A Comparative Study of Pre-defined and Automatically Discovered Concepts for Interpretability

Valerio Costa
valerio.costa@studio.unibo.it

Luca Domeniconi
luca.domeniconi5@studio.unibo.it

June 2025

Abstract

The demand for interpretability in Artificial Intelligence (AI) is growing, particularly in high-stakes domains where understanding model decisions is paramount for trust and accountability. This project proposes a comparative study of two distinct approaches to achieving interpretability in computer vision models: Concept Bottleneck Models (CBMs), which leverage human-annotated, pre-defined concepts , and Concept Recursive Activation FacTorization (CRAFT), which automatically discovers concepts from trained neural networks. We aim to explore the strengths and weaknesses of each paradigm in terms of their ability to provide understandable and actionable explanations. Furthermore, we will investigate the semantic alignment between human-engineered concepts and those learned automatically. This report details the methodology for implementing and evaluating both methods on a relevant dataset, culminating in a detailed analysis of their interpretability characteristics and a discussion of their ethical implications.

1 Introduction

The increasing integration of Artificial Intelligence (AI) into critical sectors necessitates a deeper understanding of how models arrive at their decisions. This need for transparency and accountability has fueled research into interpretability. Our project focuses on concept-based interpretability in computer vision, a method that explains model predictions in terms of high-level, human-understandable concepts.

We conduct a comparative study of two prominent paradigms:

- **Concept Bottleneck Models (CBMs)**: These models incorporate a dedicated bottleneck layer composed of predefined, human-annotated concepts and are trained to predict these concepts before using them to make the final predictions. This structure allows for understandable insights into how each concept contributes to the model's decision, improving interpretability and adding human intervention or correction.
- **Concept Recursive Activation FacTorization (CRAFT)**: Is a post-hoc interpretability method that does not require any concept annotations unlike CBMs. It analyzes instead the internal activations of a pre-trained convolutional neural network (CNN) to automatically find and extract concepts that are identified as meaningful. These discovered concepts can then be used to explain the model's decisions in interpretable way.

The primary goals of this study are to compare the interpretability provided by these two approaches, investigate the semantic alignment between their respective concepts, and analyze the ethical implications of using pre-defined versus automatically discovered concepts for ensuring transparency in AI systems.

2 Background and Motivation

2.1 The Need for Interpretability

In high-stakes domains such as medical diagnosis, autonomous vehicles or critical infrastructure management, "black box" nature of complex AI models is a major challenge to their responsible deployment.

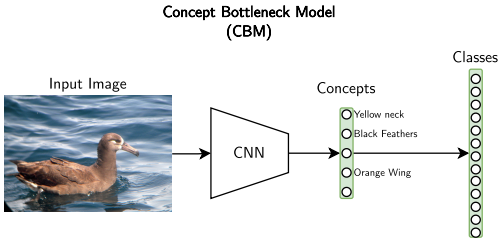


Figure 1: Architecture of a Concept Bottleneck Model (CBM). The model first predicts a set of human-defined concepts from the input image, which are then used as the sole features for the final classification. This structure enables interpretability by exposing intermediate concept-level decisions.

These systems often involve deep neural networks with millions, or even billions, of parameters making it difficult for users and developers to understand how particular decisions are made. This lack of transparency can undermine trust, impede error analysis, and raise concerns about bias, fairness, and accountability—especially when decisions have direct consequences for human well-being.

Interpretability, therefore, is not just a desirable feature but a necessary requirement in these contexts. It enables stakeholders to scrutinize model behavior, identify potential failures, understand edge cases, and ensure that decisions are made for the right reasons. Moreover, interpretability supports regulatory compliance and ethical AI development by providing insights into how models respond to different inputs and whether they treat different groups equitably.

One promising approach to achieving interpretability is through concept-based explanations, which seek to bridge the gap between abstract model representations and human reasoning. Rather than relying solely on pixel-level or feature-level importance scores, concept-based methods aim to describe model behavior in terms of meaningful, high-level concepts that humans can easily understand, such as anatomical structures in medical imaging, traffic signs in autonomous driving, or object attributes in image classification. By aligning machine representations with concepts that are semantically familiar and intuitively understandable, these explanations make it easier for users to validate and trust the model's decisions, especially in critical decision-making scenarios.

2.2 Concept Bottleneck Models (CBM)

Concept Bottleneck Models, introduced by Koh et al. [4], structure a neural network to first map an input image to a set of pre-defined, human-understandable concepts, and then use only these concept activations to make the final class prediction. This forces the model to reason in terms of the concepts provided by humans, making the decision process inherently interpretable. The model's reliance on a "concept bottleneck" allows users to directly inspect the importance of each concept for a given prediction. See Figure 1 for a graphical representation of CBM.

2.3 Concept Recursive Activation FacTorization (CRAFT)

In contrast, CRAFT, proposed by Fel et al. [2], is a method for automatically discovering concepts from a trained neural network without prior human annotation. It works by factorizing the activation tensors of intermediate layers to identify a set of concept vectors that can reconstruct the original activations. This post-hoc approach provides a way to peer inside existing "black box" models and extract meaningful concepts that the model has learned on its own, offering insights into its internal logic. See Figure 2 for a graphical representation of CRAFT.

3 Methodology

The core of our project is a parallel implementation and evaluation of a CBM and a standard CNN analyzed with CRAFT, as depicted in our conceptual workflow.

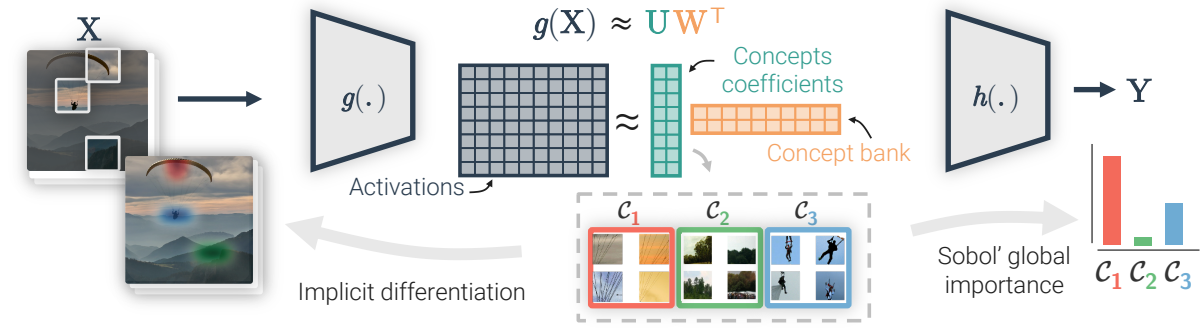


Figure 2: Architecture of the CRAFT (Concept Recursive Activation FacTorization) method. CRAFT is applied post-hoc to a trained CNN by factorizing internal activations to automatically extract meaningful visual concepts. These emergent concepts can then be analyzed for their contribution to the model's predictions, offering interpretability without requiring annotated concepts.

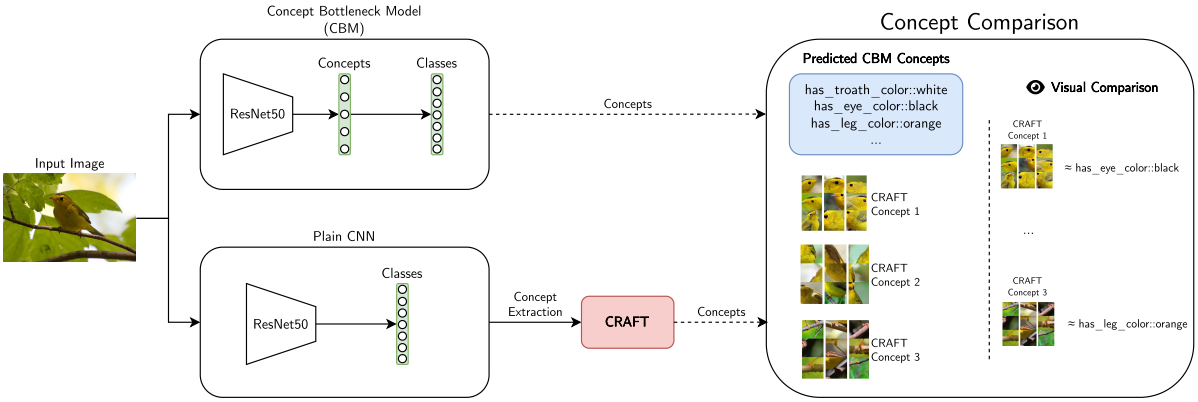


Figure 3: Overall project workflow. On the top row we use the concept naturally learned by the CBM. On the bottom row we extract the concept learned by a ResNet50 by using CRAFT. After the concept extraction, we compare the extracted concept by CRAFT with the predicted concepts from the CBM by a visual comparison.

3.1 Conceptual Workflow

As illustrated in Figure 3, our methodology involves two distinct pathways that converge for a final comparative analysis:

1. **CBM Path**: A Concept Bottleneck Model is trained using input images and their corresponding human-annotated concepts. This model produces class predictions based on an intermediate vector of concept activations.

2. **CRAFT Path**: A standard ResNet50 is trained for the same classification task, but without access to the concept annotations. After training, the CRAFT algorithm is applied to extract concepts automatically from the CNN's internal layers.

3. **Concept Comparison**: The concepts from the CBM (pre-defined) and CRAFT (discovered) are compared qualitatively. This involves extracting the most important image crops that activates a certain CRAFT concepts and visually compare them with predicted CBM concepts.

3.2 Dataset

We will use the Caltech-UCSD Birds 200 (CUB-200-2011) dataset[8]. This dataset is ideal for our study as it contains images of 200 bird species along with 312 binary attribute annotations (e.g., "has_a_red_beak", "has_a_striped_wing"). These detailed annotations provide the ground-truth, human-defined concepts required for training the CBM and serve as a basis for evaluating the semantic meaning of the concepts discovered by CRAFT. As noted by Koh et al. [4], the attribute annotations in the CUB dataset are often noisy and inconsistent, which makes training a reliable CBM particularly challenging. To mitigate this issue, we computed the frequency of each concept within each class and retained only those concepts that appeared in more than 50% of the class's images. In other words, a concept was selected if it was present in at least half of the examples for a given class. This preprocessing step ensures that all images within the same class share identical concept annotations—even though, in reality, this assumption does not always hold. While this approximation simplifies training, it introduces its own limitations: some of the selected concepts still show weak correlation with class-discriminative features (see Section 5 for examples).

3.3 Model Architectures

- **Concept Bottleneck Model (CBM)**: The CBM will consist of two main components:

  1. A **Concept Encoder**: A CNN backbone (e.g., ResNet50) that takes an input image and outputs a vector of concept activations, trained to predict the human-annotated attributes [4].
  2. A **Concept Predictor**: A simple linear layer or small multi-layer perceptron (MLP) that predicts the final class (bird species) using only the concept activations from the encoder as input.

  The model will be optimized using a joint loss function that combines concept prediction accuracy and final label prediction accuracy.

- **Classical CNN for CRAFT**: We will train a standard ResNet50 architecture on the bird species classification task, without explicitly training it on the concept annotations. After training, we will apply the CRAFT algorithm to the intermediate layers of this model to automatically discover concepts[2].

4 Evaluation and Analysis

The evaluation will be multifaceted, covering classification performance, interpretability, and ethical considerations.

Method	Accuracy	Precision	Recall	F1-score
CBM	0.666	0.708	0.666	0.659
CRAFT (Plain CNN)	**0.774**	**0.789**	**0.774**	**0.772**

Table 1: Comparison of the quality metrics for CMB and CRAFT methods, computed on the test set. Note that for CRAFT, we simply train a ResNet50 without any modification (Plain CNN).

4.1 Classification Metrics

To ensure that interpretability does not come at a significant cost to performance, we will evaluate both the CBM and the standard CNN using standard classification metrics, including Accuracy, Precision, Recall, and F1-score.

As can be seen from Table 1, the classification metrics on both methods are very similar, indicating that focusing on explainability doesn't necessarily worsen the performances.

4.2 Interpretability Metrics and Visualization

A fundamental challenge in this comparative study is that the concepts derived from the Concept Bottleneck Model (CBM) and those discovered by Concept Recursive Activation FacTorization (CRAFT) are not directly comparable in their native representations. The concept activations from each model exist as vectors within distinct, high-dimensional latent spaces that lack any inherent alignment. The CBM's concept space is explicitly defined by, and has a dimensionality equal to, the set of human-annotated attributes. In contrast, the concepts from CRAFT are discovered automatically through the factorization of a pre-trained network's activation tensors. The resulting vector space is an emergent property of the model's learned internal logic, with dimensions that have no pre-assigned meaning.

Although CBM and CRAFT operate in fundamentally different concept spaces, we initially attempted a direct comparison using attribution maps (e.g., Grad-CAM) to assess spatial alignment between their concepts. However, this approach proved to be ineffective. CBM concepts, while semantically clear, lack inherent spatial grounding, making it difficult to generate meaningful localization maps (see Section 4.3 for more details). In contrast, CRAFT concepts naturally correspond to localized visual patterns.

Due to this limitation, we shifted to a qualitative comparison strategy focused on visual and semantic correspondence. Specifically, we followed these steps:

1. Extract predicted concept activations per class from the CBM.
2. Extract discovered concepts per class using CRAFT.
3. For CRAFT only, generate attribution maps and identify the most relevant concept crops.
4. Visually compare CBM predictions with the spatial localizations of CRAFT concepts to evaluate semantic overlap.

This method enabled a more practical comparison given the representational differences between the two approaches.

4.3 Attribution maps Experiment

To compare CBM and CRAFT concepts, we explored whether they rely on similar regions of the input image by generating attribution maps for individual concepts. Using attribution methods such as GradCAM++, ScoreCAM, LayerCAM, and DeepLiftShap, we computed heatmaps for both CBM and CRAFT concepts and then measured the similarity between them using cosine similarity. This reframes the comparison from asking "Are these concepts semantically similar?" to "Do these concepts attend to the same parts of the input?", allowing us to quantitatively assess concept alignment without relying on human interpretation.

We conducted this experiment across multiple architectures (e.g., ResNet50, InceptionV3) and attribution techniques. However, results revealed a critical limitation of CBM: the attribution maps for all predicted concepts, within a single image, were nearly identical. This suggests that the CBM does not localize concepts independently. Instead, it appears to treat groups of co-occurring concepts as class proxies, focusing on broad discriminative regions rather than concept-specific areas (see Figure 4 for examples).

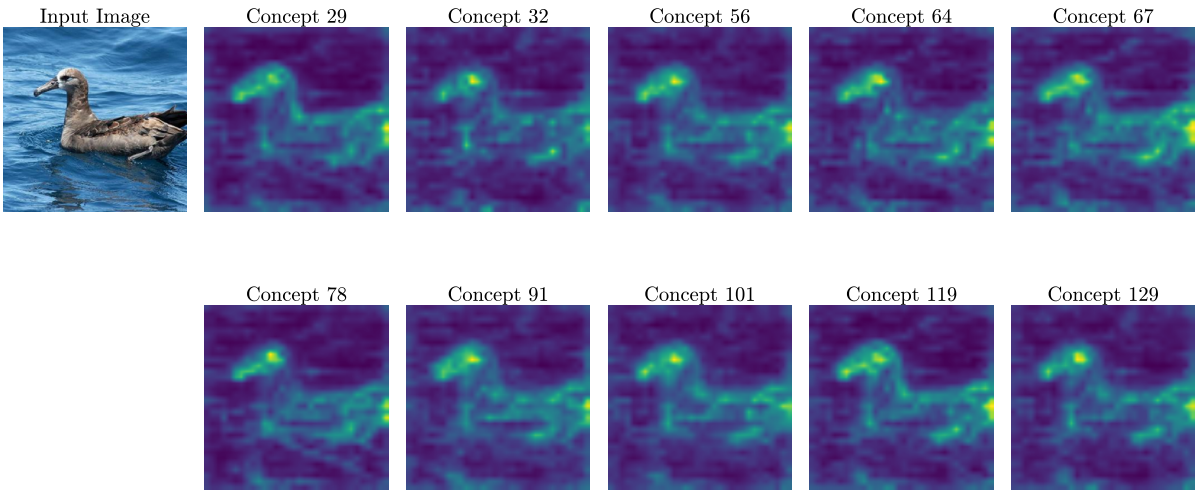


Figure 4: Attribution maps computed for different concepts for the CBM model (with ResNet50 backbone). The attributions are computed using GradCAM++ and, on average, they show a cosine similarity of 0.9825.

These findings highlight a fundamental challenge when using CBMs for spatial interpretability: **although concept predictions are interpretable at the semantic level, they lack the spatial disentanglement necessary for reliable localization.**

5 Results

The qualitative analysis described in Section 4.2 was conducted to better understand the behavior of both CBM and CRAFT during the classification task. Specifically, we examined the visual correspondence between the human-annotated concepts predicted by CBM and the localized, automatically discovered concepts produced by CRAFT across multiple input examples for each class.

The visualizations reveal that CRAFT is often able to extract concepts that closely resemble those used by CBM, while also identifying their spatial locations within the input image. Since CRAFT is not constrained by predefined annotations and is trained in a post-hoc manner on a standard CNN (e.g., ResNet50), it is free to exploit any data-driven regularities, including unintended dataset biases. For instance, if a certain bird species consistently appears against a sea background, CRAFT may treat the background itself as a meaningful concept. CBM, in contrast, cannot leverage such contextual cues, as it only reasons over a fixed set of human-defined attributes.

This behavior is evident in Figure 5, where concepts discovered by CRAFT often include background elements such as grass, water, or sky. While these features are visually consistent, they are not truly discriminative of bird species and can introduce misleading explanations.

Our observations can be summarized into the following key findings:

- CBM underperforms compared to a plain CNN trained without concept supervision (see Table 1).
- The concepts extracted by CRAFT often align semantically with those of CBM, except in cases where dataset bias or concept label noise is present.
- While CBM provides semantically intuitive explanations, it lacks spatial localization of concepts, limiting its ability to offer visually meaningful insights.
- Given the trade-off between interpretability and performance, and the added benefit of spatial localization, CRAFT appears to be a more effective tool for explainability in this specific setting.

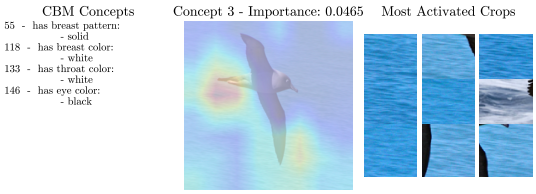
6 Ethical Considerations and Regulatory Implications

Interpretability is not only a technical objective but also an ethical and legal requirement, particularly under Article 22 of the GDPR, which grants individuals the right not to be subject to decisions based

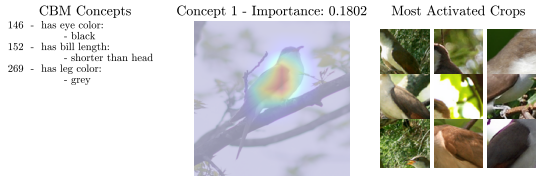
(a) Correctly recognized concept

(b) Correctly recognized concept

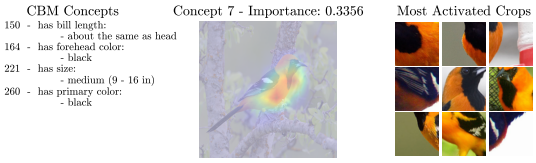
(c) Concept derived from bias



(d) Concept derived from bias



(e) Concept not present in CBM



(f) Concept not present in CBM (dataset noise)

Figure 5: 6 examples of concept extracted using CRAFT. (a) and (b) shows examples of correctly recognized concept from CRAFT. (c) and (d) shows examples of CRAFT concepts derived from the bias (flower and water). (e) is an example of CRAFT concept not present in CBM. (f) is an example of CRAFT concept not present in CBM probably caused by dataset concept noise.

solely on automated processing without meaningful explanation [3]. Concept Bottleneck Models (CBMs), by explicitly using human-annotated concepts, provide inherently interpretable decision processes that align well with this requirement. In contrast, CRAFT offers post-hoc interpretability, which, while valuable, may fall short of legal standards for transparency and accountability [7].

Both CBMs and CRAFT generate inferred data—high-level semantic representations not directly provided by the data subject. According to data protection principles, such inferred data may still be considered personal data, especially if it affects decisions about individuals [6]. CRAFT, in particular, poses ethical risks if its discovered features correlate with protected attributes, potentially resulting in indirect discrimination [1]. CBMs, by relying on explicitly chosen concepts, offer greater control and fairness assurance.

Finally, under the upcoming EU AI Act, high-risk AI systems are required to ensure transparency, human oversight, and robustness [5]. CBMs better support these principles by design, aligning with the "data protection by design and by default" doctrine (GDPR Article 25). While CRAFT can augment interpretability for black-box models, its deployment must be accompanied by rigorous oversight to meet ethical and regulatory expectations.

7 Conclusion

This project compared two concept-based interpretability methods: CBMs, which rely on human-defined concepts, and CRAFT, which discovers concepts automatically. Our experiments showed that CBMs are easier to interpret but perform worse and lack spatial information. CRAFT, while more flexible and visually informative, can be harder to control and may pick up on dataset biases. From a technical point of view, CRAFT offers better overall explainability. However, from an ethical and legal perspective, CBMs are stronger, as they provide clearer and more controlled explanations that better align with current regulations on transparency and accountability in AI systems.

References

[1] Solon Barocas and Andrew D Selbst. "Big data's disparate impact". In: *Calif. L. Rev.* 104 (2016), p. 671.

[2] Thomas Fel et al. "Craft: Concept recursive activation factorization for explainability". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2711–2721.

[3] *General Data Protection Regulation (GDPR)*. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`. Regulation (EU) 2016/679. 2016.

[4] Pang Wei Koh et al. "Concept bottleneck models". In: *International conference on machine learning*. PMLR. 2020, pp. 5338–5348.

[5] *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. European Commission. COM/2021/206 final. 2021.

[6] Andrew D Selbst and Solon Barocas. "The intuitive appeal of explainable machines". In: *Fordham L. Rev.* 87 (2018), p. 1085.

[7] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31 (2017), p. 841.

[8] Catherine Wah et al. "The caltech-ucsd birds-200-2011 dataset". In: (2011).