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

Tianchu Guo1*, Hui Zhang2, ByungIn Yoo3, Yongchao Liu4*, Youngjun Kwak3, Jae-Joon Han3

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; Krafta 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 $L_2$ 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 $L_2$ 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:

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

Where $y^n(0 \leq n < N)$ is the $n$-th component of $y$, and \text{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 $y$.

Now we focus on the ordinal classifier. The decision plane of the $n$-th classifier is denoted as $g(w_n, b_n) := w^T_n x + b_n$, for a feature $x$ 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:

$$g(w_0, b_0) \geq g(w_1, b_1) \geq \ldots \geq g(w_{N-1}, b_{N-1})$$

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 $x$, the output values, i.e. $w^T_n x + b_n, n = 0, 1, \ldots, N - 1$ 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 $w_0 \approx w_1 \approx \ldots \approx w_{N-1}$. Secondly, we make all the bias $b_n, n = 0, 1, \ldots, N - 1$ in order by introducing differential-bias constraint. The 2D geometric interpretation is shown in
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 $L_2$ 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 al. (Ruiz, Chong, and Rehg 2018) and Wang et al. (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 al. (Pan et al. 2018) computed the expectation of the prediction according to the Softmax distribution, and then used $L_2$ 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 al. (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 al. (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 al. (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-

![Figure 1: Visualization of the problem in ordinal loss and the benefit of our method. (a) shows the sequence of $w_i^T x + b_i$, $n = 0, 1 \cdots N - 1$, 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.](image)
tive space between decision planes, and the differential-bias constraint denoted as $Loss_{\Delta \text{Bias}}$, 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

$$Loss_{\text{total}} = \lambda_p Loss_{\text{Plane}} + \lambda_\Delta Loss_{\Delta \text{Bias}} + \lambda_\alpha Loss_o$$  \hspace{1cm} (2)$$

where $\lambda_p$, $\lambda_\Delta$ and $\lambda_o$ are the respective coefficients.

**Original Ordinal Loss**

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

$$Loss_o = \frac{1}{K} \sum_{i=1}^{K} \sum_{n=0}^{N-1} - (y_i^o \log(\sigma(W_n^T Net(S_i; W_{net}) + b_n)) $$

$$+ (1 - y_i^o \log(1 - \sigma(W_n^T Net(S_i; W_{net}) + b_n)))$$

where $S_i$ is a training sample, $Net(S_i; W_{net})$ denotes the feature extracted by the backbone network whose parameters are $W_{net}$, $\sigma(\cdot)$ is the sigmoid function, $w_n$ 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,

$$predict = BS \cdot \sum_{n=0}^{N-1} I((w_n^T Net(Sample; W_{net}) + b_n) \geq 0) + 0.5) + R_{\text{min}},$$

where $BS$ is $\text{BinSize}$ in Eq. 1, $I(\cdot)$ 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 $L_2$ norm of all weights to be similar. Such strong constraint helps mitigate overfitting.

Specifically, we denote all weights $w_0, w_1, \cdots, w_{N-1}$ as $W$, whose shape is $(N, C_{in})$, where $N$ is the number of the decision planes, and $C_{in}$ is the feature dimension. Let $\tilde{w}_n = \frac{w_n}{\|w_n\|_2}$ denote the normalized weight, where $n = 0, 1 \cdots N - 1$. We define $F$ as the following,

$$F = WW^T - I$$

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

$$Loss_{\text{plane}} = \left(1 - \sum_{row=0}^{N-1} \sum_{col=0}^{N-1} \frac{F_{row,col}}{N^2} \right) + \alpha_{var} \cdot Var\{\|\tilde{w}_n\|_2, n = 0, 1 \cdots N - 1\}$$

where $F_{row,col}$ is an element in $F$. Thus the first part (i.e. before $\alpha_{var}$) constrains the angles between decision planes to be small. $Var(\cdot)$ term computes the variance of the $L_2$ norm of all $w_n$, $\alpha_{var}$ is a hyper parameter to balance the angle and $L_2$ norm terms.

The advantage of adding the $Var(\cdot)$ term is two-folded. First, constraining the $L_2$ norm of all weights to be similar force the network to find a distinguishing direction for decision plane, since the value of $w_n^T x$ doesn’t depend on the weight amplitude heavily. Second, compared with fixing the $L_2$ 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 this elimination problem. In more detail, we define $\Delta b_n = b_n - b_{n+1}$, and then minimize the $Loss_{\Delta \text{Bias}}$ in the following formula,

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

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

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 $x$, both weight and bias can be learned to make $w_0^T x + b_0$ is greater than $w_{n}^T x + b_1$. If there is no upper bound limitation, the amplitude of bias $b$ may be very large, leading to limited effectiveness of weight learning.

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

$$Loss_{\Delta \text{Bias}} = \sum_{n=0}^{3} \left(\max(0, m_{\min} - \Delta b_n) + \max(0, \Delta b_n - m_{\max})\right)$$

$$= 4 \times m_{\min} - (b_0 - b_1) - (b_1 - b_2)$$

$$- (b_2 - b_3) - (b_3 - b_4)$$

$$= 4 \times m_{\min} - b_0 + b_4$$

1498
Only $b_0$ and $b_4$ are optimized and other bias terms are eliminated from $\text{Loss}_{\Delta \text{Bias}}$.

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

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

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

$$b_0 = b_2 + \Delta b_0 + \Delta b_1$$
$$b_1 = b_2 + \Delta b_1$$
$$b_2 = b_2$$
$$b_3 = b_2 - \Delta b_2$$
$$b_4 = b_2 - \Delta b_2 - \Delta b_3$$

As all the $\Delta b_n$ 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

$$\text{Loss}_{\Delta \text{wkb}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=0}^{N-2} \max(0, -((w_n^T x_i + b_n) - (w_{n+1}^T x_i + b_{n+1})))$$

Parameters are learned by optimizing $\lambda_w \text{Loss}_o + \lambda_{\Delta \text{Loss}_{\Delta \text{wkb}}}$. 

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 $w_i$. Parameters are learned by optimizing $\lambda_b \text{Loss}_o + \lambda_{\Delta \text{Loss}_{\Delta \text{Bias}}}$. 

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

$$\text{Loss}_{\text{weights}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{n=0}^{N-2} \max(0, -(w_n - w_{n+1})^T x_i)$$

Parameters are learned by optimizing $\lambda_w \text{Loss}_o + \lambda_{\text{Loss}_{\text{weights}}}$. 

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 analyzed.

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 FSNet (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 $[-99^\circ, +99^\circ]$. 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 $\lambda_{\text{var}}, \lambda_p, \lambda_{\Delta}$ to 1.0, 1.5, and 0.03 respectively. $\text{var}$ in $\text{Loss}_{\text{plane}}$ is 0.2. $m_{\text{min}}$ and $m_{\text{max}}$ in $\text{Loss}_{\Delta \text{Bias}}$ 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 $\lambda_{\text{var}}, \lambda_p, \lambda_{\Delta}$ to 1.0, 0.5, and 0.01 respectively. $\text{var}$ in $\text{Loss}_{\text{plane}}$ is 0.2. $m_{\text{min}}$ and $m_{\text{max}}$ in $\text{Loss}_{\Delta \text{Bias}}$ 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 “\( \theta \)” and phi “\( \phi \)” angles, respectively. We use the pretrained weights on ImageNet for the convolution parts of VGG-16. For hyper parameters, we set \( \lambda_\theta \), \( \lambda_\phi \) and \( \lambda_\Delta \) to 1.0, 0.5, and 0.01 respectively. \( \alpha_{\text{var}} \) in \( \text{Loss}_{\text{plane}} \) is 0.2. \( m_{\text{min}} \) and \( m_{\text{max}} \) in \( \text{Loss}_{\Delta \text{Bias}} \) 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 \( \lambda_\theta \), \( \lambda_\phi \) and \( \lambda_\Delta \) to 1.0, 1.0, and 0.01 respectively. \( \alpha_{\text{var}} \) in \( \text{Loss}_{\text{plane}} \) is 0.5. \( m_{\text{min}} \) and \( m_{\text{max}} \) in \( \text{Loss}_{\Delta \text{Bias}} \) 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/BWI 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. \( w^T x + b_n, n = 0, 1 \cdots N - 1 \) 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.

![Output values of the classifier](image)

**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 \( L_2 \) norm of the weights, whose necessity is shown in Tab. 2. The methods in “Equal \( \| w \|_2 \)” column means that the \( L_2 \) 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-
HopeNet (α = 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 x 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%.

<table>
<thead>
<tr>
<th>Method</th>
<th>Multi-Task (Han et al. 2017)</th>
<th>MV (Pan et al. 2018)</th>
<th>soft-ranking (Zeng et al. 2019)</th>
<th>MV+Ordinal</th>
<th>Ours: gain 4.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Err (years)</td>
<td>3.0</td>
<td>2.79</td>
<td>2.71</td>
<td>2.78</td>
<td><strong>2.65</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>iTracker (Krafka et al. 2016)</th>
<th>TAT (Guo et al. 2019)</th>
<th>Ordinal</th>
<th>Ours: gain 3.9%</th>
<th>Ours+TAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Err (cm.)</td>
<td>1.86</td>
<td>1.73</td>
<td>1.80</td>
<td>1.73</td>
<td><strong>1.71</strong></td>
</tr>
</tbody>
</table>

(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.

<table>
<thead>
<tr>
<th>Setting</th>
<th>MPII Err (°)</th>
<th>GazeCapture Err (cm.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal</td>
<td>Different</td>
<td>Equal</td>
</tr>
<tr>
<td>$|w|_2=1.0$ (fixed)</td>
<td>$|w|_2=S$ (learnable)</td>
<td>$\alpha_{var}=0.0$</td>
</tr>
<tr>
<td>$\alpha_{var}=0.0$</td>
<td>$\alpha_{var}=0.0$</td>
<td></td>
</tr>
<tr>
<td>4.69</td>
<td>4.61</td>
<td>4.74</td>
</tr>
<tr>
<td>1.79</td>
<td>1.75</td>
<td>1.80</td>
</tr>
</tbody>
</table>

(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.

All the results in “Equal $\|w\|_2$” column are worse than that in “Different $\|w\|_2$” with the best settings. Specifically, the results from “$\|w\|_2 = 1.0$ (fixed)” are worse than that from the learnable setting, possibly because the “$\|w\|_2 = S$ (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 $\|w\|_2, \alpha_{var} = 0.0$” realization, which means that there is no constraint on weight norm variation, performs the worst. In this situation there is no constraint on $L_2$ 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.
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.

**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.


