Supervisor's Review of Doctoral Thesis

Title of the thesis: Multimodal Summarization

Author of the thesis: Mateusz Krubiński, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

Author of the review: Pavel Pecina, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

As the title says, Mateusz's thesis focuses on multimodal summarization, the task where multimodal data in the input are summarized into an output of potentially multiple modalities. The work is highly up-to-date. Multimodality (as a broader topic) has attracted a lot of attention during the last few years in the research community with the main goal to come up with a model capable of processing multiple modalities for various tasks, summarization being one of them.

The thesis is written in English, covering 148 pages in total. It is structured into 5 numbered chapters plus an introduction, conclusion, and two appendices. The text is readable, logically structured, easy to follow and understand. Numerous figures and examples help understand the text and get a complete and clear picture. Citations are used properly, the list of references is very rich.

The main contributions of the thesis are numerous, including methods, datasets, tools and experimental findings: 1) the unified formulation of the multimodal summarization tasks with a common, encoder-decoder architecture trainable in a multi-task fashion; 2) the MLASK dataset for video-based multimodal summarization; 3) the experiments with task-specific pre-training and the study of how visual input effects the quality of textual output; 4) the framework for collecting human annotations of the quality and relevance of pictorial summaries; 5) the COMES metric for evaluating the quality of textual summaries; 6) the metric for evaluation of Machine Translation output based on questions generation and question answering. All those are very well described in the thesis, including several series of experiments and their analysis. I also appreciate the user studies and manual analysis of the results of the experiments.

The work described in the thesis is supported by five author's publications (all peer-reviewed), two of them published in Findings of the EACL conference (2023 and 2024), a major conference in the field of Computational Linguistics. Two papers were published in Proceedings of the Conference on Machine Translation, a major venue in

the field of Machine Translation. The remaining paper was published at a workshop collocated with AACL. In addition, the author published five other papers based on his work on other projects and within an internship. During the four years of Mateusz's work, his papers have already collected 36 citations (according to Google Scholar). Mateusz also contributed to the NLP research by publishing several datasets: the MLASK dataset for multimodal summarization and three datasets for machine translation of dialectal Arabic.

Mateusz participated in several (international) research projects, mainly CELL (Contextual Machine Learning of Language Translations, funded by CELSA), and WELCOME (Multiple Intelligent Conversation Agent Services for Reception, Management and Integration of Third Country Nationals, funded by EU). He also did a research internship with Amazon working on Computer vision tasks.

Conclusion

I conclude that the work of Mateusz Krubiński is a significant and innovative contribution to the state of the art in the area of Multimodal Summarization. I fully recommend the thesis to be defended.

Pavel Pecina in Prague, September 19, 2024