1 Introduction

...

2 Related work

- Cyclical learning rates for training neural networks [1].
- An empirical study of learning rates in deep neural networks for speech recognition [2].

3 Experiments

3.1 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, illumination conditions and 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 [3] 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 256 by 256 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 [4], so that the images are 256px 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.

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

3.2 The model

A convolutional neural network (CNN) has been used for both datasets. The architecture of this CNN is presented in the Table 1.

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

Table 1: 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.

Every convolutional layer is followed by an ELU activation layer [5] and a batch normalization [6]. At the output, a Proportional Odds Model (POM) is used with different link functions [7]. 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.

- *Logit.* Logit link function is the most widely used function for Proportional Odds Models.

These kind of models are also called Cummulative Logit Models. The logit link is shown in Equation 1.

$$logit[P(Y_i \leq j)] = \alpha_j + \beta' x_i, \quad j = 1, ..., c-1 \quad (1)$$

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

$$\Phi^{-}1[P(Y \leq j)] = \alpha_j + \beta' x, \quad j = 1, ..., c-1 \quad (2)$$

- *Complementary log-log.* Unlike logit and probit, 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 Equation 3.

$$log[-log[1-P(Y \leq j)]] = \alpha_j + \beta' x, \quad j = 1, ..., c-1 \quad (3)$$

3.3 Procedure

The model is optimized using a batch based first-order optimization algorithm called Adam [8]. 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.

Quadratic Weighted Kappa Loss, that J. de la Torre described in previous work, is considered as loss function for this optimizer as it gives better performance for ordinal classification problems.

Both datasets have been artificially equalised using data augmentation techniques [9][10] 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 [11]. This evaluation measure gives a higher weight to the errors that are further from the correct class.

3.4 Parameters

Three different parameters 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. 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.

4 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.

4.1 Diabetic Retinopathy

... 2

BS	LR	LF	κ_{val}	κ_{test}
5	10^{-02}	poml	0.44888	0.5009
5	10^{-02}	pomp	0.46724	0.49614
5	10^{-02}	cloglog	0.42854	0.50006
5	10^{-03}	poml	0.56496	0.5198
5	10^{-03}	pomp	0.57084	0.51114
5	10^{-03}	cloglog	0.54796	0.50808
5	10^{-04}	poml	0.52884	0.57954
5	10^{-04}	pomp	0.53802	0.5772
5	10^{-04}	cloglog	0.53824	0.57636
10	10^{-02}	poml	0.5497	0.48474
10	10^{-02}	pomp	0.52124	0.5092
10	10^{-02}	cloglog	0.4276	0.52524
10	10^{-03}	poml	0.58036	0.54692
10	10^{-03}	pomp	0.56536	0.52716
10	10^{-03}	cloglog	0.5883	0.53076
10	10^{-04}	poml	0.53158	0.59238
10	10^{-04}	pomp	0.5456	0.59372
10	10^{-04}	cloglog	0.53928	0.59242
15	10^{-02}	poml	0.55828	0.50664
15	10^{-02}	pomp	0.54112	0.50426
15	10^{-02}	cloglog	0.56676	0.4991
15	10^{-03}	poml	0.55748	0.57388
15	10^{-03}	pomp	0.58182	0.5573
15	10^{-03}	cloglog	0.5634	0.55906
15	10^{-04}	poml	0.54686	0.5994
15	10^{-04}	pomp	0.53258	0.60264
15	10^{-04}	cloglog	0.539	0.6005

Table 2: Diabetic Retinopathy results.

4.2 Adience

...

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

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

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

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

[9] 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.

[10] 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.

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