Semantically Explainable Bias Detection and Mitigation in Multimodal AI

Overview

The rise of large language models (LLMs) integrated with visual inputs has ushered in a new era of multimodal systems, exemplified by vision-language assistants (VLAs) like LLaVA and InternVL. These systems enable tasks such as image captioning, visual question answering, and multimodal reasoning, presenting transformative potential for applications ranging from accessibility tools to creative industries. However, as with any AI system, these models inherit biases present in their training data, particularly regarding sensitive attributes such as gender. Studies have revealed that VLAs exhibit biased behavior, such as over-representing men in technical occupations or disproportionately attributing positive personality traits to women (Zhao et al., 2017; Bolukbasi et al., 2016). The widespread adoption of VLAs necessitates rigorous evaluation frameworks to identify and mitigate such biases, ensuring that these systems contribute to equitable societal outcomes.

Despite progress in debiasing methods, challenges persist in balancing bias mitigation with performance retention in downstream tasks. Fine-tuning-based approaches have shown promise, achieving favorable trade-offs between fairness and functionality (Tan et al., 2020; Ravfogel et al., 2022). However, the effectiveness of debiasing often depends on a nuanced understanding of the underlying biases, which is difficult to achieve using traditional evaluation methods. Many current approaches rely on statistical metrics or simple clustering techniques, which may lack semantic interpretability for end-users (Kim et al., 2020). This lack of transparency complicates efforts to analyze and address biases in AI systems, particularly in high-stakes domains where human oversight is critical.

To bridge this gap, explainability-focused tools like "Say My Name" (SaMyNa) have emerged, offering novel methods to semantically disentangle task-related and bias-related features in deep learning models (Ghorbani et al., 2019). SaMyNa's text-based pipeline highlights how semantic insights can support both bias detection and debiasing processes, enhancing model interpretability and usability. Building on this foundation, this research seeks to adapt SaMyNa's principles to the multimodal domain, providing an explainable and actionable framework for analyzing gender biases in VLAs. By uniting advances in bias evaluation for VLAs with explainability-driven approaches, this project aims to ensure that AI systems are not only performant but also equitable and transparent in their operation.

In this project, we will be mainly working on leveraging multimodal solutions and improving their explainability/interpretability. After the study and implementation of a benchmark with all the most popular and effective approaches, we seek to design next-generation solutions towards unification of information in multimodal contexts. This exploration will also validate or raise concerns/questions of known concepts around model compression and efficiency, that can lead to the design of new and more effective approaches. The perfect candidate has a strong background in Python coding and puts in massive passion and effort. Depending on the final quality and quantity of the results achieved, a submission of a paper will be taken into consideration.

Objectives at a glance

- Study the most recent state-of-the-art approaches.
- Deeply understand forward and back-propagation mechanisms and how these impact algorithmic complexity for these.
- Replicate/reproduce state-of-the-art results, with partial or total re-implementation of the approaches.
- Study the feasibility of proposed improvements (both from the perspectives of energy consumption and interpretability), scaling from simple to more complex tasks, from simple convolutional neural networks to transformers, and ultimately to foundation models.

Tutors

The main responsible for the project is **Enzo Tartaglione**, MdC for the Multimedia equipe enzo.tartaglione@telecom-paris.fr. Through the project candidates will also receive supervision by Prof. **Stephan Alaniz**.

Pre-interview for this project is **mandatory** and can be agreed upon by email.

Curious about meeting the student's equipe and learning about my student's research? Visit the website https://enzotarta.github.io/