Using LLMs to explain AI-generated art classification via Grad-CAM heatmaps

Giovanna Castellano, Maria Grazia Miccoli, Raffaele Scaringi, Gennaro Vessio and Gianluca Zaza*

Department of Computer Science, University of Bari Aldo Moro, Italy

Abstract

The proliferation of AI-generated media, especially in art, has sparked interest in creating models that differentiate between original and AI-generated artworks. However, understanding why these models make certain decisions remains a significant challenge. This paper enhances the explainability of Vision Transformer-based classification models by using Grad-CAM to generate visual explanations of the model's focus areas, combined with Large Language Models (LLMs) to provide natural language descriptions. We evaluate three cutting-edge LLMs—LLaVa-NeXt, InstructBLIP, and KOSMOS-2—by using them to generate textual explanations for Grad-CAM visualizations applied to artwork classification. Through quantitative and qualitative analyses, we find that while InstructBLIP and KOSMOS-2 achieve higher similarity scores between generated descriptions and visual content, LLaVa-NeXt provides more insightful and coherent explanations, particularly for AI-generated art. This study demonstrates the potential of LLMs to improve the interpretability of AI decisions in complex image classification tasks, helping to bridge the gap between model decisions and human understanding in art classification.

Keywords

Explainable AI, Large Language Models, Grad-CAM, AI-generated art, Artwork classification

1. Introduction

Artificial Intelligence (AI) has achieved remarkable advancements in today's digital age, particularly in creating synthetic media. Generative models, such as GANs (Generative Adversarial Networks) and diffusion models, can produce highly realistic images, videos, and artworks, making it increasingly difficult to distinguish between AI-generated and human-created content. This growing challenge is especially critical in the domain of art, where concepts of creativity, authorship, and authenticity are deeply rooted in human expression. The ability to accurately classify and explain AI-generated versus original artworks is therefore essential for preserving the integrity of human creativity and safeguarding intellectual property.

A key concern related to AI-generated art is its potential to blur the boundaries between real and synthetic content, raising questions about originality and ownership. AI-generated media can disrupt traditional notions of being an artist as machines begin to emulate complex artistic styles and compositions with unprecedented fidelity. Beyond artistic expression, the rise of manipulated media, such as deepfakes, has further complicated the landscape by enabling the creation of highly realistic yet artificial videos and images [2]. These technologies, often indistinguishable to the human eye, pose ethical and legal challenges, particularly in cases of media manipulation and copyright infringement [3].

In response to these challenges, deep learning models have been developed to automatically classify artworks as either original (human-created) or AI-generated. These models typically leverage sophisticated neural architectures, including Convolutional Neural Networks (CNNs) and Transformer-based models, to perform classification tasks with high accuracy. For example, in a previous work [4], we

XAI.it - 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024 [1]

*Corresponding author.

giovanna.castellano@uniba.it (G. Castellano); raffaele.scaringi@uniba.it (R. Scaringi); gennaro.vessio@uniba.it (G. Vessio); gianluca.zaza@uniba.it (G. Zaza)

0000-0002-6489-8628 (G. Castellano); 0000-0001-7512-7661 (R. Scaringi); 0000-0002-0883-2691 (G. Vessio); 0000-0003-3272-9739 (G. Zaza)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3839/paper5.pdf

demonstrated the effectiveness of deep learning models, such as Vision Transformers (ViTs) [5], in distinguishing between human-created and AI-generated art. However, while these models achieve impressive performance, their decision-making processes often remain opaque, limiting user trust and understanding.

Explainable AI (XAI) methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM) [6], have been developed to address this opacity. Grad-CAM generates visual heatmaps that highlight the regions of an image that most influence a model's decision, offering a degree of interpretability by showing users where the model "looked" to make its classification. However, while Grad-CAM provides valuable visual insights, it may not be sufficient for non-expert users who need a more explicit explanation of why certain regions were emphasized in the decision-making process.

To bridge this gap, recent advancements in natural language processing (NLP), mainly through Large Language Models (LLMs), offer an opportunity to enhance the interpretability of these visual explanations. LLMs are designed to generate natural language explanations describing complex visual information. By integrating LLMs with Grad-CAM outputs, it is possible to generate textual descriptions that explain why certain areas of an artwork were highlighted during classification, improving the transparency and interpretability of AI models.

This paper proposes a framework combining Grad-CAM visualizations with advanced LLMs to enhance the explainability of deep learning models in artwork classification. Specifically, we evaluate the performance of three cutting-edge LLMs—LLaVa-NeXt [7], InstructBLIP [8], and KOSMOS-2 [9]—in generating coherent and insightful natural language explanations for classifying original versus AI-generated artworks. Through quantitative and qualitative analyses, we assess how well these models can generate meaningful explanations that align with the visual heatmaps produced by Grad-CAM.

The rest of this paper is organized as follows. Section 2 reviews related works. Section 3 details our methodology, including the integration of Grad-CAM with LLMs for the explainability of artwork classification models. Section 4 presents the experimental results. Finally, Section 5 concludes with a summary of the results and future directions for research in this area.

2. Related work

This section reviews significant research contributions in three key areas: AI-generated image classification, explainability in AI models using Grad-CAM, and leveraging LLMs to generate textual explanations for improving model interpretability.

2.1. AI-generated image classification

Differentiating AI-generated images from human-created ones has garnered increasing attention in recent times. Martin-Rodriguez et al. [10] proposed methods based on pixel-level feature extraction, such as Photo Response Non-Uniformity (PRNU) and Error Level Analysis (ELA), to train CNNs for the classification of AI-generated images versus real photographs. Similarly, Epstein et al. [11] explored real-time detection of AI-generated images using advanced neural network architectures, emphasizing the importance of fast and accurate identification in fighting synthetic content.

In the domain of art, few works have tackled this classification task. Ha et al. [12] analyzed the distinguishing features between AI-generated and human-created artworks using deep learning models. They demonstrated how neural networks could differentiate these categories based on creative elements like style and composition. More recently, we evaluated the performance of deep learning models, including VGG-19 [13], ResNet-50 [14], and Vision Transformers [5], in classifying AI-generated artworks [4], reporting classification accuracies of up to 97%. In this study, we build on our previous ViT model, which achieved the highest performance, while enhancing the interpretability of the classification process using explainability techniques.

2.2. Explainability of AI models using Grad-CAM

The need for interpretability in AI models has led to the adoption of XAI techniques, with Grad-CAM [6] being one of the most popular methods. Grad-CAM generates heatmaps highlighting image regions that influence a model's decision-making process. These visualizations are invaluable for improving the transparency of AI systems.

In the context of artwork classification, we applied Grad-CAM to visualize the decision-making processes of deep learning models, demonstrating how these heatmaps could enhance model interpretability [4]. However, while these heatmaps provide insight into where the model is "looking", they may still be challenging for non-expert users to interpret without additional guidance. Distinguishing between AI-generated and human-created art requires a nuanced understanding. Improving models' interpretability is critical in domains where creative ownership and authenticity are essential, making this a viable use case for the study of explainability.

2.3. Leveraging LLMs for explainability

Recent advances in NLP have led to the development of LLMs capable of generating coherent and detailed textual descriptions based on visual inputs. Yang et al. [15] explored using LLMs to detect sophisticated image tampering, demonstrating that advanced models could accurately identify subtle manipulations in AI-generated content. However, they also noted that current LLMs struggle with highly realistic AI-generated images, underscoring the need for further improvements. In response to these challenges, Samesh et al. [16] combined CNNs with multimodal fusion techniques to detect advanced deepfakes, integrating LLMs to enhance accuracy. Their findings highlight the potential of LLMs to offer detailed explanations that improve model transparency.

For our study, we selected three advanced LLMs to generate explanations for AI-driven artwork classification: LLaVa-NeXt [7], InstructBLIB [8], and KOSMOS-2 [9]. These models, designed to process visual and textual inputs, have shown promise in generating human-understandable explanations of complex AI decisions.

3. Methods

This section outlines the proposed framework, illustrated in Fig. 1, which integrates ViTs with Grad-CAM and LLMs to enhance the interpretability of AI-generated artwork classification.

3.1. Proposed framework

The proposed framework aims to provide human-understandable explanations of how a deep learning model classifies artworks as either original or AI-generated. The pipeline consists of three key stages:

- A ViT model [5] is first trained to classify artworks. The input consists of RGB images of both original and AI-generated artworks. We selected this model because it achieves 97% accuracy on the same dataset used in this study for this classification task.
- During the inference stage, the Grad-CAM technique [6] is applied to generate heatmaps that highlight regions of the image that most strongly influenced the ViT model's classification decision. These heatmaps provide visual explanations of the areas of the artwork that were most important for the model's decision-making process.
- To generate natural language explanations of the Grad-CAM heatmaps, three advanced LLMs—LLaVa-NeXt [7], InstructBLIB [8], and KOSMOS-2 [9]—are integrated into the pipeline. Each LLM receives the Grad-CAM overlay image and a specially designed prompt to generate a description explaining the model's focus on some areas of the artwork.

This combined approach offers visual and textual explanations, ensuring that non-expert users can better understand the classification decisions made by the AI model.

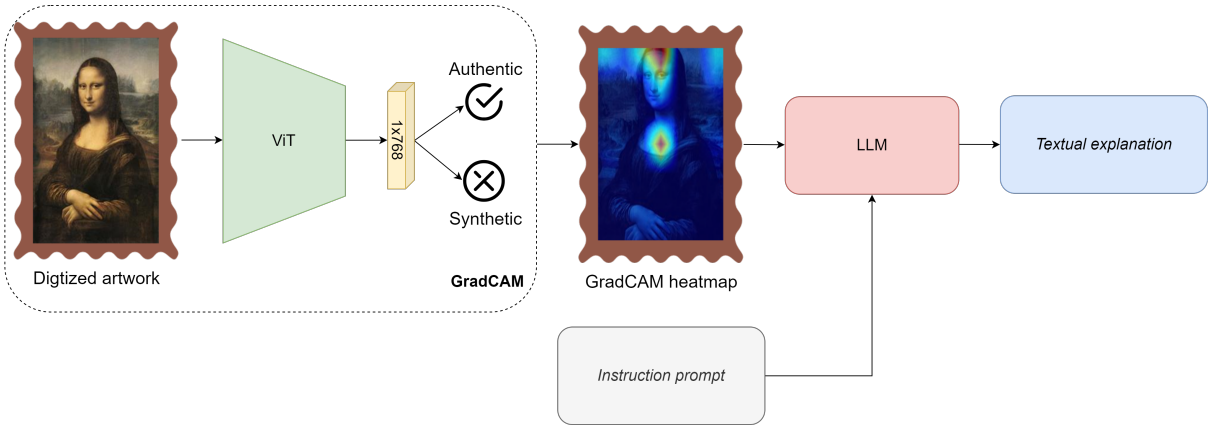


Figure 1: Proposed framework. A ViT model for classification processes an artwork. Grad-CAM is then applied to identify influential regions. Lastly, the overlayed image is fed to an LLM (LLaVa-NeXt, InstructBLIP, or KOSMOS-2), generating a human-understandable explanation.

3.2. Implementation details

The ViT model used in this study is based on the work of Dosovitskiy et al. [5]. It processes the artwork as a sequence of image patches and uses multi-headed self-attention to capture global information. We employ ViT to classify images into two categories, namely original or AI-generated. Specifically, we used ViT-B/16, pre-trained on the ImageNet dataset [17], which takes as input RGB images with a resolution of 224×224 pixels. Then, we finetuned for 30 epochs the last layer, training the model to recognize whether or not the input artwork is original or AI-generated. It is worth noting that during the training stage, we optimized a binary cross-entropy loss using the well-known Adam optimizer, with an initial learning rate of 10^{-3} and a step learning rate scheduling every seven epochs with a decay factor $\gamma = 0.1$. Furthermore, an early stopping mechanism was employed, halting after three epochs with no decrease in validation loss.

Once the ViT model is trained, Grad-CAM [6] is applied to visualize which parts of the input image influence the model's decision. Specifically, the gradients of the output class score are computed with respect to the token embeddings (representing image patches) from the final layers. These gradients are used to weigh the importance of each token, and a heatmap is generated that highlights the most influential image regions for the classification.

Lastly, we overlay the input image with the Grad-CAM heatmap. We employed three advanced LLMs to generate textual explanations based on that, each with a tailored prompt iteratively refined through multiple attempts to improve the clarity and relevance of the generated explanations. Specifically, the prompt used in this work was developed following the main guidelines of prompt engineering to ensure effective communication with the language model. In particular, five basic rules were applied:

1. Clarity of input: The model was provided with a description of the input type, specifying that it was an artwork with a superimposed heatmap.
2. Context: The context in which the input was located was provided, explaining that the heatmap indicated the areas of the artwork that the classifier considered most relevant to its decision.
3. Objective of the output: The type of output required was explicitly stated, namely, an analysis of the possible causes that led the classifier to identify those specific areas.
4. Output format: Clear instructions were given on the expected format of the output.
5. Example of output: An example was also provided to guide the model in generating the desired response.

The three LLMs employed are:

- LLaVa-NeXt [7], an evolution of LLaVA-1.5, supports image resolutions up to 672×672 pixels and enhances visual reasoning and OCR capabilities. We employed the quantized version of this model.

- InstructBLIP [8], built on the BLIP-2 architecture, specializes in zero-shot vision-language tasks. Key hyperparameters include `num_beams=5` (for beam search decoding), `max_new_tokens=250`, `min_length=1`, `top_p=0.9` (to maintain diversity), and `temperature=1` (to balance randomness). These parameters ensure coherent and concise descriptions are generated efficiently.
- KOSMOS-2 [9] is a multimodal Transformer-based model designed for visual-textual grounding. Key hyperparameters, such as `max_new_tokens=1024` and attention masks, control the generation process, ensuring the output is aligned with the visual input.

3.3. Evaluation

We employ quantitative and qualitative analyses to evaluate the quality of the generated explanations. The quantitative metrics measure aspects that may not always capture human understandability or insightfulness. The qualitative evaluations, instead, may provide deeper insights into the coherence and relevance of explanations, which the metrics might miss.

We measure the alignment between the generated descriptions and the visual content using two metrics:

- Image-to-text similarity: Using CLIP [18], we compute the cosine similarity between the image (with Grad-CAM overlay) and the generated textual description. A higher score indicates that the text better reflects the image content.
- Text-to-label similarity: We assess the consistency between the generated text and the classification label (original or AI-generated) using the S-BERT model [19] by computing the cosine similarity between the embedding of the generated explanation, and the description of the target label.

In the qualitative analysis, we manually examine the coherence, relevance, and insightfulness of the explanations generated by each LLM. This involves comparing the descriptions with the Grad-CAM heatmaps and assessing whether the explanations provide meaningful insights into the model's decision-making process.

4. Experiments

In this section, we compare the performance of the selected LLMs—LLaVa-NeXt, InstructBLIB, and KOSMOS-2—in improving the interpretability of AI models for artwork classification. The experiments consisted of two primary analyses: quantitative evaluation using similarity metrics and qualitative evaluation of the model's ability to generate coherent and relevant descriptions for selected images.

4.1. Materials

To evaluate the proposed method, we used a subset of 100 images, equally divided between 50 AI-generated artworks and 50 original artworks, drawn from the dataset described in [4]. This dataset combines elements from *ArtGraph* [20] and *ArtiFact* [21]. *ArtGraph* is a specialized knowledge graph with 116,475 artworks classified across 32 styles and 18 genres. *ArtiFact* is a large-scale dataset with 2,496,738 images, including authentic and fake images from various domains such as art, human faces, and vehicles.

Each image in our subset included the original artwork and a corresponding Grad-CAM overlay generated during classification. The experiments were run on the Google Colab platform, utilizing an Intel Xeon processor, 12 GB RAM, and an NVIDIA T4 GPU with 15 GB VRAM.

4.2. Quantitative analysis

For the quantitative analysis, we measured the similarity between the generated descriptions, images, and labels to assess the models' ability to provide relevant textual explanations.

Table 1
Similarity metrics between image-to-text, text-to-label, and the total similarity for the three LLMs.

Model	Image-to-text	Text-to-label	Total similarity
LLaVa-NeXt	0.22	0.14	0.36
InstructBLIB	0.26	0.14	0.40
KOSMOS-2	0.26	0.14	0.40

Table 1 presents the results of this analysis, showing the similarity metrics for each LLM. The results show that InstructBLIB and KOSMOS-2 achieve the highest overall similarity scores (0.40), with LLaVa-NeXt scoring slightly lower (0.36). All models perform similarly in terms of text-to-label similarity (0.14). Regarding image-to-text similarity, InstructBLIB and KOSMOS-2 score 0.26, while LLaVa-NeXt reaches 0.22.

4.3. Qualitative analysis

In addition to the quantitative analysis, we conducted a qualitative evaluation using four sample images, two AI-generated and two original artworks (Fig. 2). All images were correctly classified by ViT. The goal was to assess the ability of each LLM to generate insightful and relevant explanations for the classification of these images.

For AI-generated artworks, LLaVa-NeXt accurately identifies the relevant areas of interest and provides plausible explanations for the network's focus. However, for original artworks, it occasionally struggles to account for the model's final classification, leading to less coherent explanations. InstructBLIP identifies regions of interest but often misinterprets the heatmap as a thermal image, particularly in cases of AI-generated artworks. This results in inaccurate descriptions that fail to explain the classification decision meaningfully. KOSMOS-2 performs relatively well in identifying relevant areas for both AI-generated and original artworks, but its explanations are vague and often lack depth, leaving some interpretability issues unresolved.

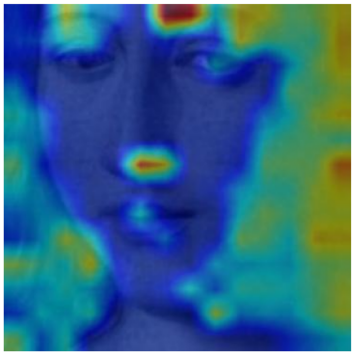
The qualitative analysis reveals that LLaVa-NeXt generally produces more intelligent and abstract explanations, particularly for AI-generated images, despite some inaccuracies with original artworks. In contrast, InstructBLIP and KOSMOS-2, while performing well in the quantitative analysis, often produce literal descriptions that fail to provide meaningful insights into the classification decision.

5. Conclusion

In this work, we aimed to enhance the interpretability of Transformer-based models for classifying artworks as original or AI-generated by integrating Grad-CAM heatmaps with Large Language Models. We evaluated the performance of three state-of-the-art LLMs—LLaVa-NeXt, InstructBLIP, and KOSMOS-2—on their ability to generate coherent and meaningful descriptions based on Grad-CAM visualizations applied to both synthetic and original artworks.

While LLaVa-NeXt outperformed the other models in generating meaningful explanations for AI-generated content, it requires further refinement to improve its accuracy when handling original artworks. Addressing this limitation would increase its versatility across a broader range of image classification tasks. InstructBLIP, on the other hand, struggled with interpreting Grad-CAM heatmaps, often providing overly literal descriptions. Future work could focus on refining its training strategies or prompt engineering to align its outputs more closely with the visual features it is meant to explain. KOSMOS-2 showed promise with its strong multimodal grounding but lacked the depth of insight necessary for robust interpretability. Further research could enhance its ability to abstract complex visual inputs into more meaningful explanations.

Ultimately, combining advanced LLMs with explainability techniques like Grad-CAM can significantly improve the transparency and trustworthiness of AI-driven image classification systems. Fine-tuning



LLaVa-NeXt

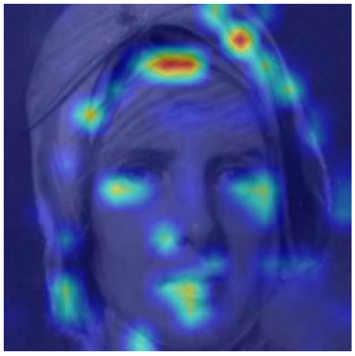
The activation map generated by the network focuses on certain areas of the work, such as the face and hair color. This suggests that the network focused on these areas to classify the work, as these areas are likely to be the parts of the work that have the greatest influence on the classification. The network may have analyzed the lines and colors to identify whether the work was artificially generated or whether it was created from an original work of art.

InstructBLIP

The activation map presents those warm colors in that area of the artwork due to the use of thermal imaging technology. Thermal imaging, also known as infrared thermography, is a non-destructive testing technique that uses infrared cameras to detect and visualize thermal energy emitted by objects. In this case, the activation map likely represents the temperature distribution of the artwork's surface, with the warm colors indicating higher temperatures and the cooler colors indicating lower temperatures. This could be caused by various factors, such as the presence of hotspots on the artwork's surface, uneven heating or cooling due to environmental factors, or even the use of certain materials that have different thermal properties.

KOSMOS-2

The classifier might have focused on the areas of the face that are most likely to be associated with emotion, such as the eyes, nose, mouth, or lips. The activation map highlights these areas, which could be the result of the classifying algorithm's attention to these features.



LLaVa-NeXt

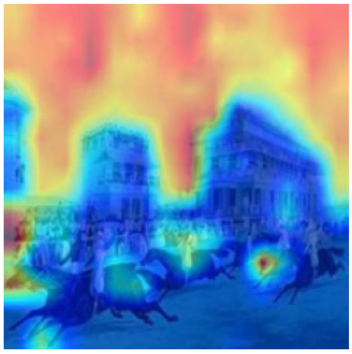
The activation map generated by the network focuses on some specific areas of the work, such as the profile of the face and the lines that define the shape of the head. These areas are likely relevant to the classification of the work, as they may be unique features that the network used to distinguish the original artwork from an artificially generated one.

InstructBLIP

The activation map presents those warm colors in that area of the artwork due to the presence of a heat signature. Heat signatures, also known as thermograms, are generated by the infrared radiation emitted by an object or a person's body. In this case, the activation map is likely capturing the heat signature of the woman's face, resulting in the display of warm colors in the affected area. This could be caused by various factors, such as the intensity of the infrared radiation, the distance between the camera and the subject, or the type of clothing or accessories worn by the subject.

KOSMOS-2

The classifier might have focused on the areas of the face that are most likely to be associated with emotion, such as the eyes, nose, and mouth. The classification could have been based on the artist's facial expression, the way they hold their head, or the way the artwork is positioned in the image.



LLaVa-NeXt

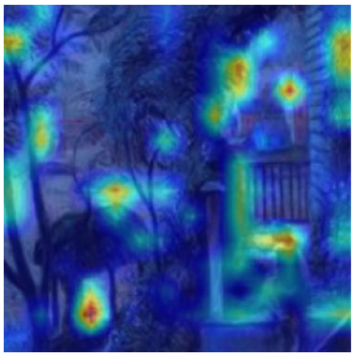
The activation map by the network focuses on some parts of the work, such as these that define the architecture of the buildings and lines that follow the contour of the buildings. The specific lines are then worked further by the generator for classification, as they may be unique features that distinguish the original work from an artificially generated network.

InstructBLIP

The abstract map presents some colors that are in contrast due to their use of thermal imaging technology. Thermal imaging captures radiation based on the intensity of infrared radiation emitted by objects, which is proportional to their temperature. In this particular artwork, thermal readings generate in sequence warm colors in terms of a figure capturing through rays resulting in a colorful and vibrant representation. The capture buttons and activation map correspond to high temperatures or clusters that yield areas of intense heat such as during building fires. This technique allows for realistic visual representation through detection of events, while also highlighting intense heat generated by these events at any given time throughout the day.

KOSMOS-2

The designer might have focused on how there is imagery at extreme ends such as paintings depicting historic events depicted in dynamic presence while others would have portrayed quiet desolation instead. Therefore, these risks are located within the city, and presence findings behind each image might have attracted attention additional answers elaboration on each one with discovery mannerism, adding aesthetic aspects but also hazard.



LLaVa-NeXt

The activation map generated by the network focuses on some areas of the work, such as the sky, the leaves of the trees and the lines that define the contours of the objects. These areas were probably chosen because they contain visual information that the network used to classify the work as "AI Generated". The lines and colors in the work may have been artificially generated to create an image that appears original, but was actually generated by a classification network.

InstructBLIP

The activation map presents those warm colors in that area of the artwork due to the presence of heat signatures. Heat signatures, also known as thermal imaging or thermography, are generated by the infrared radiation emitted by objects at different temperatures. In this case, the activation map is likely capturing the heat signatures emitted by the trees and bushes in the image, resulting in the display of warm colors in the affected area.

KOSMOS-2

The classifier might have focused on the warm colors in the artwork, such as the yellow and orange spots, as they are associated with warmth and happiness. The class might have also focused on other elements of interest, such as the tree, the house, or the person, as these elements might evoke positive emotions or evoke a sense of connection with nature.

Figure 2: Examples of artworks with overlaid Grad-CAM and corresponding explanations generated by the tested LLMs. The two artworks at the top are AI-generated, while the ones at the bottom are authentic.

and expanding the capabilities of these models will be critical to enhancing interpretability and fostering wider adoption, especially in other domains where explainability is essential.

Acknowledgments

G. Castellano and G. Zaza acknowledge funding support from the FAIR - Future AI Research (PE00000013) project, Spoke 6 - Symbiotic AI (CUP H97G22000210007), under the NRRP MUR program funded by NextGenerationEU. R. Scaringi's research is funded through a Ph.D. fellowship within the Italian "D.M. n. 352, April 9, 2022" program, under the NRRP, Mission 4, Component 2, Investment 3.3, with the Ph.D. project titled "Automatic analysis of artistic heritage via Artificial Intelligence," co-supported by Exprivia S.p.A. (CUP H91I22000410007).

References

[1] M. Polignano, C. Musto, R. Pellungrini, E. Purificato, G. Semeraro, M. Setzu, XAI.it 2024: An Overview on the Future of Explainable AI in the era of Large Language Models, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.

[2] G. Wu, W. Wu, X. Liu, K. Xu, T. Wan, W. Wang, Cheap-fake Detection with LLM using Prompt Engineering, in: 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), IEEE, 2023, pp. 105–109.

[3] S. A. Yang, A. H. Zhang, Generative AI and copyright: A dynamic perspective, arXiv preprint arXiv:2402.17801 (2024).

[4] T. Bianco, G. Castellano, R. Scaringi, G. Vessio, Identifying AI-Generated Art with Deep Learning, in: CREAI@ AI* IA, 2023, pp. 16–25.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, International Journal of Computer Vision 128 (2020) 336–359.

[7] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, Y. J. Lee, LlaVa-Next: Improved reasoning, OCR, and world knowledge, 2024.

[8] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26296–26306.

[9] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, F. Wei, KOSMOS-2: Grounding multimodal large language models to the world, arXiv preprint arXiv:2306.14824 (2023).

[10] F. Martin-Rodriguez, R. Garcia-Mojon, M. Fernandez-Barciela, Detection of AI-created images using pixel-wise feature extraction and convolutional neural networks, Sensors 23 (2023) 9037.

[11] D. C. Epstein, I. Jain, O. Wang, R. Zhang, Online detection of AI-generated images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 382–392.

[12] A. Y. J. Ha, J. Passananti, R. Bhaskar, S. Shan, R. Southen, H. Zheng, B. Y. Zhao, Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?, arXiv preprint arXiv:2402.03214 (2024).

[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[15] X. Yang, J. Zhou, Research about the Ability of LLM in the Tamper-Detection Area, arXiv preprint arXiv:2401.13504 (2024).

[16] S. E. VP, R. Dheepthi, et al., LLM-Enhanced Deepfake Detection: Dense CNN and Multi-Modal Fusion Framework for Precise Multimedia Authentication, in: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), IEEE, 2024, pp. 1–6.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[19] N. Reimers, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, arXiv preprint arXiv:1908.10084 (2019).

[20] G. Castellano, V. Digeno, G. Sansaro, G. Vessio, Leveraging Knowledge Graphs and Deep Learning for automatic art analysis, Knowledge-Based Systems 248 (2022) 108859.

[21] M. A. Rahman, B. Paul, N. H. Sarker, Z. I. A. Hakim, S. A. Fattah, ArtiFact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection, in: 2023 IEEE International Conference on Image Processing (ICIP), IEEE, 2023, pp. 2200–2204.