Classification of ordinal data in deep learning: an experimental study

Victor-Manuel Vargas, Pedro-Antonio Gutiérrez and César Hervás

***Abstract*—The abstract goes here.**

***Index Terms*—IEEE, IEEEtran, journal, LATEX, paper, template.**

I. Introduction

DEEP LEARNING, introduced by Yann Lecun [1], combines multiple Machine Learning techniques and allows computational models that are composed of numerous processing layers to learn representations of data with various levels of abstraction. These methods have dramatically improved the state-of-the-art in many domains, such as image classification [2]–[4], speech recognition [5], control problems [6], object detection [7], [8], privacy protection [9], recovery of human pose [10], semantic segmentation [11] and image retrieval [12]. Deep learning discovers complex structures in large datasets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the features in the previous layer.

Convolutional neural networks (ConvNets) are designed to process data that comes in the form of multiple arrays. ConvNets are suited for images, video, speech and audio processing, and they have been used extensively in the last years for automatic classification tasks [13]–[15]. On image classification tasks, each colour channel is represented by a 2D array. In this case, convolutional layers extract the main features from the pixels of the images and, after that, a fully connected layer classify every sample based on its extracted features. At the output of the convolutional net, a softmax function provides the probabilities of the set of classes predefined in the model for classification tasks. These outputs are compared against the correct values.

The backpropagation algorithm is a stochastic gradient descent algorithm that minimises a predefined loss function. It has been used in the last years in many works [4], [16]–[18] for training shallow and deep neural networks. It updates the layer's parameters after backpropagating the loss function gradients through the network. Learning rate hyper-parameter controls the strength of the changes that are applied to those parameters. Some works have checked the importance of finding the optimal value for this parameter [19] and have tried different approaches to try to improve the training process [20].

Batch normalization is another technique that is used for this kind of networks. It reduces the internal covariate shift by normalizing layer inputs. It was presented in 2015 [21] and gives a critical enhancement to the training phase. Batch size is an important hyper-parameter that should be adjusted as it affects the layer's parameters updates and also the normalization [22][23].

Ordinal classification problems are those where labels are ordered, and there are different inter-classes weights for each pair of classes. This kind of problem can be treated as a nominal classification problem, but when you do this, you are the discarding ordinal information. A better approach is to use specific methods that take into account this kind of information to improve the performance of the classification model. One way to use the ordinal information is to evaluate the model using an ordinal metric. Multiple metrics exist in the literature of machine learning and statistics [24], [25]. Kappa index is a well-known statistic coefficient defined by Cohen [26] to measure inter-rater agreement on classifying elements into a set of categories. Later, Weighted Kappa (WK) is a modified version of the Kappa statistic calculated to allow assigning different weights to different levels of aggregation between two variables. weighted kappa loss was described in previous work [18] and is a cost function that is based on the WK metric. These functions are indicated for problems where different inter-classes weights are assigned, and that work proved that using these functions improves the model performance and reduce the overfitting risk. This kind of problems is associated with ordinal data processing. Those weights are predefined and depend on the type of the chosen WK. In a linear penalization, the weight is proportional to the distance between the predicted and the real class. In the Quadratic Weighted Kappa (QWK) the penalization is proportional to the square of the distance. In this work, we will combine the QWK metric and the QWK loss.

Softmax function is widely used for the output layer in neural networks for classification tasks. It is a simple and efficient function for multi-class problems but not the best when working with ordinal data. In this paper, the Proportional Odds Model (POM) [27], [28] will be used instead of the softmax function. Different link functions will be explored to compare their performance. Also, the influence of other parameters like the learning rate of the optimization algorithm and the batch size will be studied. We will contrast the results obtained with statistical analysis to provide more robust results. An approximated ANOVA III test will be performed because of the limitations of the computational time required to run a higher number of executions.

The experiments will be run using two different ordinal datasets: Diabetic Retinopathy [18], that contains high-

resolution fundus images related with diabetes disease, and Adience [29], that is formed of human faces images that are assigned an age range.

This paper is organised as follows: in Section II, we take a look at previous works related to this paper, in Section V we describe the model, the experiments and the datasets used, in Section VI we present the results obtained and the statistical analysis and finally in Section VII we expose the conclusions of this work.

II. Related work

There are many works related to the application and development of convolutional neural networks models. Few works treat ordinal classification problems, though.

J. de la Torre et al. [18] proposed the use of QWK loss function for the optimization algorithm. They compared this cost function against the traditional log-loss function across three different databases, including the Diabetic Retinopathy database as the most complex one. They proved that their function could improve the results as it reduces the overfitting and the required training time. Also, they checked the importance of hyper-parameter tuning.

Christopher Beckham and Christopher Pal [29] proposed a straightforward technique to constrain discrete ordinal probability distributions to be unimodal via the use of the Poisson and binomial probability distributions. They evaluated this approach in the context of deep learning on two large ordinal image datasets, including the Adience dataset used in this paper, and they obtained promising results.

At present, there is a great discussion in terms of finding the best activation function that offers good performance and mitigates the effects of the gradients vanishing problem [30], [31]. Rectified Linear Unit (ReLU) [32] is widely used in most deep learning works, but recently, Clevert et al. proposed the Exponential Linear Unit (ELU) [33]. They proved that ELUs alleviate the vanishing gradient problem via the identity for positive values. In their experiments, ELUs lead not only to faster learning but also to significantly better generalization performance than ReLUs and LReLUs on networks with more than five layers. That's why we decided to use the ELU function for the experiments described in this paper.

Sergey Ioffe and Christian Szegedy [21] described Batch Normalization and its benefits in previous work. It reduces the internal covariate shift by normalizing layer inputs. Their method draws its strength from making normalization a part of the model architecture and performing the normalization for each training batch. It allows them to use higher learning rates and be less careful about initialization. It also eliminates the need for using regularization techniques like Dropout.

III. Ordinal problem definition

The ordinal problem consists of predicting the label y of an input vector x, where $x \in \mathcal{X} \subseteq \mathbb{R}^K$ and $y \in \mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_Q\}$, i.e. x is in a K-dimensional input space, and y is in a label space of Q different labels. The objective of the ordinal problem is to find a function $r : \mathcal{X} \to \mathcal{Y}$ to predict the labels or categories of new patterns, given a training set of N points, $D = \{(x_i, y_i), i = 1, ..., N\}$. A natural label ordering exists for ordinal problems: $\mathcal{C}_1 \prec \mathcal{C}_2 \prec ... \prec \mathcal{C}_Q$. The assumption of order between class labels makes that two different elements of $\mathcal{Y}$ can always be compared by using the relation $\prec$, which is not possible under the nominal classification setting. In regression (where $y \in \mathbb{R}$), real values in $\mathbb{R}$ can be ordered by the standard $<$ operator, but labels in ordinal regression ($y \in \mathcal{Y}$) do not carry metric information, so the category serves as a qualitative indication of the pattern rather than a quantitative one.

IV. Proportional Odds Model

Arising from a statistical background, the Proportional Odds Model (POM) is one of the first models designed explicitly for ordinal regression [34], dated back to 1980. It is a member of a wider family of models recognised as Cumulative Link Models (CLMs) [28]. CLMs predict probabilities of groups of contiguous categories, taking the ordinal scale into account. In this way, cumulative probabilities $P(y \prec Cj|x)$ are estimated, which can be directly related to standard probabilities:

$$P(y \preceq \mathcal{C}_q|x) = P(y = \mathcal{C}_1|x) + ... + P(y = \mathcal{C}_q|x),$$

$$P(y = \mathcal{C}_q|x) = P(y \preceq \mathcal{C}_q|x) - P(y \preceq \mathcal{C}_{q-1}|x),$$

with $q = 2, ..., Q$, and considering that $P(y = \mathcal{C}_1|x) = P(y \preceq \mathcal{C}_1|x)$ and $P(y \preceq \mathcal{C}_Q|x) = 1$.

The model in inspired by the notion of latent variable, considering a linear transformation $f(x) = w^T x$ of the input variables as the one-dimensional mapping. The decision rule $r : \mathcal{X} \to \mathcal{Y}$ is not fitted directly, but stochastic ordering of space $\mathcal{X}$ is satisfied by the following general model form [35]:

$$g^{-1}(P(y \preceq \mathcal{C}_q|x)) = b_q - w^T x, \quad 1, ..., Q-1,$$

where $g^{-1} : [0, 1] \to (-\infty, +\infty)$ is a monotonic function often termed as the inverse link function and b_q is the threshold defined for class $\mathcal{C}_q$. Consider the latent variable $y* = f(x)* = w^T x + \epsilon$, where ϵ is the random error component. The most common choice for the distribution of ϵ is the logistic function (which is the default function for POM). Label $\mathcal{C}_q$ is observed if and only if $f(x) \in [b_{q-1}, b_q]$, where the function f and $b = (b_0, b_1, ..., b_{Q-1}, b_Q)$ are to be determined from the data. it is assumed that $b_0 = -\infty$ and $b_Q = +\infty$, so the real line defined by $f(x), x \in \mathcal{X}$, is divided into Q consecutive intervals. Each region corresponds to a category. The constraints $b_1 \leq b_2 \leq ... \leq b_{Q-1}$ ensures that $P(y \preceq \mathcal{C}_q|x)$ increases with q [34]. In this work, we use this ordinal model with different link functions, including logit, probit and complementary log-log. These functions will be detailed in Section V-B.

V. Experiments

A. Data

We make use of two ordinal datasets appropriate for deep neural networks:

- *Diabetic Retinopathy (DR)*[1]. DR is a dataset consisting of extremely high-resolution fundus image data. The training set consists of 17563 pairs of images (where a pair consists of a left and right eye image corresponding to a patient). In this dataset, we try and predict from five levels of diabetic retinopathy: no DR (25810 images), mild DR (2443 images), moderate DR (5292 images), severe DR (873 images), or proliferative DR (708 images). The images are taken in variable conditions: by different cameras, varying conditions of illumination and different resolutions. These images come from the EyePACS dataset that was used in a Diabetic Retinopathy Detection competition that was hosted on the Kaggle platform. Also, this dataset was used in later works [18], [36] and ordinal techniques (such as an ordinal cost function) were applied in order to achieve better performance. A validation set is set aside, consisting of 10% of the patients in the training set. The images are resized to 128 by 128 pixels. Data augmentation techniques are applied to achieve a higher number of samples.
- *Adience*[2]. This dataset consists of 26580 faces belonging to 2284 subjects. We use the form of the dataset where faces have been pre-cropped and aligned. The dataset was preprocessed, using the methods described in a previous work [29], so that the images are 256 pixels in width and height, and pixels values follow a $(0; 1)$ normal distribution. The original dataset is split into five cross-validation folds. The training set consists of merging the first four folds which comprise a total of 15554 images. From this, 10% of the images are held out as part of a validation set. The last fold is used as test set.

B. The model

A convolutional neural network (CNN) has been used for both datasets. The architecture of this CNN is presented in the Table I.

Every convolutional layer is followed by an ELU activation layer [33] and a batch normalization [21]. At the output, a Proportional Odds Model is used with different link functions [28]. The logit link function is commonly used within POM. In this paper, we are comparing other link functions like probit or complementary log-log with the logit link. These three types of links are explained below and represented in Figure 1.

- *Logit.* Logit link function is the most widely used function for Proportional Odds Models. The logit link is shown in 1.

$$logit[P(y \preceq \mathcal{C}_q)] = b_q - w^T x, \quad 1, ..., Q-1 \qquad (1)$$

[1]https://www.kaggle.com/c/diabetic-retinopathy-detection/data
[2]http://www.openu.ac.il/home/hassner/Adience/data.html

TABLE I
Description of the architecture used in the experiments. For convolutional layers, Conv_N_WxH, where N is the number of filters, W the filter width and H the filter height. Stride is 1 for every convolutional layer. For max pool layers, MaxPool_SxS, where S is the pool size.

Layer	Output shape
Conv_32_3x3	254x254x32
Conv_32_3x3	252x252x32
MaxPool_2x2	126x126x32
Conv_64_3x3	124x124x64
Conv_64_3x3	122x122x64
MaxPool_2x2	61x61x64
Conv_128_3x3	59x59x128
Conv_128_3x3	57x57x128
MaxPool_2x2	28x28x128
Conv_128_3x3	26x26x128
Conv_128_3x3	24x24x128
MaxPool_2x2	12x12x128
Conv_128_4x4	9x9x128
Dense_1_output	1

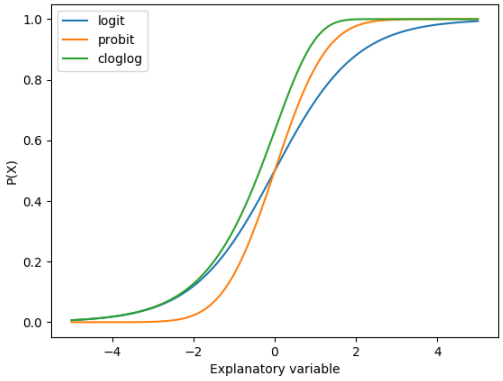

Fig. 1. Representation of link functions.

- *Probit.* Probit link function is the inverse of the standard normal cummulative distribution function (cdf). Its expression is shown in 2.

$$\Phi^{-1}[P(y \preceq \mathcal{C}_q)] = b_q - w^T x, \quad 1, ..., Q-1 \qquad (2)$$

- *Complementary log-log.* Like the logit and the probit transformation, the complementary log-log transformation takes a response that is restricted to the (0,1) interval and converts it into something in $(-\infty, +\infty)$ interval, but unlike the logit and the probit, the complementary log-log function is not symmetric. With a continuous predictor x, for example, $P(Y \leq j)$ approaches 0 at a different rate than it approaches 1. Complementary log-log expression is shown in 3.

$$log[-log[1 - P(y \preceq \mathcal{C}_q)]] = b_q - w^T x, \quad 1, ..., Q-1 \qquad (3)$$

When the given data is not symmetric in the [0,1] interval and increase slowly at small to moderate value, but increases sharply near 1, the logit and probit models are inappropriate. However, in this situation, the complementary log-log model might give satisfying results.

C. Procedure

The model is optimized using a batch based first-order optimization algorithm called Adam [37]. We study different initial learning rates in order to find the optimal one for each problem. We apply an exponential decay across training epochs to the initial learning rate.

Both datasets have been artificially equalised using data augmentation techniques [4], [38] based on image cropping and zooming, horizontal and vertical flipping, brightness adjustment and random rotations. In the case of Diabetic Retinopathy Detection, the epoch size has been fixed to 100000 images per epoch. For the Adience dataset, epoch size is the number of images in the training set.

The model is evaluated with Quadratic Weighted Kappa metric (QWK) [39]. This evaluation measure gives a higher weight to the errors that are further from the correct class. Quadratic weighted kappa loss is considered as loss function for the optimizer as it gives better performance for ordinal classification problems.

Experiments were run with the standard cross-entropy loss and the softmax function too in order to prove the performance improvement of the QWK loss and the Proportional Odds Model. The results of these experiments are analysed in Section VI-D.

D. Factors

Three different factors have been considered: learning rate, batch size and link function for the final output layer.

- *Learning rate*. Learning rate is one of the most critical hyper-parameters to tune for training deep neural networks. Optimal learning rate can vary depending on the dataset and the CNN architecture. Previous works have presented some techniques that adjust this parameter in order to achieve better performance [19], [20]. Within this work, we have considered three different values for this parameter: 10^{-2}, 10^{-3} and 10^{-4}.
- *Batch size*. Batch size is also an important parameter as it controls the number of weight updates that are made on every epoch. It can affect the training time and the model performance. In this paper, we have tried three separate batch sizes: 5, 10 and 15.
- *Link function*. Different link functions have been used for the POM at the last layer output: logit, probit and complementary log-log.

VI. Results

In this section, we present the results of the experiments. For each dataset, we show a table with the detailed experiments done training the model with each combination of parameters. Every parameter combination has been run five times. These tables show the average quadratic weighted kappa across these five executions for validation and test values.

TABLE II
Diabetic Retinopathy results. BS stands for Batch Size, LR for Learning Rate and LF for link function.

BS	LR	LF	κ_{val}	κ_{test}
5	10^{-2}	Logit	0.44888	0.41630
		Probit	0.46724	0.45972
		c log-log	0.42854	0.41448
	10^{-3}	Logit	0.56496	0.55356
		Probit	0.57084	0.56392
		c log-log	0.54796	0.53436
	10^{-4}	Logit	0.52884	0.52032
		Probit	0.53802	0.52302
		c log-log	0.53824	0.51974
10	10^{-2}	Logit	0.54970	0.53142
		Probit	0.52124	0.50806
		c log-log	0.42760	0.42270
	10^{-3}	Logit	0.58036	0.57686
		Probit	0.56536	0.55810
		c log-log	**0.58830**	**0.58150**
	10^{-4}	Logit	0.53158	0.53850
		Probit	0.54560	0.54082
		c log-log	0.53928	0.53742
15	10^{-2}	Logit	0.55828	0.55076
		Probit	0.54112	0.53430
		c log-log	0.56676	0.56428
	10^{-3}	Logit	0.55748	0.55106
		Probit	0.58182	0.57894
		c log-log	0.56340	0.55938
	10^{-4}	Logit	0.54686	0.54342
		Probit	0.53258	0.53278
		c log-log	0.53900	0.53828

A. Diabetic Retinopathy

Detailed results for the Diabetic Retinopathy dataset are presented in Table II.

The best mean QWK value was obtained with the complementary log-log link function using a batch size of 10 and a learning rate of 10^{-3}.

B. Adience

Results for the experiments made with the Adience dataset are shown in Table III.

The best mean QWK value was obtained with the probit link function using a batch size of 15 and a learning rate of 10^{-2}.

C. Statistical analysis

In this subsection, a statistical analysis will be performed in order to obtain conclusions from the results presented in this section.

The significance and relative importance of the parameters concerning the results obtained, as well as suitable values for each, were obtained using the ANalysis Of the VAriance (ANOVA) test.

The ANOVA test [40] is one of the most widely used statistical techniques. ANOVA is essentially a method of analysing the variance to which a response is subject into its various components, corresponding to the sources of variation which can be identified.

ANOVA, in this case, examines the effects of three quantitative variables (termed factors) on one quantitative response. Considered factors are the link function, which

TABLE III
ADIENCE RESULTS. BS STANDS FOR BATCH SIZE, LR FOR LEARNING RATE AND LF FOR LINK FUNCTION.

BS	LR	LF	κ_{val}	κ_{test}
5	10^{-2}	Logit	0.72072	0.68408
		Probit	0.77608	0.72574
		c log-log	0.75430	0.70742
	10^{-3}	Logit	0.71460	0.67868
		Probit	0.81906	0.77154
		c log-log	0.57196	0.55608
	10^{-4}	Logit	0.54432	0.51934
		Probit	0.54360	0.52422
		c log-log	0.54802	0.52484
10	10^{-2}	Logit	0.78872	0.72994
		Probit	0.79332	0.74750
		c log-log	0.78004	0.73260
	10^{-3}	Logit	0.49686	0.47112
		Probit	0.59778	0.58476
		c log-log	0.52016	0.49968
	10^{-4}	Logit	0.60972	0.56814
		Probit	0.54638	0.52632
		c log-log	0.75428	0.52094
15	10^{-2}	Logit	0.79104	0.74388
		Probit	**0.84254**	**0.78666**
		c log-log	0.78802	0.74624
	10^{-3}	Logit	0.56946	0.52978
		Probit	0.70262	0.65416
		c log-log	0.49242	0.47320
	10^{-4}	Logit	0.55202	0.52484
		Probit	0.54288	0.51740
		c log-log	0.53826	0.52542

can be logit, probit and complementary log-log, the learning rate for the Adam optimization algorithm (10^{-2}, 10^{-3} and 10^{-4}) and the batch size (5, 10 and 15).

Following the setup of the previous study, we performed an ANOVA III analysis and multiple comparison tests. We assume that five executions are enough to do the statistical tests because of the computational time limitations.

The test shows that there is no batch size whose results are significatively better than the results of all other batch sizes. This does not mean that these differences could not exist for specific numbers of samples per batch. So, in order to determine for each type of function whether a batch size is better than the others, we have performed an ANOVA I analysis - where the only factor is the batch size - and multiple comparison tests.

We denote by $QWK_{i,j,k,l}(i = 1, ..., 3; j = 1, ..., 3; k = 1, ..., 3)$ the value observed when the first factor is at the i-th level, the second at the j-th level and the third at k-th level. We assume that the three factors do not act independently and therefore there exists an interaction between each pair of them and between the three factors. In this case, the observations fit 4.

$$\begin{aligned} QWK_{i,j,k,l} =& \mu + L_i + P_j + B_k + LP_{i,j} + LB_{i,k} \\ &+ PB_{j,k} + LPB_{i,j,k} + \epsilon_{i,j,k,l} \end{aligned} \quad (4)$$

where μ is the fixed effect that is common to all the populations; L_i is the effect associated with the *i-th* level of the link factor (logit, probit, complementary log-log); P_j is the effect associated with the *j-th* level of the parameter factor (10^{-4}, 10^{-3}, 10^{-2}) and B_k is the effect associated with the *k-th* level of the batch size factor (5, 10, 15). The term $LP_{i,j}$ denotes the joint effect of the presence of level i of the first factor and level j of the second one; this, therefore, is denominated the interaction term between L and P factors. The same interaction effect is appreciated on $LB_{i,k}$, $PB_{j,k}$ and $LPB_{i,j,k}$. The term $\epsilon_{i,j,k,l}$ is the influence on the result of everything that could not be assigned or of random factors. $W_{i,j,k,l}$ is the quadratic weighted kappa measure, the response variable used to perform the statistical analysis.

TABLE IV
ANOVA III FOR THE ANALYSIS OF THE MAIN FACTORS IN THE DESIGN OF A CONVOLUTIONAL ORDINAL NEURAL NETWORK FOR THE RETINOPATHY DATASET.

Response variable QWK

Source	S.S.	D.F.	M.S.	F-ratio	Sig.
Model	37.860	9	4.207	1562.840	0.000
L factor	0.057	2	0.029	10.646	0.000
P factor	0.121	2	0.060	22.468	0.000
LP factors	0.057	4	0.014	5.261	0.001
Error	0.339	126	0.003		
Total	38.199	135			

We consider some hypotheses testing where the null hypothesis is proposed that each term of the above equation is independent of the levels involved. The hypotheses for the levels of the L factor are

$$H_0 \equiv L_1 = L_2 = L_3$$

$$H_1 \equiv \text{ some } L_i \text{ is different}$$

The same hypotheses are made for the other factors. In this way, we test in the null hypothesis that all of the population means are equal against an alternative hypothesis that there is at least one mean that is not equal to the others.

The hypothesis associated with the interaction between L and P is

$$H_0 \equiv LP_{i,j} = 0 \quad \forall i, j$$

$$H_1 \equiv \exists LP_{i,j} \neq 0$$

Similar hypotheses can be assumed for the interaction between the other factors.

The analysis of variance table containing the sum of squares, degrees of freedom, mean square, test statistics and significance level represent the initial study in a compact form, where non-significative factors and interactions have been removed (p-value > 0.05). These factors and interactions take part of the error component now.

The results of the ANOVA III test for the Diabetic Retinopathy dataset are resumed in Table IV. There are significative differences in the means of QWK depending on the link function and also depending on the learning rate for $\alpha = 0.05$. Moreover, an interaction between the link function and the learning rate can be recognised (p-value $= 0.001$).

As Table IV showed that there exist significative differences between the means, we are now analysing those differences. A post-hoc multiple comparison test has been performed on the mean QWK obtained. An HSD Tukey's

TABLE V
ANOVA III FOR THE ANALYSIS OF THE MAIN FACTORS IN THE DESIGN OF A CONVOLUTIONAL ORDINAL NEURAL NETWORK FOR THE ADIENCE DATASET.

Response variable QWK					
Source	S.S.	D.F.	M.S.	F-ratio	Sig.
Model	52.328	15	3.489	1426.170	0.000
L factor	0.027	2	0.014	5.582	0.005
P factor	1.031	2	0.516	210.809	0.000
B factor	0.089	2	0.045	18.253	0.000
LP factors	0.183	4	0.046	18.695	0.000
PB factors	0.124	4	0.031	12.695	0.000
Error	0.294	120	0.002		
Total	52.622	135	3.489		

test was made under the null hypothesis that the variance of the error of the dependent variable is the same between groups. The best link function is the complementary log-log, though it doesn't have significative differences with the probit function ($\alpha = 0.05$). There are differences between the complementary log-log and the probit function concerning the logit function.

The results have been obtained following a similar methodology with the different values of the learning parameter including the HSD Tukey's test and the learning parameter ranking. Significative differences can be observed in the mean values for $\eta = 10^{-3}$ with respect to the other values of the parameter ($\alpha = 0.05$). Also, there are differences for $\eta = 10^{-4}$ concerning $\eta = 10^{-2}$.

The results of the ANOVA III test for the Adience dataset are shown in Table V. The interactions where there are not significative differences in QWK means have been omitted. First, the factor associated with the link function is analysed. The differences between the means are significative (p-value = 0.005). Secondly, there are significative differences concerning the learning rate factor too (p-value = 0.000). Also, there are differences for the means of QWK for the batch size factor. Significative interactions exist between two pairs of factors: link function and learning rate, and learning rate and batch size (p-value = 0.000 for both cases).

A post-hoc multiple comparison test has been performed on the average QWK obtained for this dataset too. Under the null hypothesis that the variance of the error of the dependent variable is the same between groups, an HSD Tukey's test has been done. The results showed that the best link function is the logit one, in this case, with 0.60553 as mean QWK value. However, there are not significative differences between this function and the complementary log-log (p-value = 0.110). It has differences with the probit function though (p-value = 0.003). The complementary log-log link reported the best results after the logit function, having a mean QWK of 0.58738. However, it doesn't show significative differences with the probit function. The results have been obtained following a similar methodology with the different values of the learning parameter including the HSD Tukey's test and the learning parameter ranking. The test reported significative differences between $\eta = 10^{-2}$ (mean value 0.73378) and the rest of values of this parameter ($\alpha = 0.005$), being this one the best value for this factor. The best value after 10^{-2} is 10^{-3}, which have a mean QWK value of 0.57989. Lastly, the best value for the batch size parameter is 5 (mean value 0.63243) as it has significative differences with the rest of values. The second best value is 15 (mean 0.61129), but it has no significative differences with 10 (p-value = 0.194).

To sum up, the results showed that the best parameter configuration depends on the problem that is being solved. None of the optimal parameters is the same for Retinopathy and Adience datasets. These results highlight the importance of adjusting the hyper-parameters for each problem that is treated instead of trying to find an optimal configuration for all the datasets.

TABLE VI
NOMINAL METHOD RESULTS. BS STANDS FOR BATCH SIZE AND LR FOR LEARNING RATE.

Dataset	BS	LR	κ_{val}	κ_{test}
Retinopathy	5	10^{-2}	0.2210	0.2050
		10^{-3}	0.2807	0.2654
		10^{-4}	0.4580	0.4449
	10	10^{-2}	0.2972	0.2997
		10^{-3}	0.3972	0.3989
		10^{-4}	**0.4972**	**0.4967**
	15	10^{-2}	0.3692	0.3676
		10^{-3}	0.4107	0.4162
		10^{-4}	0.4929	0.4859
Adience	5	10^{-2}	0.0605	0.0769
		10^{-3}	0.7994	0.7313
		10^{-4}	0.8009	0.7383
	10	10^{-2}	0.0559	0.0576
		10^{-3}	0.8239	0.7436
		10^{-4}	0.7938	0.7218
	15	10^{-2}	0.0593	0.0713
		10^{-3}	**0.8309**	**0.7575**
		10^{-4}	0.7867	0.7129

D. Statistical comparison between nominal and ordinal methods

The same experiments described in Section V were repeated using the cross-entropy instead of the QWK as loss function for the optimizer and the softmax function instead of the Proportional Odds Model for the output of the network. The evaluation metric remains the same in order to be able to compare. The results for both datasets are shown in Table VI. There are some parameter configurations where the training process gets stagnated and a very low QWK is obtained. As we saw in Sections VI-A and VI-B, this problem is not present when using the ordinal method.

As we did in Section VI-C, we must analyse the effect of the batch size (with values 5, 10 and 15) and the learning rate (10^{-2}, 10^{-3} and 10^{-4}) factors have over the QWK metric. So, we make an ANOVA II analysis for each dataset.

In the Diabetic Retinopathy dataset, there are significative differences between the mean values of QWK depending on the batch size and the initial value of the learning rate. Also, no significative interaction was detected between these factors.

TABLE VII
COMPARISON BETWEEN THE BEST QWK RESULTS OF NOMINAL, ORDINAL AND PREVIOUS WORKS.

Method	BS	LR	LF	κ_{val}	κ_{test}
Diabetic Retinopathy					
Ord. proposed	10	10^{-3}	c log-log	0.58830	0.58150
Nom. proposed	10	10^{-4}	softmax	0.49720	0.49670
Best from [18]	20	10^{-4}	softmax	0.53700	-
Best from [36]	-	-	softmax	-	0.55500
Adience					
Ord. proposed	15	10^{-2}	probit	0.84254	0.78666
Nom. proposed	15	10^{-3}	softmax	0.83090	0.75750

After that, a post-hoc Tukey's test shows that the best value for the batch size factor is 15 followed by 10. However, there are no significative differences between them, but there are differences with size 5. Also, the best value for the learning rate is 10^{-4} and there are significative differences with the rest of values.

A similar study was made for the Adience dataset. There are not significative differences for the mean value of QWK w.r.t. the batch size factor. Also, there are no interactions between the batch size and the learning rate. The best results were obtained with 10^{-3} for the learning rate followed by 10^{-4}, without significative differences between them.

The conclusions obtained from the results are exposed below:

- There are no interactions between both factors.
- Batch sizes 10 and 15 obtained the best results.
- Learning rate 10^{-4} is the best choice for both datasets as it the best for Diabetic Retinopathy, with significative differences, and it is the second for Adience, without significative differences with the best.

Finally, a comparison of the best QWK for each dataset for ordinal and nominal cases is shown in Table VII. Also, it is compared with previous works. The proposed ordinal model outperforms all the other alternatives. The performance increase of POM over softmax reaches 17.07% for Diabetic Retinopathy and 3.85% for Adience dataset. The enhancement of the ordinal method for Retinopathy dataset is more notable than the rise for Adience dataset. It seems that the method proposed in this work offers a more significant improvement as the given problem complexity increases.

VII. Conclusions

In this section, the conclusions obtained from this work are exposed. The first thing that we have noticed is that the optimal values for the different parameters considered are problem-dependant.

The complementary log-log function has the best average performance across both datasets. It offers the best results in Diabetic Retinopathy dataset and the second best result in Adience dataset. These results provide an opportunity for exploring new generalised link functions.

The best value for the learning rate parameter for Diabetic Retinopathy dataset is $\eta = 10^{-3}$ while this value is the second best option for Adience dataset. It can be considered a good value for this parameter when training the model with new datasets.

Both datasets have obtained the best performance with batch size 10.

Also, the statistical tests reported that there are interactions between some parameters like the link function and the learning rate, for Diabetic Retinopathy, or those factors in addition to learning rate and batch size, for Adience dataset.

The proposed POM has improved the performance of the deep network regarding quadratic weighted kappa compared to the model that uses the softmax function. This enhancement is more notable for the Diabetic Retinopathy dataset. Also, it reduces the chance that the model gets stuck when training with some parameter configurations. So, the most significant improvements of these link functions are the QWK performance increase, the reduction of the number of parameters configurations that should be tried to find the best one and the prevention of the over-fitting and the stagnation.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[2] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," *ArXiv preprint arXiv:1202.2745*, 2012.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[7] X. Jiang, Y. Pang, X. Li, and J. Pan, "Speed up deep neural network based pedestrian detection by sharing features across multi-scale models," *Neurocomputing*, vol. 185, pp. 163–170, 2016.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[9] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "Iprivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1005–1016, 2017.

[10] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang, "Multimodal deep autoencoder for human pose recovery," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5659–5670, 2015.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[12] Z. Li and J. Tang, "Weakly supervised deep metric learning for community-contributed image retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1989–1999, 2015.

[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*, Springer, 2014, pp. 184–199.

[14] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476–3483.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[16] J. Leonard and M. Kramer, "Improvement of the backpropagation algorithm for training neural networks," *Computers & Chemical Engineering*, vol. 14, no. 3, pp. 337–341, 1990.

[17] X.-H. Yu, G.-A. Chen, and S.-X. Cheng, "Dynamic learning rate optimization of the backpropagation algorithm," *IEEE Transactions on Neural Networks*, vol. 6, no. 3, pp. 669–677, 1995.

[18] J. de la Torre, D. Puig, and A. Valls, "Weighted kappa loss function for multi-class classification of ordinal data in deep learning," *Pattern Recognition Letters*, vol. 105, pp. 144–154, 2018.

[19] A. Senior, G. Heigold, K. Yang, *et al.*, "An empirical study of learning rates in deep neural networks for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, IEEE, 2013, pp. 6724–6728.

[20] L. N. Smith, "Cyclical learning rates for training neural networks," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, IEEE, 2017, pp. 464–472.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *ArXiv preprint arXiv:1502.03167*, 2015.

[22] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *ArXiv preprint arXiv:1609.04836*, 2016.

[23] P. M. Radiuk, "Impact of training set batch size on the performance of convolutional neural networks for diverse datasets," *Information Technology and Management Science*, vol. 20, no. 1, pp. 20–24, 2017.

[24] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, and P. A. Gutiérrez, "Metrics to guide a multi-objective evolutionary algorithm for ordinal classification," *Neurocomputing*, vol. 135, pp. 21–31, 2014.

[25] N. Mehdiyev, D. Enke, P. Fettke, and P. Loos, "Evaluating forecasting methods by considering different accuracy measures," *Procedia Computer Science*, vol. 95, pp. 264–271, 2016.

[26] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[27] P. A. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez, "Ordinal regression methods: Survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2016.

[28] A. Agresti, *Analysis of ordinal categorical data*. John Wiley & Sons, 2010, vol. 656.

[29] C. Beckham and C. Pal, "Unimodal probability distributions for deep ordinal classification," *ArXiv preprint arXiv:1705.05278*, 2017.

[30] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[31] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013, pp. 1310–1318.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[33] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *ArXiv preprint arXiv:1511.07289*, 2015.

[34] P. McCullagh, "Regression models for ordinal data," *Journal of the royal statistical society. Series B (Methodological)*, pp. 109–142, 1980.

[35] R. Herbrich, "Large margin rank boundaries for ordinal regression," *Advances in large margin classifiers*, pp. 115–132, 2000.

[36] À Nebot *et al.*, "Diabetic retinopathy detection through image analysis using deep convolutional neural networks," in *Artificial Intelligence Research and Development: Proceedings of the 19th International Conference of the Catalan Association for Artificial Intelligence, Barcelona, Catalonia, Spain, October 19-21, 2016*, IOS press, 2016, p. 58.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ArXiv preprint arXiv:1412.6980*, 2014.

[38] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.

[39] A. Ben-David, "Comparison of classification accuracy using cohen's weighted kappa," *Expert Systems with Applications*, vol. 34, no. 2, pp. 825–832, 2008.

[40] R. G. Miller Jr, *Beyond ANOVA: Basics of applied statistics*. Chapman and Hall/CRC, 1997.