



HAL
open science

Beyond Human Perception: Challenges in AI Interpretability of Orangutan Artwork

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

► **To cite this version:**

Cedric Sueur, Elliot Maître, Jimmy Falck, Masaki Shimada, Marie Pelé. Beyond Human Perception: Challenges in AI Interpretability of Orangutan Artwork. 2024. hal-04714755

HAL Id: hal-04714755

<https://univ-catholille.hal.science/hal-04714755v1>

Preprint submitted on 30 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Beyond Human Perception: Challenges in AI Interpretability of**

2 **Orangutan Artwork**

3 Cédric Sueur^{1,2}, Elliot Maitre⁴, Jimmy Falck³, Masaki Shimada⁵, Marie Pelé⁶

4

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

6 2. Institut Universitaire de France, 75231 Paris, France

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

8 4. Laboratoire lorrain de recherche en informatique et ses applications (Loria -

9 CNRS/Université de Lorraine/Inria), Nancy, France

10 5. Department of Animal Sciences, Teikyo University of Science, 2525, Yatsusawa, Uenohara

11 409-0193, Yamanashi, Japan

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

13 France

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

15

16

17 Abstract: Drawings serve as a profound medium of expression for both humans and apes,

18 offering unique insights into the cognitive and emotional landscapes of the artists, regardless

19 of their species. This study employs artificial intelligence (AI), specifically Convolutional

20 Neural Networks (CNNs) and the interpretability tool Captum, to analyze non-figurative

21 drawings by Molly, an orangutan. The research utilizes VGG19 and ResNet18 models to

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

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

31 **1. Introduction**

32 The potential for anthropocentric bias emerges when deciphering the content of drawings, a
33 common focus in studies of figurative artwork. However, meanings also permeate non-
34 figurative sketches, as evidenced in creations by young children (Gardner 1981; Goodnow
35 2013; Restoy et al. 2022). Drawings are rich in information, yet reliance on a limited set of
36 manually selected features can restrict the depth of analysis. This is especially pertinent in
37 examining drawings by non-human primates, where an anthropocentric selection bias may
38 overlook features significant to other species (Saito et al. 2014; Martinet and Pelé 2021).

39 Consequently, such an approach fails to harness the full informational value of these
40 drawings. In human contexts, querying the artist about their intent offers a partial solution
41 (Martinet et al. 2021; Sueur et al. 2022), though this method falls short when artists cannot
42 verbally articulate their intentions, such as in the scribbles of young children or individuals
43 with communication impairments. This limitation extends to non-verbal non-human drawers
44 like non-human primates (Pelé et al. 2021; Martinet et al. 2023).

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

56 disease diagnosis (Zhou et al. 2021) without the need for predefined features. Convolutional
57 Neural Networks (CNNs), a prevalent deep learning tool in image analysis, autonomously
58 learn to identify relevant features, evolving from simple shapes to complex objects through
59 layers. While CNNs enhance accuracy, they struggle with interpretability, prompting the
60 development of techniques to demystify model decisions (Zhang and Zhu 2018; Carabantes
61 2020).

62 Artificial intelligence's application to drawing behavior remains very little explored (Beltzung
63 et al. 2023), despite its success in classification tasks, like identifying stroke patterns (Wu et
64 al. 2018) or categorizing drawings by object (Zhang et al. 2016). Our prior study in 2022
65 (Beltzung et al. 2022) harnessed AI to examine seasonal trends in Molly's artwork, utilizing
66 the VGG19 (Dutta et al. 2016) model for seasonal classification, achieving a 41.6% accuracy.
67 We explored how different features, from simple to complex, influenced these seasonal
68 variations, highlighting the roles of color and pattern. We found with deep learning models
69 similar results that we found using classical measures in ethology and drawing analyses. The
70 research underscored deep learning's potential to objectively analyze non-figurative art,
71 encouraging its application beyond primates to include human toddler scribbles. However, it
72 was impossible to open the black-box and to fully decipher and interpret the mechanisms
73 through which the deep learning model identifies and classifies elements within the drawings.
74 Here, we used interpretability and explicability models to reach this aim (Zhang and Zhu
75 2018; Spannaus et al. 2023). Interpretability in deep learning refers to the ability to
76 understand and explain the decision-making process of deep neural networks. Deep learning
77 models are often considered black boxes because they lack transparency and it is difficult to
78 comprehend how they arrive at their predictions (Shwartz-Ziv and Tishby 2017; Carabantes
79 2020). Captum (Kokhlikyan et al. 2020) is a novel, unified, open-source model
80 interpretability library for PyTorch. It contains implementations of various gradient and

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

97 **2. Materials and Methods**

98 *2.1. Dataset*

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

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

113 2.2. Convolutional Neural Network (CNN)

114 Acknowledging that human-selected features might not capture the intricate details in
115 drawings, our study focused on deep learning, particularly convolutional neural networks
116 (CNNs). Initially, we trained CNNs, specifically VGG19 (Beltzung et al. 2022) and ResNet
117 18 (Liang 2020), to predict the seasons depicted in drawings by Molly. To address the
118 limitations of our dataset's size, we utilized transfer learning, repurposing models pre-trained
119 on different tasks. This method enabled us to leverage the capabilities of ResNet18 and
120 VGG19, both pre-trained with ImageNet weights, enhancing our analysis. ResNet18 is
121 renowned for its efficiency and depth, achieved through residual learning, which we adapted
122 for our study. This model's architecture addresses the vanishing gradient problem with
123 shortcut connections that bypass layers, facilitating the training of deeper networks. We
124 applied this approach to both models, adjusting various parameters and hyperparameters, such
125 as learning rate and the configuration of fully connected layers, to find the optimal setup.
126 Each model's final layer was a fully connected layer with four neurons, each representing a
127 season, using softmax activation for classification. We employed a categorical cross-entropy
128 loss function and optimized the models using stochastic gradient descent (SGD) with a

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

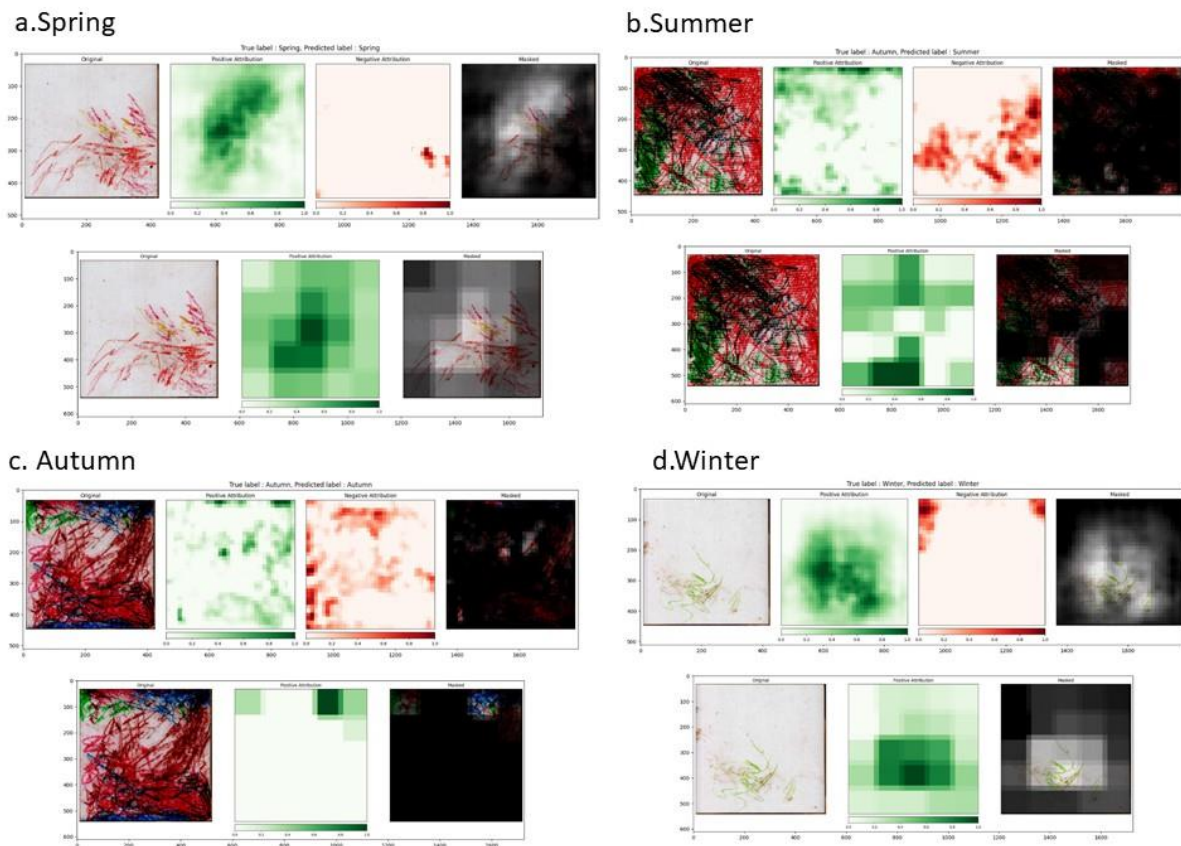
133 *2.3. Interpretability with Captum.*

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

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

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

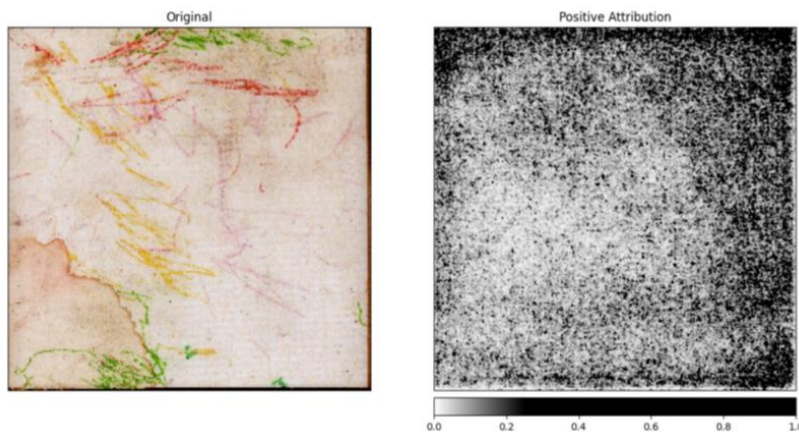
152 patterns or features that were consistently influential across Molly’s drawings, providing a
153 quantitative measure of the visual elements most associated with each season (figure 1).



154
155 Figure 1: Seasonal Examples of Occlusion Analysis Using Captum Across Various Pixel
156 Sizes. For each season, the sequence displays: the original drawing (top), followed by positive
157 attribution (in green) highlighting pixels crucial for identification and classification, negative
158 attribution (in red) indicating non-essential pixels, and finally, the differential attribution
159 showcasing the contrast between positive and negative contributions. The top row represents a
160 threshold (t) of 0, and the bottom row a threshold (t) of 0.5 for each depicted example.

161
162 Captum – Integrated Gradients: We employed Integrated Gradients (Kwon et al. 2021),
163 another interpretability technique from Captum, to attribute the prediction of the model to its
164 input features (figure 2). This method offers a way to visualize the importance of each pixel in

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



169

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

173

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

178 Application of PCA / t-SNE in Scikit learn: With the linear layer removed and raw features
179 extracted, we applied Principal Component Analysis (PCA) (Holland 2008) and t-Distributed
180 Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton 2008; Wattenberg et al.
181 2016) for dimensionality reduction. These techniques reduced the high-dimensional feature
182 space into a two- or three-dimensional space, making it possible to visualize the distribution

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

186 Clustering (Louvain Algorithm): Finally, we applied the Louvain algorithm (Combe et al.
187 2015; Emmons et al. 2016) for clustering the drawings based on their reduced-dimensional
188 features. The Louvain algorithm is a community detection method known for its efficiency in
189 large networks. By clustering the drawings, we aimed to discover natural groupings within the
190 data, which could indicate distinct styles, motifs, or themes recurring across different seasons.
191 This unsupervised learning approach provided a bottom-up perspective on the dataset,
192 potentially uncovering new insights into Molly's drawing behavior and how it varied with the
193 seasons.

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

197

198 *2.4. Statistical analyses*

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

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

207 determine if there were statistically significant differences across the seasons. The analysis
208 was performed using the R statistical software, with a significance level set at $\alpha = 0.05$.
209 Image analyses are available at <https://doi.org/10.5281/zenodo.10973649>. All codes are
210 available at <https://github.com/cedricsueur/drawinganalyses>

211 **3. Results**

212 *3.1. CNNs: VGG19 and ResNet18.*

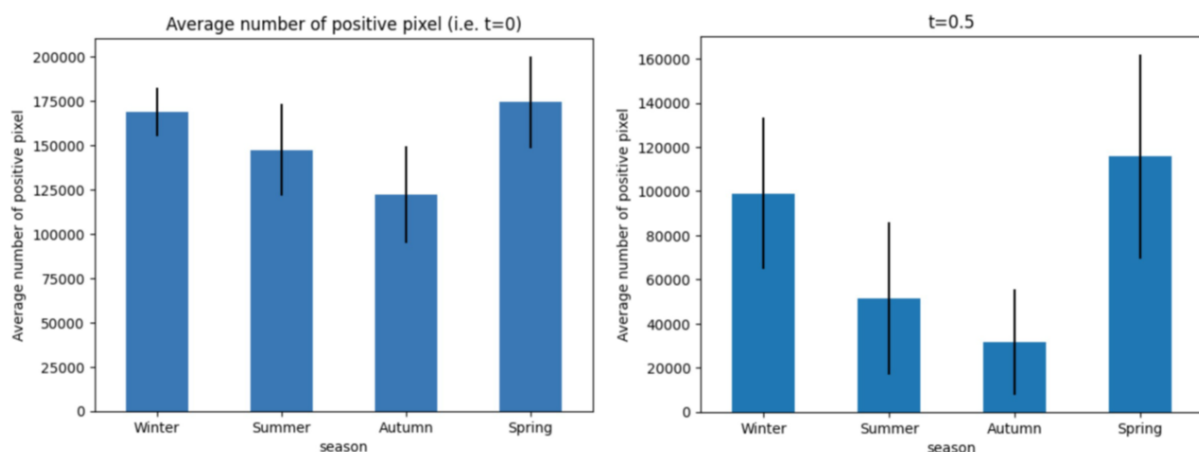
213 The VGG19 model trained to classify drawings according to seasons achieved 42% accuracy
214 on the test set. The ResNet18 reached an accuracy a bit higher of 50%. These accuracies are
215 higher than that expected by random (approximately 29%, by always classifying drawings as
216 the most common class). To provide further context, the same models were also tasked with
217 differentiating between drawings by Molly and humans drawings. In this classification
218 challenge, both models demonstrated a remarkable accuracy higher than 93%. Additionally,
219 when the model was retrained for a binary classification task—distinguishing between
220 drawings with low and high coverage by Molly—it achieved an even accuracy higher than
221 95%. These outcomes suggest that the model is highly capable of distinguishing between
222 drawings when the distinctions are pronounced. Consequently, the relatively modest accuracy
223 in the seasonal classification task likely stems from the more nuanced differences present
224 among Molly's drawings across different seasons.

225 *3.2. Captum Occlusion and integrated gradients*

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

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

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



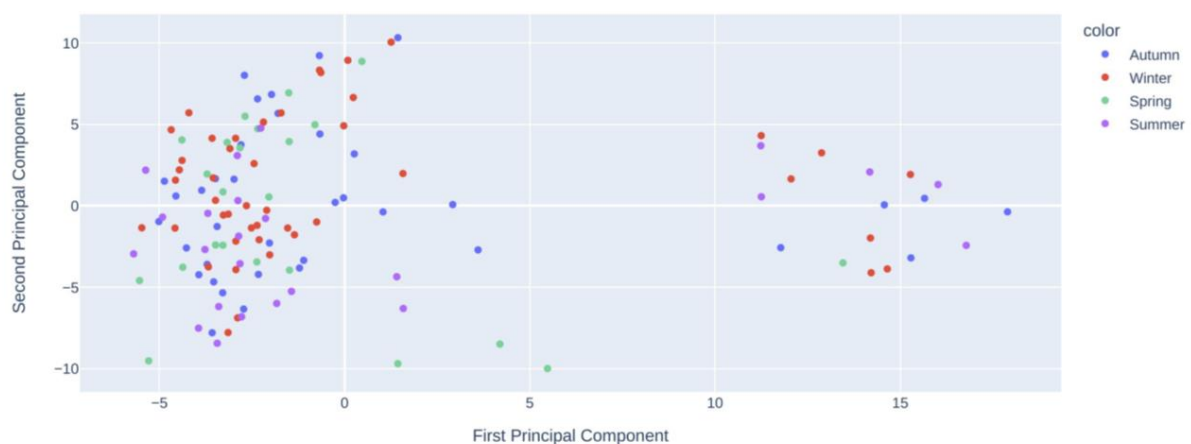
246 Figure 3: Comparative Analysis of Positive Attribution Pixel Counts. This graph presents the
247 average number of pixels significant for seasonal drawing classification at two thresholds: $t=0$
248 (indicating smaller pixels) and $t=0.5$ (representing larger pixels), highlighting the role of pixel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

250

251 3.3. PCA and Louvain Clustering

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

262



263

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

267

268 **4. Discussion**

269 In synthesizing insights from our deep learning-based analysis of non-human drawings,
270 particularly those by Molly the orangutan, with discussions on the interpretability of these
271 models, several key themes and conclusions emerge that bridge the gap between traditional
272 drawing analysis and advanced AI methodologies. The use of Captum (Kokhlikyan et al.
273 2020) specifically for the analysis of drawings and artwork appears to be a novel application.
274 While interpretability tools have been increasingly applied in various domains to understand
275 the decision-making processes of AI models—ranging from healthcare and finance to
276 autonomous vehicles and natural language processing—their application within the realm of
277 art analysis, especially in studying non-human drawings, marks a pioneering step. As Nagel
278 pointed out (Nagel 1980), it's beyond human capability to fully understand the experience of
279 being another animal. However, we believe that artificial intelligence can play a role in
280 reducing the prevalence of biases.

281 *4.1. Deep Learning Models: VGG19 and ResNet18*

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

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

294 *4.2. Interpretability: Captum Occlusion and Integrated Gradients*

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

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

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

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

316 *4.3. PCA and Louvain Clustering*

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

324 *4.4. Implications and Future Directions*

325 This study with the previous ones illustrates the complementary strengths of traditional
326 drawing analysis and AI-driven methodologies as made in other domains (Soto and Adey
327 2016; Lu et al. 2024). Where traditional analysis provides a framework for understanding the
328 thematic and stylistic components of drawings, deep learning models offer a means to
329 systematically and objectively analyze these components across large datasets. The
330 intersection of these approaches, facilitated by interpretability tools like Captum, holds the
331 promise of enriching our understanding of non-human art, offering nuanced insights that
332 neither approach could achieve in isolation. However, it is needed to think and to build a
333 framework to know how to work in complementary with traditional approaches and AI-driven
334 approaches.

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

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

348

349 **Funding**

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

354

355 **Institutional Review Board Statement**

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

358

359 **Informed Consent Statement**

360 Not applicable.

361

362 **Data Availability Statement**

363 Dataset is available at <https://doi.org/10.5281/zenodo.10973649>. All codes are available at

364 <https://github.com/cedricsueur/drawinganalyses>

365

366 **Acknowledgments**

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

368

369 **Conflicts of Interest**

370 The authors declare no conflict of interest. The funders had no role in the study design;

371 collection, analyses, or interpretation of data; writing of the manuscript; or decision to publish

372 the results.

373

374 **References**

375 Beltzung B, Pelé M, Renoult JP, et al (2022) Using Artificial Intelligence to Analyze Non-

376 Human Drawings: A First Step with Orangutan Productions. *Animals* 12:2761.

377 <https://doi.org/10.3390/ani12202761>

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

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

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

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

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

389 Dutta A, Gupta A, Zissermann A (2016) VGG image annotator (VIA). URL [Httpwww Robots](http://www.robots.ox.ac.uk/vggsoftware/via)
390 [Ox Ac Uk Vggsoftware/via](http://www.robots.ox.ac.uk/vggsoftware/via)

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

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

395 Gardner H (1981) *Artful scribbles: The significance of children’s drawings*

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

398 Goodnow J (2013) *Children drawing*. Harvard University Press

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

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

404 Holland SM (2008) Principal components analysis (PCA). *Dep Geol Univ Ga Athens GA*
405 30602–2501

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

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

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

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

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

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

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

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

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

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

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

430 Martinet L, Sueur C, Matsuzawa T, et al (2023) Tool assisted task on touchscreen: a case
431 study on drawing behaviour in chimpanzees (*Pan troglodytes*). *Folia Primatol (Basel)*
432 1:1–17

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

463 Zhang H, Liu S, Zhang C, et al (2016) Sketchnet: Sketch classification with web images. pp
464 1105–1113

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

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

470

471

472

473

474 Figure Captions

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

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

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

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

491