POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Leveraging Multimodal Large Language Models for Explainable Deepfake

Laurea Magistrale in Computer Science & Engineering - Ingegneria Informatica

Author: Lorenzo Morelli

Advisor: Prof. Mark James Carman

Co-advisors: Andrea Sassella, Paolo Bestagini

Academic year: 2023-2024

---

1. Introduction

In recent years, deepfake technology has rapidly advanced with techniques like GANs, autoencoders, and transformer-based models, and it has become increasingly difficult to distinguish between actual media and fake media. Though the initial deepfakes were noticeably flawed and contained artifacts, recent techniques have generated hyper-realistic media that can be used for identity theft, misinformation, and other malicious purposes. Though most current detectors may be effective, their black-box design limits transparency, which is a fundamental requirement in most applications like forensics. To overcome these constraints, multimodal approaches integrating visual information with natural language reasoning using Vision-Language Models (VLMs) and Large Language Models (LLMs) have the promise of more precise detection and improved explainability.

2. Background

Deepfake technology emerged as a breakthrough artificial media, from the early-days manual manipulation techniques to the current-day deep learning models. At the time, deepfakes had artifacts that were easily noticeable; however, with the emergence of GANs [5], autoencoders, and transformer-based architectures [22], their level of reality has been boosted greatly with architectures like StyleGAN [6] and Stable Diffusion [19]. This development has enabled deepfakes to render realistic imitations of real media, raising huge concerns for abuse in identity theft, political manipulation, and cybersecurity attacks.
Besides visual forgeries, deepfakes have reached the level of replicating human faces, voices, and even gestures, for entertainment and deception. Deepfake detection is then required, and new detection methods employ deep neural networks and transformers to uncover the nuanced inconsistencies and "GAN fingerprints" [14] that are symptoms of manipulation. Despite such advancements, the sophistication of today's forgeries and the need for generalizable detection procedures, i.e., zero-shot and few-shot learning [24] are a challenge.
Along with detection, explainability has also emerged as a significant area. Human-interpretable, transparent explanations are required by scientists and stakeholders to be able to trust and warrant AI decisions, especially in high-stakes areas like forensics, medicine,

and finance [4, 20]. Explainability methods include intrinsically interpretable model construction and post-hoc methods like LIME [17] and SHAP [13], along with saliency maps and layer-wise relevance propagation methods [2]. The interpretability-accuracy trade-off remains a significant challenge.

3. Related Work

3.1. Transformers

Transformers, introduced by Vaswani et al. [22], have revolutionized the way models process sequential data by employing self-attention and multi-head attention mechanisms. Unlike traditional RNNs that process data sequentially, transformers capture long-range dependencies in parallel, enabling faster training and improved performance on a wide range of tasks. Their ability to generate contextualized embeddings forms the foundation for many modern natural language processing systems.

3.2. Multimodal Language Models

Building on transformer architectures, multimodal language models integrate visual and textual data to generate richer, more comprehensive representations. Models such as CLIP [16] align image and text embeddings using contrastive learning, facilitating zero-shot classification. Additionally, frameworks like Flamingo [1] and LLaVA (Large Language and Vision Assistant) extend these ideas by combining a pre-trained language model with a vision encoder. This integration not only improves the detection of subtle anomalies in manipulated media but also supports the generation of natural language explanations, which is critical for tasks like deepfake detection.

3.3. Explainability in Multimodal Systems

Explainability is essential for building trust and ensuring transparency, especially in high-stakes applications such as deepfake detection. Transformer-based models inherently provide some interpretability through attention maps, which highlight the parts of the input that most influence the output [18]. Complementary post-hoc methods such as LIME [17] and SHAP [13] further enhance our understanding of model decisions. By combining these techniques, our approach not only achieves robust performance but also delivers clear and interpretable explanations, which are crucial for forensic validation and user trust.

4. Research Questions

The study is guided by several key research questions:

- What role do multimodal language models play in generating natural language explanations for detected manipulations?
- How does the performance of zero-shot approaches vary across different image tampering techniques?
- What impact do zero-shot and few-shot strategies have on detection outcomes and explanation quality?
- To what extent does fine-tuning a multimodal model enhance both its detection accuracy and interpretability?
- How can quantitative and qualitative evaluation metrics be integrated to systematically assess model performance?

These questions inform the experimental design and drive the analysis throughout the study.

5. Dataset Creation

A robust dataset is essential to support our deepfake detection research by providing consistent examples. Preliminary experiments with *LLaVA OneVision* showed that among various manipulation techniques, splicing produced the most distinct and predictable alterations. Motivated by this insight, we developed a custom dataset that focuses on splicing while also including genuine images to mitigate bias.

The final dataset comprises 20,000 images, divided as follows:

- 10,000 manipulated images (with splicing) from the DIS100K dataset [21].
- 5,000 genuine images from the RAISE dataset [3].
- 5,000 genuine images obtained by cropping tampered regions from DIS100K.

Each manipulated image is annotated with a dynamic caption generated using the *BLIP Large* model [10], and the pipeline for the dataset creation is shown in Figure 1. This multimodal annotation approach provides a rich foundation for evaluating, and eventually training, deepfake

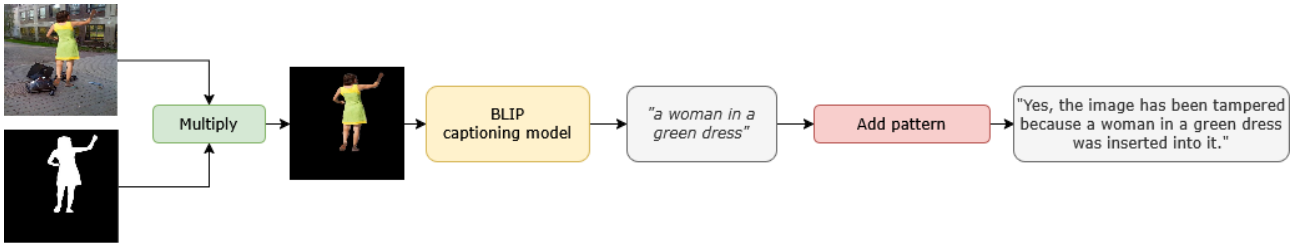

Figure 1: General pipeline of our custom dataset creation.

detection models under zero-shot, few-shot, and fine-tuned settings.

6. Methodology and Results

This chapter outlines our experimental framework and presents the key results of our deepfake detection study. We evaluated several state-of-the-art vision-language models in both zero-shot and few-shot settings, and then we fine-tuned one model using a parameter-efficient approach to enhance forensic precision.

6.1. Zero-Shot and Few-Shot Settings

We compared three multimodal models:

- **LLaVA OneVision 0.5B** [9]: An 899M-parameter model that integrates a vision encoder (SigLIP), a two-layer MLP projector, and the Qwen2 language model.
- **Llama 3.2 Vision Instruct 11B** [12]: A model optimized for detailed visual reasoning through instruction-based training.
- **DeepSeek VL2 Tiny** [23]: A computationally efficient model using a Mixture-of-Experts framework.

In the zero-shot setting, each model was given a test image along with the prompt:

```
"Has this image been
manipulated?  If so, what has
been added to the image?"
```

The outputs were manually categorized as true positives, false positives, true negatives, or false negatives, and standard evaluation metrics were computed (see Table 1).

Model	Acc	Pre	Rec	F1
Llama	0.5330	0.5107	0.9815	0.6718
LLaVA	0.6000	0.5498	0.9856	0.7059
DeepSeek	0.8659	0.9303	0.7808	0.8490

Table 1: Zero-Shot Evaluation Metrics

Model	Acc	Pre	Rec	F1
Llama	0.5080	0.4963	0.6879	0.5766
LLaVA	0.5190	0.5060	0.5175	0.5117
DeepSeek	0.5240	0.6897	0.0411	0.0775

Table 2: Few-Shot Evaluation Metrics

6.2. Fine-Tuning and Results

Despite promising initial results, the performance observed in both zero-shot and few-shot settings indicated that further improvements were required, as in our experiments the accuracy hardly reached 60%. This motivated the fine-tuning phase applied to our selected model *LLaVA OneVision 0.5B* [9], using Low-Rank Adaptation (LoRA). Only 5.8M of the 899M parameters were updated, preserving the robust language understanding of Qwen2 while adapting the visual encoder and the multimodal projector to the forensic domain.

Our fine-tuning employed the Adam optimizer [7] with a learning rate of 1×10^{-4} over 44 epochs, using 4 NVIDIA A40 GPUs, a batch size of 2, and images scaled down by 50%. A chat-based prompt (`"Has this image been manipulated? Why?"`) was used to simulate realistic interactions, with target captions generated from our dataset.

After fine-tuning, the model was evaluated on a test set of 1,000 images. The fine-tuned model achieved the results shown in Table 3.

Metric	Value
Accuracy	0.8840
Precision	1.0000
Recall	0.7618
F1 Score	0.8648

Table 3: Evaluation metrics for the fine-tuned model.

Figure 2 shows the confusion matrix where 513

genuine images were correctly identified (true negatives) and 371 tampered images correctly detected (true positives), with no false positives and 116 false negatives. A detailed analysis of true positives (Figure 3) further demonstrates the model's ability to reliably describe modifications.

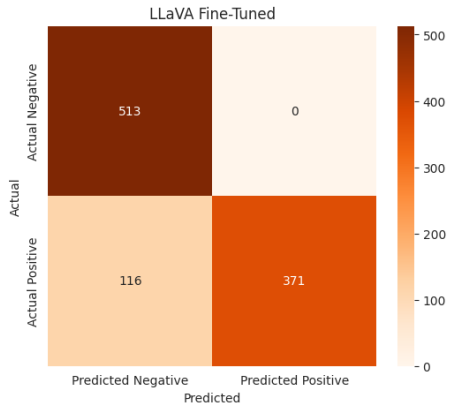

Figure 2: The confusion matrix assessing the quality of the fine-tuned model's responses.

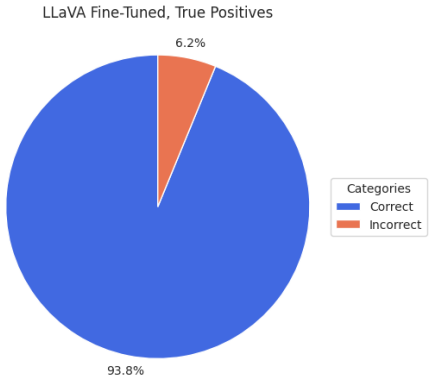

Figure 3: Pie chart showing the distribution of True Positives in the fine-tuned model.

In summary, the results show that while DeepSeek performs best in zero-shot tests, it performs poorly in few-shot settings, and Llama and LLaVA deliver subpar performance in both settings. However, the fine-tuned LLaVA model yields the highest accuracy, F1 score, and perfect precision, i.e., task-specific fine-tuning is the optimum solution for forensic deepfake detection.

7. Methods for the Analysis of the Results

In real-world forensic use, determining whether an image has been manipulated—and understanding exactly how—matters. Subtle differences between the generated caption and the ground-truth are not always captured by metrics based on simple word overlap; and manual parsing of these captions to evaluate semantic equivalence is not only time-consuming, but also prone to subjectivity.

We present an automated evaluation framework that provides high semantic accuracy ratings to captions. Our approach integrates classical NLP measures, embedding-based decisions, and LLM judges, with a *meta-classification* approach that automatically marks captions as good or bad using optimal threshold values. Since our fine-tuned model produced better results compared to other models, its generated captions are taken as the primary subject of our evaluation, compared directly with our custom dataset of ground-truth captions.

7.1. Conventional NLP Assessments

We begin by computing traditional metrics such as BLEU [15] and ROUGE [11] to quantify n-gram overlap between generated captions and their reference counterparts. Although these metrics provide a useful baseline for assessing fluency and adequacy, their reliance on surface-level similarity limits their ability to capture deeper semantic correspondence.

7.2. Embedding-Based Evaluation

To overcome the limitations of n-gram based approaches, we employed embedding-based methods using *all-MiniLM-L6-v2* and *NV-Embed-v2* [8]. Through extensive prompt engineering, we identified top-performing prompts:

- **Prompt A**: `"Assess the extent to which the two sentences convey the same meaning"`.
- **Prompt B**: `"Evaluate the degree to which the two sentences share the same meaning"`.
- **Prompt C**: `"Given two sentences, determine how semantically similar they are"`.

that yielded high average similarity scores. We then performed a greedy search over similarity thresholds (0–1) to optimize F1 scores, as validated by confusion matrices and scatter plots.

7.3. LLMs as Judges

To complement the quantitative metrics, we used *Llama-3.1-8B-Instruct* as an expert judge. The LLM was provided with pairs of generated and reference captions and prompted to assign a discrete similarity score, which was normalized to a 0–1 scale. As shown in Table 4, the resulting evaluation metrics—accuracy of 95.3%, precision of 99.63%, recall of 94.89%, and F1 score of 97.20%, aligned closely with our embedding-based assessments.

7.4. Final Evaluation

The final evaluation metrics, summarized in Table 4, demonstrate the robust performance of our methods. *NV-Embed-v2* with the selected prompts achieves high accuracy, precision, recall, and F1 scores, while the LLM-as-a-Judge approach offers valuable qualitative insights that complement the embedding-based evaluations. Overall, this comprehensive evaluation strategy, which bridges conventional metrics with human-like judgment, provides a reliable and generalizable means to assess caption quality, thereby supporting the efficacy of our deepfake detection approach.

Approach	Acc	Pre	Rec	F1
NV, A	0.9630	0.9714	0.9861	0.9787
NV, B	0.9620	0.9757	0.9803	0.9780
NV, C	0.9630	0.9714	0.9861	0.9787
Llama	0.9530	0.9963	0.9489	0.9720

Table 4: Evaluation metrics for NV-Embed-v2 using the three selected prompts compared to Llama-as-a-Judge.

8. Conclusions

Our study demonstrates that leveraging multimodal large language models can significantly enhance both the detection and interpretability of deepfake manipulations. By fine-tuning a state-of-the-art model on a meticulously annotated dataset focused on splicing techniques, we achieved notable improvements in detection accuracy and in the generation of human-understandable explanations.

Despite the encouraging results, our evaluation was confined to static images and splicing manipulations. Future work should widen the scope to include other tampering methods such as copy-move, in-painting, as well as extend the framework to video deepfakes or real-time detection.

In summary, our research underscores the potential of multimodal language models for deepfake detection and explainability, paving the way for more robust and generalizable forensic systems in real-world applications.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015.

[3] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: a raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, MMSys '15, page 219–224, New York, NY, USA, 2015. Association for Computing Machinery.

[4] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[6] Tero Karras, Samuli Laine, and Timo Aila.

A style-based generator architecture for generative adversarial networks, 2019.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[8] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models, 2025.

[9] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024.

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[11] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[12] Meta Llama. Llama 3.2-11b vision instruct. `https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct`, 2024. Accessed: March 7, 2025.

[13] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[14] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints?, 2018.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[18] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works, 2020.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[20] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.

[21] Eren Tahir. Dis100k. `https://www.kaggle.com/datasets/erentahir/dis100k`, 2024. Data set on Kaggle. Retrieved February 6, 2025.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[23] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024.

[24] Jiazhen Yan, Ziqiang Li, Ziwen He, and Zhangjie Fu. Generalizable deepfake detection via effective local-global feature extraction, 2025.