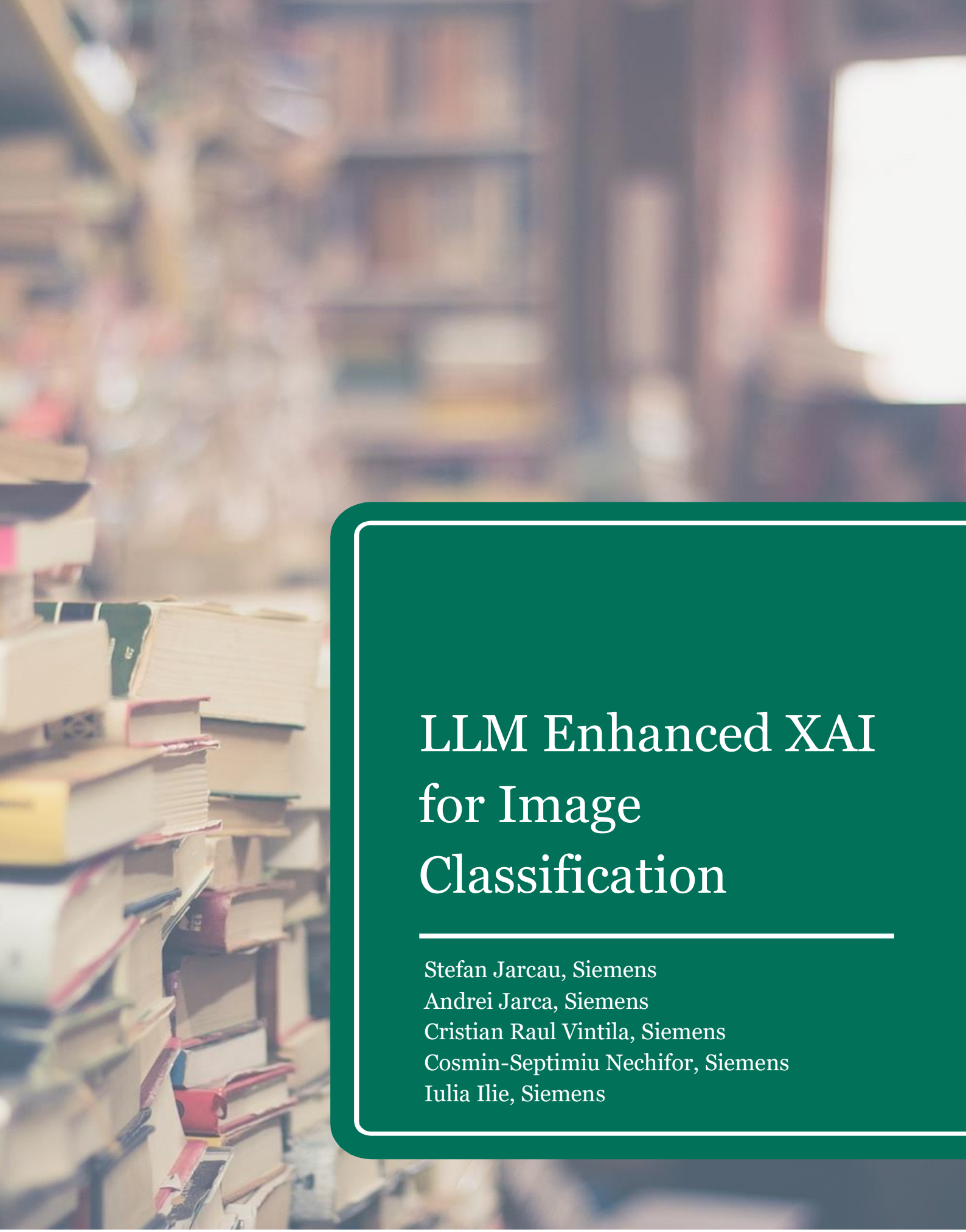

LLM Enhanced XAI for Image Classification

Stefan Jarcau, Siemens
Andrei Jarca, Siemens
Cristian Raul Vintila, Siemens
Cosmin-Septimiu Nechifor, Siemens
Iulia Ilie, Siemens

Co-funded by the European Union

This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under the Grant Agreement No. 101073909.

LLM Enhanced XAI for Image Classification

Stefan Jarcau, Andrei Jarca, Cristian Raul Vintila, Cosmin-Septimiu Nechifor, Iulia Ilie (Siemens)

We propose a framework integrating Explainable Artificial Intelligence (XAI) techniques with Large Language Models (LLMs) to advance natural language explanations for AI based image classification. This approach can be used for critical infrastructure cybersecurity by enhancing transparency in threat detection, providing cybersecurity analysts with clear insights into decision-making processes. LLM-generated natural language explanations can offer a detailed understanding of cybersecurity rationale, ensuring effective communication with diverse stakeholders. Our framework can contribute to a proactive cybersecurity stance, aligning with regulatory compliance standards and providing a robust strategy for safeguarding critical infrastructures.

1. Introduction

In the rapidly evolving landscape of machine learning and artificial intelligence applications, the need for transparent and interpretable models has become increasingly significant. The interpretability of classification algorithms for image classification is not only needed for building trust in decision-making of the automated systems but also for understanding the factors influencing their outputs. This paper introduces an approach for assessing the appropriateness of explanations produced by automated explanation algorithms for a set of image classifications. The proposed approach involves combining explainable Artificial Intelligence (XAI) techniques for deep learning-based image classification and incorporating their outputs into a Large Language Model (LLM) for a comprehensive evaluation.

The integration of XAI techniques combined with LLM into the evaluation process of an automatic classification technique enables us to extract meaningful insights into their decision-making processes. By dissecting the black-box nature of these models, we aim to shed light on the factors contributing to their outputs. Additionally, the incorporation of a LLM allows us to capture the subtle nuances in the explanations, providing a richer understanding of the decision rationale, and increase the audience to which these explanations can be delivered.

2. The Current State of Affairs in XAI for Image Classification

XAI techniques have lately been used more often to support image classification and their application to real world scenarios, where trustworthy systems are needed. These techniques help provide insights into the decision-making processes of the machine learning models used to simulate and predict necessary outcomes. Comprehensive analyses have been conducted to evaluate and compare different XAI methods for remote sensing image classification [1][2]. Such an approach involves using state-of-the-art machine learning

algorithms to classify remote sensing images and then applying XAI methods to analyze the results qualitatively and quantitatively [3]. Another study focused on developing a multi-scale scheme of LIME (Local Interpretable Model-Agnostic Explanations) to explain decisions made by Convolutional Neural Network (CNN) models in image classification [4]. Additionally, XAI frameworks have been proposed to produce multiple explanations for image classification systems based on middle-level input features extracted by autoencoders [5]. These studies provide valuable insights and recommendations for selecting appropriate XAI methods in the context of image classification.

However, these explanations are often directed to data scientists and require a high level of expertise to further employ in a significant manner. The use of LLMs in combination with XAI for explaining image classification is a very recent area of research, with notable works including [6][7][8]. Further, we believe that an interesting outcome of combining LLM and XAI for image classification could be a way to assess the appropriateness of XAI explanations for general audiences. We propose here a system to achieve these aspects.

3. The Role of Combining XAI and MLLM for Image Classification

In order to improve and assess the quality and non-expert audience reach of XAI explanation for image classification, we propose a fully AI based, automated classification-explanation-translation system. Our system employs state-of-the-art neural network architectures, such as Vision Transformer (ViT [9]), for image classification. We apply the ViT on widely accepted benchmarks like ImageNet and CIFAR-10. Then, we explain the decision-making process of classification using multiple XAI techniques, e.g. Lime, SHAP, etc. We evaluate the outputs of these techniques through multimodal LLMs (MLLMs), comparing the resulting natural language explanations after utilizing consistent prompts. This research seeks to determine if the results of LLMs are influenced by the choice of XAI methods and assess their informativeness for a general audience.

4. The Research and Development Path in ATLANTIS

The proposed approach of integrating XAI techniques with LLMs for image classification has significant potential for improving risk management in the cybersecurity of critical infrastructures in the context of the ATLATIS project. This approach can be applied to enhance threat detection and interpretability, providing cybersecurity analysts with transparent insights into the features influencing decision-making. The natural language explanations generated by LLMs contribute to a detailed understanding of the rationale behind cybersecurity decisions, facilitating informed decision-making and risk assessment.

Adapting the language used in LLM-generated explanations ensures broad audience adaptability, allowing effective communication of cybersecurity risks to non-technical stakeholders and policymakers. This is exemplified in Figure 1, where the simple visual explanation provided by LIME [10], a SOTA XAI approach for image classification is enhanced by a natural language explanation provided by an MLLM system (LLAVA [11]),

allowing the user to better understand the pixel demarcation and the AI system decision making.

Continuous monitoring of critical infrastructure systems using such an integrated approach could enable fast adaptation to evolving threats, ensuring a proactive cybersecurity posture. Possible scalability features of the integrated approach could contribute to handling large and complex datasets associated with critical infrastructures, improving resource efficiency, and reducing response times to potential threats.

In incident response and forensic analysis, our integrated approach aids in investigating and understanding cybersecurity incidents by providing transparent explanations for detected events. This can support cybersecurity professionals in identifying root causes. Additionally, the approach can contribute to regulatory compliance and reporting, helping meet cybersecurity standards and provide comprehensive insights for regulatory authorities. Overall, we believe that the proposed integrated XAI and LLM approach offers a robust framework for advancing risk management strategies in the image-based cybersecurity domain.

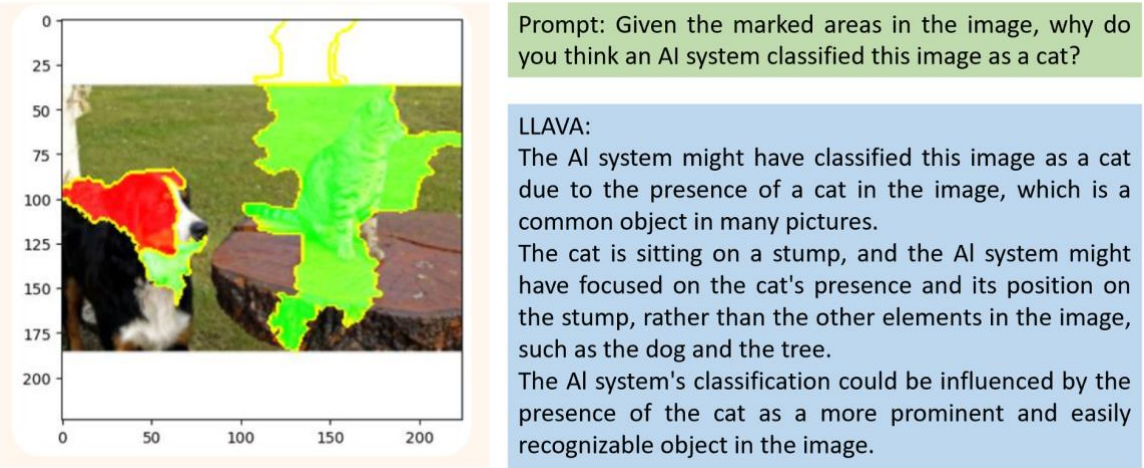

Figure 1. Prompt given ro an MLLM system (LLAVA), asking it to explain the decision making of an AI system(ViT) in classifying an image, given the extra pixel demarcation provided by LIME, and the answer returned by the MLLM system.

5. The Challenges and Barriers

Ensuring seamless compatibility between state-of-the-art neural network architectures, such as Vision Transformer, and a diverse array of XAI techniques poses a significant challenge. The inherent differences in model architectures and complexities necessitate comprehensive compatibility assessments and adaptations to facilitate integration and interoperability. Significantly interpreting the outputs generated by XAI techniques can also be a complex task, particularly when confronted with intricate and high-dimensional data from deep learning models. Developing robust visualization and interpretation strategies is going to be necessary to derive meaningful insights from the XAI outputs, to provide clear explanations of the underlying decision-making processes to general audiences.

Another important challenge in the development of our proposed approach is going to be the management of sensitive information within the employed datasets. These can raise

privacy and security challenges, particularly when deploying XAI techniques that reveal details about specific data instances. Robust data anonymization and encryption measures must be implemented to safeguard sensitive information and ensure compliance with privacy regulations.

The scalability of our proposed pipeline, especially concerning large datasets and computationally intensive XAI techniques, introduces potential challenges related to performance bottlenecks. Implementing optimization techniques and parallel processing will become essential to enhance the overall scalability and performance of the integrated system.

6. The Benefits and Impact

The benefits and possible impacts of our research include:

- Improved explanations for AI based image classification.
- Larger reach of AI based classification explanations, leading to increased chance of AI system adoption by non-experts.
- Comprehensive LLM-XAI framework for image classification that can streamline the process of selecting and applying appropriate risk management strategies.
- Improved communication of cybersecurity risks detected in images to diverse stakeholders, including non-technical audiences and policymakers.

7. Future Outlook

In the ongoing development, the current status involves the successful implementation of a state-of-the-art neural network (NN) architecture. Explanations have been generated using LIME for a set of figures, and the input from LIME has been seamlessly integrated into the LLAVA Large Language Model resulting in promising explanations.

The next phase of the project entails the implementation of additional XAI approaches. These diverse XAI methods will be applied to generate explanations, and their respective outputs will be prepared for integration into LLAVA. The objective is to systematically study the effects of changing the XAI approaches on the final natural language explanations produced by LLAVA. This iterative process of incorporating various XAI techniques aims to evaluate their individual contributions and nuances in shaping the interpretability and coherence of the explanations generated by LLAVA.

By exploring multiple XAI approaches, the project seeks to identify the most effective methods for providing transparent and meaningful insights into the decision-making processes of the neural network, ultimately contributing to the enhancement of natural language explanations.

8. Conclusions

The proposed integration of explainable Artificial Intelligence techniques with Large Language Models for image classification in the context of critical infrastructure cybersecurity can provide a significant advancement in risk management. This approach can enhance transparency in threat detection and interpretation, providing cybersecurity analysts with clear insights into decision-making processes. The natural language explanations generated by LLMs could contribute to detailed understanding of the rationale behind cybersecurity decisions, facilitating informed decision-making and risk assessment. The adaptability of language can ensure effective communication of cybersecurity risks to diverse stakeholders, including non-technical audiences and policymakers. In incident response and forensic analysis, the integrated approach aids in investigating and understanding cybersecurity incidents by providing transparent explanations. Additionally, our approach can contribute to regulatory compliance and reporting, aligning with cybersecurity standards and offering insights for regulatory authorities. Overall, this integrated XAI and LLM approach can become a comprehensive framework for bolstering risk management strategies in the cybersecurity domain.

References

[1] Aadil Ahamed, K. Alipour, Sateesh Kumar, Severine Soltani, Michael J. Pazzani. (2022). "Improving Explanations of Image Classification with Ensembles of Learners." Artificial Intelligence and Applications. https://doi.org/10.5121/csit.2022.121801

[2] Adrien Bennetot, Gianni Franchi, Javier Del Ser, Raja Chatila, Natalia Díaz-Rodríguez. (2022). "Greybox XAI: a Neural-Symbolic learning framework to produce interpretable predictions for image classification." Knowledge Based Systems, 255, 109947. https://doi.org/10.1016/j.knosys.2022.109947

[3] Akshatha Mohan, Joshua Peeples. (2023). "Quantitative Analysis of Primary Attribution Explainable Artificial Intelligence Methods for Remote Sensing Image Classification." arXiv.org. https://arxiv.org/abs/2306.04037

[4] Andrea Apicella, Francesco Isgrò, Andrea Pollastro, Roberto Prevete. (2023). "Strategies to exploit XAI to improve classification systems." arXiv.org. https://arxiv.org/abs/2306.05801

[5] Hooria Hajiyan, Mehran Ebrahimi. (2023). "Multi-scale local explanation approach for image analysis using model-agnostic Explainable Artificial Intelligence (XAI)." Proceedings Article, 12471, 124711N-124711N. https://doi.org/10.1117/12.2654307

[6] Chen, Long, et al. "Driving with llms: Fusing object-level vector modality for explainable autonomous driving." arXiv preprint arXiv:2310.01957 (2023).

[7] Han, Songhao, et al. "LLMs as Visual Explainers: Advancing Image Classification with Evolving Visual Descriptions." arXiv preprint arXiv:2311.11904 (2023).

[8] Yang, Yue, et al. "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[9] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020)

[10] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. " Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

[11] Haotian, Liu., Chunyuan, Li., Qingyang, Wu., Yong, Jae, Lee. (2023). Visual Instruction Tuning. arXiv.org, doi: 10.48550/arXiv.2304.08485

Front cover image by Eli Francis via Pixabay.
https://pixabay.com/users/elifrancis-1160677