LRP-CLIP: A Zero Shot Approach for the Explanation of the Cognitive Functions of Vision Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Cognitive explanation of deep neural networks is enabled by the use of Layer-wise Relevance Propagation (LRP) heatmaps. Heatmaps indicate where a region of high significance is, but current approaches rely on further computational training overhead to assign semantic meaning. We propose LRP-CLIP, a zero-shot framework that grounds these heatmaps in natural language using pretrained CLIP (Contrastive Language-Image Pre-Training) and thereby allowing cognitive explanations of deep learning models without further training. Our method extracts relevance-based image crops and matches them against domain-specific textual attributes, producing human-readable explanations without any additional training or supervision. Applied to bird classification with a Vision Transformer (ViT), we find that the model relies heavily on head features, while occasionally exploiting spurious background cues. This dual pattern reflects both systematic heuristics and potential biases in the model's cognitive strategy. Quantitative evaluation shows that our pipeline reliably localizes annotated parts and assigns semantically valid labels, clearly outperforming random baselines. These results demonstrate an extensible zero-shot approach to model cognition.

1 Motivation

Deep learning models achieve impressive results, but their decision-making remains opaque. These black-box systems can make accurate predictions, yet offer little insight into why they arrive at a certain conclusion. Layer-wise Relevance Propagation (LRP) highlights which parts of the input contribute most to the output, but the resulting maps are often abstract: they show where the model focuses, not what it sees or why those regions matter. To make these explanations more meaningful, we propose using external knowledge to give structure to the spatial relevance information. Our approach consists two steps: (1) decompose relevance hotspots into distinct sub-regions, and (2) reclassify these components and translate them into natural language, so a reader understands both what the model is looking at and its semantic meaning. Building on this idea, we propose LRP-CLIP, a novel framework that extends LRP-based explanations by semantically interpreting the highlighted sub-regions using CLIP (Contrastive Language-Image Pre-Training). CLIP aligns images and text in a shared embedding space for zero-shot classification. We extract relevant input regions using LRP and then classify these regions with CLIP, using domain-specific textual labels and attributes. This approach produces human-readable explanations without requiring specialized expert knowledge to interpret LRP maps or additional training, since it leverages CLIP directly. This approach bridges low-level model reasoning with natural language, offering a modular pipeline for explainability. We demonstrate the method's feasibility on the CUB-200 bird classification dataset Wah et al. [2011] using a Vision Transformer (ViT) and discuss its strengths, limitations, and future applicability to other domains, including time-series data.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Related Work

The complexity of modern models, particularly Transformers, has amplified interest in explainable AI (XAI), which seeks to make decisions more transparent and trustworthy, especially in high-stakes domains.

Explainability approaches fall into three categories: data-level, model-level, and post-hoc Sun et al. [2024]. Data-level methods jointly model multiple modalities to produce both predictions and explanations. Model-level methods use inherently interpretable architectures. Post-hoc methods, in contrast, analyze a model's outputs in relation to its inputs, independent of training.

Post-hoc methods include signal-based techniques such as LRP Achtibat et al. [2024] and Grad-CAM Selvaraju et al. [2019], surrogate models such as LIME Ribeiro et al. [2016], and hybrids that combine them Dhore et al. [2024].

Yet even high-quality relevance maps often remain inaccessible to non-experts Bove et al. [2022], highlighting the semantic gap between what the model and what humans interpret.

Recent work has turned to zero-shot models such as CLIP Radford et al. [2021], which align vision and language for label-free classification using natural language prompts.

Some studies use CLIP directly for explanations, generating zero-shot semantic descriptions Sammani and Deligiannis [2024].

We propose a pipeline that integrates LRP for visual relevance extraction with CLIP for semantic labeling of regions, bridging the gap between faithful explanations and human understanding without retraining and limited architecture complexity.

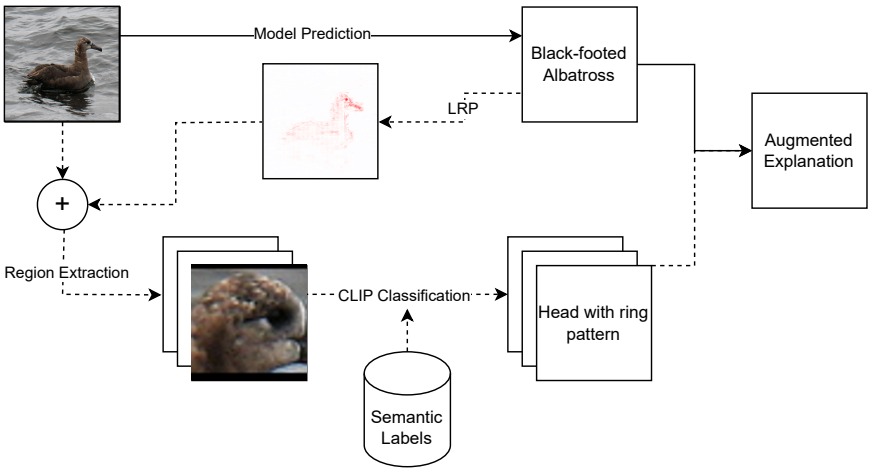

Figure 1: The general pipeline flow design of LRP-CLIP.

3 Architecture of the Pipeline

We propose a model agnostic modular architecture for generating post-hoc, human-interpretable explanations of classification decisions on image data (see Figure 1). The system consists of two main components: the Model Prediction Module and the Explanation Module. The Prediction Module carries out the core classification task. It receives an image as input and outputs a predicted class label. In Figure 1, the Prediction Module is depicted in the top part, where the image-based input is fed into a model that performs a prediction. In this example, the output is the class label *Black-footed Albatross*.The Explanation Module operates after a prediction has been made for a given input. In our approach, we use LRP to generate relevance heatmaps. These heatmaps, in combination with the original input image, are used to extract the most relevant regions. As a result, we obtain one or more image regions (from 1 to N), each of which is passed to CLIP, a vision-language model that has been primed with external knowledge (i.e. Semantic Labels). In the example shown in Figure 1, the external knowledge consists of general information about bird species.

Implementation

We fine-tuned a ViT-L (ImageNet-1000 pretrained) on the CUB-200 dataset, reaching 87.2% validation accuracy. The Explanation Module is implemented using the following methods:
The projection from raw input to relevant information is performed using LRP. We used LRP with ViT-specific settings Achtibat et al. [2024], enforcing positive relevance for more focused heatmaps Anders et al. [2021]. Regions of highest relevance were extracted, merged if adjacent, or split if complex (e.g., head vs. eye) based on high relevance regions, producing the largest meaningful crops. These were ordered by relevance and passed to CLIP. We used CLIP ViT-H/14 (LAION-2B pretrained) for zero-shot classification, comparing crops against CUB-200 attribute labels in text space Schuhmann et al. [2022], Radford et al. [2021], Ilharco et al. [2021].

4 Evaluation

We evaluate on the CUB-200 dataset Wah et al. [2011], which includes 12k images across 200 bird species with fine-grained annotations (labels, parts, and attributes). These detailed annotations make it well-suited for assessing both prediction accuracy and the semantic quality of explanations.

Findings

Figure 2: The three largest relevance-based crops are shown. Classifications include a rusty blackbird, a black-footed albatross, and a misclassification of an Acadian Flycatcher (predicted as Yeast Flycatcher, with Acadian Flycatcher second). The top CLIP labels of the nine crops are: (1) black forehead, (2) black back, (3) gray legs, (4) white forehead, (5) hooked bill, buff-colored, similar in size to head, (6) bird of medium (9–16 in) size, (7) rufous forehead, (8) broad buff-colored wings, (9) buff-colored eye.

Exemplary outputs of the pipeline (Fig. 2) show that the ViT consistently identifies the head as the most relevant region, while other highlighted areas vary by image (e.g., body, wings, or legs). CLIP generally assigns meaningful labels to these regions, though it occasionally misidentifies fine-grained attributes such as color.
One example (black-footed albatross) illustrates a limitation: highly relevant regions sometimes correspond to background patches, leading to less informative labels. In another case (Acadian Flycatcher), the model misclassifies the species but still provides interpretable relevance maps and semantically useful labels. These findings suggest that the pipeline can enhance model interpretability, even in failure cases, and may support model refinement.

Quantitative Evaluation

We evaluate whether (1) generated crops overlap annotated body parts and (2) CLIP labels match the correct part. Part annotations were mapped to CUB200 attributes (e.g., bill→beak).

Table 1: Evaluation of part matching performance.

Metric	LRP-CLIP			Random Box		
	TOP1	TOP3	TOP5	TOP1	TOP3	TOP5
Body part present in crop (%)	93.33	93.33	93.33	35.61	35.61	35.61
At least one correct CLIP label (%)	49.15	62.36	68.47	9.58	14.00	16.71

Table 1 shows that 93% of crops contain annotated parts and nearly 70% yield a correct CLIP label (top-5). Both metrics outperform a random baseline, supporting the pipeline's ability to generate semantically valid part-level explanations.

Table 2: Evaluation of part label heuristics.

Metric	LRP-CLIP			Random Box		
Top three matched labels in all parts (%)	Eye	Forehead	Breast	Back	Leg	Breast
	23.23	21.76	11.47	18.92	16.9	15
Least three matched labels in all parts(%)	Crown	Belly	Tail	Tail	Belly	Crown
	3.58	1.27	0.4	1.94	1.75	1.64

Table 2 shows the top three and least three distribution of the matched labels in all parts of all correctly labeled crops. It shows a clear bias towards the head by LRP-CLIP, while random box focuses on general big parts of the anatomy of a bird. The least three matched labels are the same between LRP-CLIP and random box. This either indicates a lack of understanding for CLIP for these specific labels or the little prevalence of these parts in bird pictures.

Discussion

Explanation Module Components LRP highlights regions that often correspond to bird parts, but cropping involves trade-offs: small crops may lose context, while large ones dilute relevance. Future work may explore rules that also capture negative relevance.
Labeling with CLIP CLIP labeling depends on crop quality and label sets. Using CUB-200 attributes limits coverage (e.g., background context is excluded), and small or ambiguous crops often lead to guesses. Nonetheless, CLIP maps abstract regions to interpretable part labels without retraining, making it a lightweight semantic layer. Future work could expand label sets or preserve more image context for robustness.
Efficiency The pipeline adds little overhead: one forward pass, one backward pass, and CLIP embedding of crops, with labels pre-embedded for efficiency.

5 Future Work

This paper presents only first results toward enhancing LRP explanations with natural language semantics. The current study is limited to a single dataset (CUB-200), one explanation method (LRP), and a basic region extraction strategy. These choices allowed us to demonstrate feasibility with minimal additional training effort, but they restrict generality.
Future work will extend the pipeline in several directions: adopting more sophisticated cropping algorithms, comparing multiple explanation methods, and evaluating on additional datasets. In the longer term, we aim to adapt the approach outside of vision domains like medical time-series classification, where interpretability is particularly challenging yet essential for trust and adoption.

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 135–168. PMLR, 21–27 Jul 2024.

Christopher J. Anders, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Software for dataset-wide xai: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy. *CoRR*, abs/2106.13200, 2021.

Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 807–819, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391443. doi: 10.1145/3490099.3511139. URL https://doi.org/10.1145/3490099.3511139.

Vaibhav Dhore, Achintya Bhat, Viraj Nerlekar, Kashyap Chavhan, and Aniket Umare. Enhancing explainable ai: A hybrid approach combining gradcam and lrp for cnn interpretability, 2024. URL https://arxiv.org/abs/2405.12175.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

Fawaz Sammani and Nikos Deligiannis. Interpreting and analysing clip's zero-shot image classification via mutual knowledge, 2024. URL https://arxiv.org/abs/2410.13016.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

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. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Shilin Sun, Wenbin An, Feng Tian, Fang Nan, Qidong Liu, Jun Liu, Nazaraf Shah, and Ping Chen. A review of multimodal explainable artificial intelligence: Past, present and future, 2024. URL https://arxiv.org/abs/2412.14056.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Juli 2011.