Beyond Human Perception: Challenges in AI Interpretability of Orangutan Artwork

Cedric Sueur, Elliot Maître, Jimmy Falck, Masaki Shimada, Marie Pelé

Beyond Human Perception: Challenges in AI Interpretability of Orangutan Artwork

Cédric Sueur[1,2], Elliot Maitre[4], Jimmy Falck[3], Masaki Shimada[5], Marie Pelé[6]

1. Université de Strasbourg, IPHC, CNRS, UMR 7178, 67000 Strasbourg, France

2. Institut Universitaire de France, 75231 Paris, France

3. Université de Toulouse - IRIT UMR5505, 31400, Toulouse, France

4. Laboratoire lorrain de recherche en informatique et ses applications (Loria - CNRS/Université de Lorraine/Inria), Nancy, France

5. Department of Animal Sciences, Teikyo University of Science, 2525, Yatsusawa, Uenohara 409-0193, Yamanashi, Japan

6. ANTHROPO-LAB - ETHICS EA 7446, Université Catholique de Lille, F-59000 Lille, France

Corresponding author: Cédric Sueur, cedric.sueur@iphc.cnrs.fr, 0033388107453.

Abstract: Drawings serve as a profound medium of expression for both humans and apes, offering unique insights into the cognitive and emotional landscapes of the artists, regardless of their species. This study employs artificial intelligence (AI), specifically Convolutional Neural Networks (CNNs) and the interpretability tool Captum, to analyze non-figurative drawings by Molly, an orangutan. The research utilizes VGG19 and ResNet18 models to

decode seasonal nuances in the drawings, achieving notable accuracy in seasonal classification and revealing complex influences beyond human-centric methods. Techniques such as occlusion, integrated gradients, PCA, t-SNE, and Louvain clustering highlight critical areas and elements influencing seasonal recognition, providing deeper insights into the drawings. This approach not only advances the analysis of non-human art but also demonstrates the potential of AI to enrich our understanding of non-human cognitive and emotional expressions, with significant implications for fields like evolutionary anthropology and comparative psychology.

Keywords: deep learning, non-human primates, primatology, apes, explicability

1. Introduction

The potential for anthropocentric bias emerges when deciphering the content of drawings, a common focus in studies of figurative artwork. However, meanings also permeate non-figurative sketches, as evidenced in creations by young children (Gardner 1981; Goodnow 2013; Restoy et al. 2022). Drawings are rich in information, yet reliance on a limited set of manually selected features can restrict the depth of analysis. This is especially pertinent in examining drawings by non-human primates, where an anthropocentric selection bias may overlook features significant to other species (Saito et al. 2014; Martinet and Pelé 2021). Consequently, such an approach fails to harness the full informational value of these drawings. In human contexts, querying the artist about their intent offers a partial solution (Martinet et al. 2021; Sueur et al. 2022), though this method falls short when artists cannot verbally articulate their intentions, such as in the scribbles of young children or individuals with communication impairments. This limitation extends to non-verbal non-human drawers like non-human primates (Pelé et al. 2021; Martinet et al. 2023).

In a previous study, we employed artificial intelligence to scrutinize the artwork of Molly, a female orangutan who produced 1,299 drawings as part of a behavioral enrichment program at Tama Zoo in Japan until her passing in 2011 (Pelé et al. 2021). Previous research on Molly's drawings revealed influences from her caretaker's identity and daily events, like the birth of peers (Hanazuka et al. 2019). Traditional ethological approaches were used to distinguish changes in Molly's drawings over time and seasonal variations, noting preferences for certain colors and line styles across different seasons. However, these manual feature extraction methods, focused on elements commonly analyzed in children's drawings, such as loops and circles, might not align with orangutan perception (Kellogg 1969). Deep learning, leveraging artificial neural networks (Jacob et al. 2021), offers a robust alternative for image analysis, excelling in tasks like microscopy image classification (Buetti-Dinh et al. 2019) and

disease diagnosis (Zhou et al. 2021) without the need for predefined features. Convolutional Neural Networks (CNNs), a prevalent deep learning tool in image analysis, autonomously learn to identify relevant features, evolving from simple shapes to complex objects through layers. While CNNs enhance accuracy, they struggle with interpretability, prompting the development of techniques to demystify model decisions (Zhang and Zhu 2018; Carabantes 2020).

Artificial intelligence's application to drawing behavior remains very little explored (Beltzung et al. 2023), despite its success in classification tasks, like identifying stroke patterns (Wu et al. 2018) or categorizing drawings by object (Zhang et al. 2016). Our prior study in 2022 (Beltzung et al. 2022) harnessed AI to examine seasonal trends in Molly's artwork, utilizing the VGG19 (Dutta et al. 2016) model for seasonal classification, achieving a 41.6% accuracy. We explored how different features, from simple to complex, influenced these seasonal variations, highlighting the roles of color and pattern. We found with deep learning models similar results that we found using classical measures in ethology and drawing analyses. The research underscored deep learning's potential to objectively analyze non-figurative art, encouraging its application beyond primates to include human toddler scribbles. However, it was impossible to open the black-box and to fully decipher and interpret the mechanisms through which the deep learning model identifies and classifies elements within the drawings.

Here, we used interpretability and explicability models to reach this aim (Zhang and Zhu 2018; Spannaus et al. 2023). Interpretability in deep learning refers to the ability to understand and explain the decision-making process of deep neural networks. Deep learning models are often considered black boxes because they lack transparency and it is difficult to comprehend how they arrive at their predictions (Shwartz-Ziv and Tishby 2017; Carabantes 2020). Captum (Kokhlikyan et al. 2020) is a novel, unified, open-source model interpretability library for PyTorch. It contains implementations of various gradient and

perturbation-based attribution algorithms for both classification and non-classification models, including graph-structured models built on Neural Networks (NN). Captum emphasizes multimodality, extensibility, and ease of use. It supports different modalities of inputs such as image, text, audio, or video, and allows the addition of new algorithms and features. Captum also introduces an interactive visualization tool called Captum Insights, which enables sample-based model debugging and visualization using feature importance metrics. Integrating Captum into our study enhances the explicability further by providing a comprehensive toolkit for model interpretability. Captum supports various interpretability algorithms, including Occlusion and Integrated gradients (Selvaraju et al. 2017), allowing researchers to not only visualize important regions but also understand the attribution of each input feature to the model's output. By applying Captum's visualizations on the validation data for our model, we can gain deeper insights into the discriminative features recognized by the models. This integration will facilitate a more nuanced understanding of model predictions, particularly in complex cases where direct interpretation of features is not straightforward. At our knowledge, this is the first time that Captum is applied to drawing and artwork.

2. Materials and Methods

2.1. Dataset

The dataset comprises 1,299 drawings created by Molly, a female orangutan who began drawing around the age of 50, from 2006 to 2011. Molly, who seldom interacted with her group members, had a daily routine: spending mornings in an enclosure (either indoors or outdoors) and afternoons in a restroom equipped with crayons, allowing her to draw at her leisure for about 2 to 3 hours, producing 1–2 drawings each day. She indicated the end of a drawing session by placing the drawing materials on the floor. She was given paperboard and

crayons daily, but as the drawing activity was not initially aimed at studying her drawing behavior, only minimal metadata were recorded. The exact dates of the drawings, marking the only external information, enabled categorization into seasons: autumn (374 drawings), summer (284), spring (269), and winter (372). The drawings were adjusted to square shapes and resized to 224x224 pixels for analysis. The dataset was divided into training (907 images) and validation sets (392 images), with seasonal distribution as follows: autumn (28.8%), spring (20.7%), summer (21.9%), and winter (28.6%). For additional information on Molly and her artwork, see references (Hanazuka et al. 2019; Pelé et al. 2021).

2.2. Convolutional Neural Network (CNN)

Acknowledging that human-selected features might not capture the intricate details in drawings, our study focused on deep learning, particularly convolutional neural networks (CNNs). Initially, we trained CNNs, specifically VGG19 (Beltzung et al. 2022) and ResNet 18 (Liang 2020), to predict the seasons depicted in drawings by Molly. To address the limitations of our dataset's size, we utilized transfer learning, repurposing models pre-trained on different tasks. This method enabled us to leverage the capabilities of ResNet18 and VGG19, both pre-trained with ImageNet weights, enhancing our analysis. ResNet18 is renowned for its efficiency and depth, achieved through residual learning, which we adapted for our study. This model's architecture addresses the vanishing gradient problem with shortcut connections that bypass layers, facilitating the training of deeper networks. We applied this approach to both models, adjusting various parameters and hyperparameters, such as learning rate and the configuration of fully connected layers, to find the optimal setup. Each model's final layer was a fully connected layer with four neurons, each representing a season, using softmax activation for classification. We employed a categorical cross-entropy loss function and optimized the models using stochastic gradient descent (SGD) with a

learning rate of 0.1 and a batch size of 16. Performance was evaluated through early stopping, based on validation set accuracy, ceasing training after no improvement over three epochs. To prevent overfitting, we froze the convolutional blocks and implemented data augmentation techniques, like horizontal and vertical flips, to improve model accuracy.

2.3. Interpretability with Captum.

Following the initial training and optimization of our convolutional neural network models using ResNet18 and VGG19 architectures, we applied advanced analytical techniques to further interpret the deep learning model's decisions and to gain insights into the features influencing the classification of seasons in Molly's drawings. These techniques included the use of Captum (Kokhlikyan et al. 2020) and Scikit learn (Kramer and Kramer 2016; Bisong and Bisong 2019; Hao and Ho 2019) for model interpretability, dimensionality reduction, and clustering algorithms.

Captum – Occlusion: We utilized the Occlusion method from Captum, a model interpretability library for PyTorch (Stevens et al. 2020; Imambi et al. 2021), to understand the impact of different regions of the drawing on the model's prediction (figure 1). By systematically occluding parts of the input image and observing the effect on the model's output, we could identify which areas of the drawings were most significant for determining the season. This method helps in pinpointing the 'positive' pixels or regions that contribute most to the classification decision.

Average Number of Positive Pixels Per Season: Building on the occlusion analysis, we calculated the average number of positive pixels per season. This step involved aggregating the pixels that positively influenced the model's seasonal classification across all drawings attributed to a specific season. By analyzing these averages, we aimed to discover seasonal

patterns or features that were consistently influential across Molly's drawings, providing a quantitative measure of the visual elements most associated with each season (figure 1).

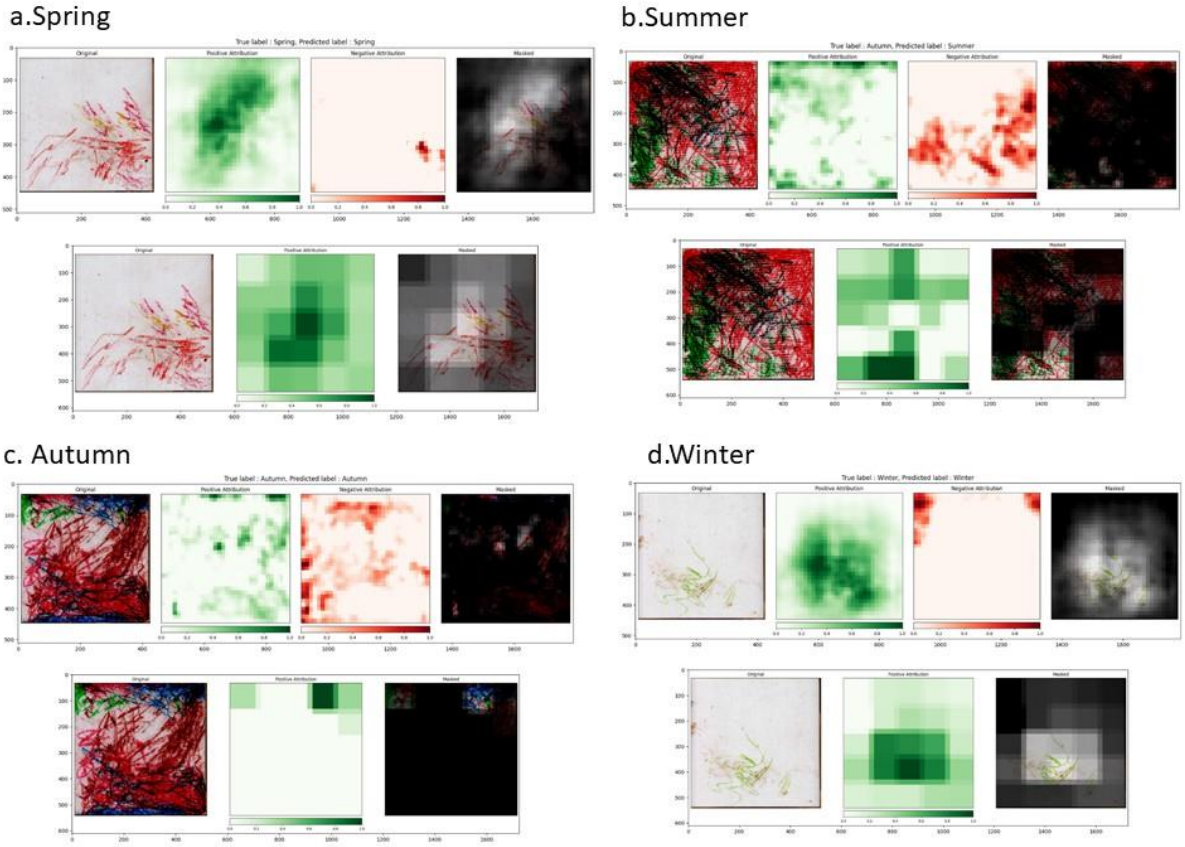

Figure 1: Seasonal Examples of Occlusion Analysis Using Captum Across Various Pixel Sizes. For each season, the sequence displays: the original drawing (top), followed by positive attribution (in green) highlighting pixels crucial for identification and classification, negative attribution (in red) indicating non-essential pixels, and finally, the differential attribution showcasing the contrast between positive and negative contributions. The top row represents a threshold (t) of 0, and the bottom row a threshold (t) of 0.5 for each depicted example.

Captum – Integrated Gradients: We employed Integrated Gradients (Kwon et al. 2021), another interpretability technique from Captum, to attribute the prediction of the model to its input features (figure 2). This method offers a way to visualize the importance of each pixel in

the original drawing for the classification decision. By highlighting the pixels and regions within the drawings that had the most significant impact on the model's prediction, Integrated Gradients provided a deeper understanding of the model's behavior and the features it deemed important.

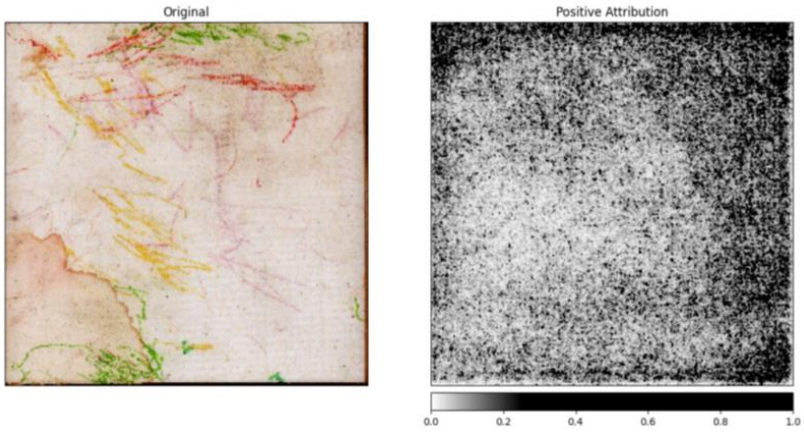

Figure 2: Integrated Gradients Analysis Using Captum. This illustration juxtaposes the original drawing on the left with its positive attribution on the right, where black denotes the pixels of utmost importance for classification.

Removal of the Linear Layer from the Network: To explore the features extracted by the CNN in a more interpretable form, we removed the final linear (fully connected) layer of the network. This allowed us to access the raw features extracted by the convolutional layers directly, which represent a high-level abstraction of the drawings.

Application of PCA / t-SNE in Scikit learn: With the linear layer removed and raw features extracted, we applied Principal Component Analysis (PCA) (Holland 2008) and t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton 2008; Wattenberg et al. 2016) for dimensionality reduction. These techniques reduced the high-dimensional feature space into a two- or three-dimensional space, making it possible to visualize the distribution

and relationships between different drawings. This visualization helped us to observe patterns and clusters within the data, potentially revealing intrinsic similarities between drawings of the same season or identifying outliers.

Clustering (Louvain Algorithm): Finally, we applied the Louvain algorithm (Combe et al. 2015; Emmons et al. 2016) for clustering the drawings based on their reduced-dimensional features. The Louvain algorithm is a community detection method known for its efficiency in large networks. By clustering the drawings, we aimed to discover natural groupings within the data, which could indicate distinct styles, motifs, or themes recurring across different seasons. This unsupervised learning approach provided a bottom-up perspective on the dataset, potentially uncovering new insights into Molly's drawing behavior and how it varied with the seasons.

Through the application of these advanced techniques, our study sought to deepen the analysis of non-human drawing, leveraging the power of deep learning and interpretability tools to uncover the nuances of seasonal variation in Molly's productions.

2.4. Statistical analyses

In our analysis, we employed the occlusion process to identify the most significant pixel in each drawing. This pixel was evaluated based on several criteria: its positional deviation from the center of the cardboard, the presence of lines, the variety and intensity of colors present, the total number of distinct colors, and the presence of additional elements. These elements included the cardboard backing, any signs of moisture damage, and traces not intended for drawing, as well as any tears or damage at the specific pixel's location.

Subsequently, we conducted a Kruskal-Wallis test for each of these dependent variables, with the season acting as the independent variable. This non-parametric test was chosen to

determine if there were statistically significant differences across the seasons. The analysis was performed using the R statistical software, with a significance level set at $\alpha = 0.05$.

Image analyses are available at https://doi.org/10.5281/zenodo.10973649. All codes are available at https://github.com/cedricsueur/drawinganalyses

3. Results

3.1. CNNs: VGG19 and ResNet18.

The VGG19 model trained to classify drawings according to seasons achieved 42% accuracy on the test set. The ResNet18 reached an accuracy a bit higher of 50%. These accuracies are higher than that expected by random (approximately 29%, by always classifying drawings as the most common class). To provide further context, the same models were also tasked with differentiating between drawings by Molly and humans drawings. In this classification challenge, both models demonstrated a remarkable accuracy higher than 93%. Additionally, when the model was retrained for a binary classification task—distinguishing between drawings with low and high coverage by Molly—it achieved an even accuracy higher than 95%. These outcomes suggest that the model is highly capable of distinguishing between drawings when the distinctions are pronounced. Consequently, the relatively modest accuracy in the seasonal classification task likely stems from the more nuanced differences present among Molly's drawings across different seasons.

3.2. Captum Occlusion and integrated gradients

Expanding upon our occlusion analysis, we computed the average number of positive pixels for each season, which are pixels crucial for correctly identifying the season depicted in the drawings. Our comparison across seasons, accounting for varying degrees of occlusion,

revealed that independent of pixel size, Winter and Spring drawings consistently had a higher count of significant pixels for accurate season identification than those from Summer and Autumn. While we also explored the use of Integrated Gradients as an interpretive tool, this method proved less definitive than Occlusion in providing clear insights. Consequently, we chose not to pursue further analysis with Integrated Gradients, focusing instead on the more revealing outcomes derived from the occlusion technique.

Our analysis revealed that 78.5% of the pixel identified as most significant per drawing contained markings attributable to the ape's drawing activities. In contrast, the average area covered by markings on the paperboard for Molly was approximately 60%. Notably, the positional data of these key pixel per drawing exhibited seasonal variation (Kruskal-Wallis chi-squared = 16.721, df = 3, p-value = 0.0008064), with pixels from drawings made in Autumn tending to be more centrally located compared to those from Spring and Winter (figure 3). However, when examining other characteristics of the most significant pixel per drawing —such as the number of colors present, specific colors used, and the presence of additional marks or traces—no seasonal differences were observed (Kruskal-Wallis chi-squared < 6.7164, df = 3, p-value > 0.08151).

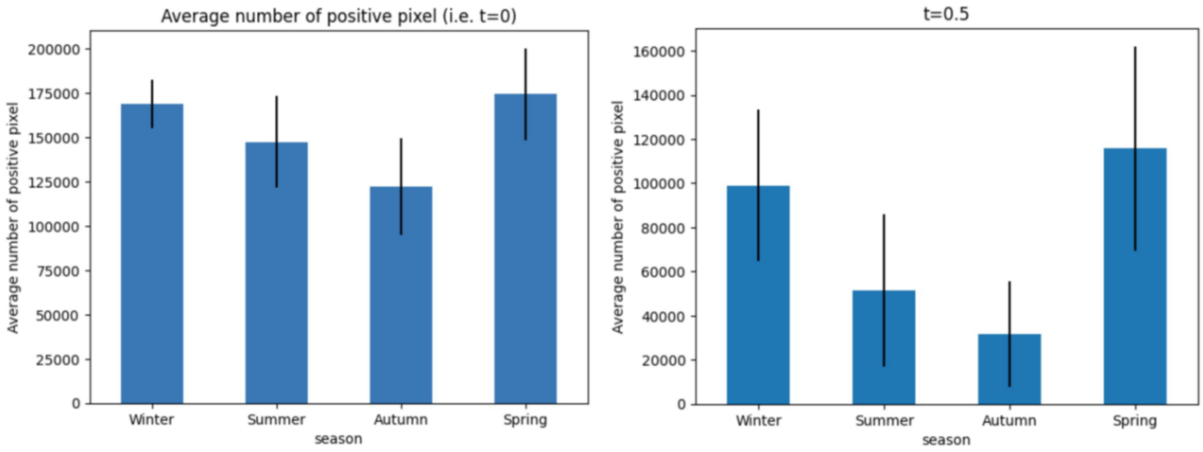

Figure 3: Comparative Analysis of Positive Attribution Pixel Counts. This graph presents the average number of pixels significant for seasonal drawing classification at two thresholds: t=0 (indicating smaller pixels) and t=0.5 (representing larger pixels), highlighting the role of pixel size in determining drawing classification importance.

3.3. PCA and Louvain Clustering

After the linear layer removed and raw features extracted, the Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction did not reveal any clusters according to the season (figure 4). We obtained similar results with the Louvain clustering algorithm. However, it revealed two clusters, one with many drawings and one with a lower number of drawings (figure 4). The small cluster revealed to be a cluster with many second drawings, meaning drawings made the same day after a first one. These second drawings have much more lines than the first drawings but are also made at the back of the paperboard which is not white as its front. Surprisingly, despite these differences, these second drawings are correctly classified with the season by the models.

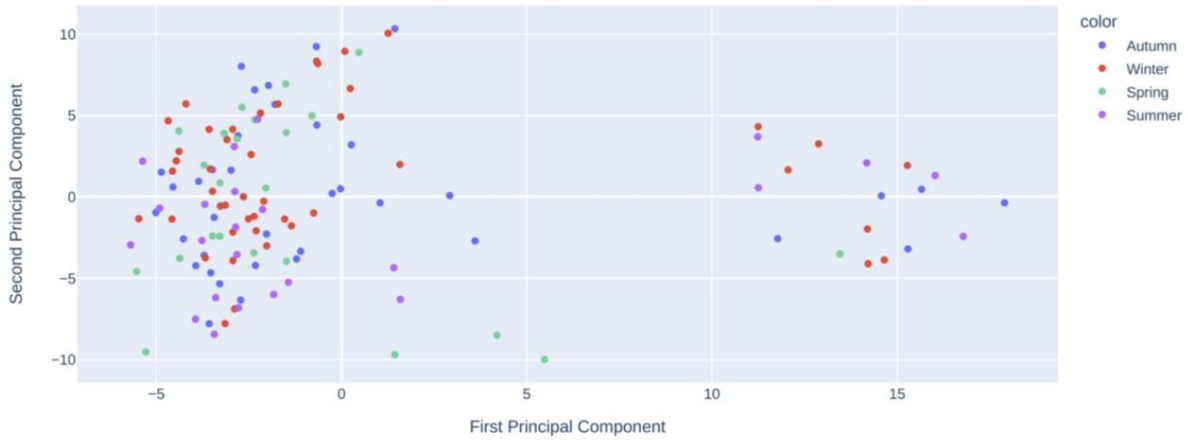

Figure 4: PCA Two-Dimensional Visualization. This graph displays clusters identified through Louvain algorithm, with each color representing a different season, illustrating the distribution and grouping of seasonal variations within the PCA-reduced feature space.

4. Discussion

In synthesizing insights from our deep learning-based analysis of non-human drawings, particularly those by Molly the orangutan, with discussions on the interpretability of these models, several key themes and conclusions emerge that bridge the gap between traditional drawing analysis and advanced AI methodologies. The use of Captum (Kokhlikyan et al. 2020) specifically for the analysis of drawings and artwork appears to be a novel application. While interpretability tools have been increasingly applied in various domains to understand the decision-making processes of AI models—ranging from healthcare and finance to autonomous vehicles and natural language processing—their application within the realm of art analysis, especially in studying non-human drawings, marks a pioneering step. As Nagel pointed out (Nagel 1980), it's beyond human capability to fully understand the experience of being another animal. However, we believe that artificial intelligence can play a role in reducing the prevalence of biases.

4.1. Deep Learning Models: VGG19 and ResNet18

The performance disparities between VGG19 and ResNet18 in classifying seasonal variations in Molly's drawings underline the nuanced nature of this task. While both models significantly outperform random chance, indicating an ability to detect some form of seasonal patterning, the modest accuracies highlight the challenge of discerning subtle distinctions within the drawings. The stark contrast in model performance when tasked with more defined classification problems, such as differentiating between non-human and human drawings or

assessing drawing coverage, reinforces the notion that deep learning excels in identifying pronounced differences. This suggests that the complexity of seasonal variation in Molly's drawings may encompass more intricate, less overt features that challenge the models' classification capabilities. Indeed, Molly may change mood and personality across days, seasons and age affecting drawings (Hanazuka et al. 2019; Pelé et al. 2021) as other apes also having these activities (Martinet et al. 2023).

4.2. Interpretability: Captum Occlusion and Integrated Gradients

The employment of Captum's interpretability tools, especially Occlusion, has shed light on the decision-making processes of our models, pinpointing certain pixels and areas as pivotal for recognizing different seasons in the drawings. This method enriches our grasp of the model's perceptual focus, but it's crucial to acknowledge a limitation: while we can identify which pixels influence seasonal classification, the underlying features leading to these distinctions remain less clear (Gilpin et al. 2018). The inherent challenge lies in discerning the specific attributes—be it color, texture, or shape—that these critical pixels represent, as the models do not explicitly reveal this.

Opting for Occlusion over Integrated Gradients was driven by the former's ability to yield more direct insights into the importance of various image regions for model predictions. However, this preference also brings to the fore the complexity of interpreting deep learning models. Even as Occlusion helps highlight influential regions within the drawings, it underscores a broader challenge in AI interpretability: understanding the 'why' behind the model's reliance on these regions (Gilpin et al. 2018; Fan et al. 2021; Rudin et al. 2022; Li et al. 2022).

This cautious approach to interpreting Occlusion results emphasizes the necessity of combining interpretability tools with domain expertise. By doing so, we can hypothesize

about the features these critical pixels may correspond to, such as seasonal color palettes or thematic elements unique to certain times of the year. Yet, the exact nature of these features often requires further investigation, possibly through additional analytical techniques or cross-referencing with domain-specific knowledge.

4.3. PCA and Louvain Clustering

The utilization of PCA, t-SNE, and Louvain clustering further extends the analytical depth of this study, revealing unexpected patterns such as the clustering of drawings based on sequential order rather than seasonal attributes. This finding suggests that Molly's drawing behavior—and potentially that of other non-human artists—may be influenced by factors unrelated to season, such as the physical context of the drawing or the events of the day. The ability of the models to accurately classify these drawings despite such confounding factors speaks to the robustness of deep learning in extracting relevant features from complex data.

4.4. Implications and Future Directions

This study with the previous ones illustrates the complementary strengths of traditional drawing analysis and AI-driven methodologies as made in other domains (Soto and Adey 2016; Lu et al. 2024). Where traditional analysis provides a framework for understanding the thematic and stylistic components of drawings, deep learning models offer a means to systematically and objectively analyze these components across large datasets. The intersection of these approaches, facilitated by interpretability tools like Captum, holds the promise of enriching our understanding of non-human art, offering nuanced insights that neither approach could achieve in isolation. However, it is needed to think and to build a framework to know how to work in complementary with traditional approaches and AI-driven approaches.

The implications of this research extend beyond the specific case of Molly's drawings, touching on broader themes in evolutionary anthropology, comparative psychology, and animal welfare. By applying deep learning to the study of non-verbal drawing behavior across different species, we can explore evolutionary trajectories of artistic expression and cognitive processes (Sueur and Pelé 2023). Additionally, the potential for deep learning to assist in the early detection of neurodegenerative diseases in apes presents a novel application of AI in enhancing animal welfare and healthcare (Pelé et al. 2021).

In conclusion, the integration of deep learning into the analysis of non-human drawings not only challenges and expands the boundaries of traditional art analysis but also opens new avenues for interdisciplinary research. As we continue to refine these models and interpretability techniques, the potential for AI to deepen our understanding of the cognitive and emotional worlds of non-human artists becomes increasingly evident, promising insights into the universal language of art and expression.

Funding

This project received financial support from the CNRS through the MITI interdisciplinary programs, from the PNRIA and from the University of Strasbourg through an IDEX Exploratory Research program. This study was partially supported by JSPS KAKENHI (grant number 20H01409).

Institutional Review Board Statement

The Tama Zoological Park Ethics Board approved this noninvasive behavioral study, which complied with the Code of Ethics of the Japanese Association of Zoos and Aquariums.

Informed Consent Statement

Not applicable.

Data Availability Statement

Dataset is available at https://doi.org/10.5281/zenodo.10973649. All codes are available at https://github.com/cedricsueur/drawinganalyses

Acknowledgments

We are grateful to the Tama Zoological Park in Japan for providing the orangutan drawings.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the study design; collection, analyses, or interpretation of data; writing of the manuscript; or decision to publish the results.

References

Beltzung B, Pelé M, Renoult JP, et al (2022) Using Artificial Intelligence to Analyze Non-Human Drawings: A First Step with Orangutan Productions. Animals 12:2761. https://doi.org/10.3390/ani12202761

Beltzung B, Pelé M, Renoult JP, Sueur C (2023) Deep learning for studying drawing behavior: A review. Front Psychol 14:992541

Bisong E, Bisong E (2019) Introduction to Scikit-learn. Build Mach Learn Deep Learn Models Google Cloud Platf Compr Guide Begin 215–229

Buetti-Dinh A, Galli V, Bellenberg S, et al (2019) Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. Biotechnol Rep 22:e00321

Carabantes M (2020) Black-box artificial intelligence: an epistemological and critical analysis. AI Soc 35:309–317

Combe D, Largeron C, Géry M, Egyed-Zsigmond E (2015) I-louvain: An attributed graph clustering method. Springer, pp 181–192

Dutta A, Gupta A, Zissermann A (2016) VGG image annotator (VIA). URL Httpwww Robots Ox Ac Uk Vggsoftwarevia

Emmons S, Kobourov S, Gallant M, Börner K (2016) Analysis of network clustering algorithms and cluster quality metrics at scale. PloS One 11:e0159161

Fan F-L, Xiong J, Li M, Wang G (2021) On interpretability of artificial neural networks: A survey. IEEE Trans Radiat Plasma Med Sci 5:741–760

Gardner H (1981) Artful scribbles: The significance of children's drawings

Gilpin LH, Bau D, Yuan BZ, et al (2018) Explaining explanations: An approach to evaluating interpretability of machine learning. ArXiv Prepr ArXiv180600069 118

Goodnow J (2013) Children drawing. Harvard University Press

Hanazuka Y, Kurotori H, Shimizu M, Midorikawa A (2019) The effects of the environment on the drawings of an extraordinarily productive orangutan (Pongo pygmaeus) artist. Front Psychol 10:2050

Hao J, Ho TK (2019) Machine learning made easy: a review of scikit-learn package in python programming language. J Educ Behav Stat 44:348–361

Holland SM (2008) Principal components analysis (PCA). Dep Geol Univ Ga Athens GA 30602–2501

Imambi S, Prakash KB, Kanagachidambaresan G (2021) PyTorch. Program TensorFlow Solut Edge Comput Appl 87–104

Jacob G, Pramod R, Katti H, Arun S (2021) Qualitative similarities and differences in visual object representations between brains and deep networks. Nat Commun 12:1872

Kellogg R (1969) Analyzing children's art. McGraw-Hill Humanities, Social Sciences & World Languages

Kokhlikyan N, Miglani V, Martin M, et al (2020) Captum: A unified and generic model interpretability library for pytorch. ArXiv Prepr ArXiv200907896

Kramer O, Kramer O (2016) Scikit-learn. Mach Learn Evol Strateg 45–53

Kwon HJ, Koo HI, Cho NI (2021) Improving explainability of integrated gradients with guided non-linearity. IEEE, pp 385–391

Li X, Xiong H, Li X, et al (2022) Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Knowl Inf Syst 64:3197–3234. https://doi.org/10.1007/s10115-022-01756-8

Liang J (2020) Image classification based on RESNET. IOP Publishing, p 012110

Lu Y, Shen Z, Shen L, et al (2024) Combining AI and traditional screening for discovery of a potent ROCK2 inhibitor against lymphoma. J Mol Struct 1303:137394. https://doi.org/10.1016/j.molstruc.2023.137394

Martinet L, Pelé M (2021) Drawing in nonhuman primates: What we know and what remains to be investigated. J Comp Psychol Wash DC 1983 135:176–184. https://doi.org/10.1037/com0000251

Martinet L, Sueur C, Hirata S, et al (2021) New indices to characterize drawing behavior in humans (Homo sapiens) and chimpanzees (Pan troglodytes). Sci Rep 11:3860. https://doi.org/10.1038/s41598-021-83043-0

Martinet L, Sueur C, Matsuzawa T, et al (2023) Tool assisted task on touchscreen: a case study on drawing behaviour in chimpanzees (Pan troglodytes). Folia Primatol (Basel) 1:1–17

Nagel T (1980) What is it like to be a bat? In: The Language and Thought Series. Harvard University Press, pp 159–168

Pelé M, Thomas G, Liénard A, et al (2021) I Wanna Draw Like You: Inter- and Intra-Individual Differences in Orang-Utan Drawings. Animals 11:3202. https://doi.org/10.3390/ani11113202

Restoy S, Martinet L, Sueur C, Pelé M (2022) Draw yourself: How culture influences drawings by children between the ages of two and fifteen. Front Psychol 13:940617

Rudin C, Chen C, Chen Z, et al (2022) Interpretable machine learning: Fundamental principles and 10 grand challenges. Stat Surv 16:1–85

Saito A, Hayashi M, Takeshita H, Matsuzawa T (2014) The origin of representational drawing: A comparison of human children and chimpanzees. Child Dev 85:2232–2246

Selvaraju RR, Cogswell M, Das A, et al (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. pp 618–626

Shwartz-Ziv R, Tishby N (2017) Opening the black box of deep neural networks via information. ArXiv Prepr ArXiv170300810

Soto BG de, Adey BT (2016) Preliminary Resource-based Estimates Combining Artificial Intelligence Approaches and Traditional Techniques. Procedia Eng 164:261–268. https://doi.org/10.1016/j.proeng.2016.11.618

Spannaus A, Hanson HA, Penberthy L, Tourassi G (2023) Topological Interpretability for Deep-Learning

Stevens E, Antiga L, Viehmann T (2020) Deep learning with PyTorch. Manning Publications

Sueur C, Martinet L, Beltzung B, Pelé M (2022) Making drawings speak through mathematical metrics. Hum Nat 33:400–424

Sueur C, Pelé M (2023) Fractals and artificial intelligence to decrypt ideography and understand the evolution of language. Behav Brain Sci 46:e254. https://doi.org/10.1017/S0140525X23000808

Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:

Wattenberg M, Viégas F, Johnson I (2016) How to use t-SNE effectively. Distill 1:e2

Wu X, Qi Y, Liu J, Yang J (2018) Sketchsegnet: A rnn model for labeling sketch strokes. IEEE, pp 1–6

Zhang H, Liu S, Zhang C, et al (2016) Sketchnet: Sketch classification with web images. pp 1105–1113

Zhang Q, Zhu S-C (2018) Visual interpretability for deep learning: a survey. Front Inf Technol Electron Eng 19:27–39

Zhou W, Yang Y, Yu C, et al (2021) Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images. Nat Commun 12:1259

Figure Captions

Figure 1: Seasonal Examples of Occlusion Analysis Using Captum Across Various Pixel Sizes. For each season, the sequence displays: the original drawing (top), followed by positive attribution (in green) highlighting pixels crucial for identification and classification, negative attribution (in red) indicating non-essential pixels, and finally, the differential attribution showcasing the contrast between positive and negative contributions. The top row represents a threshold (t) of 0, and the bottom row a threshold (t) of 0.5 for each depicted example.

Figure 2: Integrated Gradients Analysis Using Captum. This illustration juxtaposes the original drawing on the left with its positive attribution on the right, where black denotes the pixels of utmost importance for classification.

Figure 3: Comparative Analysis of Positive Attribution Pixel Counts. This graph presents the average number of pixels significant for seasonal drawing classification at two thresholds: t=0 (indicating smaller pixels) and t=0.5 (representing larger pixels), highlighting the role of pixel size in determining drawing classification importance.

Figure 4: PCA Two-Dimensional Visualization. This graph displays clusters identified through Louvain algorithm, with each color representing a different season, illustrating the distribution and grouping of seasonal variations within the PCA-reduced feature space.