Proceedings of the DataNinja sAIOnARA 2024 Conference 62-64 DOI: 10.11576/dataninja-1173

Closing the Loop with Concept Regularization

Andres Felipe Posada-Moreno ANDRES.POSADA@DSME.RWTH-AACHEN.DE
Sebastian Trimpe TRIMPE@DSME.RWTH-AACHEN.DE
Institute for Data Science in Mechanical Engineering (DSME), RWTH Aachen University, Aachen, Germany

Abstract

Convolutional Neural Networks (CNNs) are widely adopted in industrial settings, but are prone to biases and lack transparency. Explainable Artificial Intelligence (XAI), particularly through concept extraction (CE), allows for global explanations and bias detection, yet fails to offer corrective measures for identified biases. To bridge this gap, we introduce Concept Regularization (CoRe), which uses CE capabilities alongside human feedback to embed a regularization term during retraining. CoRe allows for the adjustments in model sensitivities based on identified biases, aligning model prediction process with expert human assessments. Our evaluations on a modified metal casting dataset demonstrate CoRe's efficacy in bias mitigation, highlighting its potential to refine models in practical applications.

Keywords: Explainable Artificial Intelligence, Concept Extraction, Concept Learning

1. Introduction

Convolutional Neural Networks (CNNs) are extensively used in industrial applications, yet they are opaque and prone to biases and shortcut learning. Explainable Artificial Intelligence (XAI), particularly through concept extraction (CE), offers tools to dissect these models, explain their prediction processes, and detect biases. However, a significant gap remains: CE methods can tell us if a model's predictions are based on the wrong reasons, but offer no solutions for correcting these errors.

XAI research offers two recourse paths: integrating local explanations with intensive human feedback into training loss [1], and ante-hoc techniques like concept bottleneck networks that also require specific architectures [2]. These methods are either labor-intensive or unsuitable for already trained models, with no recourse mechanisms for CE methods [3–5].

To address these limitations, we introduce the method CoRe (Concept Regularization), which uses the concept localization capabilities of ECLAD [4] combined with human feedback to integrate a regularization term in the retraining process of a model. This approach utilizes concept masks from ECLAD [4] to identify and penalize model sensitivities in undesired areas, using a single human feedback input to influence model behavior across the entire dataset.

This abstract works on preliminary results towards a concept-based alignment method. In particular, we introduce the method CoRe, extending CE to provide recourse in bias mitigation. We explore the feasibility of our approach through experiments on a modified metal casting dataset, showing significant mitigation of biases.

2. Concept Regularization (CoRe)

We introduce Concept Regularization (CoRe) as shown in Figure 1, an extension of concept-based explanation methods designed to improve model alignment with human feedback. CoRe is a teacher-student framework, which uses the frozen original model (f) as the teacher to guide the training of a new or adjusted model (f'). Initially, the teacher model is analyzed using ECLAD [4] to extract a set of concepts (C), which are then presented to users to gather feedback, forming a modified set of concepts (C^h) containing only the concepts to adjust and their importance scores ($I^h_{c_j}$). This feedback serves to identify and localize undesired biases, allowing the regularization of the sensitivity of the model in these locations during the retraining of the student model.Finally, the student model is re-evaluated to measure improvements in alignment against the teacher, employing a novel Importance Alignment Score (IAS).

In the retraining phase, we penalize the gradient of the model in the regions containing the undesired concepts, using the term below:

$$L_{\text{CoRe}}(x_i, y_i, f', C^h) = \sum_{c_j \in C^h} \frac{\|\nabla_x g(f'(x_i)) \odot m^{c_j}_{x_i}\|_2}{\|m^{c_j}_{x_i}\|_2}, \tag{1}$$

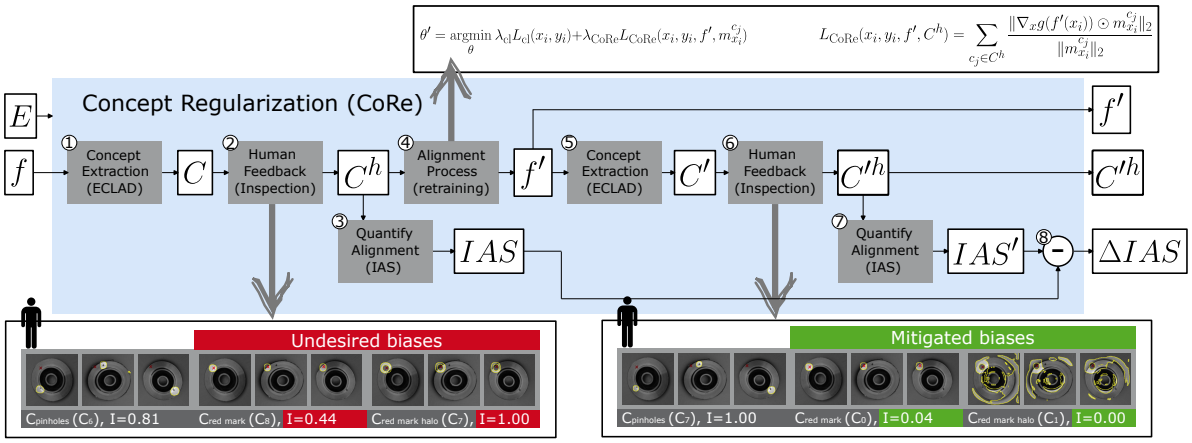

Figure 1: *Concept Regularization (CoRe).* CoRe starts with the Concept Extraction from the model f using ECLAD, followed by human feedback on these concepts to identify biases, and alignment assessment (steps 1-3). The model is then retrained to align with this feedback, and improvements are measured using the Importance Alignment Score (IAS) and its change ΔIAS (steps 4-8).

where $m_{x_i}^{c_j}$ is the concept mask of the input x_i and concept c_j computed with the teacher model f, $\nabla_x g(f'(x_i))$ is the gradient of the wrapped student model f' with respect to the input x, and g is the wrapping function $g(y) = \|y \cdot \mathbf{1}^\top - \mathbf{1} \cdot y^\top\|_2$ introduced in [5].

The IAS quantifies discrepancies between the model's sensitivity to concepts I_{c_j} and the human-assigned importance ratings $I^h_{c_j}$:

$$IAS = \frac{1}{n_c} \sum_{c_j \in C^h} \left| I_{c_j} - I^h_{c_j} \right|, \tag{2}$$

where I_{c_j} represents the importance score assigned by the model for concept c_j, and $I^h_{c_j}$ is the human-assigned importance score for that concept.

This method enables both the retraining of existing models and the use of a teacher model to guide the training of new models with different architectures, learning from the teacher's mistakes.

3. Preliminary Results

We validate the Concept Regularization (CoRe) method using a synthetic dataset and a modified metal casting dataset, presenting results primarily from the latter. The metal casting dataset was altered to include a red mark (bias) alongside desired classification features (pinholes). For our evaluation, we employed a DenseNet121 model, trained to convergence on this dataset. After model explanation and inspection, two concepts were identified as biases: the red mark and its surrounding halo. We then applied CoRe in three settings: *retraining* using the teacher's architecture and weights, *fine-tuning* the classification head only, and training a *new model* from scratch. We tested various learning rates and scaling factor of regularization loss λ_{CoRe}, presenting the best results below. We evaluated the reduction of biases and the change in IAS, with results detailed in Table 1 and the example in Figure 1.

Case	c_{pinholes}	c_{mark}	c_{halo}	Δ IAS
Teacher	0.81	0.44	1.00	-
Retraining	1.00	0.10	0.00	0.67
Fine-tuning	1.00	0.04	0.00	0.7
New model	1.00	0.12	0.51	0.40

Table 1: Importance scores and importance improvement (Δ IAS) after applying CoRe.

Our findings indicate that CoRe significantly minimized the model's reliance on the biases (mark and halo), redirecting focus towards the genuine defect characteristics (pinholes). This improvement was seen across the three settings, showing CoRe's versatility, and its effectiveness in addressing biases in industrial datasets.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy -– EXC-2023 Internet of Production -– 390621612.

References

[1] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2662–2670. AAAI Press, 2017. URL https://www.ijcai.org/proceedings/2017/371.

[2] Nishad Singhi, Jae Myung Kim, Karsten Roth, and Zeynep Akata. Improving Intervention Efficacy via Concept Realignment in Concept Bottleneck Models, 2024. URL http://arxiv.org/abs/2405.01531.

[3] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html.

[4] Andrés Felipe Posada-Moreno, Nikita Surya, and Sebastian Trimpe. ECLAD: Extracting Concepts with Local Aggregated Descriptors. *Pattern Recognition*, page 110146, 2023. URL https://www.sciencedirect.com/science/article/pii/S0031320323008439.

[5] Andrés Felipe Posada-Moreno, Kai Müller, Florian Brillowski, Friedrich Solowjow, Thomas Gries, and Sebastian Trimpe. Scalable Concept Extraction in Industry 4.0. In *Explainable Artificial Intelligence*, pages 512–535. Springer Nature Switzerland, 2023. URL https://link.springer.com/chapter/10.1007/978-3-031-44070-0_26.