Order Regularization on Ordinal Loss for Head Pose, Age and Gaze Estimation

Tianchu Guo[1*], Hui Zhang[2], ByungIn Yoo[3], Yongchao Liu[4*], Youngjun Kwak[3], Jae-Joon Han[3]
[1] Artificial Intelligence Center, DAMO Academy, Alibaba Group, Hangzhou, China
[2] Samsung Research China - Beijing (SRC-B)
[3] Samsung Advanced Institute of Technology (SAIT), South Korea
[4] Beijing ByteDance Technology Co., Ltd.
tianchu.gtc@alibaba-inc.com, hui123.zhang@samsung.com, byungin.yoo@samsung.com, liuyongchao.eric@bytedance.com, yjk.kwak@samsung.com, jae-joon.han@samsung.com

Abstract

Ordinal loss is widely used in solving regression problems with deep learning technologies. Its basic idea is to convert regression to classification while preserving the natural order. However, the order constraint is enforced only by ordinal label implicitly, leading to the real output values not strictly in order. It causes the network to learn separable feature rather than discriminative feature, and possibly overfit on training set. In this paper, we propose order regularization on ordinal loss, which makes the outputs in order by explicitly constraining the ordinal classifiers in order. The proposed method contains two parts, i.e. similar-weights constraint, which reduces the ineffective space between classifiers, and differential-bias constraint, which enforces the decision planes in order and enhances the discrimination power of the classifiers. Experimental results show that our proposed method boosts the performance of original ordinal loss on various regression problems such as head pose, age, and gaze estimation, with significant error reduction of around 5%. Furthermore, our method outperforms the state of the art on all these tasks, with the performance gain of 14.4%, 2.2% and 6.5% on head pose, age and gaze estimation respectively.

Introduction

Benefiting from the strong ability of feature representation, convolution neural network (CNN) is widely used to solve regression problems, such as head pose (Yang et al. 2019; Ruiz, Chong, and Rehg 2018), age (Li et al. 2019; Chen et al. 2017; Zhang et al. 2017b), gaze (Park et al. 2019; Krafka et al. 2016; Cheng et al. 2020), and depth estimation (Fu et al. 2018). Most researchers prefer enhanced Softmax (Gao et al. 2017) or ordinal loss (Chen et al. 2017; Fu et al. 2018) to $\mathbf{L}_2$ loss, because such loss functions quantize the continuous value to discrete value, converting the regression problem to a classification problem, which is less sensitive to outliers compared with $\mathbf{L}_2$ loss. Among them, ordinal loss is outstanding, because it preserves the property of the regression problem, which means that the farther from the ground truth the prediction, the larger the punishment.

In order to employ the ordinal loss, a continuous value gt is converted to an **ordinal label y**, which is a vector with the length of N, using the following formula:

$$y^n = \begin{cases} 1, & \text{if } (n+1) \cdot BinSize + R_{min} \leq gt \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Where $y^n (0 \leq n < N)$ is the n-th component of $\mathbf{y}$, and $BinSize$ quantizes the regression range $[R_{min}, R_{max}]$ into $N+1$ intervals. Each y^n has a corresponding binary classifier, i.e. **ordinal classifier**, and the ordinal loss is defined as the cross-entropy loss to supervise all N binary classifiers with $\mathbf{y}$.

Now we focus on the ordinal classifier. The decision plane of the n-th classifier is denoted as $g(\mathbf{w}_n, b_n) := \mathbf{w}_n^T \mathbf{x} + b_n$, for a feature $\mathbf{x}$ extracted by CNN. It judges whether the condition in Eq. 1 is satisfied. For greater regression values, more and more classifiers output 1 sequentially. Thus, intuitively there should be the following constraint:

$$g(\mathbf{w}_0, b_0) \geq g(\mathbf{w}_1, b_1) \geq \ldots \geq g(\mathbf{w}_{N-1}, b_{N-1})$$

We call it **implicit order constraint** in ordinal loss.

However, this constraint may not be satisfied in real situations. We observed that the values computed with the decision planes are not strictly in order, as shown in Fig. 1-(a). The invalid order problem may cause the classifiers easy to overfit, since the learned feature is separable rather than discriminative. Fig. 1-(b) shows the 2D geometric interpretation with a toy model consisting of three classifiers. The training samples (represented as black shapes) can be perfectly classified. However, the feature is separable rather than discriminative. Thus a test sample in star category (i.e. the red star) may be misclassified to the circle category, crossing several planes, which has larger error than misclassified to the neighbouring triangle category.

In this paper, we propose an order regularization to constrain the order of the classifiers explicitly. The basic idea is that given a $\mathbf{x}$, the output values, i.e. $\mathbf{w}_n^T \mathbf{x} + b_n, n = 0, 1 \cdots N-1$ in order can be accomplished through constraining the decision planes in order. To achieve this goal, firstly, we make the weights of all decision planes be similar by introducing similar-weights constraint, which means $\mathbf{w}_0 \approx \mathbf{w}_1 \approx \cdots \approx \mathbf{w}_{N-1}$. Secondly, we make all the bias $b_n, n = 0, 1, \cdots N-1$ in order by introducing differential-bias constraint. The 2D geometric interpretation is shown in

*Work was done when they were employed by SRC-B.

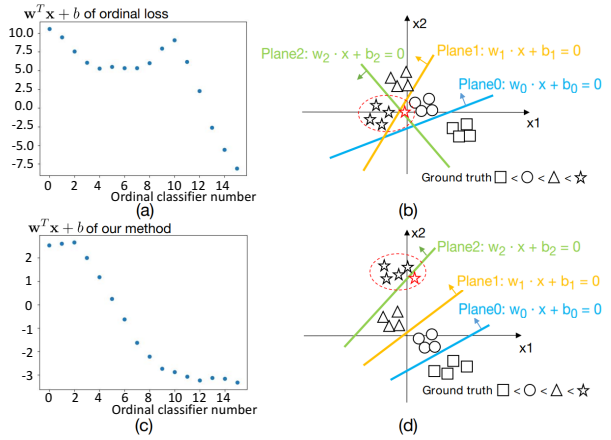


Figure 1: Visualization of the problem in ordinal loss and the benefit of our method. (a) shows the sequence of $\mathbf{w}_n^T\mathbf{x} + b_n, n = 0, 1 \cdots N-1$, which are not strictly in order, when the original ordinal loss is employed. (b) shows the 2D geometric interpretation of the feature space learned with ordinal loss. The black-coloured training samples are perfectly classified but the features are separable rather than discriminative. Thus the test sample represented by the red star may be misclassified to the circle category, which has large error. The small arrow on each plane indicates the positive side. (c) shows that the outputs of our proposed method are nearly in order. (d) is the 2D geometric interpretation of our method. The decision planes are in order, helping to learn discriminative feature. The misclassified red star has smaller error than that in (b). Best viewed in color and zoom-in.

Fig. 1-(d). The angles between the in-order planes are small, which helps to learn discriminative features. In other words, we hope to learn order-preserving feature space, where the misclassified red star is much more likely to fall in a neighbouring category with smaller error (Fig. 1-(d)), instead of falling in the category far from the ground truth (shown in Fig. 1-(b)). The output values with our method are nearly in order, as shown in Fig. 1-(c).

The proposed method has two benefits: (1) improving the interpretability of the classifiers and the feature discrimination power, (2) mitigating overfitting by adding explicit order constraint. The major contributions of this paper can be summarized as follows.

- We analyse the possible weak point of the original ordinal loss. There is no explicit order constraint, leading to invalid order of the outputs, which is easy to overfit.
- An order regularization, which consists of similar-weights constraint and differential-bias constraint, is proposed to solve this problem by making decision planes in order. Multiple variant realizations are discussed and experimentally compared to verify its novelty and effectiveness.
- Our proposed method significantly boosts the performance of the ordinal loss on multiple regression problems such as head pose, age and gaze estimation. Especially, we outperform the state of the art on all these problems without elaborate tuning for each specific task.

Related Work

CNNs are widely used to solve regression problems, e.g. head pose, age and gaze estimation. Different from classification problem, the ground truth of the regression problem is a continuous value. Though $\mathbf{L}_2$ loss is a natural choice to optimize regression problem, it is sensitive to outliers. SmoothL1 (Ren et al. 2015) and WingLoss (Feng et al. 2018) are proposed to solve this issue. By smoothing the gradient of the abnormal prediction, the training procedure is more stable. Such loss functions are widely used in object localization and landmark detection problems.

Many existing methods convert regression problem to classification problem, which is easy to optimize. Shen et al. (Shen et al. 2018) proposed a deep regression forest to divide the data with the split node. Rothe et al. (Rothe, Timofte, and Van Gool 2018) used Softmax loss to solve this problem. To get better performance, Ruiz et at. (Ruiz, Chong, and Rehg 2018) and Wang et at. (Wang et al. 2018) introduced a coarse to fine scheme. They classified the data into several coarse categories first, and then used a series of sub classifiers or regressors to refine for each coarse category. However, most of them provide the equal punishment no matter the prediction is far or near from the ground truth, which is inconsistent with the regression property. To solve this problem, Pan et at. (Pan et al. 2018) computed the expectation of the prediction according to the Softmax distribution, and then used $\mathbf{L}_2$ loss to optimize. Some researchers try to learn a distribution instead of the hard label (Gao et al. 2017; Shen et al. 2017; Yang et al. 2015).

In addition to this, ordinal loss (Chen et al. 2017; Niu et al. 2016; Chang, Chen, and Hung 2011; Guo et al. 2019) is employed benefiting from its ordered prediction. It is widely used to solve regression problems, and continuously improved in recent years. Liu et at. (Liu, Kong, and Goh 2017; Liu, Wai Kin Kong, and Keong Goh 2018) paid attention on the ordinal relationship from triplets samples to make sure the prediction ranking is correct. Besides adding constraint on the prediction ranking, some researchers also think that there is a latent function mapping the instance to a real line and there exists some bounds dividing the real line into some continuous intervals. That is similar to our problem definition. Specifically, Diaz et at. (Diaz and Marathe 2019) presented a soft target to embed ordinal information into ground truth, which outperforms the method using hard one-hot label. Liu et at. (Liu, Wang, and Kong 2019) employed Gaussian Processes regression for ordinal regression. Different from existing methods, we propose a regularization on the ordinal classifiers, which is easy to interpret and optimize. Besides, our method is evaluated on much larger datasets.

Method

Our basic idea is to make the decision planes in order. To achieve this goal, we introduce the order regularization on the original ordinal loss denoted as $Loss_o$. The proposed regularization contains two parts, i.e. the similar-weights constraint denoted as $Loss_{Plane}$, which reduces the ineffec-

tive space between decision planes, and the differential-bias constraint denoted as $Loss_{\Delta Bias}$, which enforces the decision planes in order. Both constraints are necessary and work together to achieve our goal. Finally, the total loss to be minimized is as follows

$$Loss_{total} = \lambda_p Loss_{Plane} + \lambda_\Delta Loss_{\Delta Bias} + \lambda_o Loss_o \tag{2}$$

where λ_p, $\lambda_\triangle$ and λ_o are the respective coefficients.

Original Ordinal Loss

The formula of original ordinal loss is the cross-entropy loss as follows,

$$Loss_o = \frac{1}{K}\sum_{i=1}^{K}\sum_{n=0}^{N-1} -(y_i^n log(\sigma(\mathbf{w}_n^T Net(S_i;\mathbf{W}_{net}) + b_n)) + (1-y_i^n) log(1-\sigma(\mathbf{w}_n^T Net(S_i;\mathbf{W}_{net}) + b_n))), \tag{3}$$

where S_i is a training sample, $Net(S_i;\mathbf{W}_{net})$ denotes the feature extracted by the backbone network whose parameters are $\mathbf{W}_{net}$, $\sigma(\cdot)$ is the sigmoid function, $\mathbf{w}_n$ and b_n are the decision planes' parameters (weights and bias), y_i is the corresponding ordinal label generated with Eq. 1, K is the sample number in the training set.

The prediction for a test $Sample$ is as follows,

$$predict = BS \cdot (\sum_{n=0}^{N-1} \mathcal{I}((\mathbf{w}_n^T Net(Sample;\mathbf{W}_{net}) + b_n) \geq 0) + 0.5) + R_{min}, \tag{4}$$

where BS is $BinSize$ in Eq. 1, $\mathcal{I}(\cdot)$ is indicator function.

Similar-Weights Constraint

The design idea of the similar-weights constraint is to reduce the ineffective space between decision planes. Thus, we constrain the angles between the decision planes to be small and make the $\mathbf{L}_2$ norm of all weights to be similar. Such strong constraint helps mitigate overfitting.

Specifically, we denote all weights $\mathbf{w}_0, \mathbf{w}_1 \cdots \mathbf{w}_{N-1}$ as $\mathbf{W}$, whose shape is (N, C_{in}), where N is the number of the decision planes, and C_{in} is the feature dimension. Let $\widetilde{\mathbf{w}_n} = \frac{\mathbf{w}_n}{\|\mathbf{w}_n\|_2}$ denote the normalized weight, where $n = 0, 1 \cdots N-1$. We define $\mathbf{F}$ as the following,

$$\mathbf{F} = \widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T - \mathbf{I} \tag{5}$$

where $\widetilde{\mathbf{W}}$ consists of all $\widetilde{\mathbf{w}_n}$, $\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T$ is a matrix of size $N \times N$, and $\mathbf{I}$ is the identity matrix. Note that the element in $\mathbf{F}$ computes the cosine similarity between the normal directions of two decision planes. Then we compute the plane loss $Loss_{plane}$ as the following

$$Loss_{plane} = (1 - \frac{\sum_{row=0}^{N-1}\sum_{col=0}^{N-1}\mathbf{F}_{row,col}}{N^2 - N}) + \alpha_{var} \cdot Var(\{\|\mathbf{w}_n\|_2, n = 0, 1 \cdots N-1\}) \tag{6}$$

where $\mathbf{F}_{row,col}$ is an element in $\mathbf{F}$. Thus the first part (i.e. before α_{var}) constrains the angles between decision planes to be small. $Var(\cdot)$ term computes the variance of the $\mathbf{L}_2$ norm of all $\mathbf{w}_n$. α_{var} is a hyper parameter to balance the angle and $\mathbf{L}_2$ norm terms.

The advantage of adding the $Var(\cdot)$ term is two-folded. First, constraining the $\mathbf{L}_2$ norm of all weights to be similar force the network to find a distinguishing direction for decision plane, since the value of $\mathbf{w}^T\mathbf{x}$ doesn't depend on the weight amplitude heavily. Second, compared with fixing the $\mathbf{L}_2$ norm of all weights, allowing some weight norm variation is more flexible. Experimental results shown in Section "Ablation Study" will verify the benefit of such design.

Differential-Bias Constraint

The design idea of the differential-bias constraint is to enforce the decision planes in order by making the bias values in order. We first explain the guarantee of achieving this goal. Then, we declare that its realization is not straightforward, and trivial implementation will cause that most of the bias values are eliminated. Finally, we give our solution for this elimination problem.

The Guarantee. To make the bias in order, we just make the bias in the preceding classifier is greater than the bias in the current one. In more detail, we define $\Delta b_n = b_n - b_{n+1}$, and then minimize the $Loss_{\Delta Bias}$ in the following formula,

$$Loss_{\Delta Bias} = \sum_{n=0}^{N-2}(max(0, m_{min} - \Delta b_n) + max(0, \Delta b_n - m_{max})), \tag{7}$$

where m_{min} and m_{max} are fixed hyper parameters which means the lower and upper bounds of $\Delta\mathbf{b}$.

The lower bound m_{min} is a positive value, which gives a guarantee of all the classifiers' bias in order. In addition, it also determines how large the space between the neighbouring decision planes is enough. The upper bound m_{max} restricts the bias not too large. In fact, given a $\mathbf{x}$, both weight and bias can be learned to make $\mathbf{w}_0^T\mathbf{x} + b_0$ is greater than $\mathbf{w}_1^T\mathbf{x} + b_1$. If there is no upper bound limitation, the amplitude of bias $\mathbf{b}$ may be very large, leading to limited effectiveness of weight learning.

Implementation in Practice. Traditional fully connected layer stores the bias $\mathbf{b}$ directly. Then Δb_n can be formulated as $b_n - b_{n+1}, n = 0, 1, \cdots N-2$. However, such formulation has an elimination problem in optimizing $Loss_{\Delta Bias}$. In Eq. 8 we give a sample derivation when N is equal to 5. At the beginning of the training, most Δb_n are smaller than m_{min}, so the loss is as the following

$$\begin{aligned} Loss_{\Delta Bias} &= \sum_{n=0}^{3}(max(0, m_{min} - \Delta b_n) \\ &\quad + max(0, \Delta b_n - m_{max})) \\ &= 4 \times m_{min} - (b_0 - b_1) - (b_1 - b_2) \\ &\quad - (b_2 - b_3) - (b_3 - b_4) \\ &= 4 \times m_{min} - b_0 + b_4 \end{aligned} \tag{8}$$

Only b_0 and b_4 are optimized and other bias terms are eliminated from $Loss_{\Delta Bias}$.

To solve this problem, we parametrize the bias as $\Delta b_0, \Delta b_1 \cdots b_{ref\_idx} \cdots \Delta b_{N-2}$, where $ref\_idx = Floor(N/2)$. Then bias "**b**" is computed following Eq. 9.

$$b_n = \begin{cases} b_{ref\_idx} + \sum_{j=n}^{ref\_idx-1} \Delta b_j, & \text{if } n < ref\_idx \\ b_{ref\_idx}, & \text{if } n = ref\_idx \\ b_{ref\_idx} - \sum_{j=ref\_idx}^{n-1} \Delta b_j, & \text{if } n > ref\_idx \end{cases} \quad (9)$$

We give an example to explain Eq. 9 when N is equal to 5. The proposed parametrization scheme is as the following

$$\begin{aligned} b_0 &= b_2 + \Delta b_0 + \Delta b_1 \\ b_1 &= b_2 + \Delta b_1 \\ b_2 &= b_2 \\ b_3 &= b_2 - \Delta b_2 \\ b_4 &= b_2 - \Delta b_2 - \Delta b_3 \end{aligned} \quad (10)$$

As all the Δb_n are positive constrained by Eq. 7, the computed $\{b_n\}$ are in order consequently. The benefit of this parametrization scheme will be verified in Section "Ablation Study".

Discussion

As we explained above, to make the decision planes in order, we introduce two constraints on weights and bias respectively. Here raising some questions, can the order regularization depend on only weights or bias, or without respective constraints? In this section we describe three variant methods to realize the explicit order constraint. Their experimental results will be presented in Section "Ablation Study" to show the advantage of our proposed method.

Comb. This realization employs a single, combined regularization term to constrain the decision plane outputs in order, with the following formula

$$\begin{aligned} Loss_{\Delta w\&b} = \frac{1}{K} \sum_{i=1}^{K} \sum_{n=0}^{N-2} max(0, \\ - ((\mathbf{w}_n^T \mathbf{x}_i + b_n) - (\mathbf{w}_{n+1}^T \mathbf{x}_i + b_{n+1}))) \end{aligned} \quad (11)$$

Parameters are learned by optimizing $\lambda_o Loss_o + \lambda_\Delta Loss_{\Delta w\&b}$.

B.Only. In this realization, all the decision planes are parallel and the constraint depends on only bias. In implementation, all classifiers share the same weights, denoted as $\mathbf{w}_s$. Parameters are learned by optimizing $\lambda_o Loss_o + \lambda_\Delta Loss_{\Delta Bias}$.

W.Only. In this realization, the fully connected layer's parameters only contain weights, and the order constraint is formulated as

$$Loss_{weights} = \frac{1}{K} \sum_{i=1}^{K} \sum_{n=0}^{N-2} max(0, -(\mathbf{w}_n - \mathbf{w}_{n+1})^T \mathbf{x}_i) \quad (12)$$

Parameters are learned by optimizing $\lambda_o Loss_o + \lambda_p Loss_{weights}$.

Experiments

In this section we first introduce the datasets and our evaluation protocols. After that, we will give the comparison with several state of the art methods. Then ablation study will be described, including the comparison with several different realizations of similar-weights and differential-bias constraints, as well as the variant methods described in Section "Discussion". The running time of our method is also analysed.

Dataset and Configuration

300W-LP/BIWI protocol. 300W-LP (Zhu et al. 2016) is a synthesized head pose estimation dataset which contains 122450 samples flipped from 61225 generated samples with large poses. BIWI (Fanelli et al. 2012) is a real-world head pose estimation dataset which contains about 15000 frames from 24 videos of 20 subjects captured under laboratory environment. In 300W-LP/BIWI protocol, we follow the setting of HopeNet (Ruiz, Chong, and Rehg 2018) and FSAnet (Yang et al. 2019): training on 300W-LP dataset and testing on BIWI dataset. We also follow their preprocessing by using MTCNN (Zhang et al. 2016) face detection on BIWI dataset and using only the samples within the range of $[-99°, +99°]$. Following HopeNet, we use ResNet50 (He et al. 2016) as backbone and we set the ordinal classifier's number N to 66 for all three pose angles. We train for 50 epochs with a learning rate of 1e-4 and the batch size is 16. For hyper parameters, we set λ_o, λ_p and λ_Δ to 1.0, 1.5, and 0.03 respectively. α_{var} in $Loss_{plane}$ is 0.2. m_{min} and m_{max} in $Loss_{\Delta Bias}$ are set to 0.4 and 0.8 respectively.

MORPH is a dataset for age estimation. It contains about 55000 images from about 13000 people of different races (Ricanek and Tesafaye 2006). We use the five-fold, subject-exclusive cross validation protocol (denoted as **SE**) and the VGG-16 (Simonyan and Zisserman 2014) backbone, both following (Pan et al. 2018). Note that there exists another evaluation protocol, denoted as **RS**, which uses 5-fold splitting for all images. **RS** protocol has overlapped subjects in training and testing sets, which may lead to overfitting, so it is not employed. MTCNN (Zhang et al. 2016) are employed to detect faces, and faces are aligned with the similarity transformation. The parameters of the backbone are pre-trained on IMDB-WIKI (Rothe, Timofte, and Van Gool 2018). The dimension of final feature is 4096. The hyper parameters such as learning rate also follow the settings in (Pan et al. 2018). We set the ordinal classifier's number N to 33. For hyper parameters, we set λ_o, λ_p and λ_Δ to 1.0, 0.5, and 0.01 respectively. α_{var} in $Loss_{plane}$ is 0.2. m_{min} and m_{max} in $Loss_{\Delta Bias}$ are set to 0.5 and 1.0 respectively.

MPII is a gaze dataset captured by laptop for 15 persons in everyday setting. It contains 1500 images for each person, and two eye patches are cropped from each face image. The test protocol we followed is a leave-one-person-out fashion, and single eye patch is used to predict the 3D gaze direction (Zhang et al. 2019). Note that there exists a different setting that full-faces are available as input data (Zhang et al. 2017a). Following RT-gene (Fischer, Jin Chang, and Demiris 2018), we use the VGG-16 (Simonyan and Zisserman 2014) backbone. Note that (Fischer, Jin Chang, and

Demiris 2018) used both left and right eye patches as input, processing them with two branches, while we only use one branch for single-eye input. We use the convolution parts of VGG-16 architecture followed by two fully connected layers with the size of 512 and 256, respectively. The head pose angle provided by the dataset are concatenated to the 256-dim feature. Batch normalization layer is inserted for fast convergence. We set the ordinal classifier's number N to 10 and 16 for theta "θ" and phi "ϕ" angles, respectively. We use the pretrained weights on ImageNet for the convolution parts of VGG-16. For hyper parameters, we set λ_o, λ_p and λ_Δ to 1.0, 0.5, and 0.01 respectively. α_{var} in $Loss_{plane}$ is 0.2. m_{min} and m_{max} in $Loss_{\Delta Bias}$ are set to 0.5 and 1.0 respectively.

GazeCapture is another gaze dataset captured by iphone and ipad in different orientations (Krafka et al. 2016). Following TAT (Guo et al. 2019), we use the iphone-orientation-one subset for evaluation. It contains about 400k training frames, 19k validation frames and 55k testing frames. We follow the backbone of (Guo et al. 2019). We set the ordinal classifier's number N to 15 and 28 for horizontal "x" and vertical "y" coordinates, respectively. For hyper parameters, we set λ_o, λ_p and λ_Δ to 1.0, 1.0, and 0.01 respectively. α_{var} in $Loss_{plane}$ is 0.5. m_{min} and m_{max} in $Loss_{\Delta Bias}$ are set to 0.5 and 1.0 respectively.

Comparison with State of the Art

Tab. 1a shows the head pose estimation results with 300W-LP/BIWI protocol. The experiment results show that our proposed method is a general method which can improve the original ordinal loss on different backbones. In details, we try the ordinal loss and our method on three different backbones: ResNet50 (He et al. 2016) from (Ruiz, Chong, and Rehg 2018), the FSANet architecture (Yang et al. 2019), and the EfficientNet-B0 (Tan and Le 2019). Our method boosts the performance of original ordinal loss significantly (3% to 8% for different backbones). Especially, we outperform the single-model state-of-the-art, FSANet (Yang et al. 2019), with the gain of 3.7%, by employing the same backbone. The gain is further improved to 14.4% with a more powerful backbone, EfficientNet-B0.

We give an analysis on why the performance drops when we use ordinal loss on HopeNet backbone. We guess the reason is that the feature dimension is high in HopeNet, which is 2048. High feature dimension means complex hyperplane in ordinal loss, which is difficult to optimize. The feature dimension is 48 in both FSANet and Efficient-B0 backbones (the 16-dim features from 3 stages are concatenated after capsule layer proposed in FSANet), which is more friendly for ordinal loss and our proposed method.

Tab. 1b shows the age estimation results on MORPH dataset. The accuracy comparable to state of the art is achieved with ordinal loss, demonstrating its applicability for regression problems. And our method can further improve the ordinal loss, with the error reduction of 4.7%, to set up new state of the art with the gain of 2.2%.

Tab. 1c shows the gaze estimation results on MPII dataset. Our method improves 4.7% over ordinal loss and performs the best. Also note that some reference methods employed additional information such as two-eye input (Fischer, Jin Chang, and Demiris 2018) and additional gazemap supervision (Park, Spurr, and Hilliges 2018).

Tab. 1d shows the gaze estimation results on GazeCapture dataset. Our method improves 3.9% over ordinal loss. As TAT (Guo et al. 2019) is a training scheme which is compatible with our method, we employ our loss term in the TAT training scheme, to set up the new state of the art.

We show the output values of the classifiers, i.e. $\mathbf{w}_n^T\mathbf{x} + b_n, n = 0, 1 \cdots N-1$ in Fig. 2. The output values are not strictly in order using ordinal loss. Our results are nearly in order, which demonstrates that the learned feature is discriminative.

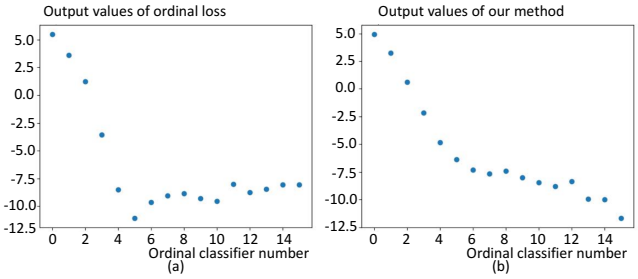


Figure 2: Output values of the classifier.(a) is the ordinal output values of phi angle from a sample in MPII dataset. They are not strictly in order. (b) is the output values of our method from the same sample. They are nearly in order. Please zoom in to view clearly.

Analysis of Fisher Criterion

The proposed order regularization enhances the Fisher criterion in the feature space along semantic dimension. Ideally, in age estimation task, feature distance between samples with age 3 and age 10 will be larger than the feature distance between samples with age 3 and age 5, because the semantics distance is larger in the first case. If the Fisher criterion between age 3 and age 5 is larger than that between age 3 and age 10, the feature distribution loses the semantic property. Thus the feature is more likely to be misclassified by other category's decision plane, as shown in Fig. 1-(b). To solve this issue, our method maintains the feature's semantic property by constraining the plane nearly parallel and thus enlarging the distance between the feature spaces along the semantic dimension. Feature distance increases gradually and Fisher criterion becomes larger for the regression values with more gap, as shown in Fig. 1-(d). Therefore, the proposed order regularization enlarges discrimination in the feature space along semantic dimension.

Ablation Study

Analysis on Variant Realizations of Similar-Weights Constraint. In Eq. 6 we allow small variation for the $\mathbf{L}_2$ norm of the weights, whose necessity is shown in Tab. 2. The methods in "Equal $\|\mathbf{w}\|_2$" column means that the $\mathbf{L}_2$ norm of all weights are the same, which includes two choices, fixed to 1.0 or equal to a learnable value S. The methods in "Dif-

Method	Yaw(°)	Pitch(°)	Roll(°)	MAE(°)
HopeNet ($\alpha = 1$) (Ruiz, Chong, and Rehg 2018)	4.81	6.61	3.27	4.90
HopeNet + Ordinal	5.51	7.56	3.68	5.59
HopeNet + Ours: gain 8%	4.96	6.90	3.54	5.13
FSA (1×1) (Yang et al. 2019)	4.78	6.24	3.31	4.31
FSA + Ordinal	4.12	5.26	3.50	4.29
FSA + Ours: gain 3%	3.97	4.87	3.62	4.15
EfficientNet-B0(Tan and Le 2019) + Ordinal	3.88	4.50	3.23	3.87
EfficientNet-B0 + Ours: gain 5%	3.68	4.36	3.02	**3.69**

(a) Head pose estimation results with 300W-LP/BIWI protocol. The performance gain vs. SOTA is 14.4%.

Method	Multi-Task (Han et al. 2017)	MV (Pan et al. 2018)	soft-ranking (Zeng et al. 2019)	MV+Ordinal	Ours: gain 4.7%
Err (years)	3.0	2.79	2.71	2.78	**2.65**

(b) Age estimation results on MORPH dataset. The performance gain vs. SOTA is 2.2%.

Method	Err (°)
MPII (Zhang et al. 2015)	6.3
GazeNet (Zhang et al. 2019)	5.5
RT-gene (1 model) (Fischer, Jin Chang, and Demiris 2018)	4.8
Pict-Gaze (Park, Spurr, and Hilliges 2018)	4.56
MeNet (Xiong, Kim, and Singh 2019)	4.9
RT-gene + Ordinal	4.71
Ours: gain 4.7%	**4.49**

(c) Gaze estimation results on MPII dataset. The performance gain vs. SOTA (RT-gene, single model) is 6.5%. Note that Pict-Gaze (Park, Spurr, and Hilliges 2018) employed additional supervision from gazemap, so it is unfair for direct performance comparison with other methods. However, we still outperform it.

Method	iTracker (Krafka et al. 2016)	TAT (Guo et al. 2019)	Ordinal	Ours: gain 3.9%	Ours+TAT
Err (cm.)	1.86	1.73	1.80	1.73	**1.71**

(d) Gaze estimation results on GazeCapture dataset. The performance gain vs. SOTA is 1.2%.

Table 1: Experimental results on four datasets: 300W-LP/BIWI, MORPH, MPII, and GazeCapture. The performance gain of our method over original ordinal loss and the gain over state of the art (SOTA) are both shown. Note that we simply update the loss term on the reference work to achieve such gain, without elaborate tuning for each specific task.

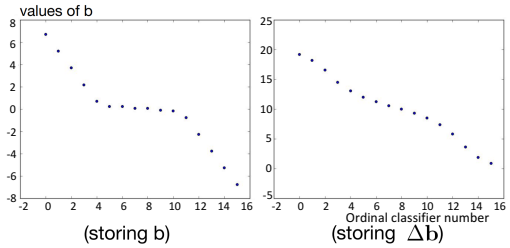
Setting	Equal $\|\mathbf{w}\|_2$		Different $\|\mathbf{w}\|_2$	
	$\|\mathbf{w}\|_2$=1.0(fixed)	$\|\mathbf{w}\|_2$= S(learnable)	$\alpha_{var} = 0.0$	α_{var} = Best Setting
MPII Err (°)	4.69	4.61	4.74	**4.49** ($\alpha_{var} = 0.2$)
GazeCapture Err (cm.)	1.79	1.75	1.80	**1.73** ($\alpha_{var} = 0.5$)

Table 2: Comparison of variant realizations of similar-weights constraint. The column "Equal $\|w\|_2$" means that the weight norms are the same, including two choices of fixed and learnable value. The column "Different $\|w\|_2$" means that the weight norms are different, including no-variation-constraint, denoted as "$\alpha_{var} = 0$", and our best setting on respective dataset.

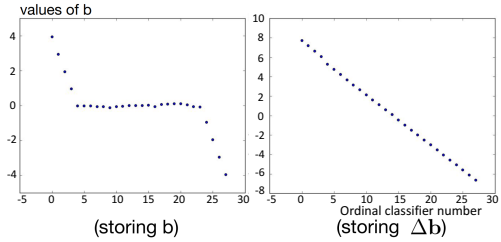
ferent $\|\mathbf{w}\|_2$" column means our method with different hyper parameter α_{var}.

All the results in "Equal $\|\mathbf{w}\|_2$" column are worse than that in "Different $\|\mathbf{w}\|_2$" with the best settings. Specifically, the results from "$\|\mathbf{w}\|_2$ = 1.0 (fixed)" are worse than that from the learnable setting, possibly because the "$\|\mathbf{w}\|_2$ = S (learnable)" realization allows the feature scale fit to the data. However, allowing variation for the weight norm, plus the punishment on too large variation with the $Var(\cdot)$ term in Eq. 6, gives the best performance. Note that the "Different $\|\mathbf{w}\|_2, \alpha_{var} = 0.0$" realization, which means that there is no constraint on weight norm variation, performs the worst. In this situation there is no constraint on $\mathbf{L}_2$ norm of the weights, which leads to the weights are trained to find an amplitude rather than the direction. In summary, the weight norm variation should be both allowed and restricted.

Analysis on Variant Realizations of Differential-Bias Constraint. In differential-bias constraint, we introduce the upper bound m_{max} in Eq. 7, and propose a new bias parametrization in Eq. 9. Their benefits are verified in Tab. 3.



(a) Bias of "ϕ" (vertical angle) on MPII.



(b) Bias of "y" (vertical offset) on GazeCapture.

Figure 3: Learned bias values with different realizations. "storing b" means the traditional realization and "storing Δb" means our realization in Eq. 9. For both datasets the bias values in the middle lose distinguishing ability with "storing b" method, which verifies the bias elimination problem mentioned in Eq. 8. This bias elimination problem may be the reason that the performance of "storing b" is worse than ours, as shown in Tab. 3. Please zoom in to view clearly.

Setting	storing **b**	storing Δ**b**	
		w/o m_{max}	Ours
MPII Err (°)	4.57	4.57	**4.49**
GazeCapture Err (cm.)	1.81	1.76	**1.73**

Table 3: Comparison of variant realizations of differential-bias constraint. "storing b" means the traditional realization and "storing Δb" means our realization in Eq. 9. "w/o m_{max}" means Eq. 7 without the m_{max} term.

The "storing **b**" means the traditional implementation which stores bias values directly. The "storing Δ**b**" means our realization in Eq. 9. The column "w/o m_{max}" means the method obtained by removing the m_{max} term in Eq. 7. The results in Tab. 3 shows that our realization performs the best.

Besides that, Fig. 3 verifies the bias elimination problem in Eq. 8. The bias in the middle has little distinguishing ability in the "storing **b**" figures on both MPII and GazeCapture datasets. We think that this is the reason why "storing **b**" performs worse than "storing Δ**b**".

In summary, we show that besides the lower bound, i.e. m_{min}, who gives a guarantee to make the bias in order, the upper bound, i.e. m_{max}, is also valuable. In addition, our new scheme for bias parametrization, as described in Eq. 9, is advantageous over the traditional bias implementation.

Comparison with Variant Methods. In Section "Discussion", we argue that both weight and bias constraints are necessary for making the classifiers in order. In Tab. 4, we will verify it with the result comparison to the variant methods defined in "Discussion". The results of "B.Only" are the worst on both MPII and GazeCapture datasets, which means that allowing some variation for the weights are important to learn distinguishing feature. The results of "W.Only" are worse than ours, which means that depending on only weights to make space between decision planes is not enough for good feature. The results of "Comb." are also worse than ours. In "Comb.", weights and bias are optimized together in a combined form, without the separate consideration like our $Loss_{plane}$ and $Loss_{\Delta Bias}$, and it is hard to converge to a good result. In summary, our proposed similar-weights and differential-bias constraints are both necessary, and they work together for discriminative feature learning.

Method	B.Only	W.Only	Comb.	Ours
MPII Err (°)	4.66	4.60	4.64	**4.49**
GazeCapture Err (cm.)	1.82	1.75	1.80	**1.73**

Table 4: Comparison with variant methods defined in Section "Discussion".

Analysis of the Running Time

Since an additional order regularization is only added on the last fully connection layer, there is no much additional time cost in the training phase. We report the following results in the head pose estimation experiments of Tab. 1a as an example. The output dimension of the fully connection layer is 66 for each of the roll, pitch and yaw angles. Experiments are conducted on a K80 GPU server, with single GPU configuration. For the time cost of one epoch, our method costs a little more time (less than one minute) than the original ordinal loss. For FSA backbone, ours vs. ordinal is 5.36 minutes vs. 4.73 minutes, +0.63 minutes (13%); and for EfficientNet backbone, ours vs. ordinal is 22.21 minutes vs. 21.35 minutes, +0.86 minutes (4%). Note that our method has the same inference algorithm as the original ordinal loss.

Conclusion

In this paper, we revisit the ordinal loss, and find its weak point, i.e. implicit order constraint leading to the invalid order of the outputs. It causes the network to learn separable feature rather than discriminative feature. To solve this issue, we propose an order regularization on ordinal loss. It consists of two parts, i.e. similar-weights constraint and differential-bias constraint. Multiple variant realizations are discussed and comparison experiments are conducted to verify its novelty and effectiveness. The proposed method improves the interpretability of the classifier and the discrimination power of the feature, and boosts the performance on various regression problems such as head pose, age and gaze estimation. Especially, we outperform the state of the art methods on all these tasks. Note that we simply update the loss term on the reference work to achieve such gain, without elaborate tuning for each specific task.

References

Chang, K.-Y.; Chen, C.-S.; and Hung, Y.-P. 2011. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR 2011*, 585–592. IEEE.

Chen, S.; Zhang, C.; Dong, M.; Le, J.; and Rao, M. 2017. Using ranking-cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5183–5192.

Cheng, Y.; Huang, S.; Wang, F.; Qian, C.; and Lu, F. 2020. A Coarse-to-Fine Adaptive Network for Appearance-Based Gaze Estimation. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Diaz, R.; and Marathe, A. 2019. Soft labels for ordinal regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4738–4747.

Fanelli, G.; Dantone, M.; Gall, J.; Fossati, A.; and Van Gool, L. 2012. Random forests for real time 3d face analysis. *International journal of computer vision* 101(3): 437–458.

Feng, Z.-H.; Kittler, J.; Awais, M.; Huber, P.; and Wu, X.-J. 2018. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2235–2245.

Fischer, T.; Jin Chang, H.; and Demiris, Y. 2018. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 334–352.

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; and Tao, D. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2002–2011.

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

Guo, T.; Liu, Y.; Zhang, H.; Liu, X.; Kwak, Y.; In Yoo, B.; Han, J.-J.; and Choi, C. 2019. A Generalized and Robust Method Towards Practical Gaze Estimation on Smart Phone. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Han, H.; Jain, A. K.; Wang, F.; Shan, S.; and Chen, X. 2017. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence* 40(11): 2597–2609.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; and Torralba, A. 2016. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2176–2184.

Li, W.; Lu, J.; Feng, J.; Xu, C.; Zhou, J.; and Tian, Q. 2019. Bridgenet: A continuity-aware probabilistic network for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1145–1154.

Liu, Y.; Kong, A. W.-K.; and Goh, C. K. 2017. Deep Ordinal Regression Based on Data Relationship for Small Datasets. In *IJCAI*, 2372–2378.

Liu, Y.; Wai Kin Kong, A.; and Keong Goh, C. 2018. A constrained deep neural network for ordinal regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 831–839.

Liu, Y.; Wang, F.; and Kong, A. W. K. 2019. Probabilistic Deep Ordinal Regression Based on Gaussian Processes. In *Proceedings of the IEEE International Conference on Computer Vision*, 5301–5309.

Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; and Hua, G. 2016. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4920–4928.

Pan, H.; Han, H.; Shan, S.; and Chen, X. 2018. Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5285–5294.

Park, S.; Mello, S. D.; Molchanov, P.; Iqbal, U.; Hilliges, O.; and Kautz, J. 2019. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 9368–9377.

Park, S.; Spurr, A.; and Hilliges, O. 2018. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 721–738.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Ricanek, K.; and Tesafaye, T. 2006. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, 341–345. IEEE.

Rothe, R.; Timofte, R.; and Van Gool, L. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision* 126(2-4): 144–157.

Ruiz, N.; Chong, E.; and Rehg, J. M. 2018. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2074–2083.

Shen, W.; Guo, Y.; Wang, Y.; Zhao, K.; Wang, B.; and Yuille, A. L. 2018. Deep regression forests for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2304–2313.

Shen, W.; Zhao, K.; Guo, Y.; and Yuille, A. L. 2017. Label distribution learning forests. In *Advances in neural information processing systems*, 834–843.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Tan, M.; and Le, Q. V. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946* .

Wang, Z.; Li, W.; Kao, Y.; Zou, D.; Wang, Q.; Ahn, M.; and Hong, S. 2018. HCR-Net: A Hybrid of Classification and Regression Network for Object Pose Estimation. In *IJCAI*, 1014–1020.

Xiong, Y.; Kim, H. J.; and Singh, V. 2019. Mixed effects neural networks (menets) with applications to gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7743–7752.

Yang, T.-Y.; Chen, Y.-T.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1087–1096.

Yang, X.; Gao, B.-B.; Xing, C.; Huo, Z.-W.; Wei, X.-S.; Zhou, Y.; Wu, J.; and Geng, X. 2015. Deep label distribution learning for apparent age estimation. In *Proceedings of the IEEE international conference on computer vision workshops*, 102–108.

Zeng, X.; Ding, C.; Wen, Y.; and Tao, D. 2019. Soft-ranking Label Encoding for Robust Facial Age Estimation. *arXiv preprint arXiv:1906.03625* .

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

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4511–4520.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017a. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 51–60.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2019. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(1): 162–175.

Zhang, Y.; Liu, L.; Li, C.; et al. 2017b. Quantifying facial age by posterior of age comparisons. *arXiv preprint arXiv:1708.09687* .

Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; and Li, S. Z. 2016. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 146–155.