CUTIE: A human-in-the-loop interface for the generation of personalised and contextualised image captions

Aliki Anagnostopoulou
DFKI & Carl-von-Ossietzky University of Oldenburg
Oldenburg, Germany
aliki.anagnostopoulou@dfki.de

Sara-Jane Bittner
DFKI
Oldenburg, Germany
sara-jane.bittner@dfki.de

Lavanya Govindaraju
DFKI
Oldenburg, Germany
lavanya.govindaraju@dfki.de

Hasan Md Tusfiqur Alam
DFKI
Oldenburg, Germany
hasan.alam@dfki.de

Daniel Sonntag
DFKI & Carl-von-Ossietzky University of Oldenburg
Oldenburg, Germany
daniel.sonntag@dfki.de

Abstract

Image captioning is an AI-complete task that bridges computer vision and natural language processing. Its goal is to generate textual descriptions for a given image. However, general-purpose image captioning often does not capture contextual information, such as information about the people present or the location the image was shot. To address this challenge, we propose a web-based tool that leverages automated image captioning, large foundation models, and additional deep learning modules such as object recognition and metadata analysis to accelerate the process of generating contextualised and personalised image captions. The tool allows users to create personalised and contextualised image captions efficiently. User interactions and feedback given to the various components are stored and later used for domain adaptation of the respective components. Our ultimate goal is to improve the efficiency and accuracy of creating personalised and contextualised image captions.

CCS Concepts

• **Human-centered computing** → *Interactive systems and tools*; • **Computing methodologies** → **Natural language generation**; **Computer vision tasks**.

Keywords

image captioning, interactive machine learning, contextualisation, personalisation

1 Introduction

Image captioning involves automatically generating textual descriptions for visual images, leveraging advancements in computer vision and natural language processing. Although current state-of-the-art models excel at producing basic image descriptions (e.g., assisting visually impaired individuals or automotive applications), they often fail when confronted with additional contextual information not captured by the image itself. This limitation is particularly pertinent when integrating user-specific details or external context, prompting the consideration of interactive and human-in-the-loop approaches that engage human participation.

Our proposed system, CUTIE, which stands for *Contextual Understanding and Tailoring for Image Explanations*, integrates interactive and contextualised image captioning within a photobook-editing-style interface. We introduce a novel tool that facilitates eliciting user-specific and contextual information to generate tailored and context-aware captions. By synergising object detection, metadata extraction, and large foundation models in an intelligent user interface, our approach effectively incorporates additional context beyond the information in the image.

2 Related work

Previous approaches in *interactive image captioning* have focused on improving general-use captions by integrating various interactive components: [9] present an interactive-predictive system for generation tasks, including image captioning, which considers user feedback and integrates online learning for adaptation. [7] involve the human-in-the-loop by providing incomplete sequences as input, in addition to each image, during inference time. [3] extend the *Show, Attend, and Tell* [14] architecture by combining high-level and low-level features, which provide explainability and beam search during decoding time. [2] propose an interactive image captioning pipeline integrating data augmentation and continual learning to avoid overfitting and catastrophic forgetting during repeated training. [13] integrate interactive prompts for improved caption inference. More recently, [5] extend LLaVA by creating a model that allows users to mark images and interact with them with visual prompts.

Contextualised image captioning considers additional context to generate an image caption that describes the image's content and includes relevant external information. The context provided is, in most cases, in text form. [4] and [12] use news articles as context; the former uses a template-based architecture, and the latter uses an end-to-end architecture, considering additional features such as face and object detection. A modified version of the model proposed by [12] is used in [8] for image captioning on Wikipedia [11].

Mensch und Computer 2025 – Workshopband, Gesellschaft für Informatik e.V., 31. August - 03. September 2025, Chemnitz, Germany

https://doi.org/10.18420/muc2025-mci-demo-261

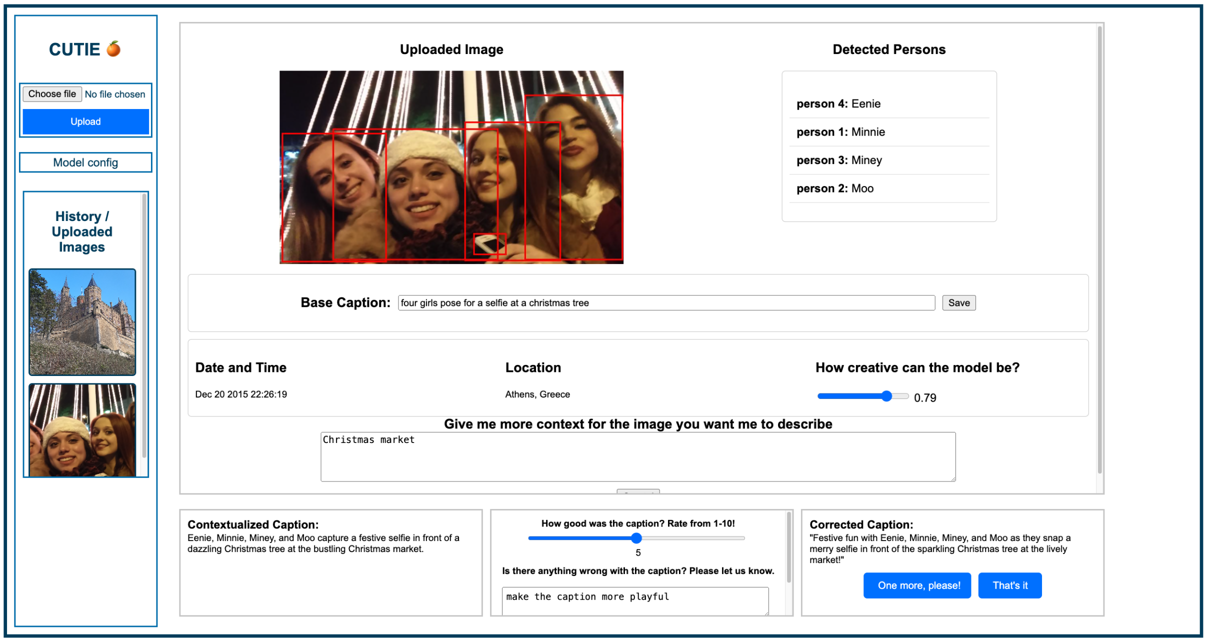

Figure 1: Screenshot of our intelligent user interface for personalised/contextualised image captioning.

3 System design

We demonstrate a web-based tool for interactive image captioning. Human-in-the-loop is essential for generating personalised and contextualised image captions. The tool allows users to process images in a photo-editing-like interface (Figure 1). We integrate various deep learning modules to extract information that the user needs to provide. Contextual information and user feedback are incorporated via large language models (LLMs) and stored for fine-tuning the deep learning components (Figure 2).

User interface. The user interface includes four main components, as seen in Figure 1: the left bar for uploading new images or selecting old ones for captioning, as well as choosing the models for image caption generation and contextualisation; the top central box, showcasing clickable object detections, which the user can then use to enter person names; the middle central box for metadata and temperature selection; and the bottom central boxes for the manual addition of context information, caption rating, and feedback incorporation. The generation of a contextualised image caption occurs in three stages. In the *first stage*, the user uploads an image (users can also re-caption existing images) and selects a model combination for captioning and contextualisation. The image is then processed for (a) object detection and (b) image captioning. In the *second stage*, the user can provide more information for personalisation and contextualisation, as well as feedback: The uploaded image is displayed on the interface, along with detected objects marked with a red bounding box. Users can click on detected persons to initiate annotation. After selecting a detected person, a text input field appears in the designated annotation panel on the right. Users can then enter the name of the person being annotated. Each time a new person is selected for annotation, an additional text input field is dynamically generated within the annotation panel. This allows multiple persons to be annotated simultaneously. The base caption generated by the image captioning component is displayed below. The user can edit and save the improved version if the initial caption contains errors. The detected metadata, namely date, time, and location, are shown in the central component. Users can adjust the generation temperature on the right part before generating the personalised and contextualised caption. Additionally, they can provide additional information relevant to the captioning process. During the *third stage*, personalised and contextualised image captioning occurs, based on person names, base captions, metadata, and further context. The initial generated caption is displayed in the left section of the bottom central component. Users can rate the quality of the generated caption on a scale from 1 to 10 and propose improvements, which are incorporated into the updated caption shown in the bottom-right section of the interface.

Implementation. Our presented tool employs multiple deep learning components to generate personalised and contextualised image captions. The two main components are an image captioning system, which extracts visual information from the input image in the form of a *base caption*, and an LLM, which leverages contextual information to transform the base caption into a *personalised/contextualised caption*. We follow the two-step contextualised caption generation procedure proposed by [1], with additional components to extract and elicit relevant information not present in the image. While this two-stage approach can, in theory, be substituted by using visual/multimodal LLMs, we argue that it provides increased controllability and interpretability and lower inference costs.

Initially, the input image is processed by both the object detection component and the image captioning one. For object detection, we

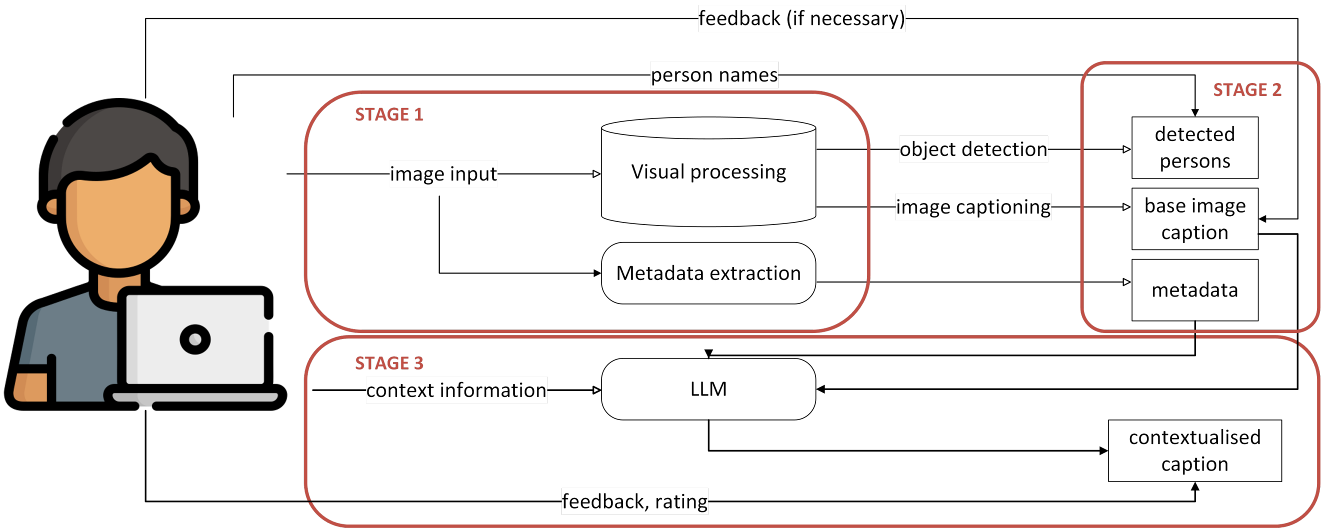

Figure 2: Overview of the architecture of our interactive image captioning system.

utilise a Faster R-CNN model[1] [10] provided by Torchvision. For image captioning, the user can select between two pre-trained models: `BLIP-2`[2] and `ViT-GPT2`[3], both provided by Huggingface. Furthermore, if the image file contains metadata, this is extracted using the `EXIF` library in Python3. To convert the information for latitude and longitude into an exact location, `Geopy` is additionally used. After the user inputs information about the people present, the correctness of the base caption, the necessity of metadata in the caption, and the temperature for generating the caption, the user feedback is used as input into the LLM chosen by the user. The user can choose between GPT-4o, provided by the OpenAI API, and llama3 [6], provided by Ollama[4]. An initial caption is generated, conditioned on the image description from the image captioning component, people's names, and additional information inferred from the image metadata or manually entered by the user. The user can rate the quality of the caption and suggest improvements or changes. The first version of the caption is passed to the LLM, along with the proposed changes, and an updated caption is generated. In parallel, user input and corrective feedback are stored in the backend. In the future, this information can be used to fine-tune the deep learning components individually.

To improve scalability and performance, the system parallelises computations using a `ThreadPoolExecutor`. It reduces redundant tasks with `Flask-Caching` backed by an in-memory cache, ensuring faster response times for multiple simultaneous image processing requests.

4 Conclusion

We designed and implemented a tool for AI caption co-creation that seamlessly integrates deep learning components with human input in an intuitive interface. The tool provides captions based on deep learning detections, which can be updated based on the user's feedback. By reducing the time and effort required for manual annotation, we aim to make the creation process more efficient and effective. We plan to conduct a user study to investigate the efficiency and effectiveness of our approach.

[1] https://pytorch.org/vision/main/models/generated/torchvision.models.detection.fasterrcnn_resnet50_fpn_v2.html
[2] https://huggingface.co/Salesforce/blip2-opt-2.7b
[3] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning
[4] https://ollama.com/

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (BMBF) under grant numbers 01IW23002 (No-IDLE) and 01IW24006 (NoIDLEChatGPT), as well as by the Endowed Chair of Applied AI at the University of Oldenburg.

References

[1] Aliki Anagnostopoulou, Thiago Gouvea, and Daniel Sonntag. 2024. Enhancing Journalism with AI: A Study of Contextualized Image Captioning for News Articles using LLMs and LMMs. arXiv:2408.04331 [cs.CL] https://arxiv.org/abs/2408.04331

[2] Aliki Anagnostopoulou, Mareike Hartmann, and Daniel Sonntag. 2023. Towards Adaptable and Interactive Image Captioning with Data Augmentation and Episodic Memory. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing, SustaiNLP 2023, Toronto, Canada (Hybrid), July 13, 2023*, Nafise Sadat Moosavi, Iryna Gurevych, Yufang Hou, Gyuwan Kim, Young Jin Kim, Tal Schuster, and Ameeta Agrawal (Eds.). Association for Computational Linguistics, 245–256. doi:10.18653/V1/2023.SUSTAINLP-1.19

[3] Rajarshi Biswas, Michael Barz, and Daniel Sonntag. 2020. Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. *Künstliche Intell.* 34, 4 (2020), 571–584. doi:10.1007/S13218-020-00679-2

[4] Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context Driven Entity-Aware Captioning for News Images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 12466–12475. doi:10.1109/CVPR.2019.01275

[5] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Making Large Multimodal Models Understand Arbitrary Visual Prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien

Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024). arXiv:2407.21783 doi:10.48550/ARXIV.2407.21783

[7] Zhengxiong Jia and Xirong Li. 2020. iCap: Interactive Image Captioning with Predictive Text. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (Dublin, Ireland) *(ICMR '20)*. Association for Computing Machinery, New York, NY, USA, 428–435. doi:10.1145/3372278.3390697

[8] Khanh Nguyen, Ali Furkan Biten, Andrés Mafla, Lluís Gómez, and Dimosthenis Karatzas. 2023. Show, Interpret and Tell: Entity-Aware Contextualised Image Captioning in Wikipedia. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 1940–1948. doi:10.1609/AAAI.V37I2.25285

[9] Álvaro Peris and Francisco Casacuberta. 2019. A Neural, Interactive-predictive System for Multimodal Sequence to Sequence Tasks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Marta R. Costa-jussà and Enrique Alfonseca (Eds.). Association for Computational Linguistics, Florence, Italy, 81–86. doi:10.18653/v1/P19-3014

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149. doi:10.1109/TPAMI.2016.2577031

[11] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2443–2449. doi:10.1145/3404835.3463257

[12] Alasdair Tran, Alexander Patrick Mathews, and Lexing Xie. 2020. Transform and Tell: Entity-Aware News Image Captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 13032–13042. doi:10.1109/CVPR42600.2020.01305

[13] Yiyu Wang, Hao Luo, Jungang Xu, Yingfei Sun, and Fan Wang. 2024. Text Data-Centric Image Captioning with Interactive Prompts. *CoRR* abs/2403.19193 (2024). arXiv:2403.19193 doi:10.48550/ARXIV.2403.19193

[14] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 2048–2057. https://proceedings.mlr.press/v37/xuc15.html