GRAD-CAM AND LLM:
UNVEILING THE SECRETS OF MODEL DECISION-MAKING.

Maria Grazia Miccoli
Department of Computer Science
University of Bari
Bari, Italy
m.miccoli45@studenti.uniba.it

July 5, 2024

ABSTRACT

This project focuses on improving the explicability of Transformer-based models used to classify images as "original" or "AI-generated", also using Grad-CAM heatmaps. We used a set of synthetic and original artwork data and evaluated three advanced large language models (LMMs): LLaVa-NeXt, InstructBLIP and KOSMOS-2. These models generated GradCAM-based descriptions applied to works of art, identifying factors that affect areas of interest in the network. Our goal was to determine how LLMs can improve the interpretability of AI models. We have conducted both quantitative and qualitative evaluation. In qualitative analysis, LLaVa-NeXt excelled in accurately identifying areas of interest and providing consistent reasons for the network's attention, especially for AI-generated artworks. However, he showed some inaccuracies with original artwork. InstructBLIP struggled in both scenarios, misinterpreting the heat map as an image of the thermal sensor. Kosmos-2 identified areas of interest but provided vague descriptions and sometimes misunderstood the task. Quantitative results showed that InstructBLIP and Kosmos-2 had higher similarity scores because of their literal descriptions, but LLaVa-NeXt was qualitatively better at providing intelligent and consistent explanations. This assessment demonstrates the potential of LLMs to improve the explicability of image classification models.

1 Introduction

The digital age is saturated with multimedia content, blurring the lines between reality and artificiality at an accelerating pace. This trend is largely driven by the latest advancements in AI and the emergence of generative models far more powerful than their predecessors. These models are capable of producing remarkably realistic synthetic content, raising worries about the authenticity of online information.
This issue impacts various facets of society, particularly multimedia content such as images and videos, where synthetic media can wield significant influence. As elucidated by Wu et al.[1] in their article, two notable examples of manipulated media are Deep-fakes and Cheap-fakes, both posing considerable risks. Deep-fakes employ advanced machine learning techniques for video manipulation, whereas Cheap-fakes utilize readily available tools ,like Photoshop, for simpler forms of manipulation such as image editing, caption alterations, or video speed adjustments. While Deep-fakes are often admired for their technical sophistication, the widespread accessibility and deceptive deployment of Cheap-fakes present a more pervasive threat. Their ease of creation and ability to deceive by using unaltered images in misleading contexts make them particularly hazardous. Detecting erroneous information based on manipulated images remains challenging, as the visual content itself remains unaltered.
Furthermore, a pressing issue concerns safeguarding the copyright of artworks originating solely from human creativity [2], and enhancing these protections with AI assistance. Additionally, distinguishing original artworks from synthetic artwork is another consequential outcome of AI's application in the art.

So, the challenge of distinguishing authentic from synthetic is pervasive in today's society. In the context of art, significant efforts have been made to deploy deep neural network models trained specifically to differentiate artworks created solely by human beings from those generated by artificial intelligence. These models are tasked with classifying artworks as either "Original" or "Generated by AI," even when confronted with previously unseen examples.
For instance, Bianco et al. [3] conducted research using transformer-based models such as VGG-19 [5] , ResNet-50 [6], and ViT [4]. Their study demonstrated these models' ability to effectively classify artworks, supplemented by an explainable-AI phase aimed at elucidating the decision-making processes inherent to each model. This phase involved generating visual maps known as Grad-CAM [7], which highlight the specific areas within an artwork that influenced the model's classification decision.

The ultimate objective of this project is to enhance the explainability of these models' decision-making processes using Large Language Models (LLMs); so, by employing LLMs, we aim to gain deeper insights into why a model makes particular decisions.

2 Related work

In this section, we review the existing state-of-art and previous studies that are relevant to our project. The related work is organized into three main areas: image classification using neural networks, explainability in AI models using Grad-Cam, and leveraging large language models (LLMs) for generating explanations. By examining these areas, we can contextualize our goal and build upon the foundations laid by prior studies.

2.1 Image Classification Using Neural Networks

The task of image classification using deep neural networks has been extensively studied, particularly in the context of distinguishing AI-generated images from human-created ones. Various models have been developed and tested to tackle this challenge.
For example, in their article, Martin-Rodriguez et al.[8] focus on distinguishing photorealistic AI-created images from real camera photos using pixel-level feature extraction techniques: Photo Response Non-Uniformity (PRNU) and Error Level Analysis (ELA). These features are used to train convolutional neural networks (CNNs) for classification. Or the Epstein et al.[9] research addresses the real-time detection of AI-generated images using advanced neural network architectures. The study highlights the importance of quick and accurate identification methods to mitigate the spread of synthetic content online. Another example is the paper of Jeong Ha et al. [10] that investigates the distinguishing features between human-created and AI-generated artworks. It employs deep learning models to analyze and classify artworks, providing insights into the capabilities of neural networks in the art domain.

The foundational work by Bianco et al.[3] titled "Identifying AI-Generated Art with Deep Learning" serves as a basis for this part of the project. This article evaluates the performance of three transformer-based models— VGG-19, ResNet-50, and ViT— in distinguishing AI-generated images from real ones, achieving accuracy rates of 95%, 96%, and 97% respectively. Our objective is not to develop a new image classification model from scratch but to enhance the explainability of its decision-making process. Therefore, we based our project on the ViT model of Bianco at al., which demonstrated the highest accuracy.

2.2 Explainability in AI Models Using Grad-CAM

Explainability in AI models, particularly in deep learning, has garnered significant attention. One of the prominent techniques for visualizing and interpreting the decisions of the model is Grad-CAM (Gradient-weighted Class Activation Mapping) [7]. Grad-CAM generates heatmaps that highlight regions of an image which contribute most to the model's decision, providing insights into the model's focus areas during classification.

This technique has also been applied to the various classifications made by the models considered in the study by Bianco et al. Consequently, our implementation is also based on their research, enabling LLMs to explain the factors that influenced the model's decision-making process.

2.3 Leveraging Large Language Models for Explainability

Large language models (LLMs) have demonstrated remarkable understanding and generation of natural language. Recent research has explored the use of these models to generate textual explanations for AI decisions, improving the interpretability of complex models. One of these, is the study conducted by Yang et al.[11] in which the capabilities of five LLMs- GPT-4, LLaVa, Bard, ERNIE Bot 4.0, and Tongyi QianwenL -in detect sophisticated tampering. The assessment focuses on two areas: detection of content generated by artificial intelligence and detection of manipulation.

Experiments show that while most LLMs can identify logically inconsistent composite images and only the most powerful models can detect visible signs of tampering. However, none of the LLMs can reliably identify accurately forged or highly realistic images generated by artificial intelligence. To combat this problem, VP et al. [12] have developed a new approach using dense convolutional neural networks (Dense CNN) and multimodal fusion and the method has worked exceptionally well across multiple datasets, effectively identifying advanced deepfake variations. But an essential component of this approach is the integration of a LLM, in particular InstructBLIB. LLM add an extra layer of control by characterizing and analyzing media segments that are susceptible to manipulation, further improving system accuracy. Considering this state of the art, within this project the following LLMs were considered:

1. LLaVa-NeXt [13]: Compared to LLaVa-1.5, released in October 2023, LLaVa-NeXt includes improvements such as higher image resolution for better detail capture, better visual reasoning and Optical Character Recognition(OCR) functionality and better visual conversation in various scenarios.
2. InstructBLIB [14]: is an advanced vision-language model designed for comprehensive vision-language instruction tuning. It leverages the pretrained BLIP-2 models and incorporates a Query Transformer to extract task-specific features. InstructBLIP excels in zero-shot performance across multiple datasets and achieves state-of-the-art results in fine-tuned tasks, demonstrating significant improvements over previous models in vision-language processing.
3. KOSMOS-2 [15]: A Multimodal Large Language Model that enhances visual and textual understanding by integrating object description perception and text grounding into the visual world. Kosmos-2 advances capabilities in multimodal grounding, expression generation, and language tasks.

2.4 Evaluation and Frameworks for Explainability

Wu et al.[1] propose a framework ,called COSMOS, for the detection of Cheap-fake images that is divided into two parts:

1. Image-to-text matching module: This module verifies the consistency between an image and its caption by calculating a consistency score using semantic textual similarity (STS) with S-BERT. If the consistency score is below a certain threshold, the image-caption pair is considered out of context.
2. Out-of-context detection module (OOC): This module further analyzes caption pairs using a natural language inference model (NLI) to evaluate the logical relationships (involvement, contradiction or neutrality) between captions.

This method has been used as a guide to carry out the evaluation phase of this project. In fact, in a first step, the transformer-based method ,called CLIP [16] ,is used to evaluate how the descriptions generated by LLMs are consistent with the image. In a second step, using the S-BERT model [17], we evaluate the consistency between the caption generated by LLMs and the label assigned by the ViT to the image.

3 Materials

The dataset used is the same as that of Bianco et al., but this project focuses on a subset composed of 50 synthetic artworks and 50 original artworks. Each image in this subset is characterized by the following attributes: the path of the artwork, the path of the Grad-Cam image (generated by the model and superimposed on the artwork), and the label associated with the artwork (0 for AI-generated artworks; 1 for original artworks). In the phase of the project where descriptions by the three LLMs are generated, these descriptions will be added to the dataset and associated with each image. For the evaluation phase, the model's label will be transformed into a categorical format, as follows:

1. 1 will become "The artwork is Original"
2. 0 will become "The artwork was generated by AI"

4 Methods

The proposed method consists of two phases: an implementation phase where LLMs are applied to Grad-Cam outputs, and an evaluation phase.

4.1 Implementation Phase

As mentioned earlier, we use LLaVa-NeXt, InstructBLIB, and KOSMOS-2. In all three cases, the Grad-CAM heatmap overlay image serves as input for LLMs. The output is a description that explains which factors may have led the network to focus on the areas of maximum heatmap activation.

4.1.1 LLaVa-NeXt

LLaVa-NeXt (LLaVA-1.6) represents a significant advancement over its predecessor, LLaVA-1.5, focusing on enhancing multimodal AI capabilities. This model introduces a quadrupled input image resolution, supporting resolutions up to 672x672 pixels across multiple aspect ratios. These enhancements enable LLaVa-NeXt to capture finer visual details crucial for different applications. Moreover, it improves Optical Character Recognition (OCR) capabilities, integrates better world knowledge, and enhances visual reasoning. In this case, the quantized version of LLaVa-NeXt is employed, and following numerous tests, the prompt that yielded the best results was in Italian:

L'immagine rappresenta un'opera d'arte a cui è stata sovrapposta una mappa di attivazione generata da Grad-CAM. Questa mappa consente di visualizzare le regioni dell'opera su cui la rete ha basato la sua decisione di classificazione. In questo caso, la mappa sovrapposta all'opera permette di associare all'opera d'arte la label "Originale" se non è stata generata artificialmente, o "Generata da AI" se è stata generata artificialmente. L'opera sottostante alla mappa è stata classificata come <label>. Descrivi cosa ha potuto far concentrare la rete in quelle aree specifiche dell'opera per la classificazione. Dammi una risposta concisa che non spieghi la classificazione, ma solo il motivo ipotetico per il quale la rete si è concentrata in quelle aree. I colori rappresentano la mappa di attivazione generata dalla rete stessa, e sotto ai colori c'è il disegno originale. Non c'è nessun pubblico. La mappa rappresenta le aree su cui la rete si è concentrata di più per classificare l'opera. #ESEMPIO DI OUTPUT: L'opera è stata classificata come Generata da AI perché molto probabilmente le linee del disegno o i colori sono discordanti dai pattern rilevati.

4.1.2 IstructionBLIB

The InstructBLIP model, based on BLIP-2 architecture, specializes in generating detailed image descriptions. It processes 26 datasets into an instruction-guided format, enabling the extraction of pertinent visual features aligned with specific instructions. This capability empowers InstructBLIP to excel in tasks that integrate visual and language inputs, showcasing high performance without extensive fine-tuning. In this case, the model was given the following prompt:

Why does the activation map present those warm colors in that area of the work? What can be the possible causes regarding that part of the artwork (without explaining the functioning of the activation map)?

And the descriptions were generated thanks to these hyperparameters:

1. num_beams: Controls the number of beams to use for beam search decoding. Set to 5, allowing the model to consider multiple sequences simultaneously.
2. max_new_tokens: Specifies the maximum number of new tokens that can be generated by the model. Set to 250, ensuring the output length is constrained.
3. min_length: Defines the minimum length of the generated sequence. Set to 1, ensuring the output is not truncated prematurely.
4. top_p: A parameter for nucleus sampling (top-p sampling) that limits the cumulative probability of tokens to sample from. Set to 0.9, ensuring diverse and fluent outputs.
5. repetition_penalty: Penalizes the model from generating repetitive sequences. Set to 1.5, discouraging repeated tokens.
6. length_penalty: Controls how much to penalize longer sequences. Set to 1.0, maintaining a balance between fluency and length.
7. temperature: Affects the randomness of sampling. Set to 1, maintaining the original distribution of logits.

These hyperparameters control the generation process and ensure the model operates effectively with the input image and text data.

4.1.3 KOSMOS-2

KOSMOS-2 is a Transformer-based multimodal language model designed to ground textual descriptions with visual understanding. It utilizes spatial coordinates from bounding boxes in image data to enhance its comprehension of object representations. By integrating these location tokens into text spans, KOSMOS-2 connects specific image regions directly to their descriptive language. This capability extends beyond traditional language models by enabling precise multimodal grounding and enhancing tasks like referring expression comprehension and phrase grounding. In this case, the model was given the following prompt:

This image displays an artwork overlaid with an activation map, where color intensity indicates the classifier's focus: warmer colors signify stronger attention, while colder colors indicate weaker attention. The map highlights the regions most influential in classifying the artwork as either 'Original' or 'AI-Generated'. In this case, the classifier determined the artwork to be 'Original'. What elements of the artwork might have caused the classifier to focus on these specific areas and classify it as Original?

1. pixel_values: Represents the pixel values of the input image.
2. input_ids: Input token IDs for the language model.
3. attention_mask: Mask to avoid performing attention on padding token indices.
4. image_embeds: Embeddings of the input image, which are set to None.
5. image_embeds_position_mask: Mask indicating positions of the image embeddings.
6. use_cache: Boolean indicating whether to use caching during generation, set to True.
7. max_new_tokens: Maximum number of tokens to generate per input sequence, set to 1024.

These hyperparameters control the generation process and ensure the model operates effectively with the input image and text data.

5 Results

In Figures 1, 2,3 and 4 we find the results obtained for the different images.

5.0.1 Qualitative analysis of the results

By making a qualitative analysis of the results obtained, we can say that:

1. LLaVa-NeXt: regarding the descriptions generated for non-original works of art, the model is very precise in identifying the areas of interest and providing plausible reasons why the network focused on those areas. However, for original works of art, while it can still identify areas of interest, it shows some inaccuracies and has difficulty considering the label given by the ViT model.
2. InstructBLIB: it seems not to work well in both cases. It can detect the areas of interest, but fails to fulfill the task because it perceives the heatmap of the work as an image generated by thermal sensors.
3. Kosmos-2: in both cases it is able to identify the areas of interest, even if with some imperfections, but gives vague descriptions and sometimes does not fully grasp the task.

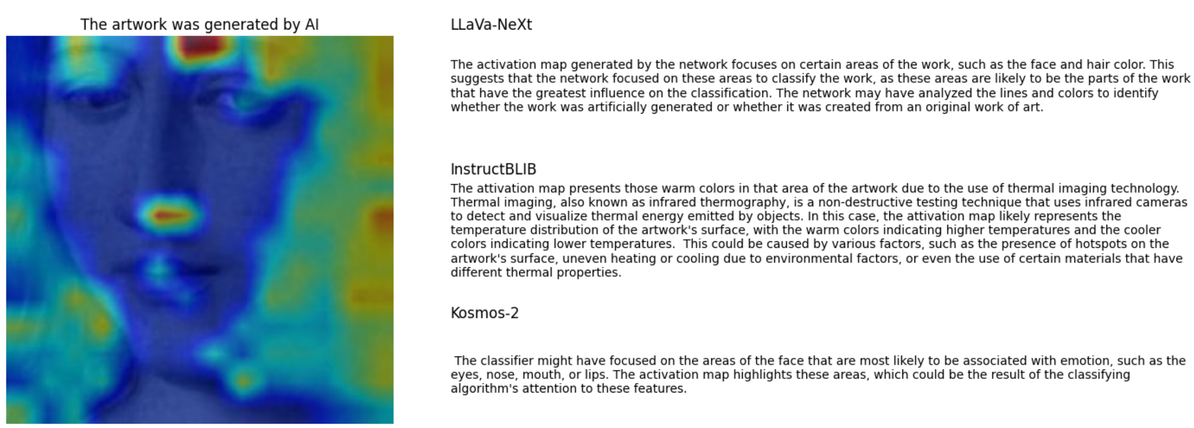

Figure 1: Descriptions associated with the artwork "Generated by AI" from the three LLMs.

Figure 2: Descriptions associated with the artwork "Generated by AI" from the three LLMs.

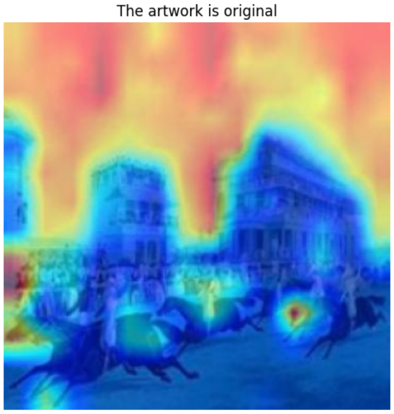

Figure 3: Descriptions associated with the artwork "Original" from the three LLMs.

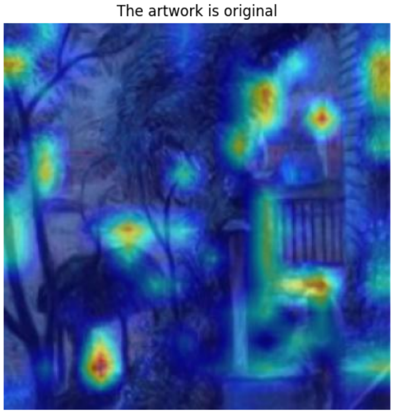

Figure 4: Descriptions associated with the artwork "Original" from the three LLMs.

In essence, from this initial qualitative evaluation, LLaVa-NeXt achieves better results by abstracting what it sees and effectively tackling the task, despite some inaccuracies. InstructBLIB fails to abstract and does not meet the task's requirements, while Kosmos-2 does abstract, but not sufficiently to fully satisfy the task.

5.0.2 Quantitative analysis of the data

As previously mentioned, the COSMOS framework was used as a reference for quantitative data analysis. Consequently, this project followed two phases:

1. Quantitative Analysis of Coherence Between Image and Generated Caption.
   To understand how well the LLMs grasp the context and content they analyze, we measure the similarity between the image and the generated caption. This is done using the CLIP model, a transformer-based model that embeds both the image and the caption, and then calculates their cosine similarity. The cosine similarity score ranges from 0 to 1, with a higher score indicating greater coherence between the two elements. Additionally, to calculate overall coherence, we considered the average similarity between the generated caption and both the Grad-Cam-applied image and the original image without Grad-Cam. This dual approach evaluates how well the LLM understands the underlying artwork and the specific task.
2. Quantitative Analysis of Coherence Between Image and Label.
   To determine how consistent the generated captions are with the labels produced by the ViT model ("The artwork is Original" or "The artwork was generated by AI"), we used the S-BERT model, which calculates cosine similarity between two texts. For each LLM, the average similarity between the captions and the image labels was calculated.

The total score is the sum of the similarities obtained from these two phases. The following results were obtained:

	Description-image Similarity	Description- label Similarity	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: Results of the quantitative analysis

From the overall similarity scores, it might seem that InstructBLIB and Kosmos-2 performed better. However, the qualitative analysis reveals a different story. LLaVa-NeXt actually performs best, as it avoids giving literal descriptions and instead abstracts what it sees to provide more intelligent and coherent explanations for the task. The higher scores of the other two models are due to their tendency to describe exactly what they see, resulting in higher similarity scores but less insightful explanations.

6 Conclusion

This project aimed to enhance the explainability of Transformer-based models used to classify images as "original" or "AI-generated" by incorporating Grad-CAM heatmaps and advanced LLMs. Our evaluation of LLaVa-NeXt, InstructBLIP, and KOSMOS-2 demonstrated varying levels of success in generating coherent and insightful descriptions based on the heatmaps applied to synthetic and original artworks.

LLaVa-NeXt emerged as the most effective model in our qualitative analysis. It identified the areas of interest and provided plausible explanations for the network's attention, particularly excelling with non-original artworks. However, it showed some inaccuracies with original artworks, indicating room for improvement. InstructBLIP, on the other hand, struggled in both scenarios, often misinterpreting the heatmap as a thermal sensor image and failing to provide relevant explanations. Kosmos-2, while capable of identifying areas of interest, tended to give vague descriptions and occasionally misunderstood the task.
The quantitative analysis revealed higher similarity scores for InstructBLIP and Kosmos-2 due to their literal descriptions, but these results did not align with the qualitative findings. LLaVa-NeXt, despite lower quantitative scores, provided more intelligent and coherent explanations, highlighting the importance of qualitative assessments in evaluating model performance.

Future work should focus on fine-tuning LLaVa-NeXt to address its inaccuracies with original artworks and to further enhance its performance. Additionally, exploring the potential of InstructBLIP with fine-tuning could improve its ability to generate meaningful descriptions. These steps are essential for improving the interpretability and practical application of AI models in image classification tasks.

References

[1] Guangyang Wu and Weijie Wu and Xiaohong Liu and Kele Xu and Tianjiao Wan and Wenyi Wang "Cheap-fake Detection with LLM using Prompt Engineering" In *arXiv,2023, https://arxiv.org/abs/2306.02776*

[2] S. Alex Yang and Angela Huyue Zhang "Generative AI and Copyright: A Dynamic Perspective" In *arXiv,2024,https://arxiv.org/abs/2402.17801*

[3] Bianco Tommaso, Castellano Giovanna ,Scaringi Raffaele, Vessio Gennaro and others. "Identifying AI-Generated Art with Deep Learning." In *CREAI@ AI* IA,2023, page 16–25*

[4] Alexey Dosovitskiy and Lucas Beyer and Alexander Kolesnikov and Dirk Weissenborn and Xiaohua Zhai and Thomas Unterthiner and Mostafa Dehghani and Matthias Minderer and Georg Heigold and Sylvain Gelly and Jakob Uszkoreit and Neil Houlsby "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" In *arXiv,2021,https://arxiv.org/abs/2010.11929*

[5] Karen Simonyan and Andrew Zisserman "Very Deep Convolutional Networks for Large-Scale Image Recognition" In *arXiv,2015, https://arxiv.org/abs/1409.1556*

[6] Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun "Deep Residual Learning for Image Recognition" In *arXiv,2015,https://arxiv.org/abs/1512.03385*

[7] Selvaraju, Ramprasaath R. and Cogswell, Michael and Das, Abhishek and Vedantam, Ramakrishna and Parikh, Devi and Batra, Dhruv "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization" In *International Journal of Computer Vision,2019 oct, Springer Science and Business Media LLC, page 336–359, vol.128, ISSN 1573-1405, numer 2 , http://dx.doi.org/10.1007/s11263-019-01228-7, DOI: 10.1007/s11263-019-01228-7*

[8] Martin-Rodriguez, Fernando and Garcia-Mojon, Rocio and Fernandez-Barciela, Monica "Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks" In *Sensor,23,2023,22, ART-NUM 9037,https://www.mdpi.com/1424-8220/23/22/9037*

[9] Epstein, David C. and Jain, Ishan and Wang, Oliver and Zhang, Richard, "Online Detection of AI-Generated Images" In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). 2023, pages 382-392*

[10] Anna Yoo Jeong Ha and Josephine Passananti and Ronik Bhaskar and Shawn Shan and Reid Southen and Haitao Zheng and Ben Y. Zhao "Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?" In *arXiv,2024,https://arxiv.org/abs/2402.03214*

[11] Xinyu Yang and Jizhe Zhou "Research about the Ability of LLM in the Tamper-Detection Area" In *arXiv,2024,https://arxiv.org/abs/2401.13504*

[12] VP, Samesh Enathe and S, Chandra Mouli and Dheepthi, R., "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),2024,pag.1-6*

[13] Liu, Haotian and Li, Chunyuan and Li, Yuheng and Li, Bo and Zhang, Yuanhan and Shen, Sheng and Lee, Yong Jae "LLaVa-NeXt: Improved reasoning, OCR, and world knowledge" In *January,2024, https://llava-vl.github.io/blog/2024-01-30-LLaVa-NeXt/*

[14] Wenliang Dai and Junnan Li and Dongxu Li and Anthony Meng Huat Tiong and Junqi Zhao and Weisheng Wang and Boyang Li and Pascale Fung and Steven Hoi "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning" In *arXiv,2023,https://arxiv.org/abs/2305.06500*

[15] Zhiliang Peng and Wenhui Wang and Li Dong and Yaru Hao and Shaohan Huang and Shuming Ma and Furu Wei "Kosmos-2: Grounding Multimodal Large Language Models to the World" In *arXiv,2023,https://arxiv.org/abs/2306.14824*

[16] Alec Radford and Jong Wook Kim and Chris Hallacy and Aditya Ramesh and Gabriel Goh and Sandhini Agarwal and Girish Sastry and Amanda Askell and Pamela Mishkin and Jack Clark and Gretchen Krueger and Ilya Sutskever "Learning Transferable Visual Models From Natural Language Supervision" In *arXiv,2021,https://arxiv.org/abs/2103.00020*

[17] Reimers, Nils and Gurevych, Iryna "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks" In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Nov 2019,Association for Computational Linguistics,https://arxiv.org/abs/1908.10084*