Focal loss improves repeatability of deep learning models

Syed Rakin Ahmed[1,2,3,4] RAKIN@MIT.EDU
[1] *Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Boston, MA, USA*
[2] *Harvard Graduate Program in Biophysics, Harvard Medical School, Harvard University, Cambridge, MA, USA*
[3] *Massachusetts Institute of Technology, Cambridge, MA, USA*
[4] *Geisel School of Medicine at Dartmouth, Dartmouth College, Hanover, NH, USA*

Andreanne Lemay[1,5] ANDREANNE.LEMAY@POLYMTL.CA
[5] *NeuroPoly, Polytechnique Montreal, Montreal, QC, Canada*

Katharina Hoebel[1,3] KHOEBEL@MIT.EDU
Jayashree Kalpathy-Cramer[1] KALPATHY@NMR.MGH.HARVARD.EDU

Editors: Under Review for MIDL 2022

Abstract

Deep learning models for clinical diagnosis, prognosis and treatment need to be trustworthy and robust for clinical deployment, given that model predictions often directly inform a subsequent course of action, where individual patient lives are at stake. Central to model robustness is repeatability, or the ability of a model to generate near-identical predictions under identical conditions. In this work, we optimize focal loss as a cost function to improve repeatability of model predictions on two clinically significant classification tasks: knee osteoarthritis grading and breast density classification, with and without the presence of Monte Carlo (MC) Dropout. We discover that in all experimental instances, focal loss improves repeatability of the resulting models, an effect compounded in the presence of MC Dropout.

Keywords: repeatability, focal loss, breast density, knee osteoarthritis, Monte Carlo dropout, medical classification, computer vision.

1. Introduction

The majority of work incorporating deep learning (DL) for clinical classification tasks focus on model accuracy and classification performance. However, in order to implement reliable and robust models on patient populations, particularly for diagnostic, prognostic and treatment classification tasks, model repeatability is of paramount importance. Repeatability refers to the ability of a model to generate near-identical predictions for the same patient under identical conditions, ensuring that the model produces precise, reliable outputs in the clinical setting. Given that small changes in an image can produce significantly different DL model predictions, it is essential that models designed for clinical deployment be specifically optimized for improved repeatability. Similar to clinical-decision making flow-diagrams where a single differential response can lead to an entirely different pathway of interventions for a patient, an incorrect, unreliable or unrepeatable model prediction would lead to a particular cascade of undesirable downstream clinical actions, that might significantly jeopardize the health and safety of a patient, and put their lives at risk.

2. Methods and Results

In this work, we utilize focal loss, with and without the presence of Monte Carlo (MC) Dropout in order to improve model repeatability. We conduct our experiments on two datasets: the publicly available longitudinal Multicenter Osteoarthritis Study (MOST) dataset for knee osteoarthritis grading, and the Digital Mammographic Imaging Screening Trial (DMIST) dataset, a multi-institutional screening dataset for breast density classification; both represent clinically significant and impactful classification tasks.

In deep learning optimization, cross-entropy (CE) loss is frequently utilized as the primary cost function across various classification tasks. The α balanced version of CE loss for a binary classification problem can be written as (Zhang and Sabuncu, 2018),

$$CE(p_t) = -\alpha_t \log(p_t) \tag{1}$$

$$p_t = \begin{cases} p & : class = 1 \\ 1-p & : otherwise \end{cases}$$

One notable property of CE loss is that even examples that are easily classified incur a loss with non-trivial magnitude, which, when summed over large numbers of easy examples, can overwhelm the rare class. For a binary problem, α_t is defined analogously as p_t with $\alpha \in [0, 1]$ for class 1 and $1 - \alpha$ otherwise, where α can be set as the inverse class frequency or as a tunable hyperparameter. While α balances the importance of positive/negative examples, it does not differentiate between easy/hard examples.

Focal Loss adds a modulating factor $(1-p_t)^\gamma$ to standard CE loss, with tunable focusing parameter $\gamma \geq 0$, thereby focusing training on hard, misclassified examples. The α balanced version of focal loss for a binary classification problem can be written as, with p_t defined as in Equation 1 above (Lin et al., 2017),

$$FOC(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{2}$$

In this work, we begin with two baseline models optimized for classification performance in each dataset, utilizing CE loss for (A) (i) the MOST dataset, and (B) (i) the DMIST dataset, with or without MC dropout and all other parameters identical for each corresponding dataset. We subsequently train models with focal loss, using parameters $\gamma = 2$ and $\alpha = 0.25$ following hyperparameter tuning, with and without MC dropout. We compare each focal loss experimental result with the corresponding MC or non-MC versions of the baseline, as highlighted in Table 1 and Figure 1. The repeatability metrics are calculated utilizing multiple images from the same patient and visit: each point on each Bland-Altman plot (Figure 1) refers to a single patient, with the y-axis representing the maximum difference in the continuous classification score across repeat images for each patient, and the x-axis plotting the mean of the corresponding scores across all repeat images per patient. We find that, in all instances, focal loss improves repeatability from the respective baseline, reflected in statistically significant decreases in the 95% limits of agreement (LoA) on the Bland-Altman plots, while not affecting classification performance, reflected in no statistically significant difference in the corresponding accuracy. This is a critical and impactful result, given the ultimate hope of obtaining identity predictions under identical conditions

for the same patient, thereby facilitating reliability and trust on clinical DL models. This improves model deployability in the clinic, particularly in critical situations where model predictions directly inform the course of treatment, and where individual patient lives are at stake.

Table 1: **Model performance overview (mean ± 95% CI)**. Values in bold indicate the better model between baseline and focal loss where a statistically significant difference ($p-value > 0.05$) was observed, for each of the MC and non-MC cases. The two first columns for each dataset measure the model repeatability where smaller values indicate better repeatability, while the two second columns represent model classification performance. LoA: Limits of agreement; Acc.: Accuracy.

	(A) MOST - Knee Osteoarthritis		(B) DMIST - Breast Density	
Model	95% LoA ↓	Acc.	95% LoA ↓	Acc.
(i) Baseline	0.170 ± 0.006	0.684 ± 0.010	0.335 ± 0.002	0.688 ± 0.005
(ii) Focal Loss	$\mathbf{0.147 \pm 0.006}$	0.668 ± 0.010	$\mathbf{0.320 \pm 0.003}$	0.690 ± 0.005
(iii) Baseline, with MC dropout	0.073 ± 0.003	0.716 ± 0.009	0.298 ± 0.005	0.708 ± 0.005
(iv) Focal Loss, with MC dropout	$\mathbf{0.062 \pm 0.003}$	0.725 ± 0.008	$\mathbf{0.258 \pm 0.004}$	0.717 ± 0.006

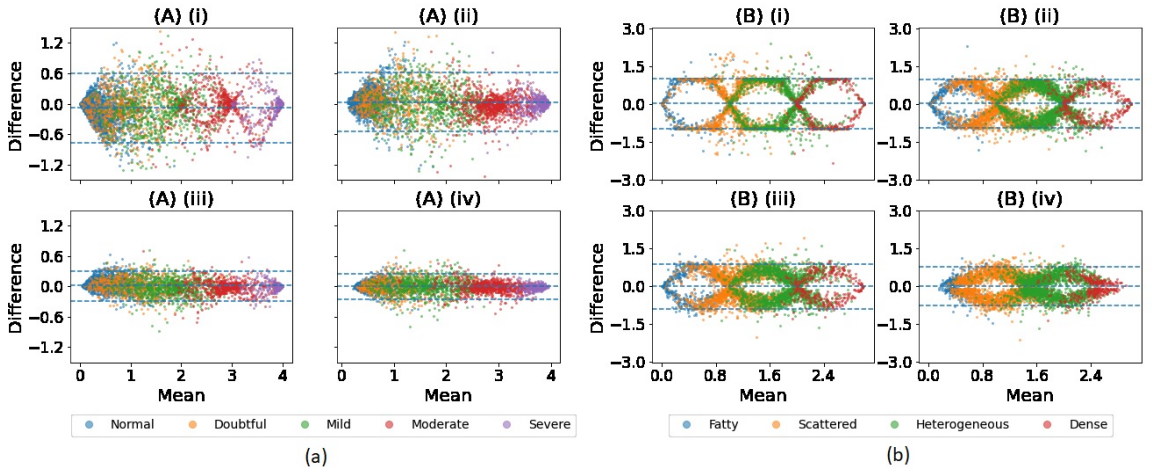

Figure 1: Bland-Altman plots on images from the same patient and visit for two clinically significant classification tasks: (a) knee osteoarthritis grading and (b) breast density classification. Each plot is labelled with the corresponding row/column label from Table 1. The legends indicate the different classes for each task.

References

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.