1ère Journée des Doctorants et Post-doctorants du programme DATA de l'I-site CAP 20-25

CAP 20-25 I-SITE CLERMONT Clermont Auvergne Project

ID de Contribution: **23** Type: **Poster**

Concept-Based Learning for Explainable AI

Field: Machine Learning / Explainable Artificial Intelligence
Affiliation: LIMOS

Deep learning models have achieved remarkable performance across a wide range of tasks, yet their lack of transparency remains a major obstacle to deployment in high-stakes and human-centered settings. My research addresses this challenge through concept-based learning, a paradigm that introduces human-interpretable concepts as intermediate representations guiding model predictions [1]. While Concept Bottleneck Models (CBMs) [2] offer an appealing framework for explainability, they often suffer from unfaithful concept–class reasoning, where predicted concepts do not meaningfully govern final decisions.

In this work, I present two complementary contributions that aim to improve the faithfulness, stability, and interpretability of concept-based models. First, I introduce KL-Guided Concept Bottleneck Models [3], which combine a transparent probabilistic module with a flexible dense classifier. The transparent module computes class probabilities from empirical concept–class associations, while the dense classifier is softly regularized using Kullback–Leibler divergence to align its predictions with this interpretable structure. This approach preserves predictive performance while encouraging semantically consistent concept usage.

Second, I propose Prior-Anchored Concept Bottleneck Models, which directly constrain the concept-to-class mapping by anchoring classifier weights to empirically estimated concept–class priors in log-odds space. A dynamic anchoring mechanism progressively transitions from annotation-based priors to prediction-based priors as concept quality improves, stabilizing training and yielding more interpretable classifier parameters.

References:
[1] Poeta, E., et al. Concept-based Explainable Artificial Intelligence: A Survey. arXiv:2312.12936, 2023.
[2] Koh, P. W., et al. Concept Bottleneck Models. ICML, 2020.
[3] El Cheikh, R., Falih, I., & Mephu Nguifo, E. KL-Guided Concept-Based Learning for Explainable Classification. XKDD @ECML PKDD 2025.

Auteur: EL CHEIKH, Rim

Co-auteurs: MEPHU NGUIFO, Engelbert (University Clermont Auvergne, LIMOS); FALIH, Issam

Classification de Session: Poster Flash Talks