Multiscale Softmax Cross Entropy for Fovea Localization on Color Fundus Photography

Yuli Wu[1], Peter Walter[2], Dorit Merhof[1]

[1]Institute of Imaging and Computer Vision, RWTH Aachen, Germany
[2]Department of Ophthalmology, RWTH Aachen, Germany
yuli.wu@lfb.rwth-aachen.de

Abstract. Fovea localization is one of the most popular tasks in ophthalmic medical image analysis, where the coordinates of the center point of the *macula lutea*, i.e. *fovea centralis*, should be calculated based on color fundus images. In this work, we treat the localization problem as a classification task, where the coordinates of the x- and y-axis are considered as the target classes. Moreover, the combination of the softmax activation function and the cross entropy loss function is modified to its multiscale variation to encourage the predicted coordinates to be located closely to the ground-truths. Based on color fundus photography images, we empirically show that the proposed multiscale softmax cross entropy yields better performance than the vanilla version and than the mean squared error loss with sigmoid activation, which provides a novel approach for coordinate regression.

1 Introduction

Outputting coordinates is common in computer vision tasks, such as in bounding-box regression for object detection and in keypoint localization for facial recognition. Regression losses are typically selected to calculate the error between the ground-truth and the prediction, *e.g.* Mean Squared Error (MSE) loss and Mean Absolute Error (MAE) loss, which measure L2 and L1 distances between the ground-truth and the prediction, respectively. In contrast, probabilistic losses are usually used in classification tasks, which includes Cross Entropy (CE) loss as one of the most popular choices (categorical cross entropy in the case of multi-class). One significant difference between these two categories is that MSE or MAE punishes incorrect predictions less, which are however close to the ground-truth, while categorical CE combined with softmax activation function treats all incorrect predictions equally to the maximum.

An accurate localization of the fovea, an important anatomical landmark in the retina, can be beneficial to the computer aided diagnosis of retinal diseases. Huang et al. [1] take advantage of the geometrical relationship between optic disc and fovea to achieve a more accurate localization and Xie et al. [2] utilize a three-stage network with coarse-fine fusion. MSE loss is used in both approaches.

Kopaczka et al. [3] combine soft-argmax loss [4] and L1 distance loss to localize and track rodents, which shows the feasibility of probabilistic loss functions in classification tasks.

In this work, we consider the localization task as two classification tasks, referring to the x- and y-axis, by using a combination of the softmax activation function and the cross entropy loss function. Trying to bridge the functional gap between the regression and probabilistic losses, we propose Multiscale Softmax Cross Entropy (MSCE), which takes the last feature map learned from the backbone convolutional neural network and combines multiple downsampled feature maps with independent softmax cross entropy.

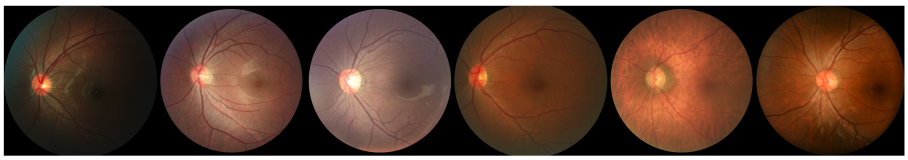


Fig. 1. Examples of color fundus images from REFUGE2 [5].

2 Materials and Methods

2.1 Dataset

The dataset of color fundus images, REFUGE2 [5], contains 1200 images with and 400 images without ground-truth annotations for training and testing, respectively. The metrics used to evaluate the predicted localization coordinates is taken from the latest Gamma Challenge[1], namely the Reciprocal of the Average Euclidean Distance (R-AED) value, which is defined as R-AED $= \frac{1}{d(\mathbf{p},\mathbf{q})+0.1}$, where the Euclidean distance is used between the normalized coordinates of ground-truth $\mathbf{p}$ and prediction $\mathbf{q}$ as $d(\mathbf{p},\mathbf{q}) = ||\mathbf{p}-\mathbf{q}||_2$.

2.2 Network Architecture

We adopt the neural network architecture from cellpose [6], which is a modified U-Net [7] with residual connections inside each convolutional block and a style vector fused to the upsampling pathway. The original image is first resized and fed into the cellpose network, which outputs the feature map of identical size. The learned feature map is pooled multiple times to generate the multiscale branches, each of which is first reduced per axis (via *e.g.* `sum` or `mean`). The multiscale loss is then calculated with independent softmax cross entropy for each branch. Finally, the final loss is aggregated with weighted sum, which we denote as Multiscale Softmax Cross Entropy (MSCE). The detailed introduction of MSCE is presented in Section 2.3. The implementation of hyperparameters during the experiments can be found in Section 2.4.

[1] https://gamma.grand-challenge.org

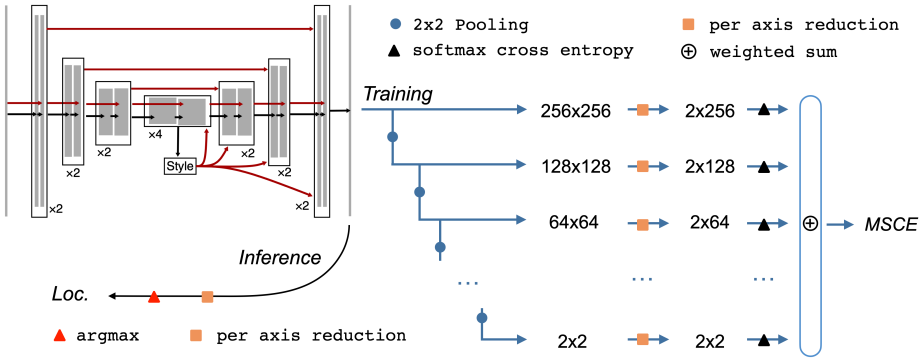


Fig. 2. Network architecture for training and inference. The network backbone is adopted from the cellpose network [6] and the corresponding figure is adapted from it.

2.3 Loss

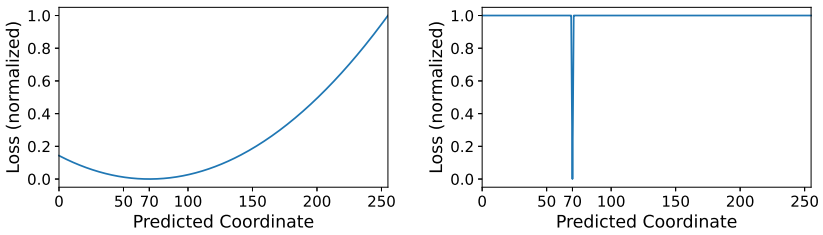
We present Multiscale Softmax Cross Entropy (MSCE), which takes two logit vectors of different sizes and calculates a weighted summation of softmax cross entropy from them.

$$SCE = -\sum_{i=1}^{C} t_i \log\left(\frac{e^{s_i}}{\sum_{j=1}^{C} e^{s_j}}\right) \tag{1}$$

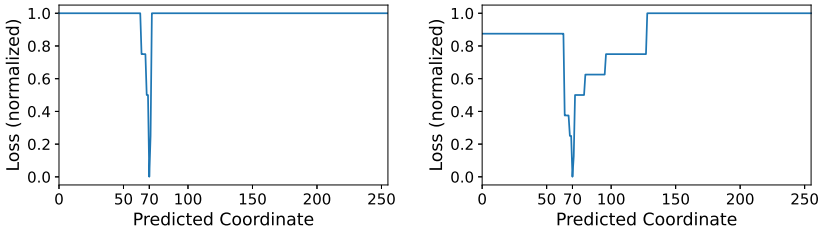
$$MSCE = \sum_{m=1}^{M} \lambda_m \cdot \Big(-\sum_{i=1}^{C_m} t_i \log\left(\frac{e^{s_i}}{\sum_{j=1}^{C_m} e^{s_j}}\right)\Big) \tag{2}$$

Based on the original Softmax Cross Entropy (SCE) in Equation 1, the multiscale version can be defined as shown in Equation 2. In both equations, s denotes the predicted logit and t_i indicates whether the i-th of total C class labels is the correct classification. In Equation 2, M denotes the number of multiscales (or the number of the branches in Fig. 2) and λ_m denotes the weights for the SCE term of each scale. In this work, we set all $\lambda_m = 1$.

In Fig. 3, different loss functions are compared with a toy example, where we assume the 70th class, *i.e.* coordinate, of a 256 dimensional vector is the ground-truth and the normalized loss values are calculated for each possible prediction. Fig. 3(a) illustrates MSE, an example from the category of regression loss, which progressively attracts the wrong predictions to the ground-truth. In the case of SCE (Fig. 3(b)), however, the incorrect coordinates have been opposed expressly and unanimously, no matter where they are located rather than the ground-truth. The proposed MSCE is expected to neutralize the characteristics of MSE and SCE, which not only distinguishes the predictions in a stepwise regressive manner but also drastically encourages the prediction to converge towards the single actual ground-truth without decreasing the reward ratio. The desired feature can be better approached, if the number of the multiscales M is set to the maximum ($M = 8$ in case of 256 classes) comparing Fig. 3(c) and Fig. 3(d).



(a) MSE (b) SCE



(c) MSCE ($M = 4$) (d) MSCE ($M = 8$)

Fig. 3. Toy experiments in 1D coordinate: loss values of different loss functions. Predicted coordinates are presented in x-axis and the normalized loss values in y-axis, with the assumption that the 70th coordinate of total 256 is the ground-truth. MSE: Mean Squared Error; SCE: Softmax Cross Entropy; MSCE: Multiscale Softmax Cross Entropy. M denotes the number of the multiscales (see Equation 2).

2.4 Hyperparameters

The *style* mechanism from the cellpose network [6] has been preserved, as it is assumed to play a role when combining the disease grading task and the fovea localization task in future work. The images are first resized to 256 by 256 and no augmentation techniques have been then applied. We use `MaxPooling` when downsampling the feature map to generate multiscales ones (blue dots in Fig. 2) and `sum` as the reduction operator to obtain the per axis logit vectors (orange squares in Fig. 2). It has been empirically shown that `MaxPooling` and `sum` yield better results than `AveragePooling` and `mean` reduction.

The training process is optimized by stochastic gradient descent that uses an exponential decay schedule with an initial learning rate of 0.01, decay steps of 400 and a decay rate of 0.9. The maximal number of epochs is set to 1000 and the `EarlyStopping` mechanism has been applied with a patience of 100 epochs in terms of overall loss values.

3 Results

Experimental results are shown in Table 1 *w.r.t.* the reciprocal of the average Euclidean distance (R-AED). It is found that `MaxPooling` with `sum` reduction

Table 1. Results from ablation experiments with different loss functions and network settings *w.r.t.* the reciprocal of the average Euclidean distance (R-AED). `Ave/mean` denotes `AveragePooling` with `mean` reduction and `Max/sum` denotes `MaxPooling` with `sum` reduction. Best results from each experimental group are marked in **bold** face.

Loss	Network	Batch Size	R-AED (↑)
Mean squared error (baseline)	`Ave/mean`	8	**5.69**
Softmax cross entropy	`Ave/mean`	8	3.45
Multiscale softmax cross entropy	`Ave/mean`	8	4.36
Mean squared error (baseline)	`Max/sum`	16	5.18
Softmax cross entropy	`Max/sum`	16	4.16
Multiscale softmax cross entropy	`Max/sum`	16	**5.31**
Mean squared error (baseline)	`Max/sum`	8	5.53
Softmax cross entropy	`Max/sum`	8	4.99
Multiscale softmax cross entropy	`Max/sum`	8	**6.12**

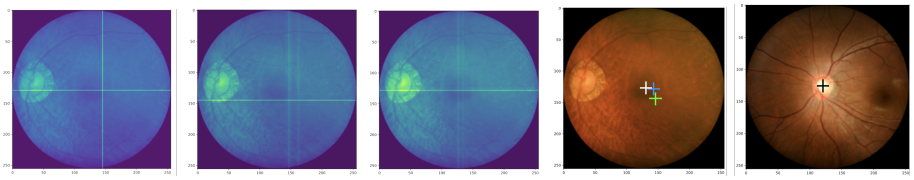
plays a significant role to boost the performance with SCE and MSCE, in which case MSCE outperforms MSE loss. Generally, the modified MSCE yields better results than the vanilla SCE, which empirically demonstrates the feasibility of probabilistic losses in the regression tasks.

Predicted fovea locations are illustrated in Fig. 4, where the final coordinate vectors are illustrated on the original fundus images with mean squared error (MSE), vanilla softmax cross entropy (SCE) and multiscale softmax cross entropy (MSCE) in (a-c), respectively. From Fig. 4(d), it can be noted that MSE (blue) and SCE (green) result in a larger offset than MSCE (white). A typical failed prediction happens if the fovea is located far away from the central region and blends into the dark marginal area, as shown in Fig. 4(e).

4 Discussion

Although the proposed MSCE loss has surpassed the commonly used MSE loss and the vanilla SCE loss based on the ablation experiments, some unstable predictions haven been noticed during the experiments. We assume that finetuned hyperparameters, including the weights λ_m in Equation 2, could mitigate this issue.

In practice, the fovea is usually localized with the help of the relative position of the optic disc by surgeons. Therefore, it is expected that fusing the relative spatial information via optic disc segmentation would achieve better results for fovea localization. Additionally, the segmentation-based feature map of our approach could further strengthen the advantages by combining different ophthalmic tasks with fovea localization based on fundus images, such as vessel segmentation, optic disc and optic cup segmentation, and disease (*e.g.* glaucoma) grading.



(a) MSE (b) SCE (c) MSCE (d) Results (e) MSCE failed

Fig. 4. Examples of predicted fovea locations with different losses. (a-c) illustrate the final coordinate vectors on the original images. (d) compares the predicted locations with crosses (blue, green, white denote MSE, SCE, MSCE, respectively). (e) shows a failed prediction, if the optic disc instead of the fovea is located in the center.

5 Conclusion

This work addresses the fovea localization task based on the feature map that is initially tailored for segmentation. Furthermore, the task of coordinate regression from logits is performed based on a probabilistic loss, which usually contributes to classification tasks. The modified multiscale version of softmax cross entropy (MSCE) has empirically shown the capability for localization tasks. The performance of MSCE surpasses both the vanilla SCE and the mean squared error loss with the identical network backbone and hyperparameter setups, which offers a novel loss alternative for fovea localization and is promising for other general coordinate regression tasks like bounding boxes in object detection.

Acknowledgements. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – grant 424556709/GRK2610.

References

1. Huang Y, Zhong Z, Yuan J, et al. Efficient and robust optic disc detection and fovea localization using region proposal network and cascaded network. Biomedical Signal Processing and Control. 2020;60:101939.
2. Xie R, Liu J, Cao R, et al. End-to-End Fovea Localisation in Colour Fundus Images With a Hierarchical Deep Regression Network. IEEE Transactions on Medical Imaging. 2021 Jan;40(1):116–128.
3. Kopaczka M, Jacob T, Ernst L, et al. Robust Open Field Rodent Tracking using a Fully Convolutional Network and a Softargmax Distance Loss. Proc BVM. 2020;.
4. Honari S, Molchanov P, Tyree S, et al. Improving landmark localization with semi-supervised learning. In: CVPR; 2018. p. 1546–1555.
5. Orlando JI, Fu H, Barbosa Breda J, et al. REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs. Medical Image Analysis. 2020 Jan;59:101570.
6. Stringer C, Wang T, Michaelos M, et al. Cellpose: A Generalist Algorithm for Cellular Segmentation. Nature Methods. 2021 Jan;18(1):100–106.
7. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–241.