalization performance, compared with those trained by the KL loss function.

7. Experimental Evaluation

The goal of our experiments is to assist in evaluating the effect of the loss function on the generalization performance of DNN based age estimation systems trained by SGD. It should be noted that none of the reported results in this paper are intended to compete with the state-of-the-arts — our goal is to demonstrate how the loss function affects the generalization performance of a DNN model trained by SGD.

7.1. Settings and Datasets

We evaluate a variety of neural network architectures trained on a number of different datasets. We study the VGG (Parkhi et al., 2015) and the ResNet50 models (Hu et al., 2018), pre-trained on VGGFace2 dataset (Cao et al., 2018) for our experiments. The last fully-connected layer in these models is replaced with a K-neurons fully-connected layer, where K is the number of the age classes. The weights of this FC layer are then randomly initialized. K is set to 101 for ages from 0 to 100. In all experiments, we train

[2] Theoretically speaking, KL is not Lipschitz. Theorem 1 states that a loss function in the form of $I(\cdot,\cdot)$ is γ-Lipschitz, if $h(\cdot)$ is γ-Lipschitz. In the case of KL, $h(x) = -\log(x)$ and so we need to find $\gamma = \sup_x |h'(x)|$ (Lemma 1). In theory, there is no finite value of γ to satisfy this condition for $x = 0$; therefore KL does not satisfy the Lipschitz property. However, from a practical point of view, we always make KL γ-Lipschitz by bounding x from below (assuming the minimum x is $1E-15$). So, we can say that this trimmed KL is γ-Lipschitz but the constant γ is very large. In contrast, GJM does not have this issue. In other words, we can state that the GJM is better in generalization due to this Lipschitzness property.

the models via SGD with the same random seed and set the following training hyper-parameters. The batch size, parameter α, weight decay and momentum are set to 64, 0.5, 0.0005 and 0.9 respectively. For the experiments with the VGG model, the learning rate is initialized as 0.001 and then scheduled with exponential decay to reach 10^{-5} at 30-th epoch. For experiments on the ResNet50 model, the learning rate is initialized as 10^{-5} and then increased to a higher value 0.07 after 5 epochs and then decreased exponentially to reach 10^{-5} after 30 epochs.

All the images used for training and testing are preprocessed by the following procedures: first, the position of the left and right middle of the eyes, the tip of the nose, the left and right edges of the mouth are extracted by utilizing the face detector proposed in (Zhang et al., 2016). By normalizing this positional information, we adjust the face at the center of the input image by the alignment method in (Wen et al., 2016). In the end, all images are reshaped to the size of 256×256 pixels and then fed to the model for training and testing. The standard data augmentation techniques, including random cropping and flipping, are preformed during the training phase. In the test stage, we use only the center-cropped images.

We evaluate the age estimation performance on 5 datasets, including Balanced AGeing (BAG) (Akbari et al., 2020a), MORPH (Ricanek & Tesafaye, 2006), FG-NET (Panis et al., 2016), FACES (Ebner et al., 2010) and SC-FACE (Grgic et al., 2011). The BAG dataset contains $200,123$ in-the-wild images. There are enough images of all ages, ranging from 0 to 100 years-old. The MORPH dataset includes $55,134$ images in the age range from 16 to 72 years-old. This dataset provides a suitable collection for analyzing the generalization performance because most of images in the dataset are African people, while this ethnic group is under represented in our training dataset. FG-NET dataset contains $1,002$ images with the age labels in the range from 0 to 69 years-old. This dataset provides large variations in pose, expression and lighting conditions. The FACES dataset has $2,052$ images with six expressions (neutrality, happiness, anger, fear, disgust, and sadness) in the age range from 19 to 80 years-old. SC-FACE dataset contains $4,160$ images in the age range from 21 to 75 years-old. We separate the SC-FACE dataset into two separate parts, namely SC-FACE-ROT and SC-FACE-SUR, which contain $1,170$ and $2,990$ images, respectively. The SC-FACE-ROT part contains 10 images for each individual captured with different head poses ranging from $-90°$ to $+90°$ in equal steps of $22.5°$. The SC-FACE-SUR part has 17 images for each subject captured with seven cameras with different shooting characteristics.

In our experiments, we use a random subset of BAG dataset, namely SubBAG, or MORPH, as the training set $\mathcal{S}$. We retain other datasets as the test sets $\mathcal{T} \in (\mathcal{X} \times \mathcal{Y})^M$. Note that under this setting, the output model $f_{\mathcal{S},\mathcal{R}}$, is statistically independent of the characteristics of images in $\mathcal{T}$s. Therefore, we can evaluate the generalization performance reliably (Akbari et al., 2020a; 2021b; 2020b).

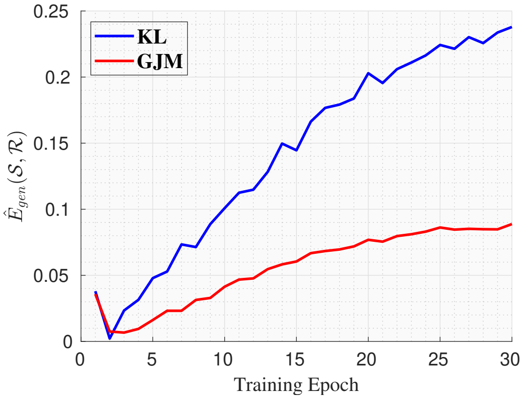

Figure 2. Generalization error curves per training epoch.

Table 1. Generalization Error of VGG Models.

	$\hat{E}_{gen}(\mathcal{S},\mathcal{R}) = \lvert R_{train}(f^{\theta}_{\mathcal{S},\mathcal{R}}) - R_{test}(f^{\theta}_{\mathcal{S},\mathcal{R}})\rvert$				
Method	FG-NET	MORPH	FACES	SC-FACE	Average
KL	0.1023	0.7321	0.8834	1.0974	0.7038
GJM	0.0097	0.2420	0.3272	0.2358	0.2037

We measure the generalization error directly in terms of the absolute difference between losses on the test and training set. Let $f^{\theta}_{\mathcal{S},\mathcal{R}}$ denote the model trained using set $\mathcal{S}$. We define $E_{gen}(\mathcal{S},\mathcal{R}) \cong \hat{E}_{gen}(\mathcal{S},\mathcal{R}) = \lvert R_{train}(f^{\theta}_{\mathcal{S},\mathcal{R}}) - R_{test}(f^{\theta}_{\mathcal{S},\mathcal{R}})\rvert$ as a measure to approximate the generalization error, where $R_{train}(f^{\theta}_{\mathcal{S},\mathcal{R}})$ and $R_{test}(f^{\theta}_{\mathcal{S},\mathcal{R}})$ denote the average loss values of the trained model $f^{\theta}_{\mathcal{S},\mathcal{R}}$ on the training and test sets, respectively. We further evaluate the generalization performance in terms of accuracy measures, including mean absolute error (MAE) and cumulative score (CS) (Guo et al., 2009), on the training and test sets. MAE is defined as $\sum_{k=1}^{M} \frac{|\hat{l}_k - l_k|}{M}$, where M is the total number of test samples and $\hat{l}_k$ is the corresponding estimated age, obtained by taking the bin index to the maximum value of the model's output. CS is defined as $\frac{M_I}{M} \times 100\%$, where K_I is the number of samples such that $|\hat{y}_k - y_k| < I$. In this paper we set I as 5.

7.2. Evaluation

In this section, we evaluate the generalization performance of the trained model with respect to the loss function adopted for training. In our first experiment, we randomly choose a subset of 50K images from the BAG dataset. 90% images of

Table 2. Generalization Performance in Terms of MAE and CS Measures (Model: VGG, Training Set: BAG).

	FG-NET		MORPH		FACES		SC-FACE-ROT		SC-FACE-SUR		Average	
Method	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)
Classification	3.20	82.14	5.50	60.34	5.33	61.60	6.07	53.59	5.44	66.76	5.10	64.88
LDL (χ^2)	5.35	59.28	4.76	66.49	4.66	65.25	**4.50**	71.07	4.89	69.80	4.83	66.37
LDL (KL)	3.08	83.83	5.27	62.43	4.72	66.76	5.25	63.93	5.46	65.71	4.75	68.53
LDL (GJM)	**2.93**	**84.43**	**4.63**	**66.03**	4.47	69.88	4.72	**71.19**	**4.78**	**71.75**	**4.30**	**72.65**

Table 3. Generalization Error in Terms of MAE and CS Measures (Model: VGG, Training Set: SubBAG).

	FG-NET		MORPH		FACES		SC-FACE-ROT		SC-FACE-SUR		Average	
Method	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)
Classification	3.57	78.94	6.54	53.38	6.59	50.83	6.45	49.32	6.19	65.05	5.86	59.50
χ^2	3.29	80.44	5.98	56.10	6.05	55.77	5.61	58.55	5.75	66.89	5.33	63.55
LDL (KL)	3.24	81.54	6.01	57.36	6.11	55.60	5.90	54.79	6.52	60.64	5.55	61.98
LDL (GJM)	**3.21**	**81.59**	**5.63**	**59.13**	**5.90**	**57.55**	**5.32**	**62.14**	5.37	67.96	**5.08**	**65.67**

Table 4. Generalization Error in Terms of MAE and CS Measures (Model: VGG, Training Set: MORPH).

	FG-NET		BAG		FACES		SC-FACE-ROT		SC-FACE-SUR		Average	
Method	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)
Classification	5.73	58.31	8.91	40.75	9.28	36.53	4.68	67.95	9.14	23.68	7.54	45.44
LDL (χ^2)	5.95	61.83	8.86	40.85	9.50	37.35	4.59	70.26	9.55	27.76	7.69	47.61
LDL (KL)	5.45	62.76	**8.41**	**42.86**	**8.43**	**40.90**	4.22	71.54	9.70	23.96	7.24	48.40
LDL (GJM)	**5.29**	**63.70**	8.62	40.99	8.73	40.45	**4.05**	**76.24**	**8.92**	**29.40**	7.12	50.15

Table 5. Generalization Error in Terms of MAE and CS Measures (Model: ResNet50, Training Set: SubBAG).

	FG-NET		MORPH		FACES		SC-FACE-ROT		SC-FACE-SUR		Average	
Method	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)
Classification	4.29	73.15	6.55	52.10	7.48	50.15	7.99	45.04	8.67	57.79	6.99	55.64
LDL (KL)	4.04	75.35	6.44	52.60	7.32	50.73	8.67	27.86	7.80	61.71	6.85	53.60
LDL (χ^2)	3.62	78.44	5.98	57.70	7.05	51.56	8.02	41.71	9.28	59.03	6.79	57.68
LDL (GJM)	**3.40**	**82.04**	**5.74**	**58.60**	**6.16**	**57.12**	**6.66**	**46.32**	**7.54**	**62.01**	**5.90**	**61.21**

this set are randomly selected for training the VGG model and the rest are used for validation. Fig. 2 illustrates the loss generalization error $\hat{E}_{gen}$ computed on the validation set using the model obtained at each epoch. Due to different scale of the loss values, we normalize the loss values to the range $[0, 1]$. The plot shows that the over-fitting issue is more severe for the model trained by the KL divergence. On the other hand, the proposed GJM loss function greatly alleviates the over-fitting issue. This behavior can also be inferred from Fig. 1. The large values of x in Fig. 1 usually occur at the beginning of training (first few epochs). At these points, the Lipschitz constant of the both GJM and KL loss functions are close to each other as seen in Fig. 1. As can also be observed from Fig. 2, the generalization error curves coincide each other for the first few epochs. However, when training continues GJM has better stability, and so it exhibits better generalization than KL at the end of training phase.

The generalization error $\hat{E}_{gen}$ of the solution achieved by SGD is further reported in Table 1 for several test datasets. It can be observed that the generalization error is lower for the model trained by the proposed GJM loss function. These observations confirm our main outcome which has been theoretically proved in Corollary 2.

In Tables 2, 3 and 4, we evaluate the generalization performance, in terms of MAE and CS measures, of various VGG models which are trained on BAG, SubBAG and MORPH datasets, respectively. Since the MORPH dataset has no images with ages outside the range 16 to 72, we removed the images with labels outside this range from the test sets. Table 5 reports the generalization performance of ResNet50 model trained on the SubBAG dataset.

Within the LDL framework, we investigate the effect of using KL divergence (Gao et al., 2017) and χ^2-statistic (Österreicher, 2002) as the loss function. The performance is also compared with the classification based age estimation method (Rothe et al., 2018), where the well-known cross entropy (CE) is used as the loss function. In this approach, the age labels are one-hot encoded. It should be noted that when $\sigma \to 0$, the LDL framework will be similar to the classification based approach. We can write $\ell_{KL}(\hat{\mathbf{y}}, \mathbf{y}) = \ell_{CE}(\hat{\mathbf{y}}, \mathbf{y}) + H(\mathbf{y})$, where $H(\mathbf{y})$ is the negative entropy of $\mathbf{y}$. For a given sample, it is a constant negative value. Other things being equal, the GJM loss achieves DNN models, whose generalization capabilities are practically distinguishable from those obtained by the other loss functions. It should be noted that the performance of models trained on the MORPH dataset is significantly lower than those trained by the BAG dataset. This reflects the characteristics of the MORPH dataset that contains images captured in a controlled environment. As the final point, it can be inferred that the choice of loss function affects the generalization performance of DNN based age estimation systems.

7.3. The Effect of the Hyper-Parameter

α is the hyper-parameter in the proposed GJM loss function which affects the performance of the trained model. From Lemma 1, $\gamma = \sup_x |h'(x)|$. Therefore, we can infer that the stability is higher for smaller α (see Fig. 1). However, the performance degrades for very small values of α, because the loss function becomes constant for small α, and is rendered a meaningless objective function for training.

In order to study the impact of α, we evaluate the generalization performance with different α values, changing from 0 to 1. Table 6 shows the age estimation accuracy of different VGG models which are trained using the GJM loss function with different values of α. We report MAEs on the FG-NET dataset. We can see that a proper α is important for the best MAE measure. Generally, $\alpha = 0.5$ is the best choice.

Table 6. The Influence of α on Age Estimation Accuracy

α	0.2	0.3	0.4	0.5	0.6	0.7	0.8
MAE	3.7	3.3	3.2	**3.2**	3.5	4	4.8

8. Conclusion and Future Work

The main goal of this study was to establish a relationship between the generalization performance of DNN based systems and the loss function. Using the notion of uniform stability, we showed that the generalization error is dependent on the properties of the loss function used for training deep neural network via the stochastic gradient descent algorithm. We proved that the model trained with a Lipschitz loss function exhibits a stronger stability, and therefore a a lower generalization error is expected. Inspired by our theoretical findings, focusing on the age estimation problem, we proposed a loss function which helps to improve the generalization capability of DNN based age estimation systems. We validated our theoretical findings experimentally by comparing the generalization error of different age estimation models (using the same DNN architecture) trained with several loss functions on large training sets.

We should emphasize that other factors, beside loss function, affect the accuracy. In this work, our goal was to show how the loss function itself affects the training process, while keeping other contributing factors fixed. In fact, we show that Lipschitzness property of the loss function is directly related to the model'stability and this gives us some insights as to how to design a loss that is more stable with respect to input changes, for instance change of illumination over the input face, etc. Considering this as the most important property of our loss, we believe that adapting GJM with state-of-the-art techniques in the age estimation, and even in other research areas, helps to improve the generalization performance in unseen scenarios. Furthermore, while this work primarily focuses on the age estimation task, the findings are applicable to other vision tasks, such as image quality assessment, recommendation systems, human pose estimation, and other tasks currently addressed with label distribution learning.

Finally, it should be noted that the total error of a model can be decomposed into the generalization error and the optimization error (difference between the expected risk and the true optimum of the empirical risk). We addressed the former component by upper bounding the error by a quantity related to the SGD stability. For a comprehensive investigation of the factors impacting on the overall performance of deep label distribution learning, we need to analyze the factors influencing the optimization error as well. Further, our theoretical framework does not directly explain how tightly the loss function upper-bounds a quantity of interest, such as MAE. A more direct linking of the loss function and measures of interest, as well as the optimization error, will be the focus of our future work.

Acknowledgements

This work was supported in part by the EPSRC Programme Grant (FACER2VM) EP/N007743/1 and the EPSRC/dstl/MURI project 0184EP/R56/1. This project made use of time on Tier 2 HPC facility JADE2, funded by EPSRC (EP/T022205/1).

References

Akbari, A., Trocan, M., and Granado, B. Image error concealment using sparse representations over a trained dictionary. In *Picture Coding Symposium (PCS)*, pp. 1–5, 2016.

Akbari, A., Trocan, M., and Granado, B. Joint-domain dictionary learning-based error concealment using common space mapping. In *International Conference on Digital Signal Processing (DSP)*, pp. 1–5, 2017a.

Akbari, A., Trocan, M., and Granado, B. Image error concealment based on joint sparse representation and non-local similarity. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 6–10, 2017b.

Akbari, A., Awais, M., Feng, Z., Farooq, A., and Kittler, J. Distribution cognisant loss for cross-database facial age estimation with sensitivity analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020a. doi: 10.1109/TPAMI.2020.3029486.

Akbari, A., Awais, M., and Kittler, J. Sensitivity of age estimation systems to demographic factors and image quality: Achievements and challenges. In *IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–6, 2020b.

Akbari, A., Trocan, M., Sanei, S., and Granado, B. Joint sparse learning with nonlocal and local image priors for image error concealment. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2559–2574, 2020c.

Akbari, A., Awais, M., Fatemifar, S., Khalid, S. S., and Kittler, J. A novel ground metric for optimal transport-based chronological age estimation. *IEEE Transactions on Cybernetics*, pp. 1–1, 2021a.

Akbari, A., Awais, M., Feng, Z., Farooq, A., and Kittler, J. A flatter loss for bias mitigation in cross-dataset facial age estimation. In *International Conference on Pattern Recognition (ICPR)*, 2021b.

Bashar, M., Akbari, A., Cumanan, K., Ngo, H. Q., Burr, A. G., Xiao, P., and Debbah, M. Deep learning-aided finite-capacity fronthaul cell-free massive mimo with zero forcing. In *IEEE International Conference on Communications (ICC)*, pp. 1–6, 2020a.

Bashar, M., Akbari, A., Cumanan, K., Ngo, H. Q., Burr, A. G., Xiao, P., Debbah, M., and Kittler, J. Exploiting deep learning in limited-fronthaul cell-free massive mimo uplink. *IEEE Journal on Selected Areas in Communications (JSAC)*, 38(8):1678–1697, 2020b.

Bousquet, O. and Elisseeff, A. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, March 2002.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In *2IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pp. 67–74, 2018. doi: 10.1109/FG.2018.00020.

Carletti, V., Greco, A., Percannella, G., and Vento, M. Age from faces in the deep learning revolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, Apr 2019.

Cha, S.-H. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(1):300–307, Nov 2007.

Devroye, L. and Wagner, T. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

Ebner, N. C., Riediger, M., and Lindenberger, U. Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42(1):351–362, Feb 2010.

Elisseeff, A., Evgeniou, T., and Pontil, M. Stability of randomized learning algorithms. *The Journal of Machine Learning Research*, 6:55–79, Jan 2005.

Gao, B., Xing, C., Xie, C., Wu, J., and Geng, X. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, June 2017.

Geng, X. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, July 2016.

Grgic, M., Delac, K., and Grgic, S. Scface – surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, Feb 2011.

Guo, G., , Fu, Y., and Huang, T. S. Human age estimation using bio-inspired features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 112–119, June 2009.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234, New York, New York, USA, Jun 2016.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *IEEE/CVF Conference on Computer Vision* *and Pattern Recognition*, pp. 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

Jakubovitz, D., Giryes, R., and Rodrigues, M. R. D. Generalization error in deep learning. *CoRR*, abs/1808.01174, 2018. URL `http://arxiv.org/abs/1808.01174`.

Khalid, S. S., Awais, M., Feng, Z. H., Chan, C. H., Farooq, A., Akbari, A., and Kittler, J. Resolution invariant face recognition using a distillation approach. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4): 410–420, 2020.

Khalid, S. S., Awais, M., Chan, C.-H., Feng, Z., Farooq, A., Akbari, A., and Kittler, J. Npt-loss: A metric loss with implicit mining for face recognition, 2021.

Lin, S. Generalization and expressivity for deep nets. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5):1392–1406, 2019.

McDiarmid, C. *On the method of bounded differences*. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge, 1989.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Skz_WfbCZ`.

Österreicher, F. Csiszár's f-divergence-basic properties. Technical report, Victoria University, Melbourne, VIC, Australia, 2002.

Panis, G., Lanitis, A., Tsapatsoulis, N., and Cootes, T. F. Overview of research on facial ageing using the fg-net ageing database. *IET Biometrics*, 5(2):37–46, May 2016.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. Deep face recognition. In *British Machine Vision Conference*, 2015.

Ricanek, K. and Tesafaye, T. Morph: a longitudinal image database of normal adult age-progression. In *International Conference on Automatic Face and Gesture Recognition (FGR)*, pp. 341–345, Apr 2006.

Rothe, R., Timofte, R., and Van Gool, L. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, Apr 2018.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, December 2010.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision (ECCV)*, pp. 499–515. Springer International Publishing, 2016.

Wu, X., Zhang, J., and Wang, F. Stability-based generalization analysis of distributed learning algorithms for big data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(3):801–812, 2020.

Xu, H. and Mannor, S. Robustness and generalization. *Machine Learning*, 86(3):391–423, Mar 2012.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503, Oct 2016.