Activation-Based Concept Extraction for Explainability in Image Classification

Matteo Bianchi; Riccardo Campi; Antonio De Santis; Sara Merengo; Marco Brambilla

POLITECNICO MILANO 1863 | DATA SCIENCE LAB

CVPR JUNE 3-7, 2026 DENVER COLORADO

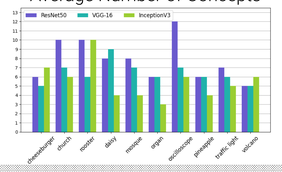
TL;DR

We propose **Activation-Based Concept extraction (ABC)**, a method that automatically extracts human-understandable visual concepts aligned with CNN activations, producing more coherent and interpretable explanations than previous methods.

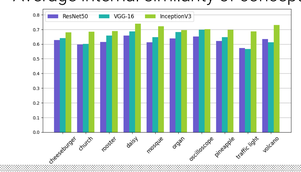
Experiments

- 3 models (VGG-16, ResNet50, InceptionV3) pretrained on ImageNet-1k
- 500 images for 10 classes overall
- Number of concepts between 7 and 20
- Fixed similarity thresholds for concept filtering and merging

Average Number of Concepts

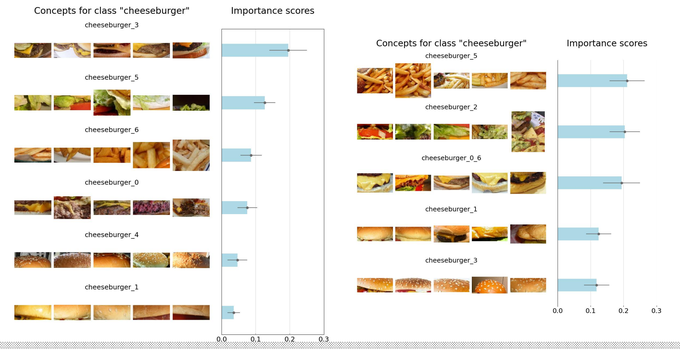


Average internal similarity of concepts

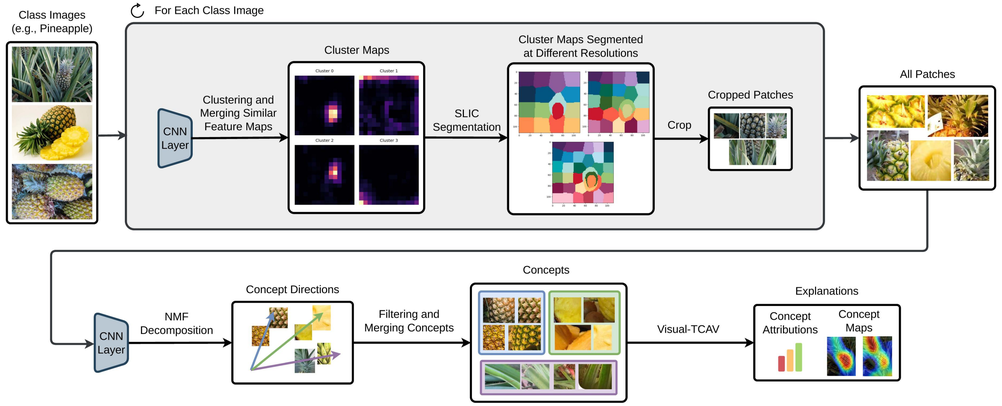


Global Explanations (ResNet50 vs. VGG-16)

ABC can produce global explanations, showing class-level explanations in terms of high-level concepts



Methodology



1. Segmentation

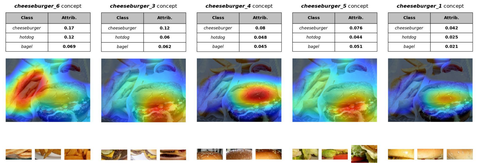
- Clustering feature maps to reduce redundancy
- SLIC segmentation on the feature maps to crop highly activated regions
- Spatial preservation of concepts using rectangular patches

2. Concept Extraction

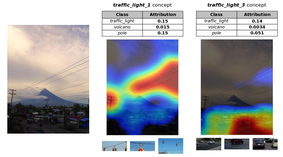
- Iterative selection of the number of concepts
- Cosine similarity of Concept Activation Vectors to merge similar concepts and remove noisy ones

Local Explanations

ABC can provide instance-level explanations, which can be useful to diagnose behaviors in specific images



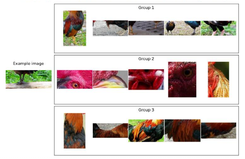
ABC can also be used to understand mispredictions, like the "volcano" image below wrongly classified as "traffic light"



The two most important concepts toward the prediction appear to be the electric cables and the cars on the road

User Evaluation: Gamified Crowdsourcing Activity

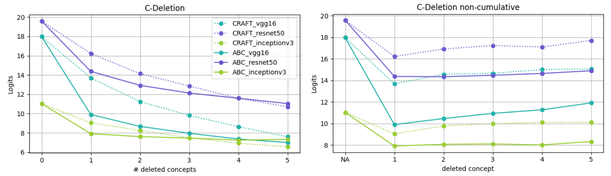
Our user study shows improvements in understandability and coherence of extracted concepts



	ABC (Ours)	CRAFT
ML expert (n=142)	**95.32** ±8.29%	69.35 ±26.11%
Not ML expert (n=74)	**94.12** ±8.83%	69.29 ±32.32%
Overall (n=216)	**94.91** ±8.65%	69.33 ±28.39%

Quantitative Evaluation: C-Deletion Curves

We evaluate the fidelity of ABC using C-Deletion curves, where removing concepts in order of importance leads to an increasing drop in the model logits



Conclusions

Key Findings

- Extraction of concepts aligned with the model's attention
- Iterative selection of the number of extracted concepts
- Refinement step to filter noisy concepts and merge similar ones
- Evaluation on understandability, coherence, and fidelity highlighting improvements over state-of-the-art methods

Future works

- Optimizing parameter selection for filtering and merging
- Extending the pipeline to Vision Transformer architectures
- Automatic concept naming and exploration of applications to ante-hoc explainability