CRAFT: Concept Recursive Activation FacTorization for Explainability

Thomas Fel[1,3,5*] **Agustin Picard**[3,6*] **Louis Bethune**[3*] **Thibaut Boissin**[3,4*]
David Vigouroux[3,4] **Julien Colin**[1,3] **Rémi Cadène**[1,2] **Thomas Serre**[1,3]

[1]Carney Institute for Brain Science, Brown University, USA [2]Sorbonne Université, CNRS, France
[3]Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France
[4] Institut de Recherche Technologique Saint-Exupery, France
[5] Innovation & Research Division, SNCF , [6] Scalian

Figure 1. **CRAFT results for the prediction "chain saw".** First, our method uses Non-Negative Matrix Factorization (NMF) to extract the most relevant concepts used by the network (ResNet50V2) from the train set (ILSVRC2012 [3]). The global influence of these concepts on the predictions is then measured using Sobol indices (right panel). Finally, the method provides local explanations through *concept attribution maps* (heatmaps associated with a concept, and computed using grad-CAM by backpropagating through the NMF concept values with implicit differentiation). Besides, concepts can be interpreted by looking at crops that maximize the NMF coefficients. For the class "chain saw", the detected concepts seem to be: • the chainsaw engine, • the saw blade, • the human head, • the vegetation, • the jeans and • the tree trunk.

Abstract

*Attribution methods, which employ heatmaps to identify the most influential regions of an image that impact model decisions, have gained widespread popularity as a type of explainability method. However, recent research has exposed the limited practical value of these methods, attributed in part to their narrow focus on the most prominent regions of an image – revealing "where" the model looks, but failing to elucidate "what" the model sees in those areas. In this work, we try to fill in this gap with CRAFT – a novel approach to identify both "what" and "where" by generating concept-based explanations. We introduce 3 new ingredients to the automatic concept extraction literature: **(i)** a recursive strategy to detect and decompose concepts across layers, **(ii)** a novel method for a more faithful estimation of concept importance using Sobol indices, and **(iii)** the use of implicit differentiation to unlock Concept Attribution Maps.*

1. Introduction

Interpreting the decisions of modern machine learning models such as neural networks remains a major challenge. Given the ever-increasing range of machine learning applications, the need for robust and reliable explainability methods continues to grow [4, 10]. Recently enacted European laws (including the General Data Protection Regulation (GDPR) [11] and the European AI act [14]) require the assessment of explainable decisions, especially those made by algorithms.

In order to try to meet this growing need, an array of explainability methods have already been proposed [5, 17, 18, 20, 21, 25–28]. One of the main class of methods called attribution methods yields heatmaps that indicate the importance of individual pixels for driving a model's decision. However, these methods exhibit critical limitations [1, 8, 22, 24], as they have been shown to fail – or only marginally help – in recent human-centered benchmarks [2, 7, 13, 16, 19, 23]. It has been suggested that their limitations stem from the fact that they are only capable of explaining *where* in an image are the pixels that are critical to the decision but they cannot tell *what* visual features are actually driving decisions at these locations. In other words, they show where the model looks but not what it sees.

A recent approach has sought to move past attribution methods [12] by using so-called "concepts" to communicate information to users on how a model works. The goal is to find human-interpretable concepts in the activation space

* Equal contribution

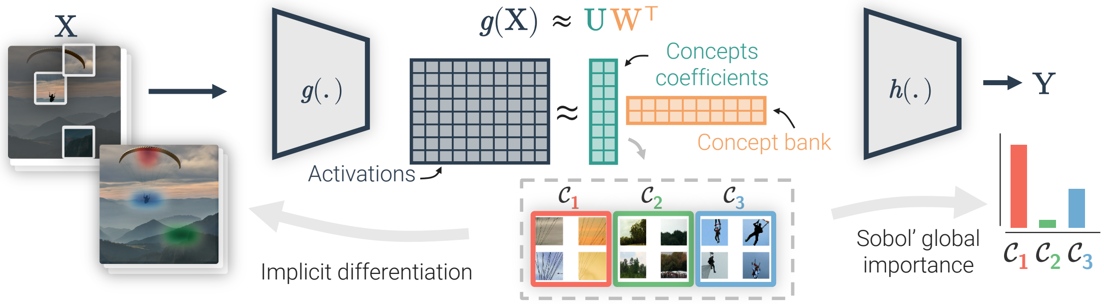

Figure 2. **Overview of CRAFT.** Starting from a set of crops $\mathbf{X}$ containing a concept $\mathcal{C}$ (e.g., crops images of the class "parachute"), we compute activations $\boldsymbol{g}(\mathbf{X})$ corresponding to an intermediate layer from a neural network for random image crops. We then factorize these activations into two lower-rank matrices, $(\mathbf{U}, \mathbf{W})$. $\mathbf{W}$ is what we call a "concept bank" and is a new basis used to express the activations, while $\mathbf{U}$ corresponds to the corresponding coefficients in this new basis. We then extend the method with 3 new ingredients: (1) recursivity – by proposing to re-decompose a concept (e.g., take a new set of images containing $\mathcal{C}_1$) at an earlier layer, (2) a better importance estimation using Sobol indices and (3) an approach to leverage implicit differentiation to generate *concept attribution maps* to localize concepts in an image.

of a neural network. Although the approach exhibited potential, its practicality is significantly restricted due to the need for prior knowledge of pertinent concepts in its original formulation and, more critically, the requirement for a labeled dataset of such concepts. Several lines of work have focused on trying to automate the concept discovery process based only on the training dataset and without explicit human supervision. The most prominent of these techniques, ACE [6], uses a combination of segmentation and clustering techniques but requires heuristics to remove outliers. However, ACE provides a proof of concept that it might be possible to discover concepts automatically and at scale – without additional labeling or human supervision. Nevertheless, the approach suffers several limitations: by construction, each image segment can only belong to a single cluster, a layer has to be selected by the user to be used to retrieve the relevant concepts, and the amount of information lost during the outlier rejection phase can be a cause of concern. More recently, Zhang et al. [29] proposes to leverage matrix decompositions on internal feature maps to discover concepts.

Here, we try to fill these gaps with a novel method called CRAFT which uses Non-Negative Matrix Factorization (NMF) [15] for concept discovery. In contrast to other concept-based explanation methods, our approach provides an explicit link between their global and local explanations (Fig. 1) and identifies the relevant layer(s) to use to represent individual concepts.

2. CRAFT

As described in Fig. 2, using CRAFT, we first perform a stage of unsupervised concept discovery by taking crops of images from the class we wish to explain, and decompose their (non-negative) activations at an intermediate layer l into two matrices $\mathbf{W}$ and $\mathbf{U}$ containing a "concept bank" and their corresponding coefficients. Once this $\mathbf{W}$ has been computed, we can apply our three ingredients to enhance our explanations:

Recursivity: By recursively exploring shallower layers of the neural network, we are able to find sub-concepts that integrate into more complex and abstract super-concepts. This can vastly improve the human understandability of the concepts extracted in the model's last layers.

Sobol' importance scores: Using Sobol' indices, we measure the global importance of each concept for the prediction of the class we wish to explain. This enables us to better understand the model's decision strategies in a per-class and per-instance basis.

Implicit differentiation: This mathematical tool can be used to estimate gradients when they are too expensive to compute via simple auto-differentiation. We propose to leverage it to unlock **Concept Activation Maps**, thus allowing us to locate individual concepts in the images we wish to explain.

3. Results

We used CRAFT to explain a ResNet50V2 [9] trained on the ILSVRC2012 [3] dataset (ImageNet), and we set up a website where we showcase our results on all the 1000 classes. In it, we discover 25 concepts for each class without any supervision, we compute their global importance, we plot the crops that activate each of them the most and we find the concepts of other classes that are the most similar to each of them. Additionally, we have integrated a novel feature visualization technique that enhances our explanation's interpretability, and applied it to each of the 25000 extracted concepts. We invite everyone to explore all these results in our Lens.

In conclusion, understanding the inner workings of the elusive ResNet has never been closer.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 1

[2] Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2

[4] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *ArXiv e-print*, 2017. 1

[5] Thomas Fel, Rémi Cadène, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[6] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[7] Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Short Papers)*, 2020. 1

[8] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 2

[10] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021. 1

[11] Margot E Kaminski and Jennifer M Urban. The right to contest ai. *Columbia Law Review*, 121(7):1957–2048, 2021. 1

[12] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 1

[13] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. HIVE: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[14] Mauritz Kop. Eu artificial intelligence act: The european approach to ai. In *Stanford - Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust and IPR Developments, Stanford University, Issue No. 2/2021. https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/*. Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust , 2021. 1

[15] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 2

[16] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[17] Paul Novello, Thomas Fel, and David Vigouroux. Making sense of dependence: Efficient black-box explanations using dependence measure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[18] 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. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[19] Hua Shen and Ting-Hao Huang. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 168–172, 2020. 1

[20] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 1

[21] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014. 1

[22] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 1

[23] Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. Do users benefit from interpretable vision? a user study, baseline, and dataset. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 1

[24] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. 1

[25] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 1

[26] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 1

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1

[28] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1

[29] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. *arXiv preprint arXiv:2006.15417*, 2020. 2