Unimodal-Concentrated Loss: Fully Adaptive Label Distribution Learning for Ordinal Regression

Qiang Li*, Jingjing Wang*, Zhaoliang Yao, Yachun Li,
Pengju Yang, Jingwei Yan, Chunmao Wang, Shiliang Pu†
Hikvision Research Institute, China
{liqiang23,wangjingjing9,yaozhaoliang,liyachun6,yangpengju,
yanjingwei,wangchunmao,pushiliang.hri}@hikvision.com

Abstract

Learning from a label distribution has achieved promising results on ordinal regression tasks such as facial age and head pose estimation wherein, the concept of adaptive label distribution learning (ALDL) has drawn lots of attention recently for its superiority in theory. However, compared with the methods assuming fixed form label distribution, ALDL methods have not achieved better performance. We argue that existing ALDL algorithms do not fully exploit the intrinsic properties of ordinal regression. In this paper, we emphatically summarize that learning an adaptive label distribution on ordinal regression tasks should follow three principles. First, the probability corresponding to the ground-truth should be the highest in label distribution. Second, the probabilities of neighboring labels should decrease with the increase of distance away from the ground-truth, i.e., the distribution is unimodal. Third, the label distribution should vary with samples changing, and even be distinct for different instances with the same label, due to the different levels of difficulty and ambiguity. Under the premise of these principles, we propose a novel loss function for fully adaptive label distribution learning, namely unimodal-concentrated loss. Specifically, the unimodal loss derived from the learning to rank strategy constrains the distribution to be unimodal. Furthermore, the estimation error and the variance of the predicted distribution for a specific sample are integrated into the proposed concentrated loss to make the predicted distribution maximize at the ground-truth and vary according to the predicting uncertainty. Extensive experimental results on typical ordinal regression tasks including age and head pose estimation, show the superiority of our proposed unimodal-concentrated loss compared with existing loss functions.

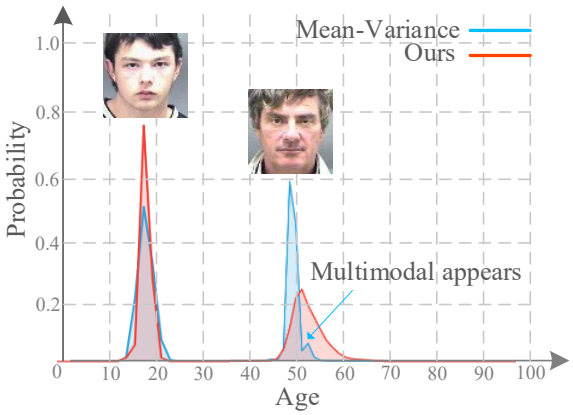

Figure 1. Distributions predicted by Mean-Variance method [22] and ours. Our predictions are optimized to be unimodal and learned according to specific instances adaptively. On the contrary, predictions of Mean-Variance are optimized to be concentrated for all instances and do not ensure unimodal distributions explicitly.

1. Introduction

Ordinal regression solves the challenging problems that labels are related in a natural or implied order. Many critical tasks are involved in the ordinal regression problem, e.g., facial age estimation, head pose estimation, facial attractiveness computation and movie ratings, which play an important role in many practical applications such as human-computer interaction, driver monitoring, precise advertising and video surveillance [10, 32].

Early classic works [11, 17, 20, 24, 37, 38] are based on ordinary classification or regression, which do not perform well due to ignoring the ordinal relationship among labels, and suffering from the ambiguous labeling. In recent years, ranking based methods [3, 21] are proposed which use multiple binary classifiers to determine the rank order. They explicitly make use of the ordinal information but they do

*Authors contribute equally to this work.
†Shiliang Pu is the corresponding author.

not consider the label ambiguity.

To address the ordinal relationship and label ambiguity, label distribution learning (LDL) [7] converts a single label to a label distribution. The label distribution covers a certain number of class labels, representing the degree to which each label describes the instance. Since the real distribution for each instance is not available and must be artificially generated with proper assumption, it can be called fixed form label distribution learning (FLDL). The typical form is the Gaussian distribution centered at the ground-truth with assumed standard deviation [1, 7, 8]. Although FLDL approaches achieve improved performance, however, they use a fixed form distribution to describe various instances which limits their expression ability.

To overcome this limitation, the concept adaptive label distribution learning (ALDL) [9] has been proposed. Among the ALDL based methods, Mean-Variance [22] is a typical work achieving the promising result, which estimates a distribution with learned mean and variance. However, it pursues a highly concentrated distribution for all instances by making the mean as close to the ground-truth as possible, and the variance as small as possible. Moreover, it can not guarantee the learned distribution is unimodal by a joint use of softmax and mean-variance loss without unimodal constraint. Therefore, we observe that the distributions learned by Mean-Variance are not fully adaptive and are multimodal for some instances, as shown in Fig. 1. We can see the learned distribution for the older man is multimodal, and the learned distributions for the two persons are similar. The learned distributions do not accord with the tendency of facial aging, which might be significantly different at different ages [9].

Obviously, current ALDL methods have not fully exploited the intrinsic properties of ordinal regression. In this paper, the following three principles are summarized for ordinal regression. First, following the empirical risk minimization, the probability corresponding to the ground-truth should be the highest in a label distribution. Second, the labels in ordinal regression tasks change gradually, and the similarity between the test instance and the class prototype decreases gradually when the label move away from the ground-truth. Therefore, the probabilities of neighboring labels accounting for the instance should decrease with the increase of distance away from the ground-truth, i.e., the distribution is unimodal. Third, the label distribution should vary with the samples changing, and even be distinct for different instances with the same label, due to the different levels of difficulty and ambiguity. In other words, the learned label distribution should be adaptive for a particular instance. To satisfy the principles above, we propose a new adaptive label distribution learning approach equipped with a unimodal-concentrated loss. Based on principle I, we directly maximize the probability at the ground-truth via concentrated loss as our primary learning objective. Based on principle II, the unimodal loss derived from learning to rank strategy (LTR) [6] is introduced to constrain the distribution to be unimodal. If two neighboring labels are ranked incorrectly, a positive loss would be output to update the trainable parameters to correct the ordinal relationship. Based on principle III, the variance of the distribution corresponding to the concentration degree is integrated and optimized jointly in the concentrated loss, which can be regarded as an indicator of data uncertainty and label ambiguity. The main contributions of this work are three-fold:

- We are the first to comprehensively summarize the intrinsic principles for learning an adaptive label distribution on ordinal regression tasks. First, the probability at the ground-truth should be the highest in the distribution. Second, the distribution should be unimodal. Third, the distribution should be adaptive to individual instances. These three principles would shed light on the design of loss functions for future works in the field of ordinal regression.

- Different from previous methods which do not fully comply the above principles, we propose a new unimodal-concentrated loss, with the unimodal part constraining the distribution to be unimodal, and with the concentrated part making the distribution concentrated at the ground-truth and fully adaptive to individual instances.

- The proposed loss can be easily embedded into existing CNNs without modifying the structure, and extensive experimental results demonstrate its superiority.

2. Related Work

Existing methods for ordinal regression can be divided into three categories: non-LDL based methods, FLDL based methods and ALDL based methods.

2.1. Non-LDL

Non-LDL methods can be grouped into regression based, classification based and ranking based. Classification based methods usually cast ordinal regression as a classification problem. For examples, age estimation was cast as a classification problem with 101 categories [27], and the angle of yaw was divided into coarse bins as class labels for head pose estimation [14, 25]. These methods treat ordinal labels as independent ones, and the cost of being assigned to any wrong category is the same, which can't exploit the relations between labels. Regression based methods directly regress the ground-truth with Euclidean loss to penalize the difference between the estimation and ground-truth mostly, which do not explicitly make use of the ordinal information. Yi et al. [38] used CNNs models to extract features

from several facial regions, and used a square loss for age estimation. Ranjan et al. [24] proposed a unified CNN network to jointly estimate facial age, head pose, and other attributes. Recently, ranking techniques are introduced to the problem of ordinal regression. Niu et al. [21] leveraged the ordinal information of ages by learning a network with multiple binary outputs, while Chen et al. [3] did this by learning multiple binary CNNs and aggregating the outputs for age estimation. However, although these methods use ordinal information for better performance, they take a single label as ground-truth without considering label ambiguity.

2.2. FLDL

Label distribution learning is proposed to address the label ambiguity issues. For FLDL based methods, distribution form is established before training and kept fixed during training. Their objective is to narrow the gap between the learned distribution and the fixed one. Geng et al. [8] firstly defined the label distribution by assigning a Gaussian or Triangle distribution for an instance. DLDL [5] adopted the normal distribution and learned the label distribution by minimizing a Kullback-Leibler divergence between two distributions using deep CNNs. Similar to DLDL, Liu et al. [19] employed three Gaussian label distributions to describe a face example in the yaw, pitch and roll domain respectively. DLDL-v2 [1] improved the DLDL by introducing an expectation loss from distribution to alleviate the inconsistency between the training objectives and evaluation metric. DFRs [30] connected random forests to deep neural networks and exploited the decision trees' potential to model any general form of label distributions. SP-DFRs [23] proposed self-paced regression forests to distinguish noisy and confusing facial images from regular ones, which alleviate the interference arising from them. However, these methods use a fixed form distribution to describe various instances which limits their expression ability.

2.3. ALDL

Different from FLDL based methods which assume fixed form label distributions, the distribution form for ALDL based methods is not assumed at the beginning and it is generated automatically during learning. Geng et al. [8] proposed two adaptive label distributions learning algorithms named IIS-ALDL and BFGS-ALDL respectively to automatically learn the label distributions adapted to different ages. He et al. [13] generated age label distributions through a weighted linear combination of the input image's label and its context-neighboring samples. Pan at al. [22] proposed the Mean-Variance loss, in which the mean loss penalizes the difference between the mean of the estimated distribution and the ground-truth, while the variance loss penalizes the variance of the estimated distribution to ensure a sharp distribution. However, we argue that existing ALDL methods have not strictly complied the intrinsic principles summarized in this work, which can not fully take the advantages of ALDL.

3. Methodology

In this section, we will first give a brief review of FLDL based methods and then detail our ALDL method, where a novel objective function, unimodal-concentrated loss, is proposed for highly flexible distribution learning.

3.1. Preliminaries

Formally, let x_i denote the i-th input instance with $i = 1, 2, ..., N$, $\hat{y_i}$ denote the predicted value by the network, and $y_i \in \{1, 2, ..., C\}$ denote the ground-truth label where N is the number of instances and C is the number of classes. Instead of regressing y_i directly, FLDL based methods transform y_i from a single class label to a label distribution and then predict $\hat{y_i}$ by label distribution learning. Gaussian distribution is commonly used in FLDL [1,5,7,9]. Instances with the same class label y_i share the identical Gaussian distribution. Taking Gaussian distribution $\boldsymbol{d} \sim N(\mu, \sigma^2)$ as an example

$$d_{i,j} = \frac{1}{S\sqrt{2\pi\sigma^2}} \exp(-\frac{(j-\mu)^2}{2\sigma^2}), j = 1, 2, ..., C, \quad (1)$$

where $d_{i,j}$ denotes the probability of x_i belongs to class j and $\sum_j^C d_{i,j} = 1$; μ equals to the ground-truth label y_i; σ is the standard deviation of $\boldsymbol{d}_i$; S is a normalization factor.

Let $\boldsymbol{z}_i = f(x_i; \Theta)$ denote the output of the last fully connected (FC) layer of a CNN model $f(\cdot)$, where Θ is the model parameter. Softmax operation is applied to turn output $\boldsymbol{z}_i$ into distribution $\boldsymbol{p}_i$. The elements $p_{i,j}$ of $\boldsymbol{p}_i$ is computed as

$$p_{i,j} = \frac{\exp(z_{i,j})}{\sum_{k=1}^{C} \exp(z_{i,k})}. \quad (2)$$

Kullback-Leibler (KL) divergence is usually adopted in FLDL as the loss function. KL loss (L_{KL}) is optimized to reduce the gap between the pre-defined distribution $\boldsymbol{d}_i$ and the predicted distribution $\boldsymbol{p}_i$. The final prediction $\hat{y_i}$ is obtained by taking the expectation of $\boldsymbol{p}_i$ as follows

$$\hat{y_i} = \sum_{j=1}^{C} j * p_{i,j}. \quad (3)$$

Thus, different instances with the same label are expected to predict similar distributions. It is against the nature that different instances with the same label should have their own distributions corresponding to their characteristics.

3.2. Proposed Approach

In order to tackle the issues above, we present a novel adaptive label distribution learning method which can produce unimodal and instance-aware distributions. Fig. 2

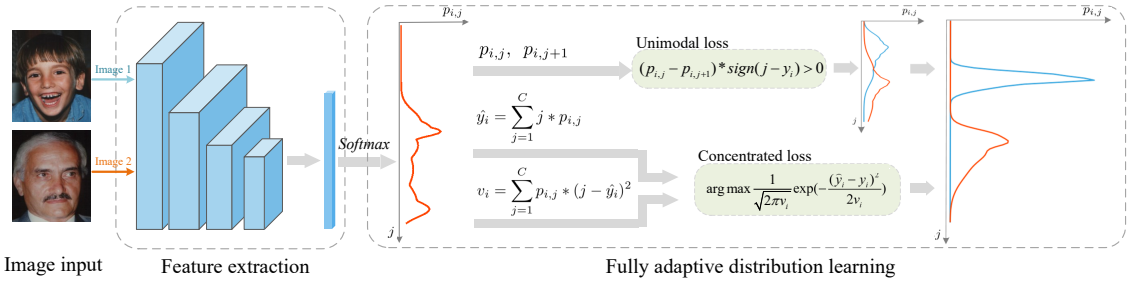

Figure 2. Overview of our proposed method. The unimodal loss makes the final predicted distribution be inclined to a mountain-liking curve with single peak, while the mean and variance of the probabilities are optimized jointly via the concentrated loss to make the predicted distribution adaptive to individual instances.

gives the overview of our approach, in which the proposed unimodel loss and concentrated loss are embedded into a exsting CNN for end-to-end learning without any additional modification on the model. The details are given below.

3.2.1 Unimodal loss

Based on the principles we have summarized previously, it is crucial to output a unimodal distribution for ordinal regression tasks. Hence, we propose a unimodal loss denoted as L_{uni}, which is formulated as follows

$$L_{uni} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C-1}\max(0, -(p_{i,j}-p_{i,j+1}) * \text{sign}[j - y_i]), \tag{4}$$

where sign$[j-y_i]$ is a sign function which equals to -1 while $j-y_i < 0$ and equals to 1 otherwise. It is desirable for value of $p_{i,j} - p_{i,j+1}$ to be negative if $j - y_i < 0$ and be positive if $j - y_i > 0$, which conforms to the properties of unimodal distribution.

Constrain distribution to be unimodal. In order to show how our unimodal loss L_{uni} performs, we take a case of $j < y_i$ (i.e. sign$[j - y_i]$= -1) for illustration, as shown in the blue region of Fig. 3, where $p_{i,j} - p_{i,j+1} > 0$. That is the adjacent probabilities are not in ascending order, and consequently the gradient of L_{uni} w.r.t. $p_{i,j}$ and $p_{i,j+1}$ can be computed respectively as

$$\frac{\partial L_{uni}}{\partial p_{i,j}} = +1, \tag{5}$$

$$\frac{\partial L_{uni}}{\partial p_{i,j+1}} = -1. \tag{6}$$

According to Eq. 5 and Eq. 6, the $p_{i,j}$ will be decreased due to its positive gradients, while $p_{i,j+1}$ will be increased due to its negative gradients. In other words, our unimodal loss L_{uni} adjusts the probabilities to make them increase monotonically before reaching the ground-truth position.

In the other direction where sign$[j - y_i]$= $+1$, our L_{uni} adjusts the probabilities to decrease monotonically after the ground-truth position. Thus, the predicted distribution will be optimized to be unimodal via L_{uni}.

Our proposed L_{uni} is superior to the softmax loss used in [22]. since L_{uni} can adjust the ranking relation within the predicted distribution while the softmax loss not. Please refer to proof in Sec. 3.2.3 for more details. Consequently, the predicted probabilities of Mean-Variance [22] are more likely to be multimodal, and the examples for comparison are given in Fig. 4.

3.2.2 Concentrated loss

According to principles discussed before, the learned distribution should maximize at the ground-truth and be adaptive for individual instances. To accomplish this goal, we propose a concentrated loss denoted as L_{con}, which integrates the difference between the estimation $\hat{y}$ and the ground-truth y and the uncertainty indictor variance of the predicted distribution together, and optimizes them jointly.

We first maximize the following likelihood for x_i

$$\Phi(\boldsymbol{p}_i; x_i, \Theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi v_i}}\exp(-\frac{(\hat{y}_i - y_i)^2}{2v_i}), \tag{7}$$

where v_i is the variance of predicted distribution $\boldsymbol{p}_i$. Based on Eq. 2 and Eq. 3, v_i can be calculated as below

$$v_i = \sum_{j=1}^{C} p_{i,j} * (j - \hat{y}_i)^2. \tag{8}$$

Then we take the negative log of $\Phi(\cdot)$ to get L_{con} as follows

$$L_{con} = -\ln(\Phi(\boldsymbol{p}_i; x_i, \Theta)) \tag{9}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\frac{1}{2}\ln v_i + \frac{(\hat{y}_i - y_i)^2}{2v_i} + \frac{1}{2}\ln 2\pi), \tag{10}$$

where constant $\frac{1}{2}\ln 2\pi$ can be omitted during optimization.

Instance-aware adaptive distribution learning. To demonstrate how it works, we take the gradient of concentrated loss L_{con} w.r.t. the variance v_i. As we all know the sample mean and variance are statistically independent of each other, for simplicity, it is computed as

$$\frac{\partial L_{con}}{\partial v_i} = \frac{1}{2v_i} - \frac{(\hat{y}_i - y_i)^2}{2v_i^2}, \tag{11}$$

where $\frac{\partial L_{con}}{\partial v_i}$ has following properties

$$\frac{\partial L_{con}}{\partial v_i} > 0, \text{ while } v_i > (\hat{y}_i - y_i)^2, \tag{12}$$

$$\frac{\partial L_{con}}{\partial v_i} < 0, \text{while } 0 < v_i < (\hat{y}_i - y_i)^2. \tag{13}$$

According to Eq. 12, the network will be optimized to decrease the intensity of v_i to make it close to $(\hat{y}_i - y_i)^2$ via its positive gradient. In this situation, $(\hat{y}_i - y_i)^2$ is an adaptive lower bound of v_i. In other words, when estimation error $(\hat{y}_i - y_i)^2$ is small which indicates an easy sample, the distribution variance v_i is decreased to be small.

According to Eq. 13, the network will be optimized to increase the intensity of v_i to make it close to $(\hat{y}_i - y_i)^2$ via its negative gradient. In this situation, $(\hat{y}_i - y_i)^2$ is an adaptive upper bound of v_i. That is to say, when estimation error $(\hat{y}_i - y_i)^2$ is large which indicates a hard sample, the distribution variance v_i is increased to be large.

Take the gradient of L_{con} w.r.t. the estimation error $\epsilon_i = (\hat{y}_i - y_i)^2$ as follows

$$\frac{\partial L_{con}}{\partial \epsilon_i} = \frac{1}{2v_i}. \tag{14}$$

According to Eq. 14, the estimation error ϵ is always optimized as small as possible via its positive gradient. Moreover, the optimization speed of ϵ is negatively correlated with the magnitude of v_i.

Finally, the estimation error and the variance of the distribution are optimized in a fully adaptive way, and consequently the learned distribution can be instance-aware. As shown in Fig. 5, the first row examples are in high quality and the second row examples are in low quality which are polluted by illumination, occlusion and heavy makeup. It is obvious that our predicted distributions can reflect the quality among faces where variances of the first row instances are small while the variances of the second ones are large.

3.2.3 Unimodal-Concentrated loss

The final objective function of our proposed approach is denoted as L_{uc} and formulated as follows

$$L_{uc} = L_{con} + \lambda * L_{uni}, \tag{15}$$

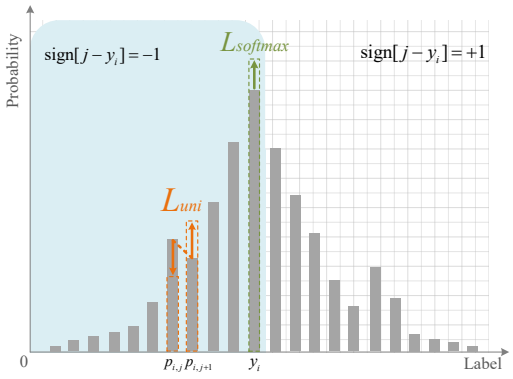

Figure 3. An illustration of how unimodal loss (orange) and softmax loss (green) affect the probability distribution respectively.

where λ is a hyper-parameter to weight the two terms.

Comparison with Mean-Variance. The Mean-Variance loss [22] can be formulated as

$$L_{m-v} = L_s + \lambda_1 L_m + \lambda_2 L_v \tag{16}$$

$$= \frac{1}{N}\sum_{i=1}^{N} -logp_{i,y_i} + \frac{\lambda_1}{2}(\hat{y}_i - y_i)^2 + \lambda_2 v_i, \tag{17}$$

where L_s is the softmax loss. To show the effect of Mean-Variance loss on the generated distribution, we take the gradient of L_{m-v} with respect to item $p_{i,j}$, $(\hat{y}_i - y_i)$ and v_i, respectively. Firstly, we take the gradient of L_{m-v} w.r.t $p_{i,j}$

$$\frac{\partial L_{m-v}}{\partial p_{i,j}} = \begin{cases} \lambda_1(\hat{y}_i - y_i)j + \lambda_2(j - \hat{y}_i)^2, & j \neq y_i \\ \frac{-1}{p_{i,j}} + \lambda_1(\hat{y}_i - y_i)j + \lambda_2(j - \hat{y}_i)^2, & j = y_i. \end{cases} \tag{18}$$

And then, we take the gradient of L_{m-v} w.r.t $\epsilon_i = (\hat{y}_i - y_i)^2$

$$\frac{\partial L_{m-v}}{\partial \epsilon_i} = \frac{1}{2}\lambda_1. \tag{19}$$

Finally, we take the gradient of L_{m-v} w.r.t v_i

$$\frac{\partial L_{m-v}}{\partial v_i} = \lambda_2. \tag{20}$$

For simplicity, we omit $\frac{1}{N}$ in Eq.18, Eq.19, Eq.20.

Base on equations above, we have three observations:

- According to Eq.18, we can see that the gradient $\frac{\partial L_{m-v}}{\partial p_{i,j}}$ and $\frac{\partial L_{m-v}}{\partial p_{i,j+1}}$ have similar expression and the direction of both $\frac{\partial L_{m-v}}{\partial p_{i,j}}$ and $\frac{\partial L_{m-v}}{\partial p_{i,j+1}}$ are irrelevant with the relative order of $p_{i,j}$ and $p_{i,j+1}$. Additionally, $\frac{\partial L_s}{\partial p_{i,j}}$ can only be non-zero when $j = y_i$ which means that the softmax loss cannot correct the wrong ordinal relationship between adjacent probabilities, see Fig. 3.

- According to Eq.20, the gradient $\frac{\partial L_{m-v}}{\partial v_i}$ is a positive constant which means that Mean-Variance loss will always optimize the variance of the predicted distribution to be small. In other words, Mean-Variance loss makes the estimated distribution as sharp as possible [22].

- According to Eq.19 and Eq.20, we can see that there is no v_i item in gradient $\frac{\partial L_{m-v}}{\partial (\hat{y}_i - y_i)^2}$ and there is no $(\hat{y}_i - y_i)$ item in gradient $\frac{\partial L_{m-v}}{\partial v_i}$. That is to say, the estimation error and variance of the predicted distribution are optimized independently without interaction.

In summary, the Mean-Variance loss does not constrain the predicted distribution to be unimodal explicitly. Besides, the minimization of Mean-Variance loss does not generate an instance-aware distribution adaptively.

4. Experiments

In this section, we will first detail the experiment settings and then compare our method with state-of-the-art works on facial age database MORPH Album II [26] and head pose databases including AFLW2000 [40] and BIWI [4].

4.1. Datasets

MORPH Album II is one of the most commonly used and largest longitudianal face databases in the public domain for age estimation, which contains 55,134 face images of 13,617 subjects and the ages range from 16 to 77 [26]. Mugshots are captured in high quality and all faces are frontal. We follow the most widely adopted evaluation protocol namely the five-fold random split (RS) protocol [1, 3, 22, 23], where 80 percent of images are randomly chosen as the training set and the remaining for testing.

IMDB-WIKI contains more than half a million labeled images of celebrities, which are crawled from IMDB and Wikipedia. Although it is the largest facial dataset with age labels, it is polluted by too much noise. Instead of using it to evaluate our method, we utilize it to pre-train our network as previous works [1, 18, 28].

AFLW2000 is one of the most commonly used benchmarks for head pose estimation [29,36,39]. The challenging AFLW2000 dataset [40] contains the first 2,000 samples of the AFLW dataset [16] which have been re-annotated with 68 3D landmarks using a 3D model for each face. The faces in the dataset have large pose variations with various occlusions, expressions as well as illumination conditions.

BIWI is collected by recording RGB-D videos of 20 different subjects across different head poses using a Kinect v2 device in a laboratory setting, and about 15,000 frames are generated with pose annotations [4] .

300W-LP dataset [40] is re-annotated from a collection of several popular in the wild facial 2D landmark datasets by fitting the 3D dense face model to the image. The database contains 61,225 samples across large poses and expands to 122,450 samples by horizontal-flipping. Following the previous works [29,36,39], we use 300W-LP dataset for network training while using AFLW2000 and BIWI for evaluation.

4.2. Implementation Details

For the age estimation task, we use VGG-16 [31] as the backbone network without modification except the dimension of the last fully-connected layer is modified to 101 for wide age range following [1, 5, 22]. All faces are cropped and resized to the 224×224 resolution. Data augmentation includes random horizontal flipping, standard color jittering and random affine transformation. The model is pre-trained on IMDB-WIKI and then fine-tuned with a learning rate lr which is initialized as lr=0.01 and decayed by a factor of 0.5 after each $10K$ iterations. the maximum number of iterations is $60K$, and batch size is set to 128. Hyper-parameter λ is 1000.

For the head pose estimation task, we directly follow the experiment settings of Hopenet [29], in which Resnet-50 [12] is chosen as the backbone network and Adam optimizer [15] is used for optimization. Please kindly refer to Hopenet for more experiment details if you need. It's worth to note that, we also make a modification like Hopenet, i.e., the output dimension is changed from 66 to 200 for the reason that the angles are in $\pm 99^{\circ}$ in fact. In this way, 31 images are discarded from AFLW2000 for their angles are out of range. Hyper-parameter λ is set to be 1000. Following [5, 22,29], we use MAE as our evaluation metric for both tasks.

4.3. Comparison with the State-of-the-arts

In this section, we compare our methods with state-of-the-art ones on Morph II, AFLW2000 and BIWI respectively. As shown in Table 1, our model for age estimation achieves 1.86 MAE with the VGG-16 backbone, which is

Table 1. Comparisons with other state-of-the-art methods on the Morph II. All results are under the five-fold random split protocol.

Method	Form	MAE
Ranking-CNN [3]	Non-LDL	2.96
BridgeNet [18]	Non-LDL	2.38
DLDL-v2 [1]	FLDL	1.97
DRFs [30]	FLDL	2.17
SPUDRFs [23]	FLDL	1.91
Mean-Variance [22]	ALDL	2.16
AVDL [35]	ALDL	1.94
Ours	ALDL	**1.86**

Table 2. Comparisons with other state-of-the-art methods on AFLW2000 and BIWI dataset. All models are trained on 300W-LP dataset.

Method	Form	AFLW2000				BIWI			
		Yaw	Pitch	Roll	Mean	Yaw	Pitch	Roll	Mean
3DDFA [40]	Non-LDL	5.40	8.53	8.25	7.39	36.17	12.25	8.77	19.06
FAN [2]	Non-LDL	6.36	12.28	8.71	9.12	8.53	7.48	7.63	7.88
Hopenet (α=2) [29]	Non-LDL	6.47	6.56	5.44	6.16	5.17	6.98	3.39	5.18
Hybrid Classification [34]	Non-LDL	4.82	6.23	5.14	5.40	-	-	-	-
FSA [36]	Non-LDL	4.50	6.08	4.64	5.07	4.27	4.96	2.76	4.00
FDN [39]	Non-LDL	3.78	5.61	3.88	4.42	4.52	4.70	**2.56**	3.93
Guo [33]	Non-LDL	-	-	-	-	**3.68**	4.36	3.02	3.69
Ours	ALDL	**3.46**	**5.24**	**3.68**	**4.13**	3.91	**3.96**	2.83	**3.57**

Table 3. The performances compared with the same backbone network but different losses.

Loss	Form	MORPH II	AFLW	BIWI
DLDL-v2	FLDL	1.90	4.20	3.80
Mean-Variance	ALDL	2.01	4.36	4.01
Ours	ALDL	**1.86**	**4.13**	**3.57**

Table 4. The results for different loss combinations of Mean-Variance and ours.

Combinations		Benchmarks		
Auxiliary	Primary	MORPH II	AFLW2000	BIWI
Softmax	Concentrated	1.92	4.25	3.61
Unimodal	Concentrated	1.86	4.13	3.57
Softmax	Mean & Variance	2.01	4.36	4.01
Unimodal	Mean & Variance	3.30	4.53	4.39

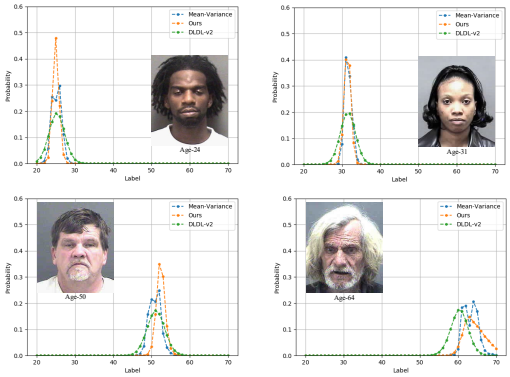

Figure 4. Age examples for comparisons with the same backbone network but different losses. Some distributions are not unimodal generated by Mean-Variance while our proposed method can ensure the unimodality of distribution. And DLDL-v2 tends to output the distributions with similar shapes.

the best performance among all methods. It is noted that compared with the FLDL based SPUDFRs [23] and ALDL based Mean-Variance [22], our result is obviously better than them which shows the effectiveness of our proposed fully adaptive label distribution learning.

As shown in Table 2, on the challenging AFLW2000 and BIWI datasets for head pose estimation, our unimodal-concentrated loss outperforms previous state-of-the-art methods such as FSA [36] and FDN [39], which further exhibits its superiority. Moreover, compared with landmark-based methods [2, 40], our method only uses pixel intensity information which is landmark-free.

Comparison with different losses. Methods for ordinal regression listed in Table 1 and Table 2 are all under different experimental settings. For fair comparison, we conduct the experiment using the VGG-16 backbone with different losses. Specifically, we choose DLDL-v2 and Mean-Variance as they achieve the convincing performances based on FLDL and ALDL respectively. As shown in Table 3, our proposed loss outperforms the DLDL-v2 and Mean-Variance loss. As viewed in Fig. 6, DLDL-v2 loss tends to output distributions with the same variance at different ages, because it learns from a fixed-form distribution with the assumed standard deviation. Mean-Variance loss generates distributions with smaller variances than ours as the distributions are optimized to be as sharp as possible. While the learned distributions of our unimodal-concentrated loss can adapt more appropriately with the facial aging. Some examples are shown in Fig. 4.

4.4. Ablation Study

4.4.1 Different Loss Combinations

Our proposed loss is composed of the unimodal loss and concentrated loss. Since it is hard to optimize the network with only one part, to demonstrate the effectiveness of each

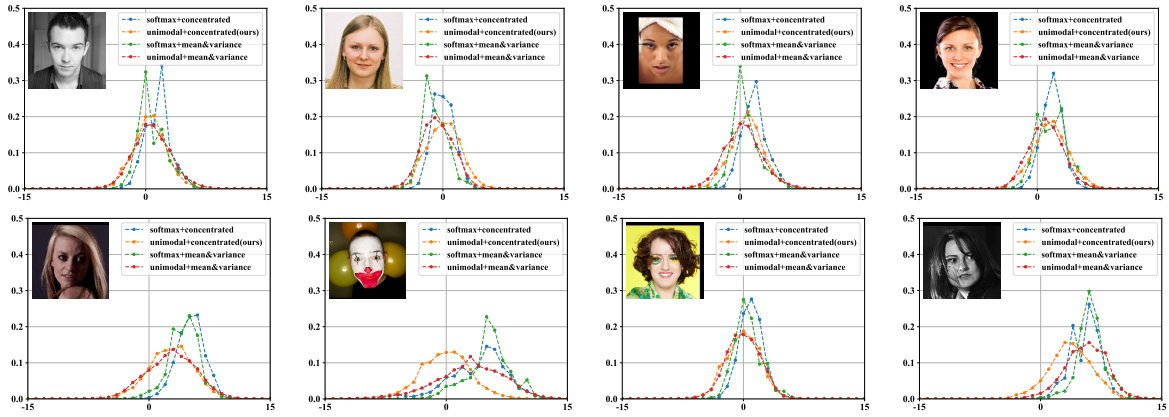

Figure 5. Examples of head pose estimation for different losses combinations when the yaw angle is zero. The first row examples are in high quality and the second row examples are in low quality which are polluted by illumination, occlusion and heavy makeup. As can be see, first, when equipped with unimodal loss, the distributions are relatively smooth and unimodal. When equipped with softmax loss, the distributions are easy to be multimodal. Second, when equipped with concentrated loss, the concentrations of distribution or the prediction uncertainties vary obviously among the samples of the same class with different quality.

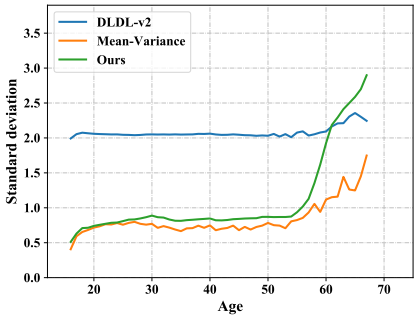

Figure 6. Average standard deviations at different labels on MORPH II. Labels with the number of samples less than 10 are discarded.

part respectively, we conduct experiments under different combinations of our loss with loss in Mean-Variance. From Table 6, we can see:

Compared with the combination of softmax and concentrated loss, the combination of unimodal and concentrated loss achieves higher performance, since the unimodal loss constrains the probabilities to be unimodal while the softmax loss not. As shown in the second image and the last image of the second row of Fig 5, the probabilities outputed by softmax+concentrated is multimodal. It verifies the effectiveness of our proposed principle II, i.e. the distribution should be unimodal.

Compared with the combination of softmax and mean loss & variance loss, the combination of softmax and concentrated loss achieves higher performance since concentrated loss takes instance-aware uncertainty into consideration instead of minimizing the mean and variance loss as small as possible in [22]. As shown in Fig 5, the confidences outputed by softmax+concentrated in the normal faces (shown in the first row) is relatively higher, and the ones in the hard faces (shown in the second row) is relatively lower. While the confidences output by softmax+mean&variance have relatively smaller difference between the two rows. It verifies the effectiveness of our proposed principle III, i.e. the label distribution should vary with the samples changing.

It is worth noting that, compared with the combination of softmax and mean loss & variance loss, although the combination of unimodal and mean loss & variance loss can make the probabilities to be unimodal as shown in Fig 5, it still gets the poorer performance. The reason is that it is hard to optimize the network with the mean & variance loss when the softmax loss is not used jointly as viewed in [22], while our concentrated loss does not have this problem. More detailed analysis can be found in the supplementary materials.

5. Conclusion

In this paper, we propose a fully adaptive distribution learning method for ordinal regression by introducing an efficient cost function called unimodal-concentrated loss. The unimodal loss ensures the unimodality of the learned distribution and the concentrated loss maximizes the probability at the ground-truth in a fully adaptive way for individual instances. Experimental results show our method outperforms previous works on MORPH II benchmark for facial age estimation, AFLW2000 and BIWI benchmarks for head pose estimation. In the future work, we would like to investigate the effectiveness of the proposed loss in other related tasks.

References

[1] Chen-Wei Xie Jianxin Wu Bin-Bin Gao, Chao Xing and Xin Geng. Age estimation using expectation of label distribution learning. In *The 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*, 2018. 2, 3, 6

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d amp; 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017. 7

[3] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 742–751, 2017. 1, 3, 6

[4] Gabriele Fanelli, Juergen Weise, Thibaut Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In *Joint pattern recognition symposium*, pages 101–110, 2011. 6

[5] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017. 3, 6

[6] Vijetha Gattupalli, Parag S. Chandakkar, and Baoxin Li. A computational approach to relative aesthetics. *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2446–2451, 2016. 2

[7] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016. 2, 3

[8] Xin Geng, Kate Smith-Miles, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:2401–2412, 2013. 2, 3

[9] Xin Geng, Qin Wang, and Yu Xia. Facial age estimation by adaptive label distribution learning. In *2014 22nd International Conference on Pattern Recognition*, pages 4465–4470, 2014. 2, 3

[10] David Gerónimo Gómez, Antonio M. López, Angel Domingo Sappa, and Thorsten Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1239–1258, 2010. 1

[11] Guodong Guo, Yun Fu, Charles R. Dyer, and Thomas S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17:1178–1188, 2008. 1

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[13] Zhouzhou He, Xi Li, Zhongfei Zhang, Fei Wu, Xin Geng, Yaqing Zhang, Ming-Hsuan Yang, and Yueting Zhuang. Data-dependent label distribution learning for age estimation. *IEEE Transactions on Image Processing*, 26(8):3846–3858, 2017. 3

[14] Du Yeong Heo, Jae Nam, and Byoung Ko. Estimation of pedestrian pose orientation using soft target training based on teacher–student framework. *Sensors*, 19(5), 2019. 2

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 6

[16] M Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE International Conference on Computer Vision Workshops*, 2012. 6

[17] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Trans Syst Man Cybern B Cybern*, 34(1):621–628, 2004. 1

[18] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. Bridgenet: A continuity-aware probabilistic network for age estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1145–1154, 2019. 6

[19] Zhaoxiang Liu, Zezhou Chen, Jinqiang Bai, Shaohua Li, and Shiguo Lian. Facial pose estimation by deep learning from label distributions. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1232–1240, 2019. 3

[20] Refik Can Malli, Mehmet Aygun, and Hazim Kemal Ekenel. Apparent age estimation using ensemble of deep learning models. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 714–721, 2016. 1

[21] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4928, 2016. 1, 3

[22] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5285–5294, 2018. 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13

[23] Lili Pan, Shijie Ai, Yazhou Ren, and Zenglin Xu. Self-paced deep regression forests with consideration on underrepresented samples. In *ECCV*, 2020. 3, 6, 7

[24] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 17–24, 2017. 1, 3

[25] Eike Rehder, Horst Kloeden, and Christoph Stiller. Head detection and orientation estimation for pedestrian safety. *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 2292–2297, 2014. 2

[26] Karl Ricanek and Tamirat Tesafaye. Morph: a longitudinal image database of normal adult age-progression. *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, 2006. 6

[27] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 252–257, 2015. 2

[28] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2016. 6

[29] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2155–215509, 2018. 6, 7

[30] Wei Shen, Yiluan Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan Loddon Yuille. Deep regression forests for age estimation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2304–2313, 2018. 3, 6

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015. 6

[32] Zheng Song, Bingbing Ni, Dong Guo, Terence Sim, and Shuicheng Yan. Learning universal multi-view age estimator using video context. *2011 International Conference on Computer Vision*, pages 241–248, 2011. 1

[33] Guo Tianchu, Zhang Hui, ByungIn Yoo, Yongchao Liu, Young jun Kwak, and Ja-Joon Han. Order regularization on ordinal loss for head pose, age and gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1496–1504, 2021. 7

[34] Yujia Wang, Wei Liang, Jianbing Shen, Yunde Jia, and Lap-Fai Yu. A deep coarse-to-fine network for head pose estimation from synthetic data. *Pattern Recognition*, 94:196–206, 2019. 7

[35] Xin Wen, Biying Li, Haiyun Guo, Zhiwei Liu, Guosheng Hu, Ming Tang, and Jinqiao Wang. Adaptive variance based label distribution learning for facial age estimation. In *ECCV*, 2020. 6

[36] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1087–1096, 2019. 6, 7

[37] Zhiguang Yang and Ai Haizhou. Demographic classification with local binary patterns. In *International Conference on Biometrics*, pages 464–473. Springer, 2007. 1

[38] Dong Yi, Zhen Lei, and S. Li. Age estimation by multi-scale convolutional network. In *ACCV*, 2014. 1, 2

[39] Hao Zhang, Mengmeng Wang, Yong Liu, and Yi Yuan. Fdn: Feature decoupling network for head pose estimation. In *AAAI*, 2020. 6, 7

[40] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and S. Li. Face alignment across large poses: A 3d solution. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016. 6, 7