Spring 3-16-2024

Comparing Generative AI And Humans In Image Feature Extraction

Md Jabir Rahman

Faruk Ahmed

COMPARING GENERATIVE AI AND HUMANS IN IMAGE FEATURE EXTRACTION

Completed Research Paper

Md Jabir Rahman
Oklahoma City University
mrahman@okcu.edu

Faruk Ahmed
The University of Memphis
mfahmed@memphis.edu

ABSTRACT

In this study, we assess the performance of Generative AI tools, such as ChatGPT, in feature labeling of unstructured image data. Our experiment, a comparative analysis between human coders and AI in identifying image features, reveals that AI-generated responses exhibit lower similarity scores, particularly in subjective feature identification, compared to human assessments. The results highlight a significant limitation in current Generative AI capabilities, emphasizing the need for further development and caution in tasks requiring nuanced interpretation. This research contributes to the Information Systems field by providing critical insights into the efficacy of Generative AI tools in processing unstructured data and guiding the prudent application and development of AI in feature extraction.

Keywords

Generative AI, ChatGPT, Image Labeling, Unstructured Data Analysis

INTRODUCTION

Recent years have witnessed a substantial surge in the growth and popularity of Generative AI tools, such as ChatGPT. Particularly notable is their capacity to comprehend inputs in a human-like manner and convert this information into practical instructions (Hariri, 2023). Numerous studies have explored the application and reliability of Generative AI tools across various domains, including but not limited to public health, global warming, programming, and nucleic acid research (e.g., Biswas, 2023; Surameery & Shakor, 2023; Chatterjee et al., 2023). While some studies reveal promising results, others raise significant concerns (e.g., Mohammad, 2023; Sallam, 2023; Chowdhury, 2023; Johnson et al., 2023; Zhu, 2023).

In the realm of the Internet of Things (IoT), users continually generate multimodal data, including images. As such, the business sector can no longer depend solely on structured data. There is a growing need to annotate, analyze, and utilize information extracted from unstructured data sources like images to enhance business outcomes. Image data is increasingly vital in various sectors, including digital marketing, online sales, social networking, medicine, security, consumer insight, and attraction (e.g., Ivasic-Kos, 2022; Cankul et al., 2022). Image mining and analysis can significantly augment business process management by extracting valuable insights potentially absent in structured data (Schmidt et al., 2016). Manual analysis, particularly for small businesses, could be more practical given the sheer volume of image data produced daily.

This study posits that leveraging Generative AI tools like ChatGPT for feature extraction could simplify the analysis of such data. No existing studies provide insights into this aspect, particularly regarding the nature of features that a model needs to identify. Thus, we conducted an experiment comparing the performance of human coders and a Generative AI tool (i.e., ChatGPT) in feature identification. Our findings reveal i) a significant disparity in the feature identification abilities between human coders and the Generative AI tool and ii) a relatively higher similarity score for objective features compared to subjective ones. Here, 'objective features' are quantifiable, as opposed to 'subjective features,' which depend on individual assessments or perceptions (Castelo et al., 2019).

This study contributes to the Information Systems (IS) field in several ways. Firstly, it is among the first to experiment with feature extraction from image data, comparing the efficacy of a Generative AI model with that of human coders. Secondly, it questions the indiscriminate reliance on Generative AI tools like ChatGPT for extracting features from unstructured data such as images. Lastly, our findings underscore the need for caution when employing these tools for tasks involving subjective judgment.

METHODOLOGY

Data Collection

Our data collection process commenced with an experimental survey, administered via Qualtrics, targeting student samples. In this ongoing research, our initial objective was to gather data on 50 images of Airbnb hosts. We ensured anonymity in the survey and incentivized participation by offering students ten bonus points upon completion. We communicated that opting out of the survey would have no adverse consequences.

Participants were randomly divided into two groups—subjective and objective—with each student evaluating a single image, randomly assigned by the survey tool. The objective group was tasked with identifying eleven distinct objective features, whereas the subjective group focused on eight features (Appendix A). The lesser number of subjective features identified is attributable to the inherently constrained scope of subjective elements recognizable by both human participants and Generative AI tools.

Upon reviewing the initial survey details, each participant made an informed decision to proceed with or exit the survey. One participant chose not to continue post-consent, and three others discontinued midway, resulting in forty-six valid responses. Additionally, six responses were excluded due to the absence of image number references, a decision to enhance our findings' confidence and integrity. Consequently, our final data set comprised forty images.

In the second phase, we sought to gather corresponding data on the same set of images using a Generative AI tool. We selected ChatGPT (version 4) for its capability to custom-build GPT models tailored to specific tasks. Two distinct GPT models were developed: one for identifying subjective features and another for objective features. To ensure consistency with human responses, we equipped the GPTs with identical questionnaires and response options. We conducted rigorous preliminary testing of these GPTs to verify their accuracy in generating the intended responses.

Data Processing

We structured our analysis around two distinct datasets: one for subjective features and the other for objective features. We developed a Python program to facilitate this analysis, leveraging the Pandas and NumPy libraries. This program incorporated custom functions for data preprocessing, such as normalization, to mitigate any discrepancies arising from textual variations, like differences in case sensitivity. In our human-annotated dataset, we took particular care in cases where two respondents evaluated the same image. Our research team meticulously reviewed these responses to ensure consistency. In divergent responses, a detailed manual review of the image was conducted, and the response most accurately reflecting the image's features was retained.

Upon completing the preprocessing stage, our custom function initiated a matching process based on 'image_id,' a unique identifier assigned to each image. This process enabled us to compute the similarity score for each image. The overall similarity score for each dataset was then determined by summing the individual similarity scores of all unique images and dividing this total by the number of images in the dataset.

Result

The table below shows the result for both datasets including their similarity scores. In summary,

	Objective	Subjective
Number of valid participants	17	23
Number of images	17	23
Number of features	11	8
Similarity score	68.8%	39.7%

Table 1. Similarity Analysis Result

for objective features, we received a 68.8% similarity score, while for subjective features, the similarity score was 39.7%.

DISCUSSION

Implications of Findings

This study has several theoretical and practical implications. First, the study contributes to our current knowledge of artificial intelligence in interpreting unstructured data. It challenges the predominant assumption that AI can match human performance or can outperform them in all forms of data analysis. The lower similarity score we received highlights the importance of a careful consideration and model refinement, especially in subjective feature identification. Second, the research underscores a fundamental limitation of Generative AI tools- difficulty in tasks that require nuanced interpretation. This further helps us refine the boundary conditions under which AI tools operate effectively. Finally, by offering a comparative analysis framework, we offer a methodological basis for future research.

Practically, our study has several important implications. First, our findings of similarity score analysis suggest that businesses and organizations should exercise high caution when implementing and accepting AI tools, especially where subjective judgement and interpretation is necessary. Second, for AI developers and researchers, we offer and differentiate areas where Generative AI tools need further improvement and attention. This could help steer research and development effort towards understanding AI's understanding and data processing to improve Generative AI's human like subjective assessment. Finally, it emphasizes the importance of transparency in AI's limitation in data processing, guiding towards a higher ethical consideration. This further help educators develop course materials and teach AI while emphasizing a realistic understanding of Generative AI's capabilities and limitations.

Limitations and Future Research Directions

Study Limitations: Our study, while insightful, has limitations. The focus on a specific Generative AI tool and a limited data set of 40 Airbnb host images may affect the generalizability of our results. Additionally, though methodologically sound, the division into subjective and objective groups might need to be more balanced with the complex nature of feature identification.

Future Research: Future studies could expand our work by exploring a broader range of AI tools and incorporating more extensive, diverse datasets. Further investigation into the nuanced differences between subjective and objective feature identification by AI tools could also provide deeper insights. Moreover, longitudinal studies examining the evolution of AI capabilities over time would be valuable in understanding the progress and potential of AI in unstructured data analysis.

Hence, we plan to collect more participant data using established survey distribution platforms like Upwork, and Prolific in the future. Moreover, once we have a larger data set with annotated features, we plan to connect this data back to our original data to identify any possible effects that those features have on a host's business.

Conclusion

This study embarked on a pivotal exploration to evaluate the efficacy of Generative AI tools, such as ChatGPT, in processing unstructured data, specifically focusing on feature labeling in image data. Our comprehensive experiment revealed a significant insight: the responses generated by these AI tools exhibit a lower similarity score compared to human assessments, particularly in subjective feature identification. These findings bear profound implications for the field of AI. They underscore a critical limitation in current Generative AI capabilities, especially in tasks requiring nuanced, subjective interpretation. This revelation is crucial for AI researchers, highlighting a pivotal area for further development and refinement within AI models. For practitioners, this study serves as a cautionary note on the reliability of AI tools in specific contexts, emphasizing the need for supplementary human oversight and verification in tasks involving subjective judgment.

Our research contributes significantly to the ongoing discourse in AI and data processing, offering valuable insights that can guide future advancements and applications in unstructured data analysis.

REFERENCES

1. Biswas, S. S. (2023). Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5), 868-869.
2. Biswas, S. S. (2023). Potential use of chat gpt in global warming. *Annals of biomedical engineering*, 51(6), 1126-1127.
3. Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5), 809-825.
4. Cankul, D., Ari, O. P., & Okumus, B. (2021). The current practices of food and beverage photography and styling in food business. *Journal of Hospitality and Tourism Technology*, 12(2), 287-306.
5. Chatterjee, S., Bhattacharya, M., Lee, S. S., & Chakraborty, C. (2023). Can artificial intelligence-strengthened ChatGPT or other large language models transform nucleic acid research?. *Molecular Therapy-Nucleic Acids*, 33, 205-207.

6. Chowdhury, M. N. U. R., & Haque, A. (2023, June). ChatGPT: Its Applications and Limitations. *In 2023 3rd International Conference on Intelligent Technologies (CONIT)* (pp. 1-7). IEEE.
7. Hankinson, G. (2005). Destination brand images: a business tourism perspective. *Journal of Services Marketing*, 19(1), 24-32.
8. Hariri, W. (2023). Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing. *arXiv preprint* arXiv:2304.02017.
9. Ivasic-Kos, M. (2022). Application of Digital Images and Corresponding Image Retrieval Paradigm. *ENTRENOVA-ENTerprise REsearch InNOVAtion*, 8(1), 350-363.
10. Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., ... & Wheless, L. (2023). Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Research square*.
11. Mohammad, B., Supti, T., Alzubaidi, M., Shah, H., Alam, T., Shah, Z., & Househ, M. (2023). The pros and cons of using ChatGPT in medical education: a scoping review. *Stud Health Technol Inform*, 305, 644-7.
12. Sallam, M., Salim, N., Barakat, M., & Al-Tammemi, A. (2023). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J*, 3(1), e103-e103.
13. Schmidt, R., Möhring, M., Zimmermann, A., Härting, R. C., & Keller, B. (2016). Potentials of image mining for business process management. In Intelligent Decision Technologies 2016: *Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016)–Part II (pp. 429-440)*. Springer International Publishing.
14. Surameery, N. M. S., & Shakor, M. Y. (2023). Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC)* ISSN: 2455-5290, 3(01), 17-22.
15. Zhu, C., Sun, M., Luo, J., Li, T., & Wang, M. (2023). How to harness the potential of ChatGPT in education?. Knowledge *Management & E-Learning*, 15(2), 133.

APPENDIX

Objective Features	Subjective Features
Estimate the Gender of the person in the photo (options: female, male, female, and male, other)	Estimate the age group of the person in the photo
Describe the hairstyle of the person in the photo (options: long, short, medium, other)	What emotion(s) can you identify in the person's expression?
Is the photo of a family/couple, or a single individual? (options: family, couple, single individual)	Describe the attire and grooming of the person in the photo by selecting and answering each option. - What type of clothing the person(s) is wearing? (options: casual, professional)
Which direction is the person's head facing? (options: forward, left, right)	How is the space used in the photo? Is there a lot of white space or is it full of objects? (options: full, lots of white space)
Is the person looking straight at the camera or at an angle? (options: straight, at an angle)	Assess the color and contrast of the photo. Options: (low/normal/high contrast)
What is the pose and posture of the person in the photo? (options: standing, sitting, other)	Rate the clarity and resolution of the photo. (options: very clear, clear, average, poor, very poor)
Does the photo appear balanced or wider than deeper or deeper than wider (wider means closer to the camera, deeper means far from the camera? (options: proportional, disproportional)	Is the photo well-lit or poorly lit? (options: well-lit, poorly lit)
Describe the context and setting of the photo. (options: indoor, outdoor urban, outdoor rural, other)	Does the photo appear to have filters or editing applied, or does it look raw and unedited? (options: raw and unedited, editing applied)
Are there other people in the photo apart from the main subject? (options: yes, no)	
Are there any significant objects in the photo, like animals (dogs, horses) or items? (options: yes, no)	
What is the race/ethnicity of the person/persons?	

Table A1: Objective and Subjective Features Used for Analysis