Low-Impact Feature Reduction regularization term: How to improve Artificial Intelligence with Explainability

Iván Sevillano-García[1,*], Julian Luengo[1] and Francisco Herrera[1]

[1]Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, 18071, Spain

Abstract

The proliferation of Artificial Intelligence in various domains has led to a growing demand for explainability. However, most modern Artificial Intelligence models are complex and opaque, making it challenging to interpret their decision-making process. The lack of transparency and interpretability of AI systems can pose significant risks, such as erroneous decisions, biased outcomes, and ethical concerns. To address these issues, various proposals have been put forth to generate explanations for black box models, both agnostic and model-dependent, and to evaluate these explanations using qualitative or quantitative measures. While these proposals have been useful in generating and evaluating explanations for AI models, none has focused on improving model quality using these evaluations. Regularization techniques, employed to introduce a quality bias into a solution, involve incorporating an additional term into the loss function to ensure that the model satisfies certain quality criteria, such as augmenting explainability. In this paper, the LIFR regularization term to improve the quality of AI models with respect to explainability is presented. We demonstrate the effectiveness of our approach on a benchmark dataset and discuss its potential applications.

Keywords

Explainable Artificial Intelligence, Interpretability, Regularization

1. Introduction

The proliferation of Artificial Intelligence(AI) in various domains has led to a growing demand for explainability, where users need to understand the reasoning behind AI systems decisions[1]. However, most modern AI models are complex and opaque, making it challenging to interpret their decision-making process. The lack of interpretability of AI systems can pose significant risks, such as erroneous decisions, biased outcomes and ethical concerns. To address these issues, various proposals have been put forth to generate explanations for black box models, both agnostic and model-dependent[2, 3], and to evaluate these explanations using qualitative[4] or quantitative measures[5]. While these proposals have been useful in generating and evaluating explanations for AI models, none has focused on improving model quality using these

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author

isevillano@ugr.es (I. Sevillano-García); julianlm@decsai.ugr.es (J. Luengo); herrera@decsai.ugr.es (F. Herrera)

0000-0002-5029-9106 (I. Sevillano-García); 0000-0003-3952-3629 (J. Luengo); 0000-0002-7283-312X (F. Herrera)

CEUR Workshop Proceedings (CEUR-WS.org)

evaluations. Regularization techniques are commonly used to bias a solution by involving an additional term to the objective function that ensure the model to fullfill some quality criterion such as increasing explainability. In this work, the Low-Impact Feature Reduction(LIFR) regularization term to improve the quality of AI models with respect to explainability is presented. We demonstrate the effectiveness of our approach and discuss its potential applications.

2. Related work

Explainable Artificial Intelligence (XAI) has gained significant attention in recent years, with researchers striving to bring transparency and interpretability to complex AI models. The field has witnessed remarkable progress in generating explanations that help users understand the decision-making process of black box models.

One of the most widely used approaches to generate explanations are LLEs, which generates an importance matrix in which each term of the matrix is associated to the influence of feature i for output j. Formally, let $f : \mathcal{X} \to \mathcal{Y}$ be a model to be explained, where $\mathcal{X} \subset \mathbb{R}^n$ is the feature space and $\mathcal{Y} \subset \mathbb{R}^m$ is the output space. Then, an explanation is a matrix $A \in \mathbb{R}^{nm}$ where each $a_{i,j}$ is the importance of the feature i for the output j.

There has been developed different criteria to be able to compare between explanations. The conciseness metric, which is developed as a metric by REVEL[5], enables to differentiate between the most important features from the less important ones. The conciseness of an explanation A is defined as $concision\,(A) = \frac{1}{m-1}\sum_{i=1}^{m} 1 - |v_i|_1$, where v_i is the column of the explanation, that is, the vector of all the importances of the feature i. However, those quality criteria has not been used to impose a bias on training to improve intelligence models.

On the other hand, regularization is a widely used method to impose quality criteria on different models in machine learning[6]. Most machine learning libraries already has regularization terms implemented as default, such as weight decay.[7]. These penalties are generally added to the cost function as a term that does not depend on the output space as it follows:

$$cost\_function(f_\Theta, X, Y) = loss(f_\theta(X), Y) + regularization(\Theta, X),$$

3. Low-Impact Feature Reduction(LIFR) regularization term.

In this work, we use the type of explanations based on importance matrix. We use the concept of derivative as the importance matrix because locally the derivative of a feature with respect to an output is the influence of this feature for this output. So, for an example x, we consider that the explanation of f for this example will be the matrix $\frac{\delta f}{\delta \mathcal{X}}(x) \subset \mathbb{R}^n \times \mathbb{R}^m$.

For the LIFR metrics development, we focus on removing the less important features. If a model uses exclusively the most important features and discards the least important ones, its behavior should be similar. Formally, we consider important features to be those whose absolute importance is greater than the conciseness. Reciprocally, the non-important features are those whose absolute importance is less than the conciseness. Once the prediction of the model and the explanation are obtained, we proceed to delete the non-important features and compare the results of both predictions. The regularization is calculated as the sum of both Kullback-Leibler

Table 1
Comparison between Efficientnet-B2 baseline and LIFR regularization sorted by Accuracy error

Experiment	Accuracy error
Efficientnet_B2_1e-05	13.78%
Efficientnet_B2_5e-05	8.98%
Efficientnet_B2_1e-03	8.75%
Efficientnet_B2_LIFR_1e-03	2.91%
Efficientnet_B2_1e-04	2.29%
Efficientnet_B2_LIFR_1e-05	1.75%
Efficientnet_B2_5e-04	1.28%
Efficientnet_B2_LIFR_5e-05	1.26%
Efficientnet_B2_LIFR_5e-04	1.05%
Efficientnet_B2_LIFR_1e-04	**0.88**%

divergences between the model predictions of the example and the model prediction with just the important features and the original one:

$$d(f(x), f(x')) = -\sum_{i=1}^{n} f(x')_i \log(\frac{f(x)_i}{f(x')_i}) + f(x)_i \log(\frac{f(x')_i}{f(x)_i}) \quad (1)$$

where x' is the example without non-important features and $f(x)$ is the prediction for x. A experimental study is presented below to show the effectiveness of LIFR in image classification.

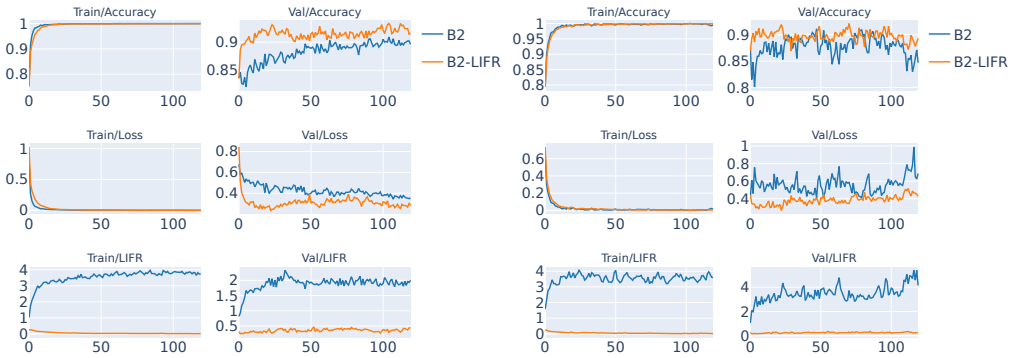
4. Experimental Setup

We use as benchmark the CIFAR 10 image dataset [8], which has 50,000 images for training and 10,000 images for testing. As base experiment, we use the EfficientNet B2[9] and used AdamW optimization [10]. We also use different learning rates ($1e-05$, $5e-05$, and $1e-04$).

5. Results

We evaluated the performance of LIFR to improve EfficientNet B2 behavior on CIFAR-10 with different learning rates. The performance of the model was measured using the classification error on the test set. The results are shown in Table 1 for the test error and Figure 1 for the training and validation scores on accuracy, loss and LIFR regularization, respectively.

As shown in Table 1, the model achieved the best performance with a learning rate of $1e-04$ with the LIFR regularization, achieving a test Accuracy error of 0.88%. Among the baseline models, the model with a learning rate of 5e-04 achieve its best performance of a 1.28% Accuracy Error. These results suggest that the choice of using the LIFR regularization has a significant impact on the performance of the model. To gain a deeper understanding of the model's performance, we also analyzed the training and validation accuracy, loss, and LIFR regularization over time of the two learning rates with best performances. Figure 1 show these metrics, respectively, as a function of the number of training epochs.



(a) Learning Rate: 1e-04 (b) Learning Rate: 5e-04

Figure 1: Comparative analysis over time of the performance of efficientnet-B2(B2) and Efficientnet-B2 + LIFR regularization(B2-LIFR) on training and validation sets. Metrics: Accuracy, loss and LIFR.

Based on the Figure 1 of the training and validation metrics, we can observe that the models without the LIFR regularization generally learn the dataset with fewer training epochs than the models with LIFR. However, as the training progresses, the regularization technique starts to catch up and eventually outperforms the baseline model in terms of validation accuracy.

6. Conclusions

In this work, we develop a theoretical basis for studying explainable artificial intelligence models. Using this base, we develop a novel regularization term called LIFR, which restricts the learning of the model and improves its stability compared to the baseline model.

Our experimental results on the CIFAR 10 image dataset shows that LIFR regularization outperforms the baseline model in terms of accuracy error. Additionally, we have shown that LIFR regularization is effective in preventing overfitting, despite the initial slower learning. As part of our ongoing development, we are currently studying the impact of LIFR on accuracy and loss. Additionally, we aim to explore the development of a new set of LIFR-like regularizations.

Acknowledgments

This work was supported by the Spanish Ministry of Science and Technology under project PID2020-119478GB-I00 financed by MCIN/AEI/10.13039/501100011033. This work was also partially supported by the Contract UGR-AM OTRI-6717 and the Contract UGR-AM OTRI-5987. and projects P18-FR-4961 by Proyectos I+D+i Junta de Andalucia 2018. The hardware used in this work is supported by the projects with reference EQC2018-005084-P granted by Spain's Ministry of Science and Innovation and European Regional Development Fund (ERDF) and the project with reference SOMM17/6110/UGR granted by the Andalusian "Consejería de Conocimiento, Investigación y Universidades" and European Regional Development Fund (ERDF).

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[2] M. Bohanec, M. K. Borštnar, M. Robnik-Šikonja, Explaining machine learning models in sales predictions, Expert Systems with Applications 71 (2017) 416–428.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[4] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, A historical perspective of explainable artificial intelligence, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (2021) e1391.

[5] I. Sevillano-García, J. Luengo, F. Herrera, Revel framework to measure local linear explanations for black-box models: Deep learning image classification case study, International Journal of Intelligent Systems (2023). doi:10.1155/2023/8068569.

[6] X. Zhang, Regularization, Springer US, Boston, MA, 2010, pp. 845–849. URL: https://doi.org/10.1007/978-0-387-30164-8_712. doi:10.1007/978-0-387-30164-8_712.

[7] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[8] A. Krizhevsky, V. Nair, G. Hinton, Cifar-10 (canadian institute for advanced research), URL http://www. cs. toronto. edu/kriz/cifar. html 5 (2010) 1.

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

[10] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. iclr. 2015, arXiv preprint arXiv:1412.6980 9 (2015).