Responsible and Reliable AI at PICUS Lab

Narendra Patwardhan[1], Lidia Marassi[1], Michela Gravina[1], Antonio Galli[1], Monica Zuccarini[1], Tannistha Maiti[2], Tarry Singh[2], Stefano Marrone[1,*] and Carlo Sansone[1]

[1] *University of Naples, Federico II, Naples, Italy*
[3] *Deepkapha AI, Assen, Netherlands*

Abstract

The PICUS Lab has been conducting research activities in the field of artificial intelligence (AI) with a focus on ethics. One area of research has been the modification of transformer-based models to make them more sustainable while maintaining performance. To achieve this goal, we have explored sustainable alternatives for the internal components of these models, such as retrieval-based techniques and diffusion modules for programmability. The aim is to pave the way for the development of ethical and sustainable AI systems that do not rely on massive computing and data, which can lead to high energy consumption and carbon footprint. Another area of research at PICUS Lab has been deepfake detection, a pressing concern due to the potential spread of false information and manipulation of public opinion. To address this issue, we developed FEAD-D (Face Expression Analysis for Deepfake Detection), a tool based on facial expression analysis. The system uses a bidirectional Long Short-Term Memory (BiLSTM) model and data from the DeepFake Detection Challenge (DFDC) to detect fake videos in about two minutes.

Keywords

Generative Models, Transformers, Sustainable AI, Foundation Models

1. Introduction

Artificial intelligence (AI) has made significant progress in recent years, yielding promising results in various downstream tasks. However, AI models often rely on massive computing and data, raising concerns due to high energy consumption and carbon footprint. To address these concerns, at the PICUS Lab we have been conducting research activities in the field of AI with a focus on ethical concerns. This paper aims to present two of these research projects carried out at the PICUS Lab: modification of transformer-based models for sustainability and deepfake detection using facial expression analysis. The importance of these topics is supported by recent studies and reports. The World Economic Forum [1] has highlighted the importance of ethical and sustainable AI systems, stating that "ethical AI can drive innovation, and sustainability should be at the heart of AI." Additionally, a report by the European Union on AI regulation highlights the need for ethical and sustainable AI systems to ensure the long-term benefit of society. These studies and reports highlight the urgency of addressing ethical and sustainability concerns in the field of AI and provide further support for the research projects presented in this paper.

In particular, in the first project we focused on the development of sustainable AI systems. The increasing use of AI models raises concerns regarding energy consumption and carbon footprint, prompting researchers to explore sustainable alternatives for the internal components of these models. Also, the increasing use of such generative models for fake news creation is posing serious ethical and sociological concerns. The researchers propose modifications to transformer-based models that maintain performance while reducing their reliance on massive computing and data, while also opening for a more ethic-by-design training strategy. Retrieval-based techniques and diffusion modules for programmability are discussed as sustainable alternatives. These modifications can pave the way for the development of ethical and sustainable AI systems.

Instead, the second project focuses on deepfake detection, a pressing concern due to the potential spread of false information and manipulation of public opinion. To address this issue, researchers at the PICUS Lab have developed a tool called FEAD-D, which stands for Face Expression Analysis for Deepfake Detection. This tool uses facial expression analysis to detect fake videos, overcoming limitations in current deepfake detection systems. The system is based on a bidirectional Long Short-Term Memory (BiLSTM) model and data from the DeepFake Detection Challenge (DFDC). FEAD-D had been founded under the CINECA ISCRA-C program (ID. HP10CMJKEO, IsC93).

The results of these studies are important in the context of the growing use of AI and its potential impact on society. The development of ethical and sustainable AI systems is crucial for the long-term benefit of soci-

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.
stefano.marrone@unina.it (S. Marrone)
0000-0002-4807-5664 (N. Patwardhan); 0000-0001-5033-9617 (M. Gravina); 0000-0001-9911-1517 (A. Galli); 0000-0001-6852-0377 (S. Marrone); 0000-0002-8176-6950 (C. Sansone)

ety. This paper highlights the importance of addressing ethical and sustainability concerns in the field of AI and presents two projects that contribute to this effort.

2. HOMINIS: Towards Sustainable Foundation Models

Artificial intelligence (AI) has emerged as a transformative force in modern society, with generative modelling serving as a key driver behind its rapid advancements. Foundation models [2] are large-scale machine learning models, pre-trained on vast amounts of diverse data, that serve as a backbone for various downstream applications through fine-tuning and adaptation. Such models, particularly those based on transformer architectures, have achieved remarkable performance in domains such as natural language processing, computer vision, and reinforcement learning. Despite their success, the development and deployment of these models have raised critical concerns regarding the ethical and environmental implications of their high computational requirements and energy consumption.

As foundation models grow in size and complexity, the search for optimal hyperparameters becomes increasingly challenging and resource-intensive. Solely relying on scaling up can lead to overfitting and may result in diminished returns on model performance improvements. It also overlooks opportunities to optimize smaller models, which could deliver similar performance with reduced costs and resource demands. The high parameter count of foundation models presents challenges for inference on standard hardware, such as personal computers and mobile devices. This limitation restricts access to the benefits of these models for a broader audience and may contribute to a digital divide between those who can afford to deploy and utilize advanced AI systems and those who cannot.

Sustainable AI refers to managing the life cycle of artificial intelligence systems in a way that minimizes negative environmental, social, and economic impacts while maximizing long-term benefits for society. This holistic approach emphasizes the importance of ethical considerations, energy efficiency, and resource optimization in AI design, as well as fostering inclusive collaboration and equitable access to AI-driven technologies.

As the world evolves rapidly, AI models must be able to adapt and learn from new information to remain relevant and useful. Finetuning large models incurs significant costs and hence programmability in AI models is necessary so that they can be updated to accommodate emerging trends, knowledge, and societal shifts. This capability is crucial for maintaining the accuracy and effectiveness of AI systems in real-world applications, as it allows them to keep pace with the dynamic nature of human societies and avoid becoming outdated or less reliable over time.

In this section, we will illustrate project Hominis, which exploits AI techniques in the realm of generative AI and foundation models, carried out at the University of Naples Federico II in collaboration with industrial partners (DeepKapha). We will highlight the innovative aspects and contributions of this project, emphasizing its significance in advancing the state-of-the-art in sustainable and programmable AI. By leveraging the expertise of both academic and industrial stakeholders, project Hominis aims to develop cutting-edge solutions and applications that harness the potential of AI to address real-world challenges and deliver tangible benefits across a wide range of sectors.

2.1. Essential Additions to Foundation Models

To optimize large-scale language models, data curation processes have shifted from human-led to heuristics-based automated filtering. However, this automation can lead to biases, memorization of private data, and vulnerability to adversarial attacks. Our research collaboration with RealAI aims to develop sustainable foundation models by addressing issues related to data sourcing, key components, and essential additions. Project Hominis aims to sanitize public datasets and develop crawling strategies for capturing diverse, multi-faceted data. This project will also develop tools for the community to analyze, curate, and critique datasets while ensuring fairness, privacy, and legality.

The transformer block, based on self-attention, is the fundamental constituent of modern foundation models. The attention mechanism plays a crucial role in the Transformer architecture, enabling the model to focus on relevant features in the input data. In this study, we will perform ablation experiments on three attention acceleration techniques: Flash Attention [3] (hardware-aware acceleration), Linear Approximations [4, 5, 6, 7] (acceleration via proxying), and Synthetic Attention [8] (acceleration through replacement). Our aim is to evaluate the effectiveness of these methods in improving model efficiency while maintaining performance levels.

The linear layer, another key component of the Transformer architecture, can be optimized using routing techniques (such as those popularized by Switch Transformer [9]) to enhance its performance without adversely affecting inference time. By employing routing strategies, we can achieve a more efficient distribution of resources within the linear layer, resulting in improved computational efficiency and reduced resource consumption without sacrificing the quality of the model's output.

Large AI models can benefit from multimodal data, which improves performance, generalization ability, and robustness. Networks that process independent modalities using a common structure exhibit synergistic effects on generalization. Additionally, diversity in data sources, particularly in low-resource languages, can improve downstream generalization.

To leverage data from diverse domains and take advantage of this synergetic effect, we rely on two primary factors, tokenization and choice of architecture. This study proposes Universal Tokenization, an innovative method that focuses on byte-level representation augmented by special tokens, enabling a unified encoding of diverse data types. The other component we rely on is cross-attention, which allows models to decouple computational complexity from sequence length, as demonstrated by the Perceiver architecture [10, 11, 12]. By using cross-attention mechanisms, these architectures can efficiently handle long sequences and large-scale inputs without incurring excessive computational costs. Our work will harness this decoupling to effectively support multimodality.

Retrieval-Augmented Generation (RAG) is a technique that combines the strengths of pre-trained language models with external knowledge sources, enabling hot updates and offering several benefits. The RETRO paper [13], which popularized this approach, showcases its potential for enhancing the performance and adaptability of AI models. By incorporating real-time information from external databases, RAG allows models to stay up-to-date with the latest developments and trends, improving their relevance and accuracy. This dynamic integration of knowledge also facilitates faster updates, reducing the need for resource-intensive retraining processes. As a result, retrieval-augmented generation contributes to more efficient and versatile AI models that can effectively address the ever-evolving needs and challenges in various domains.

We are also exploring the incorporation of diffusion as the last layer of our AI model to enhance controllability. By adding a diffusion-based mechanism, we aim to refine the generated output while preserving the structure and coherence of the model's predictions. This approach allows us to exert greater control over the model's behaviour and adjust its responses based on specific requirements or constraints.

Although our work primarily emphasizes inference time optimizations, we are also committed to reducing the carbon footprint associated with training time. To achieve this, we employ muTransfer [14], an innovative technique that enables the discovery of optimal hyperparameters by training a scaled-down model, effectively eliminating the need for resource-intensive hyperparameter searches.

Overall, this research direction aims to create more sustainable, adaptable, and responsible foundation models by addressing data sourcing, key components, and essential additions in AI systems.

3. FEAD-D

Emotions are a valuable tool for detecting deepfakes because they are difficult to replicate convincingly. This is a limitation of deepfake creation algorithms, as emotions are essential for human communication and are easily conveyed through facial expressions, voice tone, and body language. Identifying the emotional characteristics of a video has become an increasingly popular method for detecting manipulated content, as deepfakes often struggle to accurately reproduce emotions, including micro-expressions that reveal true emotions. FEAD-D (Face Expression Analysis for Deepfake Detection) exploits the inconsistencies in facial expressions introduced by deepfake creation artefacts. These videos often fail to accurately capture the full range of emotions and micro-expressions of the original subject. By analyzing a large sample of fake videos and comparing them to their original counterparts, FEAD-D identified temporal inconsistencies in the non-natural sequences of facial expressions. Figure 1 illustrates the differences in emotional patterns across consecutive frames of real and fake videos, emphasizing the importance of considering the temporal evolution of facial expressions in detecting deepfakes. To this aim, FEAD-D consists of the following modules:

- **Face detection**, in which the target video is analyzed to detect one or more faces.
- **Features Extraction**, that analyzes the texture of target images and extracts the emotion from the detected face frame by frame.
- **Features temporal analysis**, in which all the features extracted in the previous stages are analyzed together in a cross-frame fashion to spot incoherent and unnatural patterns in the emotional evolution of the target subject.

The resulting system can process a video in two minutes and is easy to adopt with minimal technical knowledge. It is worth noting that although emotional analysis is a promising approach, it presents challenges related to variations across individuals, cultures, and contexts, and the possibility of creating algorithms specifically designed to mimic emotional expressions. Further research is necessary to establish emotional analysis as a reliable method. Future work can focus on developing more robust deepfake detection methods using advanced computer vision techniques, training models to analyze both lip movements and speech signals, and investigating the effectiveness of combining multiple detection methods.

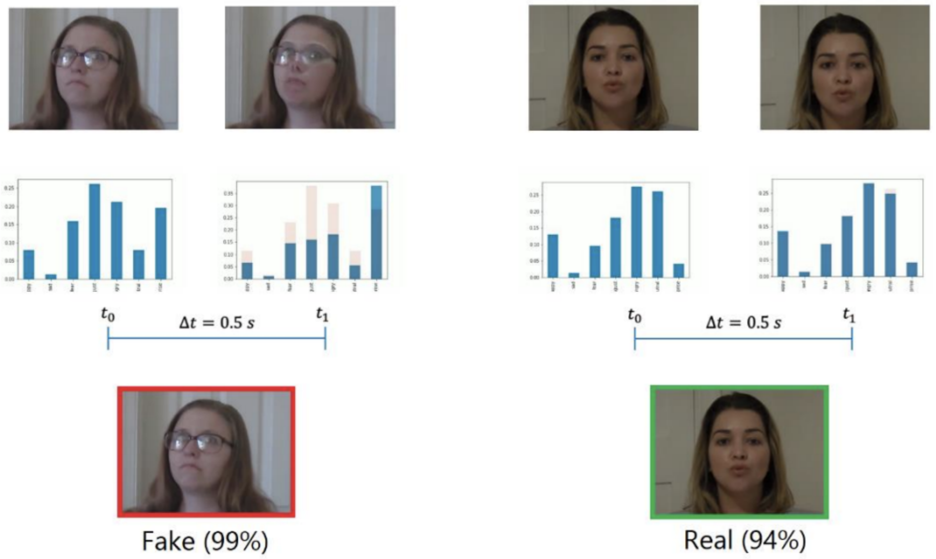

Figure 1: Illustrative example of how the emotion patterns vary across frames for a fake and for a real video. The histograms report the emotional changes across the frames in the video under analysis.

4. Conclusions

This paper presented two research projects carried out at the PICUS Lab with a focus on sustainability and ethical concerns in artificial intelligence (AI). The first project explored modifications to transformer-based models to make them more sustainable while maintaining performance. The researchers proposed sustainable alternatives such as retrieval-based techniques and diffusion modules for programmability. The second project focused on deepfake detection, a pressing concern due to the potential spread of false information and manipulation of public opinion. The researchers developed a tool called FEAD-D, which stands for Face Expression Analysis for Deepfake Detection, based on facial expression analysis. The importance of these topics is supported by recent studies and reports that highlight the urgency of addressing ethical and sustainability concerns in the field of AI. The results of these studies contribute to the development of ethical and sustainable AI systems, which are crucial for the long-term benefit of society.

In the first project, we proposed modifications to transformer-based models that maintain performance while reducing their reliance on massive computing and data, thus paving the way for the development of ethical and sustainable AI systems. In the second project, the researchers developed a tool called FEAD-D that uses facial expression analysis to detect fake videos, overcoming limitations in current deepfake detection systems. Overall, the research presented in this paper highlights the importance of addressing ethical and sustainability concerns in the field of AI and provides important contributions towards this effort. Future research in this area should continue to explore sustainable alternatives for AI models and further improve deepfake detection systems.

References

[1] Towards a more sustainable and equitable future for ai, http://www3.weforum.org/docs/WEF_Towards_a_More_Sustainable_and_Equitable_Future_for_AI_Report_2018.pdf, ????

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).

[3] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, C. Ré, Flashattention: Fast and memory-efficient exact attention with io-awareness, arXiv preprint arXiv:2205.14135 (2022).

[4] S. Wang, B. Z. Li, M. Khabsa, H. Fang, H. Ma, Linformer: Self-attention with linear complexity, arXiv preprint arXiv:2006.04768 (2020).

[5] N. Kitaev, Ł. Kaiser, A. Levskaya, Reformer: The efficient transformer, arXiv preprint arXiv:2001.04451 (2020).

[6] I. Beltagy, M. E. Peters, A. Cohan, Longformer:

The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[7] B. Chen, T. Dao, E. Winsor, Z. Song, A. Rudra, C. Ré, Scatterbrain: Unifying sparse and low-rank attention, Advances in Neural Information Processing Systems 34 (2021) 17413–17426.

[8] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences, Advances in neural information processing systems 33 (2020) 17283–17297.

[9] W. Fedus, B. Zoph, N. Shazeer, Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, ????

[10] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, J. Carreira, Perceiver: General perception with iterative attention, in: International conference on machine learning, PMLR, 2021, pp. 4651–4664.

[11] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, et al., Perceiver io: A general architecture for structured inputs & outputs, arXiv preprint arXiv:2107.14795 (2021).

[12] C. Hawthorne, A. Jaegle, C. Cangea, S. Borgeaud, C. Nash, M. Malinowski, S. Dieleman, O. Vinyals, M. Botvinick, I. Simon, et al., General-purpose, long-context autoregressive modeling with perceiver ar, arXiv preprint arXiv:2202.07765 (2022).

[13] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, et al., Improving language models by retrieving from trillions of tokens, in: International conference on machine learning, PMLR, 2022, pp. 2206–2240.

[14] G. Yang, E. J. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, J. Gao, Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, arXiv preprint arXiv:2203.03466 (2022).