Thesis Report

Mateusz Krubiński: Multimodal Summarization

The thesis entitled "*Multimodal Summarization*" authored by Mateusz Krubiński summarizes the author's research in the area of automatically summarizing multimodal texts. Although the document summarization has recently attracted somewhat less attention in the Natural Language Processing (NLP) community, there are still significant unresolved challenging problems in this research field. The thesis approaches a key challenge in the summarization domain that centers around summarizing multimodal documents, thus documents that contain more than one modality.

The author focuses on the multimodal summarization settings in which textual content is present in both the input and the output, and the textual document is summarized together with a set of images or a short video into the final summary composed of text and a single image. It is important to mention here the ubiquity of documents containing information in multiple modalities in the current news landscape, especially in news content circulating in social media, and the challenge that the different modalities pose for processing and summarizing multi-modal documents. The report below evaluates the structure and the scientific merit of the thesis.

The author makes the following contributions to this thesis:

1. Development of a large-scale dataset involving videos for the multimodal summarization task.
2. Conducting experiments to learn about the role of pretraining and the influence of visual modality on the generated summary.
3. Proposal of a human-based evaluation framework and novel quality metric for text.
4. Proposal of a new, unified multi-tasking approach for multimodal summarization.

The thesis is composed of 5 chapters, 3 of which are core chapters describing the author's key achievements with their content having been also published in the peer-reviewed research outlets of the natural language processing, and machine translation communities. In total, the research described in this thesis has appeared in 5 conference and workshop publications. At least two publication venues of these publications are of the highest standards, where the author published full papers at EACL - a Core A venue of Natural Language Processing. No journal publication has appeared yet, although this would be encouraged given the scope of the work done. In all the main 5 publications Mateusz is the first author. He has also published, as the first author 3 other papers during his PhD studies on related topics in addition to the above-discussed papers. Altogether, it is quite clear that a substantial novel contribution to the state of the art and, in particular, to our understanding of the field of multimodal document summarization, has been demonstrated through the author's research.

Regarding the structure and exposition, the thesis is generally well-written and well-structured, making it easy for a reader to follow its content. Each core chapter is preceded by an introduction connecting it with the prior chapters and summarizing the content. However, unlike typical theses, this one mixes the description of the author's proposal with the discussion of prior literature. A small remark about the thesis title is that it appears to be too generic. There were already at least several tens of papers focusing on multimodal summarization including also a recent survey while the title in its current form suggests the task to be novel and fails to specify the particular type of approaches taken or the type of modalities approached.

In **Chapter 1**, the author provides an introduction and lays the foundations for this research. The overview of the multimodal summarization field is done here on an abstract level including also extractive and abstractive summarization approaches, and the task is defined as well as vision language models are elaborated.

Chapter 2 discusses different methods for summarization of multiple modalities and further introduces the taxonomy of different approaches, especially the distinction based on the input and output modalities. The author's contributions are also briefly mentioned in this chapter. The chapter could be actually merged with Chapter 1 as being mainly the background and introduction providing section.

Chapter 3 overviews different quality evaluation metrics for the generated summaries ranging from human-based to embedding and training-based ones. Section 3.1.7 discusses the authors' solution to the evaluation called COMES which is a variant of COMET – a metric pre-trained on annotated machine translation data. This solution is rather simple and non-transparent as it relies on a neural network approach but it achieves a relatively high correlation with expert annotations. It would be beneficial to compare the performance of COMES against LLM-based approaches, especially with fine-tuned LLMs. A strong point of COMES is that it can be applied to different languages. The author has also evaluated it on German, Russian, and other languages, albeit this was done only on the translations of the SummEval data. The weakness of COMES and also other evaluation approaches is that they either focus only on text output or treat all the multimodal outputs independently, while these are clearly used by the users together. Mateusz proposes also in this chapter a cross-modal evaluation metric suited for single-image output summaries.

In **Chapter 4**, the author first explores datasets available for the multimodal summarization task. Then he proceeds to describe a large dataset created by the author based on news articles in Czech language crawled from newswire sources. The novelty of this dataset lies in its format of video modality and text forming the input, and an image as well as the title considered as an output. The problematic point of the dataset is that it is in Czech language rather than in English hence its use in the community can be rather limited, although it is clearly still important and useful to create data in less studied languages. It would make sense thus to translate the data into English. The remaining part of this chapter is devoted to describing an extension of M3LS dataset which is an English language dataset. The extension is quite simple as it relies on searching in HTML code for images annotated as key ones. Still, this preprocessing results in a quite useful resource.

Chapter 5 discusses the approach of the author for the multimodal summarization task. The method is quite technically advanced although it is unclear to what extent it is novel as the novelty seems to not be explicitly mentioned. Nevertheless, the author employs this approach along with several baselines to study the effect of visual modality on the final summary. Several interesting conclusions are made thanks to this study such as the one that the better the model is at text summarization the less effective is the use of visual clues. Next, in this section, the author discusses a unified approach for multimodal summarization that rests on concatenating vectors of different modalities and using positional embeddings for linear modalities such as video. The objective of a unified multimodal summarizer is very applaudable and the idea is quite useful, albeit the results are not that impressive. The author has rightfully noticed the lack of such unified solutions. The approach proposed is a simple encoder-decoder model but with a clever solution for selecting output images. It is however unclear how more complex settings such as the case of video as an output would work with the proposed method. Nevertheless, the authors compare their method with several baselines including also LLM-based ones.

Finally, the author concludes the thesis in the next chapter and discusses limitations as well as future work. Altogether, the **conclusions** section, albeit very important, is however a bit thin especially when it comes to the comparisons of the author's solutions with the prior literature and the future work.

To sum up, in general, the structural organization and the language of the thesis are satisfactory. Chapters 3-5 being the core part of the thesis are essentially self-contained. The objective of this work is very clear and sound, while the research undertaken is on a quite advanced level. I think that the overall quality of this work is high, the proposed approaches are novel, while the technical contributions are on a satisfactory level and the designed methods are effective as validated through experiments involving baselines, diverse datasets, and sensitivity-oriented analyses. The author demonstrated a high level of understanding of the field of multimodal summarization. Given all these factors I believe that the contributions put forward by this thesis are more than sufficient to warrant its approval, and I strongly recommend thus its acceptance.

August 25, 2024

Adam Jatowt

Adam Jatowt
Professor of University of Innsbruck, Austria
Deputy Head of the Digital Science Center
Head of the Data Science Group