UAB
Universitat Autònoma
de Barcelona

Interpretability Analysis in the Classification of Nature-Related Images in Social Media Videos

Bachelor's Thesis (TFG)

Author: Paula Feliu Criado 1630423

Profesors: Johannes Langemeyer & Ramin Solyemani

Date: March 23, 2025

Contents

Abstract

Abstract

1 Introduction and Context

The increasing urbanization and digitalization of society have significantly reduced people's direct interactions with nature, a trend linked to a growing disconnection from both the environment and broader community values (Kesebir & Kesebir, 2017; Soga & Gaston, 2016). This phenomenon, known as the "extinction of experience," describes a self-perpetuating cycle where diminished exposure to natural environments across generations weakens environmental familiarity, appreciation, and engagement (Soga & Gaston, 2016). Psychological studies confirm that reduced contact with nature not only diminishes environmental concern but also shifts individual values toward self-centered aspirations, consequently impacting generosity, community involvement, and overall social connectedness (Weinstein, Przybylski, & Ryan, 2009).

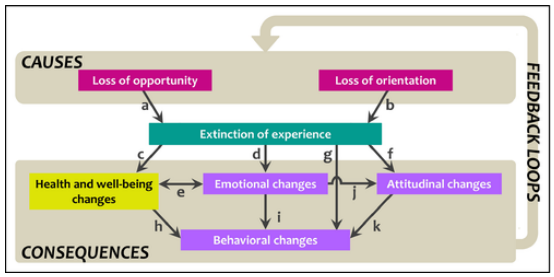

Figure 1: "Extinction of Experience"

Figure 1 illustrates the conceptual underpinnings of this extinction-of-experience process. They show how climate change, urbanization, or other environmental stressors (e.g., increased screen time, fewer outdoor opportunities) can drive a feedback loop that gradually erodes people's capabilities, opportunities, and motivations to engage with nature. Over time, these feedback loops perpetuate the cycle of disconnection, further reducing environmental concern and stewardship.

Psychological studies confirm that reduced contact with nature not only diminishes environmental concern but also shifts individual values toward self-centered aspirations, consequently impacting generosity, community involvement, and overall social connectedness (Weinstein, Przybylski, & Ryan, 2009). As we can see in the image below, in Figure 2, the first and second charts compare rates of children's outdoor activities at different points in time and highlights a marked decline in participation and time spent outside. In a related vein,the last chart demonstrates the downward trend in per capita visits to natural areas (e.g., national parks), underscoring the broader societal shift away from regular, meaningful nature experiences.

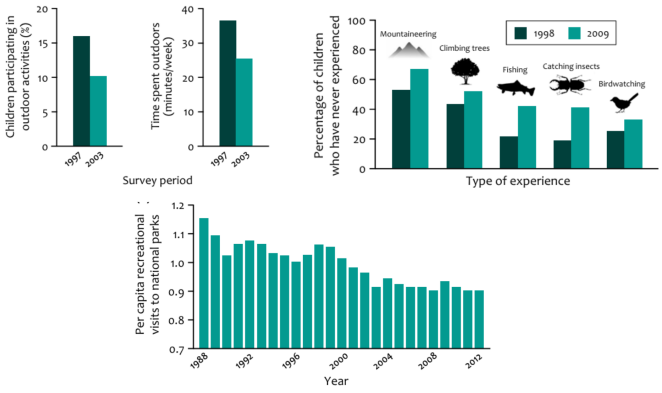

Figure 2: Statistics about the exposure to nature

In response to this challenge, the BIG-5 project investigates how digital platforms—despite their physical separation from nature—might foster meaningful human–nature connections through virtual interactions. Central to this inquiry is the concept of *Digital Relational Values* (DRV), defined as nature-related values emerging within virtual communities through indirect experiences of the natural world (Barcelona Supercomputing Center, 2021). Recent research suggests that such virtual interactions can compensate for physical disconnection by enhancing environmental appreciation, care, and stewardship, potentially spurring real-world conservation behaviors (Langemeyer & Calcagni, 2022). To systematically explore and quantify these values, the project employs artificial intelligence (AI) methodologies to process large volumes of social media images and automatically identify nature-related elements, human–nature interactions, and emotional contexts (Väisänen, Heikinheimo, Hiippala, & Toivonen, 2021).

However, the successful application of AI methods depends not only on accurate image classification but also on model interpretability. Techniques such as Class Activation Mapping (CAM), Grad-CAM, SHAP, and LIME will be used to uncover which visual cues drive AI decisions, thus ensuring transparency and reliability in the classification process. By aligning these automated insights with human perceptions, the project aims to validate and refine its analyses, ultimately bridging the gap between digital and real-world nature experiences.

2 Approach

This project is divided into two main phases, each playing a crucial role in achieving the overall goal of understanding and explaining the decision-making process of deep learning models in the context of nature-related image classification.

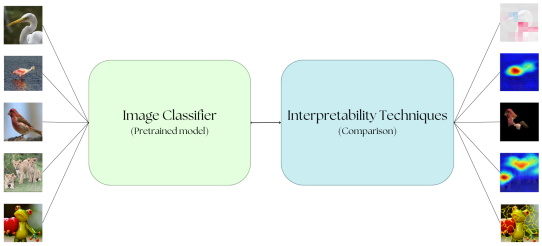

Figure 3: Approach

The first phase focuses on the development of an image classifier based on deep learning techniques. This classifier will categorize nature-related images into predefined classes, and its outputs will serve as the foundation for the second phase of the project: the comparison of interpretability techniques. The goal of this phase is to assess how well different interpretability methods can explain the decisions made by the model, particularly in terms of which image regions are most relevant for classification.

Image Classification System

The image classifier will categorize images into four distinct classes, with each class structured to ensure mutual exclusivity within its category. The four categories are:

1. **Nature vs. Non-Nature**: Classifying images as either representing nature or not.
2. **Biotic vs. Abiotic**: Identifying whether the image contains biotic (living organisms) or abiotic (non-living) elements.
3. **Material vs. Immaterial**: Differentiating between physical (material) and abstract (immaterial) aspects of the content.
4. **Landscape Type**: Classifying the image into one of five predefined landscape types: *artificial surfaces, agricultural areas, forests and semi-natural areas, wetlands, water bodies or None of them.*

Each image will be assigned one and only one label per class, ensuring exclusivity within each category. This approach follows a multiclass classification structure, where the classes are mutually exclusive, but independent of one another.

To develop the image classifier, deep learning techniques will be employed, particularly Convolutional Neural Networks (CNNs), which are highly effective for image classification tasks. The model will be trained to predict the most likely label for each of the four categories based on the visual content of the images. This phase will focus on building a robust classifier that can accurately assign each image to its appropriate category.

The first phase of the project, the development of the classifier, is crucial for answering the following key questions:

1. What regions of the image does the model consider most relevant for classification?
2. How do these regions vary across different interpretability techniques?
3. How does the model's classification criteria compare to human classification criteria?

Interpretability Techniques

The second phase of the project will focus on addressing these questions. To achieve this, the trained classifier will be used as a foundation, and various interpretability techniques will be applied to analyze the model's decision-making process. Techniques such as Class Activation Mapping (CAM), Grad-CAM, SHAP, and LIME will be used to highlight the regions of the images that are most influential in the classifier's predictions.

The goal is to evaluate how effectively these interpretability methods can explain the model's outputs and how well they align with human classification criteria. By comparing the results across these techniques, we aim to identify which methods provide the most transparent insights into the model's behavior.

An experimental approach will be followed to test and evaluate both the classifiers and the interpretability techniques using a dataset of images from social media. This comparison will help determine the strengths and limitations of each method and offer a deeper understanding of how well deep learning models align with human-driven classification.

3 State of the Art

3.1 Image Classification with Deep Learning

In recent years, image classification using deep learning has significantly advanced due to the evolution and optimization of Convolutional Neural Networks (CNNs). Landmark models such as AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGGNet (Simonyan & Zisserman, 2014), ResNet (He, Zhang, Ren, & Sun, 2016), and GoogLeNet (Szegedy et al., 2015) have drastically improved accuracy and efficiency on large-scale benchmarks like ImageNet (Deng et al., 2009). Techniques like transfer learning and fine-tuning enable these models to be adapted efficiently for specialized tasks such as classifying nature-related imagery, especially when labeled data is scarce (Tan & Le, 2019).

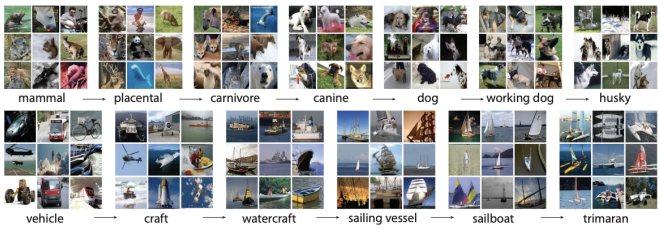

Figure 4: Imagenet

Complementary approaches involving unsupervised methods and autoencoders have been explored to extract robust features from images with high variability, providing both accurate classification and deep insights into underlying data characteristics (Goodfellow, Bengio, & Courville, 2016).

3.2 Interpretability Techniques in Classification Models

Despite impressive accuracy, CNNs often operate as "black boxes," prompting the need for interpretability methods to reveal decision-making processes.

Class Activation Mapping (CAM) and Grad-CAM

CAM and Grad-CAM (Selvaraju et al., 2017) are widely used for visually interpreting CNN decisions by generating heatmaps highlighting the regions influential for predictions. Grad-CAM uses gradients from the final convolutional layer to identify discriminative regions, effectively aligning the model's focus with human attention. Grad-CAM's effectiveness in revealing biases or dataset artifacts has made it essential for transparent AI evaluations.

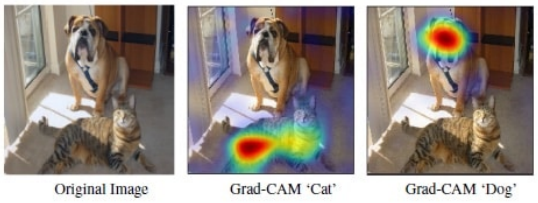

Figure 5: Grad-CAM

Local Interpretable Model-Agnostic Explanations (LIME)

LIME (Ribeiro, Singh, & Guestrin, 2016) provides local, detailed explanations by segmenting images into superpixels and evaluating their influence on model predictions through perturbations. Despite its granularity, LIME faces stability and variability challenges, though it remains valuable for individual prediction explanations.

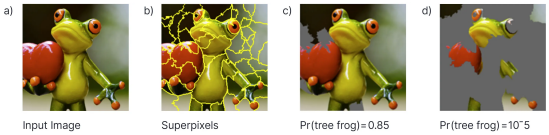

Figure 6: LIME

SHapley Additive exPlanations (SHAP)

SHAP (Lundberg & Lee, 2017) uses game theory to quantify pixel-level feature contributions to predictions, ensuring consistency and robustness in interpretations. Its unified approach integrates local and global interpretability, making it effective for validating model behavior.

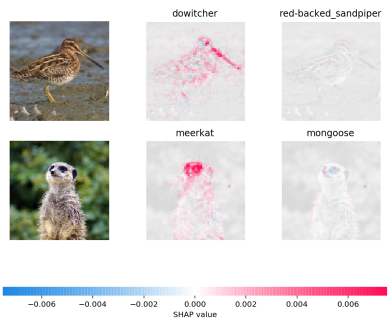

Figure 7: SHAP

3.3 Relevant Projects in Nature and Social Media Imagery

ProtoPNet for Fine-Grained Classification

The Prototypical Part Network (ProtoPNet) (Chen et al., 2019) introduced interpretability into fine-grained image classification (e.g., bird species) by learning prototypical image patches. ProtoPNet provides explanations such as "this image resembles this prototypical part," maintaining accuracy comparable to traditional CNNs while providing transparent, case-based reasoning essential for ecological research.

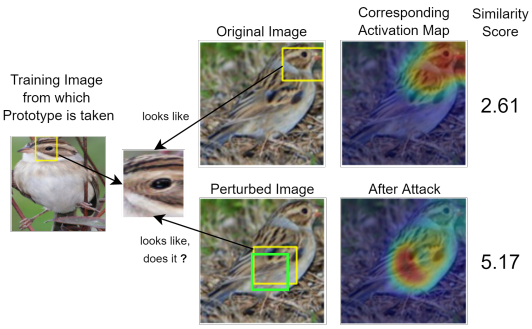

Figure 8: ProtoPNet

Bird Species Classification with XAI

Kumar and Kondaveeti (2024) developed a pipeline for bird species classification using transfer learning and LIME. Their study emphasized that accuracy alone doesn't guarantee interpretability, as high-performing models could rely on background features. By quantitatively evaluating explanations with Intersection-over-Union (IoU), they showed that EfficientNet-B0 had superior interpretability alongside high accuracy, demonstrating the critical role of interpretability in ecological research.

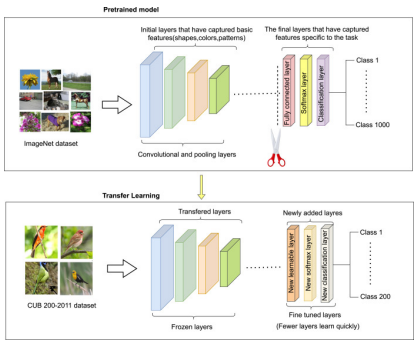

Figure 9: Model scheme

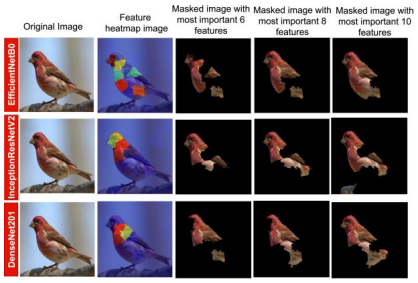

Figure 10: IoU results

Explainable Face Recognition (XFR)

The XFR project (Williford, May, & Byrne, 2020) provided a rigorous evaluation framework for face recognition interpretability. Through controlled face-triplet experiments, they benchmarked interpretability techniques like subtree Excitation Back-Propagation and DISE, significantly improving accuracy in highlighting distinguishing facial features.

They tried to answer the following question: *"What image region of the probe example is most similar to the mate example and least similar to the non-mate one?*. However, their approach has critical implications for ethical and unbiased deployment of AI in social contexts.

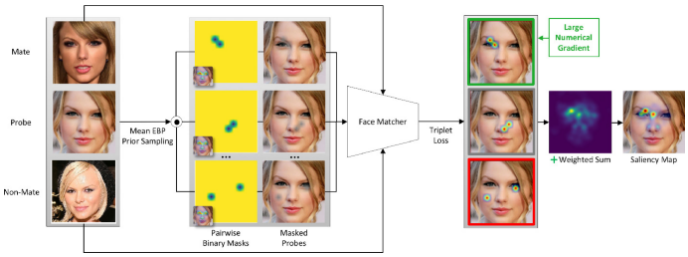

Figure 11: XFR

Integrating interpretability into CNN-based image classification significantly enhances transparency and trustworthiness, especially critical for domains like ecological monitoring and social media analysis. Techniques such as Grad-CAM, LIME, and SHAP provide complementary insights into model behavior, highlighting the importance of combining multiple interpretability methods for robust explanations. Projects like ProtoPNet and Concept Bottlenecks underline that interpretability need not compromise accuracy, while XFR emphasize interpretability's role in identifying biases and ensuring ethical AI deployments. Collectively, these studies affirm that combining accurate models with effective interpretability is essential for deploying AI systems responsibly in sensitive domains.

4 Objectives

The specific objectives of the project are:

1. **Data acquisition and processing:** Collect and preprocess images from different social media platforms to create a well-structured and labeled dataset.
2. **Classifier development:** Implement and train a neural network-based classifier to categorize nature-related visual content.
3. **Interpretability techniques:** Apply and compare methods such as CAM, Grad-CAM, SHAP, and LIME to visualize and analyze model decision-making.
4. **Comparison and evaluation:** Assess the effectiveness of interpretability techniques to determine which is most effective in this context.
5. **Conclusions and implications:** Evaluate how AI models contribute to the understanding of nature-related content and their potential impact on environmental awareness.

5 Methodology

The development of this project will be carried out in several phases, ranging from the review of previous studies to the analysis of results. Below are the key stages of the methodological process.

6.1: Literature Review

The first step will consist of an in-depth study of previous research on image classification in social media videos and interpretability methods applied to neural networks. Additionally, key differences between human and automated visual perception will be analyzed to better understand the challenges in visual content classification. This phase will help establish the foundation of the project and select the most suitable techniques for its implementation.

6.2: Data Acquisition and Processing

Videos and images will be collected from platforms such as YouTube, X, and TikTok, selecting content containing nature-related elements. Subsequently, the videos will be segmented into keyframes for individual analysis.

The data processing phase will include:

- Keyframe extrcation and image normalization to ensure consistency in visual characteristics.
- Labeling of keyframes, following a specific criterion to maximize the number of correctly classified images.
- Filtering and dataset cleaning, removing redundant or low-quality images to improve the model's accuracy.

Since data quality and diversity are key factors in classifier performance, a significant part of the project will focus on ensuring a well-structured and representative dataset.

6.3: Classifier Development

For image classification, a convolutional neural network (CNN) model will be implemented, evaluating advanced architectures such as ResNet, EfficientNet, or Vision Transformers. WE will make use of pre-trained models and fine-tuning to improve generalization with our limited data.

The model will be trained and validated using the collected data, with hyperparameter tuning and techniques applied to mitigate overfitting.

6.4: Implementation of Interpretability Techniques

To enhance the understanding of the model's predictions, interpretability techniques such as Class Activation Mapping (CAM) will be applied to visualize the regions of the image that influence classification decisions.

Additionally, the performance of Grad-CAM, SHAP, and LIME, as well as other recent methodologies, will be compared to determine which provides the best insight into the model's decision-making process in this context.

This phase will not only help assess the system's transparency but will also contribute to analyzing the consistency between the model's predictions and human perception.

6.5: Analysis and Conclusions

Finally, a comprehensive analysis of the obtained results will be conducted, addressing the following key aspects:

1. Effectiveness of interpretability methods, comparing their ability to explain the classifier's behavior.
2. Relationship between human perception and automated classification, identifying possible discrepancies and their implications.
3. Relevance of the study for future research, exploring how this approach can contribute to a better understanding of the interaction between technology and nature appreciation.

The findings will provide a critical perspective on the use of AI for image classification and its impact on the analysis of visual content in social media.

6 Planning

The project will be developed over approximately five months (including the final presentation), following a structured timeline to ensure a logical progression between phases.

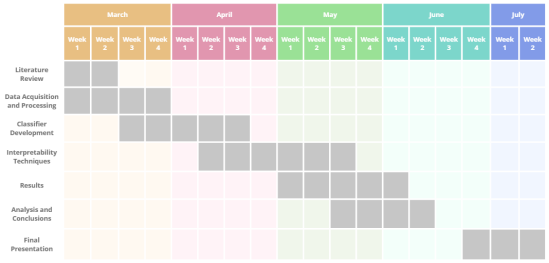

Figure 12: Gantt Chart

As it can be seen in the previous Gant Chart, the initial phase, Literature Review, spans the first three weeks of March. This period is dedicated to understanding prior research on image classification and interpretability techniques. A thorough review is essential to establish a strong theoretical foundation before moving forward.

Simultaneously, Data Acquisition and Processing will begin in the second week of March and continue until mid-April. This stage includes video collection, keyframe extraction, image labeling, and dataset cleaning. The allocated time ensures that the dataset is diverse and well-structured, which is critical for achieving accurate classification results.

Once the dataset is ready, the Classifier Development will start in the last week of March and extend through April. This phase involves selecting and implementing a CNN-based model, exploring pre-trained architectures, and performing fine-tuning. The four-week duration allows for iterative experimentation and model refinement.

Following classifier development, Interpretability Techniques will be implemented from mid-April to mid-May. This stage focuses on evaluating explainability methods, such as Grad-CAM and SHAP, ensuring that the model's decisions can be effectively analyzed.

Results analysis will take place in May and early June, using metrics to assess both classifier performance and interpretability effectiveness. The allocated time allows for an in-depth evaluation and comparison of different techniques.

The final phase, Analysis and Conclusions, will be conducted throughout June, synthesizing insights from all previous steps. This phase is crucial to contextualizing findings and identifying potential improvements.

Lastly, the Final Presentation will be prepared in early July, summarizing key results and conclusions in a structured format.

This structured planning ensures a smooth progression between tasks, balancing time allocation to prioritize data quality, model development, and result interpretation effectively.

7 Data Collection and Dataset Creation

The first step in our project involved creating a comprehensive and high-quality dataset specifically tailored for training the image classifier. Unlike projects that use publicly available datasets, our dataset was entirely created by the project team to accurately represent content shared on social media platforms. This process consisted of two main stages: data collection and data labeling.

7.1 Data Collection

The initial step was to collect relevant visual content from popular social media platforms, specifically YouTube and X. Videos were sourced from YouTube, while additional images were gathered from X, ensuring a diverse and representative dataset.

For YouTube, a systematic approach was employed to ensure reliability and consistency. We first collected various videos and subjected them to manual coding by multiple coders from the Big 5 group. Only videos that had been independently labeled by at least two coders, and where all assigned labels matched exactly, were retained. This criterion was crucial to avoid any disagreements in labeling, ensuring a reliable dataset.

After applying this quality check, we proceeded to download the approved videos for further processing.

7.2 Frame Extraction

From each validated YouTube video, we extracted representative frames to convert video content into static images suitable for classifier training. To achieve this, we adopted two strategies:

- For videos with clear and detectable scene changes, we utilized automated keyframe extraction using a scene-change detection algorithm. This process involved running each video through a frame extraction pipeline, extracting key frames whenever the algorithm detected a significant scene transition. Specifically, the process used a threshold-based method, where distinct changes between consecutive frames prompted the extraction of a new keyframe.

- In cases where no scene transitions were detected (i.e., the algorithm did not identify any significant visual variation), we manually extracted three representative frames from each video: one at the beginning, one from the middle, and one at the end. This method ensured sufficient representation even in low-variability videos.

Additionally, to enrich the dataset diversity and extend our coverage of nature-related imagery, we included supplementary images obtained directly from X. These images were collected using targeted searches focused on nature-related hashtags and verified accounts posting content aligned with the project's focus.

7.3 Data Labeling and Dataset Creation

Once the visual data was collected and processed into individual images, we began the labeling phase. To ensure consistency, the Big 5 group developed and agreed upon a detailed coding protocol. This standardized approach allowed us to classify images according to predefined categories clearly and consistently, facilitating subsequent analysis and model training.

Each coder followed the established coding protocol to classify images into four distinct, mutually exclusive categories:

1. **Nature_visual**: Nature vs. Non-Nature
2. **Nep_materiality_visual**: Biotic vs. Abiotic
3. **Nep_biological_visual**: Material vs. Immaterial
4. **Landscape-type_visual**: Artificial surfaces, Agricultural areas, Forests and seminatural areas, Wetlands, or Water bodies.

After labeling was completed, we consolidated all images into a centralized directory structure. Alongside these images, we created a file (CSV format) to systematically link each image with its assigned labels. This file recorded essential details, including:

- **Video ID**: A unique identifier corresponding to the original source video (in case of youtube data).
- **Frame name**: The specific name assigned to each extracted image frame (both youtube and X images).
- **Class Labels**: Clearly indicating the assigned labels for each of the four classification categories mentioned above.

This structured approach resulted in a highly organized dataset ready for the subsequent phases of the project. The dataset not only supports efficient training of the classification model but also ensures reproducibility and transparency in the labeling process, ultimately enhancing the reliability of the project's findings and facilitating further analyses using interpretability techniques.