Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Performance Evaluation of Deep Learning Models for Image Classification Over Small Datasets: Diabetic Foot Case Study

ABIAN HERNANDEZ-GUEDES[1,2], IDAFEN SANTANA-PEREZ[1], NATALIA ARTEAGA-MARRERO[3], HIMAR FABELO[2,4], GUSTAVO M. CALLICO[2], (Senior Member, IEEE), and JUAN RUIZ-ALZOLA[1,3], (Senior Member, IEEE)

[1]Research Institute in Biomedical and Health Sciences (IUIBS), University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain.
[2]Research Institute for Applied Microelectronics (IUMA), University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain.
[3]IACTEC Medical Technology Group, Instituto de Astrofísica de Canarias (IAC), San Cristóbal de La Laguna, Spain.
[4]Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC), Las Palmas de Gran Canaria, Spain.

Corresponding author: Abian Hernandez-Guedes (e-mail: abian.hernandez@ulpgc.es).

This work was supported in part by the Spanish Government and European Union (FEDER funds) as part of support program in the context of TALENT-HExPERIA (HypErsPEctRal Imaging for Artificial intelligence applications) project, under contract PID2020-116417RB-C42. Moreover, this work was completed while Abian Hernandez was a beneficiary of a pre-doctoral grant given by the "Agencia Canaria de Investigacion, Innovacion y Sociedad de la Información (ACIISI)" of the "Consejería de Economía, Conocimiento y Empleo", which is part-financed by the European Social Fund (FSE) (POC 2014-2020, Eje 3 Tema Prioritario 74 (85%)) and, Himar Fabelo was beneficiary of the FJC2020-043474-I funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR". The codes developed in the current study are available from the corresponding author on reasonable request.

ABSTRACT Data scarcity is a common and challenging issue when working with Artificial Intelligence solutions, especially those including Deep Learning (DL) models for tasks such as image classification. This is particularly relevant in healthcare scenarios, in which data collection requires a long-lasting process, involving specific control protocols. The performance of DL models is usually quantified by different classification metrics, which may provide biased results, due to the lack of sufficient data. In this paper, an innovative approach is proposed to evaluate the performance of DL models when labeled data is scarce. This approach, which aims to detect the poor performance provided by DL models, in spite of traditional assessing metrics indicating otherwise, is based on information theoretic concepts and motivated by the Information Bottleneck framework. This methodology has been evaluated by implementing several experimental configurations to classify samples from a plantar thermogram dataset, focused on early stage detection of diabetic foot ulcers, as a case study. The proposed network architectures exhibited high results in terms of classification metrics. However, as our approach shows, only two of those models are indeed consistent to generalize the data properly. In conclusion, a new methodology was introduced and tested to identify promising DL models for image classification over small datasets without relying exclusively on the widely employed classification metrics.

INDEX TERMS Deep Learning, Information Theory, Information Bottleneck, Diabetes, Thermal Imaging.

I. SUPPEMENTARY MATERIAL

In this work, add-hoc models were used to evaluate the IB theoretical framework and its application using a scarce dataset. In this section, a state-of-the-art architecture will be used for comparison purposes. A baseline EfficientNet, EfficientNetB0, which consists of 10 layers [1], was used. This CNN has been previously trained using ImageNet [2] and a non-thermal DFU[1] public dataset [3], i.e., RGB images of ulcers in diabetic patients which share the same target that the thermal DFU dataset used in this work, but not the same image modality.

Table 1 shows high metrics as observed in the experiments

[1]https://www.kaggle.com/datasets/laithjj/diabetic-foot-ulcer-dfu

TABLE 1. Classification metrics results using EfficientNet.

Sensitivity	Specificity	Precision	Accuracy
0.937	0.937	0.937	**0.937**

with the add-hoc models proposed. As aforementioned, these metrics are not reliable due to the reduced amount of samples from the test set. In fact, Fig. 1 demonstrates that the model is not reliable. For a better legibility of the figures, the IP estimation have been divided into two independent images, Fig. 1(b,c).

Firstly, Fig. 1(a) shows that the model is not converging to an optimal solution in the test set, exhibiting a clear overfitting to the training set depicted by the significant difference between training and testing losses that is gradually increasing. Secondly, the IP estimation depicted in Fig. 1(b,c), shows different DPI violations between layers. In the first three layers, according to Fig. 1(b) there is a clear DPI violation since the first two layers have a worse representation of input X. But this may be because the first three layers are not a compressed version of X, as they are dimensionality-expanding layers. However, Fig. 1(c) shows a clear violation that cannot be justified and goes against the IB theoretical framework.

The IB is an information theoretic concept where the compromise between compression and prediction is presented. Fig. 1(c) shows that, discarding the last layer, the $I(T;Y)$ is reduced on each subsequent layer. In this way, it can be concluded that the information relevant to the prediction is not being stored in the compressed representation of X. Furthermore, a compression phase is observed in the last layer, being associated to an overfitting.

EfficientNet is a complex model which has been developed for image classification in highly-demanding datasets. The number of parameters of this model is really high, and overfitting is more likely when using a small dataset, as demonstrated in this work. These results were expected, even though transfer learning was applied with a powerful dataset, such as the DFU dataset in the visible spectrum. Thus, it is concluded that the use of an extremely complex pretrained model with a higher quality database could not bring any improvement, and it is not possible to justify this experiment as an optimal case for DFU classification based on thermal images.

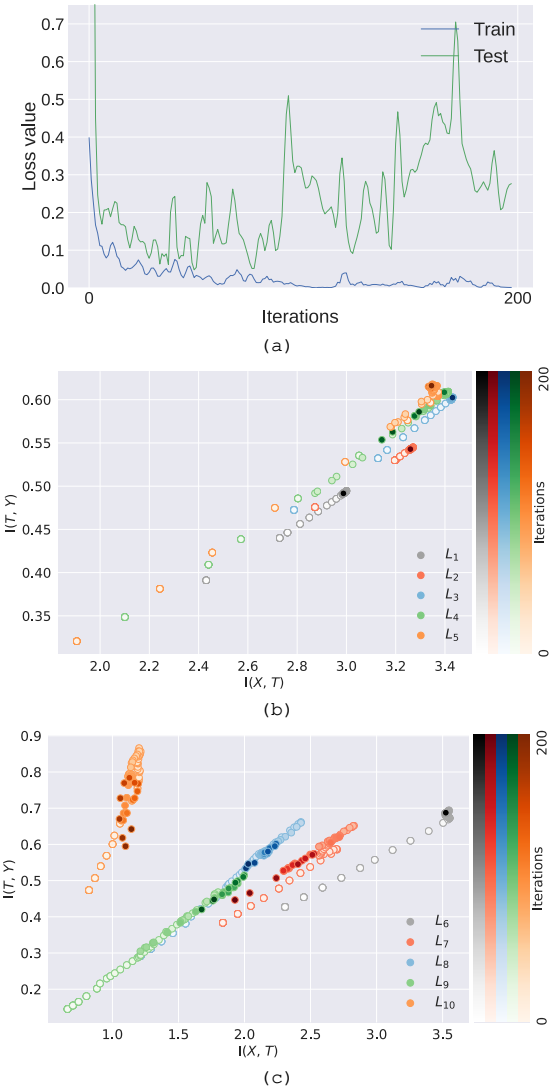

FIGURE 1. EfficientNetB0 experiment results where (a) corresponds to the CE loss values in the different iterations, and (b, c) depict the IP estimation with DFU. To improve the legibility of IP estimations, the different layers have been subdivided into two figures.

REFERENCES

[1] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International conference on machine learning, pp. 6105–6114, PMLR, 2019.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, Ieee, 2009.

[3] L. Alzubaidi, M. A. Fadhel, S. R. Oleiwi, O. Al-Shamma, and J. Zhang, "Dfu_qutnet: diabetic foot ulcer classification using novel deep convolutional neural network," Multimedia Tools and Applications, vol. 79, no. 21, pp. 15655–15677, 2020.