The contribution of computer vision in multimodal content moderation

In today's data-driven world, the problem of content moderation on online platforms has reached critical proportions. Several sources confirm a general increasing trend of people’s exposure to inappropriate or misleading pieces of content. For example, according to a national survey conducted in the UK [1], approximately 90% of people aged between 18 and 34 (most online represented age group) had at least one interaction with digital harmful content. With the increasing complexity/power of adversarial attacks towards current moderation systems, a robust and efficient solution is required. This would be essential to ensure a safe, respectful, and legally compliant digital landscape, where users can interact, share information, and engage without undue risks.

Designing a content moderation solution can be a challenging task because it involves handling various considerations, either from a technical point of view or a more social one. On one hand, content on digital platforms can be vast and constantly changing. Adding new labels and tuning the solution to the new requirements could be a very time-consuming endeavor. Moderation solutions need to be able to process and assess a high volume of content in real-time, while also maintaining a high degree of predictive performance.

On the other hand, the wide diversity present in online communities produces the requirement that the system should be able to deal with nuanced or context-dependent content like sarcasm, satire or cultural references. Moreover, the moderation process must avoid perpetuating biases and discrimination, providing a fair chance to free online communication. In addition, besides user-generated content, content generated by powerful AI models represents a novel distinct category that should be treated accordingly. Lastly, the system should not be a complete black box relative to its users. Every moderation decision should be accompanied by a concise and clear explanation for the result of the analysis.

The previous challenges could outline a set of features that the solution should have. As opposed to the textual modality, the visual one is characterized by more informationally rich samples, conveying several messages through a single set of pixels. This implies that the moderation system should have a broad understanding of cultural and social concepts to be robust against natural cultural variations but also against intentionally adversarial ones [2] (i.e., masking

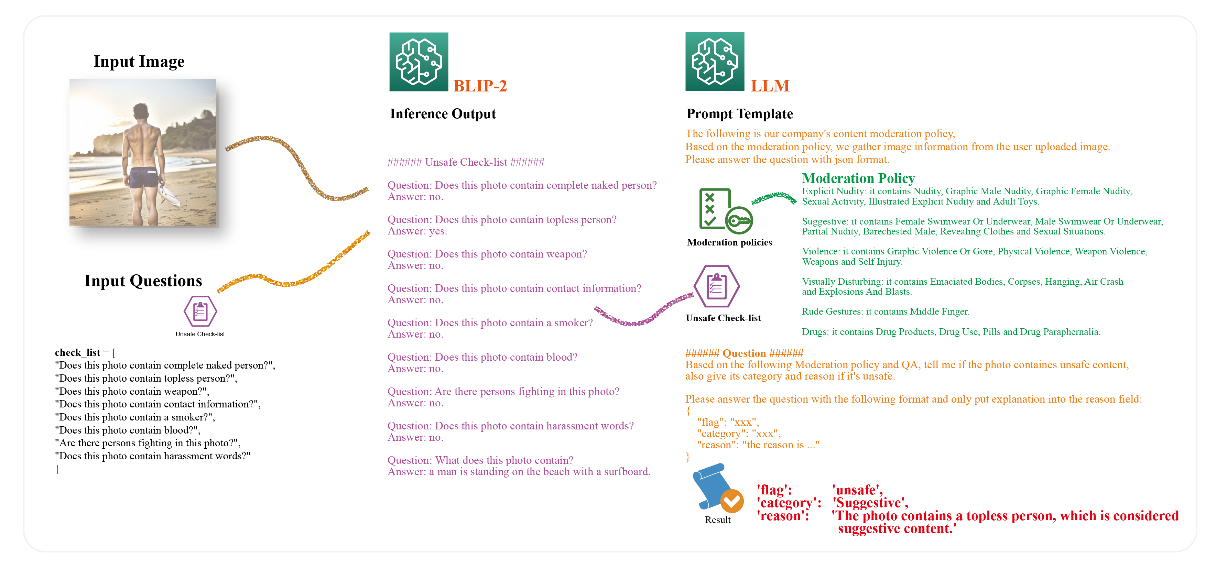


Fig. 1 Outline of a multimodal content moderation system (image from [3])

harmful or untrue content through noise injection). Also, it should have the ability to explain its classification to reinforce the result with sound and logical reasons for the decision.

One current avenue which could be pursued for a more powerful moderation system would be a specialized (i.e., fine-tuned) large multimodal model (LMM). Handling visual data, text and possibly other modalities, the model would be able to form a comprehensive understanding of what harmful or nonfactual means. It could handle multicultural content, providing that this is also representative for the data used for pretraining and fine-tuning. It can better identify attempts to evade content moderation through techniques like coded language or disguised imagery by considering various modalities together. Moreover, it could be steered towards explaining its decision or it could be accompanied by a different model that provides explanation. Lastly, regarding the discrimination of factual or nonfactual content, the common hallucination failure mode of large generative models may be mitigated by grounding the model in actual verified data that would be made available in-context during inference.

While current multimodal approaches vary( [3], [4], [5]), a concrete example in content moderation realm is built with detail by Wang et. al. [6] Their approach is based on extracting a text description of a possibly offensive image via a BLIP2 [5] instance, represented as a list of question-answer pairs(checklist) aimed to capture the intrinsic offensiveness. This caption is sent to a more capable large language model (e.g., LLaMA 2 - [7]) together with a moderation policy to check the safety of the content. Even though the designed system leans more towards text analysis capabilities, it still requires a powerful representation of visual content, facilitated by the frozen pre-trained image encoder from BLIP2.

The future of content moderation technology will likely be shaped by continuous improvements in AI, the evolving nature of online communication, and the need to strike a balance between safety, free expression, and privacy in our digital lives. Considering the increasing amount of data that needs to be processed and analyzed, originating from heterogenous sources, (either in a scientific sense or cultural one) scaling, bias reduction and global compliance could also be added to a large, expanding list of requirements needed to adapt to an ever-changing landscape.

# References

|  |  |
| --- | --- |
| [1] | F. Dennehy, "Almost 90% of young people exposed to harmful content on social media," 3 March 2023. [Online]. Available: https://www.turing.ac.uk/news/almost-90-young-people-exposed-harmful-content-social-media. [Accessed 21 October 2023]. |
| [2] | J. Zhang, Q. Yi and J. Sang, "Towards adversarial attack on vision-language pre-training models," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. |
| [3] | H. Liu, C. Li, Q. Wu and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485,* 2023. |
| [4] | A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin and J. Clark, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021. |
| [5] | J. Li, D. Li, S. Savarese and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597,* 2023. |
| [6] | G. Wang, Y. Cui and M. Li, "Build a generative AI-based content moderation solution on Amazon SageMaker JumpStart," Amazon, 5 September 2023. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/build-a-generative-ai-based-content-moderation-solution-on-amazon-sagemaker-jumpstart/. [Accessed 21 October 2023]. |
| [7] | H. Touvron and al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288,* 2023. |