POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Concept dataset extraction for Concept Activation Vectors: A study on open-vocabulary grounding and segmentation

Laurea Magistrale in Computer Engineering - Ingegneria Informatica

Author: Alessandro Cogollo, 233022

Advisor: Prof. Marco Brambilla

Co-advisors: Matteo Bianchi, Riccardo Campi, Antonio De Santis

Academic year: 2025-2026

1. Introduction

Convolutional Neural Networks (CNNs) are widely adopted AI models, even though they are considered to be black-box models. In fact, their internal functioning and the steps to make a decision are not clear. This clearly poses limitations to their adoption in many critical areas, motivating explainability (XAI), the field that design and develop frameworks and methods to understand the functioning of many AI models. This work focuses on the concept-based explainability (C-XAI) methods, that try to explain the reasons that lead a certain model to make a decision, by analyzing the concepts it focused on, and by providing a human friendly explanation. There are many C-XAI methods, the most interesting ones for this work fall into the category of the supervised post-hoc concept-based explainability methods. Among these, it is worth mentioning Testing with Concept Activation Vectors (TCAV) [4], and its extension Visual-TCAV [5]. Both the aforementioned methods require the training of CAVs, which are vectors pointing to the concept set of examples. Training CAVs requires a sufficient amount of concept pictures, that in many cases are hard to obtain, as the curation of these datasets is costly under many aspects. At the time of the writing, only a few pipelines exist to automatically extract concept images from textual prompt; one of them is proposed by Campi et al. [2], and is based on the generation of synthetic images. Astolfi et al. [1] instead show that zero-shot generation does not produce realistic concepts. Briefly, this work places itself as an alternative to the synthetic generation approach, proposing a modular pipeline to automatically extract image crops containing a target concept. This end-to-end pipeline takes as an input a textual prompt using open-vocabulary detection and segmentation models (in this case, SAM3). The main research question investigates whether it is possible to automatically obtain a dataset that correctly represents a concept, using a textual prompt and a starting set of images as input; to answer this, the following four questions have been posed. The first research question (RQ1) assesses to what extent the intrinsic quality of the manually curated dataset compares with the one of the automatically extracted set. With the second research question (RQ2), we assessed how stable and robust a CAV generated from the automatically extracted dataset is. The third research question (RQ3) examines how CAVs obtained

from the automatically extracted data are similar to the manually curated ones. Finally, the goal of RQ4 is to understand what are the most influential components of the pipeline, and to compare the extracted dataset with the manually curated one. This work is intended as an exploratory study, providing a good baseline for further experiments and studies.

2. Proposed pipeline and methodology

The proposed architecture is composed of three modules: an extractor, a cropper, and an augmenter. The extractor's goal is to identify a bounding box and a mask containing the target concept, given one or more pictures. The extractor is based on an open-vocabulary detection and segmentation module, which, in the case of this work, is the Segment Anything Model 3 (SAM3) [3] by `Meta`, which represents the state-of-the-art of open-vocabulary detection and segmentation modules. It natively works as both a grounding and segmentation model, and introduces the Promptable Concept Segmentation (PCS), permitting the localization of concepts, objects and textures by using noun phrases, or example images. Once the extraction phase is performed, the coordinates localizing the target concept are passed to the cropper, which is instead responsible for cropping the picture, using one of four cropping policies. These cropping policies are: sliding_window, bbox, center_mask or largest_bbox; all of them turn the mask provided by the extractor into a square crop. First, the `sliding_window` strategy ensures that the crop captures the biggest square in the mask. The `bbox` policy centers the square on the geometric center of the bounding box. The `center_mask` strategy instead centers the crop on the mask's center of mass. Finally, the `largest_bbox` ignores the masks and crops the starting picture around the biggest bounding box provided by SAM3. The last step of the proposed pipeline is the augmenter, which is responsible for applying standard augmentation techniques to the crops provided by the previous modules. The augmenter can be enabled optionally, and applies a random combination (at least one) of the following transformations: rigid rotations of multiples of 90°, horizontal shearing, center crop-and-zoom, horizontal and vertical flips, and changes in the colors and brightness. Its function is to increase the number of pictures if the extracted crops are not enough to reach the required number of samples, or to increase the diversity of the positive samples. Figure 1 presents a schema of the pipeline.
The proposed pipeline takes as input a pair, composed of a set of images from a chosen starting domain, and a textual prompt, describing the target concept. In turn, it outputs a set of images containing the target concept.

3. Experimental setup

This section presents the experimental choices made to both generate the extracted dataset, and to test the pipeline.
To evaluate the extracted dataset in terms of similarities with the original dataset, we adopt the following setup: first, we apply global average pooling (GAP) as well as a standardization on the shared negatives. Also, instead of using logistic regression to compute the CAV, we use difference-of-means, as it is shown to improve stability, also coherently with other works; CAVs are computed as shown in formula 1

$$v_c^l = \frac{\mu_c^l - \mu^{-,l}}{||\mu_c^l - \mu^{-,l}||_2} \tag{1}$$

where c represents the concept, at layer l, and CAV is computed as the vector linking the two centroids.
We evaluate along four different axes: intrinsic quality, stability over different samplings, similarity with the baseline layer by layer, and finally an ablation study, as will be discussed better in the next sections.
The following datasets are used to extract the concept classes considered ground truth for comparison with automatically extracted concept datasets. To extract the cropped images, we use the `ImageNet` dataset. The following datasets are instead used as manually curated datasets, useful to train CAVs on a consolidated image set, and then to compare the CAVs: `Zero-shot Text-to-Image (T2I)` Dataset, which contains also synthetically generated images, and the `Describable Textures Dataset (DTD)`, both consisting of a collection of textural images.
We use the following backbones: as a baseline, the `ResNet50`, using the `IMAGENET1K_V1`

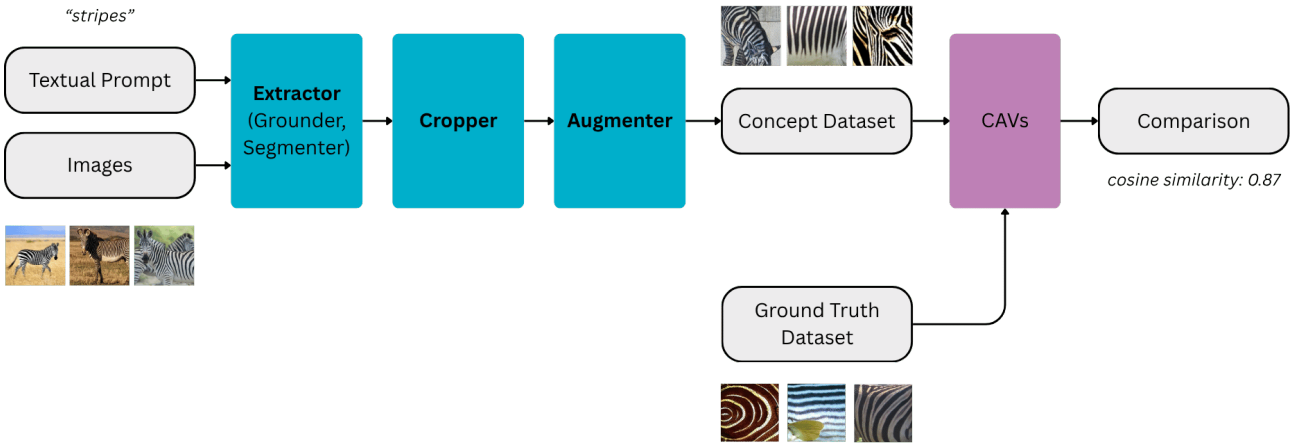

Figure 1: Schema of the proposed pipeline.

weights, and the `torchvision` implementation; to conduct ablation studies, the `VGG16` and `InceptionV3`, both using the `torchvision` implementation.

The chosen hooks to extract the activations from are four; they were chosen at increasing depths of the network architecture (one early, two intermediate, and one late), so that the CAVs quality and other metrics could be evaluated at different stages of the feature-extraction pipeline.

To perform the first experiment, the `Qwen2-VL-2B-Instruct` is chosen as the Vision Language Model (VLM) utilized as a judge. The Contrastive Language-Image Pre-training (CLIP) model instead is the CLIP `ViT-B/32`. The confidence threshold for the VLM is 0.25, as well as the threshold of SAM3, which is 0.25; below this threshold, the mask is not cropped by the cropper, as the model is not sure that the target concept is contained in the picture. SAM3 has been used with its 3.1 checkpoint, and the publicly available release on `GitHub`.

The number of images chosen to perform the experiments is 100-120 per concept class and starting domain. These numbers were chosen after conducting a stability experiment, showing that 100-120 samples allowed obtaining a stable CAV. No limit on the starting domain was set, to ensure that the maximum number of concept pictures would be available for the experiments. We ran both the extraction and the experiments on Kaggle using GPU T4x2, and on Colab using an A100 GPU.

The concepts explored in this work are nine, and are the following: `striped, chequered, dotted, wood, water, braided, bubbly, fibrous, veined`; these concepts have been chosen as they belong to different categories, enabling the exploration of behaviors of different types of concepts (e.g. *easy* concepts that can be found in many areas, as well as materials, and attributes and shapes); also, for the aforementioned concepts, we were sure of having a baseline. All of the concepts are extracted starting from three different starting domains: single-class, multi-class and random-class; the single-class domain takes as a starting class a single ImageNet class (e.g. crosswords), whereas the multi-class aggregates a set of classes (e.g. crosswords, quilt); both of these domains consider classes that are likely to contain the target concept. Finally, the random-class starting domain chooses random starting classes. The choice of the starting domain classes has been suggested by ChatGPT, as well as the prompts for both the extractor and the VLM judge. This choice is an experimental one, done to simplify the choice of the classes when dealing with many classes (e.g. ImageNet); classes were fixed for reproducibility purposes. Prompts proposed by ChatGPT follow a short noun-phrase, and a few synonyms; an example is reported for the `striped` concept class, which used the prompt: `"striped pattern", "stripes", "striped texture"`.

4. Results

In this section, the results are evaluated after the experiments have been conducted.

RQ1 - Intrinsic quality. The experiments run for the first research question show that the intrinsic quality of the extracted dataset is comparable with the manual baseline, when considering the extractions made on the single-class and multi-class starting domain. A strong degradation of the quality metrics is observed instead for the random-class starting domain. These metrics do not show an inability of the pipeline to extract the concept from random classes, but instead show that a way higher number of images in the starting domain is needed; in fact, in many cases, very few images are extracted, as concepts can be found, but with a lower frequency. For example, for the striped concept class, in which the metrics in terms of mean VLM-as-judge alignment, are 0.87 for the baseline (DTD), against ~ 0.84–0.92 for the extractions on the single-class and multi-class starting domain; these metrics drop to ~ 0.50–0.57 for the random-class starting domain, resulting in a gap of almost 0.36. The lowering of the results for the random-class can be explained by the high variance, and low presence of the concept in the starting classes, rather than by the inability of the extractor to correctly identify the concept: aggregating the concepts extracted from the random domain has a mean alignment score which is lower by $\Delta = 0.27$ on average, with respect to the single-class domain.
As proven by the intrinsic evaluation metrics, extractions performed on the multi-class starting domain also introduces greater variation between images as they are cropped from a more diverse starting domain. Intrinsic quality is measured in terms of: mean alignment, std, share below 0.25, precision at 0.5 and 0.7, and other metrics depending on a CLIP model (`ViT-B/32`) which are the repetition rate, near-duplicate pairs, participation ratio.

RQ2 — Stability and robustness. Regarding the second research question (RQ2), the test performed evaluates the intra-similarity of CAVs generated with 20 different re-trainings on different negative draws, with the goal of assessing their stability with respect to the sampling of negatives; results indicate that CAVs in the third and fourth layers reach an high intra-similarity, which is of 0.975 in the third layer (mean 0.960 across the four layers, for *striped*). Also, we conducted a test to find the minimum number of samples required to train a stable CAV, showing that intra-trial CAV cosine similarity reaches a plateau around $\sim$ 75-100 samples. To be conservative, we chose to fix 100 as the minimum for a reliably stable CAV.

RQ3 — Automatic vs. manual similarity. Experiments conducted for the third research question showed that similarities between automatic versus manual grow in the intermediate and deep layers: from ~ 0.3 in the early layer to ~ 0.65 in the intermediate/deep layers; textures reach the highest scores in the third layer, meanwhile materials (*water*, *wood*) are captured by the intermediate layer and degrade in the last one. This behavior is illustrated in Figure 2. Also, intra-similarity per concept is measured.

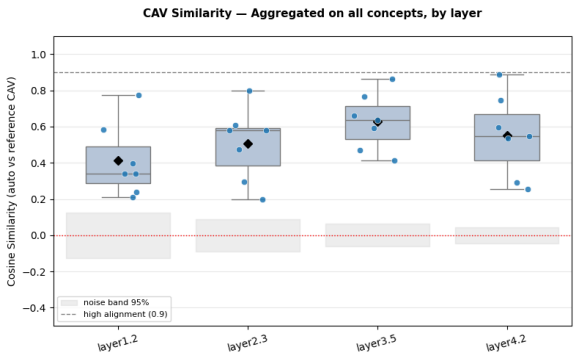

Figure 2: Cosine similarity evolution per layer, between CAVs trained on automatically extracted images, and CAVs trained on manually curated sets of images.

RQ4 — Ablation and SOTA comparison. Finally, the last experiments are dedicated to ablation studies. First, the most important driver of quality we observed is the starting domain, which is the dominant factor. In fact, the coherence and quality of the extracted dataset depend on the presence of the concept in the source domain; experiments show that the extracted dataset often does not reach the required amount of pictures before running out of samples in the starting pool. An ablation study has been conducted also to evaluate the cropping strategies; we observed that they do not lead to substantial improvements (e.g. in terms of VLM mean alignment: `bbox`: 0.66, `center_mask`: 0.70, `largest_bbox`: 0.67,

`sliding_window`: 0.65), but instead allow obtaining similar CAVs or larger crops that are better for the intrinsic metrics.
Experiments conducted to evaluate the effects of the augmentation techniques show that augmentation lowers the intrinsic VLM-alignment score by ~ 0.05 ($0.70 \rightarrow 0.65$ on the baseline config) but improves CAV similarity.
Experiments were also conducted on different backbones, showing slightly different performances, but an overall coherence between the models, meaning that the behavior observed on the CAVs using `ResNet50` is not necessarily dependent on the architecture. `InceptionV3` and `ResNet50` have higher similarities in the intermediate layers, with the `ResNet50` being the more stable.
Prompt variations have been evaluated too, comparing datasets extracted with three different prompt variations: Bare-term, in which the bare word is looked for in the picture, Surface-phrase and Texture-phrase, in which respectively a surface and a texture with the concept are looked for in the pictures; variations in terms of intrinsic quality metrics are minimal in terms of VLM mean, but can be high in terms of precision and CAVs similarity.

Metric	Bare	Surf.	Text.	Base
Mean VLM	0.700	0.669	0.717	0.687
% < 0.25	4.43	0.64	2.76	3.33
P@0.5 (%)	88.0	73.3	85.8	82.2
P@0.7 (%)	55.4	53.0	64.2	58.3
Size N	47	41	48	60

Table 1: Results show intrinsic quality metrics computed on different set of images extracted using different prompts; "Mean VLM" stands for "mean VLM alignment". N stands for the average number of samples per concept.

Finally, a comparison with the work of Campi et al. [2] has been conducted, to evaluate the performances obtained by the two pipelines. For coherence with the work of the synthetic generation pipeline, it has been conducted on 35 concepts and three backbones. The CAVs extracted on the proposed extraction pipeline reach an overall cosine of 0.61 ± 0.18 (Search 0.65, DTD 0.61, ImageNet 0.59; textures 0.64, objects 0.59), meanwhile the synthetic ones score 0.77. This gap can be partially and plausibly explained by the absence of a CLIP filter for our pipeline. Domain dependence is well illustrated by the *asparagus* case, where only 17 valid crops are extracted out of more than 4000 scanned images (a $\sim 0.4\%$ success rate). Intra-class similarity is instead very close between the two pipelines.

5. Conclusions

To conclude, we present a reproducible and well documented pipeline to automatically extract a set of concept images, as an alternative to manual curation or synthetic generation. Findings show that the extracted dataset has an overall quality comparable with the baseline, but strongly dependent on the source domain. The CAV similarity changes during the layers, and reaches the highest score in the deep layers, ($\sim 0.31 \rightarrow 0.67$), even though these results are moderate in terms of CAV similarity, but high in terms of intrinsic quality. Also, it is reasonable to hypothesize that the proposed pipeline might be preferable when the goal is to obtain a dataset without hallucinations but instead with realistic textures, or when the generation is not feasible, even though this has not been tested in this work, and is left to future works. Since the generative pipeline proposed by Campi et al. do not depend on the starting domain, in principle, it might be preferable when dealing with small datasets from which to extract the images; experiments to confirm this hypothesis are left to future works.
Several limitations to the presented pipeline are discussed here: first of all, in this experiment, CAVs are trained using full-scene random ImageNet pictures as negatives; this could lead to the CAV learning the difference between a cropped picture containing the target concept, and the full-size scene. Another limit is that the VLM, the CLIP and the SAM3 models used as judges share similar architectures, risking a circularity in the extraction and evaluation. A final note showing one of the limitations of this work is the number of concepts tested, which leads to qualitative results, instead of quantitative results.
These limitations naturally lead to listing some

future works. First, an expansion of the proposed pipeline could be, the implementation of a CLIP filter, consistently with other pipelines, to select the images before computing the CAVs. Another extension could be the implementation of an automatic class selector using a large language model. Finally, it would be interesting to extend the pipeline also to other families of architectures such as Vision Transformers.

References

[1] Giacomo Astolfi, Matteo Bianchi, Riccardo Campi, Antonio De Santis, and Marco Brambilla. A Framework for Evaluating Zero-Shot Image Generation in Concept-Based Explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3303–3311, 2026.

[2] Riccardo Campi, Santiago Borrego, Antonio De Santis, Matteo Bianchi, Andrea Tocchetti, and Marco Brambilla. Towards Synthetic Concept Activation Vectors via Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2745–2753, 2025.

[3] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts, March 2026.

[4] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), June 2018.

[5] Antonio De Santis, Riccardo Campi, Matteo Bianchi, and Marco Brambilla. Visual-TCAV: Concept-based Attribution and Saliency Maps for Post-hoc Explainability in Image Classification. *Transactions on Machine Learning Research*, December 2025.