Extending CRAFT: Cross-Architecture and Cross-Modal Concept Factorization

Laury Magne *laury.magne@telecom-paris.fr*
Télécom Paris

Josephine Bernard *josephine.bernard@telecom-paris.fr*
Télécom Paris

Phuong N-M-NGUYEN *nnguyen-24@telecom-paris.fr*
Télécom Paris

Cecile Li *cecile.li@telecom-paris.fr*
Télécom Paris

Reviewed on OpenReview: *https://openreview.net/forum?id=XXXX*

Abstract

We revisit Concept Recursive Activation Factorization (CRAFT) by Fel et al. (2023), originally introduced on ImageNet-classified images with a ResNet-50 backbone and extend its explanatory scope in three directions. First, we evaluate CRAFT on four semantic categories (animals, humans, objects, vegetation) to assess category-specific concept formation. Second, we replace the convolutional backbone with a self-attention–based Vision Transformer (ViT) and analyze the impact of global patch interactions on the resulting concept bank. Finally, we transpose the methodology to the language domain by coupling CRAFT with a BERT-type encoder, thus providing the first cross-modal study of concept factorization. Our experiments demonstrate that CRAFT remains stable across modalities, while the nature and granularity of discovered concepts depend strongly on the underlying architecture and data domain. Our code is freely available: `https://github.com/lolomasterIA/Craft.git`

1 Introduction

1.1 Post-hoc concept-based explainability: state of the art

Interpreting the decisions of deep learning models remains a key challenge. Post-hoc attribution methods are widely used to highlight input regions that influence model predictions Simonyan et al. (2014); Selvaraju et al. (2017); Sundararajan et al. (2017); Petsiuk et al. (2018). However, these methods typically reveal where the model focuses, without clarifying what abstract features or structures it actually relies on for decision-making. These limitations have been extensively discussed in recent literature, especially regarding their instability, lack of semantic content, and vulnerability to adversarial behavior Adebayo et al. (2018); Hase & Bansal (2020).

To address this, concept-based explainability has emerged as a promising alternative. Rather than attributing decisions to isolated input units, these methods aim to identify high-level, human-interpretable patterns — "concepts"— within a model's internal representations. Early works required predefined concept datasets Kim et al. (2018), while later ones introduced automatic extraction but suffered from heuristic clustering and segmentation noise Ghorbani et al. (2019). More recent methods have explored unsupervised approaches

based on matrix decomposition techniques, such as PCA, ICA, or NMF, but often lack mechanisms for quantifying the relevance of the discovered concepts or linking them back to specific parts of the input Chen et al. (2019).

CRAFT (Concept Recursive Activation Factorization) Fel et al. (2023) addresses many of these shortcomings. It extracts latent concepts using Non-Negative Matrix Factorization (NMF), ranks them by importance using Sobol indices, and optionally produces concept attribution maps through implicit differentiation. By combining unsupervised discovery, quantitative relevance, and localization, the method could be used as a general-purpose tool for post-hoc interpretability. However, although CRAFT is presented as applicable across models and modalities, it has so far only been tested on ResNet-50 models trained for image classification.

This raises the question: Does CRAFT truly generalize across architectures and data types?

1.2 Motivations

Although CRAFT is presented as broadly applicable, so far it has not been tested outside the narrow setting of CNN-based image classification. Yet today's deep learning models increasingly rely on alternative architectures, such as Vision Transformers (ViT), and are deployed across a variety of data modalities, including text.

This prompts us to investigate two key dimensions.

First, we ask whether CRAFT can still extract compact, human-interpretable concepts when applied to different semantic image categories. We also consider whether this holds when switching from convolutional networks to attention-based architectures such as ViT, where spatial hierarchies are replaced by patch-level interactions.

Second, we explore whether CRAFT can be transferred to the language domain. In this setting, the model processes sequences of token embeddings rather than pixel-based inputs.

This leads to four core research questions:

(i) Can the fundamental principle behind CRAFT generalize beyond convolutional vision models?

(ii) How do model architecture and data modality affect the interpretability and relevance of the concepts extracted?

(iii) In the case of language, can these concepts align with semantic distinctions, and reflect real decision boundaries?

(iv) Can concept extraction help explain why two classes overlap?

1.3 Contribution: cross-modal and cross-architecture evaluation of CRAFT

To address these questions, we conduct the first cross-domain evaluation of CRAFT through three experiments:

(i) We apply it to four heterogeneous visual categories using a ResNet-50 backbone to assess its robustness across varied semantic content.

(ii) We test it on the same dataset using a ViT-S/14 architecture to examine how transformer-based visual representations affect concept formation.

(iii) We transpose the method to a RoBERTa-large encoder fine-tuned on textual topic classification to explore whether interpretable concepts emerge in the language domain.

These experiments highlight how both data modality and architectural inductive biases shape the emergence, clarity, and relevance of learned concepts.

2 Overview of CRAFT

The method used by CRAFT (**C**oncept **R**ecursive **A**ctivation **F**ac**T**orization) Fel et al. (2023) seeks to translate the opaque internal activations of a deep network into a compact set of human-interpretable *"concepts"*. By combining global and local explainability, CRAFT enhances transparency, helps to understand model failures, and enforces more trust in AI systems answers.

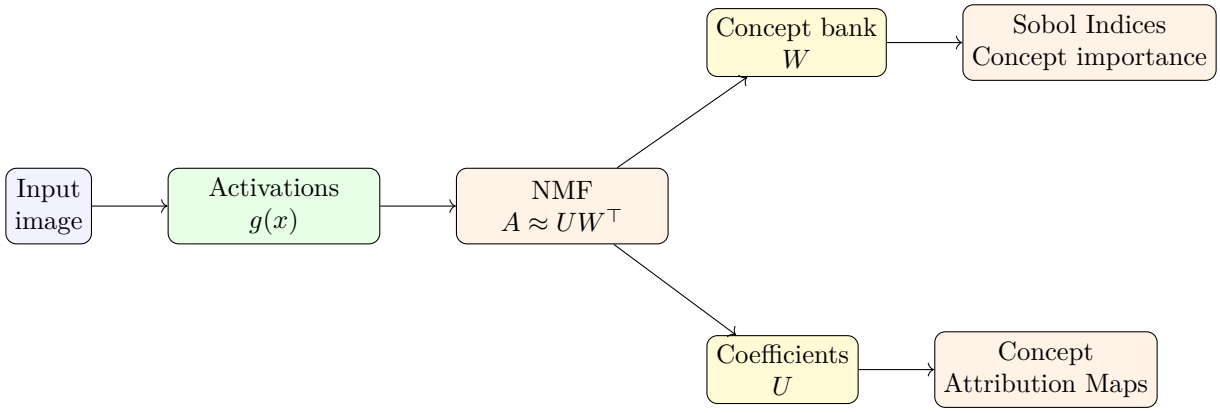


Figure 1: CRAFT overview: decomposition of internal activations into human-interpretable concepts, sensitivity analysis using Sobol indices, and spatial localization through attribution maps.

2.1 Concept Extraction via Non-Negative Matrix Factorization

Given a model $f(x) = h \circ g(x)$, where g maps inputs $x \in \mathbb{R}^d$ to intermediate non-negative activations $g(x) \in \mathbb{R}_+^p$, and h maps these activations to outputs, CRAFT applies **Non-Negative Matrix Factorization** (NMF) to the activation matrix $A \in \mathbb{R}_+^{n \times p}$ obtained from n image crops.

The NMF decomposes the positive activations A as:

$$A \approx UW^\top,$$

where $U \in \mathbb{R}_+^{n \times r}$ contains the coefficients of A, and $W \in \mathbb{R}_+^{p \times r}$ contains the Concept Activation Vectors (CAVs). This decomposition has multiple benefits: sparsity, part-based representation, and non-negativity—properties beneficial to human interpretability. Each row of W defines a direction in activation space associated with a visual pattern or structure, while rows of U provide the activation strength of each concept in a given image crop.

2.2 Recursive Concept Decomposition

A key contribution of CRAFT is its **recursive factorization strategy**, designed to resolve the semantic confusion caused by the so-called "neural collapse" of deep networks, which leads to unusable clusters. If a concept C is ambiguous or difficult to interpret, it repeatedly enhances the representation by selecting the top-k% of images that most strongly activate C, and applying NMF again to previous layers. This enables the decomposition of C into a hierarchy of sub-concepts $\{C_1, C_2, \dots\}$, revealing more refined semantic structures.

2.3 Measuring Concept Relevance with Sobol Indices

The relevance of each concept to the model's prediction is determined by CRAFT by using **Sobol indices** from global sensitivity analysis. Gradients are known to be noisy and unstable in high-dimensional spaces, so CRAFT defines the *total Sobol index* S_{T_i} of concept i as:

$$S_{T_i} = \frac{\mathbb{E}_{\mathbf{M}_{\sim i}}\left[\mathrm{Var}_{M_i}\left(h((U \circ M)W^\top) \mid \mathbf{M}_{\sim i}\right)\right]}{\mathrm{Var}(h(UW^\top))},$$

where $M \in [0,1]^r$ is a perturbation mask applied to the concept coefficients U, and $\circ$ denotes the Hadamard product. High Sobol scores suggest that modifying the concept significantly changes the model output, indicating that it has an important explanatory role in the decision process.

2.4 Localizing Concepts with Implicit Differentiation

While global concept extraction provides class-level understanding, CRAFT also enables **concept attribution maps**—local visualizations of where each concept manifests in a specific input. This is achieved by computing $\partial U / \partial x$, allowing backpropagation of concept relevance to the input space. Since U is computed via a non-differentiable optimization (NMF), CRAFT uses **implicit differentiation** of the fixed-point solution to the NMF problem, leveraging the Karush-Kuhn-Tucker (KKT) conditions to obtain the required Jacobians efficiently.

This enables compatibility with both white-box and black-box attribution methods, extending the flexibility of concept-based interpretability to a broad class of model settings.

3 Experimentation

3.1 Experiment 1: Application to different datasets

4 datasets are used to test CRAFT with ResNet50 : Siamese cats as an animal, babies as human, daisies as vegetation and car as object.

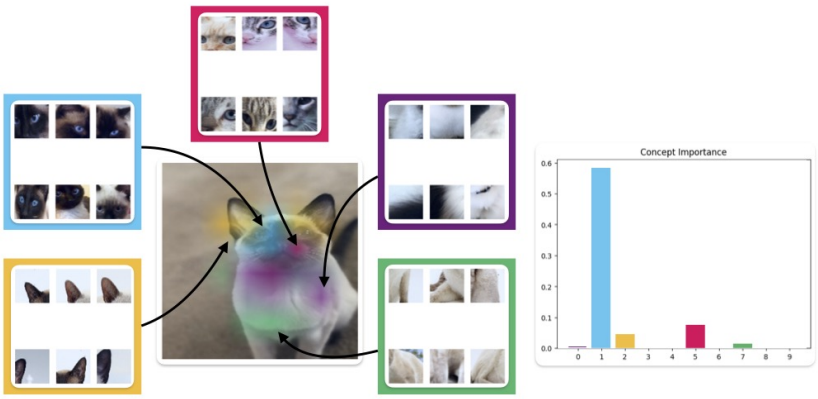


Figure 2: **Results on animal** Applying CRAFT on a Siamese's dataset shows that more than half of the extracted concepts focus on facial features (ears, muzzle, and eyes) while a minority are associated with other regions(chest).

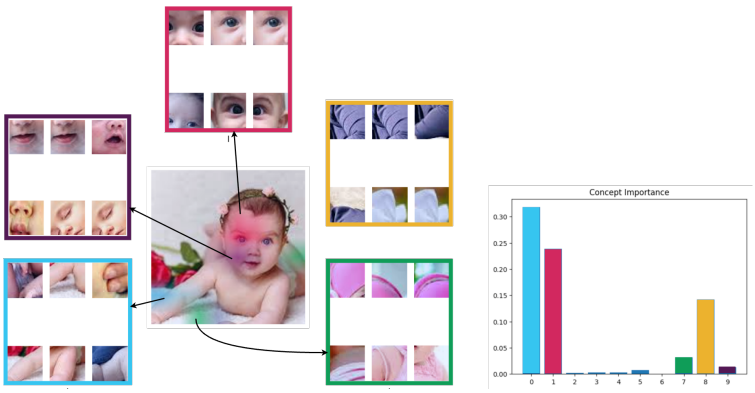


Figure 3: **Results on human** Applying CRAFT to the baby dataset reveals that the arm is identified as the most important concept, rather than the face. This unexpected result is due to the fact that the ImageNet-pretrained ResNet-50 does not include a specific class for "baby", which affects the estimation.

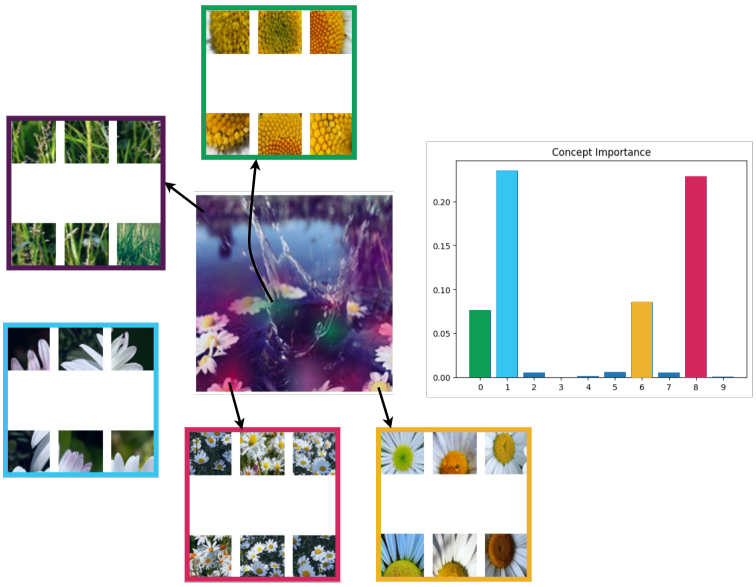


Figure 4: **Results on vegetation** Applying CRAFT to the daisy dataset reveals that the flower petals is the most important concept. There aren't any unexpected concept found.

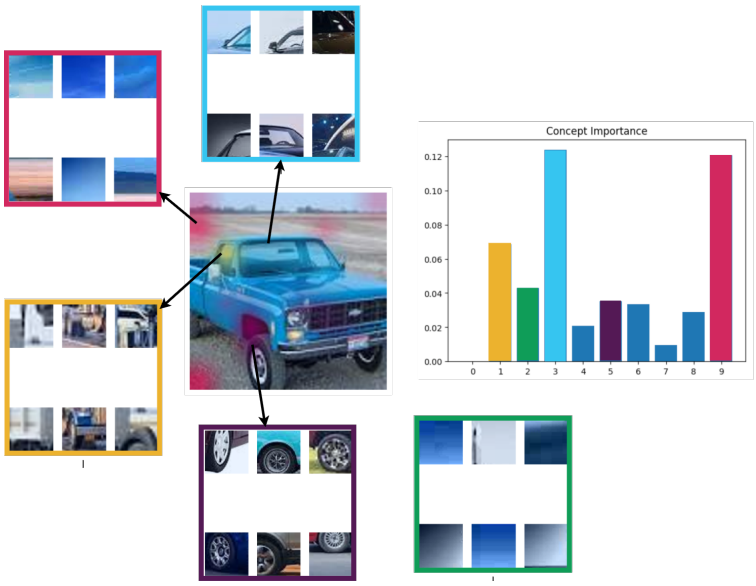


Figure 5: **Results on object** Applying CRAFT to the car dataset reveals that the sky has more importance than than the car component itself. While ResNet-50 was trained only on racing cars, applying it to diverse cars shouldn't give such unexpected output.

3.2 Experiment 2: Application with ViT-S/14 backbones

For feature extraction, we rely on **DINOv2 (ViT-S/14)** Oquab et al. (2024); the smallest model is sufficient for our work and minimizes both the memory footprint and the inference time. Each RGB image $I \in \mathbb{R}^{3\times H\times W}$ is partitioned into 14×14 patches, then linearly projected as in the original ViT encoder.

We wrap the frozen DINOv2 backbone so that it returns the normalized patch tokens $z = g(I) \in \mathbb{R}^{B\times N\times C}$ (with N patches and C = 384 channels out of the DINOv2 backbone), as these tensors are more appropriate for the concept of factorization **g(.)** than the embedding of global class tokens. A lightweight linear layer **h(z)** maps the average pooled representation to class logits.

In the explanatory framework, therefore, we set **g(.)** *"input-to-latent"* function and **h(z)** *"latent-to-logit"* function. This modular design isolates vision feature learning from task-specific decision making, while ensuring that the latent representation remains positive (ReLU applied post-token reshaping) and thus compatible with CRAFT's non-negative matrix factorization.

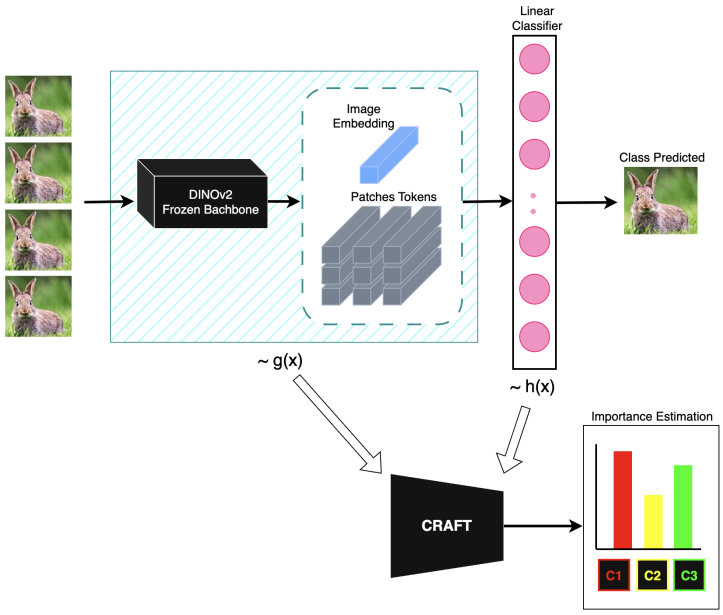


Figure 6: **Overview of DINOv2 applications.** Starting from a set of images, we use a DINOv2 frozen backbone to extract image embeddings and patch tokens, and then create a classification head to predict the image class. At the input of CRAFT, the DINOv2 frozen backbone and its patch tokens are composed as g(.), and the classification layer is now an h(.) function. We estimate the importance of concepts by using Sobol indices.

The input dataset consists of a set of rabbit images with dimensions [B, C, H, W], where B represents the total number of images, C = 3 for RGB, and H and W denote the height and width, respectively. This is the same test dataset with ResNet50, as announced in the article. CRAFT successfully recognizes the main concepts in the image; however, the closer the patch sizes are to DINOv2's patch sizes, the less significant the concepts become.

We also performed a patch-size sensitivity test with the DINOv2 model. The test is intended to verify the role and relationship between craft. Fit and the model used (DINOv2). DINOv2 declares a fixed patch size of 14, but this size makes the crops very fuzzy, and the concepts are no longer psychologically correct. Depending on the patch size indicated (different or larger than that of the model), CRAFT cuts the input dataset of size [B, C, H, W] into [B x L, C, P, P] where L is the number of patches per image and P is the patch size indicated Fig.8. In this step, each crop are classified by DINOv2 backbone.

Limitations. CRAFT can be applied to a ViT model, but it is necessary to organize the pre-processing of the inputs and the two functions, g and h. ViT models attempt to cut images into smaller patches to optimize task performance; however, this approach makes CRAFT confusing.

Figure 7: **Qualitative Results.** CRAFT results on rabbit class of DINOv2 pre-trained with ImageNet 1k with classification head. The results showcase the two most important concepts related to rabbits. The following three concepts from the top 5 raise the classes of vegetation rather than the rabbit.

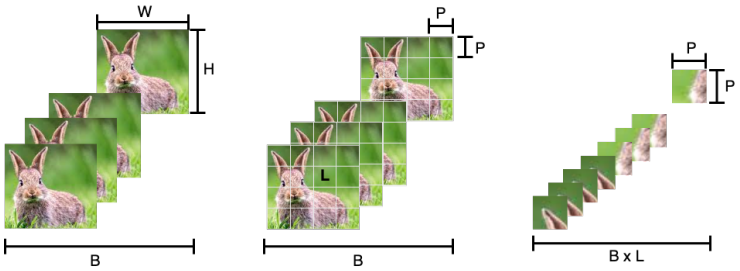


Figure 8: Scheme of patch-size processing by CRAFT

3.3 Experiment 3: Adaptation to BERT

In this experiment, we adapted CRAFT to a text classification task based on a fine-tuned RoBERTa-large model. The model was trained on quotes extracted from product reviews in the cosmetics domain, annotated with 22 marketing topics. While the overall classification accuracy reached approximately 80%, a detailed inspection of the per-class F1-scores revealed a wide variance—from as low as 40% to as high as 95%. A confusion matrix highlighted a strong overlap between certain classes, motivating a deeper analysis via concept attribution.

The original CRAFT implementation targets image explainability (patch extraction, heavy GPU batches). For large-scale text classification we faced two blockers:

- Manage texts sequences instead of images

- Batch et chunk processes to preserve cpu ram

The python class craft torch.py from the original paper has been modified in this way.

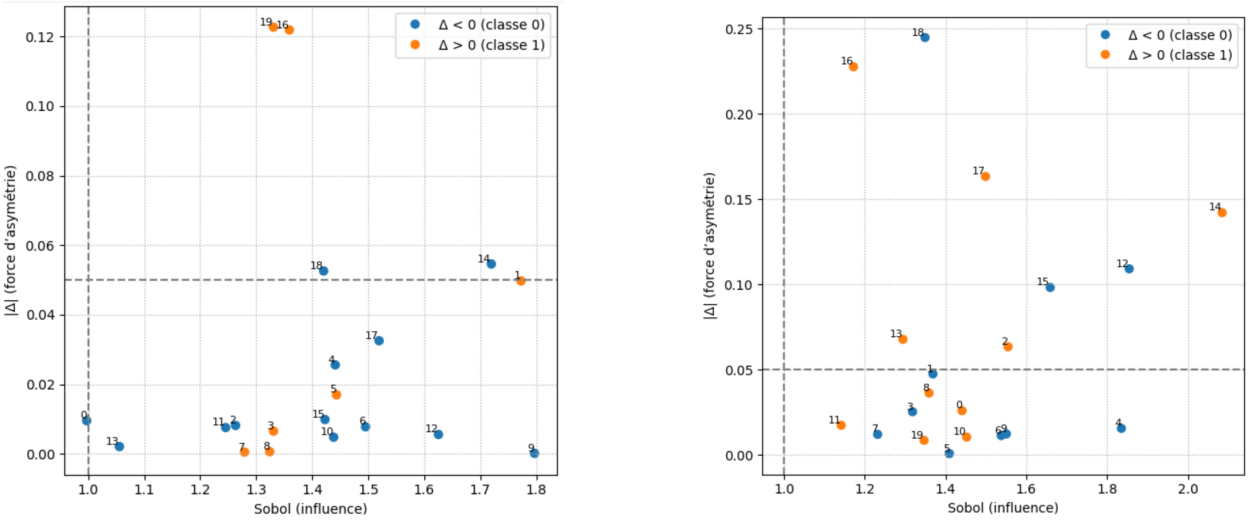
We selected two pairs of classes for this experiment: one pair known to overlap significantly (e.g., *presentation* vs. *design*) and another pair with distinct semantics (e.g., *price* vs. *ingredients*). CRAFT was applied using the following adaptations:

- The concept extractor $g(x)$ was instantiated by pooling the last hidden states of RoBERTa-large (dimension 1024). We added a ReLu function to keep only positive values (mandatory for MNF)
- Sobol-based sampling was used over the projected latent representations to estimate the contribution of concepts.

We use two metrics to evaluate the overlaping between two topics:

- Delta: “Is the concept k more present in one class than in the other?" Average difference in concept activation between classes
- Sobol indice: “If I vary the concept k, does the algorithm really change its prediction?” Sensitivity (explained variance) of the model output to the concept

Situation	$\lvert\boldsymbol{\Delta_k}\rvert$	$\text{Sobol}_{\boldsymbol{k}}$	**Lecture pratique**
specific concept for a classe	$>$ threshold	high	Powerful discriminating signal
Shared but used concept	≈ 0	high	overlap (be careful, non-linear lever)
discriminating concept but underutilized	$>$ threshold	low	Underutilized information
Anecdotal concept (noise)	≈ 0	low	Negligible impact

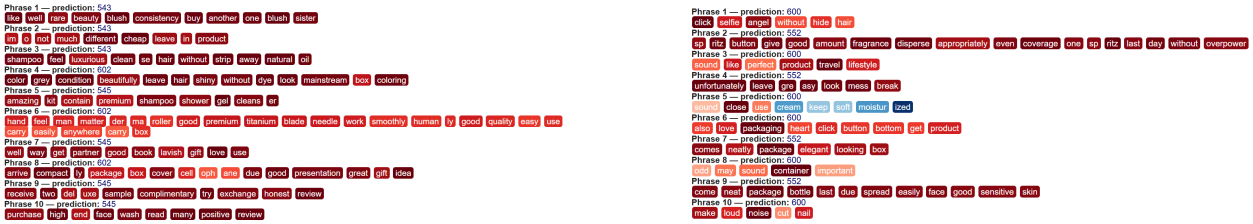


(a) Overlaped classes

(b) Separated classes

Figure 9: Comparison of concept influence (Sobol index) and class asymmetry ($|\Delta|$) for two classification cases.

On the overlaped side, only concepts 16 and 19 have a delta $>$ 0.6 and 1.2, but with a Sobol score between 1.3 and 1.4. Concept 9 has weight but does not help distinguish the two classes (delta 0). Overall, the Sobol indices and deltas are low. This explains the classifier’s difficulty in distinguishing the two classes. On the well-separated side, the values are much more differentiated. Indeed, concepts 16 and 18 alone allow the prediction to be tilted from one class to the other (delta $>$ 20 and sobol between 1.1 and 1.4). We also see concept 14: sobol $>$ 2 and delta 1.5.



(a) Overlaped classes

(b) Separated classes

Figure 10

Which concepts activate which tokens? The figure 7 shows hows two representative attribution maps for overlapping and non-overlapping class predictions. In the overlapping classes, concept activations shared significant lexical triggers (e.g., *"bottle", "design", "look"*), making class separation harder. In contrast, the well-separated classes showed minimal overlap in concept activations. For example, we see that concept 18 activates words in texts of class 600 and remains very weak for those of class 552 as also shown in the graph in figure 6.

Limitations et perspectives One critical limitation observed is the opacity of latent concepts. CRAFT concepts are latent vectors; they capture co-activations of tokens but do not preserve order or syntax. With only K = 20 concepts and without hierarchy (no recursive layer-by-layer extraction), we necessarily undersample the richness of a RoBERTa encoding (D = 1024): several linguistic phenomena are then mixed in the same concept.While CRAFT successfully identifies high-variance directions in representation space, these do not always translate to interpretable word groups. Moreover, the projection from concept space back to token space is sensitive to sentence structure and may vary across instances. As a result, human understanding of concepts is still limited, and manual intervention is required to relate latent dimensions to linguistic features.

4 Conclusion and Discussion

This study demonstrates that CRAFT-style factorization remains applicable to transformer architectures in both vision and language; it yields compact, class-sensitive concept banks and delivers quantitative insight through Sobol importance scores. Yet several methodological limits emerged.

Limitation of concept In the case of vision models, the lack of context causes CRAFT to define incorrect concepts in some cases (e.g., a rabbit in the forest or detected objects scattered throughout the image). Whereas for language models, the granularity of the concepts found is too low compared to the model's prediction.

Human interpretability. Although concept activation maxima reveal coherent clusters, a concept rarely maps to a single natural-language phrase. Manual labeling or weak supervision could help anchor latent factors to semantic descriptors.

Nature of representations. CRAFT builds concepts from latent vectors that reflect patterns of co-activation in the model, but these vectors do not preserve the order, structure, or meaning present in the input. In language models especially, using a fixed number of flat concepts without hierarchy leads to mixed and hard-to-interpret abstractions.

In summary, our results provide clear answers to the questions that guided this study: CRAFT can be extended across architectures and modalities, and its core mechanism remains effective in various settings. However they also raise a deeper issue: how much can we trust post-hoc methods like CRAFT to truly explain model behavior? While such techniques can be helpful to summarize what a model reacts to, they may also give a false sense of understanding.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, 2018. URL https://arxiv.org/abs/1810.03292.

Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition, 2019. URL https://arxiv.org/abs/1806.10574.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability, 2023. URL https://arxiv.org/abs/2211.10154.

Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations (ace), 2019. URL https://arxiv.org/abs/1902.03129.

Peter Hase and Mohit Bansal. Evaluating explainable ai: Which explanations help users predict model behavior?, 2020. URL https://arxiv.org/abs/2005.01831.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018. URL https://arxiv.org/abs/1711.11279.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018. URL https://arxiv.org/abs/1806.07421.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization, 2017. URL https://arxiv.org/abs/1610.02391.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.