Self-improving Scene Understanding with Vision-Language Knowledge Integration [Extended Abstract]

Aliki Anagnostopoulou[1,2], Hasan Md Tusfiqur Alam[1] and Daniel Sonntag[1,2]

[1]German Research Center for Artificial Intelligence (DFKI), 26129 Oldenburg, Germany
[2]Carl von Ossietzky University of Oldenburg, 26129 Oldenburg, Germany

Abstract

We propose an approach for personalised and contextualised image captioning. As pre-trained vision-language systems fail to capture details about the user's intent, occasion, and other information related to the image, we envision a system that addresses these limitations. This approach has two key components for which we need to find suitable practical implementations: multimodal RAG and automatic prompt engineering. We outline our idea and review different possibilities to address these tasks.

Keywords

image captioning, personalisation, multimodal RAG, large foundation models

1. Introduction

Image captioning has seen immense progress in the last few years. However, general-purpose systems often fail to provide personalised, context-aware captions tailored to individual users. In this work, we investigate the task of *personalised* and *contextualised* image captioning by leveraging foundational models, including large language models (LLMs) and large multimodal models (LMMs). Our proposed framework integrates a *multimodal retrieval-augmented generation (mRAG)* system and an *automatic prompt engineering* component to incorporate user feedback and enhance personalisation. The system follows a two-stage pipeline: (1) base image captioning for vision understanding, and (2) retrieval-enhanced contextualisation, where relevant multimodal knowledge—spanning user history, domain-specific corpora (e.g., journalism archives, Wikipedia), and real-time feedback —is dynamically retrieved and integrated into the caption generation process.

We aim to answer the following research questions: **RQ1:** How do multimodal retrieval systems and stage-wise context integration impact contextualised image captions' accuracy, relevance, and personalisation? **RQ2:** How does fine-tuning foundation models or employing automated prompt engineering enhance the contextualisation and personalisation of image captions, and how do these two approaches compare in terms of performance and adaptability?

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy
aliki.anagnostopoulou@dfki.de (A. Anagnostopoulou); hasan.alam@dfki.de (H. M. T. Alam); daniel.sonntag@dfki.de (D. Sonntag)

RQ3: How effectively do standardised metrics capture the quality and relevance of contextualised and personalised captions compared to user feedback in iterative evaluation processes?

2. Approach

Our interactive learning approach for image captioning aims to improve a model based on user feedback, focusing on feedback and context integration rather than developing a new model architecture for image captioning.

Baseline: off-the-shelf foundation models We follow the two-stage method described in Anagnostopoulou et al. [1] as our baseline, which leverages a conventional image captioning component for the extraction of visual information from the image in the form of a *base caption*, and an LLM, which can generate a contextualised caption using the base caption and the context information provided. We plan to use BLIP-2 [2], OFA [3], and mPLUG [4] as image captioning models and a combination contextualising LLMs, both proprietary, such as GPT-4o [5, 6], and open-source, such as llama3 [7]. To alleviate the information bottleneck in our two-stage pipeline, we employ **dense captioning**, generating captions for detected objects using the `fasterrcnn_resnet50_fpn_v2` detection model from PyTorch [8]. We additionally plan to fine-tune llama3 for the tasks of *context extraction* and *contextualised caption generation*.

Multimodal RAG A multimodal retrieval augmented generation (mRAG) [9, 10, 11] system searches for relevant data. The goal is to extract additional context or information that enhances the caption by making it more personalized or contextually accurate. In the case of *user-specific data*, past captions, user preferences, and feedback to the caption generation process are considered. We leverage multi-agentic [12] approaches to employ task-specific REasoning and ACTing (REACT) [13] agents to enhance information retrieval. In the case of *domain-specific context*, relevant knowledge from sources such as journalism archives or Wikipedia articles can be retrieved.

Automatic prompt engineering We are additionally implementing a version of automatic prompt engineering based on APE [14] to enhance the performance of the LLMs, including (1) instruction induction given initial input-desired output pairs, (2) instruction summarization and paraphrasing for the diversification of prompts, (3) instruction refinement, and (4) storing of high-scoring candidates for future use by the specific user. We compare the performance improvement of our automatic prompt engineering method to that of fine-tuned models.

Evaluation To evaluate the contribution of each component separately, we add components for each experimental setting. (1) Baseline: off-the-shelf models; (2) CIC-only: Addition of dense captioning; (3a) Addition of mRAG; and (3b) Addition of automatic prompt engineering. We plan to evaluate our pipeline on contextualised image captioning datasets such as GoodNews [15] and WiT [16], with automated metrics such as BLEU [17], ROUGE [18], METEOR [19], and BertSCORE [20]. In addition, we plan a comprehensive user study assessing contextual relevance, personalisation accuracy, and iterative improvements.

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] A. Anagnostopoulou, T. Gouvea, D. Sonntag, Enhancing journalism with ai: A study of contextualized image captioning for news articles using llms and lmms, 2024.

[2] J. Li, D. Li, S. Savarese, S. C. H. Hoi, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 19730–19742. URL: https://proceedings.mlr.press/v202/li23q.html.

[3] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 23318–23340. URL: https://proceedings.mlr.press/v162/wang22al.html.

[4] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, J. Zhang, S. Huang, F. Huang, J. Zhou, L. Si, mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 7241–7259. URL: https://aclanthology.org/2022.emnlp-main.488/. doi:10.18653/v1/2022.emnlp-main.488.

[5] OpenAI, GPT-4 Technical Report, 2023. _eprint: 2303.08774.

[6] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Madry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. L. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, Gpt-4o

system card, CoRR abs/2410.21276 (2024). URL: https://doi.org/10.48550/arXiv.2410.21276. doi:10.48550/ARXIV.2410.21276. arXiv:2410.21276.

[7] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Rozière, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. M. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, et al., The llama 3 herd of models, CoRR abs/2407.21783 (2024). URL: https://doi.org/10.48550/arXiv.2407.21783. doi:10.48550/ARXIV.2407.21783. arXiv:2407.21783.

[8] Y. Li, S. Xie, X. Chen, P. Dollár, K. He, R. B. Girshick, Benchmarking detection transfer learning with vision transformers, CoRR abs/2111.11429 (2021). URL: https://arxiv.org/abs/2111.11429. arXiv:2111.11429.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[10] H. M. T. Alam, D. Srivastav, M. A. Kadir, D. Sonntag, Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic rag, 2024. URL: https://arxiv.org/abs/2412.16086. arXiv:2412.16086.

[11] W. Chen, H. Hu, X. Chen, P. Verga, W. W. Cohen, Murag: Multimodal retrieval-augmented generator for open question answering over images and text, arXiv preprint arXiv:2210.02928 (2022).

[12] J. Li, Q. Zhang, Y. Yu, Q. Fu, D. Ye, More agents is all you need, arXiv preprint arXiv:2402.05120 (2024).

[13] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, arXiv preprint arXiv:2210.03629 (2022).

[14] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, Large language models are human-level prompt engineers, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/forum?id=92gvk82DE-.

[15] A. F. Biten, L. Gómez, M. Rusiñol, D. Karatzas, 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, 2019, pp. 12466–12475. URL: http://openaccess.thecvf.com/content_CVPR\protect\discretionary{\char\hyphenchar\font}{}{}2019/html/Biten_Good_News_Everyone_Context_Driven_Entity-Aware_Captioning_for_News_Images_CVPR\protect\discretionary{\char\hyphenchar\font}{}{}2019_paper.html. doi:10.1109/CVPR.2019.01275.

[16] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, M. Najork, WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2443–2449. URL: https://doi.org/10.1145/3404835.3463257. doi:10.1145/3404835.3463257.

[17] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th Annual Meeting of the ACL, ACL, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[18] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[19] A. Lavie, A. Agarwal, METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments, in: C. Callison-Burch, P. Koehn, C. S. Fordyce, C. Monz (Eds.), Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 228–231. URL: https://aclanthology.org/W07-0734/.

[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.