Contribution ID: **109** Type: **Študenti informatika**

Contrasting Human and Emergent Concepts in Image Classifiers

Wednesday, November 26, 2025 11:12 AM (1 minute)

In the age of AI becoming an everyday partner and support tool in both professional and private domains, users and stakeholders are increasingly confronted with the question of interpretability. This work contributes to the search for answers by exploring hidden meanings in the internal layers of convolutional neural networks trained on image classification tasks. Combining supervised concept-based and unsupervised learning paradigms, the goal is to discover semantically meaningful representations that can be contrasted with human-defined concepts (defined once per dataset). More specifically, we extracted layer-wise network-inherent clusters using hierarchical agglomerative clustering. To evaluate their semantic fidelity, we trained auxiliary classifiers on concepts as well as cluster memberships, and evaluated across all layers.
Our experiments reveal a higher classification accuracy for clusters extracted from each layer compared to human-defined concepts, indicating better separability and indication that clusters may capture patterns beyond human labels. Additionally, the classification accuracy increases for both clusters and concepts toward the output layer. Beyond quantitative evaluation, we provide qualitative insights using visualization techniques such as UMAP projections and Concept Localization Maps. Our findings highlight the potential of hybrid approaches for post hoc explainability and point to promising directions in uncovering emergent structures within deep neural networks.

Pracovisko fakulty (katedra)/ Department of Faculty

Katedra aplikovanej informatiky

Tlač postru/ Print poster

Budem požadovať tlač /I hereby required to print the poster in faculty

Authors: FARKAŠ, Igor; BILA, Tamara (Comenius University Bratislava)

Session Classification: Poster session + káva: prezentácie študentov informatika

Track Classification: Poster session + káva: prezentácie študentov: Poster session + káva: prezentácie študentov informatika